https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

Using Random Forests to Predict Extreme Sea-Levels at the Baltic Coast at Weekly Timescales

Kai Bellinghausen[1], Birgit Hünicke[1], and Eduardo Zorita[1]

[1]Institute for Coastal System Analysis and Modelling, Helmholtz-Zentrum Hereon

Correspondence: Kai Bellinghausen (kai.bellinghausen@hereon.de)

Abstract.

We have designed a machine-learning method to predict the occurrence of daily extreme sea-level at the Baltic Sea coast with lead times of a few days. The method is based on a Random Forest Classifier and uses spatially resolved fields of daily sea level pressure, surface wind, precipitation, and the prefilling state of the Baltic Sea as predictors for daily sea level above the 95% quantile at each of seven tide-gauge stations representative of the Baltic coast.

The method is purely data-driven and is trained with sea-level data from the Global Extreme Sea Level Analysis (GESLA) data set and from the meteorological reanalysis ERA5 of the European Centre for Mid-range Weather Forecasting.

Sea-level extremes at lead times of up to 3 days are statisfactorily predicted by the method and the relevant predictor regions are identified. The sensitivity, measured as the proportion of correctly predicted extremes is, depending on the stations, of the order of 70%. The proportion of false warnings, related to the specificity of the predictions, is typically as low as 10 to 20%. For lead times longer than 3 days, the predictive skill degrades; for 7 days, it is comparable to a random skill. These values are generally higher than those derived from storm-surge reanalysis of dynamical models.

The importance of each predictor depends on the location of the tide gauge. Usually, the most relevant predictors are sea level pressure, surface wind and prefilling. Extreme sea levels in the Northern Baltic are better predicted by surface pressure and the meridional surface wind component. By contrast, for stations located in the south, the most relevant predictors are surface pressure and the zonal wind component. Precipitation was not a relevant predictor for any of the stations analysed.

The Random Forest classifier is not required to have considerable complexity and the computing time to issue predictions is typically a few minutes on a personal laptop. The method can, therefore, be used as a pre-warning system triggering the application of more sophisticated algorithms to estimate the height of the ensuing extreme sea level or as a warning to run larger ensembles with physically based numerical models.

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

1 Introduction

Storm surges are extreme and short-lived increases in sea level mainly induced by extreme atmospheric conditions of wind (e.g. storms) and low-pressure systems (Wolski and Wisniewski, 2021; Field et al., 2012; WMO, 2011; Weisse and von Storch, 2010; Harris, 1963). They are a major natural hazard for coastal societies as they not only can cause severe damage to infrastructure at coastlines but also the loss of human lives. Hence, monitoring and forecasting systems for storm surges are important to prevent societal damage and inform decision-makers. This study explores the possibility of short-term predictions (lead time of a few days) of storm surges in the Baltic Sea using a purely data-driven machine-learning approach. Technically, the storm-surge problem is an air-sea interaction problem, where the atmosphere forces the water body, not necessarily directly at the coast, which in turn responds with oscillations of the water level at various frequencies and amplitudes. While the atmosphere and its wind-field influence the currents and wave dynamics of the sea, the currents in turn influence the wave dynamics, which in turn may alter the wind field (Gönnert et al., 2001). Hence, the underlying processes of storm surges are highly non-linear and often non-local, which makes predicting them a complex problem.

Operational forecasting systems of storm surges rely on numerical dynamical ocean-atmosphere models (WMO, 2011; Gönnert et al., 2001). In the Baltic Sea, a few regional models are in operation, like the BSHcmod from the Bundesamt für Schifffahrt und Hydrographie (BSH), which is a hydrostatic ocean circulation model. While those dynamical models generate reasonable estimations for general water level elevations, they often underestimate extreme (storm surge) events (Muis et al., 2016; Vousdoukas et al., 2016). This is explained by an insufficient grid-resolution (Muis et al., 2016), which leads to a mis-representation of e.g. wind fields (WMO, 2011) and extratropical cyclones (Rutgersson et al., 2022), as well as the underlying ocean bathymetry. Furthermore, the effect of mesoscale weather systems is not well represented in current storm surge models as no meteorological networks provide data at these spatial scales. (WMO, 2011). Usually, the data of meteorological fields are interpolated in time and fed to the ocean model (von Storch, 2014), which may lead to too smooth short variability of the atmospheric forcing, which in turn may result in extreme events being underrepresented in the simulations. According to Muis et al. (2016), the underestimation of extreme events can also be explained by the insufficient or missing non-linear coupling between storm-surge-relevant processes in dynamical models.

Alternatively to dynamical models, forecasting methods can be based on data-driven algorithms. These algorithms are not based on equations representing the physical dynamics but, instead, try to identify the relevant predictor patterns in a data set that appear associated with a specific predictand. This is achieved by analyzing observational data sets of the forcing (atmospheric and/or oceanic) and of the response (storm surge). This makes them computationally more efficient than dynamical models (Harris, 1962), at the expense of being a method oblivious of the underlying physical mechanisms and often more difficult to interpret. Besides the classical statistical methods based on simplified statistical models of the underlying processes, Machine Learning (ML) is one example of data-driven algorithms and is becoming more popular in climate sciences. ML algorithms are usually more complicated than classical statistical methods and do not attempt to explicitly represent physical processes but rather try to identify recurring patterns in the data that may be used for predictions. Those complex and not obvious links between predictors and predictands contribute to their growing application. However, this very complexity makes them more

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

difficult to interpret than classical methods. Also special care is therefore needed to avoid statistical pitfalls, such as overfitting. Several studies applied ML-methods in order to analyse and predict storm surges with promising results (Tiggeloven et al., 2021; Bruneau et al., 2020; Tadesse et al., 2020; Gönnert and Sossidi, 2011; Sztobryn, 2003). Statistical and machine learning models were compared when simulating daily maximum surges on a quasi-global scale based on either remotely sensed predictors or predictors obtained from reanalysis products like ERA5-Interim data (Tadesse et al., 2020). The storm surge predictand was derived from two data sets, the observed hourly sea level data from the GESLA-2 database and other *in situ* data of daily maximum surges. They compared linear regression models to a machine learning method called Random Forest (RF)s. The authors found that data-driven models work well in extratropical regions, e.g. the Baltic Sea, and that the ML methods generally performed better than linear regression. Storm surge prediction on a global scale has also been the focus of several ML-models Bruneau et al. (2020). They show that ML – in this case Artificial Neural Network (ANN)s – reconstructed storm surges with significant skill, but still struggled to represent the strongest extreme events. Bruneau et al. (2020) explained this by unavoidable limitations of the training data, as extreme events are only a small fraction of the available data set. Because ANNs are trained with a procedure that is ill-designed for outliers and biased towards the representation of the average dynamics, extreme surges are difficult to be reliably reproduced. Tiggeloven et al. (2021) use a variety of deep learning methods, a subbranch of ML, to investigate storm surges at 736 tide stations globally. The overall result showed that ML approaches to capture the temporal evolution of surges and outperform a large-scale hydrodynamic model. However, extreme events were underestimated due to similar reasons as found by Bruneau et al. (2020).

Most approaches using ML-methods are global and, hence, lack specificity for the Baltic Sea basin. The only study (to our knowledge) that applied ANNs specifically to the Polish coast of the Baltic Sea was undertaken by Sztobryn (2003), using preceding mean sea level as well as wind speed and wind direction as predictors of high water levels. She showed that neural networks can be successfully integrated into operational forecast services, possibly reducing their average error. Similar to the global studies, the study by Sztobryn (2003) showed an underestimation of extreme water levels. Altogether, a thorough application of ML to predict extreme storm surges at several tide-gauging stations in the Baltic Sea is missing in the current literature. Hence, we will render a simple RF to the specific storm-surge drivers of the Baltic Sea in order to predict extreme storm surges defined by the top five per cent highest measurements of sea-level taken from the Global Extreme Sea Level Analysis (GESLA)3-project (Haigh et al., 2021). The Baltic Sea is known for broad coverage by atmosphere and ocean measurements (Rutgersson et al., 2022), thus being a very good testbed for ML-models.

For the reader that is unfamiliar with the Baltic Sea, we will introduce its specific characteristics when looking at storm surge events. In Section 2 we will further specify the underlying datasets of this study as well as their preprocessing. In Section 3 the model architecture is presented and the basic principles of a RF are discussed. Furthermore, we will specify how the model was tuned and evaluated. In Section 4, we describe all conducted experiments and their rationale, while Section 5 summarizes their results. We end the study with a discussion and conclusion.

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

1.1 Specific characteristics of the Baltic Sea

Apart from the atmospheric forcing, the amplitude of storm surges also substantially varies with specific local conditions like the topography of the ocean basin, the extent of ice cover, as well as the direction of the storm track crossing the basin and the shape and orientation of the coastline (Muis et al., 2016; WMO, 2011; Weisse and von Storch, 2010; Gönnert et al., 2001).

Hence, understanding the local characteristics of the Baltic Sea is necessary when building and interpreting a storm-surge model. In the following paragraphs, we provide a brief background of the main physical processes that lead to storm surges in the Baltic Sea for the reader unfamiliar with this region. The Baltic Sea is a semi-enclosed intracontinental sea of the Atlantic Ocean that ranges from around 10°E - 54°N to 29°E - 65°N in Northern Europe (Weisse and Hünicke, 2019) as depicted in Fig. 1. It is connected to the North Sea and thus the Atlantic via the Straits of Denmark and the Kattegat. This connection plays an important role in the context of storm surges and tides. The Straits of Denmark block tidal waves and allow mainly internal tides of only a few centimetres within the Baltic Sea (Rutgersson et al., 2022; Wolski and Wisniewski, 2021). Due to the very narrow connection to the Atlantic, storm surges are only internally induced (Weisse and Hünicke, 2019). The risk of storm surges considerably depends on the location due to the large meridional extent of the Baltic Sea and the different orientation of coastlines in combination with trajectories of pressure systems and wind directions (Hünicke et al., 2015; Weisse, 2014; Rutgersson et al., 2022; Wolski and Wisniewski, 2020; Holfort et al., 2014).

Seasonally, the strongest increase in water levels is expected from September to February. Those winter half-year surges are mainly driven by processes that alter the volume of the Baltic Sea, e.g. prefilling, and by the ones that redistribute internal water masses of the basin, e.g. effects of wind (Weisse and Hünicke, 2019; Weisse, 2014; Hünicke and Zorita, 2006; Chen and Omstedt, 2005).

Due to the Baltic Sea's semi-enclosed basin, specific drivers of storm surges are added to the general drivers like wind stresses and atmospheric pressure. In the following sections, we provide a brief overview of those drivers.

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

Figure 1. Subbasins of the Baltic Sea as indicated in Wolski and Wisniewski (2020).

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

1.1.1 Wind-Effect

Storm surges generated by the impact of wind stress are called *wind-driven storm surges*. If a wind blows consistently over several days, it deforms the sea surface and causes drift currents and wind setup, which eventually lead to a storm surge (Wolski and Wisniewski, 2021; Harris, 1963). Wind conditions in the Baltic Sea are mainly governed by the Westerlies and the cyclonic activity in the Northern Europe-Baltic Sea area. This is true especially during the winter months, where the winds are blowing (on average) from south-western directions (Weisse, 2014; Leppäranta and Myrberg, 2009). In periods when the strong westerlies weaken or stop blowing, the elevated sea surface in the northeastern parts of the Baltic Sea relaxes and water masses flush back towards the southern and southwestern coasts. These seiches raise the water levels in the corresponding coasts (Weisse and von Storch, 2010). Furthermore, south-westerly winds, if maintained for several days, can cause a strong inflow of water masses into the Baltic Sea via the Straits of Denmark, leading to a condition of prefilling (Gönnert et al., 2001). Hence, the wind direction is an important indicator for the onset of storm surges at specified coastlines (Andrée et al., 2022; Wolski and Wisniewski, 2021).

1.1.2 Atmospheric pressure

In the Baltic region, low-pressure systems are mostly associated with regions of less than 980 hPa (Wolski and Wisniewski, 2021; Holfort et al., 2014). Those low-pressure systems lift up the sea surface by the *inverted barometer effect* (Weisse and von Storch, 2010), which eventually induce a *baric wave* travelling along the trajectory of the system (Wolski and Wisniewski, 2021). For instance, in hydrostatic equilibrium, a drop in surface air pressure of 1 hPa lifts the sea level by about 1 cm (Wolski and Wisniewski, 2021; Harris, 1963). As low-pressure systems in the Baltic Area usually move from the (South-)West towards the (North-)East during winter, the water surface is more frequently elevated in the North and depressed in the South (Wolski and Wisniewski, 2021). Wind and pressure combined may amplify the storm surge and increase its intensity, or they may oppose each other out and decrease the severity of the storm surge (Wolski and Wisniewski, 2021).

1.1.3 Prefilling of the Baltic Sea

The changing total volume of the Baltic Sea is also important for storm surges. The Baltic Sea contains an averaged volume of 20.900 km^3 (Eakins and Sharman, 2010) that is constantly altered due to different in and outflows (Weisse and Hünicke, 2019). The main inflow is the saltwater exchange of the North Sea and the Baltic Sea via the Straits of Denmark, which is approximately 1180 $km^3 a^{-1}$ (Leppäranta and Myrberg, 2009). On a daily basis, up to 45 km^3 are exchanged between the basins in both directions. Evenly distributing this water mass over the whole Baltic Sea would correspond to a sea level change of 12 cmd^{-1} (Mohrholz, 2018)).

If net water exchange persists over longer periods, mean Baltic sea-level can raise or drop accordingly by larger amounts. If the water level of the Baltic Sea is elevated 15 cm above the mean sea level for more than twenty consecutive days due to increased inflow via the Straits of Denmark, Mudersbach and Jensen (2010) speak of a *prefilling* or *preconditioning* of the Baltic Sea. Usually, water levels at the tide gauging stations in Landsort (Sweden) or Degerby (Finland) are used as proxies

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

for measuring prefilling (Weisse, 2014; Janssen et al., 2001). With a high degree of prefilling, storm surges can become more likely and extreme as less wind is needed to induce wind setup (Weisse and Weidemann, 2017; Weisse, 2014). It is mainly the already mentioned south-westerly wind direction that, when blowing over extended periods, leads to an increased inflow of water masses to the Baltic Sea through the Kattegat (Wolski and Wisniewski, 2021; Hünicke et al., 2015; Weisse, 2014). But also a sequence of fast-moving low-pressure systems coming from the West and travelling to the North-East of the Baltic Sea can result in strengthened inflows (Wisniewski and Wolski, 2011). According to Leppäranta and Myrberg (2009) peak months of inflow are during winter, especially from November to January. Combined with the effects of stronger winds and rainfall in winter, the preconditioning is an important driver of storm surges.

1.1.4 Precipitation

Finally, when low-pressure systems and corresponding cyclones move over the Baltic Sea, they usually bring precipitation along (Leppäranta and Myrberg, 2009; Harris, 1963). Extreme precipitations associated with low-pressure systems are most frequent in winter (Rutgersson et al., 2022). As stated by Weisse and Hünicke (2019), heavy precipitation increases the total volume of the Baltic Sea and changes the density due to a change in salinity profiles, which combined may lead to an increased overall water level. Therefore, the influence of precipitation is not directly related to storm surge magnitudes but rather alters preconditions like the prefilling of the Baltic Sea and the filling of rivers and estuaries (Gönnert et al., 2001). Hence, indirect effects of precipitation combined with the onset of a storm surge can lead to severe compound floodings in the Baltic Sea, especially in low-lying coastal areas (Rutgersson et al., 2022; Bevacqua et al., 2019).

2 Data

The Baltic Sea provides one of the densest tide-gauge networks with records starting in the 19th century (Hünicke et al., 2015), which is part of the record compilation of the Global Extreme Sea Level Analysis (GESLA) dataset.

2.1 Area of research

The investigated area ranges from 5°W to 30°E and 40°N to 70°N as depicted in Fig. 2 and includes the Baltic Sea (BS). We intentionally selected a broad region around the BS to account for the non-local links between the drivers of storm surges and the locality of the event itself. More specifically, seven stations were selected for model analysis. These stations are part of the GESLA data set. Station codes are provided in Table A1. This set of stations should be representative of the coastal orientations and bays of the Baltic Sea.

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

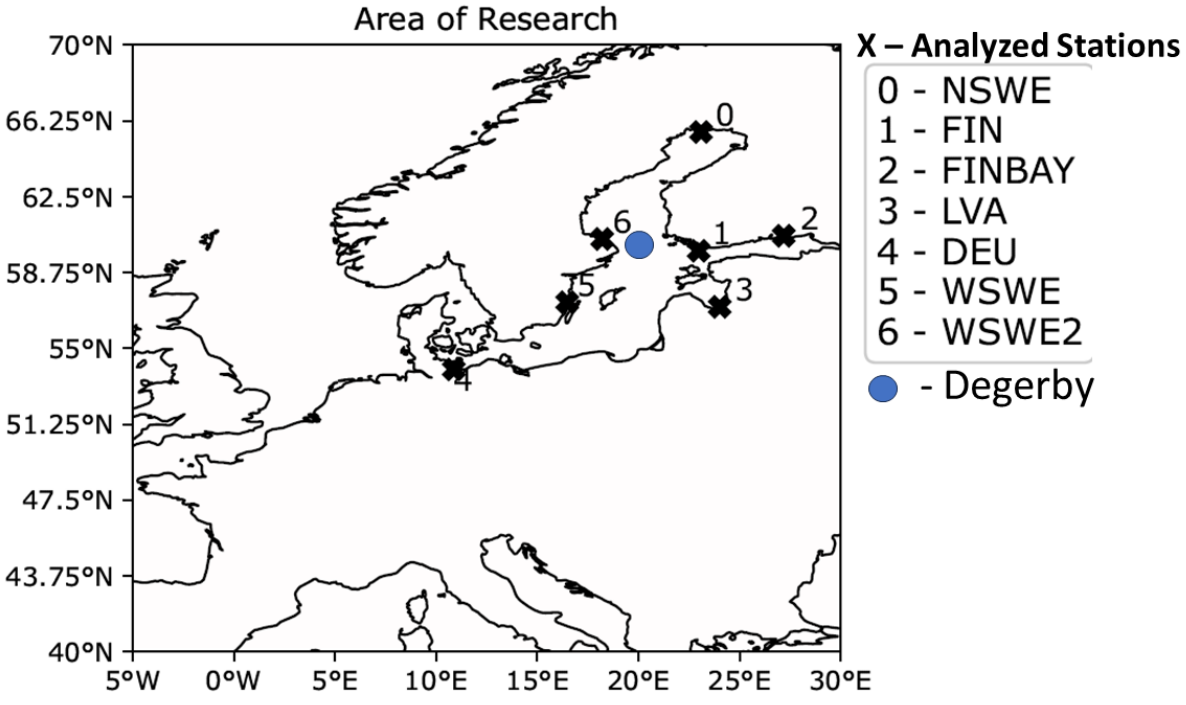


Figure 2. Map of whole research area. Black crosses and numbers indicate the analysed stations (predictand) within the Baltic Sea. Closest cities to the stations are Kalix (0), Hanko (1), Hamina (2), Riga (3), Travemünde (4), Oskarshamn (5) and Forsmark (6). The blue circle indicates the position of station Degerby, which is used as a proxy of the predictor prefilling.

.

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

2.2 Predictand

The GESLA dataset provides a global set of high frequency (at least hourly) sea level data with integrated quality control flags (Haigh et al., 2021). Height units of all stations were converted to metres and the time zone was adjusted to Coordinated Universal Time (UTC). A more thorough description of the compilation can also be found in Woodworth et al. (2016) and Haigh et al. (2021). The data is publicly accessible at http://www.gesla.org.

All stations we selected for model analysis contain hourly data, covering the period from 1960 to 2020. This sea level data is later used to derive a daily timeseries of a categorical binary predictand at respective stations after preprocessing (see Section 2.4). The categories of the predictand are either *no occurence of a storm surge* (0) or an *occurence of storm surge* (1).

2.3 Predictors

We use spatially resolved fields of daily total precipitation, daily mean wind fields (zonal and meridional), daily mean sea-level-pressure from the European Re-Analysis (ERA5) data provided by the European Centre for Medium-Range Weather Forecasts (ECMWF) and hourly or daily prefilling of the BS, depending on the combination. The ERA5 dataset ranges from 1959 to present with hourly estimates of atmospheric variables and is spatially resolved on a 30km (approximately 0.27 degrees) grid covering the Earth (Guillory, 2017). We select the period from 1999 to 2020 for this study, and an area that broadly encompasses the Baltic Sea, North Sea and part of the Eastern North Atlantic, which should include the main known drivers of Baltic storm-surges. All variables of ERA5 used as predictors are shown and briefly described in Table A2. They are surface pressure (SP), total precipitation (TP), eastward wind at 10m height (U10), northward wind at 10m height (V10). Each variable is extracted from the two-dimensional field depicted in Fig. 2. Additionally, we implemented a predictor of prefilling by using the GESLA timeseries of sea-level data at the station of Degerby. The station is situated at about 60°N and 20.38°E (see blue circle in Fig 2). The hourly waterlevel at the station of Degerby from the GESLA dataset is used as a proxy for prefilling. If PF is combined with ERA5 predictors, the daily time-frequency of ERA5 is used. Hence, we reduced the hourly waterlevels to daily data by using the maximum recorded water level of a given day as an entry. If PF is used as a sole predictor, the time-frequency is kept hourly.

2.4 Preprocessing

When preprocessing data of the predictand, we define the day of occurrence of storm surge as those days in which at least one hourly reading is above the 95% percentile of the pre-processed distribution of hourly sea-level (see Fig. 3). This definiton is station-dependent, as the 95% percentile depends on each station. To derive a daily time series of storm surge occurrence, we first select data labelled by the GESLA project as *analysis data* and the season of interest, e.g. winter months DJF. In order to obtain a *stationary process* we temporally detrended the timeseries by subtracting a linear trend in time. We then classify the data into storm surges (1) and no storm surges (0) by using the 95th-quantile q as a threshold. We set every entry of a sea level strictly below q to 0 (no storm surge) and all remaining entries to 1 (storm surge). Hence, we obtain a timeseries of hourly temporal resolution for each station, where 5% of the data is (per definition) classified as an extreme storm surge. We convert

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

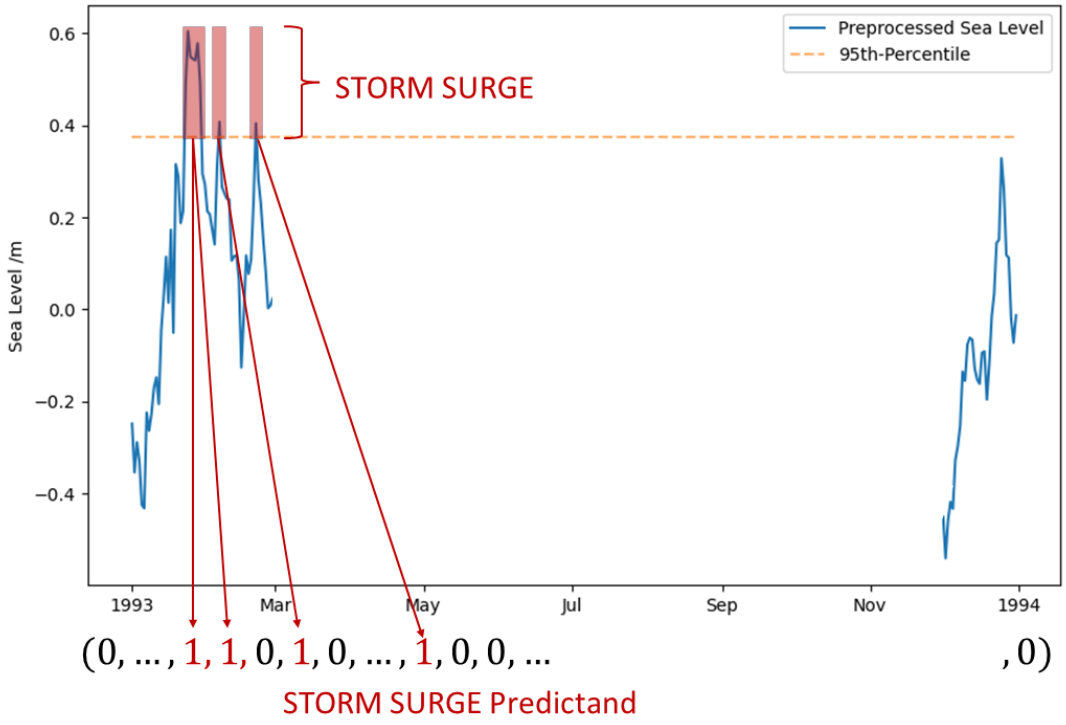


Figure 3. Transforming a continuous sea level to a categorical predictand by using a percentile-threshold based definition of extreme storm surges. The predictand is a vector, where 1 indicates a storm surge and 0 its absence.

this hourly categorical variable into a daily timeseries by attributing a storm surge to a specific day if only one hour of that day exceeds the 95th quantile.

Following this preprocessing we have 2-dimensional spatial maps as predictors, leading to a total dimensionality $n_{pred} \times n \times n_{lon} \times n_{lat}$, where n_{pred}, n, n_{lon} and n_{lat} are the number of predictors (drawn from SP, TP, U10, V10), number of samples (days or hours), longitudes and latitudes, respectively. The predictand is a categorical binary variable indicating the occurrence of a storm surge with dimensionality $n \times n_{station}$, where $n_{station} = 1$ as we analyze each station separately.

The timeseries of the predictor and predictand are intersected by date to put both on the same time-domain. For some experiments, we introduced a timelag Δt between predictor and predictand by de-aligning the timing of predictors and predictand. In these cases, the number of samples reduces to $n - \Delta t$ for both, the predictor and predictand. Hence for a timelag of Δt any timepoint $t \leq n$ of the predictand is predicted by using predictors at prior times $t - \Delta t$. If Not-a-Number (NaN) entries occur, they are replaced by its mean over all dimensions.

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

3 Methods

The overall structure of the algorithm is sketched in Fig. 4. Before passing data to the model, we split our predictor and predictand datasets into two subsets $\mathcal{M}_{\mathcal{T}}, \mathcal{M}_{\mathcal{V}}$. The first one ranges from 1999 to 2008 and is used for training and testing the model. The second one covers the period from 2009 to 2018 and is used to evaluate the generalisation of the model. Note that for stations 3 and 4 the predictand timeseries only starts in 2005, which is why time periods of $\mathcal{M}_{\mathcal{T}}$ and $\mathcal{M}_{\mathcal{V}}$ for those stations are 2009-2018 and 2019-2020, respectively.

After splitting the data we feed the model, a Random Forest (RF), with identical combinations of predictors for each station. The RF then processes the atmospheric predictors (denoted features in ML parlance) by leveraging the predictions of several Decision Tree (DT)s. Finally, the RF provides a deterministic, binary prediction of extreme storm surges (predictands, also called labels) - indicating whether a storm surge occurs (1) or not (0).

Commonly all possible predictors are initially used as inputs for the model. The model can then itself derive the most important features, which comes with additional computational costs. To evade this circumstance, we only tested combinations of predictors that were in line with the theoretical explanation of storm surges.

Our algorithm is publicly accessible on GITHUB (Bellinghausen, 2022) and is based on the *scikit-learn* library of Python.

In the following sections we will explain basics of the RF, its tuning and evaluation.

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

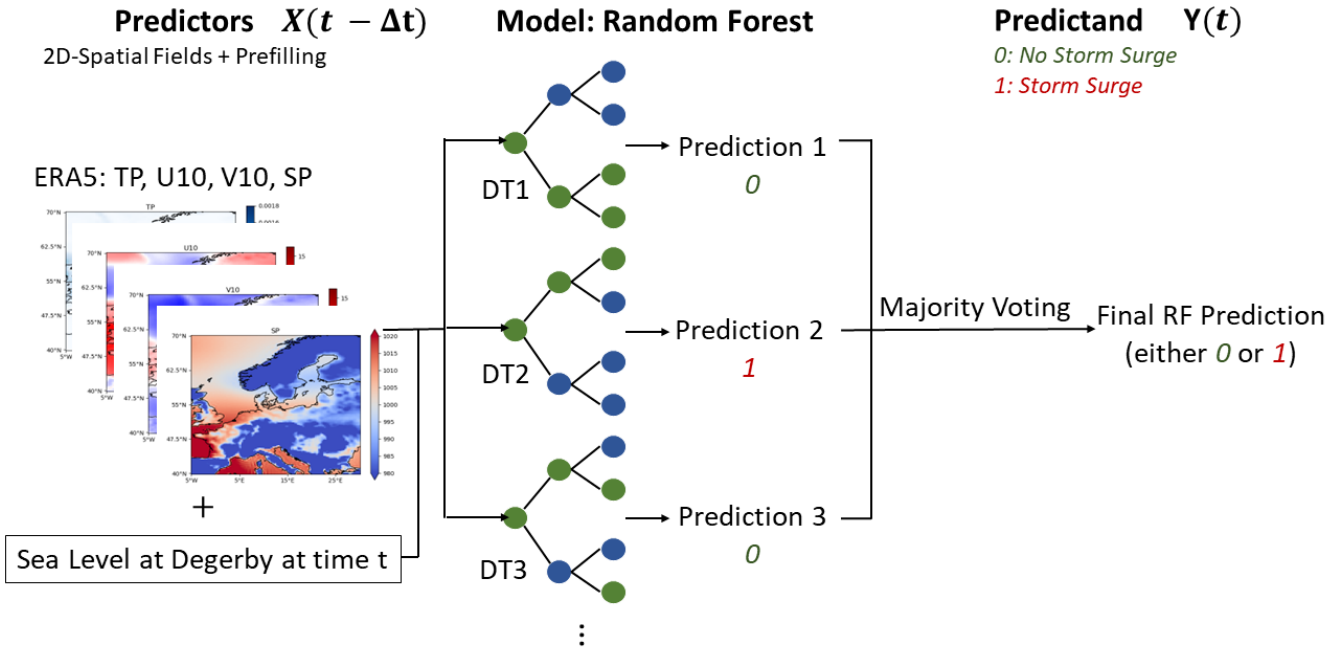


Figure 4. Software architecture as a blueprint. A random forest is used to predict storm surges categorically by using atmospheric predictors represented as 2D-spatial fields.

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

3.1 Random forests

As a classifier, we used the *RandomForestClassifier* from *scikit-learn*. A thorough description of RFs can be found in Müller (2017) and Géron (2017), from which we will briefly discuss the most important points.

The model architecture of a RF is based on an ensemble of Decision Tree (DT)s (see Fig. 4). DTs rely on a hierarchy of if/else-questions in order to conclude with a prediction. A simplified example is shown in Fig. 5. In this case, the DT formulates sequential if/else-questions about the predictors U10, SP and PF. The grey nodes indicate a path of input data, where each question is answered positively, hence leading to the prediction of an extreme storm surge. In reality, the questions in each node are more complex, testing for continuous values of the predictor at hand (e.g. u10 $> 17\text{ms}^{-1}$ as a test for strong westwind at a specific grid-point within the research area). When fitting the structure of a decision tree to fit a training data set, the algorithm uses a concept called *gini-impurity* to find the best sequence of if/else- questions for a prediction. A prediction based on new predictor data is then made by sifting through the optimized DT, answering all if/else questions.

In a RF (Fig. 4) as an ensemble of Decision Tree (DT)s, each of those DTs processes a random sample of the given data in order to conclude with a prediction. In general, those predictions are then aggregated in order to get the overall prediction of the RF. For a binary classification problem, as in our case, this aggregation is done via a majority voting.

Because DTs are based on an if/else hierarchy, RFs belong to the realm of interpretable as they provide a parameter f named *feature importance* (FI). The FI assigns a value between 0 and 1 to each feature (predictor), with higher numbers indicating greater importance. The importance of one predictor is estimated by computing the predictive loss of the algorithm when that predictor is omitted. The value of the importance is normalized by requiring that the sum of all feature importances within a DT is 1. The feature importance describes which predictors improved the prediction the most during the training process.

In this paper, predictors are atmopsheric variables on a grid, which leads to a dimensionality of the feature importance of $n_{pred} \times n_{lon} \times n_{lat}$. Hence, each grid cell is associated with a feature importance for each climate predictor (wind, pressure, ...) and we can utilize it to filter regions on the grid that are important to predict a storm surge. It is important to note, that f should not be mistaken with the causality of a predictor and only represents a correlation detected by the RF. Hence, we analyse if the feature importance resolves regions and atmospheric patterns that are coherent with theoretical drivers of storm surges described in the previous sections.

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

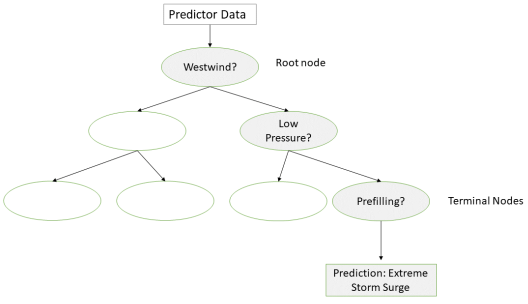


Figure 5. Simplified model architecture of a Decision Tree (DT). Grey Nodes indicate the path of test-data while sifting through the DT. Right-pointing arrows refer to positive answers.

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

3.2 Model tuning

The *RandomForestClassifier* can be tuned in several ways by altering its hyperparameter (HP)s. We will use the subset $\mathcal{M}_{\mathcal{T}}$ for this purpose and separate it into two further sets – a training- and test-set – by using the *train_test_split* method from scikit-learn, witholding 75% of the data for training and 25% for testing. Hence, the models HPs are optimized based on the training-set and the model accuracy is evaluated on the test-set. For a RF the most important HPs control the amount of DTs used (*n_estimator*), the maximum depth of each DT (*max_depth*) and the number of features used when calculating the best split (*max_features*). In general, a larger value for *n_estimator* will lead to a more robust ensemble due to less overfitting as the results of many DTs are averaged (Müller, 2017). (Breiman, 2001) show that the generalization error of RFs converges for a growing number of DTs, again indicating less overfitting. With increasing *max_depth* the DTs get more complex, hence overfitting is more likely. The *max_features* controls the randomness of each DT with a smaller value reducing overfitting (Müller, 2017). While we set *max_features* to its default value of $\sqrt{n_{\text{pred}}}$, we varied the other two.

In addition to those HPs we altered the *class_weight* and *random_state*. The *class_weight* is used to associate weights with classes. This is particularly important in this study as we deal with extreme storm surges. Hence the predictand dataset is unbalanced as there are many more days of class 0, without a storm surge, than of class 1, with extreme storm surge. Setting the *class_weight* to "balanced" adjusts weights inversely proportional to class frequencies in the input data, i.e. the model will penalize more heavily wrong predictions about class 1 days than wrong predictions about normal conditions. We set the *random_state* to 0, which gives us and the reader the possibility to reproduce results.

The HPs *n_estimator* and *max_depth* were automatically depicted by the algorithm as the best combination of HPs using *RandomSearchCV*, an optimisation procedure within *scikit-learn*. One can pass a list of values for each HP and *RandomSearchCV* automatically selects the best combination by optimising the validation score of the training set based on cross-validation. This comes with the advantage that the initial split into training and test set is sufficient and no additional validation set is needed (for more details we recommend to consult the sci-kit learn documentation).

Although the number of effective predictors is substantial, we did not reduce the dimensionality of the predictor fields (by principal components analysis or an autoencoder) to avoid loosing any regional details that could be relevant for each station. We preferred in this case to limit the depth of the Random Forest algorithm to avoid overfitting, drawing only from the list $[1, 2, 3]$ for the *max_depth* parameter. For the *n_estimator* we used either 333, 666, or 1000.

All settings are summarized in Table A3 for replication purposes.

3.3 Model evaluation

A common tool to evaluate binary classification models is the Confusion Matrix (CFM) (see Fig. 6). It summarizes the accuracy of a model in terms of success or failure rates. For our study, we aim for a high *True Positive Rate (TPR)*, which relates the absolute number of correctly predicted extreme storm surges n_1 to all incidences of storm surges n_ϵ in the underlying data. In Fig. 6 for example $n_1 = 29$ out of $n_\epsilon = 40$ extreme storm surges were correctly predicted, leading to a TPR of $\frac{n_1}{n_\epsilon} = 72.50\%$. A high TPR automatically leads to a low False Negative Rate (FNR) since their sum equals one. The FNR indicates how often

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

the model actually fails to predict a storm surge. With a high FNR, the model can not be trusted as it very likely produces false predictions of security, i.e. it is too insensitive. Especially for extreme events, this can lead to devastating damage to societies when protection measures rely on model predictions with a high FNR, as eventually no measures are taken due to a model prediction of "no storm surge" but in reality, an extreme surge appears.

The CFM can be evaluated on training and test data as well as on the validation set. If model predictions are correct almost always on training data, i.e. a TPR and True Negative Rate of around 100%, the model tends to overfitting. In practice, the CFM of test-data and the validation set is more interesting as it shows the performance of a model when confronted with data, that is uninvolved in the training process.

The second tool we use is a combination of the Feature Importance (FI) and a Predictor Map (PM). For each model, the importance of each feature is displayed by weights between 0 and 1 with all weights summing up to 1. Using FI lets us compare the overall importance amongst predictors when a combination of predictors is passed to the model. Furthermore, we can deduce which specific regions within the research area are important for model decisions for each predictor. We only show the top 1% area of importance (computed by the 99th percentile of FI for each predictor) and depict those regions by grey squares (see Fig. 7).

Unlike a correlation coefficient, the FI does not encode the magnitude nor sign of the feature that is indicative of the storm surge Müller (2017), e.g. whether it is low or high pressure within the area of importance that is related to a storm surge. Hence in the results we also include an averaged value of the predictor for all cases of a storm surge. This leads to two separate types of PMs; one where the model correctly predicts the storm surge, i.e. true positive predictions (TPPs) and another where the model predicts no storm surge even though there is one in the observations, i.e. false negative predictions (FNPs). Those PMs are compared amongst each other as in Fig. 7 (a) and (b) and their difference (bias) is showcased in (c). For instance, when PMs for TPP cases show low-pressure systems in the importance region while the FNP PMs only display high-pressure systems, this suggests that the model heavily relies on low-pressure systems to forecast a storm surge. By contrast, it also suggests that in some cases storm surges are induced even though there are high-pressure systems in the area of high FI.

As it is sufficient to only show maps for TPPs and the difference to FNPs we will do so in the results section.

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

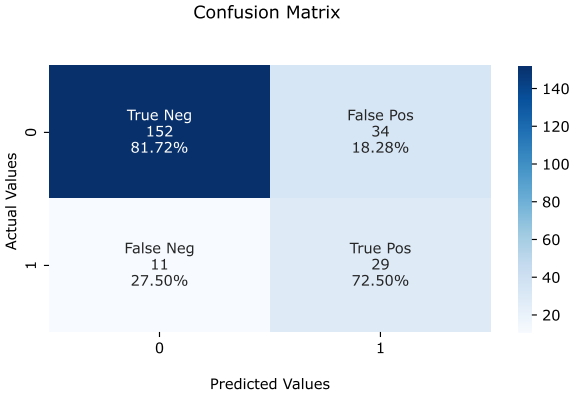


Figure 6. Confusion matrix for a binary classification model with absolute and relative values. The colour bar shows the maximum count of instances for all cases.

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

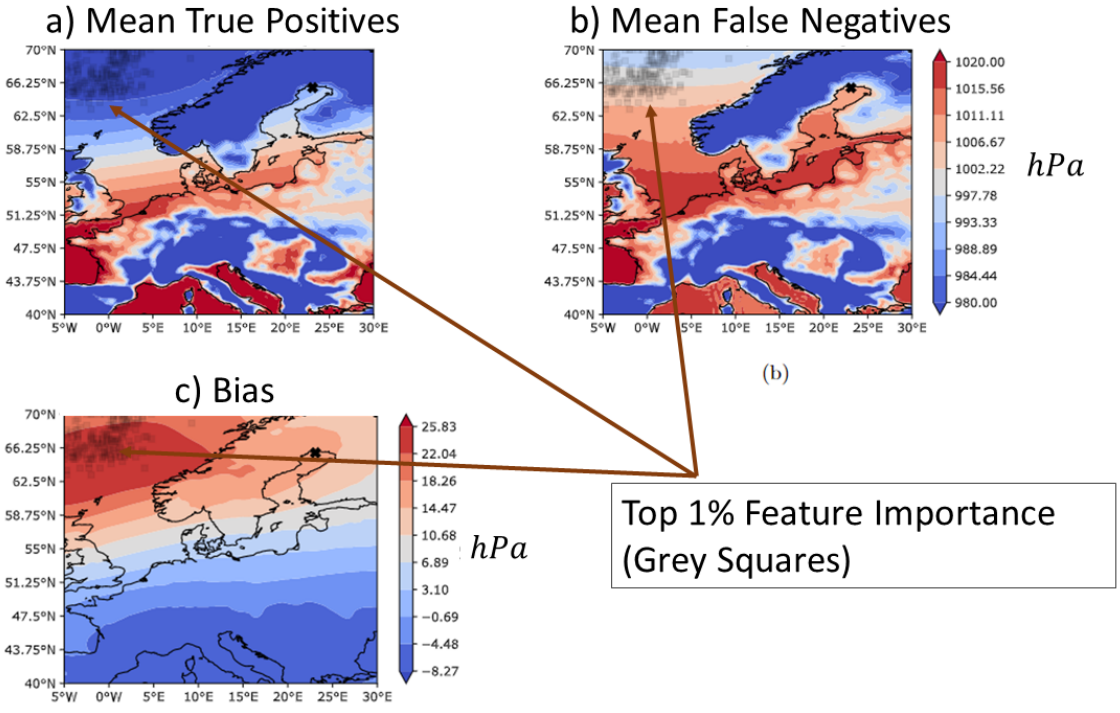


Figure 7. Mean Predictor Maps of SP all with timelag $\Delta t = 1$ for station 0 (NSWE). (a) mean True Positive Prediction (TPP)s, (b) mean False Negative Prediction (FNP)s and (c) difference of both means FNP - TPP. Note the different scaling of the colour bar for the difference maps.

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

4 Model configurations

We build six overarching model configurations (**A – F**). For each configuration, we undertake subsets of model runs which are denoted by *run_ids*. All model runs are applied to each station, i.e. similar predictors and initial HP lists were used when building the model for each station (note though that the fitted model can be different for each station due to the automatic optimization of hyperparameters). As a starting point, we analyzed the predictive skill of each predictor individually with time lags up to a week (**A**). Among those, time lags up to three days showed promising results. These time lags are interesting in order to predict storm surges in advance. Hence, we analyzed a combination of all predictors (ERA5 and PF) with time lags of 1 and 2 days, eventually getting insight into which predictors are most important (**B**). In experiment **C**, we investigate the coupling of strong winds and moving low-pressure systems by combining SP and U10 with various time lags. Because the west wind is an important driver of storm surges in the Baltic Sea due to the connection with the North Sea via the Kattegat and the possible wave build-up in the north-eastern region, we combined multiple time lags of U10 in experiment **D**. In **E** we looked into cumulative rain (TP with several time lags and U10), wind-induced waves in combination with prefilling and the state of prefilling induced by wind (both using U10 and PF). Since we use the water-level records at the Degerby station as a proxy for prefilling and not the rolling mean of 20 consecutive days like Mudersbach and Jensen (2010), we combined several time lags (up to 30 days) of PF in experiment **F**.

All model runs are summarized in Tables A4 – A9.

5 Results

We selected promising results based on a combination of the TPR of the test dataset from $\mathcal{M}_{\mathcal{T}}$ and validation datasets $\mathcal{M}_{\mathcal{V}}$, labelled as TTPR and VTPR, respectively. When interpreting those rates, we will indicate the total amount of extreme storm surges for the specific dataset with n_ϵ in parentheses. In contrast to ERA5-predictors, using prefilling as a sole predictor contains more instances of storm surges due to the hourly recording time, hence looking at TTPR instead of VTPR is also sensible.

The prediction skills described in the following subsections will be later compared in subsection 5.7 with a storm-surge reanalysis obtained with a global comprehensive dynamical model driven by atmospheric forcing from global meteorological reanalysis (Muis et al., 2016). Before that, we can provide a basic benchmark by indicating the prediction skill of a simple *uninformed* prediction scheme. One such scheme could be to always predict storm surges. This scheme would display a true positive rate of 100% but also a false positive rate of 95%. It would be very sensitive but very unspecific. A prediction scheme that always predicts 'no storm surge' would display true positive and false positive rates of 0% each. It would be totally insensitive. Both schemes are obviously not useful. A slightly more sophisticated scheme would issue a storm surge prediction randomly in 5% of the occasions. The true positive rate would amount to 5% and the false positive rate to 5%. It would still be rather insensitive. A prediction scheme has to clearly improve this random sensitivity and specificity.

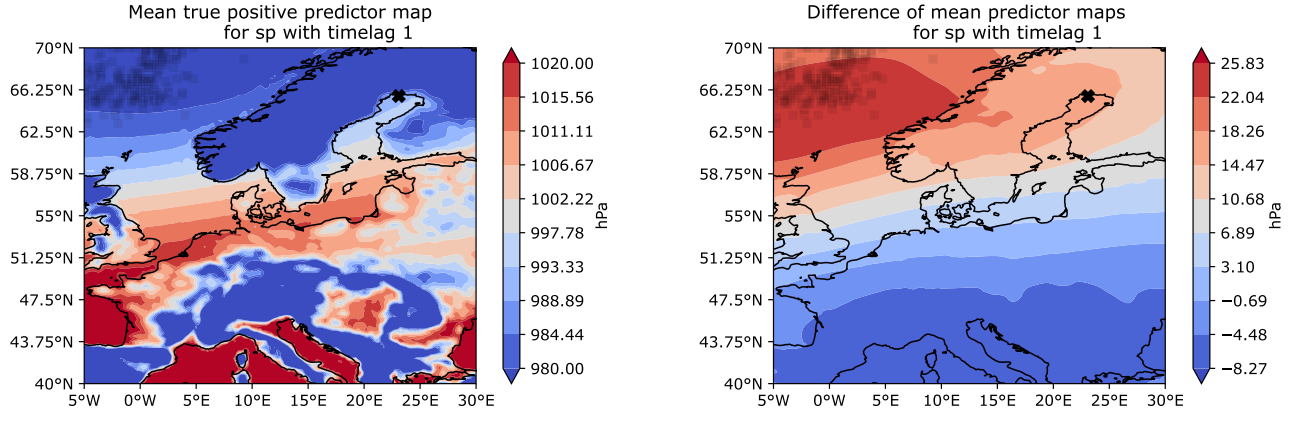
https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

5.1 A - Single Predictors and Multiple Timelags

We aim to find the best predictors for each station and analyze what time lags are most useful. Furthermore, we want to investigate how physical patterns of predictors change depending on the station location by investigating their feature importance.

5.1.1 Surface Pressure SP

In terms of VTPRs, the SP leads to good results for all stations except DEU. For instance VTPRs were at 70.67% ($n_\epsilon = 75$) and 78.9% ($n_\epsilon = 109$) for time lags of one and zero days for stations NSWE and FINBAY, respectively. Only for station DEU the SP was not such a useful predictor (VTPRs below 50%). We could observe that for almost all stations, time lags up to two days work reasonably well, while longer time lags generally reduce the VTPR. Independent of the station, low-pressure systems are important within the AoI but for cases of FNPs the pressure rises several hPa. At station NSWE, for example, the AoI of SP is within the region 5°W – 5°E and 62.5°N – 70°N (see Fig. 8). Here, mainly low-pressure systems with less than 980 hPa lead to a correct prediction of a surge. The model tends to FNPs once the pressure in the AoI increases by a mean of around 25 hPa. In some cases, high-pressure systems of more than 1020 hPa occurred in the AoI for FNPs. This behaviour repeats for several stations.



(a) (b)

Figure 8. Mean Predictor Maps of SP with $\hat{t} = 1$ for station 0 (NSWE) for (a) TPPs and (b) the difference of FNPs and TPPs. Note the different scaling of the colour bar for the difference maps.

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

5.1.2 Zonal-Wind U10

Model results show that a positive zonal wind U10, i.e. westwind, is mostly useful for stations located in the eastern parts of the Baltic Sea. While the VTPRs for station NSWE are greater than 60%, VTPRs of 73.39%, 72.48% and 69.62% ($n_\epsilon = 109$) were measured for station FINBAY. Those results were deduced for time-lags zero, one and two, respectively, indicating the short-term wind fields as the most relevant predictors. The AoI strongly depends on the location of the station as well as on the chosen time lag (see Fig 9). For instance, for station FINBAY, the AoI clusters around the Gulf of Riga and only covers the Region of the Kattegat lightly when using no time lag. This is interesting, as one would expect important short-term wind fields close to the station itself or at least close to the entrance of the Gulf of Finland, s.t. wind setup can be induced. Conversely, the AoI for time lags of one and two days locates more around the Kattegat area and the Southern Coast of the Baltic Sea (see Fig. 9a – 9c). When looking at the AoI of station NSWE for a time lag of one day, it is not anymore the area around the Kattegat that is of importance but rather the west winds close to the UK over the North Sea (Fig. 9d).

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

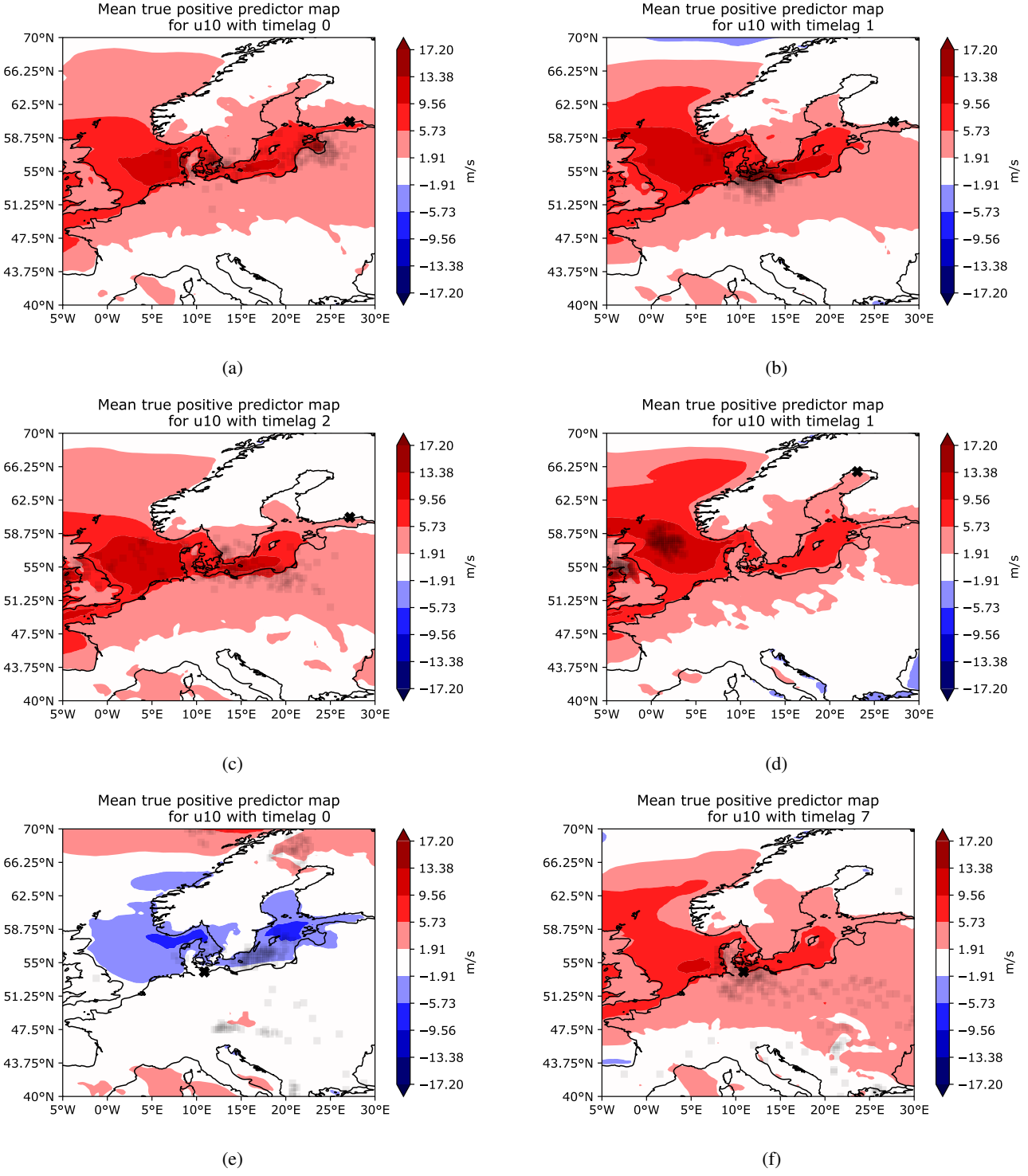


Figure 9. Mean predictor maps for TPPs using U10 with time lags 0, 1, 2 for stations FINBAY (a – c), NSWE (d) and DEU (e, f).

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

Regardless of the AoIs location, the main wind direction is eastward. For instance, at station FINBAY, mean west wind speed of around 12ms^{-1} occurs in parts of the AoI for TPPs, especially around the Danish Straits. When looking at PMs separately, windspeeds of 17ms^{-1} and more (i.e. storms) could be detected. Comparing the maps of TPPs to the ones of FNPs one can see that the model generally leads to false predictions when west wind fields become weaker. The difference maps show a mean decrease of west wind speeds of around 7ms^{-1} in parts of the AoI (not shown). Hence, the model is not as reliable when west winds are not strong and no winds or even east winds occur. This behaviour repeats for other stations except for station DEU. Here, the east wind is used for model predictions (see Fig. 9e), which is an expected result. There are fewer (positive) storm surges at the German Coast of the BS compared to other Bays as usually southwesterly winds lower the water level in those regions (Weisse and von Storch, 2010). It is interesting to see, though, that important short-term winds are mostly westward close to the station. This is also theoretically explained by the induced pile-up effect of wind at this station. Another mentioned driver for storm surges along the German Baltic Coast is seiches. These might be induced by the pronounced west wind around the station and over the Baltic Sea for a time lag of seven days (Fig. 9f). The long time lag could be sufficient for a wave growth towards the opposing coast, which in turn leads to seiches once the wind turns westward or stops blowing.

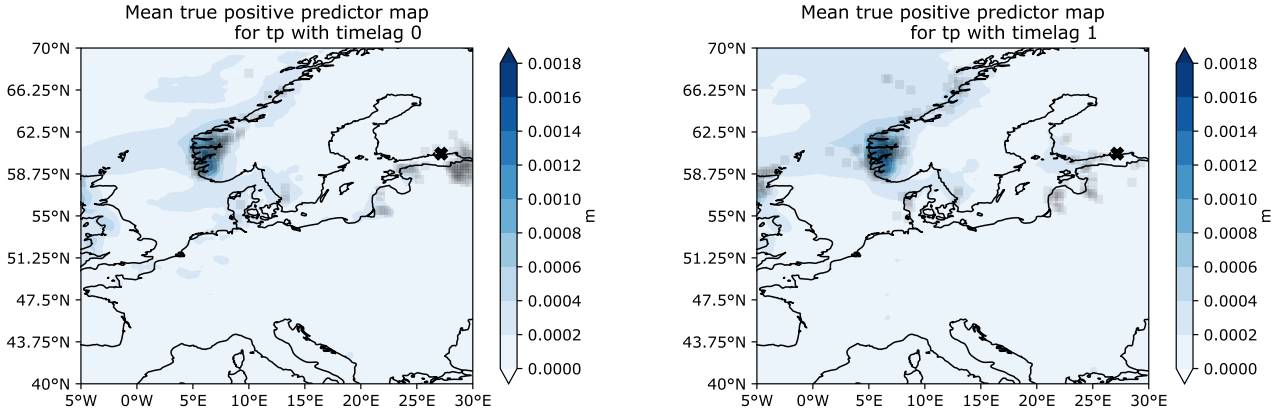
5.1.3 Meridional-Wind V10

While the zonal wind is a good predictor for stations located at zonal boundaries of the Baltic Sea, the meridional wind component V10 is a good predictor for stations located at the northern extent of the Baltic Sea. For instance, at station NSWE the meridional wind with a time lag of one day leads to VTPRs of 73.33% ($n_\epsilon = 75$). For all other stations, meridional wind was not a good predictor. For instance, WSWE and WSWE2 only had VTPRs of around 50% using V10. At NSWE though mainly light southwinds of around 6 ms^{-1} are used by the model for True Positive Prediction (TPP)s, while it struggles when no meridional wind is blowing in the AoI.

5.1.4 Total Precipitation TP

Total precipitation was (in most cases) not a promising predictor. Except for station FINBAY, where it resulted in VTPRs of 69.72% and 72.48% (both $n_\epsilon = 109$), for no time lag or a time lag of one day, respectively. The AoI for TP without a time lag is close to the station itself. When increasing the time lag by only one day, the AoI shifts towards the area around Bergen, sometimes showing connecting patterns of importance across the North Sea towards the United Kingdoms (see Fig. 10). This behavior repeated for other stations but VTPRs were lower. Nevertheless, throughout all experiments, TP was not showing any consistent patterns in terms of PMs.

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024



(a) (b)

Figure 10. Mean Predictor Maps of TPPs for predictor TP with time lags 0, 1.

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

5.1.5 Prefilling PF

Despite the good performance of ERA5-predictors, the most important predictor for all stations in terms of TTPR is prefilling. For instance, at station NSWE, time lags of two and seven days lead to TTPRs of 85.71% and ($n_\epsilon = 2254$) and 84.12% ($n_\epsilon = 2337$), respectively. For all other stations TTPRs were also above 70% when using no timelag at all. It is also worth noting that the TTPR of prefilling shrinks consistently when increasing the time lag.

From experiment **A**, we can conclude that the choice of predictors depends on the station at hand. Depending on their location, values of predictors in the AoIs vary, especially when considering wind fields. For SP the model uses always low-pressure systems in order to achieve TPPs. The most valuable predictors are SP and U10, showing mainly low-pressure and west wind fields in AoIs. In general, time lags up to three days were best. Choosing longer time lags often leads to worse results. Overall, PF was the most useful predictor for all stations. The results are summarized in Fig. 11.
We did see that, in some cases, storm surges occur even though predictors show values one would not expect based on theory, e.g. high-pressure fields in the AoI. Physically, it is not straightforward to explain. However, note that we used each predictor only in isolation. Hence, it might be possible that a combination of other predictors, e.g. a strong prefilling in combination with west winds, is inducing the storm surge. Therefore, we will analyze combinations of predictors in the following experiments.

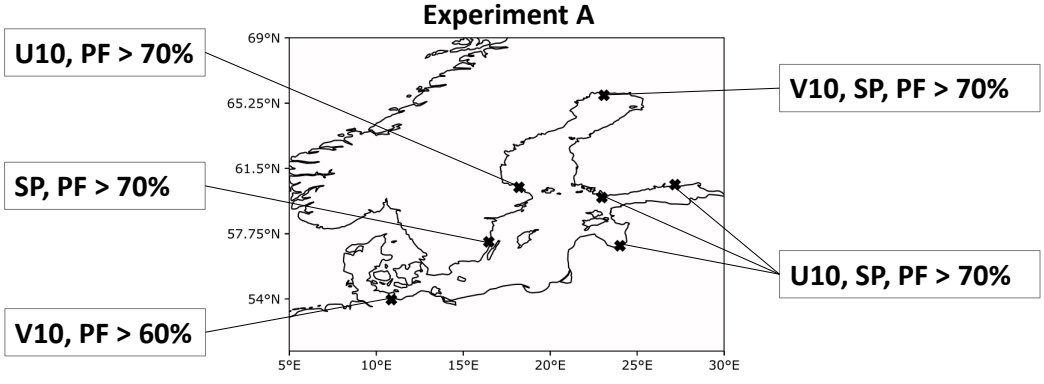


Figure 11. Summary of best predictors per station for experiment A. The percentage indicates the corresponding VTPR or TTPR.

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

5.2 B - Combination of all predictors

In this experiment, we combined all ERA5-predictors in order to rank them by feature importance and look at the behaviour of their corresponding PMs (see Table A5). For almost all stations SP and U10 are the most important predictors. They again show pronounced low-pressure fields (below 980 hPa) and strong west winds (greater than 15ms^{-1}) in their respective AoI. This behaviour switched only for the stations at the Baltic Sea's meridional extents, and V10 becomes important as well. In terms of PMs the physical components did not change compared to experiment A.

Nevertheless, using the predictors in combination showed an order of importance as depicted in Figure 12. We can see that SP and U10 are mostly used by the models, but it also switches depending on the station. In terms of the maximum VTPR achieved at each station, using isolated predictors leads to better results.

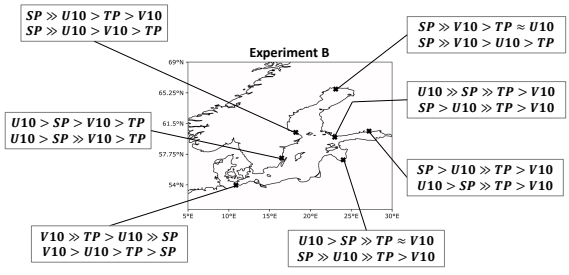


Figure 12. Order of predictor importance for experiment B with run_ids 1, 3. First and second row show timelags of one and two days, respectively. The $>>$ sign indicates that the feature importance was almost double as high, the $\approx$ indicates approximate equality.

5.3 C - Coupling of U10 and SP

We already noted that SP and U10 are important predictors. In theory, resonance coupling of strong winds and moving weather systems (low-pressure systems) leads to extreme storm surges as well. Hence, we will now investigate two sets of combinations of those predictors as shown in Table A6. One set uses similar timelags for both predictors, another uses a shorter time lag for SP compared to U10 as we expect the effects of U10 on storm surge to be slower than the influence of low-pressure systems. This is due to the fact that U10 needs to transfer kinetic energy to the ocean's surface first in order to induce waves.

Combining both predictors with a similar time lag did not improve results compared to using them in isolation. For some stations (NSWE), it even leads to worse VTPRs, perhaps an indication of overfitting at those stations. Nevertheless, short-term combinations with time lags up to three days seem to work best. Only on a few occasions (station 1: FIN) time lags up to 5 days also produce acceptable results.

Best results could be observed for station 2 (FINBAY) with the highest VTPR of 75.23% using no time lag. For this station, time lags of one or two days lead to VTPRs above 70%. All other stations had similar VTPRs above 60%, mainly for time lags up to three days. The only (expected) exception was station 4 (DEU), for which both predictors can not be used (TTPRs below

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

45%).

The PMs mainly showed similar behavior as for using isolated predictors.

Comparing VTPRs across all stations of both subsets of this experiment, we deduce that similar VTPRs over 60% and, in best cases, even up to 70% could be achieved. Using a difference in time lags of SP and U10 did lead to more stable results when altering the time lag of U10. In other words, the VTPR did not diminish quickly when increasing the time lag of U10.

In total, this experiment showed that, for most stations, a combination of short and long-term data, as well as a positive difference in time lags between U10 and SP, leads to good results in terms of VTPRs (see Fig. 13).

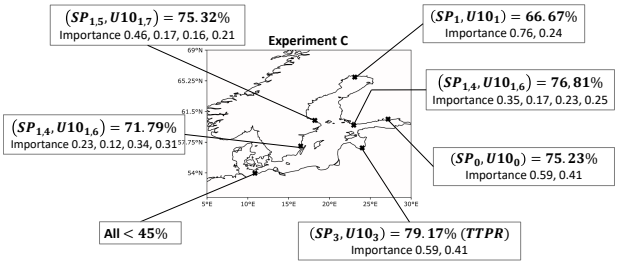


Figure 13. Best combinations of predictors for experiment C. Timelags are indicated as subscripts of the predictor. Importances are given in the order of subscripted time-lags. Depending on the station VTPR or TTPR is shown.

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

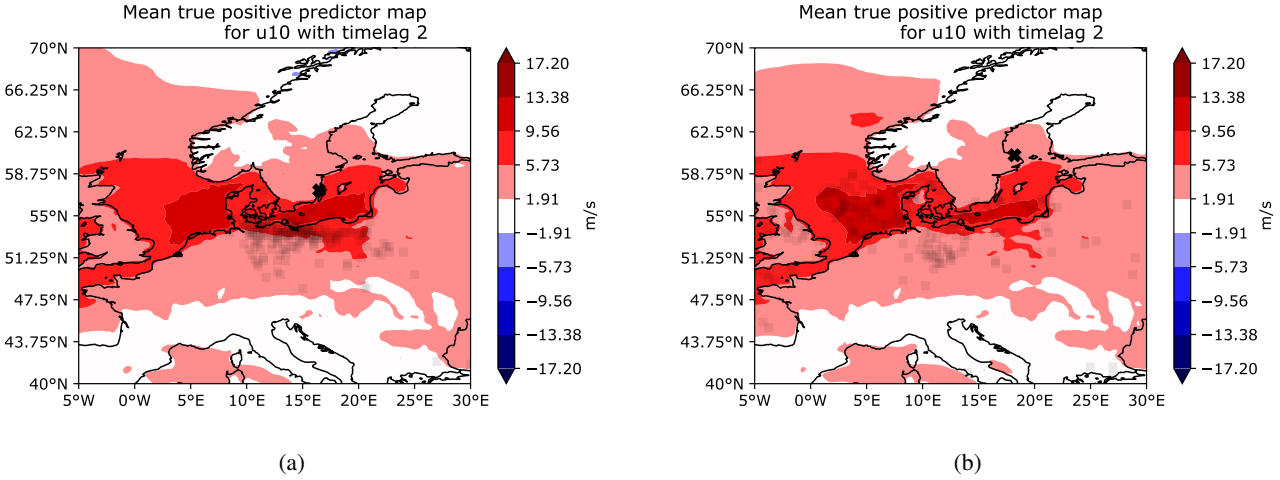
5.4 D - Combinations of Westwind-Timelags

As was already shown, west winds are an important driver of storm surges. If those winds blow consistently over several days, it deforms the sea surface and causes drift currents. Hence, in this experiment, we will investigate U10 with several time lag combinations as shown in Table A7.

We already found that a time lag of one or two days works well for U10 due to previous experiments. Hence, we combined those short-term time lags with longer ones in the first subset of this experiment (run-ids 0 – 3). In a second subset, we investigated timelags up to a week, comparing short and long-term combinations of the time lag (run-ids 4 – 7). Finally, we spread the time lags over a whole week and even over a whole month for run-ids 8 – 11.

Overall, using combinations of only U10 worked quite well for almost all stations, with VTPRs above 70%. Only for stations 0 (NSWE) and 5 (WSWE), the best VTPRs were just above 60%. As expected, station 4 (DEU) showed poor results. Mostly, time lags up to four days worked the best for all stations.

AoIs and PMs show again similar behaviour to experiments before, i.e. mainly strong west winds mostly in regions around the Danish Straits or Southern Baltic Coastline. Depending on the location of the station AoIs vary their area. They do so even for slight positional changes of stations, for instance, stations 5 and 6. Figure 14 depicts this for a time lag of two days. While for station 6, west winds around the North Sea entrance of the Danish Straits are important, this is not the case for station 5. The whole AoI shifts more towards the East. One explanation might be that west winds can not induce a direct wind-setup for station 6, as its coastline is oriented towards the North, hence sheltered from the winds. The opposite is true for station 5. The coast here is facing towards the South and South-West. Hence, west winds may induce strong wind buildup for station 5 while for station 6 a state of prefilling is induced by the wind around the Danish Straits which indirectly leads to a storm surge.



(a) (b)

Figure 14. Mean Predictor Maps for TPPs using U10 with a time lag of 2 days at stations (a) WSWE and (b) WSWE2

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

455 In summary, combinations of U10 can be used for most stations as a good predictor when focusing on time lags up to four days (see Fig. 15). For some stations, time lags up to a week also lead to good predictions. Even longer time lags should not be used as they are mostly disregarded by the model.

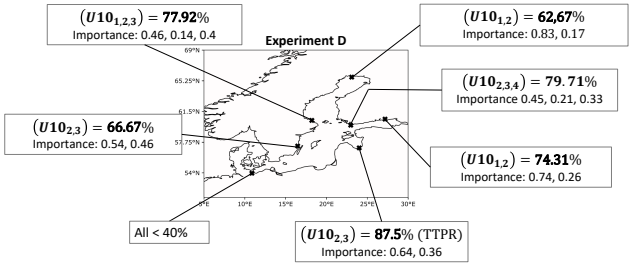


Figure 15. Best combinations of predictors for experiment D. Depending on the station VTPR or TTPR is shown. Importances are ordered as subscripted time-lags.

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

5.5 E - Predictor Combinations from Theory

We tried to emulate the effect of cumulative rain and looked into how information on prefilling changes the behaviour of the west wind for model predictions. The combinations of predictors can be found in Table A8 and results of VTPRs and TTPRs are summarized in Fig. 16

5.5.1 Cumulative Rain

Best results for the cumulative rain combination were observed for station 6 (WSWE2), with a VTPR of 74.03% ($n_\epsilon = 77$). For stations FIN, FINBAY and WSWE also good results around 60% VTPR were calculated. However, when looking at the importance, one can see that mostly U10 is used for model predictions. For all stations, except station NSWE, the sum of TP feature importances is smaller than the feature importance of U10.

5.5.2 Effect of prefilling on zonal wind

The most interesting observations could be made when looking at station 1 (FIN) for combinations of U10 and PF. As the theory suggests, with a state of prefilling in the Baltic Sea, weaker west winds are needed to induce storm surges compared to times without prefilling. The PMs of TPPs for an isolated U10 (as in experiment A), a combination of U10 and PF (run-id 1), as well as their difference, are depicted in Fig. 17. In both cases, U10 was implemented with a time lag of three days. Comparing the area around the Danish Straits, i.e. 2°E – 15°E and 54°N – 57°N, shows that in the case of no prefilling, west winds blow up to 5 ms^{-1} stronger in specific areas. Hence, the model predicts (on average) more storm surges with weaker west wind correctly when it has access to information on the prefilling.

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

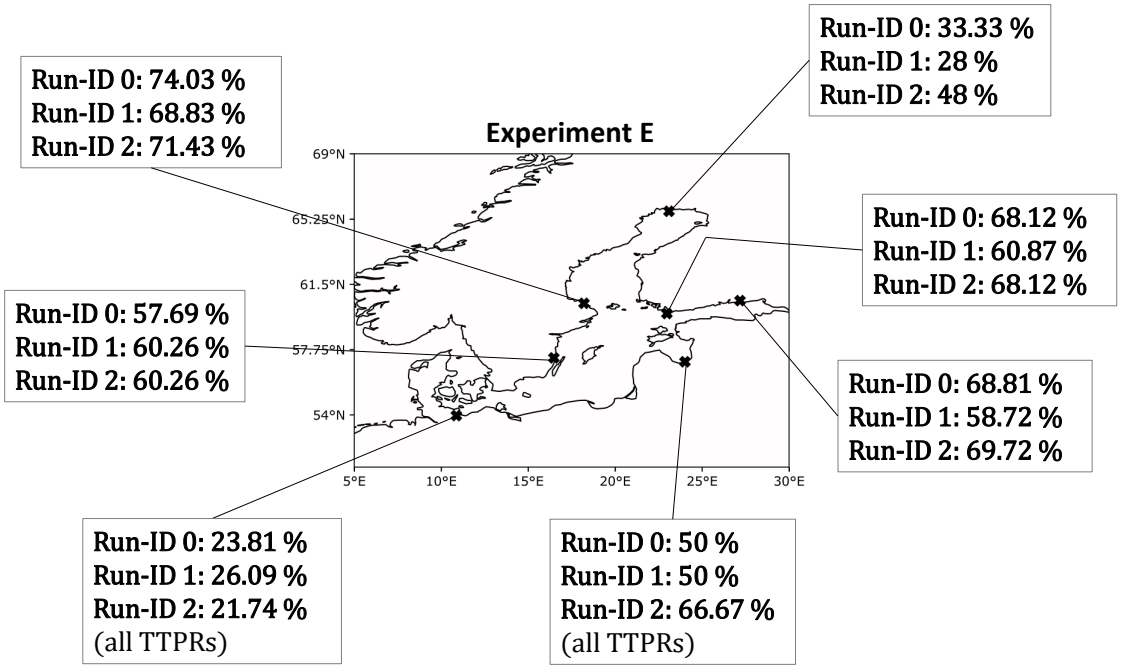


Figure 16. TTPR or VTPR results for all run-ids of experiment E.

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

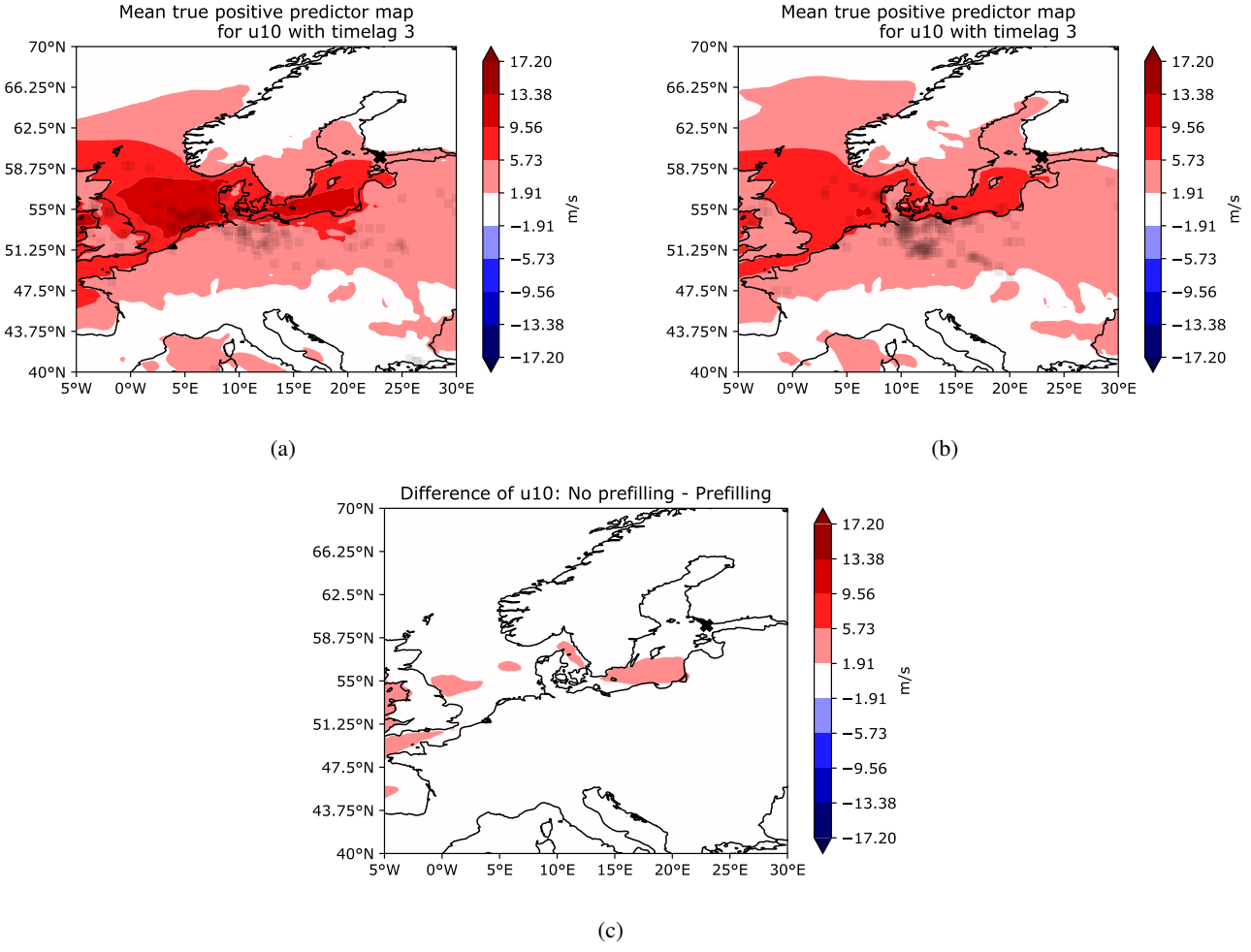


Figure 17. Mean Predictor Maps of U10 at station 1 (FIN) and a time lag of 3; (a) Without information on prefilling, (b) with information on prefilling, (c) difference of (a) and (b).

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

5.6 F - Combinations of prefilling-timelags

The prefilling of the Baltic Sea is strongly influenced by the strong west wind. While Weisse (2014) and Mudersbach and Jensen (2010) define the prefilling as the rolling mean of the water levels at Degerby over 20 consecutive days, we will use a time lag of the records of water level at Degerby as the predictor. In this experiment, we investigate PF as an isolated predictor for time lags up to a month as well as combinations of PF which include short-term (up to a week) as well as long-term (up to a month) information on the water levels. All combinations can be found in Table A9.

When using isolated predictors, results show that shorter time lags work better in terms of TTPR than longer ones. For instance, TTPRs of station 2 (FINBAY) for time lags of 10, 15, 20, and 25 days were 71.12%, 63.63%, 49.85%, 61.61%, respectively. Combining the information from several days leads to the best results for all stations. The overall highest TTPR of 91.15% ($n_\epsilon = 2869$) could be achieved for station 1 (FIN) when using the time lag-combination of 3, 14, 21 and 30 days. In this case, the feature importance for three days was significantly higher than the one for 14 days, which itself was more than the doubled feature importance of a time lag of 21 days. This indicates that the model heavily relies on the most recent water level recordings in order to provide TPPs, a behaviour that is generally repeated for all stations.

Overall, prefilling seems to be a good predictor for almost all stations when combining information from the previous water records with records up to two weeks old. Independent of the station, combining several time lags of information works better than using the information in isolation.

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

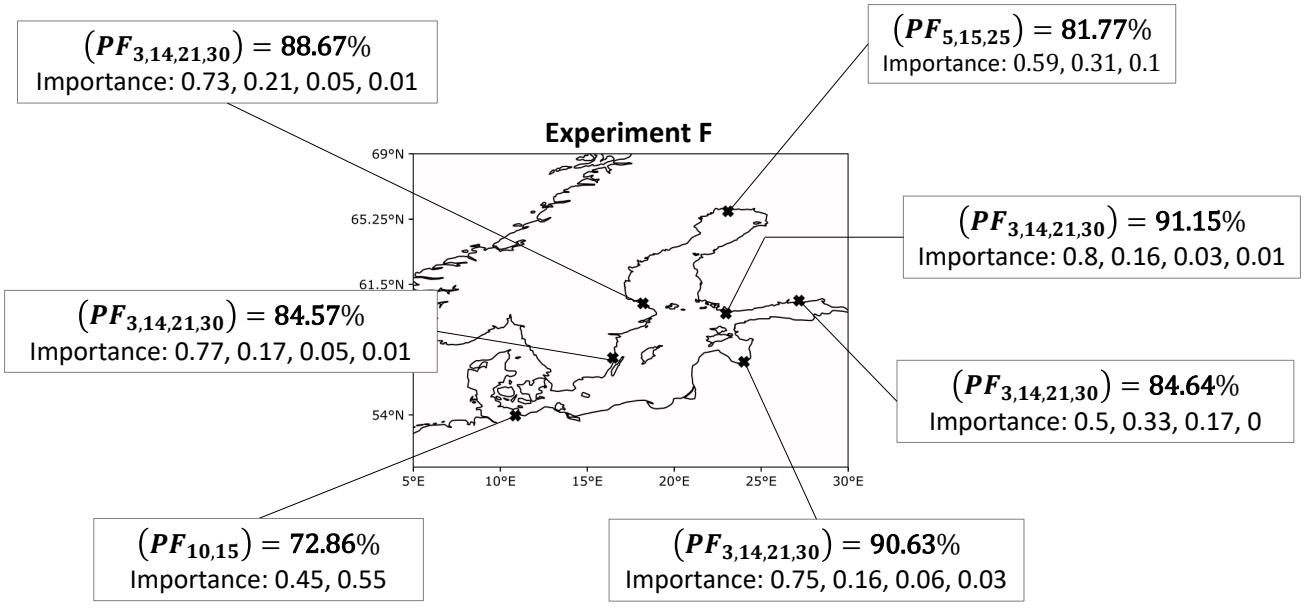


Figure 18. Best combinations in terms of TTPR of PF for experiment F. Importances are ordered as subscripted timelags.

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

5.7 Benchmark

Ideally, we should compare the performance of the Random Forest algorithm with storm-surge predictions obtained with hydrodynamical models driven by the atmospheric forcing. There are, however, several obstacles to this comparison. For a fair comparison, the hydrodynamical predictions for day d should be driven only with information available up to day d *minus timelag*, with varying time lags, or at most including numerical weather predictions up to day d. Those predictions, probably conducted by the respective hydrographic services of the different Baltic countries, are not available to us. Instead, we benchmarked our Random Forest algorithm against hydrodynamical modelling storm-surge reanalysis that, in principle, should be superior since the storm-surge dynamical model uses all the information available without any time restrictions, even after day d.

The storm-surge reanalysis that we used as benchmark are the global reanalysis of storm surges done by Muis et al. (2016). Based on hydrodynamic modelling, they presented the first global reanalysis of storm surges and extreme sea levels (GTSR data set). Their model is driven by meteorological reanalysis ERA-Interim. The spatial resolution of ERA-Interim is very close to the resolution of ERA5 from which we extract the atmospheric predictors. We will preprocess the GTSR dataset and consider it as a prediction of extreme storm surges, comparing it to our preprocessed categorical GESLA-dataset as a ground truth by computing corresponding CFMs.

The GTSR dataset consists of several model grid cells along the global coastline with daily temporal resolution. For each of the stations in our research, we selected the closest model grid cell within the GTSR dataset. From the GTSR we only selected the winter months of December, January and February (DJF) for comparison, as September was missing from the autumn months (SON). First, we intersect the GTSR time period with the GESLA dataset to ensure both datasets are on the same time domain. We then proceed by linearly detrending the datasets and finally classify them based on their respective 95th quantile. This leads to two datasets with daily temporal resolution, containing categorical entries indicating whether a storm surge occurs at a specific time or not. We consider the preprocessed GTSR dataset as predictions of extreme storm surges and the GESLA dataset as the ground truth when computing the CFMs. The resulting TPRs of GTSR are compared to the best TPR of our study in Table 1. The results confirm the previous findings by Muis et al. (2016), namely that extreme storm surges are often underestimated by GTSR. For all stations analysed in this study, our ML approach performs better than the GTSR.

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

Station	GTSR TPR	Best RF TPR (ERA5)
0 (NSWE)	27,87%	73,33% (V10)
1 (FIN)	63,36%	71% (U10)
2 (FINBAY)	48,85%	78,9% (SP)
3 (LVA)	41,98%	82,61% (U10)
4 (DEU)	35,11%	60.87% (V10)
5 (WSWE2)	7,63%	70% (SP)
6 (WSWE)	10%	71,43% (U10)

Table 1. Comparison of the TPRs for the GTSR-prediction and the RF-prediction based on the validation set $\mathcal{M}_{\mathcal{V}}$. Note that for stations 3 and 4 the TPRs for the RF are based on the testset $\mathcal{M}_{\mathcal{T}}$ and not the validation set.

6 Discussion

The theory indicates that one predictor alone should not be sufficient to describe storm surges. The main features are the wind stress and the low-pressure systems (below 980 hPa) as well as their speeds. While our models also showed good results when using isolated predictors, they showed more robust results when using them in combination. Furthermore, for almost all stations (except station 4) surface pressure and west wind were the most important ERA5-predictors. Our model results suggest that mostly low-pressure fields below 980 hPa and strong (mean) west winds of 10 ms^{-1} around the area of the Danish Straits lead to TPPs, especially for stations located in the Northeast of the Baltic Sea. For those stations the AoI of U10 was situated South of the Danish Straits reaching inland towards mid-Germany. This can actually be explained by predominant South-Westerly winds in winter months, which eventually push water masses towards the Northeast. Furthermore, PMs showed (when looking beyond the AoI) that those strong west winds often acted on a large horizontal distance, which according to Weisse and von Storch (2010) increases the potential of storm surges. It is this wind direction that leads to the fact that U10 as well as SP did not lead to any good predictions for station 4 (DEU). This is theoretically sound as for stations in the Southwest of the Baltic Sea water is pushed away towards the Northeast due to winds and baric waves. By contrast, those stations should be more subject to negative storm surges, which we did not investigate in this study.

For southern stations, north-easterly wind should rather be a predominant factor. We saw this for station DEU, where the most important predictor in terms of feature importance was V10.

According to Leppäranta and Myrberg (2009), the largest amount of precipitation is found on the eastern coast of the Baltic Sea due to the winds blowing mostly eastward in wintertime. We could not recover this for our model. If any structure at all could be obtained from AoIs of TP, it was the importance around the area of Bergen and the UK. Also corresponding PMs of TP did not show stronger rain in the eastern coast of the Baltic Sea. By contrast, Gönnert et al. (2001) states that the influence

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

of precipitation is not directly related to storm surge magnitudes but rather alters preconditions like the prefilling of the Baltic Sea and the filling of rivers and estuaries. Together with the fact that west winds are not as strong when a condition of prefilling exists, prefilling itself should be of great usage as a predictor. For almost all stations, this was actually true. Compared to other ERA5-predictors PF was generally leading to better TPRs on the validation set.

Sometimes, our model showed patterns for AoI and PMs, though, that were hard to explain by theory. For instance, for station NSWE, low-pressure fields in the European North Sea were of great importance instead of low-pressure systems close to the station (see Fig 8). This behaviour showed up mainly when using time lags of several days. Theoretically, low-pressure systems in those areas move towards the East, i.e. in the direction of the station, which might be one possible explanation. Additionally, for some cases, storm surges were observed but not predicted (FNPs) - for instance under the presence of high-pressure fields. This behavior was not due to SP being a sole predictor as we could observe the same behaviour when accounting for a combination of all predictors. The timelag of the predictors might have been too long, such that high-pressure systems could turn into low-pressure systems before the actual day of the storm surge. One idea to overcome this problem is to use hourly gradients of atmospheric pressure as predictors, which indicate a rapid (de-) intensification of low-pressure systems (similar to Bruneau et al. (2020)).

Nevertheless, we saw that time lagging the predictors improved model results. This is in alignment with Tyralis et al. (2019), who showed that Random Forests worked better when time-lagged predictors were used. In general, timelags up to 4 days worked quite reasonably, while longer timelags did not add much value to VTPRs. For instance, a time lag of 2 or 3 days for U10 was often the best choice. This is what we expected, especially for north-eastern stations, as deep-water waves need approximately two days to travel across the Baltic Sea. Furthermore, we saw that implementing a longer time lag for U10 compared to SP leads to good results in terms of VTPRs when using both in combination. For PF, mostly short-term time lags work best, but still, it was possible to even increase the time-lag up to a week. This contradicts the actual definition of prefilling, and one might argue against the usage of the plain time-series of water recordings at Degerby as a plausible predictor.

Some caveats of our model need to be mentioned. First of all, we only use a period of 3 months over nine years to generate train and test data. But Bruneau et al. (2020) showed that for Machine Learning, specifically Artificial Neural Networks, 6-7 years of daily training data is necessary. We only used a total of 18 months, though. In order to overcome this, one could extend the dataset to longer time periods. Using more data increases computing time, which is one reason why we did not implement it. Our main objective was to design a relatively simple prediction scheme that would not need heavy computing resources. However, in view of the results obtained, the algorithm could be trained using more data with more powerful resources.

Furthermore, algorithms trained with predictors based on remotely sensed data outperformed algorithms trained with predictors obtained from reanalysis data by (Tyralis et al., 2019). We used only reanalysis data as predictors. If data sources with remotely sensed data are available, testing the algorithm on them would be better.

For future studies within this context, it would be interesting to alter and specify some of the predictors. For instance, instead of only using U10 and V10 one could actually calculate all of the wind stresses, i.e. the wind direction, wind velocity and its duration. Our dataset did not involve the duration, which is especially important for the generation of surface waves and swell.

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

Furthermore, we did not use wind directions per se as a predictor but rather the zonal and meridional wind speeds. One could calculate wind directions of those datasets and use it as a new predictor. Similarly, if low-pressure systems move at relatively high velocities, i.e. greater than 16 ms^{-1}, a sub pressure-driven storm surge occurs (Wolski and Wisniewski, 2021), because the effect of the baric wave is stronger than the one of the wind. We did not use the speed nor the trajectory of a low-pressure system as model input. But this can be important as it induces resonance coupling and gives direction to the induced baric wave. Another physical change that can be made is to look at negative storm surges instead of positive ones and see if behaviours of U10 and SP change for stations like DEU. For instance, the bays of Mecklenburg and Kiel ran into strong negative storm surges due to water outflow caused by low-pressure systems moving towards the East (Wolski and Wisniewski, 2020).

From a technical perspective, one could adjust the definition, i.e. the binary encoding of storm-surges, to represent the alarming levels of specific stations instead of using percentiles. It would also be highly interesting to extend the usage of the RF to Random Regression Forests in order to investigate and predict actual heights of water level during storm surges. Further Tiggeloven et al. (2021) showed promising results using Deep-Learning-Methods when those models are tailored for specific regions. Additionally, a more common approach in the ML literature is to supply all the considered input predictors to the RF model and let the model itself decide which combinations and connections are important. We did not apply this and rather backed the choice of predictors with the underlying theory of storm surge development. Hence, we only tested combinations of predictors that were in line with the theoretical explanation of storm surges. We then wanted to infer the spatial patterns of physical predictors within the research area and their importance compared to each other. Nevertheless, complicating the model architecture can also lead to promising results and broaden the application purposes of this study.

7 Conclusion

In this study, we designed a prediction scheme for the occurrence of storm surges, i.e. the top daily 5% coastal water levels, for seven stations across the Baltic Sea. The prediction horizon is a few days, and the method is based on a Random Forest used as a binary classifier. The method was tested on records of the water level at respective stations from GESLA3 and atmospheric predictors were taken from the ERA5 dataset, from which we choose variables of surface pressure (SP), zonal (U10) and meridional (V10) windspeeds at 10 meters above the Earth's surface and total precipitation (TP). Despite its relative simplicity, the purely data-driven Random Forest binary classifier is able to predict the occurrence of storm surges in the Baltic Sea with a few days lead time with high sensitivity. The method is able to identify the relevant predictors and the relevant regions among a set of atmospheric variables, agreeing with physical expectations. The RF method is able to discriminate the predictors according to the station location. For stations at zonal extends of the Baltic Sea, U10 and SP were the most important predictors, showing strong west winds and pronounced low-pressure systems when modelling extreme storm surges. For stations at meridional extends, the importance of V10 increases.

Westwind around the Danish Straits often indicated the onset of an extreme storm surge, probably due to its influence on the Baltic Seas prefilling. We could also recover the fact that with increased prefilling the importance of west winds tends to be weaker in cases of storm surges. Increasing predictor lead times decreased model accuracy. The method works well for

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

lead times of up to three days. Combining several time lags of information works better than using the different lead time information in isolation.

Hence this study shows that the drivers of storm surges across the Baltic Sea depend on the locality of the event. Due to its brief computing time it can be used as an auxillary model that informs on the necessity to run more complex operational, numerical models.

Future research could extent the model for instance by using more sophisticated predictors, changing the predictand to a continuous waterlevel or switching the predictive scheme from deterministic to probabilistic in order to evaluate uncertainties of predictions.

Code and data availability. The code is available at the reference Bellinghausen (2022). The ERA5 data are publicly available from reference Hersbach et al. (2018). The GESLA data are available at https://gesla787883612.wordpress.com

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

Appendix A: Tables

Number of station	GESLA code	Identifier
0	"kalixstoron-kal-swe-cmems"	NSWE
1	"hanko-han-fin-cmems"	FIN
2	"hamina-ham-fin-cmems"	FINBAY
3	"daugavgriva-dau-lva-cmems"	LVA
4	"travemuende-tra-deu-cmems"	DEU
5	"oskarshamn-osk-swe-cmems"	WSWE
6	"forsmark-for-swe-cmems"	WSWE2

Table A1. Number of stations as in Fig 2 and corresponding code in GESLA dataset.

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

Name	Units	Short Description
sp	Pa	Pressure (force per unit area) of the atmosphere on the surface of land, sea and in-land water. It is measured by the weight of total air in a vertical column above the area of the Earth's surface.
tp	m	Accumulated liquid and frozen water that falls to the Earth's surface. It represents the sum of large-scale precipitation and convective precipitation. The units indicate the depth the water would have when evenly spread over the grid box.
u10	ms^{-1}	Eastward component of the 10m wind, i.e. the horizontal speed of air moving towards the east at a height of ten metres above the Earth's surface.
v10	ms^{-1}	Northward component of the 10m wind, i.e. the horizontal speed of air moving towards the north at a height of ten metres above the Earth's surface

Table A2. Variables of ERA5 dataset used as predictors. Description of data is taken from the parameter database of the official ECMWF website.

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

Parameter	Value	Short Description
n_estimator	[333, 666, 1000]	Number of DTs used within a RF.
max_depth	[1, 2, 3]	Depth of each DT.
class_weight	"balanced"	Associated weighting of each class.
oob_score	"True"	Calculating out-of-bag sample scores for each DT.
optimizer	"RandomSearchCV"	Functionality to find best combination of hyperparameters. Optionally "GridSearchCV" can be used.
k	3	k-fold cross-validation used by optimizer.
n_iter	100	Number of parameter settings that are sampled by "RandomSearchCV". Trades off runtime against quality of the solution.

Table A3. Parameters used to find optimal hyperparameters of the random forest. When multiple values are given, the optimizer chooses the best combination amongst those.

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

Experiment: **A**		
Run_Id	**Predictors**	**Timelags (in days)**
0 – 4	SP, TP, U10, V10, PF	no timelag, i.e. 0
5 – 9	SP, TP, U10, V10, PF	all with timelag 1
10 – 14	SP, TP, U10, V10, PF	all with timelag 2
15 – 19	SP, TP, U10, V10, PF	all with timelag 3
20 – 24	SP, TP, U10, V10, PF	all with timelag 4
25 – 29	SP, TP, U10, V10, PF	all with timelag 5
30 – 34	SP, TP, U10, V10, PF	all with timelag 6
35 – 39	SP, TP, U10, V10, PF	all with timelag 7

Table A4. Parameters and timelags used for experiment **A**. All predictors are used in isolation, no combinations are used.

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

Experiment: **B**		
Run_Id	**Predictors**	**Timelags (in days)**
0	(SP, TP, U10, V10)	(1, 1, 1, 1)
1	(SP, TP, U10, V10, PF)	(1, 1, 1, 1, 1)
2	(SP, TP, U10, V10)	(2, 2, 2, 2)
3	(SP, TP, U10, V10, PF)	(2, 2, 2, 2, 2)

Table A5. Parameters and timelags used for experiment **B**. Parentheses indicate that predictors are used in combination.

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

Experiment: **C**		
Run_Id	**Predictors**	**Timelags (in days)**
0, 1, ... , 7	(SP, U10), (SP, U10), ... , (SP, U10)	(0, 0), (1, 1), ... , (7, 7)
8, 9, 10	(SP, U10), (SP, U10), (SP, U10)	(2, 3), (2, 4), (2, 5)
11	(SP, SP, U10, U10)	(1, 3, 1, 5)
12	(SP, SP, U10, U10)	(1, 4, 1, 6)
13	(SP, SP, U10, U10)	(1, 5, 1, 7)

Table A6. Parameters and timelags used for experiment **C**. Parentheses indicate that predictors are used in combination.

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

Experiment: **D**		
Run_Id	**Predictors**	**Timelags (in days)**
0 – 3	all (U10, U10)	(1, 2), (2, 3), (2, 4), (3, 6)
4 – 7	all (U10, U10, U10)	(1, 2, 3), (2, 3, 4), (3, 4, 5), (5, 6, 7)
8 – 11	all (U10, U10, U10, U10)	(1, 2, 3, 4), (4, 5, 6, 7), (1, 3, 5, 7), (1, 7, 14, 21)

Table A7. Parameters and timelags used for experiment **D**. Parentheses indicate that predictors are used in combination.

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

Experiment: **E**		
Run_Id	**Predictors**	**Timelags (in days)**
0	(TP, TP, TP, U10)	(7, 5, 2, 2)
1	(U10, PF, PF, PF)	(3, 7, 5, 2)
2	(U10, U10, PF)	(5, 2, 7)

Table A8. Parameters and timelags used for experiment **E**. Parentheses indicate that predictors are used in combination.

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

Experiment: **F**		
Run_Id	**Predictors**	**Timelags (in days)**
0 – 3	all PF	10, 15, 20, 25
4 – 6	all (PF, PF)	(5, 10), (10, 15), (20, 25)
7, 8	all (PF, PF, PF)	(5, 15, 25), (7, 14, 21)
9	(PF, PF, PF, PF)	(3, 14, 21, 30)

Table A9. Parameters and timelags used for experiment **F**. Parentheses indicate that predictors are used in combination.

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

Author contributions. All authors contributed to develop the original research goal, analysed and discussed the results. K.B. coded the software, carried out the data analysis, and drafted the initial versions of the manuscript. B.H and E.Z contributed to the later and final version of the manuscript.

Competing interests. We declare that we have no competing interests.

Acknowledgements. The ERA5 data (Hersbach et al., 2018) were downloaded from the Copernicus Climate Change Service (C3S) Climate Data Store. The results contain modified Copernicus Climate Change Service information 2020. Neither the European Commission nor ECMWF is responsible for any use that may be made of the Copernicus information or data it contains.

We also thank the GESLA project (https://gesla787883612.wordpress.com/) for making the extreme sea level data sets available for the scientific community.

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

References

Andrée, E., Drews, M., Su, J., Larsen, M. A. D., Drønen, N., and Madsen, K. S.: Simulating wind-driven extreme sea levels: Sensitivity to wind speed and direction, 36, 100 422, https://doi.org/10.1016/j.wace.2022.100422, 2022.

Bellinghausen, K.: A Random Forest for Extreme Storm Surge Prediction in the Baltic Sea, https://doi.org/10.5281/zenodo.7409633, 2022.

Bevacqua, E., Maraun, D., Vousdoukas, M. I., Voukouvalas, E., Vrac, M., Mentaschi, L., and Widmann, M.: Higher probability of compound flooding from precipitation and storm surge in Europe under anthropogenic climate change, Science Advances, https://doi.org/10.1126/sciadv.aaw5531, publisher: American Association for the Advancement of Science, 2019.

Breiman, L.: Random Forests, Machine Learning, 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

Bruneau, N., Polton, J., Williams, J., and Holt, J.: Estimation of global coastal sea level extremes using neural networks, Environmental Research Letters, 15, 074 030, https://doi.org/10.1088/1748-9326/ab89d6, publisher: IOP Publishing, 2020.

Chen, D. and Omstedt, A.: Climate-induced variability of sea level in Stockholm: Influence of air temperature and atmospheric circulation, 22, 655–664, https://doi.org/10.1007/BF02918709, 2005.

Eakins, B. W. and Sharman, G. F.: Volumes of the World's Oceans from ETOPO1, https://www.ngdc.noaa.gov/mgg/global/etopo1_ocean_volumes.html, publisher: U.S. Department of Commerce, 2010.

Field, C. B., Barros, V., Stocker, T. F., and Dahe, Q., eds.: Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation: Special Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, https://doi.org/10.1017/CBO9781139177245, 2012.

Guillory, A.: ERA5, https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5, 2017.

Géron, A.: Hands-On Machine Learning with Scikit-Learn and TensorFlow, 2017.

Gönnert, G. and Sossidi, K.: A new approach to calculate extreme storm surges: analysing the interaction of storm surge components, pp. 139–150, Naples, Italy, https://doi.org/10.2495/CP110121, 2011.

Gönnert, G., Dube, S. K., Murty, T., and Siefert, W.: Die Küste, 63 Global Storm Surges, Boyens Medien GmbH & Co. KG, Heide i. Holstein, 2001.

Haigh, I. D., Marcos, M., Talke, S. A., Woodworth, P. L., Hunter, J. R., Hague, B. S., Arns, A., Bradshaw, E., and Thompson, P.: GESLA Version 3: A major update to the global higher-frequency sea-level dataset, n/a, https://doi.org/10.1002/gdj3.174, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/gdj3.174, 2021.

Harris, D. L.: THE EQUIVALENCE BETWEEN CERTAIN STATISTICAL PREDICTION METHODS AND LINEARIZED DYNAMICAL METHODS, 90, 331–340, https://doi.org/10.1175/1520-0493(1962)090<0331:TEBCSP>2.0.CO;2, publisher: American Meteorological Society Section: Monthly Weather Review, 1962.

Harris, D. L.: Characteristics of the Hurricane Storm Surge, 1963.

Holfort, J., Wisniewski, B., Lydeikaite, Z., Kowalewska-Kalkowska, H., Wolski, T., Boman, H., Hammarklint, T., Giza, A., and Grabbi-Kaiv, S.: Extreme sea levels at selected stations on the Baltic Sea coast, Oceanologia; 2014; Vol. 56; Iss. 2, https://journals.pan.pl/dlibra/publication/115021/edition/100074, publisher: Instytut Oceanologii PAN, 2014.

Hünicke, B. and Zorita, E.: Influence of temperature and precipitation on decadal Baltic Sea level variations in the 20th century, Tellus A: Dynamic Meteorology and Oceanography, 58, 141–153, https://doi.org/10.1111/j.1600-0870.2006.00157.x, 2006.

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

Hünicke, B., Zorita, E., Soomere, T., Madsen, K. S., Johansson, M., and Suursaar, : Recent Change—Sea Level and Wind Waves, in: Second Assessment of Climate Change for the Baltic Sea Basin, edited by The BACC II Author Team, Regional Climate Studies, pp. 155–185, Springer International Publishing, Cham, https://doi.org/10.1007/978-3-319-16006-1_9, 2015.

Janssen, F., Schrum, C., Hübner, U., and Backhaus, J.: Uncertainty analysis of a decadal simulation with a regional ocean model for the North Sea and Baltic Sea, Climate Research, 18, 55–62, https://doi.org/10.3354/cr018055, 2001.

Leppäranta, M. and Myrberg, K.: Physical oceanography of the Baltic Sea, Springer Praxis books geophysical sciences, Springer, Berlin Heidelberg, 2009.

Mohrholz, V.: Major Baltic Inflow Statistics – Revised, 5, https://www.frontiersin.org/article/10.3389/fmars.2018.00384, 2018.

Mudersbach, C. and Jensen, J.: Küstenschutz an der Deutschen Ostseeküste - Zur Ermittlung von Eintrittswahrscheinlichkeiten extremer Sturmflutwasserstände, 5, https://doi.org/10.3243/kwe2010.03.003, 2010.

Muis, S., Verlaan, M., Winsemius, H. C., Aerts, J. C. J. H., and Ward, P. J.: A global reanalysis of storm surges and extreme sea levels, Nature Communications, 7, 11 969, https://doi.org/10.1038/ncomms11969, 2016.

Müller, A. C.: Introduction to Machine Learning with Python, 2017.

Rutgersson, A., Kjellström, E., Haapala, J., Stendel, M., Danilovich, I., Drews, M., Jylhä, K., Kujala, P., Larsén, X. G., Halsnæs, K., Lehtonen, I., Luomaranta, A., Nilsson, E., Olsson, T., Särkkä, J., Tuomi, L., and Wasmund, N.: Natural hazards and extreme events in the Baltic Sea region, Earth System Dynamics, 13, 251–301, https://doi.org/10.5194/esd-13-251-2022, 2022.

Sztobryn, M.: Forecast of storm surge by means of artificial neural network, Journal of Sea Research, 49, 317–322, https://doi.org/10.1016/S1385-1101(03)00024-8, 2003.

Tadesse, M., Wahl, T., and Cid, A.: Data-Driven Modeling of Global Storm Surges, Frontiers in Marine Science, 7, 260, https://doi.org/10.3389/fmars.2020.00260, 2020.

Tiggeloven, T., Couasnon, A., van Straaten, C., Muis, S., and Ward, P. J.: Exploring deep learning capabilities for surge predictions in coastal areas, Scientific Reports, 11, 17 224, https://doi.org/10.1038/s41598-021-96674-0, 2021.

Tyralis, H., Papacharalampous, G., and Langousis, A.: A Brief Review of Random Forests for Water Scientists and Practitioners and Their Recent History in Water Resources, 11, 910, https://doi.org/10.3390/w11050910, number: 5 Publisher: Multidisciplinary Digital Publishing Institute, 2019.

von Storch, H.: Storm Surges: Phenomena, Forecasting and Scenarios of Change, Procedia IUTAM, 10, 356–362, https://doi.org/10.1016/j.piutam.2014.01.030, 2014.

Vousdoukas, M. I., Voukouvalas, E., Annunziato, A., Giardino, A., and Feyen, L.: Projections of extreme storm surge levels along Europe, Climate Dynamics, 47, 3171–3190, https://doi.org/10.1007/s00382-016-3019-5, 2016.

Weisse, D. R.: Klimatologie der Ostseewasserstände: Eine Rekonstruktion von 1948 bis 2011, p. 132, 2014.

Weisse, R. and Hünicke, B.: Baltic Sea Level: Past, Present, and Future, in: Oxford Research Encyclopedia of Climate Science, Oxford University Press, https://doi.org/10.1093/acrefore/9780190228620.013.693, 2019.

Weisse, R. and von Storch, H.: Marine Climate and Climate Change, Springer Berlin Heidelberg, Berlin, Heidelberg, https://doi.org/10.1007/978-3-540-68491-6, 2010.

Weisse, R. and Weidemann, H.: Baltic Sea extreme sea levels 1948-2011: Contributions from atmospheric forcing, Procedia IUTAM, 25, 65–69, https://doi.org/10.1016/j.piutam.2017.09.010, 2017.

Wisniewski, B. and Wolski, T.: Physical aspects of extreme storm surges and falls on the Polish coast, Oceanologia, 53, http://yadda.icm.edu.pl/yadda/element/bwmeta1.element.dl-catalog-470a75c2-d847-417b-91a4-10e0bbfcb266, publisher: -, 2011.

https://doi.org/10.5194/egusphere-2024-2222
Preprint. Discussion started: 13 August 2024

WMO: Guide to storm surge forecasting, WMO, Geneva, oCLC: 1075529493, 2011.

Wolski, T. and Wisniewski, B.: Geographical diversity in the occurrence of extreme sea levels on the coasts of the Baltic Sea, Journal of Sea Research, 159, 101 890, https://doi.org/10.1016/j.seares.2020.101890, 2020.

Wolski, T. and Wisniewski, B.: Characteristics and Long-Term Variability of Occurrences of Storm Surges in the Baltic Sea, Atmosphere, 12, 1679, https://doi.org/10.3390/atmos12121679, number: 12 Publisher: Multidisciplinary Digital Publishing Institute, 2021.

Woodworth, P. L., Hunter, J. R., Marcos, M., Caldwell, P., Menéndez, M., and Haigh, I.: Towards a global higher-frequency sea level dataset, 3, 50–59, https://doi.org/10.1002/gdj3.42, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/gdj3.42, 2016.