FACULTY OF SCIENCES

FORECASTING TIDAL SURGE IN THE LOWER SEA SCHELDT USING MACHINE LEARNING TECHNIQUES

Bob De Clercq
Student number: 19920280

Supervisor: Prof. Dr. ir. Willem Waegeman
Co-supervisor: Dr. ir. Jiri Nossent
Tutor: Dr. Christina Papagiannopoulou

A dissertation submitted to Ghent University in partial fulfilment of the requirements for the degree of Master of Science in Statistical Data Analysis

Academic year: 2018 - 2019

GHENT UNIVERSITY

Ghent, June 21, 2019

The promotor,	The author,
Prof. Dr. ir. Willem Waegeman	Bob De Clercq

FOREWORD

This thesis describes the conducted research on tidal surge forecast modelling on the Lower Sea Scheldt, commissioned by Flanders Hydraulic Research. The latter is a division of the department of Mobility and Public Works of the Government of Flanders. They kindly provided the necessary data for this task.

I would like to thank Prof. Willem Waegeman of the Department of Data Analysis and Mathematical Modelling at the Faculty of Bioscience Engineering for the opportunity to submerge in the fascinating world of machine learning and artificial neural networks. Big thanks to Willem and Christina for their advice and help. My gratitude also goes to Jiri Nossent of Flanders Hydraulic Research for providing the data, the follow-up and his guidance w.r.t. the practical application of the forecast model.

Finally, I would like to dedicate this thesis to my dearest Ellen, Rune and Mara for their support and patience during the exciting journey to obtain the degree of Master in Statistical Data Analysis.

CONTENTS

ABSTRACT

Throughout history, the Scheldt estuary had a large economic value and drove the industry around Antwerp and the hinterland. Besides economic benefits, it has a large impact on nature too. Water levels are not only driven by upstream river discharges, but mainly by the tidal influences because of the Scheldt's connection to the North Sea. Spring tide accompanied with strong north-western winds may lead to extreme high water levels and thus potentially endangering Antwerp city with flooding. Flood protection walls along the quays therefore need to keep the city dry. Hence, in-time closure of these protection walls requires good storm surge forecasts. Ideally, alert messages are sent out 24 hours in advance to the responsible people in case of flood risk, such that the necessary mitigating actions can be taken; 6 hours in advance alerts are considered as the minimum time frame.

Flanders Hydraulic Research, a division of the department of Mobility and Public Works of the Government of Flanders, delivers forecast modelling tools for many purposes. Scheldt water level forecasts use nowadays physically-based hydrodynamic models with wind effects, and astronomical tides are applied as boundary conditions. These variables and others, like atmospheric pressure, river discharges, etc. are known to result in discrepancies between the astronomical and observed water levels. Note that not only the tidal amplitude may differ but also the high and low water time of occurrence. This is called skew surge. The underlying causal processes have been reported in literature but not yet all quantified w.r.t. the Lower Sea Scheldt.

This research therefore aimed at improving the water level forecasts near Antwerp by using predictive data-driven models. For this purpose, decades of time series data were available covering water levels, wind speed and direction, atmospheric pressure, air and water temperature, and river discharges into the Scheldt. Not only measurements but also predictions were available. Further, water levels were measured at multiple locations along the estuary reflecting the hydrodynamic properties of the estuary such as confinement, bends and side channels.

Different machine learning models were explored w.r.t. tidal surge forecasting. For this task, a time series spanning 12 years in total, with a 10-minute temporal resolution, was used. The forecast horizon covered both 6 and 24 hours in accordance to the previously mentioned high water alert trigger times. Past observed surge levels and environmental variables were used as predictors. Predicted wind speed and tides additionally fed the models in the forecast horizon where observations were absent. A major improvement in high tide forecast performance was obtained by sample weighting. By putting more weight on the upper 5% highest water levels during model training, the storm surge forecasts majorly improved. Alert events, here defined as high water tides exceeding 6.3 m TAW (TAW=*Tweede Algemene Waterpassing*, the reference height used in Belgium for height measurements), were correctly 6-hours ahead forecasted in 45% of the cases, whereas the astronomical tide (as base forecast model) did not reveal any alert event. Note though that 47% of the true alert events were not identified. Alert events only occurred in 0.5% of all tides.

Any practical use of forecast models implicates knowledge of the model uncertainty. Uncertainty on the water level forecasts was computed by means of the so-called conformal prediction framework. The methodological advantage here is its independence of the machine learning algorithm and underlying multivariate distribution of the data. Proposals have been made to deal with time dependent data, and to generate locally variable prediction intervals. With these interval predictions, the alert identification performance was revisited. No or almost no alert events were missed by the forecasts, but the other side of the coin was that many more events were falsely classified as alert.

Several issues deserve more attention and continued research. Many other model approaches exist that may further improve the surge forecast performance. In this respect, this thesis should be seen as a first attempt towards setting up a surge forecast model for the Lower Sea Scheldt.

CHAPTER 1

RESEARCH OBJECTIVES AND OUTLINE

1.1 Introduction & problem statement

The Scheldt estuary flows through the Netherlands and Belgium and is connected to the Scheldt river, having its spring in the north-west of France. The Scheldt on the Belgian territory is divided in the Upper and Lower Sea Scheldt, where the latter is characterised by tidal influences. Generally speaking, one may say that this boundary is located near the town Schelle, where the Rupel river discharges in the Scheldt.

The Sea Scheldt has a large economic value and drives the industry around Antwerp and the hinterland. Its accessibility largely depends on the tidal nature of the river though. Besides nautical aspects, it also impacts flooding safety in e.g. Antwerp city. Spring tides accompanied with strong north-western winds may lead to extreme high water levels and thus potentially endangering Antwerp city with flooding. Along the quays, flood protection walls need to keep the city dry (Figure 1-1, left). Hence, in-time closure of these protection walls is crucial, for which good storm surge forecasts are necessary. Water levels are forecasted and alert messages are automatically sent to the responsible people for taking mitigating actions, see Table 1-1.

Table 1-1: Storm-tide alert levels in Antwerp

water levels in Antwerp [m TAW]	alert level	description
6.3	pre-alert	nothing really happens
6.6	alert	start of the “storm-tide” procedure (several measures are taken for protection and precaution)
6.7	-	the gates of the parking lot next to the river Scheldt in Antwerp are closed because water can overtop the embankments due to waves
7	dangerous storm-tide	the parking lots next to the river Scheldt will be flooded
7.3	alarm	alarm

Water level in the Scheldt estuary shows a diurnal pattern with long-term frequencies too, such as spring and neap tide with a period of around 14 days. Other oscillations exist such as the 18.6 years period related to the inclination of the moon. These are the result of the movement of the moon around earth and earth around the sun. The resulting gravitational forces lead to the so-called astronomical tides. The latter can thus be predicted, but discrepancies occur due to wind speed and direction, atmospheric pressure and river discharge (especially at low water), see e.g. Rajasekaran *et al.* (2008) and Roberts *et al.* (2015). While the astronomy and meteorology are external drivers for the water level, the water level change along the river stretch is also determined by the river’s constriction. In this respect, Figure 1-1 (right)

demonstrates the large discrepancy of astronomical tides with observed water levels in Antwerp during the storm conditions in December 2013.

The Scheldt river is already extensively studied by Flanders Hydraulic Research, being a department of Mobility and Public Works of the Government of Flanders (Belgium). Water level forecasts use nowadays physically-based hydrodynamic models with wind effects. However, storm surges in the Scheldt estuary are phenomenologically not yet fully understood or predictable. In this context, this thesis tries to demonstrate the performance of machine learning techniques w.r.t. surge forecasting.

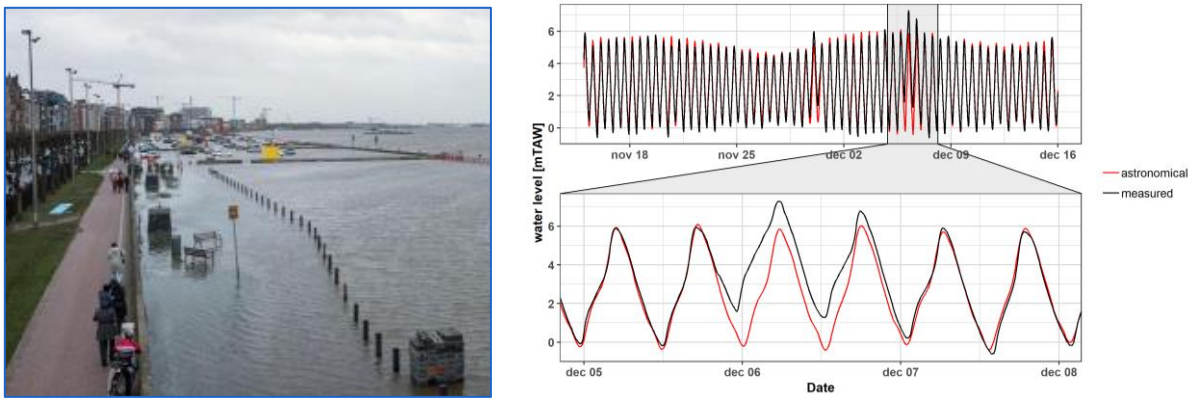


Figure 1-1: Flood protection walls along the Antwerp quays (left) (De Standaard, 2018). During the Santa Claus storm in December 2013, the astronomical tide largely deviates from the true water levels (right)

1.2 Objectives of this research

Storm surge forecasting in the Lower Sea Scheldt river is a rather new research area. The aim of this research is therefore to explore several methodologies w.r.t. their forecasting performance. For this task, an enormous amount of time series data is available to set up mathematical forecast models. This will require some dedicated mathematical tools though. Non-linear system behaviour may further complicate the forecasting task. Because the Scheldt system behaviour is not fully understood, the idea exists to let the data speak for itself by applying machine learning (ML) techniques. Penalization methods allow the selection or weighting of the most important surge driving variables. To assess flooding risks, the uncertainty on the storm surge forecasts need to be known; hence, some attention in this work is devoted to statistical inference of time series forecasts as well. The forecast horizon should ideally cover 24 hours to properly initiate mitigation measures in practice; 6 hours is considered as the lowest acceptable forecasting time.

This thesis will thus explore several ML techniques, accounting as much as possible for all available system information. Both linear and non-linear techniques are applied. Finally, the best predictive model is to be selected, and a first attempt to give a prediction interval is given.

1.3 Outline: the roadmap through this thesis

This thesis will start with a chapter on data and methods. It gives an overview of the available data and how it is pre-processed and cleaned. In addition, we will discuss the surge forecast modelling approaches as found in literature and those being applied. Chapter 3 deals with data exploration giving some insights in the river system dynamics. These insights will also guide the setup of the forecast models as discussed in Chapter 4. In this chapter, both linear and non-linear model approaches are discussed and evaluated against each other w.r.t. forecasting performance. Uncertainty on the surge forecasts is to be discussed in Chapter 5, and some general conclusions (Chapter 6) finalise this work.

CHAPTER 2

DATA & METHODS

2.1 Introduction

This chapter firstly discusses the data made available by Flanders Hydraulics Research for the surge forecast task. Data originates from different sources and therefore requires adequate pre-processing and cleaning. Secondly, a short literature overview is given of possible and frequently used ML techniques for surge forecasting. A selection is made for evaluation in this work. In addition, model evaluation techniques are discussed and a general framework for assessing forecast uncertainty in the field of ML is introduced. Finally, the applied software is described.

2.2 Description of the data

Because the effect of storm surges in Antwerp depends on many factors, it is crucial to consider all these in our mathematical model. In this respect, data provided by Flanders Hydraulics Research consisted of:

- measured water levels at Antwerp and Vlissingen
- wind speed and direction, both measured (Vlakte van de Raan and Hansweert) and forecasted (Hansweert and Terneuzen)
- air and water temperature at Melsele and Prosperpolder respectively
- barometric pressure (Melsele)
- river discharges measured at different locations upstream of Antwerp

The measurement locations in the Scheldt estuary are shown in Figure 2-1. Data originated from different sources (*Hydrological Information Centre* HIC, *Royal Netherlands Meteorological Institute* KNMI and *Flanders Environmental Agency* VMM) and required some cleaning and other preparation prior to their use for modelling purposes. Duplicated data had to be removed, and extreme/unrealistic values were replaced by the neighbour averages. Depending on the variable, different imputation techniques were applied. An overview is given in Table A 1. The provided data showed time intervals between 5 and 60 minutes; conducted interpolations allowed a time resolution of 10 minutes common for all variables in the final data set. Note that the river discharges were lumped into one discharge value for modelling purposes; all rivers discharge upstream of Antwerp, so there was no need to consider them separately.

Note that future water levels need to be predicted, hence observations cannot be used in the forecast horizon. Fortunately, wind and water level forecasts were available, being assumed having a significant impact on the forecast accuracy. Remark that the predicted water levels are based on astronomical forecasts and are thus not related to any surge-related effects.

Data were available for the period January 1998 till August 2018, however many data were missing as shown in Figure A 1. The period was therefore restricted to January 11, 2006 till August 14, 2018. Due to the large data set available, only complete cases were considered for modelling.

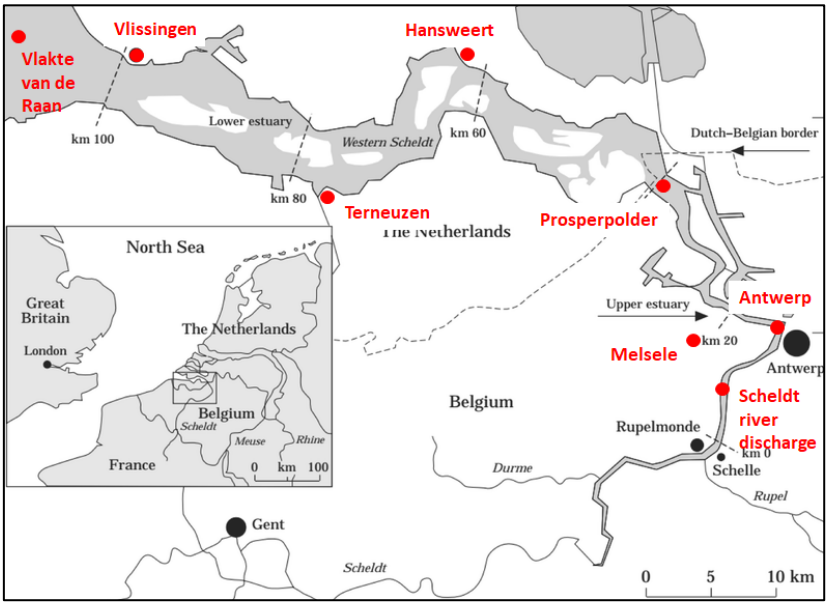


Figure 2-1: Geography of measurement locations

2.3 Methods

2.3.1 Model approaches

Many studies have already been conducted on time series analyses of water levels in the Western Scheldt. Generally speaking, Fourier analyses were conducted and the different tide/astronomical components were identified and studied. As such, the evolution of the components were studied as function of time, i.e. over a period of 3 to 4 decades (Stoorvogel & Habets (2004), Wang & Winterwerp (2013)), to identify the long-term impact of dredging and sand dumping works. These impacts seemed, however, rather negligible on the long-term. Wang and Winterwerp (2013) showed that the tide in the estuary adversely evolved, i.e. the difference between low and high tide increased over time and the location of its maximum moved inland as well. Gerritsen & van den Boogaard (1998) additionally performed a principal component analysis on the 12 most important harmonic components (over a period of 27 years), aiming at discriminating patterns in tidal amplitudes and frequencies between different measurement stations. Finally, Stoorvogel & Habets (2002) conducted a time series analysis on the Western Scheldt water levels by means of autoregressive moving-average (ARMA) models in order to predict 1-hour ahead water levels at some measurement station based on levels at the estuary mouth. The water levels were detrended by subtracting a linear trend and the main Fourier components. There was only 7.3% of unexplained variance left. Even with a relatively simple moving-average (MA) model, a large part of the variability at some measurement station was to be explained by another station. Note that the final autocorrelation

function still showed important correlation peaks at fixed distances from each other. Stoorvogel & Habets (2002) attributed this to the too high low water level forecasts. White noise was thus definitely not obtained and would require further research.

Setting up ARIMAX models (see e.g. Shumway & Stoffer, 2017), i.e. ARIMA models including exogenous predictors X, is not an easy task for estuary tidal data. In economics, where ARIMA models are very common, trends and seasonal effects are easily determined. In the present application, the superposition of many harmonic components in the water level time series severely complicates making the time series stationary. This will be discussed further. Going beyond ARIMA models, literature on surge modelling is very diverse. In general, autoregressive models are applied with lagged feature variables, so temporal autocorrelation is accounted for. When much data is available, the central idea is to include multiple lags and let the ML method decide by supervised learning what features should be included or not in the model. Literature demonstrates that non-linear models are needed to properly predict the surge. Penalized linear regression techniques, such as Lasso, are outcompeted by techniques such as random forests, support vector regression and artificial neural networks. References include e.g. Rajasekaran *et al.* (2008), Nguyen *et al*. (2015) and Mafi & Amirinia (2017).

In this work, the modelling of the tidal surge is considered, i.e. the difference between the observed and astronomical water levels. Note that not only the tidal amplitude differs but also the high and low water time of occurrence. This is called skew surge, see Figure 2-2. When the high water levels are focused upon, one may model both the level and time as target variables. Here, the entire water level time series was modelled, as it was believed to contribute to the forecasting performance. Nevertheless, this work should be seen as a first exploration of forecasting techniques from which future work can be deduced (cf. extreme value analysis on the high water levels).

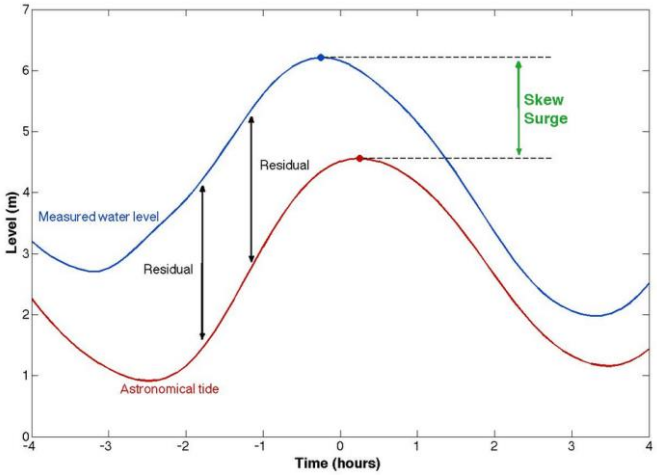


Figure 2-2: Definition of skew surge

Time series analysis generally deals with one-step ahead forecasting. In this study, multi-step forecasting was considered due to the interest in both 24-hours and 6-hours ahead forecasts. A sliding window based time series analysis was applied, i.e. prior time steps were used to predict the next time step. Sometimes, it is also called a lag method. The number of previous time steps is called the window width or size of the lag. The application of this method turns the time series data set into a supervised learning method. In our application, we not only have p-lagged observational data, but also forecasts between the present time t and the q-step

ahead forecast time. This is a modification of the default method and is schematically presented in Figure 2-3.

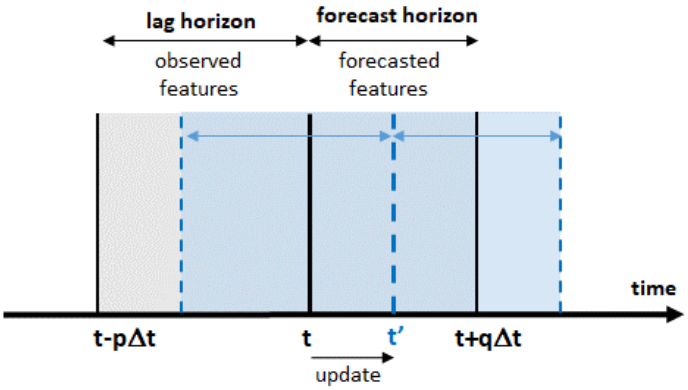


Figure 2-3: Sliding window based time series analysis

The time series model is still of autoregressive nature and can be written in the general form

$$y_{t+q} = f\left(y_t, y_{t-1}, \dots, y_{t-p}, \boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \dots, \boldsymbol{x}_{t-p}, \boldsymbol{x}_{t+q}, \boldsymbol{x}_{t+q-1}, \dots, \boldsymbol{x}_{t+1}\right)$$

where y_t is the surge at time instant t and $\boldsymbol{x}_t$ is a vector of observed or predicted environmental variables at time t depending on the time instant, i.e. whether the time instant is situated in the lag or forecast horizon.

The surge forecasting performance will not only depend on which environmental variables are considered and the lag/forecast horizon extent, but also on the model structure f, i.e. the target function. We evaluated different model specifications, going from linear model approaches such as ordinary and penalized (Lasso, Ridge and elastic net) linear regression, to non-linear techniques as random forests, support vector regression (SVR) and artificial neural networks. In this thesis, only multiple layer perceptrons (MLP) as feedforward artificial neural networks (ANNs) were examined (one and two hidden layers). Note that the computation of the SVR kernel function was very intensive for a large data set as in this work, so this method was abandoned and replaced by the less computationally intensive least-squares SVR (LS-SVR). Although the LS-SVR loses the advantage of support vector sparseness (De Brabanter, 2011), i.e. every observation now becomes a support vector, it had been successfully applied to forecasting river flows (Londhe & Gavraskar, 2015).

2.3.2 Model evaluation & selection

In supervised machine learning, an algorithm learns a model from training data. The goal of any supervised machine learning algorithm is to best estimate the target function f for the output variable y given the input data. The target function is the function that a given supervised machine learning algorithm aims to approximate. The central idea is now to achieve a low bias and a low variance such that a good forecasting performance is obtained. The parameterization of learning algorithms is often a battle to balance out bias and variance, and can be examined or optimized in different ways. As such, forecasting models are trained, validated and tested on independent data. Model hyperparameters (i.e. parameters supplied to the model but which cannot be learnt during training) are tuned on the training data, and

evaluated on the validation data set. Different models are subsequently ranked by confronting them with another independent data set, i.e. the test data, after calibrating the model with the tuned hyperparameters on both training and validation data. The highest forecasting performance then identifies the best model (with optimized hyperparameters). Selecting the best model with respect to forecasting requires some evaluation methodology. Commonly used techniques are the validation (or hold-out) set approach and cross-validation.

The validation set approach is a very simple method and considers a random split of the data in training, validation and test data. Time series data, however, should consider blocked data due to autocorrelation issues. On the other hand, cross-validation (CV) is one of the most widely used methods to assess the generalizability of algorithms in classification and regression. However, also here, care needs to be taken when dealing with time series data. When highly autocorrelated errors occur, standard CV may overfit the data (Opsomer *et al*., 2001). For that reason, different techniques of CV have been developed for such dependent cases. In this respect, Bergmeir & Benitez (2012) performed an empirical study on the impact of different CV approaches; they suggested blocked CV where non-interrupted time series are used as validating subsamples. A number of time instances h between each block are excluded, so they become independent. Bergmeir *et al*. (2018) theoretically showed that for purely autoregressive models, the use of standard K-fold CV is possible provided the models considered have uncorrelated errors. This is quite common when ML methods are used for prediction, and where CV can control for overfitting the data.

Initial training efforts with h-blocked CV were abandoned in this work due to the large computational burden for specific modelling approaches; hence, the validation set approach was applied throughout this work (see Figure 2-4 for principle). This is allowed because many data can be accessed. The following time periods were retained for this purpose (at approx. 60 / 20 / 20% of the entire data set):

- training data: from 2006-01-11 04:00 to 2013-07-01 00:00
- validation data: from 2013-07-01 00:00 to 2016-01-01 00:00
- test data: from 2016-01-01 04:00 to 2018-08-14 11:10

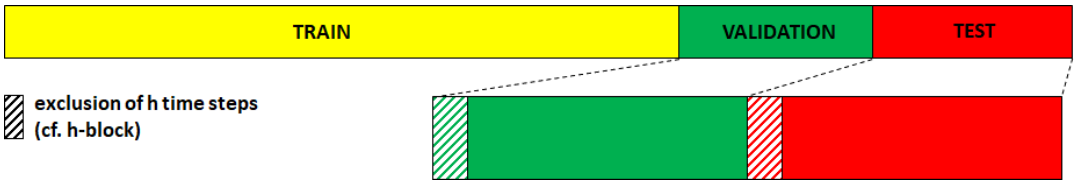


Figure 2-4: Principle of h-block validation set approach

The advantage of these evaluation methods is the direct estimation of the test error. Different performance metrics are used for time series analysis, but the root-mean-squared error ($RMSE$) and the coefficient of determination (R^2) are popular (e.g. Royston *et al.* (2012), Nguyen *et al.* (2015), James *et al.* (2017)). Whereas the first is a frequently used measure for the differences between values predicted by a model $\hat{y}$ and the values observed y, the second measure reflects the proportion of the variance in the target or outcome variable that is predictable from the features or independent variable(s). Both are defined as follows ($\bar{y}$ is the mean of the observed data):

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

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

These performance metrics account for discrepancies between forecasts and observations of the entire tidal cycle. This research primarily aims at predicting storm surges w.r.t. flooding risks, i.e. at high water levels. For this reason, the performance metrics were not only determined for the complete data series, but also for the upper 5-percentile of the surge reflecting the accuracy at which the model is able to predict extreme situations. This peak-over-threshold (POT) water level was based solely on the training data and measured 5.41 m TAW (TAW=*Tweede Algemene Waterpassing*, the reference height used in Belgium for height measurements). Note again that any model training was based on all samples and not on the POT samples.

One of the goals of the forecast model may be the initiation of alert messages w.r.t. flooding risks as mentioned in Table 1-1. Alert messages based on water level forecasts can be correctly (cf. true positives *TP*) but also falsely (cf. false positives *FP*) initiated. Obviously, the same counts when no specific alert needs to be set (cf. true negatives *TN* and false negatives *FN*). This "classification" performance was evaluated on the test data by the following measures:

$$PPV = \frac{TP}{TP + FP} = positive\ predicted\ value = precision$$

$$FDR = 1 - PPV = false\ discovery\ rate = fall - out$$

$$FNR = \frac{FN}{FN + TP} = false\ negative\ rate = miss\ rate$$

Because of the 10-minute frequency of the time series data, many water level observations or forecasts originating from the same tide will attribute to the same and/or different classes. An alert event is related to a specific tide though, so all observations or forecasts of the same tide may only count once in the evaluation to make it fair and objective. Therefore, the evaluation will be based on the high water level of each tide only.

Note that classification problems are generally related with some form of probability modelling, i.e. the probability is estimated at which a sample can be attributed to a specific class. A set probability threshold defines subsequently the attributed class (here: alert level class). For an ordinal classification problem, cumulative logit models and discriminant analysis are some example techniques that can be used. Such classification models could be constructed, but the primary goal of this work is water level, and not risk, modelling. Hence, no separate model had been set up and model results for the continuous target variables were simply projected to the alert level class definition for further analysis. As the interpretation of 6 alert categories might be hard, averaging techniques exist to create one overall metric and thus facilitates interpretation (see e.g. Manning *et al*., 2009). A macro-average will compute the metric

independently for each class and then takes the average (hence treating all classes equally), whereas a micro-average will aggregate the contributions of all classes to compute the average metric. If we are interested in the number of correct predictions, the micro-average is preferred. In this work, the latter will be used to group alert classes.

2.3.3 Conformal inference

In general, prediction intervals are constructed based on some distributional characteristics of the data. As such, normally distributed data is a prerequisite for many inference techniques. When data is time dependent, the correlation between subsequent samples should be considered as well. In this respect, the Gaussian process technique models the variance-covariance structure of the data (Rasmussen & Williams, 2006). Hence, besides the key assumption in Gaussian process modelling that the data can be represented as a sample of a multivariate Gaussian distribution, the reliability of the solution also depends on how well one may select the covariance function. Preliminary testing on the actual data set showed that this technique demanded too many computational resources and was therefore considered inapplicable, at least in this big data setting.

Besides the consideration of time-dependent data, the inference technique should thus be computationally efficient and, preferably, be applicable to any machine learning algorithm. Quite recently, the so-called conformal prediction for inference purposes was introduced. Confidence intervals for, e.g., ARIMA and linear regression are based on multivariate cumulative distribution functions (CDF) for the residuals or error terms. This is only possible when the correct circumstances are satisfied. Conformal prediction brings this idea to the world of machine learning in general. When used correctly, it guarantees to give confidence intervals with a specified probabilistic tolerance. An introduction can be found in Shafer & Vovk (2008) and Linusson (2017); a short summary is given below.

Conformal prediction uses past experience to determine precise levels of confidence in predictions. It can be used with any method of point prediction for classification or regression. Essentially, one constructs a so-called nonconformity measure α, which measures how unusual a sample or observation looks relative to previous samples. This non-conformity function can be anything; in our prediction setting, this function can be defined as, e.g., the error rate. Given a method for making a point prediction $\hat{y}$, the conformal prediction algorithm turns this nonconformity measure into a prediction region, a set Γ^{ε} that contains y with a probability of at least 1-ε.

Suppose we have a data set with n samples $\{\boldsymbol{z}_i\}_{i=1}^{n}$, where $\boldsymbol{z}_i = (\boldsymbol{x}_i, y_i)$. The p-dimensional space of possible features is called the feature space, $\boldsymbol{x}_i \in \boldsymbol{X}$, and the space of possible targets or responses is called the target space, $y_i \in \boldsymbol{Y}$; hence, we may write $\boldsymbol{z}_i \in \boldsymbol{Z} = \boldsymbol{X} \times \boldsymbol{Y}$.
We want to predict m new responses $y_{i=n+1}^{n+m}$ from new features $\boldsymbol{x}_{i=n+1}^{n+m}$. The conformal algorithm was originally developed for an on-line setting, in which one predicts the targets successively, seeing each true target value after one has predicted it and before one predicts the next one. So, the prediction $\hat{y}_{n+1}$ of the observation y_{n+1} may use observed features $\boldsymbol{x}_{n+1}$ and the preceding samples $(\boldsymbol{x}_1, y_1)$, …, $(\boldsymbol{x}_n, y_n)$. One may obviously also apply this principle to a batch setting, such as in this work.

It can be shown that this principle is valid, not only under the strong assumption of independence, but also the weaker assumption that the samples are probabilistically exchangeable. A series of samples n is exchangeable when all $n!$ permutations of this series are equally likely to occur. The conformal algorithm is applicable under exchangeability, no matter what the probability distribution of the samples is and no matter what nonconformity measure is used to construct the conformal prediction region. However, the efficiency of the conformal prediction will depend on the probability distribution and the nonconformity measure (Shafer & Vovk, 2008). In our setting of dependent data, some modifications to the technique have been proposed in literature and will be dealt with later in this section.

To perform conformal prediction, we define the nonconformal function $\alpha: \boldsymbol{Z} \rightarrow \mathbb{R}$ that measures the dissimilarity of a sample $\boldsymbol{z}_i = (\boldsymbol{x}_i, y_i)$ with previously collected samples. The sequence $\{\alpha_i\}_{i \in \mathbb{N}}$ is called the nonconformity measure; the nonconformity score denotes one instance. A common choice of α for regression problems is the absolute error of the model, i.e. $\alpha = |y - \hat{y}|$. Hence, when the nonconformity score is large, the predicted target is considered to be "strange". The nonconformity measure α can be calculated for all training samples. Whether a sample is "too nonconforming" w.r.t. the others, its nonconformity score is evaluated with the observed nonconformity measure distribution. For this purpose, one introduces the p-value for a sample $\boldsymbol{z}_n$ as the probability of nonconformal scores being larger or as large as the considered score. It is computed as the sample fraction

$$p_z = \frac{\#|\{i = 1, \dots, n: \alpha_i(\boldsymbol{z}) \geq \alpha_n(\boldsymbol{z})\}}{n},$$

which varies between $1/n$ and n/n for the largest and smallest nonconformity measures respectively. Consider now the augmented data set where the new sample $\boldsymbol{z}_{n+1} = (\boldsymbol{x}_{n+1}, y_{n+1})$, with y_{n+1} still being unknown, is added to the training sample set. Based on the known nonconformity measure distribution, we can now define a prediction region for $\boldsymbol{z}_{n+1}$ consisting of the set $\Gamma^{\varepsilon}_{n+1} = \{y | p_z > \varepsilon\}$. The algorithm tells to form a prediction region consisting of the $\boldsymbol{z}$ that are not among the fraction ε, i.e. a set of y excluding ε% of the largest possible absolute model errors.
This can be framed as an application of the Neyman-Pearson theory for hypothesis testing and confidence intervals (Lehmann, 1986). In the Neyman-Pearson theory, we test a hypothesis H using a random variable T that is likely to be large only if H is false. Once one observes $T = t$, we compute the probability $p_H = P(T \geq t | H)$ and reject the null hypothesis at level ε if $p_H \leq \varepsilon$. Because this happens under the hypothesis H with a probability of less than ε, one may declare with 100(1-ε)% confidence that the true hypothesis H is among those not rejected. In our setting, H and T can be defined as:

- H reflecting the distribution formed by the training samples, and
- the test statistic T representing the random value of α_{n+1}

Hence, under H, that is, conditional on the training samples, the 100(1-ε)% confidence interval for α_{n+1} is defined as those $\boldsymbol{z}$ for which $p_z > \varepsilon$. For a 95% prediction interval, one chooses a significance level $\varepsilon = 0.05$. The 95-percentile of computed nonconformity measures, denoted as α_s, is thus associated with the boundaries of the prediction interval $\hat{y} \pm \alpha_s$ in a regression setting, cf. $\alpha = |y - \hat{y}|$. Obviously, new targets will be covered by the prediction interval as long as the underlying model remains valid.

Many modifications of the *original conformal* framework exist. Worthwhile to note is the *split conformal* prediction (Lei *et al.*, 2018). In the original conformal prediction method both the point prediction model and the residual CDF for the prediction intervals are performed on the training data. However, in a high-dimensional setting (cf. big data) and with computationally demanding target estimators (e.g. neural networks), the inference might not be efficient. For that reason, an alternative approach has been proposed in literature. The *split conformal* method, also known as inductive conformal inference, separates the model fitting and residual CDF steps using sample splitting. Hence, the data is now split in training and calibration data on which, respectively, the point prediction model is trained and the residual CDF is determined.

From the above, it is clear that the conformal prediction bands are constant for all samples. In some scenarios this may be correct, but in other situations the residual variance may be non-constant and vary with the features $\boldsymbol{x}$. *Locally-weighted conformal* inference is an extension to the standard approach, where the fitted residuals are scaled inversely with the estimated error spread. The mean absolute deviation (MAD) can be used for this purpose in a regression setting (Lei *et al.*, 2018). The definition of the nonconformity measure then becomes

$$\alpha = \frac{|y - \hat{y}|}{\hat{\rho}}$$

where $\hat{\rho}$ denotes an estimate of the conditional MAD of $[y - \hat{y}(x)]|\boldsymbol{x}$. Hence, one not only needs a point prediction model for $\hat{y}$ but also for $\hat{\rho}$. For the latter, k-nearest neighbor (KNN) regression was selected in this work but the model for $\hat{\rho}$ resulted, however, in a very poor performance. For that reason, an alternative approach was applied. Instead of using the entire calibration data set for defining the nonconformity measure distribution, one now selected a subset of n calibration data being representative for the considered test sample, i.e. one selected the n closest samples in the high-dimensional feature space. Hence, for every sample in the test data set, n "similar" samples were searched for in the calibration data set by the KNN algorithm. The prediction intervals are thus tailored on the test feature characteristics. This method will be referred to as *local split conformal* inference.

In this work, time series are considered where the exchangeability condition is violated and the conformal predictors theoretically cannot guarantee the calibrated error rates. To deal with time-dependency, some modifications to the conformal framework were proposed by Balasubramanian *et al.* (2014) under the assumption that a sample only depends on observations within a given time lag or time window W. More theoretical aspects and applications can be found in Dashevskiy & Luo (2008) and Chernozhukov *et al.* (2018). Consider further a time series with samples $\boldsymbol{a}_1$, $\boldsymbol{a}_2$, $\boldsymbol{a}_3$,… from a q-dimensional feature space. The objective is thus to forecast $\boldsymbol{a}_i$ given $\boldsymbol{a}_1$,…, $\boldsymbol{a}_{i-1}$. Two options were presented by Balasubramanian *et al.* (2014), both based on some transformation of the original sample data.

Method 1: one considers W-lagged variables but the data are allowed to overlap. They are thus not independent but the dependency is aimed to be accounted for via the features. The following rule is used (n is the time series length):

$$\forall i, W+1 \leq i \leq n\colon \boldsymbol{z}_{i-W} = (\boldsymbol{x}_{i-W}, y_{i-W}) := \big((\boldsymbol{a}_{i-W}, \dots, \boldsymbol{a}_{i-1}), \boldsymbol{a}_i\big)$$

For example, if $n = 6$ and $W = 2$, the new transformed data will be

$$\{\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3, \boldsymbol{z}_4\} = \{((\boldsymbol{a}_1, \boldsymbol{a}_2), \boldsymbol{a}_3), ((\boldsymbol{a}_2, \boldsymbol{a}_3), \boldsymbol{a}_4), ((\boldsymbol{a}_3, \boldsymbol{a}_4), \boldsymbol{a}_5), ((\boldsymbol{a}_4, \boldsymbol{a}_5), \boldsymbol{a}_6)\}$$

Method 2: one again considers W-lagged variables but the data are now completely separated by filtering the variables with a W window. Now, the rule becomes:

$$\forall i, 0 \leq i \leq \left[\frac{n}{W+1}\right] - 1 \colon \boldsymbol{z}_{i+1} = (\boldsymbol{x}_{i+1}, y_{i+1}) := \left((\boldsymbol{a}_{n-i(W+1)-W}, \dots, \boldsymbol{a}_{n-i(W+1)-1}), \boldsymbol{a}_{n-i(W+1)}\right)$$

The example above would then give

$$\{\boldsymbol{z}_1, \boldsymbol{z}_2\} = \{((\boldsymbol{a}_4, \boldsymbol{a}_5), \boldsymbol{a}_6), ((\boldsymbol{a}_1, \boldsymbol{a}_2), \boldsymbol{a}_3)\}$$

Note that the chronological order of the original data is reversed in the transformation process. This has no influence though, because we deal now with independent data. The performance and applicability of the different conformal prediction approaches, incl. time-dependency considerations, is discussed in Chapter 5.

2.3.4 Applied software

Analyses were performed in R, version 3.4.3 (2017-11-30). Most calculations were performed on a pc with 64 bit operating system and Intel® Core™ i7-7820HQ CPU @ 2.9 GHz. A RAM memory of 16 GB was installed. For large calculations, the computational resources of the Flemish Supercomputer Center (VSC) at Ghent University were accessed.

An overview of the applied R packages is given in Table 2-1.

Table 2-1: Overview of applied R packages for modelling

R package	version	application
elmNNRcpp	1.0.1	extreme learning machines
e1071	1.7-0	support vector regression
glmnet	2.0-16	penalized linear regression (Lasso, Ridge, elastic net)
keras	2.2.4	R interface to Keras (ANN)
liquidSVM	1.2.2	least-squares SVR
ranger	0.10.1	random forest regression
tensorflow	1.4	R interface to TensorFlow
tideharmonics	0.1-1	harmonic analysis of tides
tides	2.1	maxima and minima of quasi periodic time series

CHAPTER 3

DATA EXPLORATION

3.1 Introduction

This chapter will explore the data with respect to our forecast goal. Obviously, the surge level is the central theme seen our forecast effort. The surge, i.e. the difference between true and astronomical water level, has to be known. Because the provided astronomical (predicted) water level time series showed some unexplainable aberrations, we doubted the data quality and the astronomical levels were first determined. Wind direction and speed are known to be strongly associated with surge. A proper treatment and preparation of this wind vector may improve modelling and thus deserves some more attention. Subsequently, the autocorrelation is examined and how the surge is correlated with the other predictors or features. All this will clarify some modelling decisions at a later stage. With respect to any model evaluation based on the alert levels in Table 1-1, their prevalence is examined as last topic. Some conclusions finalize this chapter.

3.2 Harmonic analysis

Flanders Hydraulics Research provided the astronomical water levels, but some inconsistencies existed at year transitions. A tidal harmonic analysis was therefore first performed on the measured water levels (Jan 1, 1998 until Aug 14, 2018), so adequate surge level data could be obtained. As a first stage, the water levels were detrended by linear regression. The mean water level measured 2.60 m TAW with a 95% confidence interval of [2.57, 2.63] m TAW. The annual water level change was not significantly different from zero, i.e. $-2.47\ 10^{-11}$ m/year with a 95% CI of [$-4.95\ 10^{-11}$, $1.01\ 10^{-13}$]). Subsequently, 114 tidal components were estimated in this study and summarized in Table B 1 for completeness. Every component has two parameters, reflecting the cosine and sine constituent. Note that harmonic regression is based on data not only reflecting tidal effects but also wind and atmospheric effects. Hence, the resulting harmonics will not be the true components and errors on the true surges will occur. This may obviously impact the surge modelling performance.

3.3 Wind decomposition

As mentioned, wind is a very important driving force for storm surges. Both observed and predicted data were available and each was provided as paired data of speed and angle. Figure B 1 shows that wind speeds at both Hansweert and Vlakte van de Raan (at sea) may go up till almost 30 m/s in the period 2006-2018. The dominant wind direction at both locations was south-west to west but wind speeds at Vlakte van de Raan were more prominent. This does however not indicate that these directions have the largest impact on the surge generation. Indeed, the Scheldt estuary is oriented (i) west to east between the estuary mouth

and Bath with a large bend near Hansweert, and (ii) north-west to south-east between Bath and Antwerp (see also Figure 2-1). Hence, winds along these directions are expected to have the largest impact, i.e. increased wind shear along a straight estuary trajectory (and also atmospheric pressure) results in a large water set-up.

For modelling purposes, the wind vector was decomposed in x and y Cartesian components. Their axis orientation was chosen such that a maximal cross-correlation with the surge occurred in the period 2006-2018. This is depicted in Figure B 2. From this, the wind vector was most optimally decomposed by selecting a x-axis angle of 289° and 295° for Hansweert and Vlakte van de Raan respectively; the respective (maximal) cross-correlations measured 0.59 and 0.61.

In machine learning and the big data world, all features are typically retained in the forecasting model and the learning algorithm will decide what features are important. However, knowing in advance that features are highly correlated allows one to reduce the high-dimensional feature space and thus the computational effort. As such, only the wind data of Vlakte van de Raan was selected for model building.

3.4 Autocorrelation & cross-correlation

Time series of water levels are highly correlated. A forecast model best performs when data is decorrelated such that correlation does not have to be accounted for. Decorrelation starts with detrending being performed by subtracting the previously determined astronomical water levels from the observed ones. Knowing to what extent the surge data is autocorrelated will elucidate how many time lags one needs to consider for the lagged regressor definition (or the lag horizon). For this purpose, we considered only the model training data set. In this respect, the autocorrelation function (ACF) in Figure 3-1 clearly shows peaks returning every 78 lags (of 10 minutes), or every 13 hours. This corresponds to one high-low tide cycle. Differencing is traditionally applied to remove autocorrelation (Shumway & Stoffer, 2017) but the approach was abandoned after continuing efforts to obtain stationarity. Because of the complex harmonic residual signal, this does not seem surprising though. It also demonstrates that ARIMAX models are not easily applied in this context. In addition, as we will discuss later, our modelling effort will consider 24-hours and 6-hours ahead forecasts where observations lack between present time and forecasting time. Hence, back-calculating stationary differenced data seems rather impossible (in general, time series forecasting predicts the observation at the next time step only).

In addition, the cross-correlation function (CCF) between the surge level and the other observed variables / features was examined for the training data set too. The results are shown in Table 3-1 and Figure B 3. As shown before, wind was correlated strongest with surge (correlation of ~0.6), followed by atmospheric pressure. Whereas all other features were positively correlated with surge, the pressure showed a negative correlation of -0.36. This makes sense, as storms are generally associated with low-pressure situations (with winds as a consequence). Scheldt water levels had a correlation of around 0.22, i.e. surges were more prominent at high water levels. Increased river discharges also resulted in larger surges (correlation of 0.23). Note that most features were leading the surge, i.e. the maximum (absolute) cross-correlation occurred at positive lags in the CCF, so the surge was maximally

correlated with features from the past. Only the water level at Antwerp, the water temperature and the river discharge were lagging, i.e. the maximum cross-correlation occurred at negative lags. This means that the latter three were lagging behind in time and were not drivers for the surge levels. Figure B 3 shows that care needs to be taken with the interpretation of the water temperature, as a second large peak arose at around -600 lags.

The auto- and cross-correlation analysis is important for setting up autoregressive models with leading predictors, as discussed in Chapter 4 (see also Shumway & Stoffer, 2017). The results further indicate that all predictors were maximally correlated with the surge level within approx. 130 10-min lags.

In §3.3, we already discussed the wind location feature selection as a result of nearly identical CCFs of surge with measured wind speeds at Hansweert and Vlakte van de Raan. This is also demonstrated in Figure B 4, depicting the wind CCF between these two locations. Vlakte van de Raan was kept for the wind variable, as it showed the largest cross-correlation (Table 3-1). W.r.t. water level, the same can be mentioned about the locations Antwerp and Vlissingen. Nevertheless, the latter two were retained because we believed that the difference between their cross-correlation functions reflects the hydrodynamic properties of the estuary such as confinement, bends and side channels.

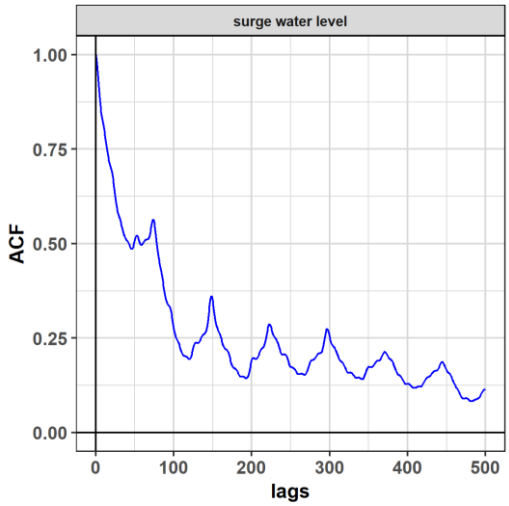


Figure 3-1: Autocorrelation function (ACF) for the surge level in Antwerp (period 2006-2013)

Table 3-1: Maximal cross-correlations between surge level and observed features (period 2006-2013), see also Figure B 3

features	max. cross-correlation	# 10-min lags
water level Antwerp	0.2250	-1
water level Vlissingen	0.2224	10
atmospheric pressure	-0.3628	69
air temperature	0.0939	121
water temperature	0.0562	-4
river discharge	0.2320	-50
X wind Hansweert	0.5928	12
Y wind Hansweert	0.3134	125
X wind Vlakte van de Raan	0.6186	17
Y wind Vlakte van de Raan	0.2939	128

3.5 Prevalence of alert events

The model performance will be evaluated on, among others, the alert messaging of Flanders Hydraulics Research. Ideally, the model should be able to correctly predict the water levels and trigger the corresponding alerts. The forecasting ability of the model largely depends on the supervised learning procedure and, as such, on the training data. If no rare events occur in the training phase, one cannot expect the model to forecast this correctly. Hence, the prevalence of these alert events is important to be examined.

As discussed in §2.3.2, the alert classification will be performed on the tidal high waters only. In this respect, Table 3-2 summarizes the number of events categorized by the alert level cut-off values; in addition, Table B 2 lists the corresponding dates of these extreme high water situations. Clearly, the alert categories represented rare events. Even more, Table B 2 shows that these events were not evenly distributed over training, validation and test periods. This will have an impact on the learning process. As such, one may expect that a good modelling performance for these events will be a hard nut to crack when the modelling focuses on the entire tidal cycle and not only on the high tides.

Table 3-2: Prevalence of alert (extreme) events in the period Jan 11, 2006 until August 14, 2018 (analysis based on high water levels of raw data, so prior to complete cases analysis)

alert levels	high water level range [m TAW]*	# tides
normal	]-∞, 6.3[	17610
pre-alert	[6.3, 6.6[	52
storm tide	[6.6, 6.7[	5
(gate lockdown)	[6.7, 7[	8
dangerous storm tide	[7, 7.3[	3
alarm	[7.3, ∞[	0

* notation: *[a, b[* denotes the range including the lower limit *a*, and excluding the upper limit *b*

3.6 Conclusions

Many measurement data are available for the Western Scheldt estuary that can be used for surge modelling purposes. Raw data covered the period Jan 1, 1998 until Aug 14, 2018. However, the period Jan 11, 2006 until August 14, 2018 was selected for model building and evaluation.

The surge time series was obtained by subtracting the estimated astronomical water levels from the observed water levels. Obviously, the surge was strongly autocorrelated with a repeating pattern every 78 10-minute lags (or 13 hours) in the autocorrelation function. The largest cross-correlations with surge were found for the variable wind. Because the correlations were very similar for Vlakte van de Raan and Hansweert, only the former was selected as feature for further modelling (because it showed slightly larger correlations). Atmospheric pressure was also strongly, but negatively, correlated followed in importance by the river discharge and the water levels. In general, all environmental variables were maximally correlated with the surge level within approx. 130 10-min lags. These observed auto- and cross-correlations will be helpful for defining the lag and forecast horizons in Chapter 4.

With respect to model performance evaluation based on correctly forecasting alert events, it could be observed that the predefined alert classes were characterized by a very low prevalence. This is important to acknowledge when discussing the forecast performance of these rare events in §4.4.

CHAPTER 4

SURGE FORECAST MODELLING

4.1 Introduction

Flanders Hydraulics Research primarily requested a 24-hours ahead forecast model for surge levels. As such, this chapter starts by describing this modelling effort, i.e. an evaluation of the several model approaches. The most optimally performing model structure will subsequently be retrained and applied for 6-hours ahead forecasts to examine any improvements of the forecast performance. Next to the impact of the forecast horizon, a small sensitivity analysis was also performed on the extent of the lag horizon. Some remarks on the final model and performance are given before ending this chapter with some conclusions.

4.2 24-hours ahead surge forecast model

4.2.1 General model structure & input data

The basic model for 24-hours ahead forecasts consists of the sliding window multi-step forecasting method (see also Figure 2-3) where the samples include the target values, i.e. the 24-hours ahead surge levels, and (i) lagged observations included in the lag horizon and (ii) predictions in the forecast horizon.

The following environmental variables were considered as input data in the lag and forecast horizons:

- Lag horizon (consisting of lagged observations)
  - surge level
  - water level in Antwerp
  - water level in Vlissingen
  - wind speed at Vlakte van de Raan, x component
  - wind speed at Vlakte van de Raan, y component
  - atmospheric pressure in Melsele
  - air temperature in Melsele
  - water temperature at Prosperpolder
  - river discharge upstream of Antwerp

- Forecast horizon (consisting of predicted variables)
  - astronomical water levels in Antwerp
  - wind speed in Terneuzen, x component
  - wind speed in Terneuzen, y component

From §3.4, it is clear that including 130 lags of 10 minutes in the model would account for all major cross-correlations between the surge, as target variable, and the features. However, with a 24-hours ahead forecast one has 144 time steps without any data, hence loosing important cross-correlations, what may affect the model performance. As one considers an autoregressive model, lagged surge instants augment the feature data set. Although the surge autocorrelation structure is not significantly included in the input data too, it was decided to include 13h of lagged variables (based on the 78 lags to cover the surge autocorrelation peak, see §3.4) such that a tidal cycle of high water and low water was covered. How sensitive the forecast results are w.r.t. the extent of the lag horizon will be dealt with in §4.3.

Note that several forecast methods demanded considerable computation times. For that reason, it was decided to have forecasts every 20 minutes instead of 10 minutes. In addition, the interval between subsequent time instants in the lag and forecast horizon was set at 30 minutes. This time resolution was considered enough to characterize high water or flood levels.

4.2.2 Linear models

The surge modelling quest was initiated with some linear model approaches: ordinary linear regression and its regularized forms Lasso, Ridge and elastic net. Regularization or penalization techniques are known to generalize better to unseen data than ordinary linear regression if correlated predictors or features occur. Because the number of observations was largely exceeding the feature number, the main advantage of Lasso compared to Ridge regression is its sparse solution, i.e. the regression coefficients are forced to zero for highly correlated features. Note that this feature selection procedure tends to select, rather randomly, one feature among the group of highly correlated features, and thus complicates any model interpretation at a later stage. Elastic-net regression is a combination of Ridge and Lasso regression and returns a sparse solution with potentially more predictors than the rank of the feature matrix; it was included in this comparative study for completeness.

The performance of these linear models on the test data is presented in Table 4-1. More performance details and parameter settings can be found in Table C 1. The table clearly shows that any regularization does not outperform ordinary linear regression; all methods look similar.

Table 4-1: Performance metrics for linear 24-hours ahead forecast models

	training data		test data	
	RMSE	R^2	RMSE	R^2
Ordinary linear regression	0.214	0.450	0.236	0.261
Lasso regression	0.214	0.445	0.235	0.269
Ridge regression	0.214	0.445	0.235	0.269
Elastic-net regression	0.214	0.445	0.235	0.269

From Table C 1 one can conclude that every regularization parameter λ approaches zero and thus these techniques are similar to ordinary linear regression. This convergence of ordinary and penalized regression techniques is further demonstrated by considering 9-fold h-block CV for Lasso regression in Figure 4-1. The penalization parameter clearly asymptotically approaches zero with a negligible change in validation error rate, assessed as MSE. For model selection, it is common practice to select the largest value of λ such that the validation error is

within one standard error of the minimum MSE (James *et al.*, 2017). Evaluation of this feature-reduced model (only 9 features remain) on the test error, however, resulted in a very bad performance (data not shown), indicating a too large model sparseness.

This demonstrates that all features considered in the model seem needed to improve the model performance and that a typical minimum in validation error rate is not obtained, i.e. the point at which overfitting occurs with further decreasing λ values. This can be attributed to the extremely long time series data available with its multitude of environmental variable interactions, resulting in as many unique surge dynamics. In this respect, a small sensitivity analysis on the lag horizon will be performed in §4.4.1 for the final model selected. Note that not only the lags and kinds of model features matter but also the target function f. Table 4-1 shows a low coefficient of determination R^2 of 0.269, which may indeed indicate some non-linear system behavior. Hence, non-linear models may improve the forecasting performance. This is discussed in the next section.

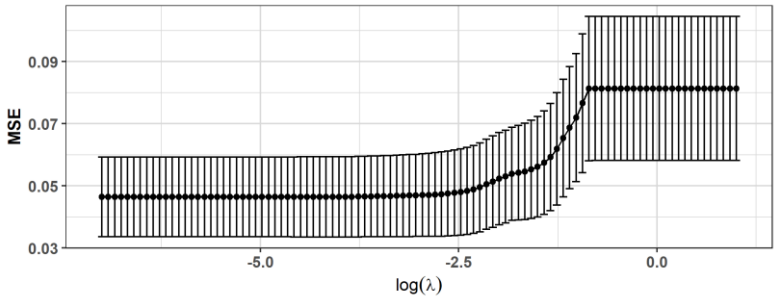


Figure 4-1: Validation error rate as function of the Lasso regularization parameter λ for 9-fold h-block cross-validation

4.2.3 Non-linear models

In order to evaluate whether surge level forecasting can be improved by applying non-linear models, four different model approaches were considered: random forest regression, least-squares support vector regression (LS-SVR), extreme learning machines (ELM) and multiple layer perceptron (MLP) models. The latter two belong to the class of artificial neural networks (ANNs).

With respect to random forest regression, the number of trees was optimized to a value of 300, at which the validation error rate stabilized when further increasing the tree number. The number of variables randomly sampled as candidates at each split was set at 65, minimizing the validation error. Similar to the linear models, the loss function (expressed as validation MSE) was rather flat such that any change in this number of variables resulted in only slight changes of MSE.

With LS-SVR, the feature space is mapped onto an augmented feature space using some fixed (non-linear) mapping, and then a linear model is constructed in this feature space. The mapping is performed by a kernel function. Selecting a particular kernel type and related kernel function parameters is usually based on application-domain knowledge. For surge modelling, one mainly applied in literature the Gaussian or radial basis function kernel (Rajasekaran *et al.* (2008), Londhe & Gavraskar (2015), Elgohary *et al.* (2017)). This kernel was retained in this

work too. The bandwidth of this Gaussian kernel was optimized at a value of 450. The LS-SVR loss function consisted of the residual sum of squares fitting error and a l_2 regularization term. The inverse of this regularization parameter, denoted as C in Table C 1, was determined as 300.

Besides random forest regression and LS-SVR, ANNs were explored as well. Simple MLP models were applied with up to two hidden layers, as they performed well in literature (e.g. Steidley *et al.* (2005), Tsai & You (2014)). Data from the input layer are sent throughout the different layers using the feed-forward method. Defining the network architecture and how to avoid overfitting are major issues when applying ANN in practice. For that reason, a small sensitivity study was performed. From one layer to another, the nodal values are transformed by the activation function, of which the most popular ones had been examined with respect to the out-of-sample error: linear, tanh, relu, elu and sigmoid. In addition, the number of nodes in each layer, the type of optimization algorithm, learning rate and number of necessary epochs were studied. For fully connected layers, nodes may develop co-dependency amongst each other during training, which curbs the individual power of each neuron leading to overfitting of training data. Two ways of dealing with this are dropout and regularization. The former method randomly drops a predetermined percentage of individual nodes out of the network. The aim of this dropout is the reduction of node interdependencies. Activity regularization, where the output of a layer is penalized, has been tested as well. In the end, a single-hidden layer model turned out to perform best on the out-of-sample data. A linear activation function was applied on the input to 1000 layer nodes. The input weights were randomly chosen from a uniform distribution. Further on, a learning rate of 0.01 and 200 epochs were needed to train the MLP. In this respect, similar to the linear model training, Figure 4-2 shows that overfitting does not occur because both training and validation error stabilize. Epochs between 50 and 1000 did not influence the test error (data not shown).
Defining the best MLP network is a tedious task with many degrees of freedom and is an art on itself. For that reason, an ELM model was applied, as it is a single-hidden layer feed-forward neural network with randomly chosen input weights and hidden bias weights. It does not require any setting of learning rate, epochs, etc…, making it very appropriate to check the applicability of a single-hidden layer ANN before exploring the multitude of possibilities of MLP. Actually, only two parameters need to be defined, i.e. the activation function and the number of nodes. A combination of the triangular basis activation function with 389 (i.e. the number of features) nodes performed best.

The hyperparameter settings for random forests and LS-SVR can be found in Table C 1; it also summarizes the examined ELM settings. For the MLP models, please refer to Table C 2.

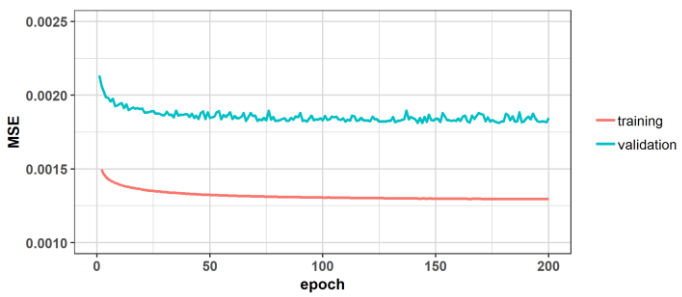


Figure 4-2: Training and validation error rates as function of the epoch number for the selected MLP model

An overview of the performance metrics ($RMSE$ and R^2) of the final selected models is given in Table 4-2. When first looking at the MLP, it is somewhat surprising that this model performed worst. It allows the most degrees of freedom but this may be the largest disadvantage too w.r.t. ANN optimization. Notably, the ELM was very comparable to MLP w.r.t. performance, although the activation function and number of nodes were different. Because of the abundant use of ANN in literature for storm surge modelling, dedicated research here should allow better forecasting performance. Present work should be seen as a first exploration though. In addition, recurrent neural networks are typically used for temporal sequences and deserve more attention. Also a further improvement such as long-short term memory (LSTM) networks are worthwhile to be further explored. One restricted to MLP only because, otherwise, the input data frame had to be reorganised.

An improvement in forecasting performance could be observed when moving to LS-SVR, see Table 4-2. The R^2 increased from 0.284 to 0.326 when comparing LS-SVR with ELM. Nevertheless, all methods were outcompeted by random forest regression. It performed best on both the training and test data. An R^2 of 0.568 was obtained for the test data.

Table 4-2: Performance metrics for non-linear 24h ahead forecast models

	training data		test data	
	RMSE	R²	RMSE	R²
Random forests	0.030	0.989	0.180	0.568
LS-SVR	0.177	0.623	0.225	0.326
ELM	0.218	0.428	0.233	0.284
MLP	0.218	0.428	0.235	0.271

4.2.4 POT weighting for extreme event modelling

From §4.2.2 and §4.2.3 it became clear that random forest regression performed best. Performance evaluation was based on the $RMSE$ and R^2, computed on all observations. One of the practical model applications is flood alert messaging. As such, the modelling effort may put more focus on correctly forecasting the high water levels and giving the low levels, the low tides, less priority. For this purpose, water levels above the 95-percentile water level (i.e. 5.39 m TAW based on the training data set) were weighted more than others in model training. This peaks-over-threshold (POT) approach should allow a better forecast of extreme events.

During construction of the random forest model, bootstrap samples were taken from the training data set in order to create multiple trees. This bootstrap aggregation, or bagging, resulted in an ensemble surge forecast. Here, POT water levels were given more weight or a higher probability to be selected in the bootstrap sample. The results of this further random forest optimization is shown in Table 4-3; note that a weight of 1 refers to the default random forest model, as obtained in §4.2.3. In addition to the default $RMSE$ and R^2, we computed these metrics on the POT water levels only to examine the improved forecasting performance for extreme high water events.

From Table 4-3 it becomes clear that any weighting decreased the training $RMSE\ POT$ to zero and the $R^2 POT$ to one, i.e. a perfect prediction. Obviously, predicted non-POT water levels now show a larger bias as reflected by the increased $RMSE$ and decreased R^2. It becomes

more interesting when one moves to the out-of-sample performance though. The $RMSE$ remained almost constant, whereas the $RMSE\ POT$ decreased with increasing weights; a weight of 30 resulted in the largest performance gain. Note also that the unweighted random forests returned a R^2POT of 0.639, being larger than the R^2 value based on all observations. This would indicate that the model, even unweighted, already performed worse on other stages of the tidal cycle than high water. Also here, a weight of 30 resulted in the largest R^2POT increase to 0.688, while roughly maintaining the R^2 value based on all test data. As already mentioned before, we dealt with an extensive time series data set, hence, improving the forecast of 5% of the samples will not deteriorate the $RMSE$ or R^2 too much. The weighted random forest model was selected as the best compromise between global surge and storm surge forecast performance.

Table 4-3: Performance metrics for weighted random forest regression (24h ahead forecast)

	training data				test data			
weight	RMSE	RMSE POT	R^2	R^2 POT	RMSE	RMSE POT	R^2	R^2 POT
1	0.030	0.037	0.989	0.986	0.180	0.184	0.568	0.639
30	0.070	0.000	0.940	1.000	0.181	0.171	0.567	0.688
60	0.096	0.000	0.888	1.000	0.180	0.172	0.568	0.684
90	0.113	0.000	0.846	1.000	0.181	0.169	0.567	0.693

We may further examine the forecast performance by analysing the model residuals. Figure C 1 clearly shows that the model forecasted surge levels best in the range -0.5 and 0.5 m. Outside this surge level range, absolute surges were underpredicted by the weighted random forest regression. The largest underpredictions occurred for water levels higher than 1.25 m TAW, with a maximum underprediction of ~2.3 m for water levels around 2.5 m TAW. These large discrepancies occurred during the transitional phase between high-low tides, and vice versa, where the water level quickly changed. Obviously, any small temporal shift in tide from skew surges results in large deviations (note that an improperly performed harmonic analysis of the astronomical tide may result in deviations too). At POT water levels, surges were more underpredicted than overpredicted as well.

4.3 6-hours ahead surge forecast model

Flanders Hydraulics Research maintains several forecasting models with different purposes. The shortest forecast horizon measures 6 hours. Hence, a 6-hours ahead tidal surge forecast was considered as the shortest forecast horizon possible. The implications should be beneficial towards forecast performance because now only 6 hours without measurement data occurred.

The best performing 24-hours forecast model, i.e. the random forest regression model, was selected and re-trained on a modified feature data frame. Again, a lag horizon of 13 hours was applied but the forecast horizon now measured 6 hours; time steps of 20 minutes were here considered. The number of trees was kept at a value of 300 and the number of variables randomly sampled as candidates at each split being optimized at 100 by a grid search.

Table 4-4 summarizes the performance metrics for both the unweighted and weighted random forest regression models. In comparison to the 24-hours ahead forecast model, the R^2 and

R^2POT increased to 0.67 and 0.71 respectively for the unweighted case. A weight of 30 further increased the R^2POT to a value of 0.75. The model residuals in Figure C *2* now showed smaller underpredictions of absolute surge levels compared to the 24-hours ahead forecasts. In addition, the POT water levels were better forecasted as well. Similar to the 24-hours ahead forecast model, water levels were extremely underpredicted around 2.5 m TAW with a maximum underprediction of ~2 m. These occurred during tidal transition.

Table 4-4: Performance metrics for (un-)weighted random forest regression (6h ahead forecast)

	training data				test data			
weight	RMSE	RMSE POT	R²	R² POT	RMSE	RMSE POT	R²	R² POT
1	0.036	0.041	0.984	0.983	0.157	0.162	0.670	0.712
30	0.075	0.000	0.933	1.000	0.158	0.152	0.668	0.747

4.4 Some remarks

In this section, some further discussion about the random forest regression model is given. For this purpose, the 6-hours ahead forecast model was selected. Its model performance was expected to depend on the extent of the lag horizon and is to be discussed in §4.4.1. In relation to this, some interpretation to the regression model by the variable or feature importances is given. Finally, §4.4.2 evaluates the model w.r.t. alert classification.

4.4.1 Influence of lag horizon and related features

The auto- and cross-correlation functions in §3.4 all showed a maximum within a lag period of 21.3 hours. These functions declined at a different rate with increasing lags. Hence, a larger lag horizon was expected to have a beneficial effect on the forecast performance.

In this respect, Table 4-5 demonstrates the effect of a 0-hours, 13-hours and 42-hours lag horizon on the performance metrics. In all cases, the number of trees was kept at a value of 300 and the number of variables randomly sampled as candidates at each split measured 25, 100 and 600 with increasing lag horizons (based on a grid search). The table remarkably shows that the lag horizon has hardly any effect on the performance.

Table 4-5: Influence of lag horizon on random forest performance metrics (6h ahead forecast)

		training data				test data			
random forests	lag horizon	RMSE	RMSE POT	R²	R² POT	RMSE	RMSE POT	R²	R² POT
unweighted	0h	0.038	0.042	0.982	0.982	0.162	0.166	0.647	0.699
	13h	0.036	0.041	0.984	0.983	0.157	0.162	0.670	0.712
	42h	0.036	0.040	0.985	0.983	0.161	0.165	0.661	0.712
weighted	0h	0.081	0.011	0.920	0.999	0.162	0.151	0.648	0.752
	13h	0.075	0.000	0.933	1.000	0.158	0.152	0.668	0.747
	42h	0.074	0.000	0.933	1.000	0.161	0.156	0.661	0.745

Why did the lag horizon have such a small influence on the performance metrics? To answer this question, the feature importances of each case were examined. A graphical representation of these feature importances can be found in Figure C 3. It seemed that the actual wind speed at Vlakte van de Raan, measured along the river stretch, mostly determined the tidal surge, independent of the lag horizon extent. For the 13- and 42-hours lag horizons, the observed surge at a 6.3-hours lag (i.e. lag 38) had the second largest influence. Accounting for a forecast horizon of 6 hours, this observation thus lagged 12.3 hours to the forecast time, being very close to the period of the semidiurnal tidal cycle (i.e. one low-high-low tide). When no lag horizon was considered, one noticed that the actual measurements played a crucial role. Besides the strong increase in importance of the measured wind speed at Vlakte van de Raan, also the actual measured surge in Antwerp and the atmospheric pressure at Melsele gained now in importance. Due to the lack of observational information, the predicted wind speed at Terneuzen became more important too. Nevertheless, the measured wind speed outcompeted both the predicted one and measured surge in Antwerp, thus stressing its importance.

4.4.2 Forecast model in practice

One of the forecast model applications is triggering alert messages under flood risk conditions. Hence, the model should be accurate enough to attribute the forecasted water level to the correct alert class, cf. Table 1-1. Based on Table B 2, the alert events were considered as being rare. The model classification performance w.r.t. alert classes with only a single or a few events was a priori deemed to be bad. A more objective and realistic method included the combination of different alert events into a single alert class, next to the class of normal events. These two classes were separated by a threshold water level of 6.3 m TAW, see Table 1-1. As flood risk conditions were considered, only high water levels were examined w.r.t. the alert classification performance. The results are shown in Table 4-6. The performance based on astronomical tides is supplemented for comparison. The latter resulted from the harmonic analysis in §3.2. Note that it does not consider any forecast horizon.

During the period Jan 2016 - August 2018 only 17 alert events occurred, i.e. 0.5% of all observed high water levels. Table 4-6 indicates that the 24-hours ahead forecast model identified 7 alert events. However, 14.3% or 1 event was falsely identified as extreme and 64.7% of the observed ones were missed during identification, i.e. 11 alert events. On the other hand, the 6-hours ahead forecast model identified 20 alert events, being larger than the observed number. Now, 11 were falsely discovered and 8 observed alert events were missed. Hence, the 6-hours ahead forecast model correctly classified three more alert events than the 24-hours ahead model. However, false alert messages were triggered as well though.

The model results were obviously better than the astronomical tidal forecasts where no alert events were identified. Nevertheless, the alert messaging needs further improvement. The forecast model was trained on the tidal surges and not only storm surges, although some improvement towards POT water levels was introduced by weighting the observations. It is thus not very surprising that discrepancies occur at these extreme events, occurring only in 0.5% of all high water levels.

Table 4-6: Alert classification performance on test data (from Jan 1, 2016 until August 14, 2018) for weighted random forest regression and harmonic analysis (cf. astronomical tides)

			high water levels (m TAW)	
forecast horizon	lag horizon		normal]-∞, 6.3[	alert [6.3, ∞[
24	13	# observations	3148	17
		# forecasts	3158	7
		positive predicted value, PPV (precision)	0.997	0.857
		false discovery rate, FDR (fall-out)	0.003	0.143
		false negative rate (miss rate)	0	0.647
6	13	# observations	3137	17
		# forecasts	3134	20
		positive predicted value, PPV (precision)	0.997	0.45
		false discovery rate, FDR (fall-out)	0.003	0.55
		false negative rate (miss rate)	0.004	0.471
astronomical tides		# observations	3137	17
		# forecasts	3134	0
		positive predicted value, PPV (precision)	0.997	0
		false discovery rate, FDR (fall-out)	0.003	1
		false negative rate (miss rate)	0.004	1

4.5 Conclusions

It is believed that including exogenous, environmental variables in water level forecast models may explain discrepancies between the astronomical and true water levels. This chapter therefore made an evaluation of different modelling approaches for tidal surge forecasts. Both linear and nonlinear autoregression techniques were applied of which random forest regression performed best on an independent out-of-sample surge time series. The forecast model focused on the tidal surge and not specifically on the high water levels only, being more important for storm surges. Because one of the applications consists of storm tide forecasting with related alert message triggering, the forecast performance towards high water levels had been improved by weighting extreme observations, being defined as water levels exceeding the 95-percentile (peaks-over-threshold or POT water levels). As such, evaluated performance measures were not only based on the entire tidal cycle but also solely on the POT water levels. The finally selected random forest regression model put 30 times more weight on the POT observations. Further, a lag horizon of 13 hours was retained. Both a 24-hours and 6-hours ahead forecast model was trained.

Whereas the 24-hours ahead forecast model returned a R^2 of 0.57, the 6-hours ahead forecast model obviously gave better results, i.e. R^2 measured 0.67. The R^2POT was determined as 0.69 and 0.75 respectively. In general, the model underpredicted the absolute surge levels. The largest errors on the forecasted surge levels were observed for water levels around 2.5 m TAW and measured up to more than 2 m. These large discrepancies occurred during the transitional phase between high-low tides, and vice versa, where the water level quickly changed. A small temporal shift in tide from skew surges resulted in these large deviations.

Random forests can be used to estimate feature importances, and to rank which features have the largest predictive power. This may be considered as some form of explanatory analysis. Other than that, random forests are very opaque but deliver good predictions. An examination of the feature importances revealed the actual wind speed at Vlakte van de Raan, measured along the Lower Sea Scheldt river stretch, as the primary predictor in the 6-hours forecast model. The observed surge in Antwerp at a 6.3-hours lag was secondly ranked.

The presented forecast model makes a point prediction of the water level in time. However, any interpretation of this point prediction requires an assessment of the related model uncertainty. This is less trivial for nonparametric techniques for time series analysis and will be dealt with in Chapter 5.

CHAPTER 5

CONFORMAL PREDICTION

Machine learning techniques can forecast the water level in Antwerp, as demonstrated in Chapter 4. However, in general, they do not provide any information on how close their forecasts are to the real water levels. Knowledge of the uncertainty on the results is of prior importance w.r.t. the model's practical use. E.g. a (point) water level forecast can be underpredicted and thus not triggering any alert message. However, when a 95% prediction interval is to be determined, an alert can still be initiated when the alert level is covered by the interval estimate.

Conformal prediction or inference provides a general framework, as explained in §2.3.3. The advantages are manifold, because one does not make any prior assumptions about the multivariate distribution of the data and the forecasting tool. The conformal prediction framework and some extensions will be demonstrated on the present case of surge forecasting. The basic assumption is exchangeable data, which is definitely not valid in case of time series analysis though. Again, §2.3.3 discussed some options based on the work of Balasubramanian *et al.* (2014). These methods will be applied to the present case of surge forecasting.

5.1 Conformal prediction for time series analysis

Balasubramanian *et al.* (2014) described two methods to deal with dependent data in the conformal inference framework. The first method uses an autoregressive model, as described in the previous chapter; one thus not mind violating the exchangeable requirement. The observations or samples are intrinsically dependent because different rows of the feature matrix contain the same features. As such, one cannot theoretically guarantee for validity. The second method filters the data such that samples become independent, i.e. any feature only resides in one single sample; see §2.3.3 for more information.

To obtain an independent data set, as proposed in the second method, the data was filtered with a certain time frame such that between-sample correlations were eliminated. This time frame was identified by gradually increasing the frame while sequentially conducting a Ljung-Box independence test (with a significance level of 0.05). This resulted in selecting samples every 491 time intervals of 10 minutes each. As a result, only 1352 samples remained for the period January 2006 till August 2018, which showed to be insufficient to build a predictive model with high performance. It was therefore decided to continue with the first method, expanded with a coverage probability check on the test data, i.e. we checked whether e.g. 95% of the predictions were covered by the 95% prediction interval.

5.2 Evaluation of conformal prediction and its extensions

In this work, the original conformal prediction framework is presented, together with three extensions or modifications. Their performance is assessed by the coverage probability of the 95% conformal prediction interval on a test data set, being independent of the training and calibration data sets. This is shown in Table 5-1 for the different methods. In addition, the prediction intervals will be visually expected based on the storms of December 2013 (Santa Claus storm) and January 2018. These can be found in Figure D 1 - Figure D 7. Note that the weighted random forest regression model from Chapter 4 is used below. The conformal prediction evaluation was performed for both the 6-hours and 24-hours ahead forecast models.

Firstly, the original conformal framework is discussed. To conduct the original framework, we trained the point forecast model on the previously set training and validation data sets, i.e. from 2006-01-11 to 2016-01-01. The nonconformity measures were determined on the same data set (denoted as calibration data) and the coverage probability was checked on the test data from 2016-01-01 to 2018-08-14. Table 5-1 shows a 95% coverage probability on the calibration data set, which was expected seen the nonconformal measure definition on this data. Whereas the prediction interval only covered 77.7% of the predictions for the 6-hours ahead forecast model, the coverage improved to 86.1% for the 24-hours ahead forecast model. These discrepancies from the set value of 95% indicate that the forecast model was not properly generalized and thus resulted in lower coverage probabilities w.r.t. the test data. From Figure D 1 and Figure D 5, one notices that the prediction intervals were very narrow as a result of the large R^2 on the training data (see Table 4-3 and Table 4-4). The R^2 on the test data was much lower, reflecting some overfitting of the forecast model on the training data. This suggests that it might be better to determine the nonconformity measures on a separate and independent data set, such that the prediction interval width becomes more realistic. This is applied in the split conformal framework.

Table 5-1: Coverage probability of 95% conformal prediction intervals

	6-hours forecast		24-hours forecast	
	calibration	test	calibration	test
Original conformal prediction	0.950	0.777	0.950	0.861
Split conformal prediction	0.950	0.963	0.950	0.960
Locally-weighted conformal prediction	0.950	0.961	n.a.	n.a.
Local split conformal prediction	0.949	0.932	0.957	0.941

Secondly, the split conformal framework was originally developed to cope with large data sets and computationally demanding target estimators. For that reason, the method separates the model fitting and residual ranking steps (to determine the nonconformity measure) using sample splitting. Hence, the data is now split in training (from 2006-01-11 04:00 to 2013-07-01 00:00) and calibration (from 2013-07-01 00:00 to 2016-01-01 00:00) data on which, respectively, the point prediction model is trained and the residual ranking is performed.
The performance results are again shown in Table 5-1, and the visual inspection is to be found in Figure D 2 and Figure D 6 for the 6-hours and 24-hours ahead forecast models respectively. The table clearly shows that the nonconformity measure distribution was now more realistic w.r.t. test data: 96% of the predicted water levels were covered by the prediction interval, being

close to the set 95%. Due to the sample split, the 6-hours ahead forecast model was now only able to be trained on one major storm and, as such, was incapable to properly forecast the 2013 storm; the 95% prediction interval did not cover the observed water levels (note that the 24-hours ahead forecast model performed much worse). Surprisingly, the 2018 storm was well predicted (probably because of the more moderate low tide level, similar to the storm the model was trained on). The nonconformity measures were now determined on more realistic, cf. unknown, predicted water levels and thus led to wider prediction intervals, as observed in the figures and the increased coverage probabilities.

Thirdly, locally-weighted conformal prediction was applied to make the prediction intervals variable and dependent on the features. For this purpose, the fitted residuals were inversely scaled with the mean absolute deviation (MAD). The latter was attempted to be forecasted by both a random forest and simple linear regression model. The MAD forecast performance was so bad that this research track was left. For completeness, results for the 6-hours ahead forecast model are given in Table 5-1 and Figure D 3.

Lastly, an alternative to the former method was examined in order to obtain feature-dependent prediction intervals: the local split conformal framework. Instead of using the entire calibration data set for defining the nonconformity measure distribution, we now selected a subset of n calibration data being representative for the considered test sample, i.e. the n closest samples in the high-dimensional feature space (in terms of Euclidean distance). The optimal subset size n was determined as 200, when the coverage probability leveled off at increasing size (data not shown). Hence, for every sample in the test data set, 200 “similar” samples were searched for in the calibration data set by the KNN algorithm. The prediction intervals are thus tailored on the test feature characteristics. The coverage probabilities on the test data are summarized in Table 5-1. Both the 6-hours and 24-hours ahead forecast models were characterized by slightly lower coverage probabilities than 95%; they are still acceptable. As demonstration, Figure 5-1 now shows a locally variable prediction interval for the 6-hours ahead forecast model; see Figure D 4 and Figure D 7 for the other storms and forecast horizon. The figures indicate not only wider intervals around low and high tide, but also around the tide transitions. The latter agrees with the surge residuals largely deviating from the observed ones in these specific tidal phases.

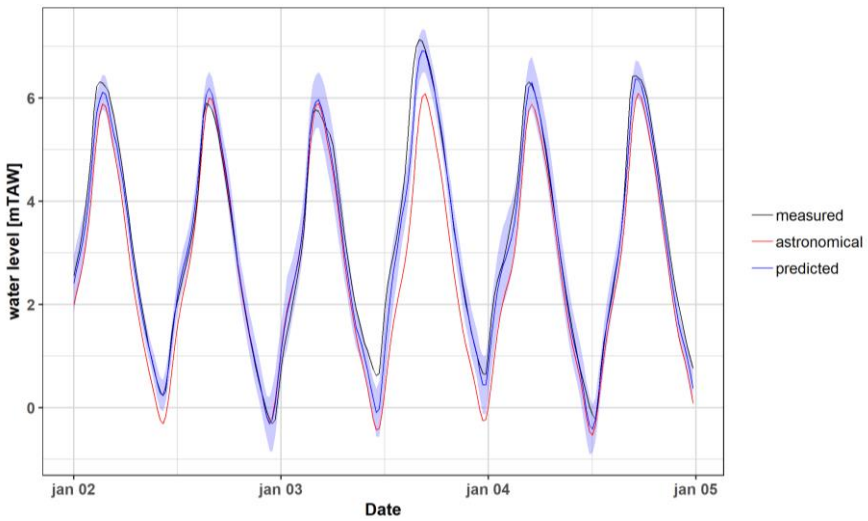


Figure 5-1: Local split conformal prediction intervals for the January 2018 storm, supplemented to the 6-hours ahead weighted random forest regression model (the astronomical tide is shown too)

5.3 Alert classification revisited

In §4.4.2, we evaluated the random forest regression model w.r.t. its alert classification performance on high water levels. The two considered classes, i.e. “normal” and “alert”, were separated by a threshold water level of 6.3 m TAW. Besides the fact that some events were wrongly identified as alert, 65 and 47% of the alert events were missed for the 24-hours and 6-hours ahead forecast models respectively. The latter can be improved by considering the uncertainty on the point prediction. As such, Table 5-2 revisits the classification performance but now based on the upper 95% local split conformal prediction limit.

From Table 5-2, we may conclude that now 6 (i.e. 1 event) and 0% of the 17 true alert events were missed during the period Jan 2016 - August 2018. This is a major improvement, but the other side of the coin is that the false discovery rate increased. Instead of 7 alert events with the 24-hours ahead forecast random forest model, we now identified 99 events with the local split conformal prediction approach. This number decreased to 92 for the 6-hours ahead prediction. Despite the large frequency of falsely triggered alerts, the algorithm did not miss any true alert event. Similar to §4.4.2, we may conclude that the alert messaging needs further improvement. Note, however, that these events were very rare and occurred only in 0.5% of all high water levels, hence, demanding high model performances.

Table 5-2: Alert classification performance on test data (from Jan 1, 2016 until August 14, 2018) for the upper 95% local split conformal prediction limit of weighted random forest regression

			high water levels (m TAW)	
forecast horizon	lag horizon		normal]-∞, 6.3[	alert [6.3, ∞[
24	13	# observations	3144	17
		# forecasts	3062	99
		positive predicted value, PPV (precision)	1	0.162
		false discovery rate, FDR (fall-out)	0	0.838
		false negative rate (miss rate)	0	0.059
6	13	# observations	3137	17
		# forecasts	3062	92
		positive predicted value, PPV (precision)	1	0.185
		false discovery rate, FDR (fall-out)	0	0.815
		false negative rate (miss rate)	0	0

5.4 Conclusions

Any practical use of forecast models implicates knowledge of the model uncertainty. In this respect, the forecasted water levels from the weighted random forest regression model were supplemented with 95% prediction intervals. These uncertainty bands were computed by means of the conformal prediction framework. The advantage here is that the methodology can be applied independent of the machine learning algorithm and underlying multivariate distribution of the data. Prudence is important when applying the conformal prediction framework because it assumes exchangeable data, which is not a priori valid in time series

analysis. For this reason, the coverage probability was checked on test data to make sure the approach was acceptable.

To obtain locally variable prediction intervals, the nonconformal measure for a test sample was based on a subset of 200 calibration data being representative for this sample, i.e. being closest in the high-dimensional feature space. The KNN algorithm was applied for this purpose. In comparison to some other alternative conformal methods, this local split conformal prediction framework resulted in the best results.

Once an interval prediction can be performed, the alert classification performance from Chapter 4 can be revisited. Alert events were now identified by the upper 95% conformal prediction limit. No or almost no alert events were missed by the classification effort, but the other side of the coin was that many more events were falsely classified as alert.

CHAPTER 6

CONCLUSIONS

Throughout history, the Scheldt estuary had a large economic value and drove the industry around Antwerp and the hinterland. Besides economic benefits, it has a large impact on nature as well. Water levels are not only driven by upstream river discharges but mainly by the tidal influences, because of the Scheldt's connection to the North Sea. Spring tide accompanied with strong north-western winds may lead to extreme high water levels and thus potentially endangering Antwerp city with flooding. Flood protection walls along the quays therefore need to keep the city dry. Hence, in-time closure of these protection walls requires good storm surge forecasts.

Flanders Hydraulic Research, a division of the department of Mobility and Public Works of the Government of Flanders, delivers forecast modelling tools for many purposes. Scheldt water level forecasts use nowadays physically-based hydrodynamic models with wind effects, and astronomical tides are applied as boundary conditions. These variables and others, like atmospheric pressure, river discharges, etc. are known to result in discrepancies between the astronomical and observed water levels. Note that not only the tidal amplitude may differ but also the high and low water time of occurrence. This is called skew surge. This research therefore aimed to improve the water level forecasts near Antwerp by feeding these environmental variables into a data-driven tidal surge forecast model.

The Scheldt estuary had been extensively measured over the last decades, resulting in huge time series data sets. As such, data was available from January 1998 until August 2018 covering water levels, wind speed and direction, atmospheric pressure, air and water temperature, and the river discharge upstream of Antwerp. Not only measurements but also predictions were available for the wind speed and direction. Further, some variables were measured at multiple locations along the estuary. Due to the very similar cross-correlation function of wind at Vlakte van de Raan and Hansweert with surge level, the former was selected based on the maximum correlation value. Water levels of Antwerp and Vlissingen were both retained, as the difference between their cross-correlation functions reflects the hydrodynamic properties of the estuary such as confinement, bends and side channels. Note that data originated from different sources and required the necessary data preparation, cleaning and imputation of missing values.

This tidal surge forecasting task is essentially a time series analysis problem. It was opted to model the surge and not the water levels themselves though. For this purpose, a harmonic analysis was first performed to subtract the astronomical water levels from the observed ones. A sliding window multi-step forecast model was applied, i.e. prior time steps were used to predict multiple time steps ahead. A lag horizon of 13 hours was withheld because longer horizons did not significantly improve the model performance anymore. In addition, both a 6-hours and 24-hours ahead forecast model were trained to examine their difference in performance. In the present application, we not only have lagged observational data, but also forecasts between the present time and the 6-hours or 24-hours ahead forecast time.

Different supervised learning techniques were subsequently evaluated on their surge forecast performance. Linear regression techniques covered ordinary linear regression and its penalized variants Lasso, Ridge and elastic net. Regularization or penalization primarily aimed at properly dealing with correlated predictors or features in order to avoid overfitting. Because of estuarine nonlinear system behaviour, increased forecast performances were expected by applying nonlinear models. In this respect, random forest regression, least-squares support vector regression, extreme learning machines and multiple layer perceptron models were evaluated.

These learning methods were trained with the validation set approach; this approach was expected to be allowed because of the extent of the available time series. Model training happened on 6.5 years of data, whereas the validation period consisted of 2.5 years. The forecast performance was subsequently evaluated on an independent out-of-sample data set of another 2.5 years. Contiguous time periods were considered here with excluded data at the period's start to account for autocorrelation. If a learning technique required hyperparameters, they were tuned by a grid search approach.

From this comparative study, random forest regression resulted in the best tidal surge forecast performance. Whereas the 24-hours ahead forecast model returned a R^2 of 0.57, the 6-hours ahead forecast model obviously gave better results with a R^2 of 0.67. Because one of the model applications consists of storm tide forecasting with related alert message triggering, the forecast performance towards high water levels had been improved by weighting extreme observations. The latter were defined as water levels exceeding the 95-percentile (peaks-over-threshold or POT) water levels. Evaluated performance measures were therefore not only based on the entire tidal cycle but also solely on the POT water levels. The finally selected random forest regression model put 30 times more weight on the POT observations. The R^2 POT was determined as 0.69 and 0.75 respectively. In general, the model underpredicted the absolute surge levels. The largest errors on the forecasted surge levels were observed for water levels around 2.5 m TAW and measured up to more than 2 m. These large discrepancies occurred during the transitional phase between high-low tides, and vice versa, where the water level quickly changed. A small temporal shift in tide by skew surges resulted in large deviations.

Storm surge alert situations were defined as tides where the high water level exceeded 6.3 m TAW. For proper alerting, and subsequent initiation of mitigation measures, predicted tides should be properly classified as a normal or alert event. Only 17 alert events occurred during the test period Jan 2016 - August 2018, i.e. 0.5% of all observed high water levels, illustrating how rare such situations were. The 24-hours ahead forecast model was capable of identifying only 7 of these alert events. Additionally, 14.3% or 1 event was falsely identified as extreme and 64.7% of the observed ones were missed during identification, i.e. 11 alert events. On the other hand, the 6-hours ahead forecast model identified 20 alert events, being larger than the observed number. Now, three more alert events than the 24-hours ahead model were correctly classified. False alert messages were triggered as well though. The model results were obviously better than the astronomical tidal forecasts where no alert events were identified.

Nonparametric models such as random forests cannot be used for inference but some explanatory analysis can be based on feature importances. Feature importance ranking gave some indication of which features had the largest predictive power. As a result, the actual wind speed at Vlakte van de Raan, measured along the Lower Sea Scheldt river stretch, was

identified as the primary predictor in the 6-hours ahead forecast model. The observed surge at a 6.3-hours lag was secondly ranked.

Any practical use of forecast models implicates knowledge of the model uncertainty. Uncertainty on the water level forecasts was computed by means of the conformal prediction framework. The methodological advantage here is its independence of the machine learning algorithm and underlying multivariate distribution of the data. Conformal prediction assumes exchangeable data though, which is not a priori valid in time series analysis. For this reason, the coverage probability was checked on test data to make sure the approach was acceptable. After evaluating several conformal prediction method alternatives, local split conformal prediction returned the best results. Once an interval prediction could be performed, the alert classification performance was revisited; alert events were now identified by the upper 95% conformal prediction limit and not by the point prediction itself. No or almost no alert events were missed by the classification effort, but the other side of the coin was that many more events were falsely classified as alert. Hence, the alert classification needs further improvement.

Further research

This work should be seen as a first exploration of different surge forecasting techniques. From the conducted work, several topics were identified that deserve more attention and effort in future research.

The surge forecast performance strongly depends on the quality of the surge data, being defined as the difference between observed and astronomical water levels. These astronomical water levels were based on a harmonic analysis of the entire water level time series. According to IMDC (2013), the maximum difference between high and low tide increases on the long term, and it moves more upstream of the Scheldt estuary too. In addition, Gerritsen & van den Bogaard (1998) documented a periodic and sometimes more erratic evolution of tidal components in their analyses. This indicates that the long-term harmonic analysis, as conducted in this work, averaged out these trend. A harmonic analysis based on, e.g., one-year water level time series is expected to improve the surge data quality, and thus the model forecast performance.

In addition to the previous, the model performance may also be improved by applying the so-called *rolling origin* approach. Rolling origin is an evaluation technique according to which the forecast origin is updated successively and the forecasts are produced from each origin. As such, the model is estimated on the training data, and n-step ahead forecasts are subsequently made. When new observations become available, the model is re-trained and forecasts are again made, but now starting from the new origin. This method would thus allow the consideration of evolving estuary system dynamics. In this respect, on-line learning techniques may give improvements as well.

Autoregressive models generally require stationary data. This data requirement was not fulfilled in this thesis though. Attempts had been made to properly transform the data by differencing, but stationarity was not obtained and it was decided to continue with non-

stationary data. It may be worthwhile to further examine the impact of the degree of differencing on the model performance. We may also include several degrees of differences in the forecast model and let the learning model select the necessary predictors or features.

Further, the surge forecast models in this work were trained with an objective function based on data of the entire tidal cycle; the model performance is thus averaged over all data. One of the model applications covers extreme high water forecasting so an objective function based on only the maximum high water levels may improve these forecasts. When the high water levels are focused upon, the simultaneous modelling of the maximum high water level and their time of occurrence by multivariate modelling techniques can be a worthwhile alternative approach. Correct alert classification is important and was dealt with in §4.4.2 and §5.3. Dedicated models could improve the classification performance, such as autoregressive multinomial logit models (Augustin *et al*., 2008), support vector machines and artificial neural networks.

Finally, multiple layer perceptron models were applied in this work as a nonlinear modelling technique. Training these models is not an easy task and requires time and effort. In this respect, artificial neural networks should deserve extra attention in future research, especially dedicated techniques such as recurrent neural networks and its long-short term memory (LSTM) variant for time series analysis.

REFERENCES

Augustin N.H., Beevers L. and Sloan W.T. (2008). Predicting river flows for future climates using an autoregressive multinomial logit model. Water Resources Research, 44(7).

Balasubramanian V.N., Ho S.-S. and Vovk V. (2014). Conformal prediction for reliable machine learning theory, adaptations and applications. Morgan Kaufmann, 334 pp.

Bergmeir C. and Benitez J.M. (2012). On the use of cross-validation for time series predictor evaluation. Information Sciences, 191, 192-213.

Bergmeir C., Hyndman R.J. and Koo B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. Computational Statistics and Data Analysis, 120, 70-83.

Chernozhukov V., Wüthrich K. and Zhu Y. (2018). Exact and robust conformal inference methods for predictive machine learning with dependent data. Proceedings of Machine Learning Research, 75, 1–17.

Dashevskiy M. and Luo Z. (2008). Network traffic demand prediction with confidence. In: Proceedings of the Global Communications Conference (GLOBECOM 2008), New Orleans, USA, 30 November - 4 December 2008.

De Brabanter K. (2011). Least-squares support vector regression with applications to large-scale data: a statistical approach. PhD thesis, KU Leuven.

De Standaard (2018). http://www.standaard.be/cnt/dmf20180103_03278019

Elgohary T, Mubasher A. and Salah H. (2017). Significant deep wave height prediction by using support vector machine approach (Alexandria as case of study). International Journal of Current Engineering and Technology, 7(1), 135-143.

Gerritsen H. en van den Boogaard H. (1998). Getijanalyse Westerschelde. Datarapport getijcomponenten. Toepassing van Principal Component Analysis. WL|delft hydraulics i.o.v. Rijkswaterstaat, the Netherlands.

James G., Witten D., Hastie T. and Tibshirani R. (2017). An introduction to statistical learning with applications in R. New York, Springer, 426 pp.

Lei J., G'Sell M., Rinaldo A., Tibshirani R.J. and Wasserman L. (2017). Distribution-free predictive inference for regression. Journal of the American Statistical Association, 113(523), 1094-1111.

Linusson H. (2017). An introduction to conformal prediction. Totorial Day, The 6th Symposium on Conformal and Probabilistic Prediction with Applications (COPA 2017), Stockholm, Sweden, 14-16 June 2017.

Londhe S. and Gavraskar S.S. (2015). Forecasting one day ahead stream flow using support vector regression. Aquatic Procedia, 4, 900-907.

Mafi S. and Amirinia G. (2017). Forecasting hurricane wave height in Gulf of Mexico using soft computing methods. Ocean Engineering, 146, 352-362.

Manning C.D., Raghavan P. and Schütze H. (2009). An introduction to information retrieval. Cambridge University Press, Cambridge, UK, 544 pp.

Nguyen T.-T., Huu Q.N. and Li M.J. (2015). Forecasting time series water levels on Mekong River using machine learning models. 7th International Conference on Knowledge and Systems Engineering, October 8-10 2015, Vietnam.

Opsomer J., Wang Y. and Yang Y. (2001). Nonparametric regression with correlated errors. Statistical Science, 16(2), 134-153.

Rajasekaran S., Gayathri S. and Lee T.-L. (2008). Support vector regression methodology for storm surge predictions. Ocean Engineering, 35, 1578-1587.

Rasmussen C. and Williams C. (2006). Gaussian processes for machine learning. MIT Press, USA, 248 pp.

Roberts K.J., Colle B.A., Georgas N. and Munch S.B. (2015). A regression-based approach for cool-season storm surge predictions along the new York-New Jersey coast. Journal of Applied Meteorology and Climatology, 54, 1773-1791.

Royston S.J., Horsburgh K.J. and Lawry J. (2012). Application of rule based methods to predicting storm surge. Continental Shelf Research, 37, 79-91.

Shafer G. and Vovk V. (2008). A tutorial on conformal prediction. Journal of Machine Learning Research, 9, 371-421.

Shumway R.H. and Stoffer D.S. (2017). Time series analysis and its applications. Springer International Publishing AG, Fourth Ed., 562 pp.

Steidley C., Sadovski A., Tissot P., Bachnak R. and Bowles Z. (2005). Water level prediction with artificial neural network models. In: Proceedings of the 6th WSEAS international conference on Automation & information, Buenos Aires, Argentina, March 1-3, 2005.

Stoorvogel A.A. and Habets L.C.G.J.M. (2002). Rapport tijdreeksanalyse Westerschelde. ICTOO/TU Delft: Eindhoven, the Netherlands.

Tsai C.-P. and You C.-Y. (2014). Development of models for maximum and time variation of storm surges at the Tanshui estuary. Nat. Hazards Earth Syst. Sci., 14, 2313-2320.

Wang Z.B. and Winterwerp H. (2013). Data-analyse waterstanden Westerschelde: basisrapport grootschalige ontwikkeling G-1. Instandhouding vaarpassen Schelde Milieuvergunningen terugstorten baggerspecie. IMDC, in cooperation with Deltares, Svasek Hydraulics and ARCADIS.

LIST OF ABBREVIATIONS

ACF	autocorrelation function
ANN	artificial neural network
ARIMA	autoregressive integrated moving average
ARIMAX	ARIMA model with exogenous variables X
CCF	cross-correlation function
CDF	cumulative distribution function
CV	cross-validation
ELM	extreme learning machine
FDR	false discovery rate (fall-out)
FNR	false negative rate (miss rate)
FP	false positives
HIC	Hydrological Information Centre
KNMI	Royal Netherlands Meteorological Institute
KNN	k-nearest neighbor
LOCF	last observation carried forward
LS-SVR	least-squares support vector regression
LSTM	long-short term memory
MAD	mean absolute deviation
ML	machine learning
MLP	multiple layer perceptron
MSE	mean squared error
POT	peak-over-threshold
PPV	positive predicted value (precision)
R^2	coefficient of determination
RMSE	root-mean-squared error
SVR	support vector regression
TAW	Tweede Algemene Waterpassing
TP	true positives
VMM	Flanders Environmental Agency

APPENDIX A

PROVIDED DATA

Table A 1: Overview of observed and predicted data, provided by Flanders Hydraulic Research

data	location	type	start date	end date	time interval	provider	data cleaning
tide	Antwerpen	M	1/01/1998	14/08/2018	10 min	HIC	
	Vlissingen	M	1/01/1998	14/08/2018	10 min	HIC	
wind speed & direction	Vlakte van de Raan	M	1/06/1998	14/08/2018	10 min	HIC	
	Hansweert	M	1/01/1998	14/08/2018	10 min	HIC	
		P	11/01/2006	3/07/2018	• 10 min • forecast horizon of 50h, updated every 6h	KNMI	
	Terneuzen	P	11/01/2006	3/07/2018	• 10 min • forecast horizon of 50h, updated every 6h	KNMI	• removal of duplicates (with slightly different values) • linear interpolation for intermediate times, unless missing data for more than 12h
air temp.	Melsele	M	1/01/2014	14/08/2018	15 min	HIC/ VMM	• removal duplicates • replace extreme values, i.e. ΔT between subsequent times is >2°C, by average of neighbours • linear interpolation for intermediate times
barometric pressure	Melsele	M	1/01/2014	14/08/2018	15 min	HIC/ VMM	• removal duplicates • replace extreme values, i.e. ΔP between subsequent times is >10Pa, by average of neighbours • linear interpolation for intermediate times
river discharge	Demer	M	1/01/2002	25/09/2018	60min/15min /5min	HIC	• missing values for max. 1h are linearly interpolated due to slowly changing discharges
	Dender	M	1/01/2002	25/09/2018	24h/15min/5 min	HIC	• 24h-meas are kept constant, i.e. LOCF • linear interpolation for 15min/5min measurements
	Dijle	M	1/01/2002	2/10/2018	15 min	HIC	• missing values for max. 1h are linearly interpolated due to slowly changing discharges
	Grote Nete	M	1/01/2002	25/09/2018	60min/15min /5min	HIC	• missing values of max. 1h are linearly interpolated
	Kleine Nete	M	1/01/2002	25/09/2018	60min/15min /5min	HIC	• missing values are linearly interpolated
	Zeeschelde	M	1/01/2002	25/09/2018	24h/15min/5 min	HIC	• 24h-meas are kept constant, i.e. LOCF • linear interpolation for 15min/5min measurements
	Zenne	M	1/01/2002	25/09/2018	60min/15min /5min	HIC	• missing values are linearly interpolated

data	location	type	start date	end date	time interval	provider	data cleaning
water temp.	Prosperpolder	M	1/1/2002	24/09/2018	5 min	HIC	• missing values for max. 1h are linearly interpolated due to slowly changing discharges

Notes:

- Type: *M* and *P* refer to *measured* and *predicted* variables respectively.
- Missing values are imputed for all variables, but only 1 value of 10 min each is imputed due to two reasons:
  1. data is collected nowadays at a frequency of 5-15 min so one acknowledges that this is the necessary frequency to capture all necessary dynamics
  2. imputing only one value drastically lowers the number of missing data periods, hence enlarging our observation data matrix
- Wind direction is given in degrees.

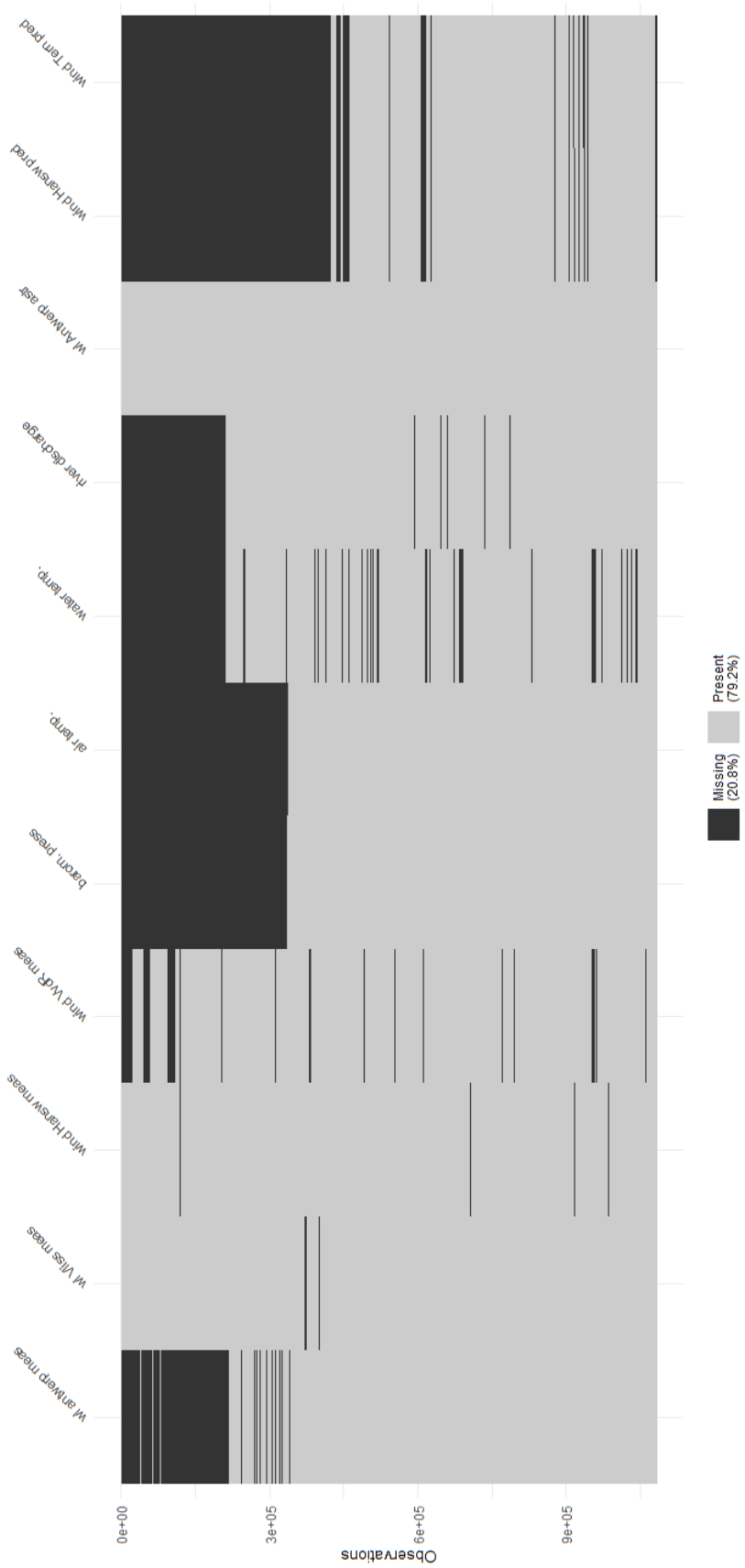


Figure A 1: Overview of missing data between 1998 and 2018

APPENDIX B

DATA EXPLORATION

Table B 1: Fitted tidal component parameters with the R package TideHarmonics (For the naming scheme of the harmonic constituents, please refer to: https://CRAN.R-project.org/package=TideHarmonics)

component names	value	95% CI
Sa_S	-0.03143	[-0.03235 , -0.03052]
Ssa_S	-0.00133	[-0.00224 , -0.00041]
Mm_S	0.012497	[0.01159 , 0.01341]
MSf_S	0.039562	[0.03865 , 0.04048]
Mf_S	0.013074	[0.01227 , 0.01388]
`2Q1_S`	-0.00199	[-0.00288 , -0.0011]
sig1_S	0.002735	[0.00185 , 0.00363]
Q1_S	0.003493	[0.0026 , 0.00438]
rho1_S	6.64E-05	[-0.00082 , 0.00096]
O1_S	0.089575	[0.08868 , 0.09046]
`1M.1p1_S`	-0.00431	[-0.00523 , -0.0034]
M1_S	0.002002	[0.00144 , 0.00257]
chi1_S	0.001752	[0.00086 , 0.00265]
pi1_S	0.001659	[0.00074 , 0.00257]
P1_S	-0.02172	[-0.02264 , -0.02081]
S1_S	-0.00755	[-0.00847 , -0.00664]
K1_S	-0.05628	[-0.05719 , -0.05538]
psi1_S	0.003372	[0.00246 , 0.00429]
phi1_S	-0.0004	[-0.00129 , 0.0005]
the1_S	-0.00045	[-0.00135 , 0.00044]
J1_S	-0.00083	[-0.00172 , 0.00007]
`1S.1o1_S`	0.003383	[0.00249 , 0.00427]
OO1_S	-0.00028	[-0.00109 , 0.00052]
`1o1q.2_S`	-0.01247	[-0.01331 , -0.01163]
`1M1N.1S2_S`	-0.02951	[-0.03042 , -0.02859]
`2N2_S`	0.038124	[0.03721 , 0.03904]
mu2_S	-0.17234	[-0.17326 , -0.17143]
N2_S	0.342162	[0.34125 , 0.34308]
nu2_S	0.133623	[0.13271 , 0.13454]
`1o1p.2_S`	-0.02477	[-0.02566 , -0.02388]
M2_S	1.927675	[1.92676 , 1.92859]
`1M1K.1S2_S`	-0.0304	[-0.03128 , -0.02952]
lam2_S	0.065446	[0.06453 , 0.06636]
L2_S	0.143448	[0.14257 , 0.14433]
T2_S	0.012247	[0.01133 , 0.01316]
S2_S	0.04164	[0.04072 , 0.04256]
R2_S	-0.00386	[-0.00478 , -0.00295]
K2_S	0.017291	[0.01642 , 0.01816]
`1M1S.1N2_S`	0.00855	[0.00764 , 0.00946]
`1k1j.2_S`	0.001876	[0.00101 , 0.00275]
`2S.1M2_S`	0.01597	[0.01551 , 0.01643]
`1M1o.3_S`	0.046986	[0.04609 , 0.04788]
M3_S	0.00994	[0.00903 , 0.01085]
`1S1o.3_S`	0.00191	[0.00102 , 0.0028]
`1M1k.3_S`	-0.03848	[-0.03939 , -0.03757]
`1S1k.3_S`	-0.00973	[-0.01063 , -0.00882]
`1M1N.4_S`	-0.01678	[-0.01769 , -0.01586]
M4_S	-0.08546	[-0.08637 , -0.08455]
`1S1N.4_S`	-0.00512	[-0.00603 , -0.0042]
`1M1S.4_S`	-0.07438	[-0.07529 , -0.07346]
`1M1K.4_S`	-0.02093	[-0.02181 , -0.02005]
S4_S	-0.0025	[-0.00342 , -0.00159]
`1S1K.4_S`	-0.00048	[-0.00135 , 0.00039]
`2M1N.6_S`	-0.00966	[-0.01011 , -0.0092]
M6_S	0.012569	[0.01166 , 0.01348]
`1M1S1N.6_S`	0.014278	[0.01337 , 0.01519]
`2M1S.6_S`	0.046094	[0.04564 , 0.04655]
`2M1K.6_S`	0.012561	[0.01212 , 0.013]
`2S1M.6_S`	0.009262	[0.0088 , 0.00972]
`1M1S1K.6_S`	0.011764	[0.01088 , 0.01264]
`2M1N.2S2_S`	-0.00125	[-0.00147 , -0.00102]
`3M.1S1K2_S`	-0.00073	[-0.00102 , -0.00044]

component names	value	95% CI
Sa_C	-0.05426	[-0.05517 , -0.05334]
Ssa_C	-0.0169	[-0.01782 , -0.01599]
Mm_C	0.006714	[0.0058 , 0.00763]
MSf_C	0.05299	[0.05208 , 0.0539]
Mf_C	0.004213	[0.00341 , 0.00502]
`2Q1_C`	0.004013	[0.00312 , 0.0049]
sig1_C	-0.00516	[-0.00605 , -0.00427]
Q1_C	0.032883	[0.03199 , 0.03377]
rho1_C	0.007159	[0.00627 , 0.00805]
O1_C	0.054434	[0.05354 , 0.05532]
`1M.1p1_C`	-0.00226	[-0.00318 , -0.00135]
M1_C	6.27E-05	[-0.0005 , 0.00063]
chi1_C	-0.00045	[-0.00135 , 0.00045]
pi1_C	-0.00187	[-0.00279 , -0.00096]
P1_C	-0.03552	[-0.03644 , -0.03461]
S1_C	-0.00788	[-0.00879 , -0.00696]
K1_C	-0.04844	[-0.04935 , -0.04754]
psi1_C	-0.00217	[-0.00309 , -0.00126]
phi1_C	-0.00102	[-0.00191 , -0.00012]
the1_C	-0.00094	[-0.00183 , -0.00004]
J1_C	0.005359	[0.00446 , 0.00626]
`1S.1o1_C`	0.005193	[0.0043 , 0.00608]
OO1_C	0.005586	[0.00478 , 0.00639]
`1o1q.2_C`	-0.01774	[-0.01857 , -0.0169]
`1M1N.1S2_C`	-0.03734	[-0.03825 , -0.03642]
`2N2_C`	0.000697	[-0.00022 , 0.00161]
mu2_C	-0.17193	[-0.17285 , -0.17102]
N2_C	-0.09272	[-0.09363 , -0.0918]
nu2_C	0.010361	[0.00945 , 0.01128]
`1o1p.2_C`	-0.0029	[-0.00379 , -0.00201]
M2_C	-1.17283	[-1.17374 , -1.17191]
`1M1K.1S2_C`	0.008489	[0.00761 , 0.00937]
lam2_C	-0.05323	[-0.05415 , -0.05232]
L2_C	-0.11107	[-0.11195 , -0.11019]
T2_C	-0.02967	[-0.03059 , -0.02876]
S2_C	-0.57296	[-0.57388 , -0.57205]
R2_C	-0.00234	[-0.00325 , -0.00142]
K2_C	-0.16855	[-0.16942 , -0.16768]
`1M1S.1N2_C`	0.044548	[0.04363 , 0.04546]
`1k1j.2_C`	-0.0036	[-0.00447 , -0.00273]
`2S.1M2_C`	0.025969	[0.02551 , 0.02643]
`1M1o.3_C`	-0.01298	[-0.01388 , -0.01208]
M3_C	0.001789	[0.00088 , 0.0027]
`1S1o.3_C`	-0.02284	[-0.02373 , -0.02195]
`1M1k.3_C`	-0.00478	[-0.00569 , -0.00387]
`1S1k.3_C`	0.008924	[0.00802 , 0.00983]
`1M1N.4_C`	-0.03593	[-0.03684 , -0.03502]
M4_C	-0.08294	[-0.08386 , -0.08203]
`1S1N.4_C`	0.000143	[-0.00077 , 0.00106]
`1M1S.4_C`	0.006459	[0.00554 , 0.00737]
`1M1K.4_C`	0.002442	[0.00156 , 0.00332]
S4_C	0.002873	[0.00196 , 0.00379]
`1S1K.4_C`	0.002345	[0.00147 , 0.00321]
`2M1N.6_C`	0.035236	[0.03478 , 0.03569]
M6_C	0.134897	[0.13399 , 0.13581]
`1M1S1N.6_C`	0.01566	[0.01475 , 0.01657]
`2M1S.6_C`	0.042549	[0.04209 , 0.04301]
`2M1K.6_C`	0.011134	[0.01069 , 0.01157]
`2S1M.6_C`	-0.00157	[-0.00203 , -0.00112]
`1M1S1K.6_C`	-0.00209	[-0.00297 , -0.00121]
`2M1N.2S2_C`	0.004001	[0.00377 , 0.00423]
`3M.1S1K2_C`	0.012289	[0.012 , 0.01258]

component names	value	95% CI
`3M.2S2_S`	-0.00134	[-0.00149 , -0.00119]
`1M1N1K.2S2_S`	0.001447	[0.001 , 0.00189]
`1S1N.1K2_S`	0.002452	[0.00157 , 0.00333]
`2S.1K2_S`	0.001062	[0.00063 , 0.0015]
`2M1S.2N2_S`	-0.00019	[-0.00042 , 0.00003]
`1M1q.3_S`	0.018192	[0.0173 , 0.01909]
`2M.1p3_S`	0.003854	[0.0034 , 0.00431]
`2M.1q3_S`	0.001174	[0.00073 , 0.00162]
`3M.1K4_S`	-0.00013	[-0.00043 , 0.00016]
`3M.1S4_S`	-0.00091	[-0.00122 , -0.00061]
`2M1S.1K4_S`	-0.0003	[-0.00074 , 0.00014]
`3M.1k5_S`	0.003439	[0.00314 , 0.00374]
M5_S	-3.5E-05	[-0.00095 , 0.00088]
`3M.1o5_S`	-0.00441	[-0.0047 , -0.00411]
`2M2N.1S6_S`	0.001938	[0.00171 , 0.00217]
`3M1N.1S6_S`	0.007242	[0.00694 , 0.00755]
`4M.1K6_S`	0.001302	[0.00108 , 0.00152]
`4M.1S6_S`	0.006425	[0.0062 , 0.00665]
`2M1S1N.1K6_S`	0.000475	[0.00003 , 0.00092]
`2M1V.6_S`	-0.00673	[-0.00719 , -0.00628]
`3M1S.1K6_S`	0.000794	[0.0005 , 0.00109]
`4M.1N6_S`	-0.00194	[-0.00217 , -0.00171]
`3M1S.1N6_S`	-0.00997	[-0.01027 , -0.00966]
`1M1K1L.6_S`	0.005271	[0.00442 , 0.00612]
`2M2N.8_S`	-0.0005	[-0.00073 , -0.00028]
`3M1N.8_S`	0.003315	[0.00301 , 0.00362]
M8_S	0.032043	[0.03114 , 0.03295]
`2M1S1N.8_S`	0.013942	[0.01349 , 0.0144]
`3M1S.8_S`	0.024013	[0.02371 , 0.02432]
`3M1K.8_S`	0.006208	[0.00592 , 0.0065]
`1M1S1N1K.8_S`	0.002574	[0.00169 , 0.00346]
`2M2S.8_S`	0.005521	[0.00529 , 0.00575]
`2M1S1K.8_S`	0.006245	[0.0058 , 0.00668]
`4M1S.10_S`	0.001956	[0.00173 , 0.00218]
`3M2S.10_S`	0.000202	[0.00005 , 0.00035]
`4M1S1N.12_S`	-0.00379	[-0.00402 , -0.00356]
`5M1S.12_S`	-0.00301	[-0.00319 , -0.00283]
`4M2S.12_S`	-0.00028	[-0.00039 , -0.00016]
`1M1V.1S2_S`	-0.01551	[-0.01643 , -0.0146]
`2M.1K2_S`	-0.03122	[-0.03166 , -0.03078]
MA2_S	0.034275	[0.03336 , 0.03519]
MB2_S	-0.02005	[-0.02097 , -0.01914]
`1M1S.1V2_S`	0.005054	[0.00414 , 0.00597]
`1S1K.1M2_S`	0.016515	[0.01564 , 0.0174]
`2M1N.1S4_S`	-0.00125	[-0.00171 , -0.0008]
`1M1V.4_S`	-0.00512	[-0.00603 , -0.00421]
`3M.1N4_S`	0.003856	[0.00355 , 0.00416]
`2M1S.1N4_S`	0.005106	[0.00465 , 0.00556]
NA2_S	0.005323	[0.00441 , 0.00624]
NB2_S	0.005974	[0.00506 , 0.00689]
`1M1S1o.5_S`	0.004149	[0.00325 , 0.00505]
`1M1S1k.5_S`	-0.0038	[-0.00471 , -0.00289]

component names	value	95% CI
`3M.2S2_C`	0.010755	[0.0106 , 0.01091]
`1M1N1K.2S2_C`	0.00117	[0.00073 , 0.00161]
`1S1N.1K2_C`	0.005662	[0.00478 , 0.00654]
`2S.1K2_C`	-0.00178	[-0.00222 , -0.00135]
`2M1S.2N2_C`	-0.0003	[-0.00053 , -0.00007]
`1M1q.3_C`	0.008011	[0.00711 , 0.00891]
`2M.1p3_C`	-0.0019	[-0.00236 , -0.00144]
`2M.1q3_C`	0.002906	[0.00246 , 0.00335]
`3M.1K4_C`	0.001703	[0.00141 , 0.002]
`3M.1S4_C`	0.00392	[0.00362 , 0.00423]
`2M1S.1K4_C`	0.003088	[0.00265 , 0.00353]
`3M.1k5_C`	-6.7E-05	[-0.00037 , 0.00024]
M5_C	1.89E-06	[-0.00091 , 0.00091]
`3M.1o5_C`	-0.00337	[-0.00367 , -0.00308]
`2M2N.1S6_C`	-6.9E-05	[-0.0003 , 0.00016]
`3M1N.1S6_C`	-0.00242	[-0.00273 , -0.00212]
`4M.1K6_C`	-0.00217	[-0.00239 , -0.00195]
`4M.1S6_C`	-0.00412	[-0.00434 , -0.00389]
`2M1S1N.1K6_C`	-0.00184	[-0.00229 , -0.0014]
`2M1V.6_C`	0.012547	[0.01209 , 0.013]
`3M1S.1K6_C`	-0.00362	[-0.00391 , -0.00333]
`4M.1N6_C`	-0.00593	[-0.00616 , -0.0057]
`3M1S.1N6_C`	-0.00498	[-0.00528 , -0.00467]
`1M1K1L.6_C`	0.001751	[0.0009 , 0.0026]
`2M2N.8_C`	0.003907	[0.00368 , 0.00413]
`3M1N.8_C`	0.012938	[0.01263 , 0.01324]
M8_C	0.042541	[0.04164 , 0.04345]
`2M1S1N.8_C`	0.006594	[0.00614 , 0.00705]
`3M1S.8_C`	0.006023	[0.00572 , 0.00633]
`3M1K.8_C`	0.000948	[0.00065 , 0.00124]
`1M1S1N1K.8_C`	-0.00194	[-0.00282 , -0.00105]
`2M2S.8_C`	-0.00417	[-0.00439 , -0.00394]
`2M1S1K.8_C`	-0.00497	[-0.00541 , -0.00453]
`4M1S.10_C`	-0.00111	[-0.00134 , -0.00088]
`3M2S.10_C`	-0.00109	[-0.00124 , -0.00094]
`4M1S1N.12_C`	0.001292	[0.00106 , 0.00152]
`5M1S.12_C`	0.002415	[0.00223 , 0.0026]
`4M2S.12_C`	0.001759	[0.00164 , 0.00187]
`1M1V.1S2_C`	-0.03584	[-0.03675 , -0.03493]
`2M.1K2_C`	-0.02154	[-0.02198 , -0.0211]
MA2_C	-0.00265	[-0.00356 , -0.00173]
MB2_C	-0.01895	[-0.01986 , -0.01803]
`1M1S.1V2_C`	-0.01012	[-0.01103 , -0.00921]
`1S1K.1M2_C`	0.028952	[0.02807 , 0.02983]
`2M1N.1S4_C`	0.002937	[0.00248 , 0.00339]
`1M1V.4_C`	-0.01358	[-0.01449 , -0.01267]
`3M.1N4_C`	0.002337	[0.00203 , 0.00264]
`2M1S.1N4_C`	-0.00197	[-0.00243 , -0.00152]
NA2_C	-0.00175	[-0.00267 , -0.00084]
NB2_C	0.00105	[0.00014 , 0.00197]
`1M1S1o.5_C`	-0.01093	[-0.01182 , -0.01003]
`1M1S1k.5_C`	-0.00972	[-0.01063 , -0.00881]

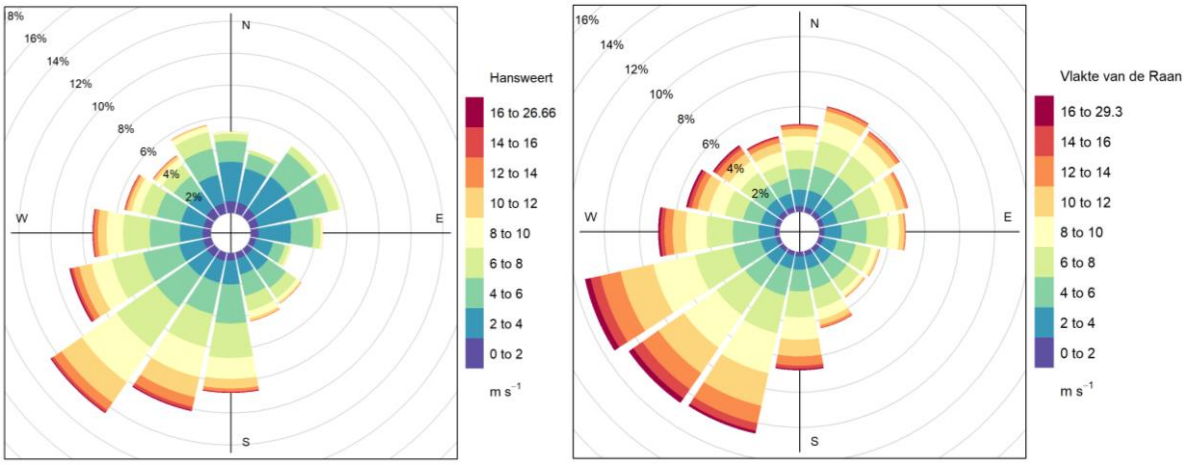


Figure B 1: Wind roses for Hansweert (left) and Vlakte van de Raan (right), based on the period 2006-2018

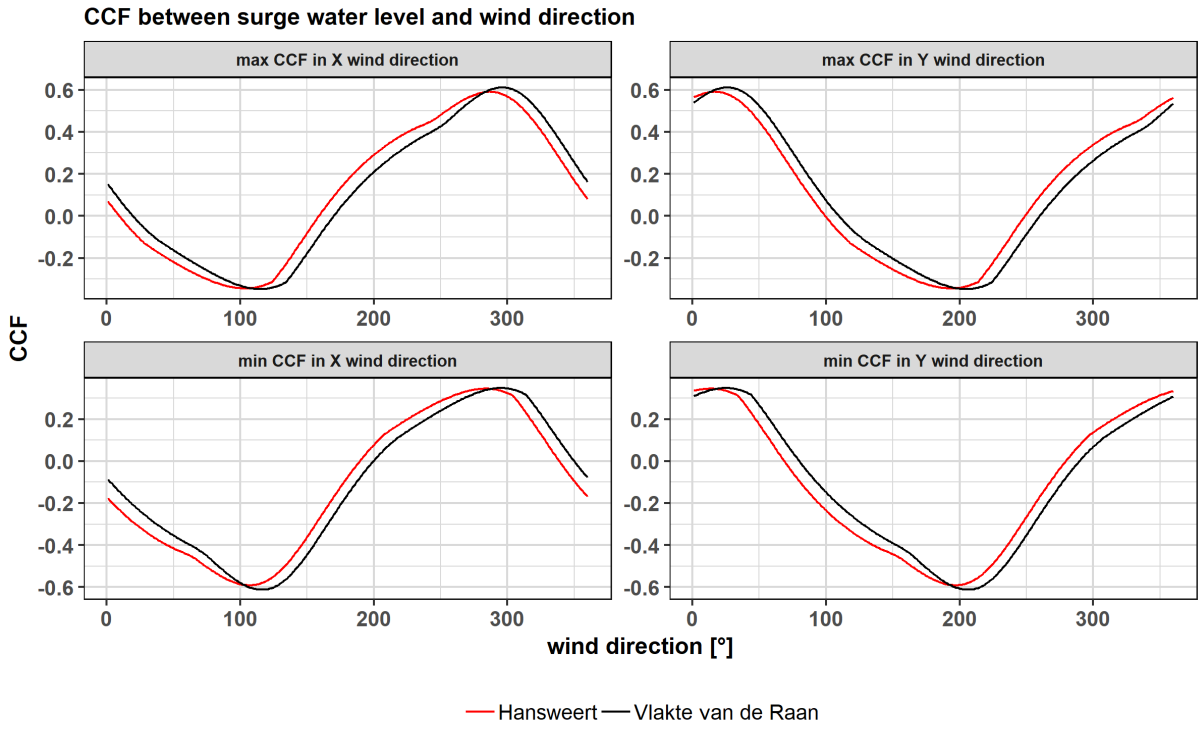


Figure B 2: Minimum and maximum value of the cross-correlation function (CCF) between surge and x/y wind speed as function of the wind directional decomposition. Results are shown for both Hansweert and Vlakte van de Raan, based on the period 2006-2018

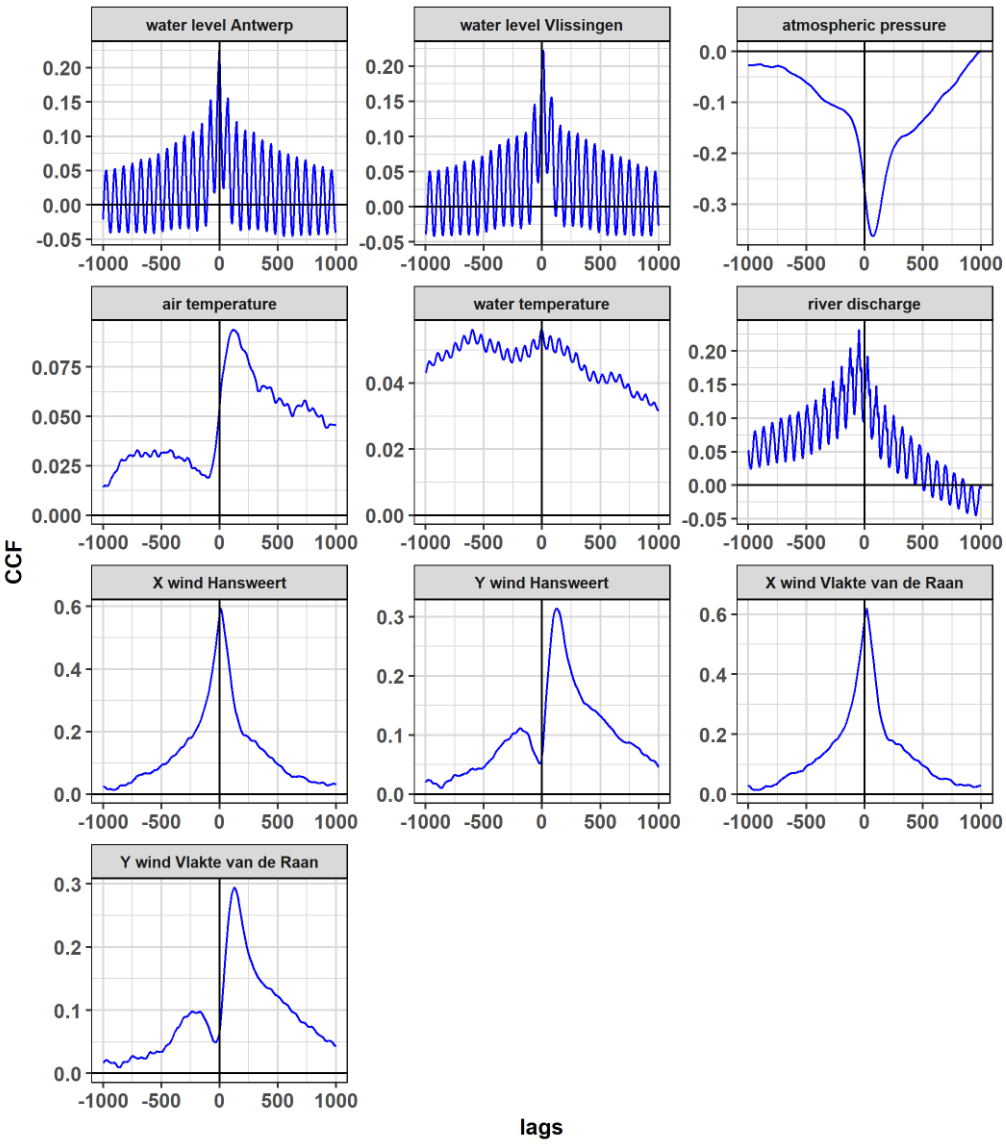


Figure B 3: CCF between the surge level and other observed variables or features (period 2006-2013). Features are leading/lagging the surge when the maximum (absolute) cross-correlation occurs at positive/negative lags

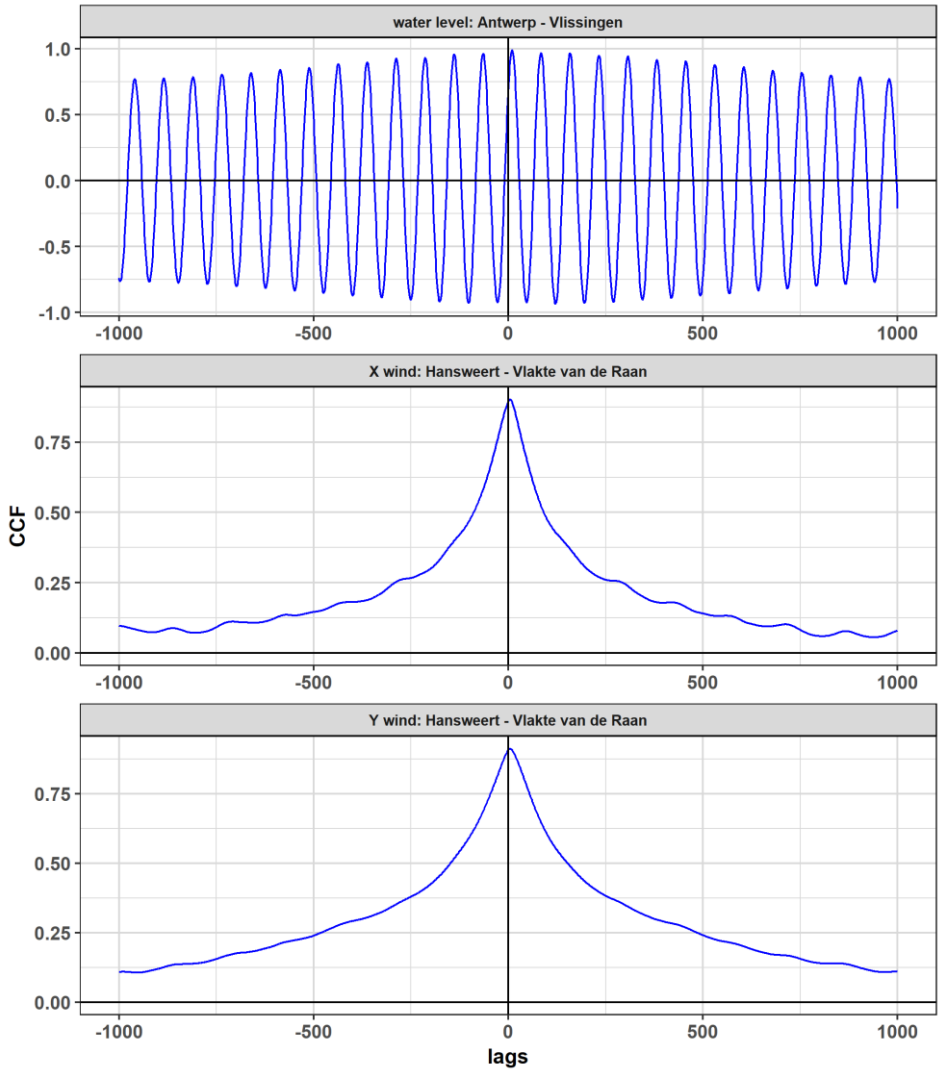


Figure B 4: CCF between different locations for water level (top) and wind (middle and bottom) for the period 2006-2013

Table B 2: Prevalence of alert events for the training, validation and test data sets

#	high water level dates					
	pre-alert	storm tide	gate lockdown	dangerous storm tide	alarm	
1	28/02/2006 16:00	9/11/2007 15:20	18/03/2007 15:20	9/11/2007 4:00		training
2	1/03/2006 4:20	28/02/2010 16:40	25/11/2007 15:40	6/12/2013 5:20		validation
3	1/03/2006 16:40	13/01/2017 3:20	21/03/2008 15:40	3/01/2018 16:00		test
4	31/03/2006 4:20	14/01/2017 4:40	10/02/2009 16:00			
5	31/03/2006 16:40	2/03/2017 5:40	6/12/2013 18:00			
6	7/10/2006 15:20		22/10/2014 2:40			
7	8/12/2006 18:20		28/11/2015 5:00			
8	18/01/2007 15:20		15/01/2016 7:00			
9	21/01/2007 17:00					
10	19/03/2007 3:20					
11	19/03/2007 16:00					
12	20/03/2007 16:40					
13	21/03/2007 5:00					
14	21/03/2007 17:00					
15	7/12/2007 14:00					
16	12/03/2008 6:00					
17	12/03/2008 18:40					
18	13/03/2008 7:00					
19	21/03/2008 3:20					
20	22/03/2008 4:00					
21	25/03/2008 5:40					
22	1/10/2008 16:40					
23	31/01/2010 16:20					
24	1/02/2010 17:00					
25	2/02/2010 17:40					
26	1/03/2010 3:40					
27	1/03/2010 16:00					
28	2/03/2010 16:40					
29	30/08/2010 6:40					
30	9/12/2011 15:00					
31	16/12/2011 19:40					
32	24/12/2011 2:40					
33	31/08/2012 3:20					
34	6/11/2013 4:40					
35	5/01/2014 6:00					
36	24/01/2015 18:20					
37	21/03/2015 16:20					
38	30/11/2015 6:20					
39	13/01/2016 5:20					
40	10/02/2016 4:20					
41	10/02/2016 16:40					
42	12/01/2017 3:00					
43	13/01/2017 16:00					
44	14/01/2017 17:00					
45	1/03/2017 5:00					
46	8/12/2017 6:40					
47	2/01/2018 3:00					
48	4/01/2018 4:40					
49	4/01/2018 17:00					
50	1/02/2018 16:00					
51	2/02/2018 4:40					
52	1/05/2018 4:00					

APPENDIX C

SURGE FORECAST MODELLING

Table C 1: Model performance measures for ordinary and penalized linear regression, random forests, LS-SVR and ELM

				training data		test data		training data		test data	
	lambda	alpha		RMSE	RMSE POT	RMSE	RMSE POT	R^2	R^2 POT	R^2	R^2 POT
OLS				0.214	0.229	0.236	0.270	0.450	0.463	0.261	0.220
Lasso	1.00E-07	1		0.214	0.231	0.235	0.256	0.445	0.453	0.269	0.298
Ridge	1.00E-07	0		0.214	0.231	0.235	0.256	0.445	0.453	0.269	0.298
elastic net	1.00E-07	0.5		0.214	0.231	0.235	0.256	0.445	0.453	0.269	0.298
	ntree	mtry	weight								
Random forests	300	65	1	0.030	0.037	0.180	0.184	0.989	0.986	0.568	0.639
	300	65	30	0.070	0.000	0.181	0.171	0.940	1.000	0.567	0.688
	300	65	60	0.096	0.000	0.180	0.172	0.888	1.000	0.568	0.684
	300	65	90	0.113	0.000	0.181	0.169	0.846	1.000	0.567	0.693
	C	gamma									
LS-SVR	300	450		0.177	0.190	0.225	0.264	0.623	0.629	0.326	0.256
	activation	no units									
ELM ANN	sig	389		0.223	0.250	0.242	0.275	0.398	0.361	0.222	0.192
	hardlim	389		0.241	0.269	0.258	0.282	0.298	0.256	0.120	0.151
	hardlims	389		0.241	0.269	0.258	0.282	0.298	0.256	0.120	0.151
	satlins	389		0.213	0.239	0.245	0.277	0.450	0.416	0.204	0.180
	tansig	389		0.231	0.257	0.249	0.276	0.356	0.322	0.181	0.184
	tribas	389		0.218	0.237	0.233	0.249	0.428	0.422	0.284	0.337
	relu	389		0.215	0.238	0.238	0.265	0.440	0.418	0.250	0.248
	purelin	389		0.215	0.231	0.237	0.270	0.444	0.455	0.256	0.224
	tribas	100		0.228	0.253	0.255	0.279	0.371	0.341	0.137	0.169
	tribas	200		0.223	0.247	0.236	0.264	0.397	0.376	0.264	0.257
	tribas	389		0.218	0.237	0.233	0.249	0.428	0.422	0.284	0.337
	tribas	600		0.215	0.233	0.233	0.254	0.439	0.444	0.281	0.310
	tribas	800		0.214	0.230	0.237	0.258	0.449	0.456	0.254	0.288
	tribas	1000		0.213	0.230	0.236	0.267	0.454	0.455	0.262	0.241
	tribas	2000		0.209	0.226	0.235	0.268	0.472	0.478	0.270	0.233
	tribas	3000		0.216	0.273	0.241	0.333	0.438	0.236	0.233	-0.184
	satlins	100		0.231	0.254	0.248	0.270	0.357	0.340	0.186	0.220
	satlins	200		0.225	0.255	0.240	0.286	0.389	0.333	0.237	0.126
	satlins	389		0.213	0.239	0.245	0.277	0.450	0.416	0.204	0.180
	satlins	600		0.210	0.228	0.238	0.252	0.466	0.465	0.252	0.320
	satlins	800		0.206	0.223	0.230	0.271	0.490	0.490	0.301	0.213
	satlins	1000		0.205	0.222	0.233	0.285	0.494	0.493	0.277	0.132
	satlins	2000		0.197	0.212	0.232	0.272	0.530	0.541	0.286	0.209
	satlins	3000		0.194	0.209	0.244	0.302	0.544	0.550	0.214	0.025

Table C 2: Model performance measures for MLP-ANN

layer 1				layer 2						training data		test data		training data		test data	
dropout	activation	activity regularizer	no. units	dropout	activation	no. units	optimizer	LR	epoch	RMSE	RMSE POT	RMSE	RMSE POT	R^2	R^2 POT	R^2	R^2 POT
	linear		389				ADAM	0.001	500	0.218	0.239	0.237	0.251	0.428	0.414	0.258	0.324
	linear		389				SGD	0.01	500	0.218	0.240	0.237	0.252	0.426	0.409	0.253	0.324
	linear		389				RMSPROP	0.001	500	0.220	0.246	0.238	0.256	0.414	0.378	0.248	0.297
	linear		1000				SGD	0.01	50	0.219	0.244	0.236	0.255	0.419	0.388	0.264	0.303
	linear		1000				SGD	0.01	200	0.218	0.236	0.235	0.248	0.428	0.429	0.271	0.343
	linear		1000				SGD	0.01	1000	0.216	0.233	0.232	0.252	0.435	0.441	0.288	0.322
	sigmoid		1000				SGD	0.01	200	0.229	0.254	0.242	0.266	0.367	0.339	0.223	0.242
	relu		1000				ADAM	0.001	200	0.199	0.205	0.239	0.255	0.523	0.568	0.241	0.306
	tanh		1000				ADAM	0.001	200	0.192	0.198	0.242	0.297	0.555	0.597	0.226	0.059
	elu		1000				ADAM	0.001	200	0.226	0.251	0.290	0.386	0.383	0.351	-0.113	-0.595
0.2	linear		1000	0.4	linear	1000	ADAM	0.001	200	0.243	0.263	0.262	0.278	0.290	0.289	0.092	0.173
0.2	linear		1000	0.4	relu	1000	ADAM	0.001	200	0.243	0.266	0.255	0.282	0.287	0.272	0.139	0.150
0.5	linear		1000	0.5	relu	1000	ADAM	0.001	200	0.250	0.266	0.258	0.280	0.246	0.275	0.119	0.164
	linear		1000		linear	1000	ADAM	0.001	200	0.219	0.244	0.237	0.253	0.420	0.388	0.256	0.317
	tanh		1000		tanh	1000	ADAM	0.001	200	0.227	0.205	0.252	0.230	0.376	0.570	0.156	0.437
	tanh		1000				SGD	0.01	200	0.192	0.198	0.242	0.297	0.555	0.597	0.226	0.059
	linear	Lasso	1000				ADAM	0.001	3	0.234	0.285	0.247	0.293	0.341	0.168	0.189	0.080

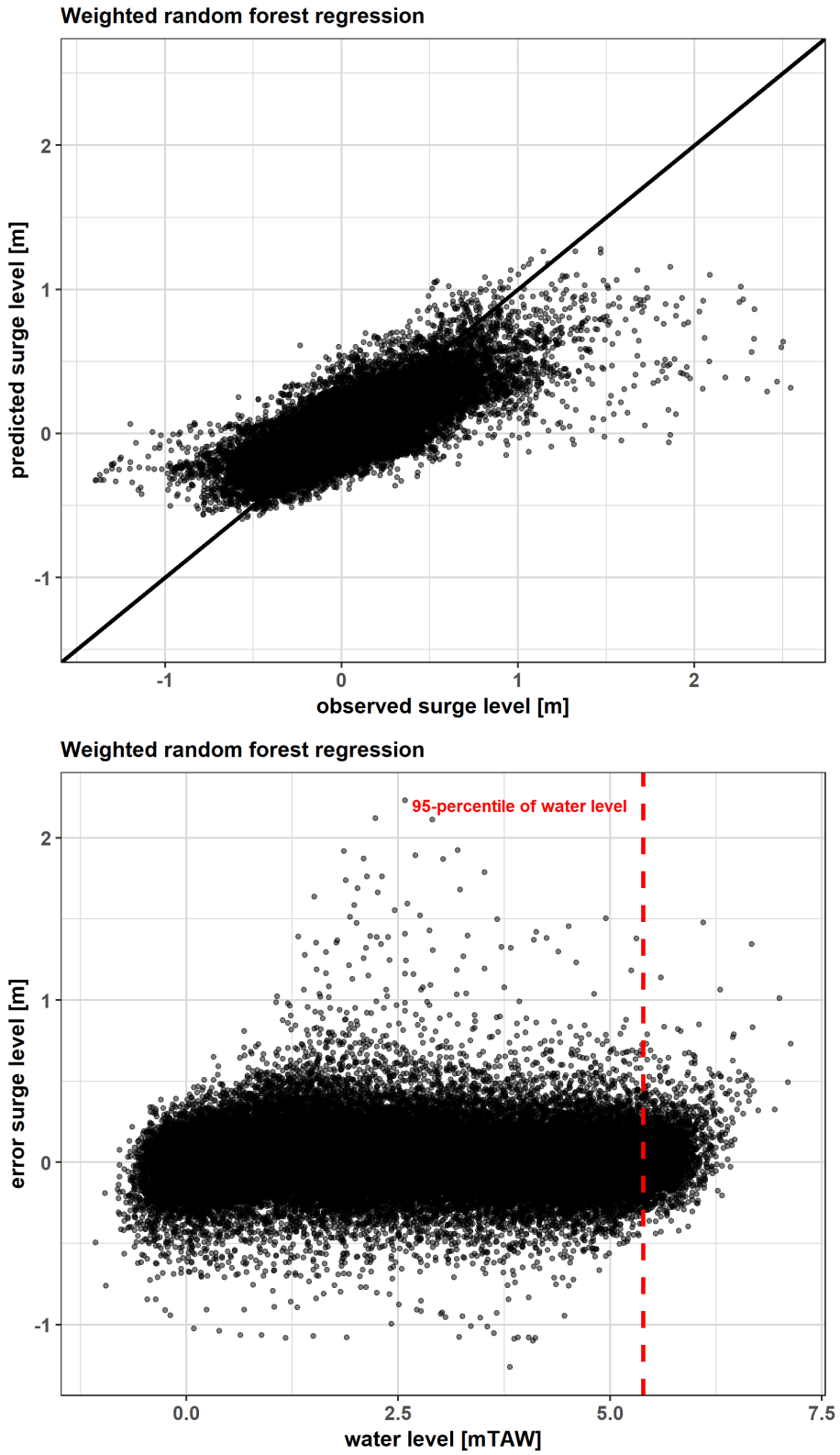


Figure C 1: Examination of observed vs predicted surge level (top), and surge errors as function of water level (bottom) for the 24-hours ahead weighted random forest forecast model (weight = 30)

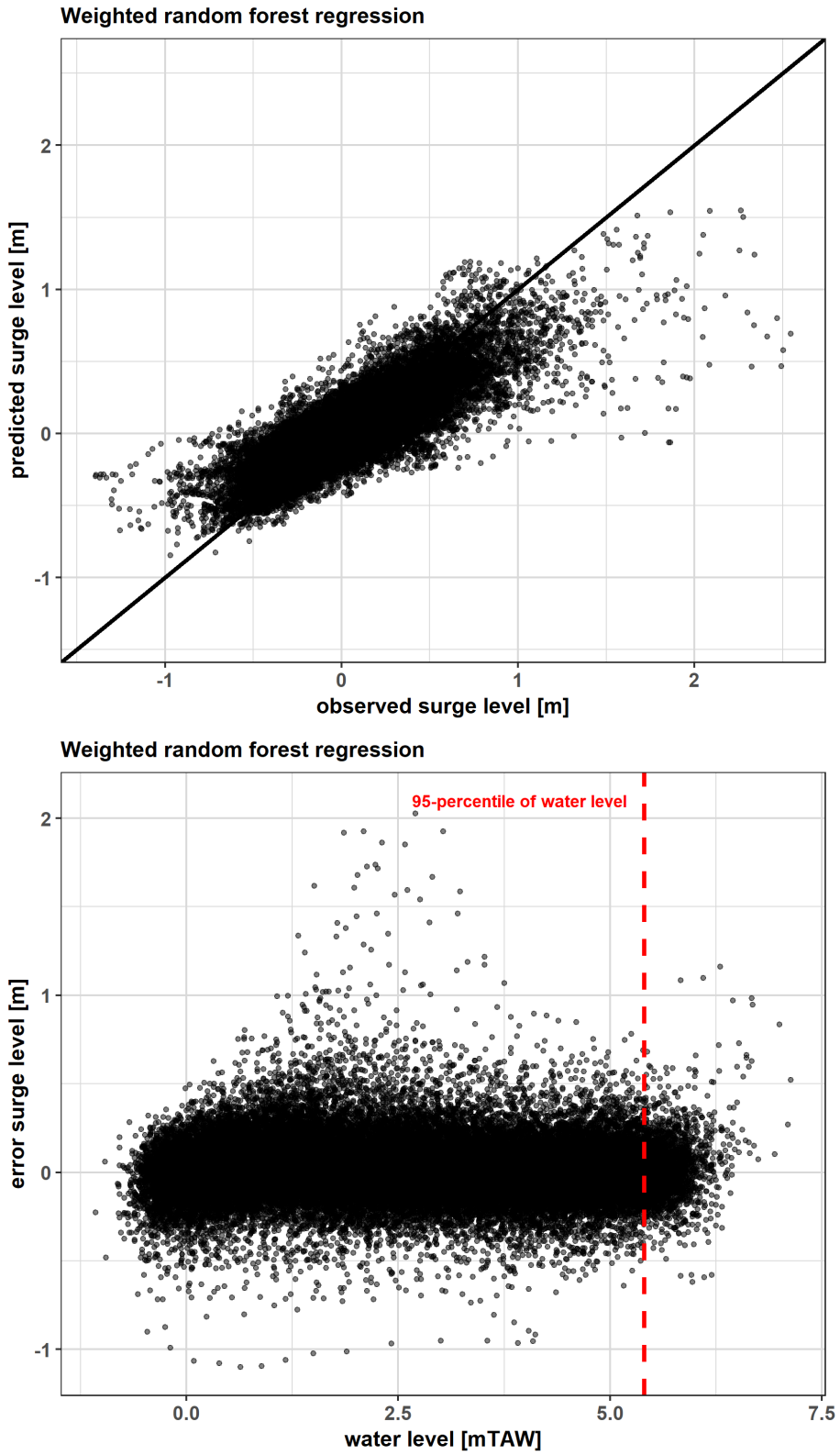
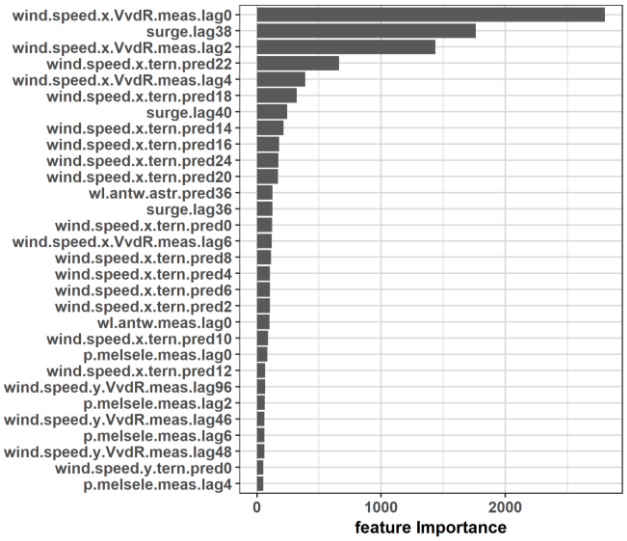
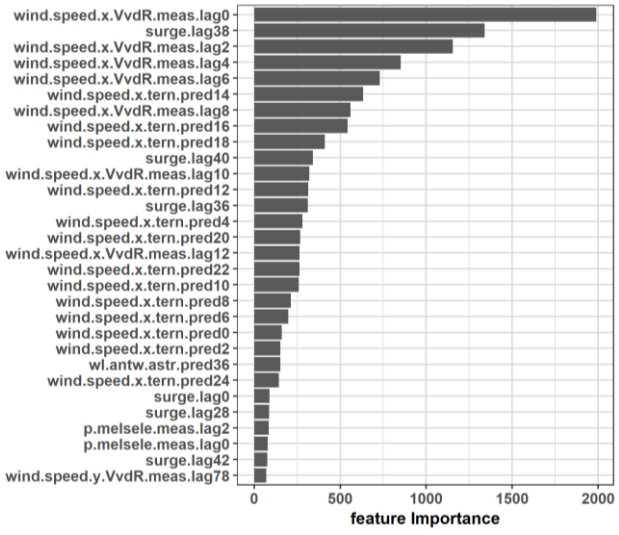


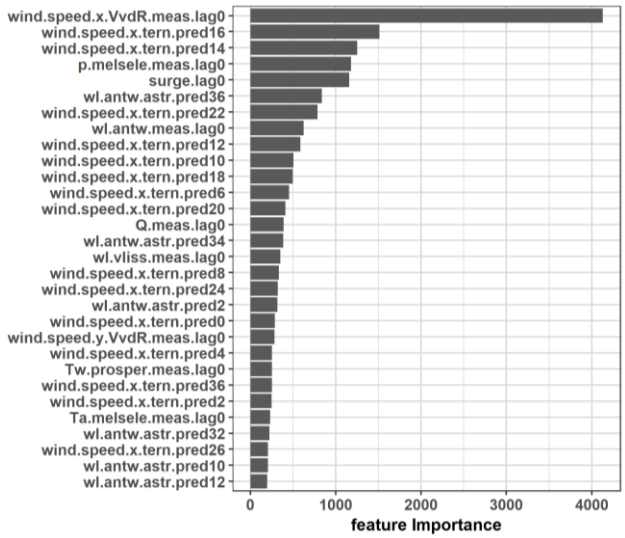
Figure C 2: Examination of observed vs predicted surge level (top), and surge errors as function of water level (bottom) for the 6-hours ahead weighted random forest forecast model (weight = 30)



42-hours lag horizon



13-hours lag horizon



0-hours lag horizon

Figure C 3: Feature importance for weighted random forest 6-hours ahead forecast model with several lag horizons: 42 hours (top), 13 hours (middle) and 0 hours (bottom). The feature names end with the number of time steps X in the lag horizon (cf. *lagX*) or forecast horizon (cf. *predX*)

APPENDIX D

CONFORMAL INFERENCE

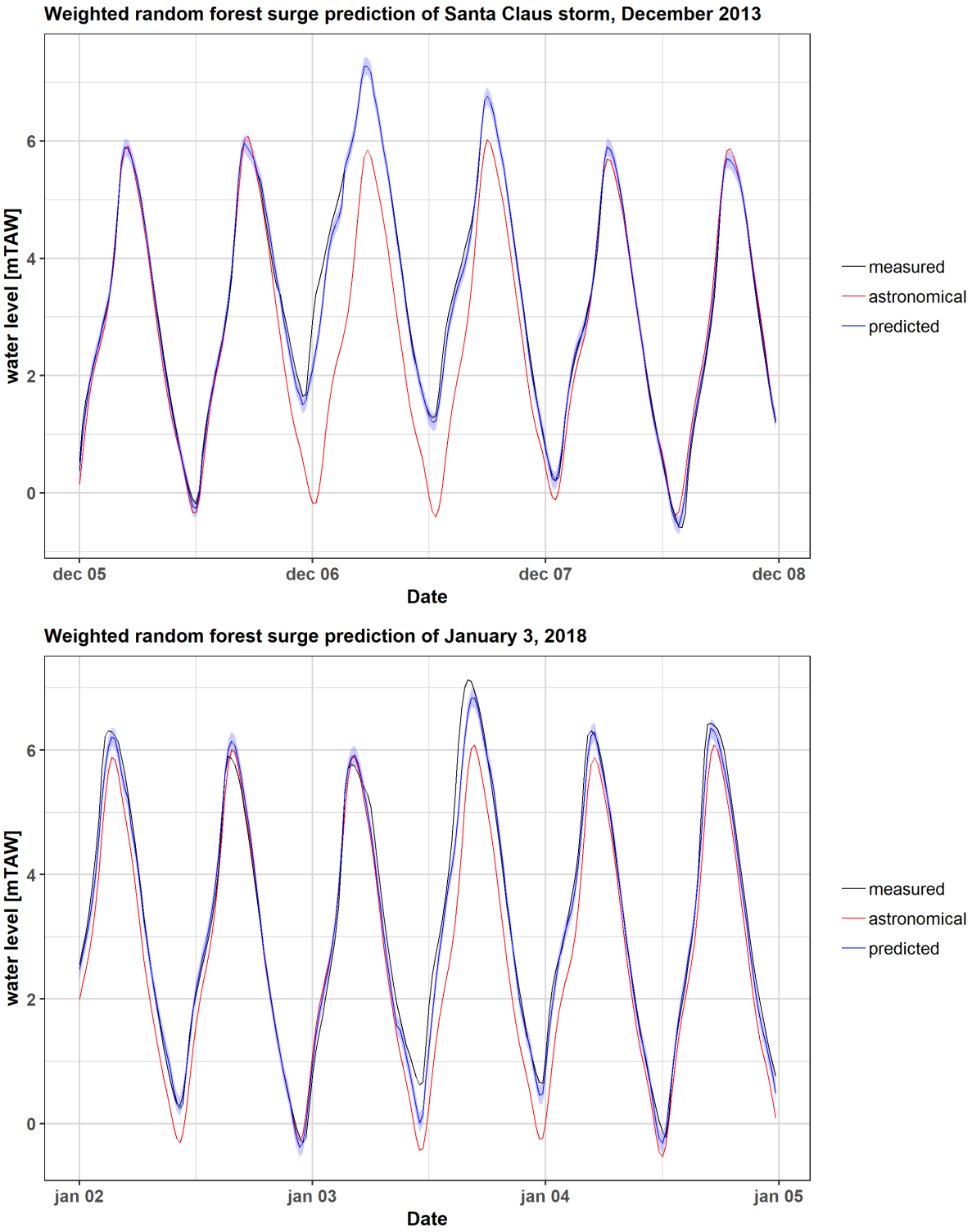


Figure D 1: Original conformal prediction intervals for two extreme events 6-hours ahead forecasted: December 2013 (top) and January 2018 (bottom)

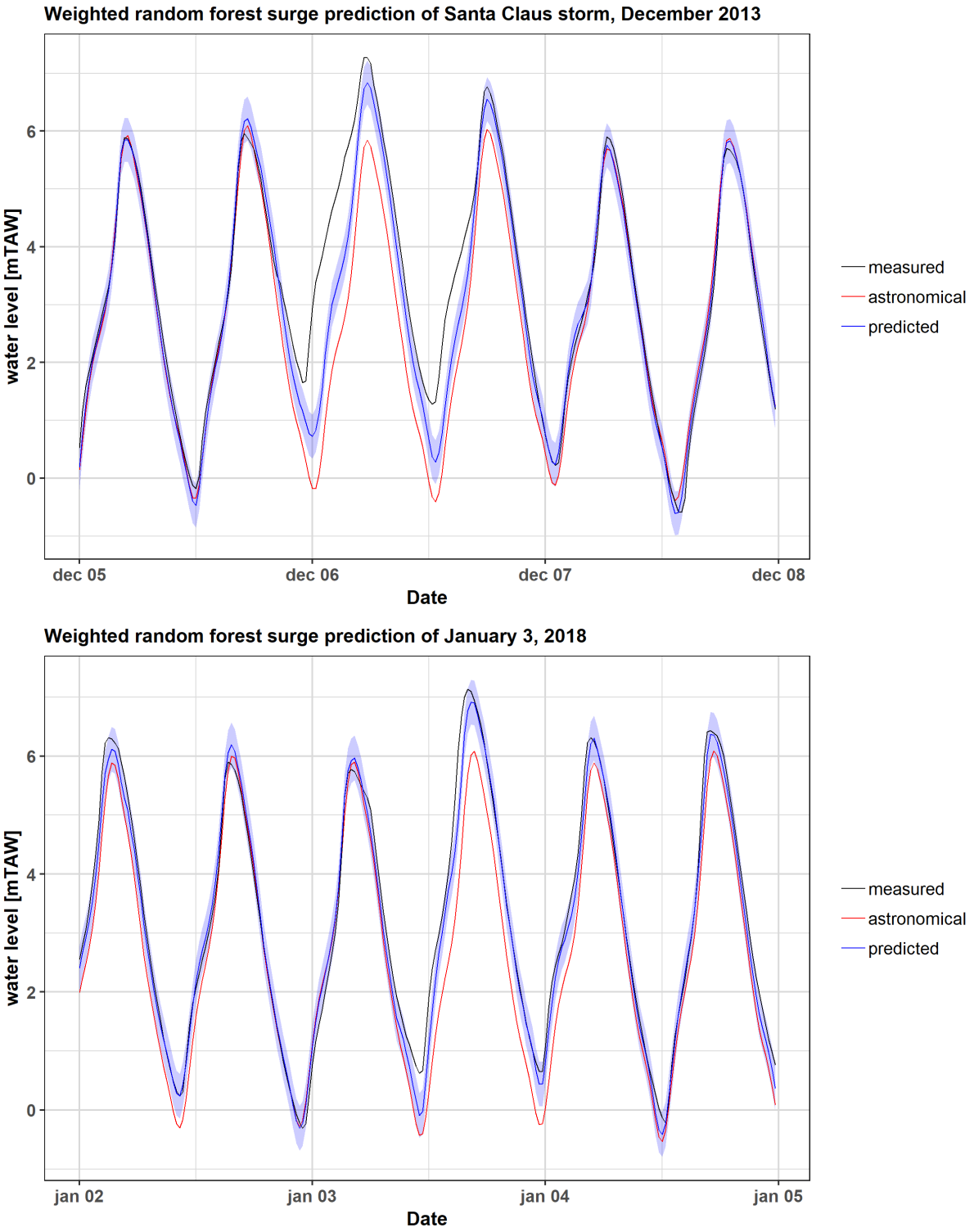


Figure D 2: Split conformal prediction intervals for two extreme events 6-hours ahead forecasted: December 2013 (top) and January 2018 (bottom)

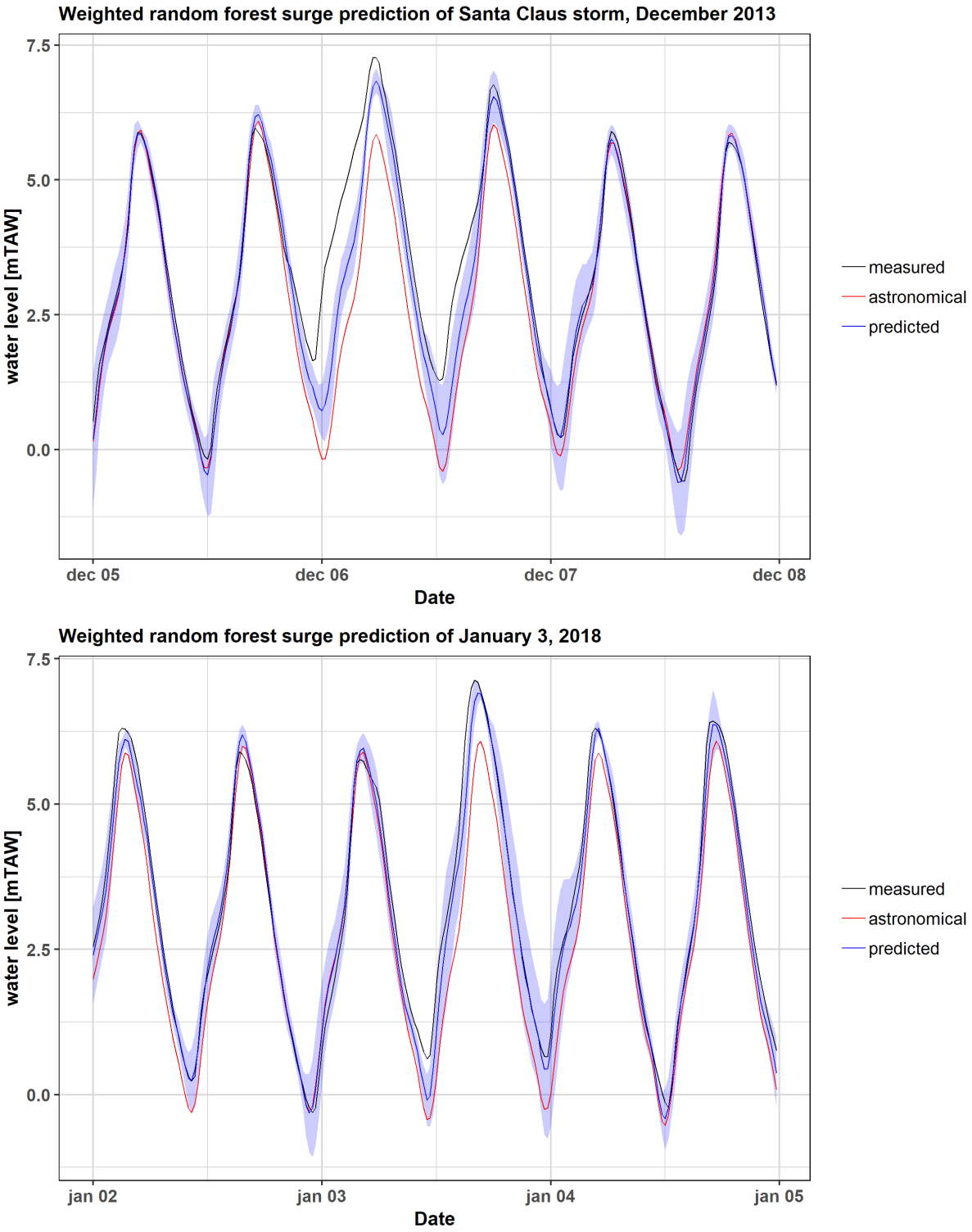


Figure D 3: Locally-weighted split conformal prediction intervals for two extreme events 6-hours ahead forecasted: December 2013 (top) and January 2018 (bottom)

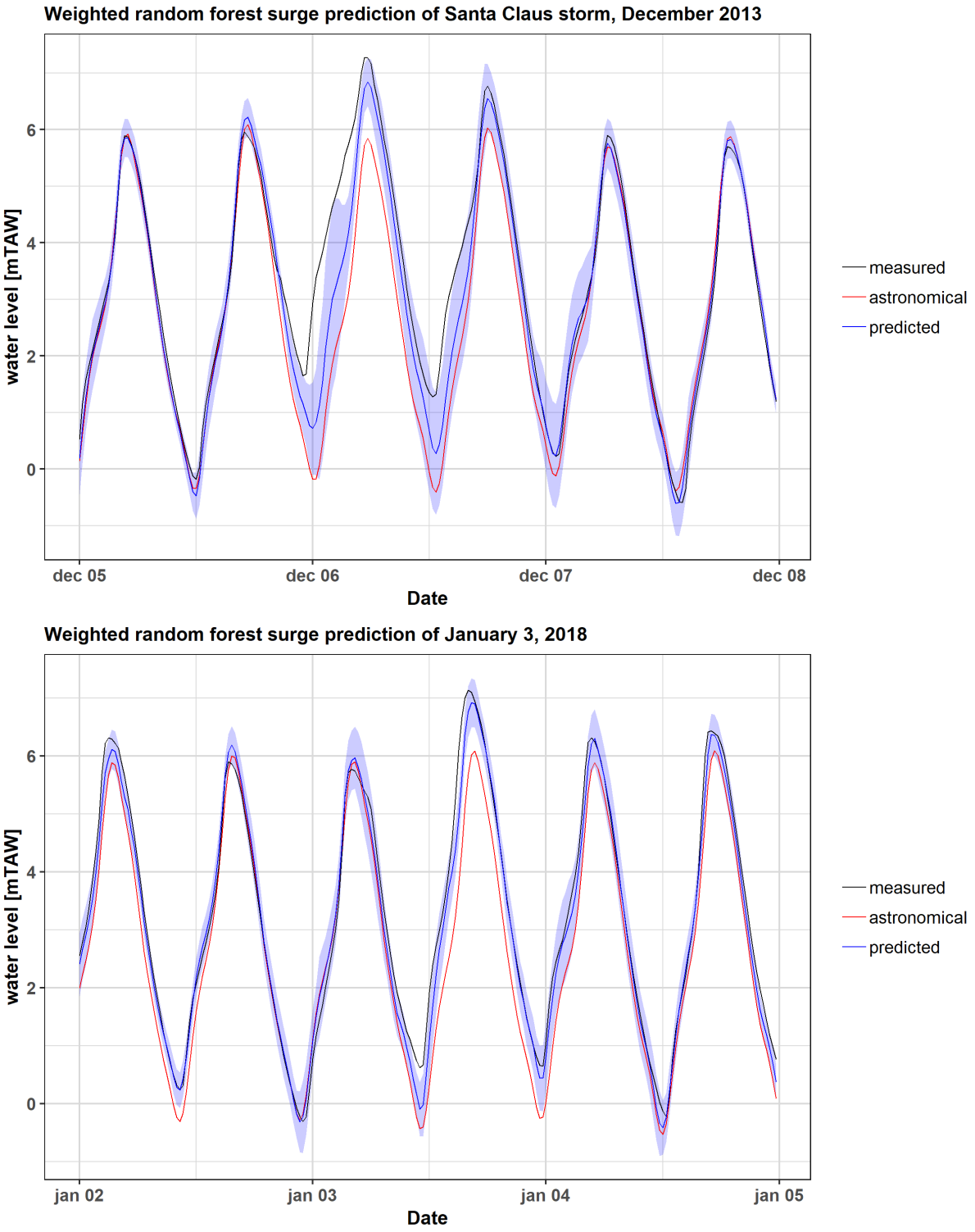


Figure D 4: Local split conformal prediction intervals for two extreme events 6-hours ahead forecasted: December 2013 (top) and January 2018 (bottom)

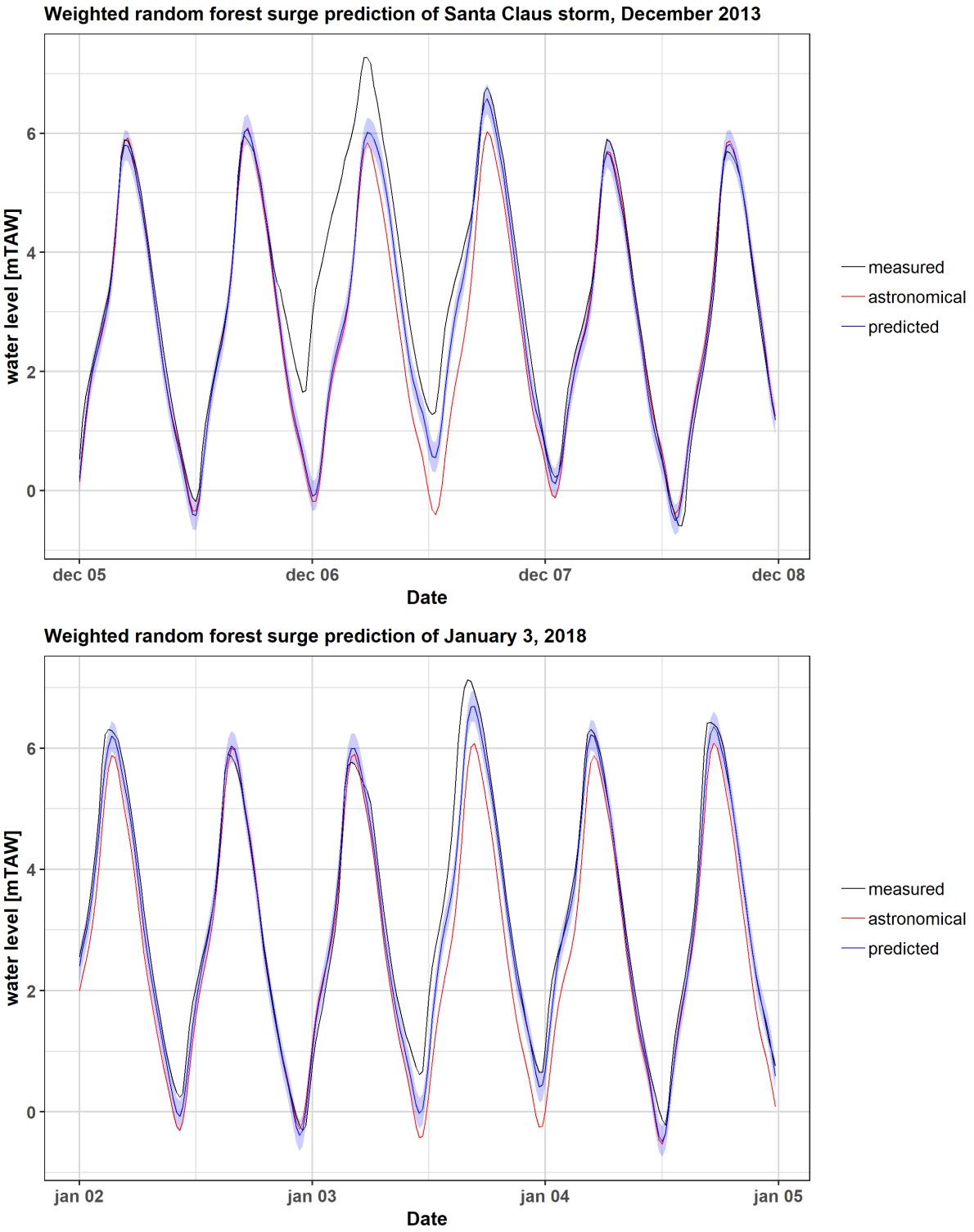


Figure D 5: Original conformal prediction intervals for two extreme events 24-hours ahead forecasted: December 2013 (top) and January 2018 (bottom)

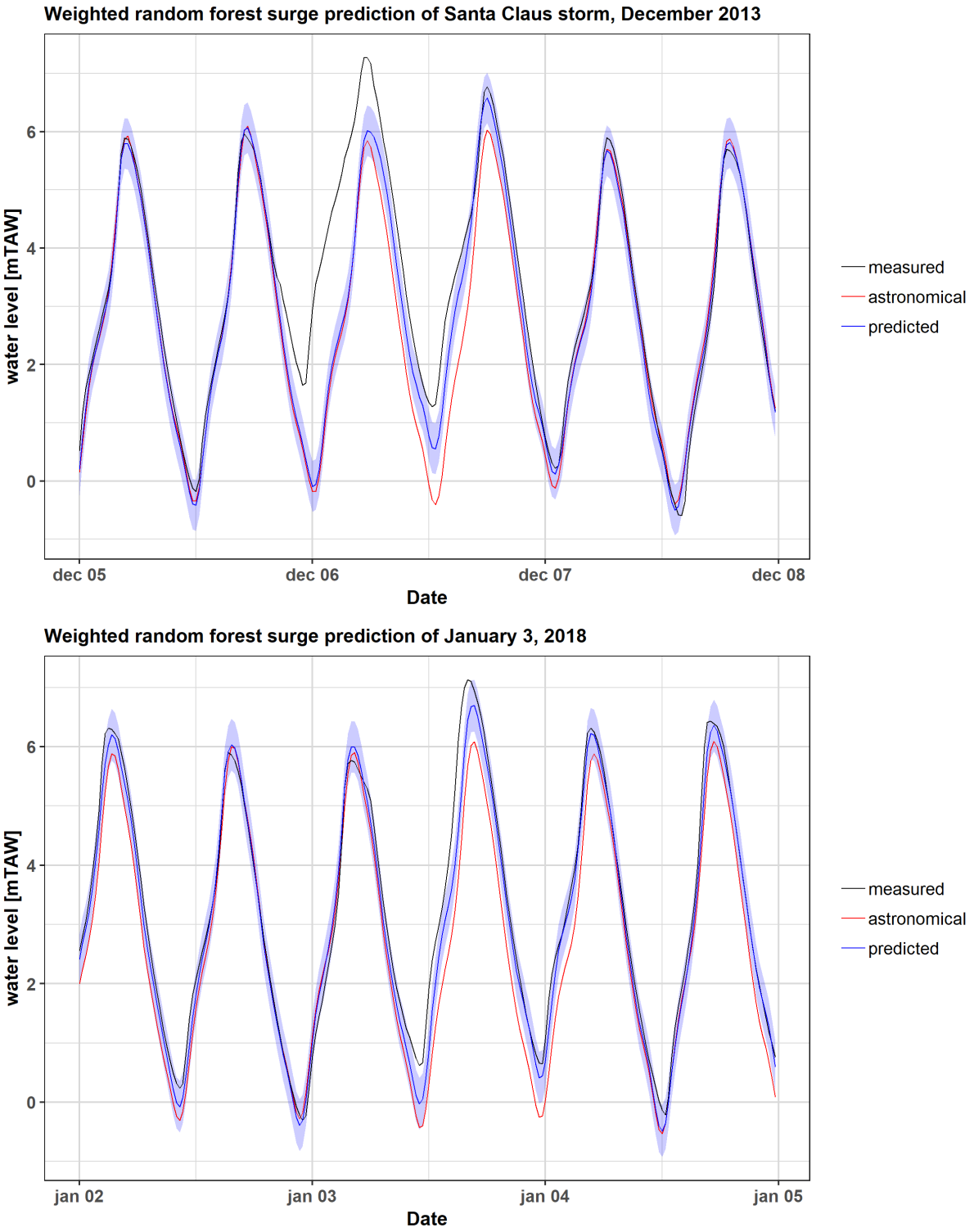


Figure D 6: Split conformal prediction intervals for two extreme events 24-hours ahead forecasted: December 2013 (top) and January 2018 (bottom)

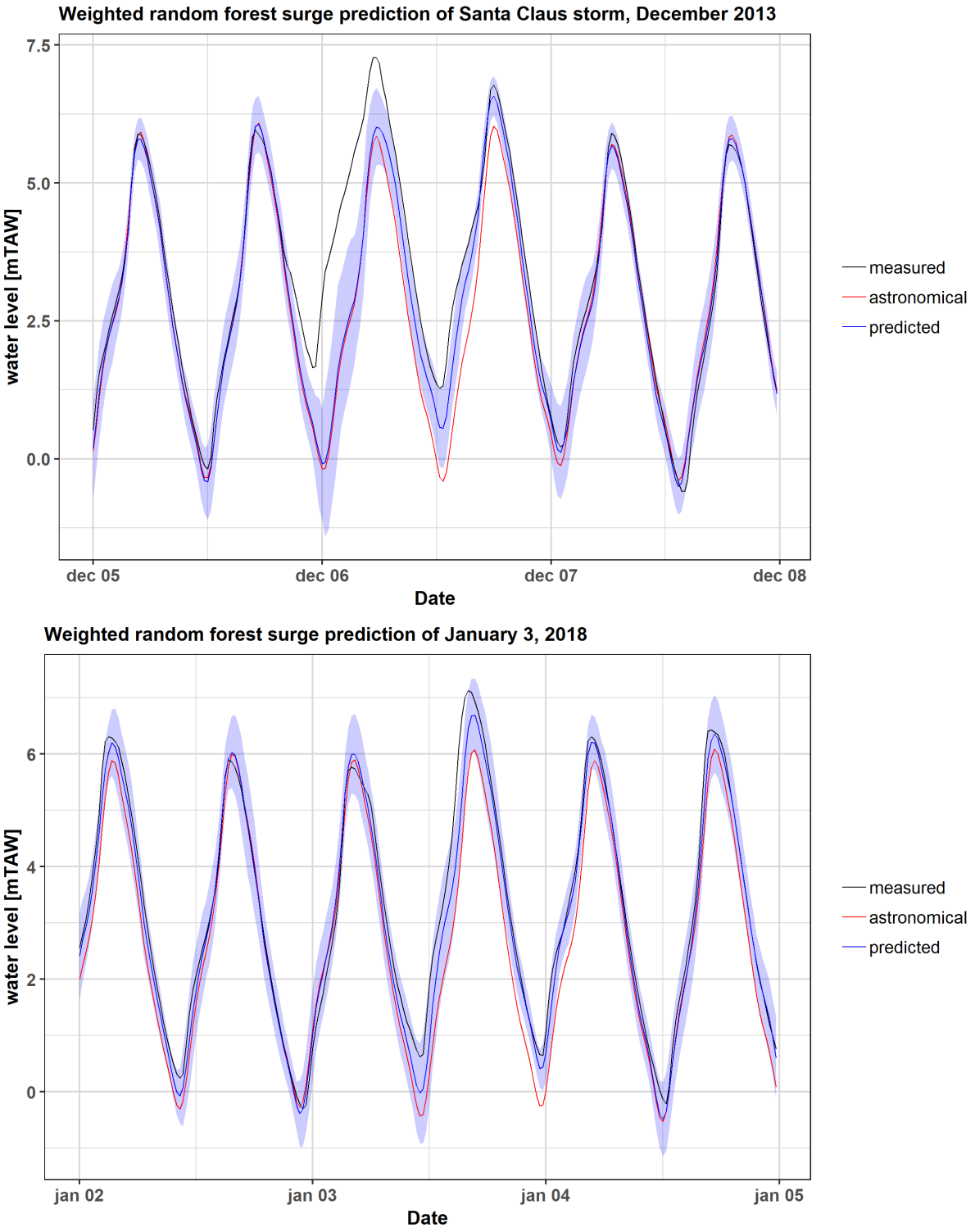


Figure D 7: Local split conformal prediction intervals for two extreme events 24-hours ahead forecasted: December 2013 (top) and January 2018 (bottom)