Estimating Extreme Storm Surge Height in the Baltic Sea by Using Various Machine Learning Methods Based on Atmospheric Data

—

A binary classification based on Random Forests

Kai Bellinghausen (Matrikelnr.: 7383766)

Supervisors: Corinna Schrum, Birgit Hünicke, Eduardo Zorita

Institute: Institute of Analysis and Modelling of Coastal Systems, Helmholtz-Zentrum Hereon

October 30, 2022

Contents

1 Introduction

Storm surges, i.e. an extreme increase in sea level due to storms, are a major natural hazard for coastal societies. They do not only impose severe damage to infrastructure at coastlines but can also be deadly to humans. Hence, monitoring and forecasting systems for storm surges are applied to coastal regions in order to inform decision makers and act as a warning system for societies.

Usually, these forecasting systems are relying on dynamical ocean-atmosphere models. While those models can predict the general sea level quite well, they struggle with forecasting extreme events. One key reason for this is the fact, that the energy transfer between ocean- and atmosphere-models is not yet well incorporated into dynamical models. Hence, the effect of extreme storm events, e.g. the kinetic energy transfer from the atmosphere to the ocean surface, is not well described.

Alternatively to dynamical models, forecasting systems can be based on data-driven models. These models are not representing the physical dynamics of a process but try to extract insight about it in form of patterns in the underlying data. The advantage of data-driven models is not only that they often are computationally more efficient than dynamical models, but they also incorporate non-linearities of physical processes.

Machine Learning (ML) is one example of data-driven modelling and is becoming more popular in climate sciences. Several studies applied ML-methods in order to analyze storm surges with promising results (Bruneau et al. (2020), Tadesse et al. (2020), Tiggeloven et al. (2021), Gönnert and Sossidi (2011), Sztobryn (2003)). Though most of these studies did not specifically analyze extreme sea levels.

Hence, we will apply ML-methods to predict extreme storm surges measured by the 95th-percentile of sea-level measurements taken from the Global Extreme Sea Level Analysis (GESLA)3-project (Haigh et al. (2021)). As by nature data-driven models need large data-sets, we will investigate the area of the Baltic Sea. This area is known for a broad coverage in atmosphere- and ocean-measurements.

As predictors we will use ERA5-reanalysis data, specifically the wind-field, total precipitation and surface pressure. Additionally, we will use a proxy for the filling of the Baltic Sea as a predictor, given by the current sea level at the tide-gauge station at Degerby taken from GESLA3. Ideally, we want to provide an operational forecast for storm-surge height in the Baltic Sea. Due to time restrictions though, this thesis only provides a binary classification based on random forests.

The remainder of this study is structured as follows. In Section 2 we will describe the characteristics of sea level variability in the Baltic Sea and the underlying physical properties. In Section 3 we will introduce current operational forecasting systems for the Baltic Sea and review to what extent ML was used to model storm surges in general. We will then introduce our modelling approach in Section 4 and show the results in Section 5. Our results will be discussed and related to current studies and the theory on storm surges in Section 6. The study ends with a conclusion.

2 Storm Surges and the Baltic Sea

2.1 Definition of Storm Surges

Storm Surges are defined in several ways, ranging from very broad definitions to more specific ones. Field et al. (2012) define storm surges as the excess of sea level above the expected tidal sea level at a specific time in a certain location. They see a storm surge as a temporary elevation of sea level forced by extreme atmospheric conditions, mainly low atmospheric pressure and strong winds. This definition is also used by Harris (1963) and the Encyclopaedia of Coastal Science (Wolski et al. (2016)). Weisse and von Storch (2010) refer to this definition as the non-tidal or meteorological residuals but add to this the influence of the Mean Sea Level (MSL) as follows

$$h_t = t_t + s_t + l_t. \tag{1}$$

Here h_t, t_t, s_t, l_t are the observed sea level height, the tides, surges and changes in MSL, respectively. Note that this leaves room to distinguish between positive and negative storm surges, where an extreme increase of observed sea level follows from the first and a significant decrease of sea level from the latter. In this case, and also in other definitions of storm surges, wind waves are ignored by incorporating the MSL, which is calculated by comparing the actual sea level to a fixed benchmark (the relative sea level) and averaging over multiple years.

The sea level variations due to a storm surge are generally short term, lasting for only minutes up to several days (Wisniewski and Wolski (2011), WMO (2011)). It is well researched that storm surges are accompanied by strong winds and low-pressure systems (Weisse and von Storch (2010), Wolski and Wisniewski (2021), WMO (2011)). Hence, the forcing of storm surges ranges from synoptic scales (cyclones) to mesoscale systems like squall lines (WMO (2011)). While wind pushes water-masses towards or away from coastlines, a low pressure field lifts up the sea-surface. When the forcing stops, the relaxation may lead to oscillating motions of the sea-level. Gönnert et al. (2001), Wisniewski and Wolski (2011) and WMO (2011) use this idea of oscillations to broaden the definition of storm surges as oscillations of the sea level within coastal areas and inland water masses.

Technically, the storm-surge problem is an air-sea interaction problem, where the atmosphere forces the water body, 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 again alter the wind-field (Gönnert et al. (2001)). Hence, the underlying processes of storm surges are highly non-linear. Figure 2 shows the ocean wave spectrum, from which one can classify surges as long gravity waves with a period of 3 hours (Gönnert et al. (2001)). The period and amplitude of storm surges vary substantially with specific local conditions like the topography of the ocean basin, the wind-speed and its direction, the extent of ice-cover, total precipitation as well as the direction of the storm track surpassing the basin and the shape of coastal estuaries (Gönnert et al. (2001), WMO (2011), Weisse and von Storch (2010), Muis et al. (2016)). For instance, in deep water a storm surge travels faster than the weather system in the atmosphere. When the storm surge approaches shallower water, it slows down (Harris (1963)) and eventually matches the speed of the weather system (WMO (2011). From this resonance coupling occurs, which transfers energy from the atmosphere to the oceans surface and thus amplifies the storm surge. Additionally, the energy

of the storm surge is compressed into a shorter, vertical water-column once it enters the shallow waters from the deeper parts of the basin (WMO (2011)). Because local characteristics are important, storm surge models are usually fitted to them. Hence, in the following section, we will further investigate the characteristics of the Baltic Sea.

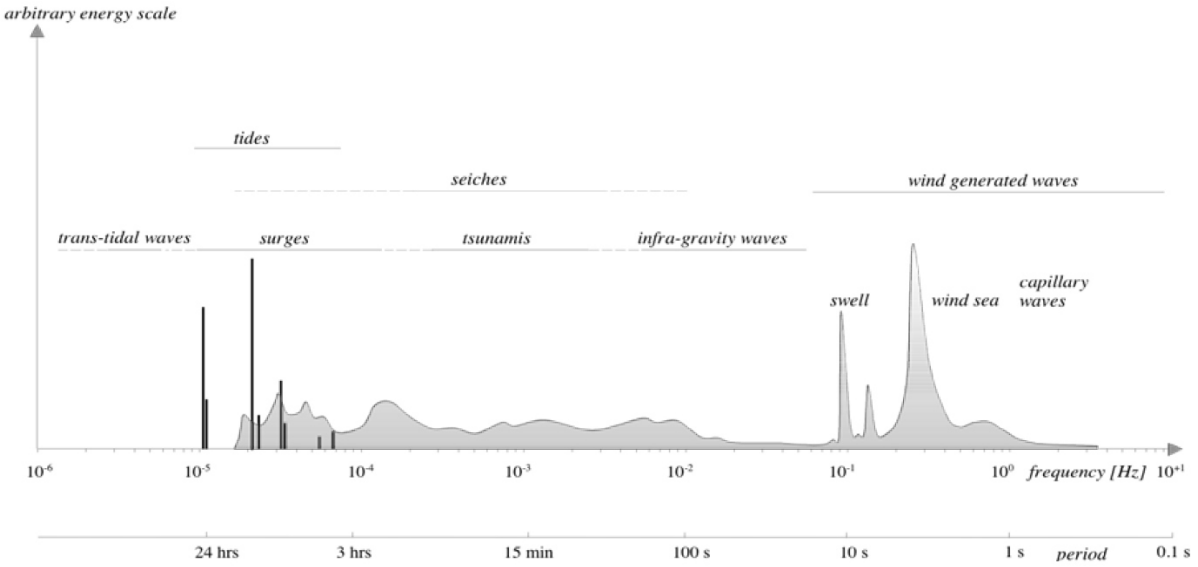


Figure 2: Frequencies and periods of the vertical motions of the ocean surface taken from Gönnert et al. (2001).

2.2 Characteristics of the Baltic Sea

2.2.1 Spatial Extent

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)). 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 leave only internal tides of a few centimeter within the Baltic Sea (Rutgersson et al. (2021), Wolski and Wisniewski (2021)). Due to the very narrow connection to the Atlantic, storm surges are only induced internally (Weisse and Hünicke (2019)).

The surface area of the Baltic Sea is of about 377.000 km^2 to 412.500 km^2, depending on whether one includes the Kattegat or not. According to Eakins and Sharman (2010) the volume of the Baltic Sea is of about 20.900 km^3. Note though that the volume is strongly influenced by the inflow of water through the Straits of Denmark. Depending on the filling of the Baltic Sea, its depth changes. The average depth is about 55m, which is due to many shallow coastal areas (Weisse and von Storch (2010), Leppäranta and Myrberg (2009)). Countries directly connected to the Baltic Sea (shown in Figure 3) are hence prone to storm surges. The risk of storm surges among those countries varies considerably due to the large meridional extent of the Baltic Sea and the different orientation of coastlines (Hünicke et al. (2015)).

Figure 3: Subbasins of the Baltic Sea as indicated in Wolski and Wisniewski (2020)

.

2.2.2 Practical Definitions of Storm Surges in the Baltic Sea

In the Baltic Sea the definition of a storm surge varies with the alarming system of particular coastlines, because a similar elevation of sea level impacts coastlines differently due to their altering shapes of bathymetry. For instance the Polish coastal protection service defines a storm surge by a dynamic increase of the sea level above the alarm or warning level due to the action of wind and atmospheric pressure (Wisniewski and Wolski (2011)). The warning level is set to 570cm N.N. corresponding to an increase of more than 70 cm above the local tide gauge zero. The alarming level is set to 600 cm N.N.. In comparison, for the German coast 600 cm N.N. is the warning and not the alarming level (Holfort et al. (2014)). The measurements of sea level are based on the European Vertical Reference System (EVRS) (Ihde et al. (2002)) in order to have a consistent interpretation of sea level elevation amongst all tide gauges in the Baltic Sea. As Wolski et al. (2016) pointed out, there was no common geodetic reference datum for the Baltic Sea MSL before the EVRS, which made interpretation and comparison of tide gauge data difficult (see also Ekman (2009)). Using this standardized reference datum allows to evaluate the accuracy of operational hydrodynamic storm surge models and thus plays an important role for consistent

storm surge forecasting in the Baltic Sea.

2.2.3 Examples of Storm Surges

Historically, the number of storm surges are consistently rising (Wolski and Wisniewski (2021)) but amongst those, two were exceptionally strong.

One of the strongest and most severe storm surges in the Baltic Sea was the wind storm Gudrun in January 2005, which lasted for almost two days. The storm reached the strength of a hurricane, leading to abnormal significant wave heights of 9m (Soomere et al. (2008)). Record surge heights were observed in several coasts, for example 2.75m above MSL in Parnu, a city in Estonia. These extreme surges were explained by the resonance coupling of strong winds in combination with the weather system moving over shallow waters as explained in Section 2.1 (Leppäranta and Myrberg (2009), Suursaar et al. (2006b)). Even though storm Gudrun imposed severe damage on the Estonian coast, main features of the storm surge could be forecasted and excess damage could be prevented (Leppäranta and Myrberg (2009)).

In contrast to Gudrun hitting the east coast of the Baltic Sea, a (positive) storm surge in 1872 was flooding Danish, German and Polish coastlines in the southwest of the basin (Weisse and von Storch (2010)). Extreme high water levels of more than 3.2 m above the long-term mean were observed, which was never seen before (Weisse and Hünicke (2019)). In general, a positive storm surge is unusual in southwestern parts of the Baltic Sea due to the specific drivers of storm surges, which will be discussed in Section 2.3.

2.2.4 Sea-Level Variability

From 1961 to 2021 the duration of high sea level in the Baltic Sea has increased by 1/3 and the average number of storm surges did rise from 3.1 to 5.5 annually (Wolski and Wisniewski (2021)). One might think that also extreme storm surges, like the ones discussed in the examples above, increase in number. Those examples already indicate that meteorological factors are a main driver of extreme sea levels in the Baltic Sea. This intuition is backed by Weisse and Hünicke (2019) and Weisse et al. (2021) with the addition that also astronomical drivers (tides) might play a role to some degree. Leppäranta and Myrberg (2009) showed that the effect of the windfield on the sea level variability is of great importance. They state as well that up to 80% of annual sea level variations are due to the connection of the Baltic Sea and the North Sea via the Straits of Denmark. We will later see that this is a very important remark as it leads to a condition of prefilling in the Baltic Sea (see Section 2.3). Similar to this prefilling, Arns et al. (2015) also note the importance of a rising MSL, as it changes the depth of the Baltic Sea, which again leads to changes in the propagation of storm surges. They further state that MSL-variations interact in a non-linear manner with storm-surges. The MSL has a strong seasonal cycle, with minima occuring in spring and maxima mainly during winter (Chen and Omstedt (2005)). In southern parts of the Baltic Sea maxima occur in late summer rather than in winter (Barbosa and Donner (2016)). Further drivers of the variability in sea level are the atmospheric and seawater temperature, total precipitation, melting of sea-ice and the air pressure (Hünicke and Zorita (2006), Weisse and Hünicke (2019)). In conclusion, the Baltic Sea level varies seasonally and across a broad spatial range. The processes involved can be separated into the ones that alter the volume of the Baltic Sea, e.g. prefilling, and the ones that redistribute water

masses within the basin, e.g. effects like wind (Weisse and Hünicke (2019)).

2.2.5 Distribution of Storm Surges and Potential Drivers

Because we want to develop a model for the whole Baltic Sea, we will further investigate the historical distribution of water masses and its cause within the basin.
Long-Term water levels of the Baltic Sea and their seasonal variations were analyzed by Weisse (2014) and are summarized in Figure 4. This analysis shows that north-eastern subbasins like the Gulf of Riga, the Gulf of Bothnia and the Gulf of Finland are most likely to experience storm surges. This is explained by the shape of these subbasins in combination with the eastward trajectories of low-pressure systems and strong westerly winds (Rutgersson et al. (2021), Wolski and Wisniewski (2020), Holfort et al. (2014)). By contrast the central parts of the Baltic and the Swedish coast do not undergo strong variations in extreme water levels (Rutgersson et al. (2021), Wolski and Wisniewski (2020)). Variations in the southwestern water-levels of the Baltic can lead to positive and negative storm surges. For instance, the bays of Mecklenburg and Kiel ran into strong negative storm surges ($\leq -70cm$) due to water outflow caused by low-pressure systems moving towards the East (Wolski and Wisniewski (2020)). Seasonally the strongest increase in water-levels is expected from September to February. This seasonality strongly depends on variations in meteorological factors over the Atlantic and the Baltic Sea and the prefilling of the Baltic Sea (Weisse (2014)). Holfort et al. (2014) also refer to the influence of the bathymetry and morphology of the basins as well as the location of the tide-gauges on the diversity of the (extreme) sea levels. They also conclude that the occurrence of extreme sea levels depends mainly on three factors; the action of wind-stresses (meaning the wind-direction, wind velocity and its duration), the prefilling of the Baltic Sea, and the passage of low-pressure systems which in turn lead to seiche-like oscillations. Rutgersson et al. (2021) agree with this view but stress the importance of compound events when evaluating the risk of coastal flooding due to storm surges. According to them, every cyclone can generate a storm surge but the severity of a flooding event depends on the prefilling of the Baltic Sea. The interplay of many factors leading to storm surges was also proposed by Leppäranta and Myrberg (2009) and was further investigated by Hünicke and Zorita (2006), who showed that also temperature and precipitation may influence the sea level variability. Hence, it is important to investigate and understand all drivers of storm surges in the Baltic Sea.

2.3 Physical properties of Storm Surges in the Baltic Sea

Apart from general drivers of storm surges like wind, atmospheric pressure, the tides, the effect of waves and earth's rotation there are processes of storm surges specific to the Baltic Sea. These processes are due to the Baltic Sea being a semi-enclosed basin and range from the current state of water volume (prefilling) over the effect of the bays to the extent of ice-sheet coverage (Wolski and Wisniewski (2021)). Even though all these factors are interlinked, we will discuss them separately in the following subsections in order to get a better understanding of their specific influence on storm surges.

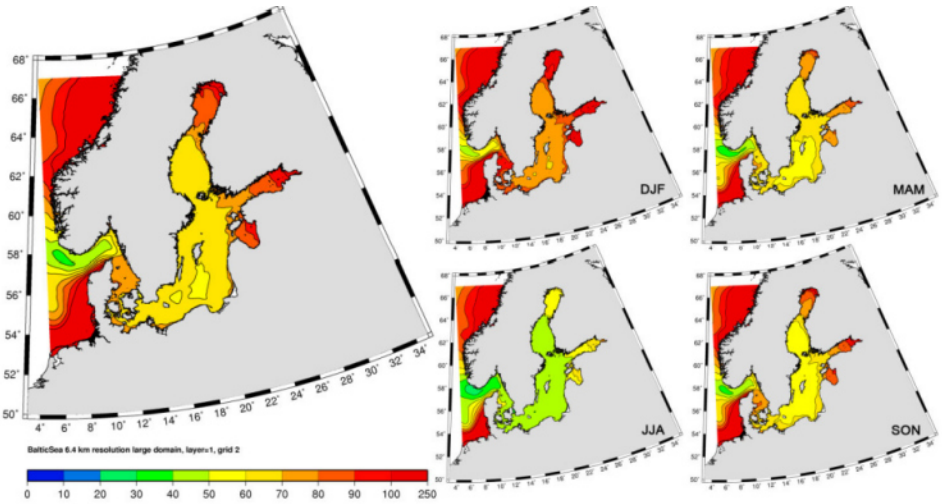


Figure 4: Left: Long-Term high water in cm above N.N. (99%-Percentiles, 1948 – 2011). Right: Seasonal high water in cm above N.N. for DJF (December, January, February), MAM (March, April, May), JJA (June, July, August), SON (September, October, November) from 1948 – 2011, taken from Weisse (2014)

.

2.3.1 Wind-effect

The already discussed seasonality of extreme water levels in the Baltic Sea is strongly connected to the stormy weather from September to April in this area, with peak values in storminess between October and February (Gönnert et al. (2001), Weisse (2014), Leppäranta and Myrberg (2009)). For the Baltic Sea Niros et al. (2002) calculated mean wintertime wind speeds of 8 ms^{-1} to 10 ms^{-1} from 1991-1999. Storm surges generated by the impact of wind-stress are called *wind-driven storm surges*. The wind-stress within a given area is further specified by the winds direction, its velocity, duration and fetch (Weisse (2014)) as well as the compensatory water flows in the inshore zones (Holfort et al. (2014)). If a wind blows consistently over several days, it deforms the sea surface and causes drift currents. Both combined may lead to a storm surge (Wolski and Wisniewski (2021)). The effect of the wind-stress on the water surface was partially described by Weisse and von Storch (2010) via

$$\frac{\partial \zeta}{\partial x} = \frac{\rho_a}{\rho} \frac{c_d u^2}{Dg}, \tag{2}$$

where $\zeta, x, \rho_a, \rho, D, c_d, g$ and u represent the sea surface height, the fetch (i.e. the horizontal distance over which the the wind forces the sea surface), the densities of the atmosphere and seawater, the water depth, the drag coefficient, acceleration of gravity and wind-speed, respectively. This equation is a simplification as it neglects the Coriolis-force due to the rotation of the Earth. Considering this influence of rotation, the net water transport would not be parallel to the wind direction but angled due to the process of *Ekman transport*. This process is especially important when winds are directed parallel to coastlines as surface currents are deflected approximately 45° to the right of the winds direction (Harris (1963)). Further modifications of Equation 2 would be due to channeling effects in large estuaries, shoaling of water masses due to changing water depth and sea-ice coverage reducing the effective fetch (Weisse and von Storch (2010)). Nevertheless, the equation provides insight into the basic effects of wind-stress on the sea surface. Usually the elevation of the sea surface is proportional to the wind-speed and increases with increasing fetch. Hence, strong winds acting on a large horizontal distance have an increasing potential

to induce storm surges. In contrast, the water depth is inversely proportional related to the sea surface elevation. This is especially important in the shallow Baltic Sea, as the mean water depth is only 55m due to the many Bay-areas and estuaries. WMO (2011) note that with decreasing water depth, surface drift currents align with the wind direction. Harris (1963) adds that this true except to areas near the coast, where the water-flow is guided by the depth contours of the bathymetry. Combined with the Ekman transport, surges become therefore largest, when the wind is blowing directly towards a coastline at shallow areas and parallel to it in deep-water regions.

Theoretically any wind-direction can be decomposed into two components, one that is normal to the coastline and the other being parallel to it. If the wind is blowing towards the coast, the normal component of the wind induces a direct *wind set-up*, which increases the sea level at the upwind shore. In contrast the water level at the downwind shore is decreased (Harris (1963)). This behavior explains the distribution of the extreme water levels in the Baltic Sea described in Section 2.2.5, where strong south westerly winds increase the water levels in north-eastern parts of the Baltic Sea and decrease it in the south-west (Weisse and von Storch (2010)). The onshore surface currents induced by the wind set-up lead to a bottom return current as mentioned by Holfort et al. (2014) and Gönnert and Sossidi (2011). As the water level rises, the strength of the return current increases from which the water level again decreases. This is one reason why storm surge events should be described by non-linear physical processes.

Wind conditions in the Baltic Sea are mainly governed by the Westerlies and the cyclonic activity in the Northern Europe-Baltic Sea area. Both of them are strongly influenced by the North-Atlantic-Oscillation (NAO)-index, which also explains parts of the variability of wind-conditions in the Baltic (Donat et al. (2010)). This is true especially during the winter month, where the winds are blowing (on average) from south-western directions (Leppäranta and Myrberg (2009), Weisse (2014)). The intensity of the western circulation is most pronounced if the meridional temperature gradient between marine and continental air masses is strong (Weisse (2014)). It is this wind direction that, if maintained for several days of a storm event, can lead to a massive inflow of water masses through the Straits of Denmark to the Baltic Sea (Gönnert et al. (2001)). This inflow can lead to an increased likelihood of storm events, which will be further discussed in Section 2.3.9. When strong westerlies stop blowing, the elevated sea surface in the north-eastern parts of the Baltic Sea relaxes and water masses flush back towards the southern and southwestern coasts. This seiche-like fluctuations may raise the water levels in the corresponding coasts (Weisse and von Storch (2010)) and are further discussed in Section 2.3.3. Similarly, a change of wind direction to north-northeasterly winds may cause extreme high water in the southern Baltic (Gönnert et al. (2001)). In a recent study Andrée et al. (2022) showed that with northerly to easterly wind directions large volumes of water masses are driven toward the Straits of Denmark, where they eventually pile up due to the bathymetry and landmasses restricting the flow. They furthermore conclude that depending on the winds duration different wind directions lead to extreme water levels at specific coastlines. For example in the Bothnian Bay the most hazardous wind direction is directed towards the coast (i.e. south-westerly) on short time scales but rotates to a stricter southerly component on longer time scales. The importance of the wind direction is also highlighted by Wolski and Wisniewski (2021). They state that the maximum of a wind driven surge will be attained faster if a weaker wind blows perpendicular to a coastline compared to a stronger wind blowing at a lower angle relative to the coast. In studies investigating surges at specific coastlines (see Wolski and Wisniewski (2021), p. 16), the effect of the

wind direction was further analyzed. For instance the highest water levels in the Estonian coast are reached for strong southwestern and western winds directed to the coast at specific angles (Jaagus and Suursaar (2013)). They also state that these winds and storm surges are connected to deep low pressure systems.

Harris (1963) also showed that the wind-stress is, due to its relation to the wind-speed, indirectly related to the pressure gradient in the atmosphere. The concept of so called *geostrophic wind* for instance assumes proportionality between the wind-speed and the pressure gradient in equilibrium conditions. In extreme storm events, this relation might be proportional to the square root of the pressure gradient. Hence, the direct wind-effect might not be sufficient to fully explain the onset of storm surges. This was recently backed by Andrée et al. (2022), who isolated wind-driven effects on storm surges and showed that only using these effects underestimated actual water levels in many cases. They conclude that other drivers are needed to explain surge water levels and the wind-effect might not even be the only main driver of storm surges. Holfort et al. (2014) also argue that besides the wind-effect, negative pressure systems ($\leq$ 980 hPa) contribute to storm surges in the Baltic Sea, if they move at a certain velocity. Therefore it is important to look at pressure-fields in the atmosphere.

2.3.2 Pressure-effect

The Baltic Sea is situated within the location of the *circumpolar low-pressure zone* (Gönnert et al. (2001)). Here, the warm Westerlies rise above the cold Polar Easterlies and as a consequence produce stormy weather due to strong temperature gradients and regions of low-pressure, especially in winter (Leppäranta and Myrberg (2009)). Usually the atmospheric sea level pressure is of 1013hPa (Weisse and von Storch (2010)). Hence, low-pressure systems are mostly associated with regions of $\leq$ 980hPa (Wolski and Wisniewski (2021), Holfort et al. (2014)). If a system like that moves at relatively high velocities ($\geq 16\text{ms}^{-1}$) a *subpressure-driven storm surge* occurs (Wolski and Wisniewski (2021)). This is due to the physical effect of the low-pressure region on the sea surface. Similar to a high-pressure system lowering the underlying sea level, a low-pressure system rises the sea surface (Weisse and von Storch (2010)). This effect is known as the *inverted barometer effect*, which is derived from the assumption of equilibrium-response of the sea surface to atmospheric pressure forcing (Weisse and von Storch (2010)) stated as

$$p_a + \rho g \zeta = \text{const.}. \tag{3}$$

Here, p_a denotes the atmospheric pressure at the sea surface (sea level pressure), ρ is the density of seawater, g is gravitational acceleration and ζ is the sea surface height. Harris (1963) noted earlier that this assumption is only true for open oceans and only if the pressure change is not too rapid, such that the total pressure beneath the surface water remains constant. If Equation 3 holds, a change in sea level pressure Δp_a will then alter the sea surface according to

$$\Delta\zeta = -\frac{\Delta p_a}{\rho g}. \tag{4}$$

Using Equation 4 one can calculate a rule of thumb, where a drop in pressure of 1hPa increases the sea level by 1cm (Wolski and Wisniewski (2021), Harris (1963)). Assuming a sea level pressure of 980hPa

the sea level would rise about 33cm relative to the height of the surface at an average pressure of 1013hPa. Compared to the quadratic relationship between the sea surface elevation and the wind-forcing (see Equation 2), the pressure-effect is sometimes argued to be of secondary importance (WMO (2011), Weisse and von Storch (2010)).

In general, the pressure-effect should not be considered to be isolated from the wind-effect. As Wolski and Wisniewski (2021) state, there is synchronous activity of both, pressure and wind, during every storm surge and they label these situations as *mixed surges* or *subpressure-wind surges*. On the one hand these two forces combined may amplify the storm surge and increase its intensity, on the other hand they may cancel each other out and decrease the severity of the storm surge (Wolski and Wisniewski (2021)). The static Equation 4 then needs to be adjusted for a more dynamic one, that includes the effects of the wind on the sea surface. In Wisniewski and Wolski (2011) the dynamic inverse barometer effect is described by the (shallow-water) wave-speed C, which represents the propagation speed of surface gravity waves induced by wind forcing, and the velocity of the pressure system V_L (see Figure 5).

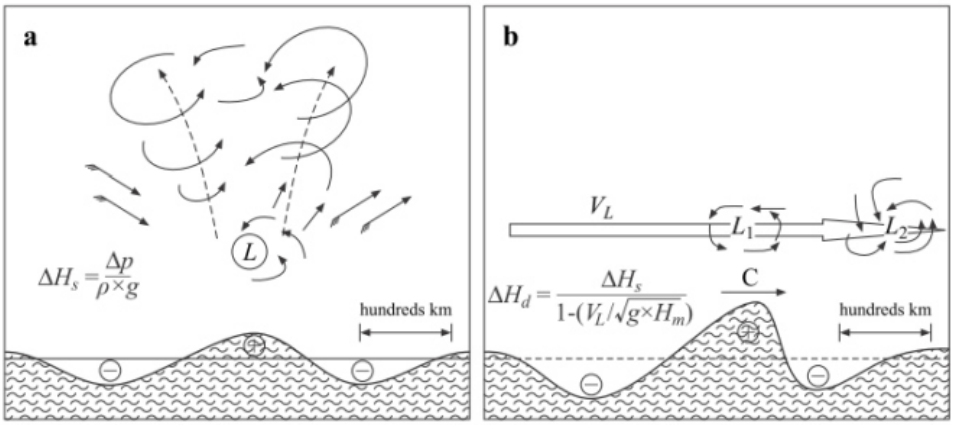


Figure 5: Sea surface deformation caused by a subpressure system a) static sea surface deformation, b) dynamic sea surface deformation with $C = \sqrt{gH_m}$ and H_m as average water depth. Taken from Holfort et al. (2014) based on Wisniewski and Wolski (2011).

.

Hence, important characteristics of the pressure-effect are the pressure in the center of the low-pressure system, its trajectory and its velocity (Wolski and Wisniewski (2021)).

According to the dynamic inverse pressure effect, the elevated water cushion moves along the trajectory of the low-pressure area as a so called *baric wave* (Wolski and Wisniewski (2021), Holfort et al. (2014)). As Wisniewski and Wolski (2011) note it is the velocity of the baric low-pressure system, that determines whether the elevation of the sea surface is mainly driven by the wind-effect or the baric-wave action. With high progressive velocities of the pressure-field, the baric wave action is predominant whereas with slow-moving low-pressure areas, the wind-effect is stronger. If both are synchronized, e.g. on-shore winds and rapid baric wave crests, storm surges are exceptionally high at the (Polish) coastline. The same was generally stated by Wolski and Wisniewski (2021) and Holfort et al. (2014); a baric wave, that moves with a similar progressive-velocity as the low-pressure system, amplifies surges, which in turn leads to serious flooding hazards at effected coasts. As low-pressure systems in the Baltic Area usually move from the (South-)West towards the (North-)East, the water surface is mostly elevated in the North and depressed in the South (Wolski and Wisniewski (2021)). This is why they see a sea level decrease at Western Baltic

tide gauging stations as an indicator for storm surges generated by low-pressure systems.
Once the storm and the low-pressure field abate, the sloped sea surface relaxes. This relaxation may in turn lead to seiche-like oscillations (until equilibrium is restored), which can again contribute to surges at opposite coastlines (Wisniewski and Wolski (2011), Wolski and Wisniewski (2021)).

2.3.3 Seiches

Seiches are standing waves that can occur in enclosed basins like the Baltic Sea (Weisse (2014), WMO (2011)). They can be generated by various natural processes but the main drivers in the Baltic Sea might be the wind-forcing and atmospheric pressure disturbances (Weisse (2014), Chapman and Giese (2001), WMO (2011)). As has been discussed in the previous section, with strong westerly winds and low-pressure systems moving towards the North East of the basin, the sea level slopes downward from North to South (Wolski and Wisniewski (2021), Weisse (2014)). Once the storm ends, the sea surface relaxes and the water masses flow back towards the southern parts of the basin until the sea surface is at rest again (Weisse and von Storch (2010), Holfort et al. (2014)). The relaxation excites resonant oscillations (seiches) until they eventually disappear due to friction (Weisse (2014)). According to Gönnert et al. (2001), the period of seiches is between 26 to 39 hours, depending on the shape of the basin (Weisse and Hünicke (2019)). An example where seiches could have played a role is given in Figure 6. It shows the starting point of an oscillation in bays of Estonia and Finland which returns within one day to the southern most coastlines of the Baltic Sea. Often the increase of water levels in the South are strengthened by winds that turned its direction from southwesterly to northeasterly after the storm (Weisse (2014)).
If the effect of seiches is interacting with already present wind-surges or is in resonance with low-pressure systems, they may contribute to extreme storm surges at Baltic coastlines (Suursaar et al. (2006a), Weisse and Weidemann (2017), Wolski and Wisniewski (2020)). For instance, the Umweltministerium Mecklenburg-Vorpommern explained a rise in water level of almost one meter along the German and Danish coastlines by the effect of seiches (Weisse and von Storch (2010)). This is in accordance with a numerical calculation done by Magaard and Rheinheimer (1974), who showed maximal oscillations of one meter within bay areas. Those calculated seiches are rarely observed in the real world due to the superposition of seiches, wind-forcing and inflowing water masses from the North Sea (Weisse (2014)).
Seiches and their contribution to storm surges are highly debated in the scientific community. A thorough discussion of corresponding literature can be found in Weisse (2014). While some scientists argue that seiches do not add any significant elevation of sea surface to storm surges (Bork and Müller-Navarra (2009)) others conclude that seiches are an important driver of storm surges within the Baltic basin (Meinke (1999), Meinke 2003 in Baerens et al. (2003) see Weisse (2014) p.22). Meinke 2003 (in Baerens et al. (2003)) proved the contribution of seiches for 50% of storm surges analyzed in her study. She notes that seiches might be sufficient to explain storm surges above the warning level of 1m (above NN) without the preconditioning of wind-surges. In contrast, Koppe (2002) state that a combination of seiches and wind-surges is primarily driving storm surges at German coastlines. Weisse (2014) calculated a positive correlation (0.6) between years with high numbers of occurrences of seiches and extreme water levels (99% percentiles). Nevertheless they attributed the influence of seiches on storm surges to only 37% of the analyzed 183 storm surges in Wismar. When considering only extreme storm surges, they could

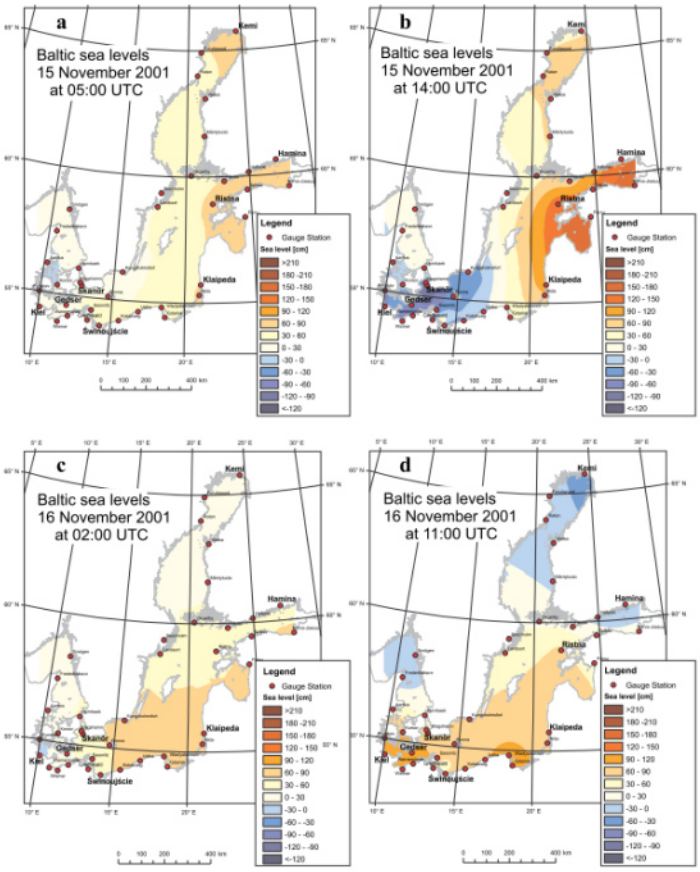


Figure 6: Sea level (in cm) at selected tide gauging stations during a storm surge event from 15.11.2001 (a – b) and the day after the storm on 16.11.2001 (c – d) at specific times. The figure indicates the behavior of seiche-like oscillations. Taken from Holfort et al. (2014).

.

not find any significant contribution of seiches. For an extreme water level of more than 149cm (above NN) only 2.55cm of sea surface elevation (on average) where due to seiches. Furthermore, the maximal contribution of seiches was measured to be time-delayed and appeared several hours after the highest water level during a storm surge. This is why the planning for coastal security in Schleswig-Holstein does not account for the effect of seiches (Hofstede (2022)).

In our model we will not directly use Seiches as a predictor but rather try to discover if seiche-like patterns occur in model results.

2.3.4 Precipitation and Surge-River Interaction

When low pressure systems and corresponding cyclones move over the Baltic Sea, they usually bring precipitation along (Harris (1963), Leppäranta and Myrberg (2009)). Extreme precipitations associated to low-pressure systems are most frequent in winter (Rutgersson et al. (2021)). Depending on the region, the rainfall per year can peak at up to 750mm (Leppäranta and Myrberg (2009)). The largest amount of precipitation is found at the eastern coast of the Baltic Sea due to the winds blowing mostly eastward in winter time (Leppäranta and Myrberg (2009)). 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)). As Harris (1963) state, precipitation can lead to above normal water levels in estuaries which alters the gradient in the river water-level and thus leads to an accumulation of rainwater in the river bed. Additionally, if the river bed is relatively flat, the storm surge can easily penetrate the riverbanks upstream for tens or even hundreds of kilometers (WMO (2011)). In the worst case, extreme surges can propagate over marshes and expand them. Winds blowing over these open waterbodies generate waves and hence transport even more water inland. The backflow of this water to the Baltic Sea is generally very slow (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. (2021), Bevacqua et al. (2019)). In their study Bevacqua et al. (2019) calculated return periods of 10 - 200 years (depending on the geographical location) for compound floodings in the Baltic Sea based on ERA5-Interim data. According to the IPCC (SMP) (2021), the frequency and intensity of heavy precipitation events increased since the 1950s due to human activity and is projected to increase further (with medium to high confidence, see also Rutgersson et al. (2021)). Despite these projections, the spatial and temporal variability of precipitations remains to be a challenge for current climate models (Rutgersson et al. (2021)). This is why we included it as a predictor in order to investigate if our model can find interesting patterns that lead to storm surges.

2.3.5 Bay Effect

As has been shown in Section 2.2.5 the bays of the Baltic Sea, especially the Gulf of Finland, Riga an Bothnia in the East as well as the Bay of Mecklenburg in the Southwest, are most prone to experience (extreme) storm surges (Holfort et al. (2014)). Theoretically, this is explained by the geomorphological characteristics and the shallow bathymetry of the bays (Wolski and Wisniewski (2020)). This *Bay-Effect* describes the increase in extreme water levels towards the interior of the bays as the bottom-topography becomes shallower and the bay itself narrower (Wolski and Wisniewski (2020), Holfort et al. (2014)). The main reason for this effect is the elongated shape of the Baltic Sea and the relatively large area of open water compared to the length of the coast and the width of the bay (Hünicke et al. (2015), Holfort et al. (2014)). Additionally, many bays are directly exposed to the predominant westerly wind direction (Hünicke et al. (2015)), making it easier for water masses to be pushed into them.

The Bay-Effect was directly observed by Holfort et al. (2014), who showed that the number of storm surges at tide gauges situated in mentioned bay-areas was higher over the period from 1960 to 2010 compared to measurements in other coastal areas. Wolski and Wisniewski (2020) further add that those extreme sea level were more intense and lasted longer farthest inland of the Bays.

2.3.6 Effect of Waves

The wind stress during a storm also generates waves by penetrating the sea surface (WMO (2011)). These waves generally move in the same direction as the wind blows (Harris (1963)). If the waves approach the coastline, where the topography gets shallower, the wave height increases (wave shoaling) and they start to break. This releases momentum due to dissipation and eventually considerable water-masses are

transported shoreward (Weisse et al. (2021), Harris (1963)). A long sequence of breaking waves can then lead to an increase in the mean water level in the coastal zone, because the water transported shoreward cannot flow back to the open sea as easily as it was pushed towards the bay (Weisse et al. (2021), Harris (1963)). This phenomenon is called a *wave-setup* and can contribute to short-term sea level extremes (WMO (2011), Harris (1963), Weisse et al. (2021)).

The effect of wave-setup can increase the water level up to one meter in the surf zone (region of breaking waves), hence wave-setups should be considered when evaluating flooding risks during storm surges (WMO (2011), Gönnert et al. (2001), Vousdoukas et al. (2017)). How strong a wave-setup is, depends on the beach topography, the wave-conditions and the bathymetry (Gönnert et al. (2001), Harris (1963)). Peak values of wave-setup are expected in coastal regions with steep slopes in bathymetry. This way, waves can travel close to the shore in deep waters and then release their energy in a sudden transfer of momentum in the surf zone as they break, a so called shore break (Harris (1963)).

2.3.7 Climate Modes

Extreme sea levels strongly depend on the wind-stress and low-pressure-fields moving over the Baltic Sea. Both are influenced by climate modes, especially the North-Atlantic-Oscillation (NAO). This oscillation is characterized by the difference in atmospheric pressure between the *Islandic Depression* and the *Azores High*. A strong Azores-High and a pronounced Islandic-Depression are then indicated via a positive NAO-Index, which leads on average to an increased circulation of the Westerlies (Weisse (2014)). Hence, the NAO can contribute to stronger westwinds, which is the predominant wind-direction for storm surges in the Baltic Sea.

Woodworth and Marcos (2018) showed, a positive correlation between extreme water levels in the Baltic Sea and the NAO-Index (see also Weisse (2014)). As Hünicke and Zorita (2006) stated earlier, this correlation variates spatially, being weaker in the southwestern Baltic Sea compared to central or northern parts. This is explained by the winds blowing westerly, hence not leading to any wind-surges in the southwest (Weisse (2014)). Furthermore, Woodworth and Marcos (2018) removed the long-term mean sea level changes in their analysis but still found a positive correlation between NAO and extreme water levels. This suggests, that the NAO may also contribute to the effect of prefilling, i.e. an increase of the Baltic Seas volume, and to locally generated wind surges (Weisse and Hünicke (2019)). Leppäranta and Myrberg (2009) add, that with a positive NAO-Index, winters are warmer and ice-sheets melt, again adding to this volume. Altogether the baseline for extreme storm surges shifts and less wind is required on time scales shorter than one month to induce extreme surges (Weisse and Hünicke (2019), Weisse et al. (2021)). This conclusion is in accordance with results of a model study, which used coupled North and Baltic Sea models forced by wind and sea level pressure only (Weisse and Weidemann (2017)). The study showed, that lower wind speeds were needed to sustain high sea level extremes when the volume of the Baltic Sea was above normal and those situations appeared more often during phases of a positive NAO-Index.

Despite these connections, Hünicke and Zorita (2006) point out strong decadal variations in the correlation and Weisse and Hünicke (2019) conclude from this that the NAO is not the the optimal atmospheric pattern describing sea level variability in the Baltic Sea. A possibly better index was suggested by

Karabil et al. (2018), the Baltic Sea and North Sea Oscillation (BANOS), which is based on the pressure differences between the Bay of Biscay and Tromsø.

2.3.8 Earth Rotation and Tidal Effects

In the general analysis of storm surges, tides do play an important role as they can elevate the sea surface up to several meters within a range of hours. Within the enclosed basin of the Baltic Sea though tides are fairly negligible and just contribute to an elevation of up to a few centimeters (Weisse and von Storch (2010), Weisse and Hünicke (2019)). For most regions in the Baltic Sea the tidal variation is only of about 3cm (Schmager et al (2008) in Weisse and Hünicke (2019)), with the exception of the western basin, where up to 30cm of variation can be observed (Leppäranta and Myrberg (2009)). For this reason, we neglect tides within this research.

The rotation of the earth leads to the *Coriolis force*, which effectively describes a deflection of tracers to the right in the Northern Hemisphere. This influences the current generated by wind-stress and may lead to a rise in water level to the right of a current hitting the coastline in order to balance the Coriolis force (Harris (1963)). Of course the Coriolis force is incorporated in dynamical storm surge models via the Navier Stokes equation and is represented indirectly by data-sets in data-driven models (see Section 3 and Equations 13 – 15). Nevertheless it finds only very little resonance in scientific literature on storm surges (in the Baltic Sea).

2.3.9 Prefilling

The Baltic Sea contains an averaged volume of 21.205 km^3 (Weisse (2014)) that is constantly altered due to different in- and outflows (Weisse and Hünicke (2019)), shown in Figure 7. According to Leppäranta and Myrberg (2009), annually 215km^3 are added per precipitation while 175km^3/a evaporate. The freshwater influx through rivers is 440km^3/a. This corresponds to a net inflow of 480km^3 per year which is balanced by the saltwater exchanged with the North Sea. The inflow of saltwater into the Baltic Sea via the Straits of Denmark is approximately 1180km^3/a and the outflow 1660km^3/a (Leppäranta and Myrberg (2009)). On a daily basis up to 45km^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 12cm/d (Mohrholz (2018)) or 320cm averaged over a whole year (Leppäranta and Myrberg (2009)).

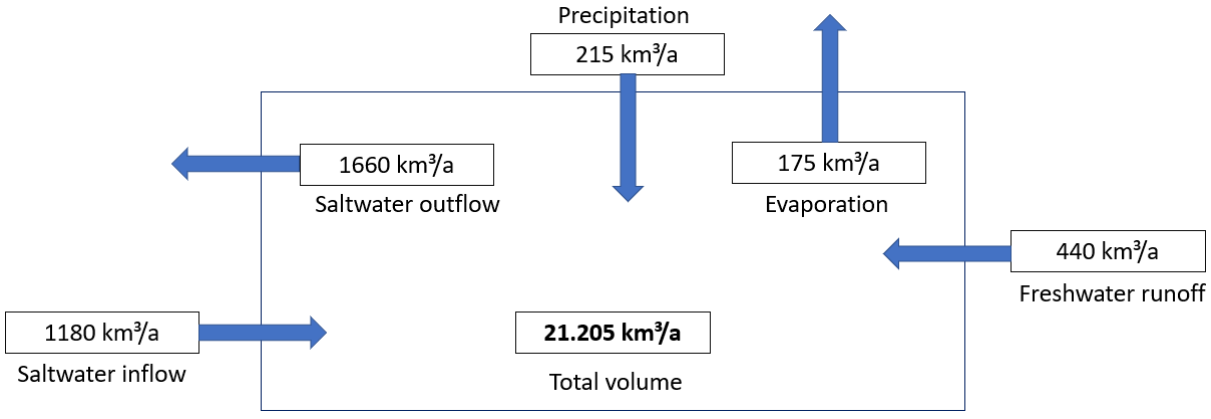


Figure 7: Annual water budget of the Baltic Sea based on Leppäranta and Myrberg (2009)

.

If the water level of the Baltic Sea is elevated 15cm above the mean sea level for more than 20 consecutive days due to increased inflow via the Straits of Denmark, Mudersbach and Jensen (2010) speak of a *prefilling* of the Baltic Sea (see also Kowalewska-Kalkowska 2012 in Wolski and Wisniewski (2021)). In other contexts, this situation is also termed *preconditioning*. The degree of filling is then given by the averaged water level of the Baltic Sea (Weisse (2014)). Usually the tide gauging stations in Landsort (Sweden) or Degerby (Finland) are used as proxies for measuring prefilling of the Baltic Sea (Weisse (2014), Janssen et al. (2001)).

Depending on the degree of prefilling storm surges can become more likely. For instance, Weisse (2014) state that less wind is needed to induce storm surges during times of strong preconditioning compared to situations without prefilling. Hence, it is important to understand the factors contributing to preconditioning.

As mentioned, wind blows mainly (south-) westerly in winter times due to climate modes and associated pressure gradients. It is mainly this 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), Weisse (2014))). Mietus et al (2004, in Wolski and Wisniewski (2021)) showed that the inflow is becoming even stronger as the Western air mass inflow is becoming more stable and intense in future projections. As the inflow is predominantly depending on the strength of the Westerlies, which themselves are subject to the NAO-Index, years of a strong positive NAO-Index tend to be associated with higher numbers of prefilling in the Baltic Sea (Weisse (2014)). On shorter time-scales, the water exchange is mainly driven by atmospheric conditions (Leppäranta and Myrberg (2009)). For example, a sequence of fast-moving low pressure systems coming from the West and travelling to the North-East of the Baltic Sea resulted 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, with a monthly inflow of 120km^3. Minima occur in May with only 70km^3 of inflow. Combined with the effects of stronger winds and rainfall in winter, the preconditioning is an important driver of storm surges.

In their study Weisse (2014) analyzed timeseries from 1948 to 2011 of waterlevels in Wismar and connected them to situations of prefilling. Preconditioning was calculated by taking the 20 day rolling mean of hourly water levels in Landsort and datapoints showing values 15cm above the mean were considered timepoints of prefilling. They found, that not only less wind is needed to induce storm surge during preconditioning but also conclude that the direction of the wind becomes less important (at least in Wismar). Furthermore they showed, that 54% of the storm surges in Wismar during the study-period happened during states of preconditioning. A positive correlation (0.55) was deduced between prefilling and (hourly) extreme water levels (above 99%-percentiles). No effect of prefilling on the prolonged duration of a storm surge could be established but in times of preconditioning, storm surges become more frequent (within the German Bight). Weisse and Weidemann (2017) also state that a higher Baltic Sea Volume leads to a higher amplitude in extreme water levels under similar wind conditions when compared to a lower prefilling. For instance, during the storm Gudrun in 2005 the background sea level of the Baltic Sea was of about 70cm, hence contributing to the extremeness of the surge (Suursaar et al. (2006b)). These extreme degree of prefilling built up during the preceding month of the storm, where strong cyclonic activity pushed additional water masses through the Danish Straits to the Baltic Sea (Hünicke et al. (2015)).

The outflow occurs less frequent than the inflow (Wolski and Wisniewski (2021)) but is favored by easterly winds (Leppäranta and Myrberg (2009)). The main constituent for outflow though is the sea-level difference between the Baltic Sea and the North Sea (Leppäranta and Myrberg (2009)). This difference strongly depends on the salinity profile of the Baltic Sea, which has a strong gradient from the north-east to the south-west. This is due to the inflow of more saline and dense waters from the North Sea (Leppäranta and Myrberg (2009)). This profile (on average) leads to a sloping sea surface through density differences and explains the natural variations of water level accross the Baltic Sea (Ekman and Mäkinen (1996) in Weisse et al. (2021), Leppäranta and Myrberg (2009)). Besides the large (annual) natural variability of the inflow, which is not only dictated by the sea-level difference but also the air pressure and wind distributions (Leppäranta and Myrberg (2009)), there is no significant long-term trend of prefilling (Weisse (2014)).
This section shows the importance of preconditioning when analyzing extreme storm surges and a proxy should be implemented in respective models.

2.3.10 Ice-Sheets

The wind-stress can only transfer energy from the atmosphere to the sea when the latter is ice-free. If the surface is ice-covered, kinetic-energy from the wind-stress dissipates and the ice acts as a barrier to meteorological forcing (WMO (2011), Weisse et al. (2021)). One should distinguish the influence of the ice-cover on negative and positive storm surges. While ice cover can reduce the amplitude of the latter, no contribution to the amplitude of the former was yet observed (WMO (2011)). In the Baltic Sea the maximum ice extent is largely driven by the NAO and its inter-annual variablity correlates with NAO-Index (Omstedt and Chen (2001), Vihma and Haapala (2009)). As Leppäranta and Myrberg (2009) mentioned, a positive NAO-Index is related to warm winters, and as a consequence less ice-coverage. They also note that the annual ice-extent does not show any temporal structure and can thus be compared to white noise with large variance. In averaged climate projections the Baltic Sea ice experiences shrinking and thinning (BACCI (2008) , BACCII (2015) in Rutgersson et al. (2021)), which gives the wind-stress more open water to act on. Altogether this may increase the number of experienced storm surges in the future. Despite this, WMO (2011) notes that the influence of sea-ice on sea-level oscillations was not taken into account in many storm surge models.

3 Modelling Storm Surges

In the realm of forecasting storm surges two general types of models can be distinguished, dynamical models and statistical models. While dynamical models are based on governing equations of underlying physical processes of fluid dynamics, statistical models use large data-sets in order to explain the onset of storm surges over reoccurring patterns within the data.
We will give a brief overview of dynamical and statistical models in the oncoming sections, based on the more comprehensive description of WMO (2011), where further references can be found.

3.1 General Remarks on Storm Surge Forecasting Models

Most data-sets used in forecasting models are based on meteorological and hydrological measurements taken from tide gauging stations, which is true especially in the Baltic Sea (WMO (2011)). Surface winds (10m above sea level), data on sea surface stress and mean pressure at sea level are mainly used as meteorological input, while river runoff and total precipitation is considered as hydrological input (WMO (2011)). From the meteorological datasets, the surface winds and the mean pressure at sea level are mainly used in two dimensional storm surge forecasting. In general, datasets used in forecasting models should be quality controlled. This is especially the case for meteorological input, because the magnitude of surges is at least proportional related to the square of the wind-velocity (see Equation 2) (WMO (2011)). Hence, small errors in observations or in the output of atmospherical models providing the meteorological data, may lead to incorrect predictions of total water level or surge heights. These predictands are usually used in surge forecasting models, which provide either field-outputs fixed in time or time-series over a tide-gauging station (WMO (2011)).

WMO (2011) report that approximately 75% of operational or pre-operational storm surge predictions are two-dimensional dynamical numerical models. A review of those models is also found in Gönnert et al. (2001). Hence, we start with explaining briefly the underlying principles of these models.

3.2 Dynamical Storm Surge Models

Most numerical models for storm surges are based on the linearized version of the vertically averaged and integrated Navier-Stokes equations in a right handed Cartesian coordinate system (see Appendix Equations 13 - 15 and Gönnert et al. (2001), von Storch (2014)). These equations can be solved numerically on a grid and describe the transport of the sea water due to forcing functions like the atmospheric pressure gradients and the wind-stress (Gönnert et al. (2001)), taken from atmospherical models or observations. The wind-stress and bottom-friction act as a source and sink for momentum respectively and are usually parameterized, due to the lack of model-resolution (von Storch (2014)). For instance, one could approximate the surface wind stress as a simple function of wind speed similar to Equation 2. Note though that this is a caveat of those models, generally resulting in a lack of accuracy. The solution of the numerical models strongly depends on the initial state and the implemented boundary conditions. Both of them are set by data from operational weather forecasts of the current hydrodynamic and meteorological state (von Storch (2014)). In order to ensure accurate initial states of the dynamical model run, a *hindcast* is done prior the actual forecast (WMO (2011)). With now accurate initial conditions, the solution provided gives insight into the surge amplitude or the current water level of the basin. The actual height of a storm surge is mostly calculated by running two simulations, one with meteorological forcing and one without it. Subtracting those simulations results in the amplitude of the storm surge (WMO (2011)). These outputs are diverse in a sense of format. Some models result in a time-series of surge level for specific tide-gauging stations with a forecasting period of usually 36 to 72 hours (WMO (2011)). Note here, that the accuracy of these time-series forecasts shrinks drastically if the surge event develops rapidly and lasts less than 24 hours. Situations where the development of a surge take 48 or more hours are better reproduced (WMO (2011)). Other models provide spatial maps fixed in time, local peak and maxima charts or an overview of flooded coastal areas (WMO (2011)).

The accuracy and complexity of models is closely linked to the underlying structure of the grid. The first models used rectangular or squared grids in order to resolve the area of interest. Rectangles though usually do not describe the mostly curved shapes of the coastlines sufficiently. Hence, the usage of those grids resulted in artificial subbasins with own seiche-like oscillations which contaminated the computation of storm surges in the main basin (WMO (2011)). This is why most current models use irregular triangular grids, which better represent the coastal geometry and bathymetry of shallow waters (WMO (2011)). In regions of no expected surges, the triangular grid can be coarse and in coastal regions resolution can be higher due to smaller triangles within the grid. For regional models, this resolution ranges from 10 to 20km but a coastal resolution of up to 1km or finer can be achieved when using nested grids (WMO (2011)). This is why nowadays approximately 75% percent of operational models are using this approach based on irregular triangles and *finite elements* methods (WMO (2011)).

In the Baltic Sea there are a couple of regional models in operation. For instance the BSHcmod from the Bundesamt für Schifffahrt und Hydrographie (BSH) is a hydrostatic circulation model with a grid-resolution of about 5km in the Western Baltic. In Denmark a two dimensional hydrodynamic model is used with resolutions ranging from 16km to even 600m (for specific regions) depending on the nesting (WMO (2011)). How specific dynamical models can get is explained on behalf of a numerical model of joint water-ice dynamics for the Gulf of Finland (WMO (2011)). This is a nested model, specifically designed for representing the characteristics of the complicated, enclosed morphology of the Gulf. Three different grids are used, one resolving the Baltic Sea with 30km, the other adjusted for the Gulf of Finland with a grid-size of 5km and the last one for the eastern part of the Gulf with a resolution of 1km (WMO (2011)). Due to this nesting, the specific weather systems can be better described but also the model structure gets more complex.

In general nesting leads to faster computations but compared to statistical models, computation time is still long. But this is not the only drawback of dynamical models. For instance Muis et al. (2016) showed, that dynamical models underestimate extreme storm surges. They argue, that the resolution of the atmospherical models is not high enough to describe the onset of storms, and hence lead to a less accurate forcing of storm surge models. In WMO (2011) it is also stated that due to resolution problems in coastal areas wind-speeds are underestimated, which may in turn lead to a weaker modelled storm surge. Rutgersson et al. (2021) recently backed this and argue that the intensity of cyclones is generally underestimated as well. The concept of *downscaling* is proposed by von Storch and Woth (2008) in order to overcome these challenges without increasing the resolution of atmospherical models. Nevertheless WMO (2011) state, that the effect of mesoscale (10-100km) weather systems is not represented in current storm surge models as there are no networks providing data at these scales and hence they are not reflected in operational coastal flooding forecasts. At the same time, flooding events driven by mesoscale atmospheric events, like storms, happen frequently (WMO (2011)). In addition to the resolution problem of atmospherical models and storm surge models, the coupling between those models is highly complex. In order to transfer the energy imposed by the atmosphere on the Baltic Sea basin and force the surge model, both have to be linked at the boundary layer through a moving and changing interface (Gönnert et al. (2001)). Usually the data of meteorological fields is interpolated to the surge model grid (von Storch (2014)), which may lead to smoothing of the data. Once again, this smoothing may lead to an underrepresentation of extremes within the model. Smoothing was also used as an explanation by Muis

et al. (2016) for underestimating extreme storm surges, when ERA5-interim data of wind and pressure fields were flattened due to the temporal and spatial resolution. Nevertheless it should be stated that dynamical storm surge models are mostly accurate when forecasting the water level. It is mainly the extreme events that are often underestimated. According to Muis et al. (2016) this may also be due to the fact, that dynamical models do not incorporate the non-linear interactions between the drivers of storm surges.

It is thus of interest in the scientific community to model storm surges and their extremes via statistical models, where especially the methods of Machine Learning seem to be a good tool for storm surge forecasting.

3.3 Statistical Storm Surge Models

Harris (1962) was among the first to investigate statistical methods for storm surge predictions. In contrast to dynamical models, the statistical approach can not describe the underlying physical processes involved in a storm surge. The advantage of statistical models is though that they make most efficient use of the input data and forecasts are produced with less computational expense (Harris (1962)). Additionally, statistical models are not subject to the same initial and boundary conditions as numerical models and instead may use implicit information given in the dataset, like correlations not clearly recognized in physical processes, which may improve the accuracy of the prediction (Harris (1962)). In order to capture those hidden patterns, large datasets are needed.

The Baltic Sea is a unique testbed in terms of long-term data. A dense observational network is covering the Baltic coastlines and the basin providing observational data since the 1950s with hourly measuring frequencies (Rutgersson et al. (2021)). Hence, sufficient data is given to establish statistical models for the Baltic Sea and also validate them on independent test data. The latter is very important due to the tendency of statistical models to (over)fit the input data, i.e. representing only patterns within the given data. Thus testing the model on data that was not incorporated into the building process of the model is mandatory (Harris (1962)) and is common practice when evaluating models based on machine learning methods. WMO (2011) and von Storch (2014) also note that most statistical models can therefore only be applied to the region they were developed for and a transition to other regions might be complicated or results in less accurate forecasts. Similarly, if the underlying geophysical conditions of storm surges within the region change considerably in the future, the model has to be adjusted as it is only based on data sets of past events (von Storch (2014)).

Further requirements needed when assessing storm surge statistics are given by von Storch (2014), namely,

1. **Homogeneous data:** A change in statistics has to be independent of changes in local environments, measurement technology and reporting practice of the data set.

2. **Length of time-series:** The dataset must cover long time-periods, for instance more than sixty years, such that robust statistics can be calculated based on two separate 30 year segments. Having an insufficient amount of data may lead to underspecification, i.e. a lack of details involved and the forecast may not represent important specifics of the storm surge.

3. **Sound Statistics:** When building empirical models, one must ensure that the linkage between an-

alyzed datasets is physically sound and statistically correct. For instance, in some *semi-empirical* models changes in temperature were related to changes in sea level. This was later shown to be statistically unsound, because it only compared two trends.

Once these requirements are met, there are many approaches that can be taken in order to derive data-driven, statistical models. A brief, and by no means complete, overview of data-driven models is now sketched.

Siek and Solomatine (2010) showed that storm surges can be seen as deterministic chaos with certain limits to its predictability. They explain this chaotic behavior due to nonlinear coupling of storm surge drivers within a dynamical system. In their study, they investigated storm surges along the Dutch coast in the North Sea via chaotic models based on chaos theory and compare it to an Artificial Neural Network (ANN), a deep learning method of machine learning. Both methods were forecasting time periods of 1 to 12 hours. The ANN is easier to implement and shows similar results as the chaotic model for normal storm surges but extreme storm surges were better represented by the chaotic model. While this is an interesting approach, we could not find a similar setting for the Baltic Sea.

Gönnert and Sossidi (2011) also used empirical methods to investigate extreme storm surges in the North Sea but at a single tide gauging station in Cuxhaven. Their method respects the hydrodynamics of storm surges and especially analyses the components of tides, wind surges and external surges as well as their non-linear interactions. Extreme storm surge values are based on daily maximum values of wind surge, tides and external surges. They conclude that respecting non-linear interactions in the model leads to a lower water level compared to superimposing the effects linearly. Another approach in Cuxhaven was done by Dangendorf et al. (2014), who used wind-surge formulations and implemented the relationship between surge, wind and sea level pressure in the model, which resulted in correlations to observations of 0.91 and a Root Mean Squared Error (RMSE) of 13.9cm. Wahl and Chambers (2016) specified simple and multiple linear regression models to describe the relation between large-scale climate variability along the United States coastline and multidecadal extreme sea levels. Besides statistical models, applications of machine learning methods also become more popular due to their efficient capability of linking predictor and predictand data (Tadesse et al. (2020)). Similar to Siek and Solomatine (2010), Bezuglov et al. (2016) also used ANNs to predict storm surges along the coastline of North Carolina.

A coupled approach of machine learning and numerical models was used by French et al. (2017), who combined ANNs with a two dimensional hydrodynamical model to predict the flood extent at the Port of Immingham, United Kingdom. This approach showed higher accuracy compared to only using the national numerical tide-surge model. This already indicates, that machine learning techniques can provide additional value to numerical models.

Tadesse et al. (2020) come to a similar conclusion in their study, highlighting data-driven models as "a powerful and computationally cheap complementary way to simulate storm surges in addition to process-based but computationally expensive numerical models". They compared statistical and machine learning models 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 within a time period from 1998 to 2014. All predictors (wind-field, sea level pressure etc.) were filtered through a

Principal Component Analysis (PCA) and only the principal components explaining 90% of the variance were used, which reduces model complexity drastically. To ensure the time-delayed effect of predictors, they were lagged 30h prior to the time, where the daily maximum surge occured. The storm surge predictand was derived from two data-sets, the observed hourly sea level data from the GESLA-2 database and in situ data of daily maximum surges taken from the Global Tide and Surge Reanalysis (GTSR) as mentioned in Muis et al. (2016). They compared linear regression models to a machine learning method called Random Forest (RF)s (see Section 4.5). RFs are fast and easy to use and can capture non-linear dependencies between predictor variables and the predictand. Usually, these models are not prone to *overfitting*, i.e. not heavily specializing on the data they are based on, and at the same time are suitable for high dimensional data (Tyralis et al. (2019)). The authors find that data driven models work better in extratropical regions, e.g. the Baltic Sea, compared to the tropics. Furthermore, models trained with predictors based on remotely sensed data outperformed models forced with predictors obtained from reanalysis data. More specifically, when the linear regression as well as the RF were based on remotely sensed data and when they were time-lagged, they showed higher accuracy compared to models without a time lag. Both methods also reveal mean sea level pressure as the most important predictor to model daily maximum surge, at least for 70% of the analyzed tide-gauges. They also analyzed extreme storm surge and models based on reanalysis data showed an average correlation of 0.51 with the predictand in extratropical regions.

A direct comparison of a simple linear regression to an ANN was done by Bruneau et al. (2020), again on a global scale. They fitted an ensemble of ANNs to over 600 tide gauges around the world and predicted general and extreme non-tidal residuals of the sea water level. These ensembles generate a probabilistic forecast at each tide gauge. Sea level anomalies were calculated based on the GESLA-2 database and predictors (10m wind, mean sea level pressure and significant wave height) were taken from ERA5-reanalysis. From the atmospheric predictors proceeding 3h gradients were calculated in order to represent late intensification or de-intensification of for instance rapidly developing low-pressure systems. It was shown that ANNs generally outperformed multivariate linear regressions (in terms of Continous Ranked Probability Score (CRPS) and correlations) on a global scale. This is the case for the general variability of the water level as well as for extreme events. While ANNs could reconstruct extreme surges with significant skill, they still struggle with strongest extreme events. Bruneau et al. (2020) explain this by the choice of training data, where extreme events are only a fraction of the data-set. Because ANNs are trained with a procedure that is ill-designed for outliers and biased for average dynamics, extreme surges and can not be reproduced reliably. This finding gives us reason to adjust the training set in our approach accordingly, s.t. it weights extreme events more heavily. A further interesting finding of Bruneau et al. (2020) is, that 6-7 years of training data are sufficient to get reasonable results, at least with ANNs. They further state, that multivariate linear regressions may capture the main effects of the atmospherical drivers on sea level, the non-linear interactions and atmospheric-driven intensification could only be deduced by ANNs. This again shows, how powerful simple machine learning models can be when predicting (extreme) storm surges. They further tried to use a more complex method, a so called Long Short-Term Memory (LSTM) network, which, at least in their study, did not lead to any improvements.

The LSTM is a machine learning method, that is able to selectively store long-term information about cor-

relations in a timeseries and it predicts the next timestep by using informations from previous timesteps. Tiggeloven et al. (2021) used LSTMs and came to a different conclusion than Bruneau et al. (2020), namely a good skill of LSTMs compared to other neural networks. Overall they used different deep learning methods, a further branch of the machine learning realm, including ANNs, Convolutional Neural Network (CNN)s, LSTMs and Convolutional Long Short-Term Memory (ConvLSTM) to predict surges at 736 tide stations globally. As predictand they used surges extracted from the GESLA-2 dataset. Predictors were obtained from the highly resolved European Re-Analysis (ERA5) and hence the time-series ranges from 1979 to 2019. Within this period, they selected all stations containing at least seven years of continuous data. They increased the complexity of the models by subsequently adding predictors like the Mean Sea Level Pressure (MSLP), the hourly gradient of MSLP, meridional and zonal 10m wind components and the wind speed magnitude as well as other variants of those. Unfortunately the forecasting period is not clearly stated in the paper. The results were compared to the GTSR surge time-series obtained from a hydrodynamic model forced with ERA5 reanalysis data.

When using the same hyperparameters across different methods, the LSTM outperformed other types of neural networks. If the model-architecture is fine tuned, i.e. the spatial extent of the predictor values is extended and the number of hidden layers is increased, CNN outperformed LSTMs but at a huge cost of computation time (increasing almost 15-fold). This result holds in general, where a larger spatial footprint for predictor selection increases model accuracy but in contrast increasing the number of hidden layers does not necessarily lead to better forecasts. Depending on the local tide gauge, model characteristics like the number of hidden layers need to be adjusted in order to optimize forecasts. Furthermore they showed that the best performance in general is achieved when using the 5 predictors already stated or the quadratic wind components in addition.

The overall result shows, that machine learning approaches capture the temporal evolution of surges and outperform the large-scale hydrodynamic model. Only extreme events are underestimated due to similar reason as in Bruneau et al. (2020).

As this review indicates, most approaches using ML-methods are global and hence lack in specification for the Baltic Sea basin in regards of hyperparameters and predictors. The only study (to our knowledge) that applied ANNs specifically to the Polish coast of the Baltic Sea was done 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 could be successfully integrated in operational forecast services and possibly reduces their average error. This result was achieved though by using a numerical formulation of the winds impact on coastal areas within the ANN. Similar to the global studies, this study shows an underestimation of extreme water levels.

In conclusion, all of the discussed models did not capture extreme storm surges well due to the choice of training data. Hence, these studies motivated us to develop ML models that are tailored to the characteristics of the Baltic Sea in order to predict extreme storm surges occurring at Baltic coastlines.

4 Our Model

We aim to predict extreme storm surge events (*predictand*) around the Baltic Sea by using atmospheric *predictors* like surface pressure, total precipitation and wind-fields as well as one hydrological predictor, the prefilling. As a machine learning method we will use a *Random Forest (RF)*. Those forests provide binary classifications of the investigated data-set. In the following we will give insight to the research area and how the model was developed to give the reader the possibility of replication.

4.1 Area of Research

The area of research ranges from 5°W to 30°E and 40°N to 70°N, which includes the Baltic Sea (BS) (see Figure 8). More specifically seven stations were selected for model analysis. These stations are part of the Global Extreme Sea Level Analysis (GESLA) data-set. Station codes are provided in Table 1. We choose those stations in order to cover all coastal orientations and bays of the BS.

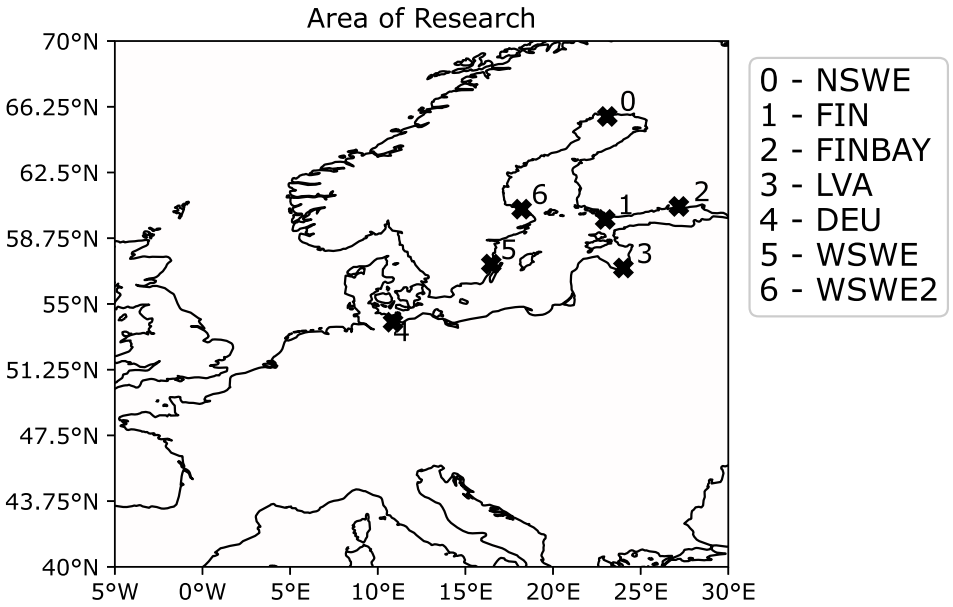


Figure 8: Map of whole resarch area. Crosses and numbers indicate the analysed stations within the Baltic Sea.

.

4.2 Intended Use

The overall structure of the software is sketched in Figure 9. After preprocessing the datasets, they are separated into training and test data used to fit and validate the model respectively. We fit the model with similar combinations of predictors for each station and interpreted a rise of waterlevel above the 95th percentile of the measured waterlevel at a given station as an extreme storm surge (see Section 4.4). The model then processes the atmospheric predictors in order to provide a binary prediction about extreme storm surges.

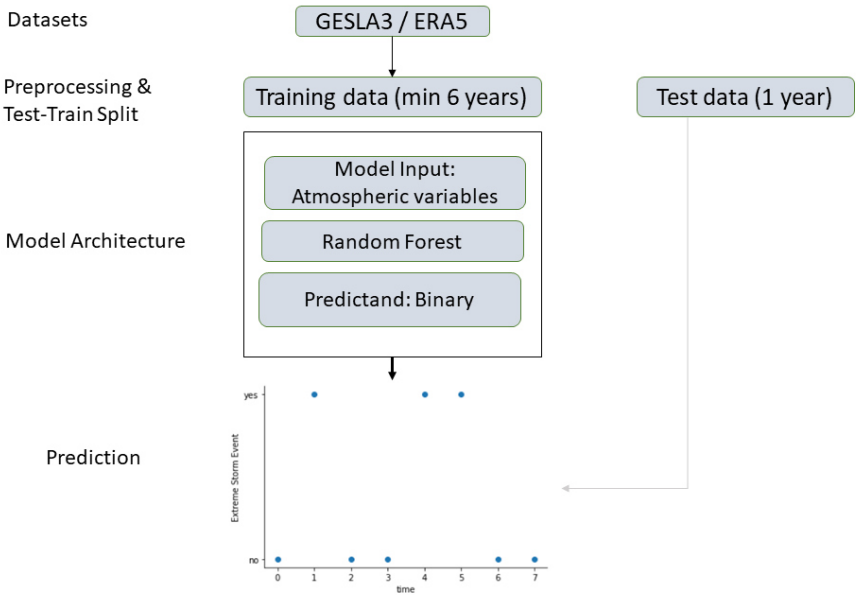


Figure 9: Software architecture as a blueprint based on: Tadesse et al. (2020)

.

4.3 Data Sources and Model Input

Machine Learning methods usually rely on big data sets, including so called *features* or *predictors*, from which they deduce underlying patterns in order to conclude with predictions. The BS provides one of the most dense 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. Together with the vast European Re-Analysis (ERA5) dataset from Hersbach et al. (2018) it makes the sea a perfect test-bed for machine learning.

The project of Global Extreme Sea Level Analysis (GESLA) provides a global set of high frequency (at least hourly) sea level data with integrated quality control flags. Several updates were applied to the dataset, the most recent one in 2021 (Haigh et al. (2021)). Height units of all stations were converted to metres and the time zone was adjusted to Coordinated Universal Time (UTC). The frequency of each station remained untouched and was at least hourly. 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 and can be downloaded on their website (`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 (after preprocessing, see Section 4.4) as a predictand for extreme storm surges at respective stations.

As mentioned before extreme surges are induced by low pressure systems, precipitation, wind-fields and prefilling of the BS. Except prefilling, we incorporated each of those drivers as predictors into our model setup based on atmospherical European Re-Analysis (ERA5) data provided by the European Centre for Medium-Range Weather Forecasts (ECMWF). The reanalysis combines model and observational data leading to a dataset with a high temporal and spatial resolution.

The ERA5 ranges from 1959 to present with hourly estimates of atmospheric variables and is spatially

resolved on a 30km (aprroximately 0.27 degrees) grid covering the Earth (see Guillory (2017)). The dataset used for this study only covers the period from 1999 to 2022.
All variables of ERA5 used as predictors are shown and briefly described in Table 2. 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 lon-lat-field described in Section 4.1 and depicted in Figure 8.
Additionally, we implemented a predictor of prefilling by using the GESLA timeseries of sea-level data at the station of Degerby as a proxy (see Weisse (2014), Janssen et al. (2001)) further explained in the following Section.

4.4 Preprocessing of Data

The machine learning algorithm expects a certain format of input data, i.e. for the predictand and the predictors. Usually, the predictand Y is given as a vector containing n_{samples} entries called *samples*. The predictor connects to every sample n_{features} *features*, yielding a $n_{\text{samples}} \times n_{\text{features}}$ matrix. In this study values of the predictors over the whole research area in form of a longitudinal-latidunial map are used as features. For example, when usings *Pythons numpy ndarrays* the algorithm expects the following formatting

$$Y = (n_{\text{samples}},) \tag{5}$$

$$X = (n_{\text{samples}},\ n_{\text{features}}) \tag{6}$$

Ultimately we need to provide the format above before passing it to the machine learning algorithm.

We deduce the predictand from the GESLA dataset, where we only selected seasonal data of winter months December, January and February. We did so, because the strongest increase in water-levels is expected from September to February and we decided to split this into the seasons of autumn (September to November) and winter. Note, that the software also has the option to select autumn as a season but this is not analyzed in the results section due to lack of time.
From this seasonal selection we only incorporated data labeled by the GESLA team as "analysis data". This data is either not quality controlled (control flag 0), has correct values (control flag 1) or is an interpolated value (control flag 2). Let $\tilde{Y} = \left(\tilde{Y}_t : t \in T_{\tilde{Y}}\right)$ denote this time-series of sea-level data at any station with $T_{\tilde{Y}}$ being the set of all record-times of a particular station.
We then need to detrend $\tilde{Y}$ in order to obtain a *stationary process*, which is later passed to the model. For detrending we used the implemented linear least-squares regression $\mathcal{L}$ to the data of the *scipy*-library. This method essentially subtracts a linear regression fitted to the data from the data itself, removing trends in sea-level records leading to

$$Y = \left(\tilde{Y}_t - \mathcal{L}(\tilde{Y}_t)\right), \text{ for all } t \in T_{\tilde{Y}}. \tag{7}$$

Next, so called *one-hot-encoding* is applied to Y, i.e. data-points with values above the 95th-percentile are converted to 1 (i.e. extreme storm surge), all others are set to 0 (i.e. no extreme storm surge). This

implies our definition of extreme storm surges being the top 5% highest hourly recorded water-levels at one particular station.

Finally we converted the hourly recording frequency of Y to daily time-frequency. Therefore, we took the maximum value of Y per day, i.e. if there is only one incidence of an extreme storm surge during the day, the whole day is marked as an incidence of extreme storm surge. (Note, that this matching of time-frequency is done when intersecting recording time-periods of predictor and predictand within the software).

When downloading the ERA5 datasets for all predictors the time-periods were initially separated into the intervals 1999 to 2008, 2009 to 2018 and 2019 to 2022. The last period was not used in this study, since it does not intersect with the recording times of the predictand at some stations (which ends in 2019).

According to Bruneau et al. (2020) six years of daily input data is sufficient for some machine learning algorithms to produce reliable predictions. Hence, for each predictor the years 1999 to 2008 were selected to train the model. Only for stations 3 and 4 the period from 2009 to 2018 was chosen, because the records originate from 2005.

We used *Climate Data Operators (CDO)* to select the lon-lat-map mentioned in Section 4.1 and to further calculate daily averages of the hourly recorded data. Finally, the units of the predictor SP are converted from Pa to hPa.

The same procedure was applied to the data of period 2009 to 2018 which is later used as completely unknown data – in a sense that it was never used in the training process of the model – to validate the model predictions. We refer to this data as *validation set*.

The only predictor that is not based on the ERA5 dataset and hence was treated differently during preprocessing is the prefilling (PF) of the Baltic Sea. This predictor is loaded from the GESLA dataset at the station of Degerby (GESLA-code: "degerby-deg-fin-cmems"). When PF is combined with ERA5 predictors, the time-frequency of ERA5 is used. Hence, we reduced the hourly data of the station to daily data by using the maximum recorded water level of that day as an entry. If PF is used as a single predictor, the time-frequency is kept hourly.

We represent the predictor dataset now by

$$X = (X_t : t \in T_X)\,, \tag{8}$$

where T_X contains all date entries of the timeseries.

Both timeseries X, Y of the predictor and predictand respectively are intersected by date, yielding $T_S = T_X \cap T_Y = \{t_1, t_2, \dots, t_n\}$. In some cases, we introduced a timelag $\hat{t} \in \mathbf{N}$ to the predictor data by shifting T_S as follows

$$\hat{T}_Y = \{t_{\hat{t}+1}, t_{\hat{t}+2}, \dots, t_n\} \tag{9}$$

$$\hat{T}_X = \{t_1, t_2, \dots, t_{\hat{t}+1}, \dots, t_{n-\hat{t}}\}, \tag{10}$$

where $\hat{T}_Y, \hat{T}_X$ are the timelagged dates of timeseries X, Y.

Note, that X contains the information of a lon-lat map per date entry (e.g. a *numpy-ndarray*-shape of $(n_{\text{samples}},\ n_{\text{lon}},\ n_{\text{lat}})$), while Y is only evaluated at each date (e.g. a *numpy-ndarray*-shape of (n_{samples})). Hence, we need to reduce the dimension of X by flatten the longitudinal-latitudinal structure, such that the shape of X becomes $(n_{\text{samples}},\ n_{\text{features}})$, with $n_{\text{features}} = n_{\text{lon}} \cdot n_{\text{lat}}$. At last we set non-existing values, i.e. Not-a-Number (NaN) entries, to -999. This ensures that the machine learning algorithm will treat the entries as outliers and ultimately neglects them. Now the data is ready to be passed to the model.

4.5 The Model - Random Forests

Originally the idea was to use multiple classifiers and regressors like ANNs, Support Vector Machine (SVM) and Random Regression Forest (RRF). Due to lack of time and complexity though, we only focused on more simple Random Forest (RF)s and binary classification.

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 depicted in Figure 10. A RF consists of an ensemble of Decision Tree (DT)s. Each of those DTs processes a random sample of the test-data in order to conclude with a prediction. Those predictions are then averaged in order to get the overall prediction of the RF. For classification purposes this averaging is done via a *soft voting* strategy, which attaches a probability to each prediction of individual DTs. These probabilities are averaged and the class with the overall highest probability is chosen as a prediction of the RF. Hence, the generation of a prediction within a RF depends on how DTs form predictions.

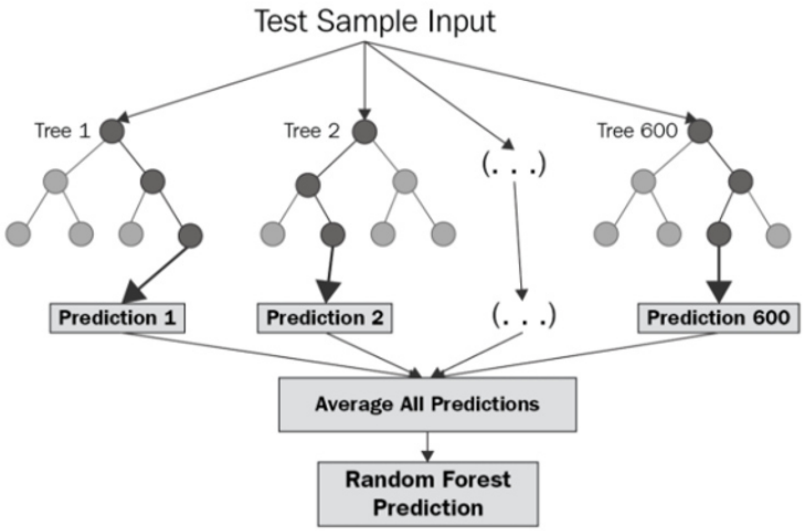


Figure 10: Model architecture of a Random Forest (RF) taken from `https://levelup.gitconnected.com`

.

DTs rely on a hierarchy of if/else-questions in order to conclude with a prediction. A simplified example is shown in Figure 11. 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 lon-lat combination within the research area). The rectangular node is called *terminal node* or *leaf* and provides the prediction. When fitting data to a decision tree, the algorithm essentially finds the best sequence of if/else- questions to get to the true answer. A prediction on unknown predictor data is then made by sifting through the optimized DT, answering all if/else-questions.

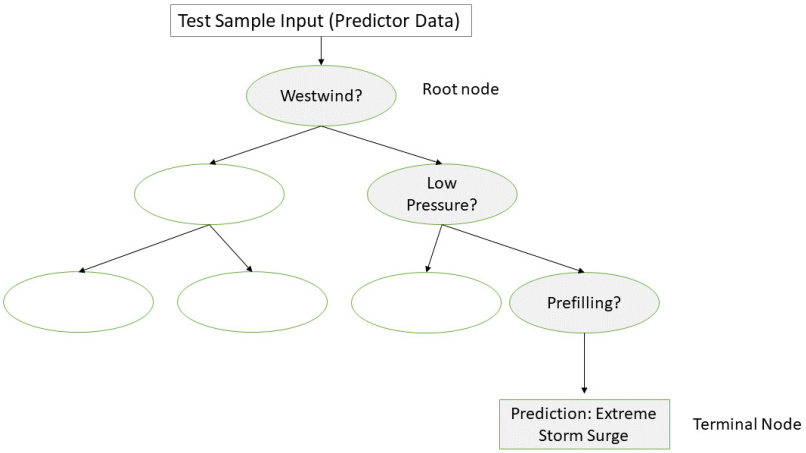


Figure 11: Simplified model architecture of a Decision Tree (DT). Grey Nodes indicate the path of test-data while sifting through the DT. Right pointing arrows refer to positive answers.

Especially for smaller datasets with a small number of features DTs are a good tool to visualize the decision process of the algorithm. With bigger datasets, a visualisation of the tree itself mostly contains too many branches and nodes in order to be interpreted easily. One can still get valuable insight on which predictors the DT relied the most via the concept of *feature importance*. This concept assigns a value between 0 and 1 to each feature, with higher numbers indicating a greater importance. The sum of all feature importances within a DT is 1. We will use this concept later in order to investigate which regions within the research area are of importance for individual predictors (see Section 4.8).

The problem of DTs is that they are prone to *overfitting*, i.e. they are representing the training data almost perfectly, while performing badly on unknown data. This can be overcome by using RFs, which average the results of multiple independent and randomly built DTs. Using more DTs when building a RF leads to less overfitting. The randomness of DTs is achieved by a twofold strategy, namely randomly selecting data samples to build the tree and features in each if/else-question.

The number of DTs used within a RF and the number of features analyzed when looking for the best split are just two examples of so called *hyperparameters* that can be tuned when building a RF.

4.6 Software and Model Tuning

A pseudocode of the software is provided in Listing 1. On a metalevel it consists of three parts; loading and preprocessing of predictor (ERA5, PF) and predictand (GESLA) datasets, building the RF and evaluating the results.

We have already discussed the loading and preprocessing part in Section 4.4 and looked at the theoretical functioning of a RF in the previous Section 4.5. Before we built the model we (randomly) split the predictor and predictand datasets into training- and test-data, containing 75% and 25% of the total data

respectively. The training-data is then used to build and fit the model, while the test-data is used later on to evaluate the model.

In order to build the model, i.e. the RF, we used the *scikit-learn* package within *Python*. The standard classifier of this package is the *RandomForestClassifier (RFC)* from the *sklearn.ensemble* module (a documentation can be found online under `https://scikit-learn.org`).

This classifier can be tuned in several ways by altering its hyperparameter (HP)s. 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 amount of features used when calculating the best split (*max_features*). In general a larger value for *n_estimator* will lead to less overfitting as the results of many DTs are averaged. 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 $\sqrt{n_{\text{features}}}$, we varied the other two.

In addition to those HPs we altered the following; *class_weight*, *oob_score* and *random_state*. The *class_weight* is used to associate weights with classes. This is particular important in this study as we deal with extreme storm surges. Hence the predictand dataset is unbalanced as there are by definition 95% of data points of class 0 (no extreme storm surge) and only 5% of class 1 (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 mistakes made in extreme storm surge predictions (true positives) heavier than mistakes made in predicting no surge (true negative). The *oob_score* is set to "True" meaning that *out-of-bag*-samples are used to calculate scores for each DT within the RF. We also set the *random_state* to 0, which gives us and the reader the possibility to reproduce results. All other HPs were left at their default value.

For the HPs *n_estimator* and *max_depth* we started out to manually depict their value based on the concept of *validation curves*. As this method got to cumbersome we switched to automatically finding the best combination of HPs using *GridSearchCV* and *RandomSearchCV*, two optimization processes within *scikit-learn*. One can pass a list of values for each HP to these functionalities and they search for the best combination automatically using cross-validation. While *GridSearchCV* searches among all possible combination of HPs and their value lists, *RandomSearchCV* subsamples these lists and only uses a fixed number of parameter settings. This optimization is computational expensive, especially when using *GridSearchCV*. Hence, we decided to only use *RandomSearchCV* in order to make multiple forecasts for all stations.

After several attempts we realized that many predictions were prone to overfitting, hence we strongly reduced the values of the *max_depth* parameter, drawing only from the list $[1, 2, 3]$. For the *n_estimator* we used either 333, 666 or 1000.

All settings are summarized in Table 3 for replication purposes. The software is publicly accessible on GITHUB (`https://github.com/bkaib`), together with a more elaborated description and comments. The evaluation part will be described further in Section 4.8.

When all HPs have been optimized the model can be fit to the training data, extracting information from predictor data to conclude with a prediction.

4.7 Conducted experiments

Before fitting a model, we can control what information should be included in the predictor data-set. This gives room for several try-outs to investigate which (combination of) predictors and timelags work best in order to predict extreme storm surges.

We build 6 overarching experiments (**A** − **F**) each containing 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 each predictor individually with timelags up to a week (**A**). From this, timelags up to three days showed promising results (see Section 5). Additionaly those timelags are interesting in order to predict storm surges in advance. Hence, we analyzed a combination of all predictors (ERA5 and PF) with timelags of 1 and 2 days, eventually getting insight on which predictors are most important (**B**). Further interesting combinations of predictors and timelags were drawn from the theory Section 2 and are analyzed in experiments **C** – **E**. In experiment **C** we investigate the resonance coupling of strong winds and moving low-pressure systems by combining SP and U10 with various timelags. Because the westwind is an important driver of storm surges in the BS due to the connection with the North Sea via the Kattegat and the possible wave-setup in north-eastern region, we combined multiple timelags of U10 in experiment **D**. In **E** we looked at cumulative rain (TP with several timelags 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 timelags (up to 30 days) of PF in experiment **F**.

All experiments and model runs are summarized in Tables 4 – 9.

4.8 Evaluation Methods

A common tool to evaluate binary classification models is the Confusion Matrix (CFM) (see Figure 12). It summarizes the accuracy of a model in terms of 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 Figure 12 for example $n_1 = 29$ out of $n_\epsilon = 40$ extreme storm surges were correctly predicted, leading to a TPR of $\frac{n_1}{n_\epsilon} = 72.50\%$. Note, that n_ϵ varies amongst model runs even for the same station, because the one-hot encoding is applied before intersecting time-intervals of GESLA and ERA5 data.

A high TPR automatically leads to a low False Negative Rate (FNR) since they are linked via

$$\text{TPR} + \text{FNR} = 1. \tag{11}$$

The FNR indicates how often 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 predicitions of security. Especially for extreme events this can lead to devastating damage to societies when protection measures rely on model predictions with a high FNR. Eventually no measures are taken due to a model prediction of "no storm

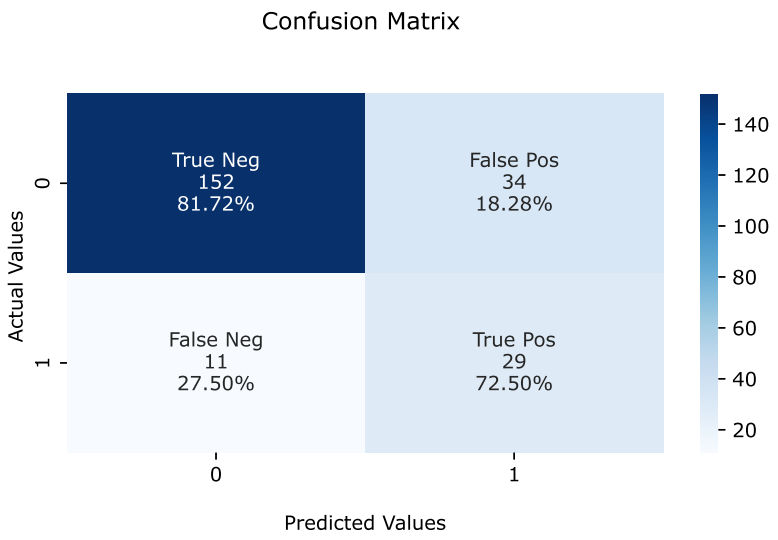


Figure 12: Confusion matrix for a binary classification model with absolute and relative values. The colorbar shows the maximum count of instances for all cases.

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 to (unknown / new) data. While for our study mostly the test-data set contains less cases of extreme storm surges than the validation set, the latter is more interesting to look at. Note that for stations 3 and 4 no validation sets were used.

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 of the research area are important for model decisions for each predictor.

Unlike a correlation coefficient the FI does not encode which class a feature is indicative of (see Müller (2017)), e.g. whether low or high pressure systems are important within the area of importance is not explained. Hence, we combined those areas with actual values of a predictor on the whole research area leading to a Predictor Map (PM). Only the top 1% area of importance regions is shown (grey squares), which was calculated by the 99th percentile of FI for each predictor. We investigated mainly two cases of PMs, namely True Positive Prediction (TPP) and False Negative Prediction (FNP), which were compared amongst each other as in Figures 13a and 13b. 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. In contrast it also suggests, that in some cases storm surges are induced even though there are high-pressure systems around.

To investigate further we calculated mean PMs for each of those cases and looked at the average difference of both maps as depicted in Figure 13. As it is sufficient to only show maps for TPPs and the difference to FNPs we will do so in the results section.

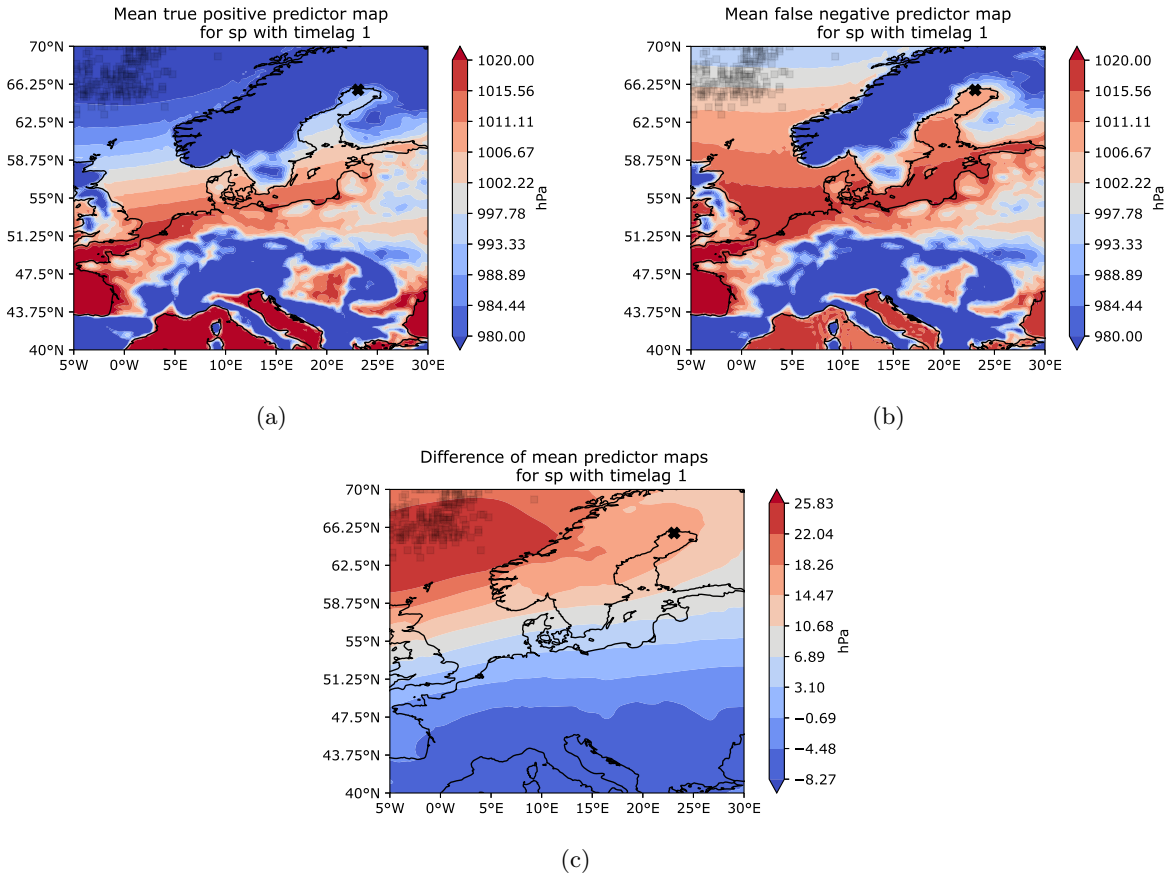


Figure 13: Mean Predictor Maps of SP all with $\hat{t} = 1$ for station 0 (NSWE). (a) mean True Positive Prediction (TPP)s, (b) mean False Negative Prediction (FNP)s and (c) difference of both means FNP - TPP. Note the different scaling of the colorbar for the difference maps.

5 Results

As we need to conclude several model runs for all stations, we will not display all results but rather focus on the most interesting ones for experiments **A** − **F**. We selected promising results based on a combination of the TPR of the test and validation datasets, labeled 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 parenthesis. In contrast to ERA5-predictors the prefilling dataset contains more instances of storm surges due to the hourly recording time, hence looking at TTPR instead of VTPR is also sensible.

5.1 A - Single Predictors and Multiple Timelags

In order to get an idea of how each predictor is influencing storm surges at specific stations, we used them in isolation while adding timelags up to a week.

For station NSWE, located in the most northern part of the BS, the best ERA5 predictors are SP and V10 both with a timelag of one day. While the TTPRs for SP and V10 are 56.52% and 60.87% ($n_\epsilon = 23$), the more important VTPRs are significantly higher with 70.67% and 73.33%, respectively ($n_\epsilon = 75$).

In Figure 14 the corresponding mean PMs are shown. For SP the Area of Importance (AoI) is within the region 5°W – 5°E and 62.5°N – 70°N. Here, mainly low pressure systems with less than 980 hPa lead to

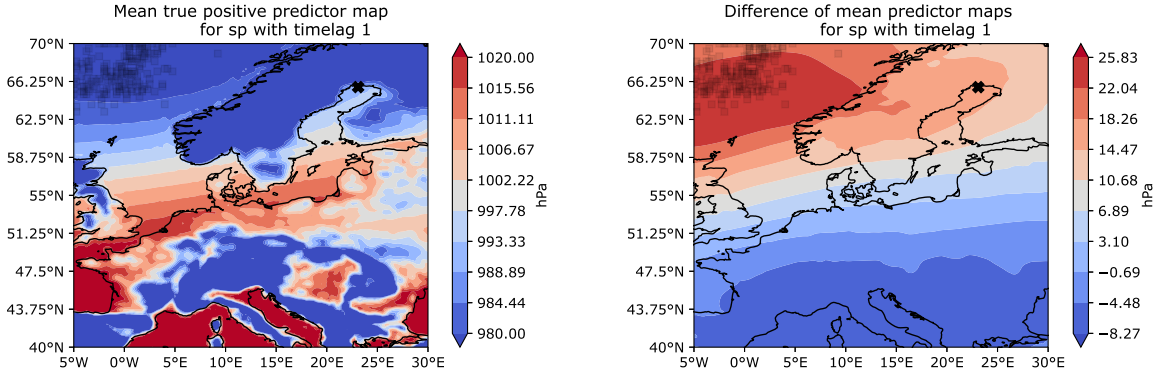
a correct prediction of a surge. The model tends to False Negative Prediction (FNP)s once the pressure in the AoI increases by a mean of around 25 hPa. In some cases high pressure systems of more than 1020 hPa occurred in the AoI for FNPs. The AoI of V10 is located close to the boarder of Sweden and Norway and stretches towards the north east. Mainly light southwinds (i.e. winds blowing from the south towards the north) 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.

Interestingly the southwind does not seem to be important right over the BS, even though the general southerly wind field of the mean PM is quite large and covers the whole BS. Even more curiously the AoI of SP is not even close to the BS but rather covers parts of the European North Sea.

Additionally U10 showed reasonable results with VTPRs greater than 60% for zero, one or five days of timelag. While for $\hat{t} = 1$ the AoI clusters in the North Sea close to the entrance to the Danish Straits (see Figures 14e - 14f), it spreads more over the Danish Straits and into the BS for longer timelags like $\hat{t} = 5$ (not shown). Regardless of the AoIs location, the main wind-direction is eastward. For TPPs of $\hat{t} = 1$ the westwind-fields are up to 8 ms^{-1} stronger (on average) than for FNPs. This behavior generally repeats for other timelags.

All mentioned predictors show behaviors that were expected by literature, when only looking at TPPs. Amongst those are low-pressure fields (baric waves), southwind (pushing watermasses towards the station) and (strong) westwinds (pushing water through the Danish Straits). Nevertheless, the cases of FNPs show that there is also different behavior for some extreme storm surges, which might be due to the isolated usage of predictors (see Section 6).

Despite the good performance of ERA5-predictors the most important predictor in terms of TTPR is prefilling. For timelags of two and seven days the TTPRs are 85.71% and ($n_\epsilon = 2254$) and 84.12% ($n_\epsilon = 2337$), respectively.



(a) (b)

Figure 14: Mean Predictor Maps of SP (a, b), V10 (c, d), U10 (e, f) all with $\hat{t} = 1$ for station 0 (NSWE). Left column shows TPPs, right column shows the difference of FNPs and TPPs. Note the different scaling of the colorbar for the difference maps.

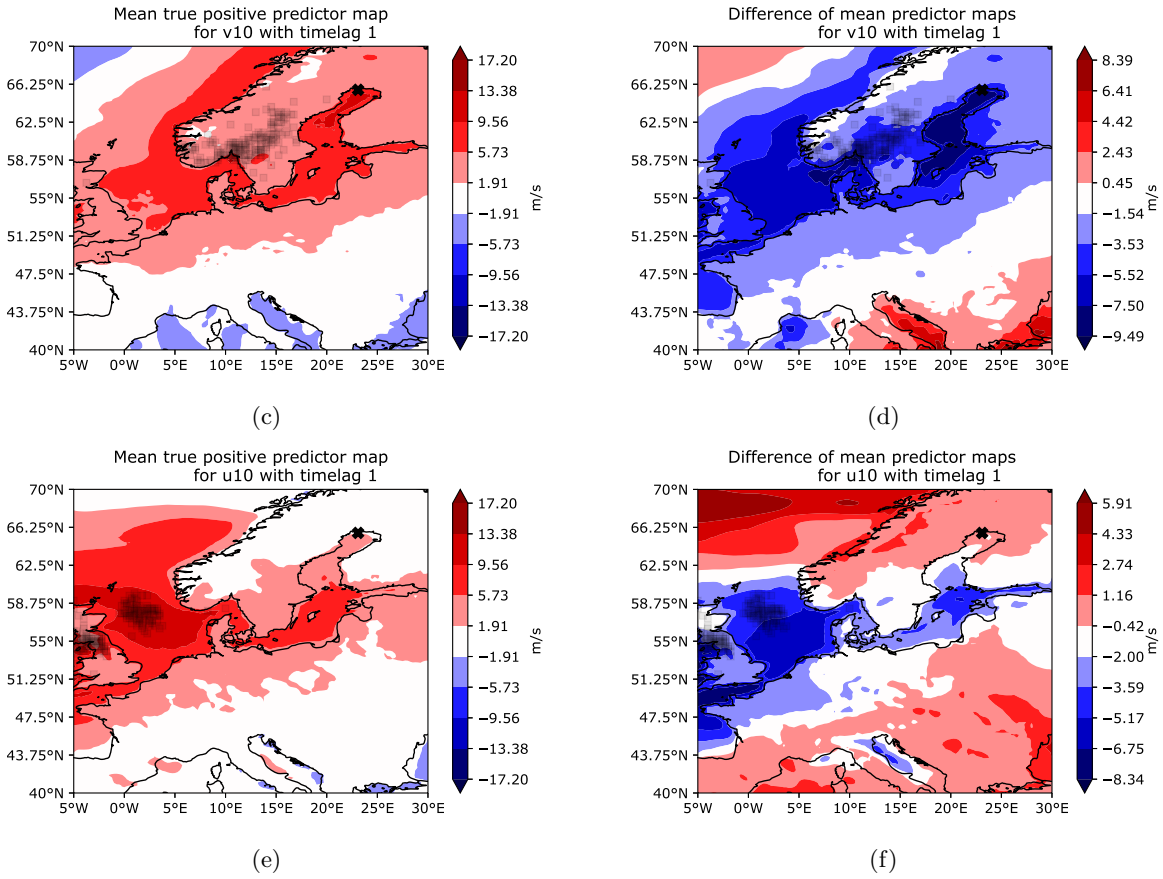
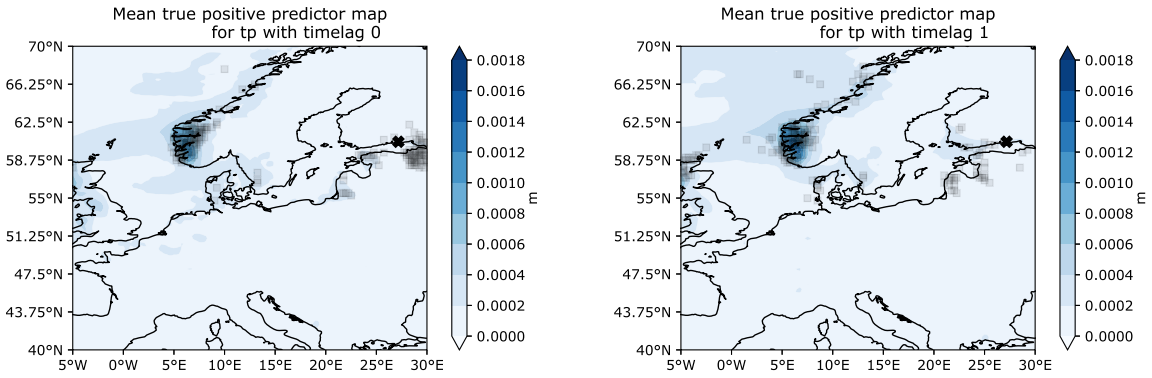


Figure 14: Mean Predictor Maps of SP (a, b), V10 (c, d), U10 (e, f) all with $\hat{t} = 1$ for station 0 (NSWE). Left column shows TPPs, right column shows the difference of FNPs and TPPs. Note the different scaling of the colorbar for the difference maps (continued).

Slightly different behavior in terms of important predictors is observed for station 2 (FINBAY). In contrast to station NSWE the meridional wind is not as important, rather the zonal wind-component U10 is. With VTPRs of 73.39% 72.48%, 69.62% (all with $n_\epsilon = 109$) for timelags zero, one and two respectively, the short term wind-fields are most important. Increasing the timelag up to a week lead to worse results of VTPRs around 60%. The best performing ERA5-predictor is again SP with a VTPR of 78.9% ($n_\epsilon = 109$) for timelag zero. Again, timelags of one or two days showed good results, having VTPRs around 70%. In general though the VTPR drops with increasing timelags. Also quite promising results could be achieved when using TP without any timelag or $\hat{t} = 1$. It resulted in VTPRs of 69.72% and 72.48%, respectively (both $n_\epsilon = 109$). For total precipitation the VTPR drops quite fast to around 50% when increasing the timelag. Again the overall best predictor seems to be the PF. The highest and lowest TTPRs are 88.52% and 74.45% ($n_\epsilon = 3699$) for $\hat{t} = 0$ and $\hat{t} = 7$, respectively (TTPRs drop consistently when increasing the timelag).

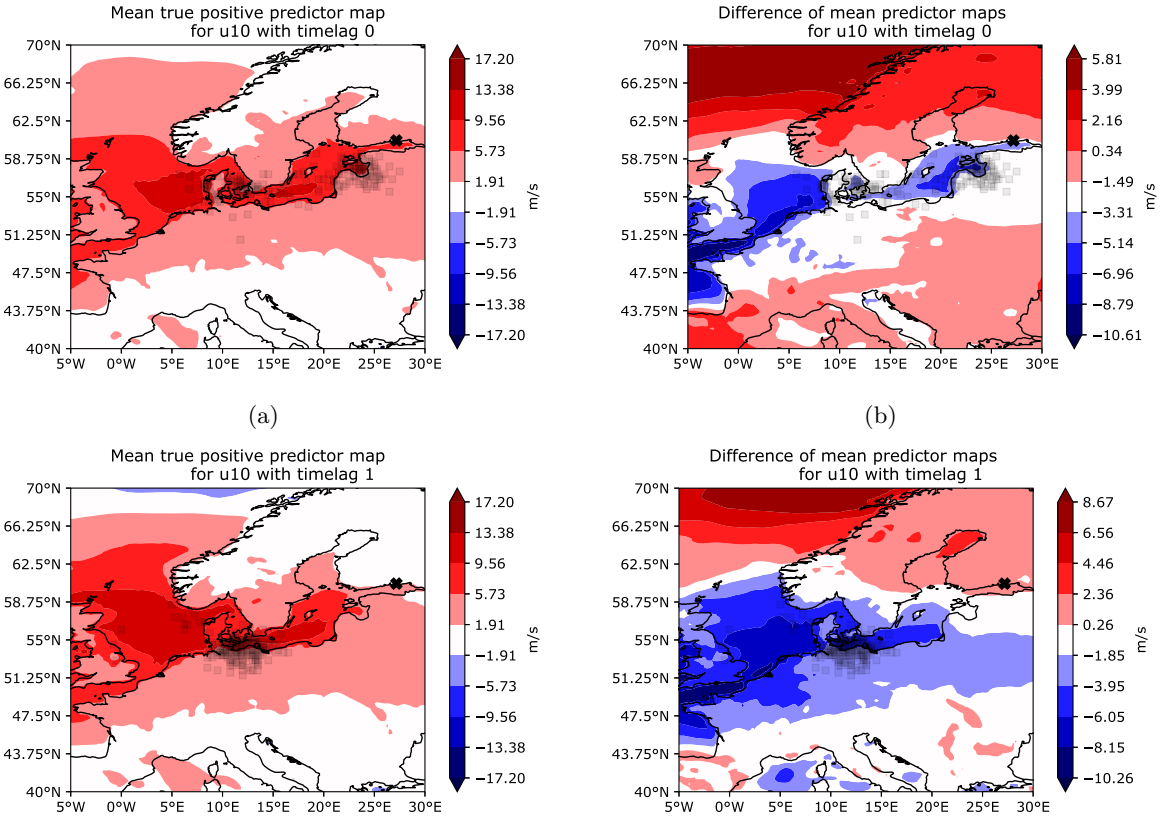
Similar to station NSWE the AoI of TP for $\hat{t} = 0$ is close to the station itself. When increasing the timelag by only one day, the AoI shifts toward the area around Bergen, sometimes showing connecting patterns of importance across the North Sea towards the United Kingdoms (see Figure 15).



(a) (b)

Figure 15: Mean Predictor Maps of TPPs for predictor TP with timelags (a) 0, (b) 1.

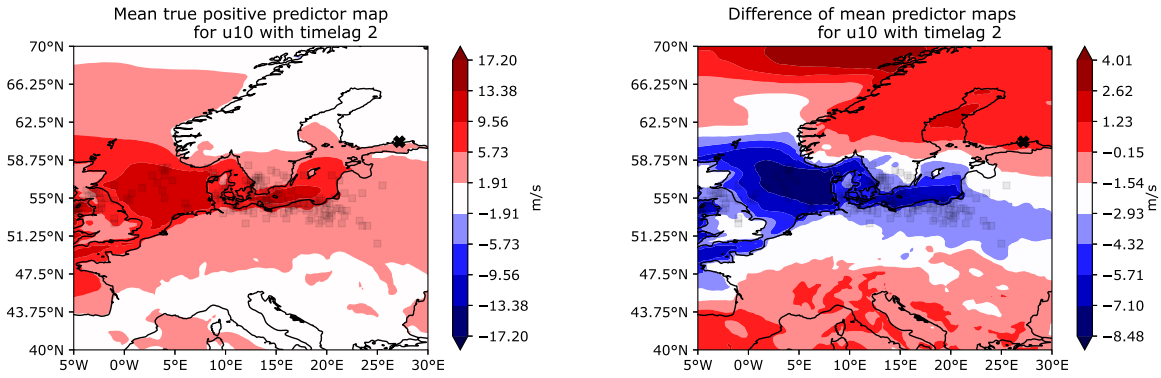
The PMs of SP show similar patterns as for station NSWE (not shown). For short timelags up to a couple of days, the AoI is located in the Eurpoean North Sea around 70°N again, covering almost the whole zonal extent of the research area. Interestingly the AoI clusters more around the station FINBAY when increasing the timelag for instance to five days. Regardless of the timelag though, there is always low pressure systems in the AoI for TPPs, which increase several hPas when looking at FNP maps.



(a) (b)

(c) (d)

Figure 16: Mean Predictor Maps of U10 with timelags 0, 1, 2 per row (top to bottom). Left column shows TPPs, right column shows the difference of FNPs and TPPs. Note the different scaling of the colorbar for the difference maps.



(e) (f)

Figure 16: Mean Predictor Maps of U10 with timelags 0, 1, 2 per row (top to bottom). Left column shows TPPs, right column shows the difference of FNPs and TPPs. Note the different scaling of the colorbar for the difference maps (continued).

The PMs of U10 are shown in Figure 16 for timelags of zero, one and two days. For $\hat{t} = 0$ the AoI clusters around the Gulf of Riga and only covers the Region of the Kattegat lightly. This is interesting, as one would expect important short term windfields 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 timelags of one and two days locates more around the Kattegat area and the Southern Coast of the Baltic Sea. The former is especially true for $\hat{t} = 1$, while the latter can be seen for $\hat{t} = 2$. Mean westwind speed of around 12ms^{-1} occur 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 observed. Comparing the maps of TPPs to the ones of FNPs one can see, that the model generally leads to false predictions, when westwind fields become weaker. The difference maps show a mean decrease of westwind speeds of around 7ms^{-1} in parts of the AoI. Hence, the model is not as reliable when westwinds are not strong, no winds or even eastwinds occur.

Qualitatively station 1 (FIN) can be compared to station 2 (FINBAY). While TP and SP are good predictors, the overall best predictor is again PF with TPRs over 90%. The meridional wind is not important, but again strong westwinds are, leading to U10 as the most important ERA5-predictor for this station.

In contrast eastwind is used for model-predictions at station 4 (DEU) for $\hat{t} = 0$ (see Figures 17a – 17b). When increasing the timelag to seven days, this behavior switches and westwind becomes important again (see Figures 17c – 17d). Furthermore, SP and TP are not good predictors for this station, having VTPRs below 50%. Note though, that for this stations there were just few surges to analyze. Hence, model training is impaired.

Regardless of impaired training, this resembles with theory though. 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 induced wind setup at this station. Another mentioned driver for storm surges along the German Baltic Coast are Seiches. These might be induced by the pronounced westwind around the station and over the BS for $\hat{t} = 7$. The long timelag could be sufficient for a wave-setup at the opposing coast which

in turn leads to seiches once the wind turns westward or stops blowing.

Even though only few instances of storm surges could be analyzed, the predictor PF lead to good results with TTPRs around 70%. Compared to other stations, this number is slightly decreased though.

Similar to station 4, also station 3 (LVA) only dealt with a small number of storm surges. It showed similar results as stations 0 – 2.

Note though, that no validation sets were used for both stations, hence the interpretation of TPRs must be taken cautiously.

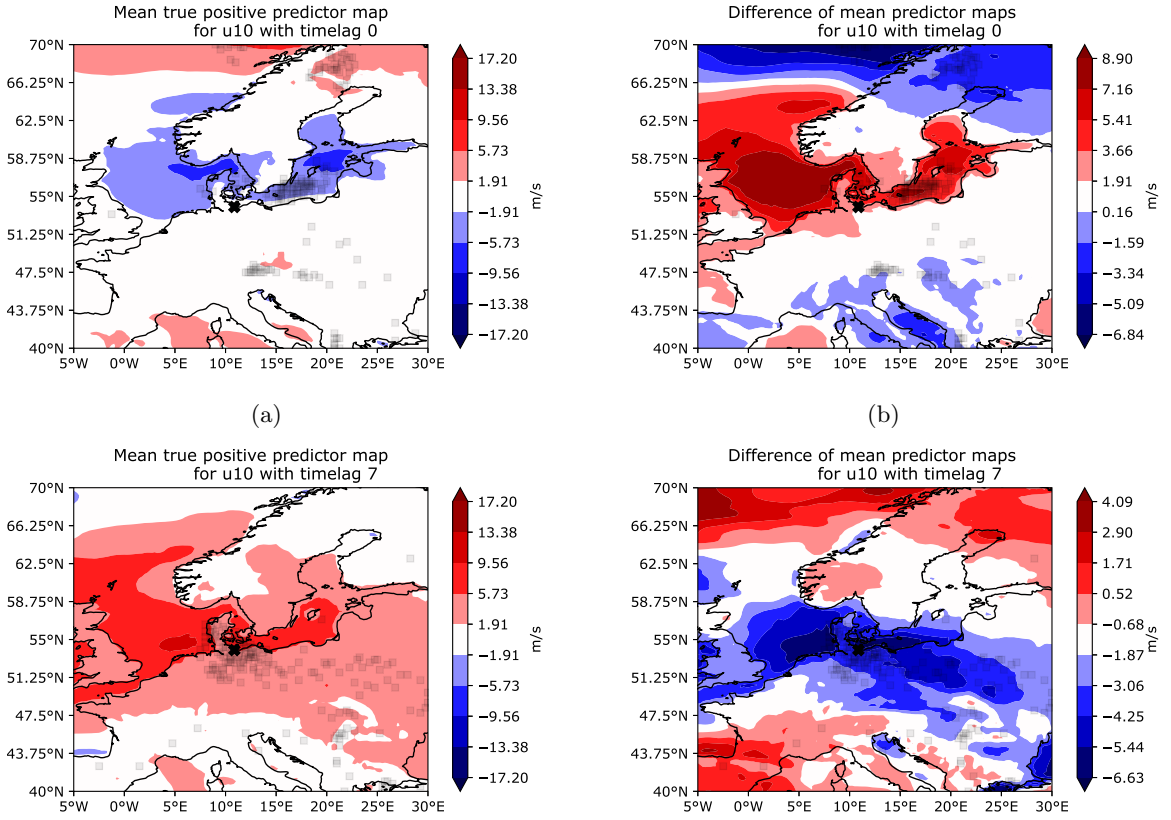


Figure 17: Mean Predictor Maps of U10 with timelags 0, 7 per row (top to bottom) at station 4 (DEU). Left column shows TPPs, right column shows the difference of FNPs and TPPs. Note the different scaling of the colorbar for the difference maps.

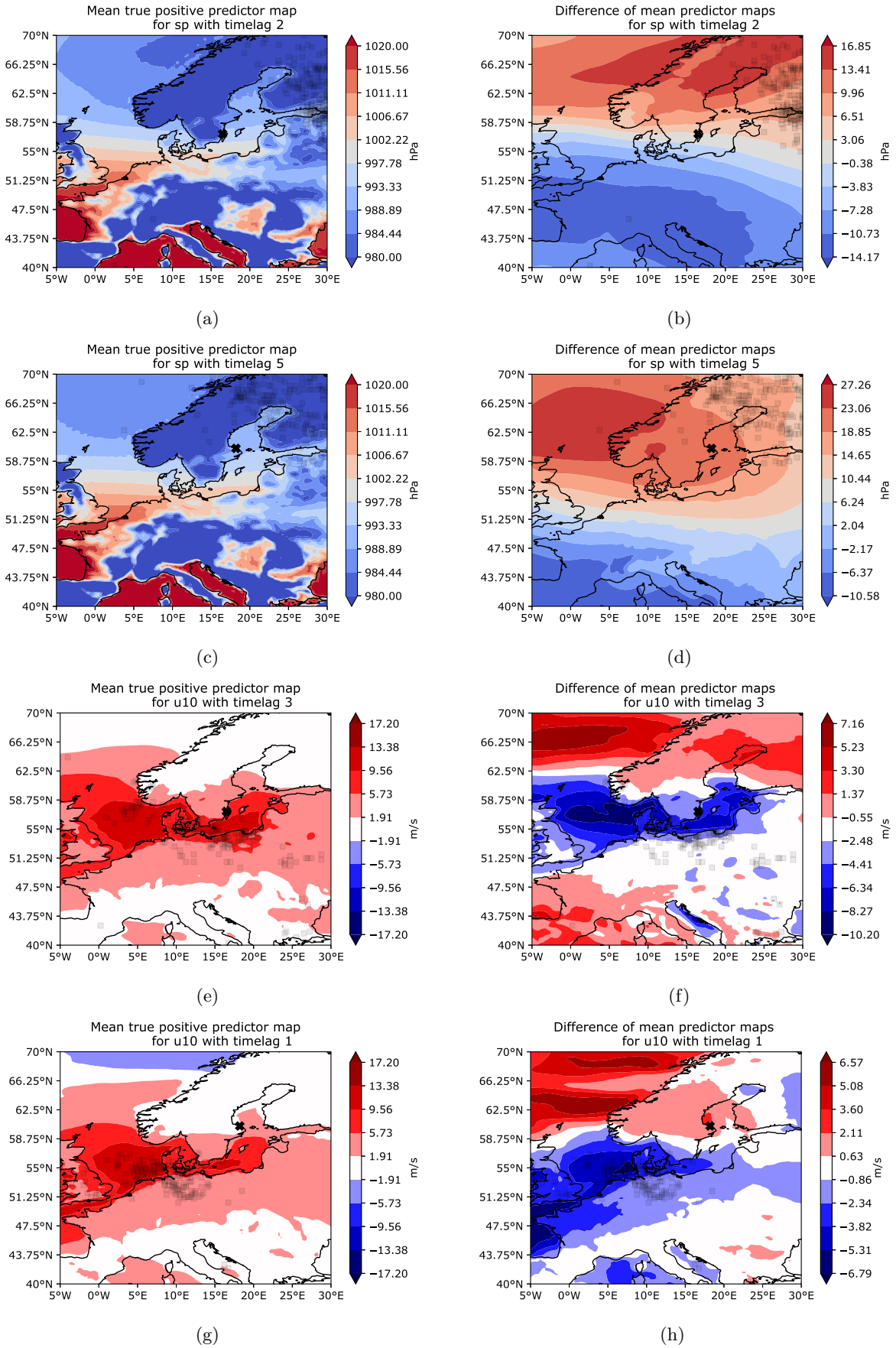


Figure 18: Mean Predictor Maps of SP (timelags 2, 5), U10 (timelags 3, 1) starting with stations 5 (WSWE) alternating in rows (top to bottom) with station 6 (WSWE2). Left column shows TPPs, right column shows the difference of FNPs and TPPs. Note the different scaling of the colorbar for the difference maps.

Stations 5 (WSWE) and 6 (WSWE2) showed similar qualitative results in terms of PMs for almost all predictors, i.e. low-pressure systems, strong westwinds and heavier rainfall within the AoIs of both stations. The AoIs for SP were mostly important around the Gulf of Finland and Gulf of Bothnia. If high pressure fields occured in those regions, the model tends to false predictions (see Figures 18a – 18d). For U10 the area around the Danish Straits is of importance for both stations. The AoIs mostly start close to the North Sea entrance of the Danish Straits and stretch along the southern coast of the BS, ranging almost 2 degrees land-inwards towards the South. Similar to the stations before, strong westwinds in those areas are leading to TPPs and lighter winds to more FNPs (see Figures 18e – 18h). For both stations TP seems to be important mainly around Bergen with TP > 0.0014m for TPPs.

Both stations differ in their quantitative results. While WSWE relies more on SP with timelags of one or two days (VTPRs around 70%), WSWE2 shows its best VTPRs of 71.43% ($n_\epsilon = 132$) for U10 with a timelag of one day. Nevertheless both work for each station quite reasonably. Precipitation can only be used for WSWE2 (VTPRs around 65%), while for WSWE the VTPRs are below 50%. This is interesting as both stations are very close to each other. The VTPR of V10 generally lies below 50% for both stations, except for a timelag of four days at station WSWE2. In both cases the PF is the overall best predictor with TTPRs of around 88%.

From this experiment 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 looking at windfields. For SP the model uses always low-pressure systems in order to conclude with TPPs.

For most stations SP and U10 play the most important role, showing mainly low-pressure and westwind fields. In general timelags up to three days were used. Choosing longer timelags often lead to worse results. Overall PF was the most useful predictor for all stations. The results are summarized in Figure 19.

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, why this is happening. Keep in mind though, 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 westwinds, is inducing the storm surge. Therefore we will analyze combinations of predictors in the upcoming experiments.

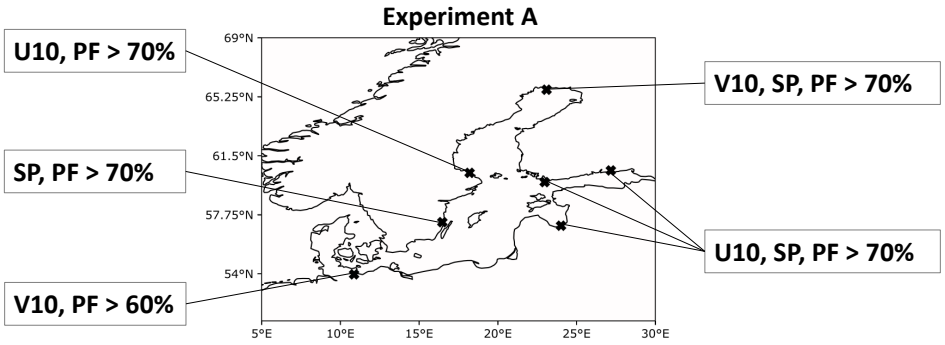


Figure 19: Summary of best predictors per station for experiment A. The percentage indicates the corresponding VTPR or TTPR.

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 behavior of their corresponding PMs.

Altogether the predictors show quite similar results and patterns in PMs as in Section 5.1, where we looked at each predictor in isolation. For almost all stations SP and U10 are the most important predictors. They again show pronounced low pressure fields (below 980 hPa) and strong westwinds (greater than 15ms^{-1}) in their respective AoI.

Only for the most southern and northern stations, this behavior switched and V10 important as well. More strikingly SP can not be used for the station in Germany at all. Theoretically this can be explained due to the fact, that pressure systems usually travel from the South-West towards North-East in the area of research, hence carrying water masses via a baric wave away from the station. This would also explain why SP is an important predictor for station NSWE.

For stations 0 – 2 a timelag of one day leads to better results (VTPRs around 70%) compared to a timelag of two days. This behavior switches for stations 5 and 6.

In Figure 20 we compared westwind PMs of stations 1 – 3, all with a timelag of $\hat{t} = 1$. While the VTPRs of station 1 (FIN) and 2 (FINBAY) were above 70%, station 3 (LVA) only resulted in a TTPR of around 60%. From the PMs one can readily see the importance of pronounced westwind fields for the model to provide TPPs together with weaker performance when the westwind slows down or even turns eastward. What is more interesting in this case though is that only slight changes in the location of the station lead to bigger changes in the AoI. This can be seen in Figures 20a and 20e, when switching from station 1 at the entrance to the Gulf of Finland to station 3, which is deep in the Gulf of Riga. While westwind fields around the German Baltic Coast (and further inland from 8°E – 16°E and 51.25°N to 53°N) are important for the former, the latter relies on windfields much closer to the station. This difference might

be due to the orientation of the coastline, where the respective station is located.

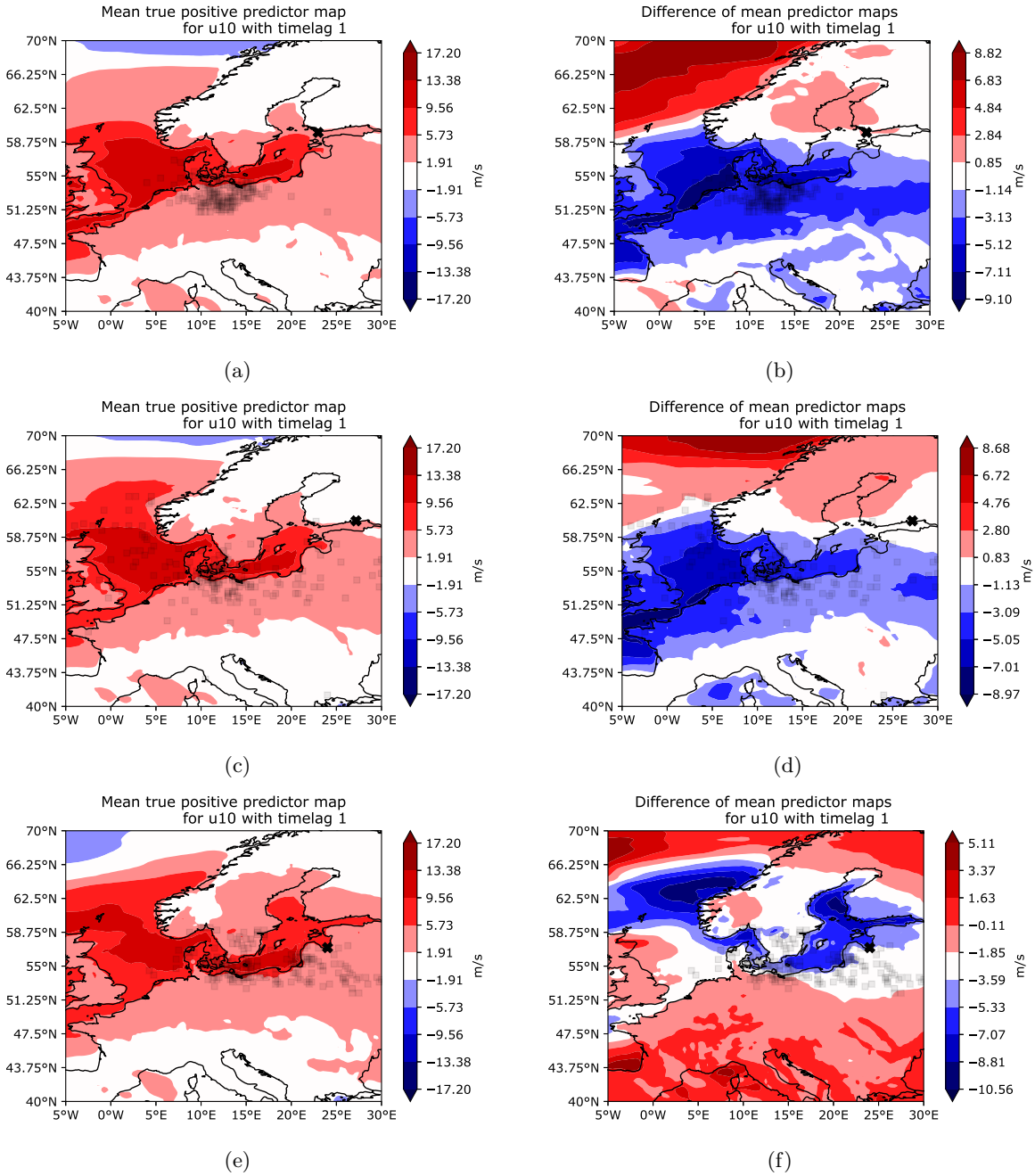


Figure 20: Mean Predictor Maps of U10 with $\hat{t} = 1$ and stations 1 – 3 in rows from top to bottom. Left column shows TPPs, right column shows the difference of FNPs and TPPs. Note the different scaling of the colorbar for the difference maps.

While station 1 directly faces the open waters of the BS, station 3 is more sheltered. Hence, windfields around the Southern Baltic that blow towards the North East may induce a wave setup at station 1. Even when comparing station 1 and 2, i.e. Figures 20a and 20c, one can see slight differences. The main cluster of AoI remains unmoved but for the station located deeper in the Gulf of Finland, the model relies stronger on the North Sea entrance towards the Kattegat. This subtle difference could indicate that for this station not only the wavesetup is important but also the prefilling due to watermasses being pushed into the BS.

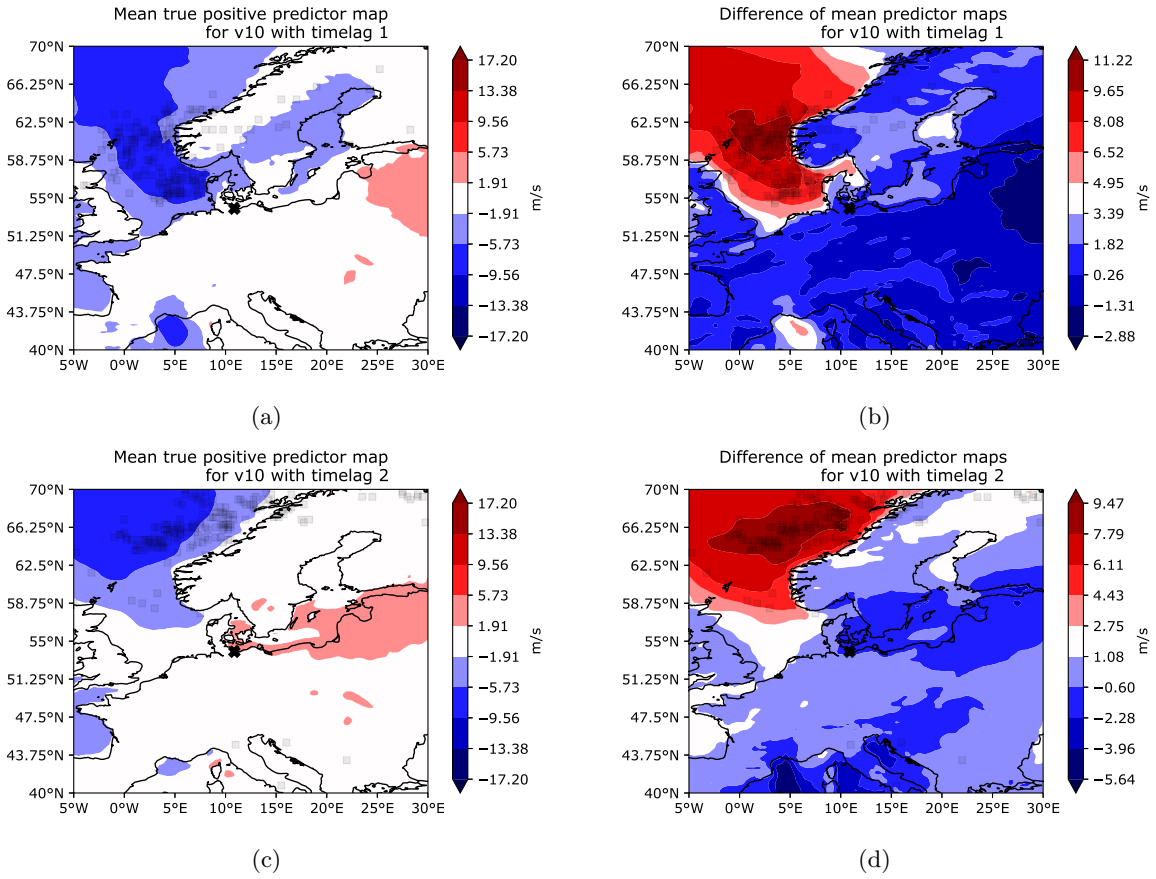


Figure 21: Mean Predictor Maps of V10 at station 4 (DEU) with timelags of one and two days for rows from top to bottom. Left column shows TPPs, right column shows the difference of FNPs and TPPs. Note the different scaling of the colorbar for the difference maps.

Another interesting shift of AoIs, in this case due to a change of timelag from one to two days, can be observed for station 4 (DEU). Initially a huge area within the European North Sea is covered (see Figure 21c) with winds blowing towards the South. The shorter the timelag becomes, the more the AoI clusters towards the North Sea entrance of the Danish Straits (see Figure 21a). Still Northwinds are important in this area. A prevailed wind blowing from the North might also push water masses through the Danish Straits due to wave refraction. Despite this interesting behavior the TTPRs are quite low with 36.67% and only 32% for timelags one and two respectively.

Nevertheless using the predictors in combination showed an order of importance as depicted in Figure 22. We can see, that SP and U10 are mostly used by the model, but it also switches depending on the station. In terms of VTPR accuracy, using isolated predictors lead to better results in some cases.

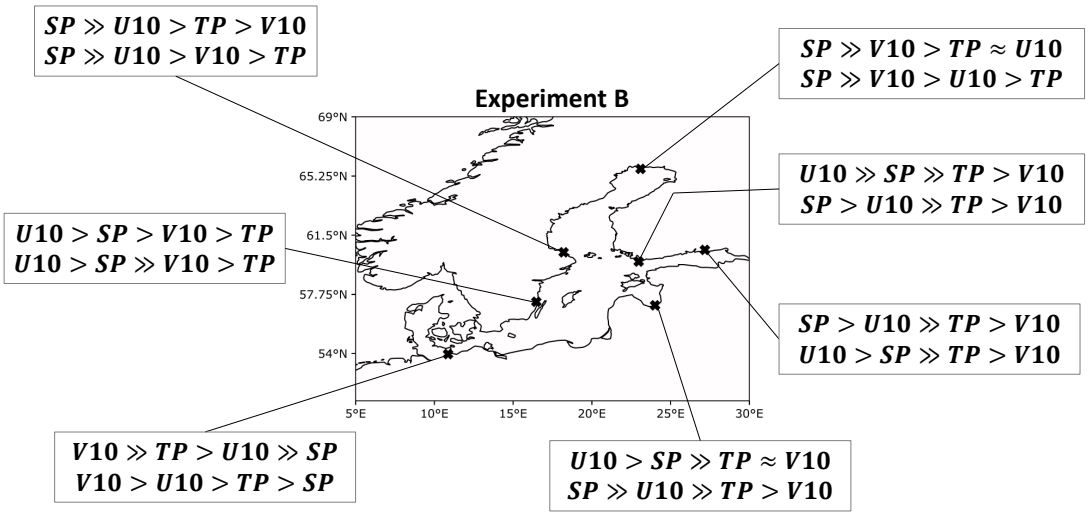


Figure 22: Order of predictor importance for experiment B with run_ids 1, 3. First and second row show $\hat{t} = 1, 2$, respectively. The >> sign indicates that the feature importance was almost double as high, the $\approx$ indicates approximate equality. For PF feature importance was close to zero.

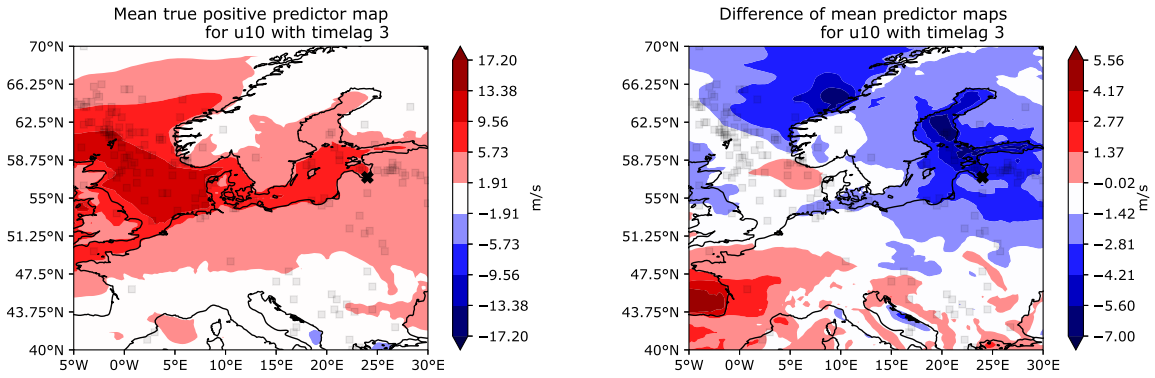
5.3 C - Coupling of U10 and SP

In the previous experiments we already observed that SP and U10 are important predictors. In theory resonance coupling of strong winds and moving weather systems (low-pressure systems) lead to extreme storm surges as well. Hence, we will now investigate several combinations of those predictors as shown in Table 6.

Combining both predictors with a similar timelag did not improve results compared to using them in isolation. For some stations (NSWE) it even lead to worse VTPRs. Nevertheless short term combinations with timelags up to 3 days seem to work best. Only in few occasions (station 1: FIN) timelags up to 5 days work.

Best results could be observed for station 2 (FINBAY) with the highest VTPR of 75.23% for $\hat{t} = 0$. For this station timelags of one or two days lead to VTPRs above 70%. All other stations had similar VTPRs above 60%, mainly for timelags up to three days. The only (expected) exception was station 4 (DEU) for which both predictors can not be used (TTPRs below 45%).

The PMs mainly showed similar behavior as for using isolated predictors. Mainly low pressure systems and strong westwinds were observed in the AoI. For most stations, models produce FNPs when higher pressure systems or no pronounced westwind fields are settling in the AoI. One interesting difference could be observed for station 3 (LVA) though. Here, the AoIs for U10 with $\hat{t} = 3$ are mainly clustering in the European North Sea and close to the station itself. For TPPs only westwinds occur in both areas. Interestingly though for FNPs these westwinds prevailed in the AoI around the North Sea, which eventually leads to prefilling via watermass-transfer to the Danish Strait. For the area close to the station (25°E – 30°E and 57°N), no westwinds prevailed, but rather the winds stopped blowing (see Figure 23).



(a) (b)

Figure 23: Mean Predictor Maps of U10 at station 3 (LVA) with $\hat{t} = 1$. Left column shows TPPs, right column shows the difference of FNPs and TPPs. Note the different scaling of the colorbar for the difference maps.

Furthermore, an increased timelag leads to shifted AoIs as similar to Section 5.2.

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. Hence, we used a shorter timelag for SP compared to U10 in the next set of the experiment.

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 timelags of SP and U10 did lead to more stable results when altering the timelag of U10. In other words the VTPR did not diminish quickly when increasing $\hat{t}$ of U10.

In terms of AoIs and PMs no differences to experiment A could be observed. Still low-pressure fields and westwinds are important factors.

When increasing the timlag of U10 already lead to more stable results in terms of VTPR, does it also help to incorporate short-term ($\hat{t} = 1$) and long-term (timelags up to 7 days) information of predictors into the model while maintaining a shorter timelag for SP compared to U10? The last set of runs tries to answer this question.

While for station 0 (NSWE) random AoIs for predictors with longer timelags suggest that the model relies on short timelags only, comparing the actual feature importance for other stations showed that the shorter timelag was just slightly more important. In general, most of the stations showed similar behavior in terms of AoIs and PMs as experiment A. One interesting behavior could be observed for station 1 (FIN) when using timelags 1, 5 and 1, 7 for SP and U10 respectively. While AoIs remained similar to previous experiments, the PMs vary at least for SP. For $\hat{t} = 1$ the difference-PM indicates a strong increase in pressure when comparing TPPs and FNPs. This is not true anymore when increasing the timelag to $\hat{t} = 5$ (see Figures 24b and 24d).

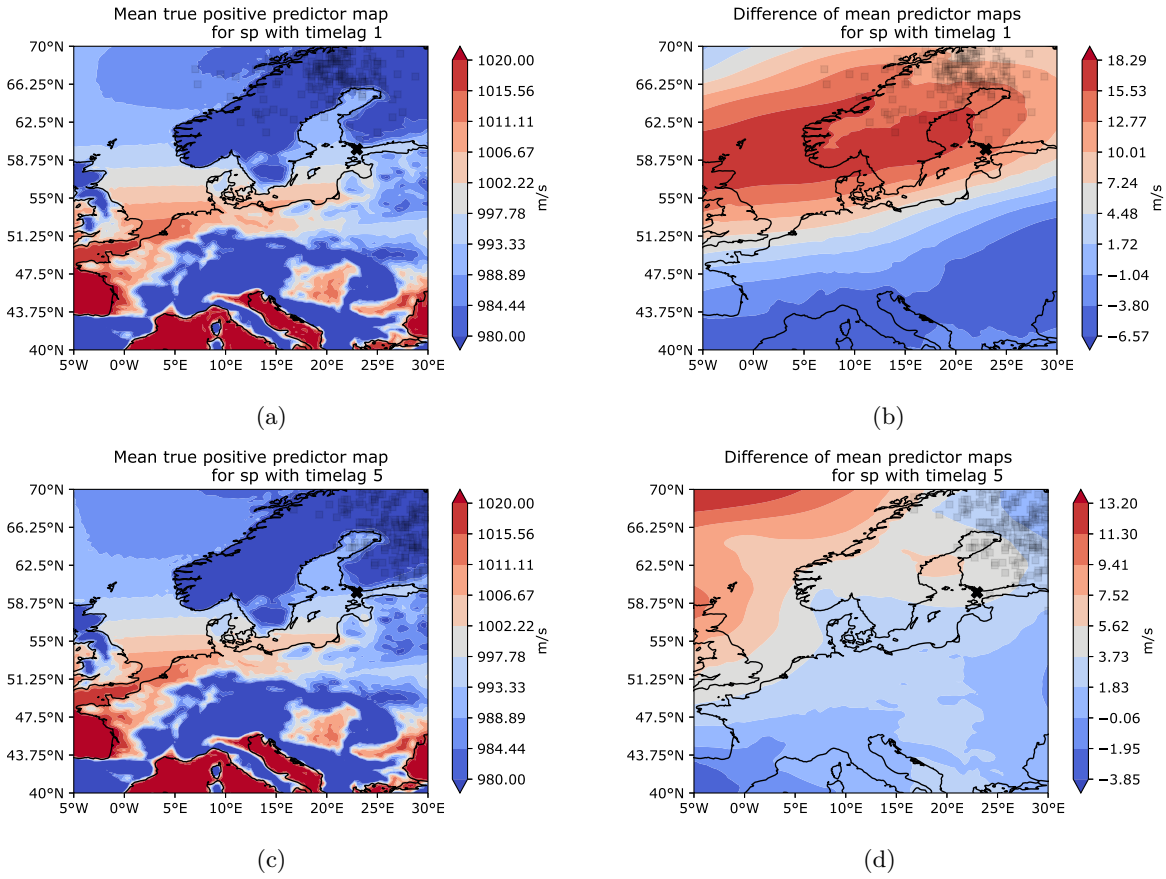


Figure 24: Mean Predictor Maps of SP at station 1 (FIN) with timelags one and five for rows from top to bottom. Left column shows TPPs, right column shows the difference of FNPs and TPPs. Note the different scaling of the colorbar for the difference maps.

In total this experiment showed, that for most stations a combination of short and longterm data as well as a positive difference in timelags between U10 and SP leads to good results in terms of VTPRs (see also Figure 25). The experiments analyzed up to now furthermore showed the consistent importance of prevailed westwinds around the area of the Danish Straits, which will be investigated in the next Section.

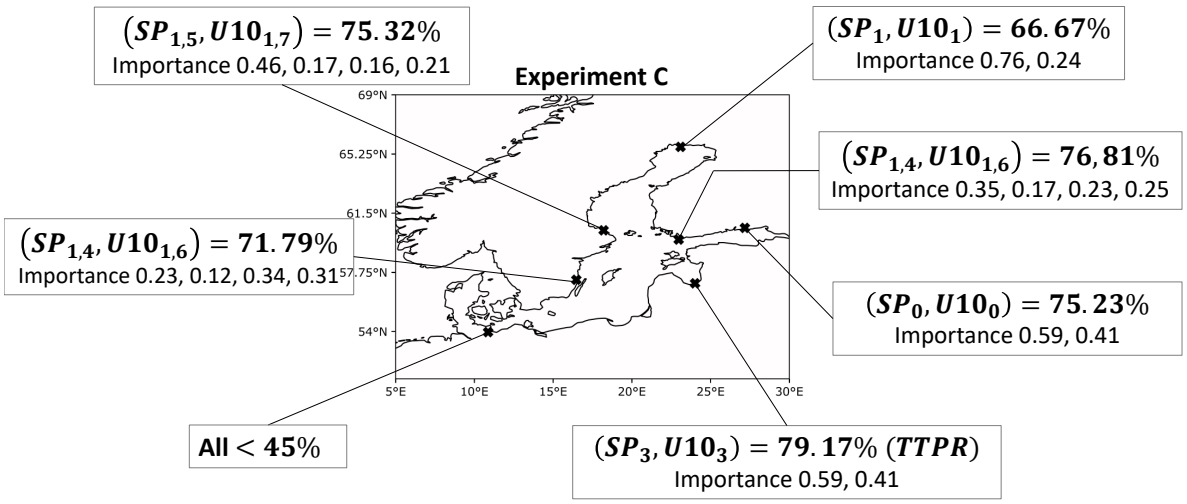


Figure 25: Best combinations of predictors for experiment C. Timelags of predictors are indicated in the subscript. Depending on the station VTPR or TTPR is shown.

5.4 D - Combinations of Westwind-Timelags

As was already shown, westwinds 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 timelag combinations as shown in Table 7.

We already know that a timelag of one or two days works well for U10 due to previous experiments. Hence we combined those short term timelags 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 $\hat{t}$ (run-ids 4 – 7). Finally we spread the timelags 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) was showing poor results.

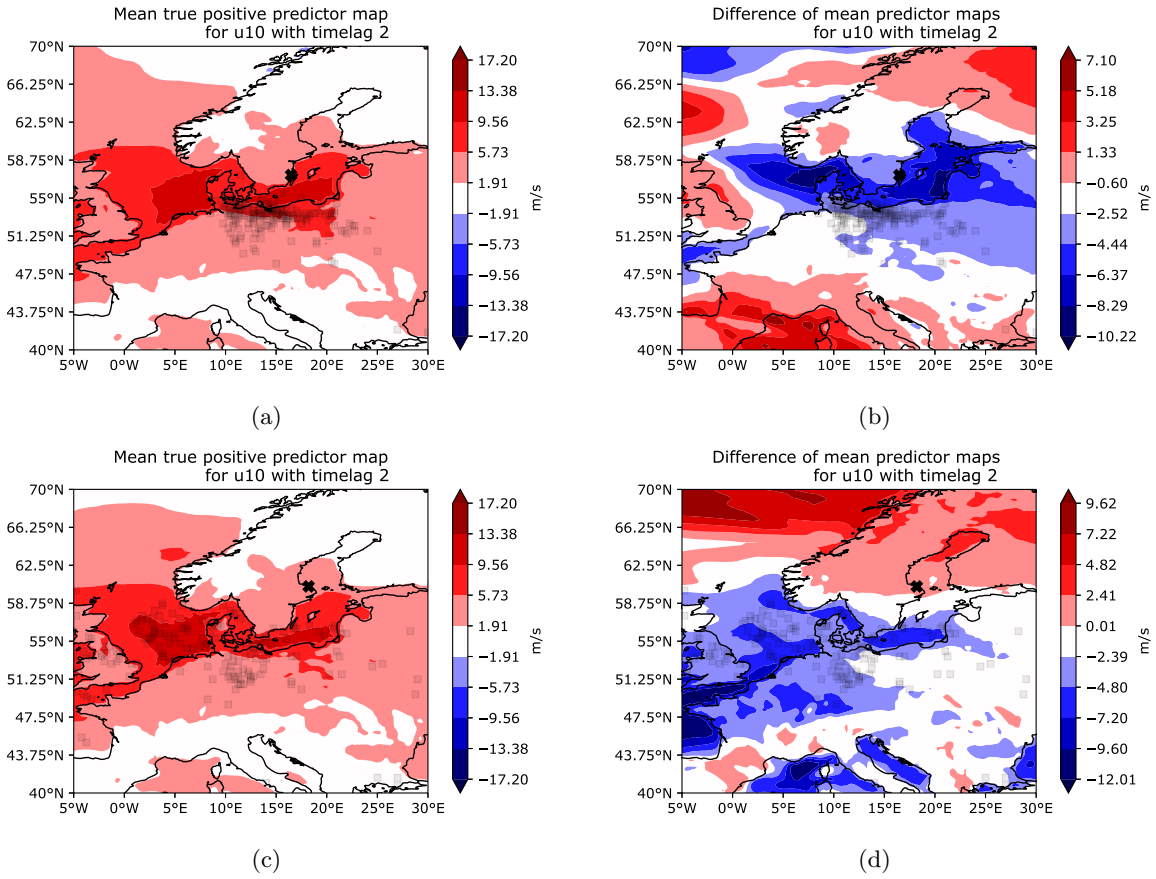


Figure 26: Mean Predictor Maps of U10 with timelag 2 at stations 5 (WSWE) and 6 (WSWE2) for rows from top to bottom. Left column shows TPPs, right column shows the difference of FNPs and TPPs. Note the different scaling of the colorbar for the difference maps.

Mostly timelags up to four days worked the best for all stations. For instance for station 1 the combination of timelags two, three and four days resulted in a VTPR of 79.71% ($n_\epsilon = 69$). When increasing the timelag up to a week, VTPRs of station 5 (WSWE) diminish. On the other hand a combination of short and long-term timelags (1, 3, 5, 7) gives promising results again for stations 1, 2 and 6. For the latter, for example, a VTPR of 77.92% ($n_\epsilon = 77$) was observed for this combination.

AoIs and PMs show again similar behavior to experiments before, i.e. mainly strong westwinds 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 like stations 5 and 6. Figure 26 depicts this for a timelag of two days. While for station 6 westwinds 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 westwinds 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. It is completely directed towards the South and South-West. Hence, westwinds may induce strong windsetup here and prefilling, i.e. the wind around the Danish Straits, is not as important for model predictions.

In summary combinations of U10 can be used for most stations as a good predictor when focusing on timelags up to four days (see Figure 27. For some stations timelags up to a week also lead to good predictions. Even longer timelags should not be used as they are mostly disregarded by the model.

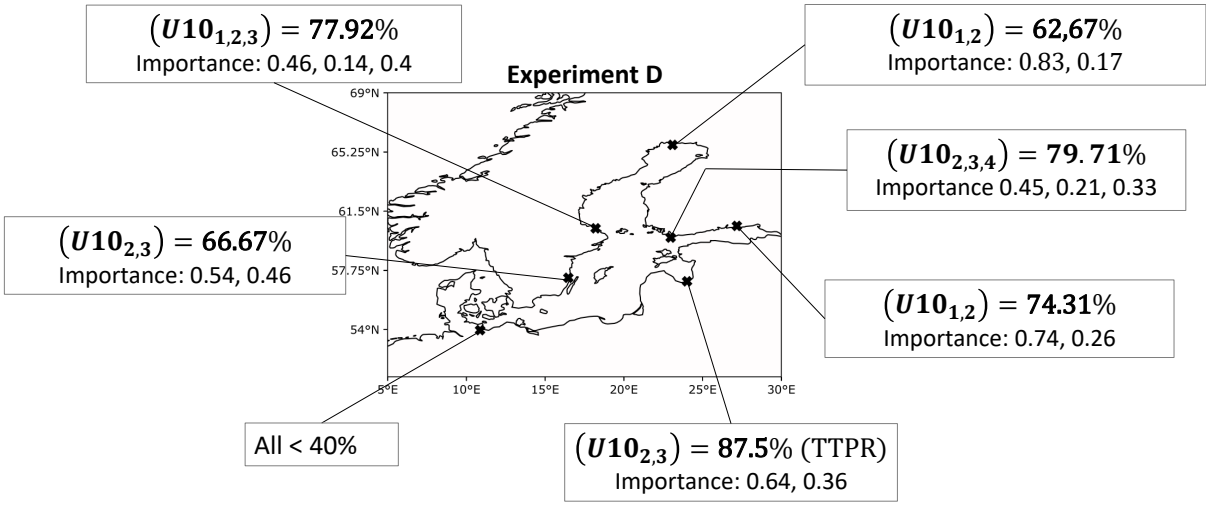


Figure 27: Best combinations of predictors for experiment D. Depending on the station VTPR or TTPR is shown. Importances are ordered as subscripted timelags.

5.5 E - Predictor Combinations from Theory

In this experiment we tried to emulate the effect of cumulative rain and looked at how information on prefilling changes the behavior of the westwind for model predictions. The combinations of predictors can be found in Table 8 and accuracy-results are summarized in Figure 28

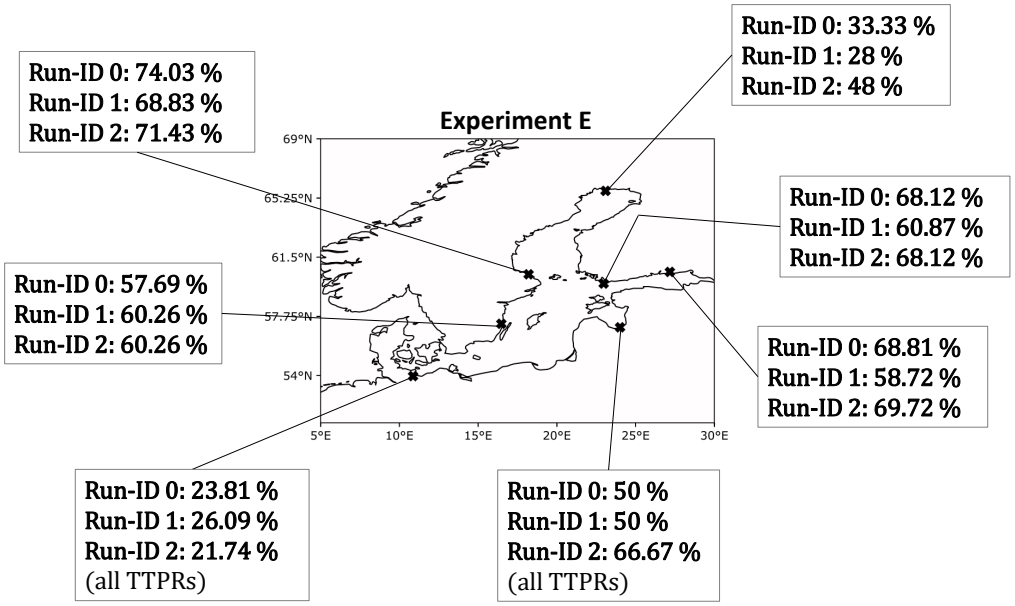


Figure 28: TTPR or VTPR results for all run-ids of experiment E.

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. When looking at the importance though, one can see that mostly U10 is used for model predictions. For all stations except station NSWE the sum of TP feature importance is smaller than the feature importance of U10. This insight is further more backed by the PMs and AoIs of TP which do not show consistent structures.

In general the VTPRs in this experiment were always larger than corresponding TTPRs, which is kind of an interesting feature. The latter were mostly below 60% and for some stations even way below 50%. This is why stations LVA and DEU were not analyzed thoroughly. Also for station 0 (NSWE) VTPRs were consistently below 50%.

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 BS weaker westwinds 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 Figure 29. In both cases U10 was implemented with a timelag of $\hat{t} = 3$. Comparing the area around the Danish Straits, i.e. 2°E – 15°E and 54°N – 57°N, shows that in the case of prefilling westwinds are blowing up to 5ms^{-1} stronger in specific areas. Hence, the model predicts (on average) more storm surges with weaker westwind correctly, when passing information of the prefilling to it.

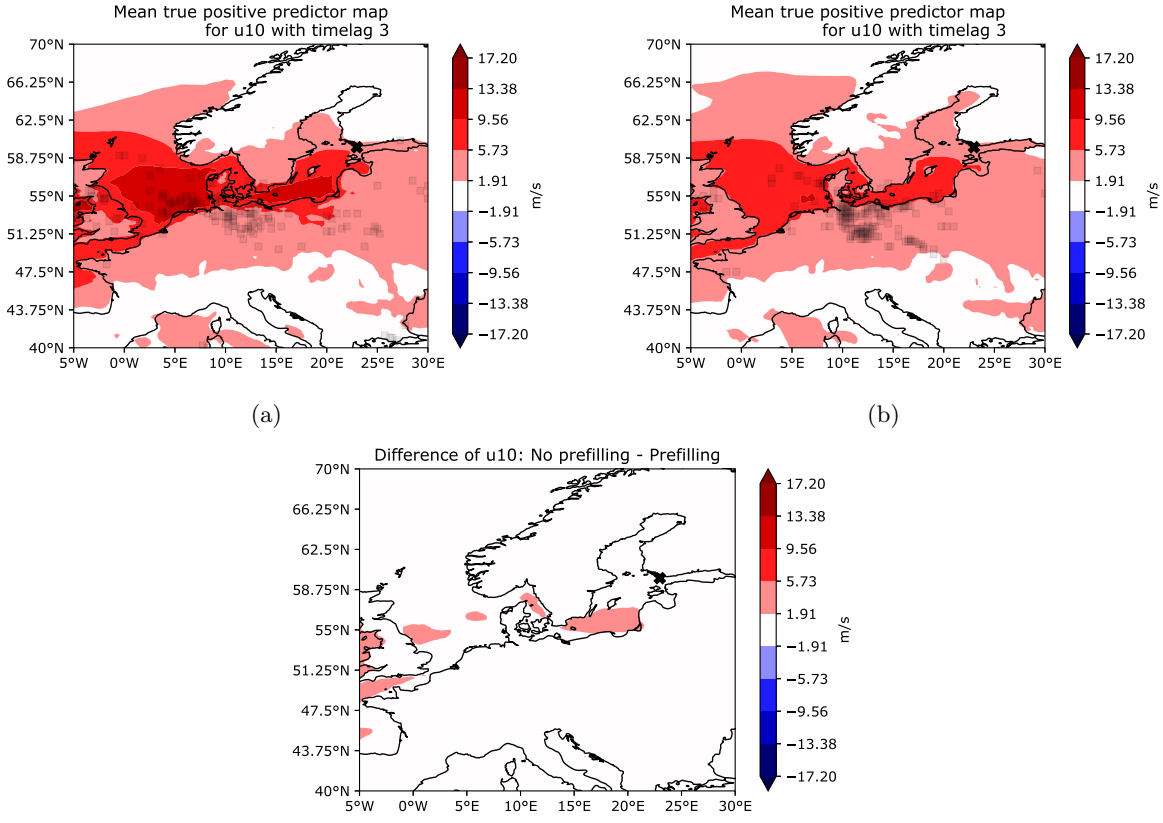


Figure 29: Mean Predictor Maps of U10 at station 1 (FIN) and a timelag of 3; (a) Without information on prefilling, (b) with information on prefilling, (c) difference of (a) and (b).

5.6 F - Combinations of Prefilling-Timelags

Strongly influenced by strong westwind is the prefilling of the BS. While Weisse (2014) and Mudersbach and Jensen (2010) define the prefilling as the rolling mean of the waterlevels at Degerby over 20 consecutive days, we will use a timelag of the records of waterlevel at Degerby as the predictor. In this experiment we investigate PF as an isolated predictor for timelags 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 waterlevels. All combinations can be found in Table 9.

When using isolated predictors, results show that shorter timelags work better in terms of TTPR than longer ones. For instance TTPRs of station 2 (FINBAY) for timelags of 10, 15, 20, and 25 days were 71.12%, 63.63%, 49.85%, 61.61%, respectively.

Combining information of several days lead to 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 timelag-combination of 3, 14, 21 and 30 days. In this case, the feature importance for 3 days was significantly higher than the one for 14 days, which itself was more than the doubled feature importance of $\hat{t} = 21$. This indicates that the model heavily relies on most recent waterlevel recordings in order to provide TPPs. This behavior repeated generally, with greater feature importance for shorter timelags independent of the station at hand. All other stations (except DEU) showed TTPRs above 80% using this very combination of timelags as well. Altogether prefilling seems to be a good predictor for almost all stations when combining information of the previous water records with records up to two weeks. Independent of the station, combining several timelags of information works better than using the information in isolation.

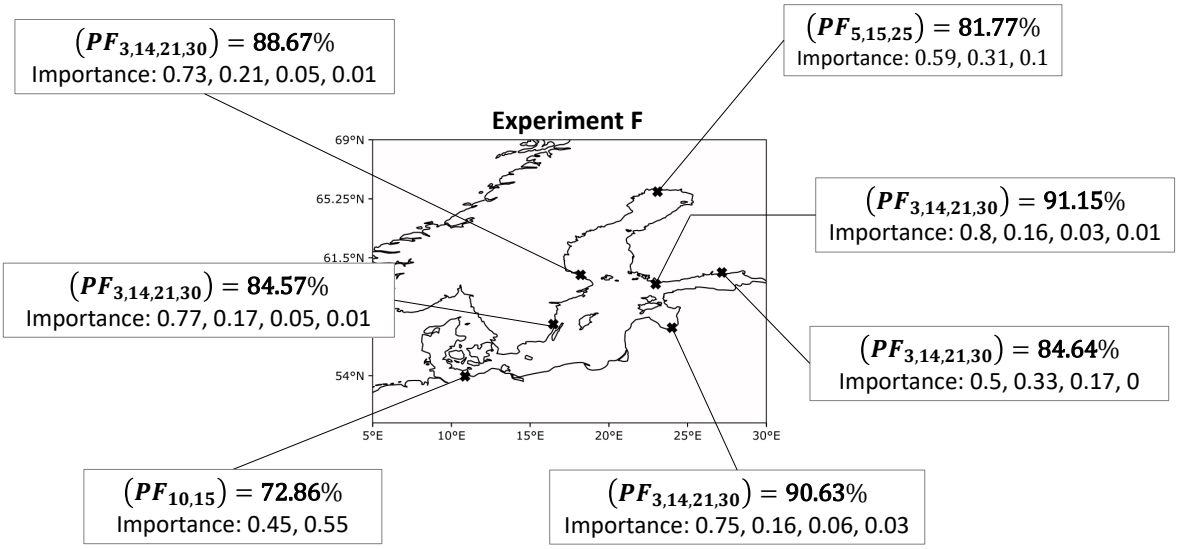


Figure 30: Best combinations in terms of TTPR of PF for experiment F. Importances are ordered as subscripted timelags.

5.7 A Brief Comment on Runtime

We calculated the runtime of the software excluding the computing time for plotting the PMs. The model was run on a conventional personal computer with 4 kernels and 8 processors.
While the software only needs a maximum of three minutes to calculate solutions when using isolated ERA5-predictors, this almost triples when using PF alone (up to 10 minutes of runtime). Using combined ERA5-predictors like in experiment B lead to runtimes up to 10 minutes. The longest calculated runtime was 25 minutes when using a combination of PF as predictors.

6 Discussion

We will now connect theoretical expectations based on Section 2 with the model results from Section 5. The theory indicates that one predictor alone should not be sufficient to describe storm surges. Main features are the wind-stress and the low-pressure systems (below 980 hPa) as well as their speeds. While our models showed also good results when using predictors isolated, they showed more robust results when using them in combination. Furthermore for almost all stations (except station 4) surface pressure and westwind were the most important ERA5-predictors. We could verify that mostly low-pressure fields below 980 hPa and strong (mean) westwinds 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 westwinds often acted on a large horizontal distance, which according to Weisse and von Storch (2010) (see Equation 2) increases the potential of storm surges. It is this very 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. In contrast, those stations should be more subject to negative storm surges, which we did not investigate in this study.
For southern stations rather north-easterly wind should be a predominant factor. We saw this for station DEU, where the most important predictor in terms of feature importance was V10.
According Leppäranta and Myrberg (2009) the largest amount of precipitation is found at the eastern coast of the Baltic Sea due to the winds blowing mostly eastward in winter time. We could not recover this for our model. If any structure could be obtained from AoIs of TP it was 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. In contrast Gönnert et al. (2001) states that 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. Together with the fact that westwinds 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 VTPRs.
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. This behavior showed mainly when using timelags 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 (FNPs) there were storm surges even though high pressure fields were present. We first argued that this might be due to the lack of combining predictors in model runs but the same behavior reappeared when using combinations of predictors. At this point we do not have any valid explanation for this behavior. One idea is that the involved timelag of predictors was long enough, such that high pressure systems could turn into low pressure systems before the actual day of the storm surge. One could account for this when calculating hourly gradients of atmospheric pressure, which indicate a rapid (de-) intensification of low-pressure systems (similar to Bruneau et al. (2020)).

Nevertheless we saw that timelagging the predictors increased model results. This is in alignment with Tyralis et al. (2019), who showed that Random Forests worked better when timelagged predictors were used. In general timelags up to 4 days worked quite reasonable, while longer timelags did not add much value to VTPRs. For instance a timelag 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 2 days to travel accross the Baltic Sea (see Equation 16). Furthermore we saw that implementing a longer timelag for U10 compared to SP leads to good results in terms of VTPRs when using both in combination. For PF mostly short-term timelags work best but still it was possible to even increase the timelag up to a week. This is contradicting the actual definition of prefilling and one might argue against the usage of the plain timeseries 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 month over 9 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 month though. In order to overcome this, one could extent the dataset to longer time periods. Using more data increases computing time though, which is one reason why we did not implement it. Furthermore, models trained with predictors based on remotely sensed data outperformed models forced with predictors obtained from reanalysis data (see Tyralis et al. (2019)). We used only reanalysis data as predictors. If data-sources with remotely sensed data are available it would be better to test the model on them.

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. Furthermore we did not use wind directions per se as a predictors 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 ($\geq$ 16ms$-$1) a subpressure-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. Closely connected to these topics are climate modes like the NAO. In theory, the NAO-index is correlated to prefilling and the strength of

westerlies. It would be interesting to use the NAO-Index or BANOS-Index as a predictors as well. Another physical change that can be made is to look at negative storm surges instead of positive ones and see if behaviors of U10 and SP change for stations like DEU. For instance, the bays of Mecklenburg and Kiel ran into strong negative storm surges (≤ -70cm) 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 one hot encoding, to represent the alarming levels of specific stations instead of using percentiles. It would also be highly interesting to extent 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. Hence, changing the model architecture to more complex ones can may also lead to promising results.

7 Conclusion

In this study we predicted extreme storm surges, i.e. the top 5% water-levels, for 7 stations across the Baltic Sea via machine learning. More specifically we used Random Forests as a binary classifier.

The Baltic Sea is a semi-enclosed basin and hence has special characteristics regarding the onset of storm surges. They are only induced internally. Main drivers of storms surges are the pressure- and wind-effect as well as a condition labeled as prefilling, an increase of the Baltic Seas volume due to water-mass exchange with the North Sea via the Danish Straits. Local characteristics like orientation of the coastline are also important for the risk of storm surges. Cities within the Gulf-regions are subject to more storm surges than cities located in the West of the Baltic Sea. It is important to note, that no isolated driver leads to a storm surge. Rather it is the combination of multiple physical processes through coupling effects that induces surges by non-linear effects. For instance, the intensity of westwinds leading to a storm surge in North-Eastern parts of the Baltic Sea depends on the current volume of the sea itself. With increased volume, weak westwinds can already lead to a surge event at corresponding coastlines.

In order to model storm surges, two model-architectures are distinguished; dynamical models and data-driven models. Within the Baltic Sea mainly dynamical models are in operation. These model rely on the underlying physical principles of the storm surge process. While this leads to accurate results when predicting the water level, it comes with the caveat of intense computing times. Furthermore these models use linear equations which are incapable of capturing non-linear effects efficiently. Data-driven models try to overcome this caveat by learning statistics and patterns off historical datasets containing information about storm surges. Most often, this is faster than solving the system of equations used for dynamical models and eventually capture non-linearities inherent to the dataset. One promising realm of data-driven models is machine learning. Most approaches using machine learning to predict (extreme) storm surges are global and hence lack in specification for the Baltic Sea. For both, data-driven and dynamical models, extreme storm surges are rarely predicted well.

Hence in this study we focused on predicting extreme storm surges within the Baltic Sea. As most storm surges occur seasonally and mainly during winter, we used the month December to February for analysis. We defined extreme storm surges as the top 5% water levels of an investigated station. Records of the water-level at respective stations were obtained from GESLA3, a compilation of global, hourly

water-level records. Based on our definition continuous measures of water-levels were transformed to a binary predictand for extreme storm surges. 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 meter above the Earth's surface and total precipitation (TP). These atmospheric predictors were analyzed spatially on a longitudinal-latitudinal map from 5°W to 30°E and 40°N to 70°N. In addition to those four predictor variables, we used the water-level at Degerby as a proxy for the prefilling (PF) of the Baltic Sea. We choose 7 stations across the whole Baltic Sea to account for various exposures of coastlines. For each station we conducted several experiments, where we altered the combinations of predictors and introduced various timelags.

We analyzed the results in terms of Confusion Matrices and their respective rates. Mainly we focused on the Validation True Positive Rate (VTPR), which measures how often a storm surge was reliably predicted within the validation set. To investigate on which predictors the model relies the most, the feature importance was investigated. Finally, we looked at the actual values of predictors within the area of research in order to observe what physical patterns are responsible for model decisions. We took the mean over all predictor values of True Positive Prediction (TPP)s and False Negative Prediction (FNP)s and looked at their differences to concise results.

In the first experiment we used all predictors isolated from each other while introducing timelags up to a week. For most stations we could achieve VTPRs above 70% in doing so. These rates could be achieved by using PF as a predictor for all stations, while the best atmospheric predictors varied depending on the stations location. We could also observe that in general short timelags worked better than longer ones. In a second experiment we combined all predictors amongst each other including timelags of one and two days. While VTPRs above 60% were obtained for most stations, the experiment revealed an order of importance. Mainly SP and U10 were most important. The only exception was for the station DEU, which is situated in the far South-West of the Baltic Sea. Here, V10 was most important. The next experiment combined the SP and U10, both with various combinations of timelags. While the best VTPRs were above 70%. This experiment showed that using a combination of short term and long term information on both predictors increases the VTPR. It also indicates, that using a shorter timelag for SP compared to U10 is reasonable. We also observed in another experiment that combinations of only U10 can be used for most stations as a good predictor when focusing on timelags up to four days. Furthermore TP is most often not a good predictor, while information on prefilling leads to weaker westwinds when looking at Predictor Map (PM)s of TPPs. Also prefilling seems to be a good predictor for almost all stations when combining information of the previous water records with timelags up to two weeks. Independent of the station, combining several timelags of information works better than using the information in isolation.

From PMs we could observe that mainly low-pressure fields are used by the model for TPPs. We could also see that for many stations, the westwind around the Danish Straits was an important driver. Both observations are in accordance with theory, which suggest that low-pressure fields induce baric waves and lift the water levels while prevailed westwinds lead to either wind-setup at coastline or increase the prefilling of the Baltic Sea. We could also recover the fact that with increased prefilling the westwinds tend to be weaker in cases of storm surges. We also observed a change of predictor importance depending on the orientation of the coastline at which a station is located. For instance for the station in Germany,

SP and U10 were not used as predictors but V10 was. In general no obvious patterns could be deduced from V10 and TP predictor maps.
Some caveats of our model need to be mentioned. We only used a period of 3 month over 9 years to generate training data while 6-7 years of daily training data is suggested by literature. For predictors we used reanalysis data but according to Tyralis et al. (2019) models based on remotely sensed data outperformed models forced with predictors obtained from reanalysis data. We also used the most basic predictors. Due this we neglected for example all possible wind-directions and the duration of wind-fields. For surface pressure we did not include information about its trajectory and velocity which theoretically are also important for the onset of storm surges. Hence there are several possibilities of extending this study. From a physical perspective, one could include more information on the ice-sheet coverage or the salinity of the Baltic Sea. Even the use of climate mode indices like the NAO- or BANOS-Index as further predictors is reasonable. From a technical perspective one could adjust the definition, i.e. the one hot encoding, to represent the alarming levels of specific stations instead of using percentiles. It would also be highly interesting to extent the usage of the Random Forest to Random Regression Forests in order to investigate and predict continuous heights of water level during storm surges.

Altogether we showed that a simple machine learning method like a random forest produces reasonable results when predicting extreme storm surges within the Baltic Sea. These results represent underlying theoretical expectations in many cases. With respect to the very short computing time of these models (up to 10 minutes per station), further investigations of machine learning methods in the context of climate-extreme event predictions are recommended. Extending experiments as previously discussed would be valuable not only for the scientific community but for our society in general.

References

E. Andrée, M. Drews, J. Su, M. A. D. Larsen, N. Drønen, and K. S. Madsen. Simulating wind-driven extreme sea levels: Sensitivity to wind speed and direction. 36:100422, 2022. ISSN 22120947. doi: 10.1016/j.wace.2022.100422. URL https://linkinghub.elsevier.com/retrieve/pii/S2212094722000135.

A. Arns, T. Wahl, S. Dangendorf, and J. Jensen. The impact of sea level rise on storm surge water levels in the northern part of the German Bight. *Coastal Engineering*, 96:118–131, Feb. 2015. ISSN 0378-3839. doi: 10.1016/j.coastaleng.2014.12.002. URL https://www.sciencedirect.com/science/article/pii/S0378383914002191.

C. Baerens, B. B. Braudler, H., H. Birr, H. J. Dick, S., F. Kleine, L. W. Lampe, R., I. Meinke, M. Meyer, M. Mueller, S. Mueller-Navarra, G. Schmager, K. Schwarzer, and T. Zenz. Water levels at the baltic sea coast. trends – storm surges – climate change. 2003.

S. M. Barbosa and R. V. Donner. Long-term changes in the seasonality of baltic sea level. 68(1):30540, 2016. ISSN null. doi: 10.3402/tellusa.v68.30540. URL https://doi.org/10.3402/tellusa.v68.30540. Publisher: Taylor & Francis _eprint: https://doi.org/10.3402/tellusa.v68.30540.

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

A. Bezuglov, B. Blanton, and R. Santiago. Multi-output artificial neural network for storm surge prediction in north carolina, 2016. URL http://arxiv.org/abs/1609.07378. type: article.

I. Bork and S. H. Müller-Navarra. Simulation und analyse extremer sturmhochwasser an der deutschen ostseeküste. page 77, 2009.

N. Bruneau, J. Polton, J. Williams, and J. Holt. Estimation of global coastal sea level extremes using neural networks. *Environmental Research Letters*, 15(7):074030, July 2020. ISSN 1748-9326. doi: 10.1088/1748-9326/ab89d6. URL https://doi.org/10.1088/1748-9326/ab89d6. Publisher: IOP Publishing.

D. C. Chapman and G. S. Giese. Seiches. 2:344–350, 2001.

D. Chen and A. Omstedt. Climate-induced variability of sea level in stockholm: Influence of air temperature and atmospheric circulation. 22(5):655–664, 2005. ISSN 1861-9533. doi: 10.1007/BF02918709. URL https://doi.org/10.1007/BF02918709.

S. Dangendorf, S. Müller-Navarra, J. Jensen, F. Schenk, T. Wahl, and R. Weisse. North sea storminess from a novel storm surge record since AD 1843. 27(10):3582–3595, 2014. ISSN 0894-8755, 1520-0442. doi: 10.1175/JCLI-D-13-00427.1. URL https://journals.ametsoc.org/view/journals/clim/27/10/jcli-d-13-00427.1.xml. Publisher: American Meteorological Society Section: Journal of Climate.

M. Donat, G. Leckebusch, J. Pinto, and U. Ulbrich. European storminess and associated circulation weather types: Future changes deduced from a multi-model ensemble of GCM simulations. 42:27–43, 2010. doi: 10.3354/cr00853.

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

M. Ekman. *The changing level of the Baltic Sea during 300 years: a clue to understanding the earth.* Summer Institute for Historical Geophysics, 2009. ISBN 978-952-92-5241-1.

C. B. Field, V. Barros, T. F. Stocker, and Q. Dahe, editors. *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, 2012. ISBN 978-1-139-17724-5. doi: 10.1017/CBO9781139177245. URL http://ebooks.cambridge.org/ref/id/CBO9781139177245.

J. French, R. Mawdsley, T. Fujiyama, and K. Achuthan. Combining machine learning with computational hydrodynamics for prediction of tidal surge inundation at estuarine ports. 25:28–35, 2017. ISSN 2210-9838. doi: 10.1016/j.piutam.2017.09.005. URL https://www.sciencedirect.com/science/article/pii/S2210983817301682.

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

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

G. Gönnert and K. Sossidi. A new approach to calculate extreme storm surges: analysing the interaction of storm surge components. pages 139–150, Naples, Italy, Apr. 2011. doi: 10.2495/CP110121. URL http://library.witpress.com/viewpaper.asp?pcode=CP11-012-1.

G. Gönnert, S. K. Dube, T. Murty, and W. Siefert. *Die Küste, 63 Global Storm Surges*. Boyens Medien GmbH & Co. KG, Heide i. Holstein, 2001. ISBN 978-3-8042-1054-7.

I. D. Haigh, M. Marcos, S. A. Talke, P. L. Woodworth, J. R. Hunter, B. S. Hague, A. Arns, E. Bradshaw, and P. Thompson. GESLA version 3: A major update to the global higher-frequency sea-level dataset. n/a, 2021. ISSN 2049-6060. doi: 10.1002/gdj3.174. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/gdj3.174. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/gdj3.174.

D. L. Harris. THE EQUIVALENCE BETWEEN CERTAIN STATISTICAL PREDICTION METHODS AND LINEARIZED DYNAMICAL METHODS. 90(8):331–340, 1962. ISSN 1520-0493, 0027-0644. doi: 10.1175/1520-0493(1962)090⟨0331:TEBCSP⟩2.0.CO;2. URL https://journals.ametsoc.org/view/journals/mwre/90/8/1520-0493_1962_090_0331_tebcsp_2_0_co_2.xml. Publisher: American Meteorological Society Section: Monthly Weather Review.

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

H. Hersbach, B. Bell, P. Berrisford, G. Biavati, A. Horányi, J. Muñoz Sabater, J. Nicolas, C. Peubey, R. Radu, I. Rozum, D. Schepers, A. Simmons, C. Soci, D. Dee, and J.-N. Thépaut. Era5 hourly data on single levels from 1959 to present. copernicus climate change service (c3s) climate data store (cds). (accessed on 03-mar-2022). 2018. doi: 10.24381/cds.adbb2d47.

D. J. M. Hofstede. Generalplan Küstenschutz des landes schleswig-holstein fortschreibung 2022. page 110, 2022.

J. Holfort, B. Wisniewski, Z. Lydeikaite, H. Kowalewska-Kalkowska, T. Wolski, H. Boman, T. Hammarklint, A. Giza, and S. Grabbi-Kaiv. Extreme sea levels at selected stations on the Baltic Sea coast. *Oceanologia; 2014; Vol. 56; Iss. 2*, 2014. ISSN 0078-3234. URL https://journals.pan.pl/dlibra/publication/115021/edition/100074. Publisher: Instytut Oceanologii PAN.

B. Hünicke and E. Zorita. Influence of temperature and precipitation on decadal Baltic Sea level variations in the 20th century. *Tellus A: Dynamic Meteorology and Oceanography*, 58(1):141–153, Jan. 2006. ISSN 1600-0870. doi: 10.1111/j.1600-0870.2006.00157.x. URL https://www.tandfonline.com/doi/full/10.1111/j.1600-0870.2006.00157.x.

B. Hünicke, E. Zorita, T. Soomere, K. S. Madsen, M. Johansson, and [U+FFFD] Suursaar. Recent Change—Sea Level and Wind Waves. In The BACC II Author Team, editor, *Second Assessment of Climate Change for the Baltic Sea Basin*, Regional Climate Studies, pages 155–185. Springer International Publishing, Cham, 2015. ISBN 978-3-319-16006-1. doi: 10.1007/978-3-319-16006-1_9. URL https://doi.org/10.1007/978-3-319-16006-1_9.

J. Ihde, W. Augath, and M. Sacher. The vertical reference system for europe. pages 345–350. 2002. ISBN 978-3-642-07701-2. doi: 10.1007/978-3-662-04683-8_64.

IPCC (SMP). Climate Change 2021 - The Physical Science Basis. A Summary for Policymakers. 2021.

J. Jaagus and [U+FFFD] Suursaar. Long-term storminess and sea level variations on the estonian coast of the baltic sea in relation to large-scale atmospheric circulation. 62(2):73, 2013. ISSN 1736-4728. doi: 10.3176/earth.2013.07. URL https://kirj.ee/?id=22308&tpl=1061&c_tpl=1064.

F. Janssen, C. Schrum, U. Hübner, and J. Backhaus. Uncertainty analysis of a decadal simulation with a regional ocean model for the North Sea and Baltic Sea. *Climate Research*, 18:55–62, 2001. ISSN 0936-577X, 1616-1572. doi: 10.3354/cr018055. URL http://www.int-res.com/abstracts/cr/v18/n1-2/p55-62/.

S. Karabil, E. Zorita, and B. Hünicke. Contribution of atmospheric circulation to recent off-shore sea-level variations in the baltic sea and the north sea. 9(1):69–90, 2018. ISSN 2190-4979. doi: 10.5194/esd-9-69-2018. URL https://esd.copernicus.org/articles/9/69/2018/. Publisher: Copernicus GmbH.

B. Koppe. Hochwasserschutzmanagement an der deutschen ostseeküste. page 218, 2002.

M. Leppäranta and K. Myrberg. *Physical oceanography of the Baltic Sea*. Springer Praxis books geophysical sciences. Springer, Berlin Heidelberg, 2009. ISBN 978-3-540-79702-9.

L. Magaard and G. Rheinheimer. *Meereskunde der Ostsee*. Number vol. 1. Springer Verlag, Berlin, 1974.

I. Meinke. Sturmfluten in der südwestlichen ostsee – dargestellt am beispiel des pegels Warnemünde. 1999.

V. Mohrholz. Major baltic inflow statistics – revised. 5, 2018. ISSN 2296-7745. URL https://www.frontiersin.org/article/10.3389/fmars.2018.00384.

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

S. Muis, M. Verlaan, H. C. Winsemius, J. C. J. H. Aerts, and P. J. Ward. A global reanalysis of storm surges and extreme sea levels. *Nature Communications*, 7(1):11969, Sept. 2016. ISSN 2041-1723. doi: 10.1038/ncomms11969. URL http://www.nature.com/articles/ncomms11969.

A. C. Müller. *Introduction to Machine Learning with Python*. 2017.

A. Niros, T. Vihma, and J. Launiainen. Marine meteorological conditions and air-sea exchange processes over the northern baltic sea in 1990s. 38, 2002.

A. Omstedt and D. Chen. Influence of atmospheric circulation on the maximum ice extent in the baltic sea. 106:4493–4500, 2001. ISSN 01480227. doi: 10.1029/1999JC000173. URL http://doi.wiley.com/10.1029/1999JC000173.

S. Rikka. ESTIMATION OF WAVE FIELD PARAMETERS FROM SAR IMAGERY IN THE BALTIC SEA. page 46, 2014.

A. Rutgersson, E. Kjellström, J. Haapala, M. Stendel, I. Danilovich, M. Drews, K. Jylhä, P. Kujala, X. Guo Larsén, K. Halsnæs, I. Lehtonen, A. Luomaranta, E. Nilsson, T. Olsson, J. Särkkä, L. Tuomi, and N. Wasmund. Natural Hazards and Extreme Events in the Baltic Sea region. *Earth System Dynamics Discussions*, pages 1–80, Apr. 2021. ISSN 2190-4979. doi: 10.5194/esd-2021-13. URL https://esd.copernicus.org/preprints/esd-2021-13/. Publisher: Copernicus GmbH.

M. Siek and D. P. Solomatine. Nonlinear chaotic model for predicting storm surges. *Nonlinear Processes in Geophysics*, 17(5):405–420, Sept. 2010. ISSN 1023-5809. doi: 10.5194/npg-17-405-2010. URL https://npg.copernicus.org/articles/17/405/2010/. Publisher: Copernicus GmbH.

T. Soomere, A. Behrens, L. Tuomi, and J. Nielsen. Wave conditions in the baltic proper and in the gulf of finland during windstorm gudrun. 8, 2008. doi: 10.5194/nhess-8-37-2008.

[U+FFFD] Suursaar, J. Jaagus, and T. Kullas. Past and future changes in sea level near the estonian coast in relation to changes in wind climate. 11:123–142, 2006a.

[U+FFFD] Suursaar, T. Kullas, M. Otsmann, I. Saaremäe, J. Kuik, and M. Merilain. Cyclone gudrun in january 2005 and modelling its hydrodynamic consequences in the estonian coastal waters. 11:17, 2006b.

M. Sztobryn. Forecast of storm surge by means of artificial neural network. *Journal of Sea Research*, 49(4):317–322, June 2003. ISSN 13851101. doi: 10.1016/S1385-1101(03)00024-8. URL https://linkinghub.elsevier.com/retrieve/pii/S1385110103000248.

M. Tadesse, T. Wahl, and A. Cid. Data-Driven Modeling of Global Storm Surges. *Frontiers in Marine Science*, 7:260, 2020. ISSN 2296-7745. doi: 10.3389/fmars.2020.00260. URL https://www.frontiersin.org/article/10.3389/fmars.2020.00260.

T. Tiggeloven, A. Couasnon, C. van Straaten, S. Muis, and P. J. Ward. Exploring deep learning capabilities for surge predictions in coastal areas. *Scientific Reports*, 11(1):17224, Dec. 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-96674-0. URL https://www.nature.com/articles/s41598-021-96674-0.

H. Tyralis, G. Papacharalampous, and A. Langousis. A brief review of random forests for water scientists and practitioners and their recent history in water resources. 11(5):910, 2019. ISSN 2073-4441. doi: 10.3390/w11050910. URL https://www.mdpi.com/2073-4441/11/5/910. Number: 5 Publisher: Multidisciplinary Digital Publishing Institute.

T. Vihma and J. Haapala. Geophysics of sea ice in the baltic sea: A review. 3-4(80):129–148, 2009. ISSN 0079-6611. doi: 10.1016/j.pocean.2009.02.002. URL https://www.infona.pl//resource/bwmeta1.element.elsevier-78c303a5-5aa4-31eb-86d3-bbfef7e6ba96.

H. von Storch. Storm Surges: Phenomena, Forecasting and Scenarios of Change. *Procedia IUTAM*, 10:356–362, Jan. 2014. ISSN 2210-9838. doi: 10.1016/j.piutam.2014.01.030. URL https://www.sciencedirect.com/science/article/pii/S2210983814000315.

H. von Storch and K. Woth. Storm surges: perspectives and options. *Sustainability Science*, 3(1):33–43, Apr. 2008. ISSN 1862-4057. doi: 10.1007/s11625-008-0044-2. URL https://doi.org/10.1007/s11625-008-0044-2.

M. I. Vousdoukas, L. Mentaschi, E. Voukouvalas, M. Verlaan, and L. Feyen. Extreme sea levels on the rise along Europe's coasts. *Earth's Future*, 5(3):304–323, 2017. ISSN 2328-4277. doi: 10.1002/2016EF000505. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/2016EF000505. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/2016EF000505.

T. Wahl and D. P. Chambers. Climate controls multidecadal variability in u. s. extreme sea level records. 121(2):1274–1290, 2016. ISSN 2169-9291. doi: 10.1002/2015JC011057. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/2015JC011057. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/2015JC011057.

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

R. Weisse and B. Hünicke. Baltic Sea Level: Past, Present, and Future. In *Oxford Research Encyclopedia of Climate Science*. Oxford University Press, Apr. 2019. ISBN 978-0-19-022862-0. doi: 10.1093/acrefore/9780190228620.013.693. URL https://oxfordre.com/climatescience/view/10.1093/acrefore/9780190228620.001.0001/acrefore-9780190228620-e-693.

R. Weisse and H. von Storch. *Marine Climate and Climate Change*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. ISBN 978-3-540-25316-7 978-3-540-68491-6. doi: 10.1007/978-3-540-68491-6. URL http://link.springer.com/10.1007/978-3-540-68491-6.

R. Weisse and H. Weidemann. Baltic Sea extreme sea levels 1948-2011: Contributions from atmospheric forcing. *Procedia IUTAM*, 25:65–69, Jan. 2017. ISSN 2210-9838. doi: 10.1016/j.piutam.2017.09.010. URL https://www.sciencedirect.com/science/article/pii/S221098381730175X.

R. Weisse, I. Dailidiene, B. Hünicke, K. Kahma, K. Madsen, A. Omstedt, K. Parnell, T. Schöne, T. Soomere, W. Zhang, and E. Zorita. Sea level dynamics and coastal erosion in the Baltic Sea region. *Earth System Dynamics*, 12(3):871–898, Aug. 2021. ISSN 2190-4987. doi: 10.5194/esd-12-871-2021. URL https://esd.copernicus.org/articles/12/871/2021/.

B. Wisniewski and T. Wolski. Physical aspects of extreme storm surges and falls on the Polish coast. *Oceanologia*, 53((1-TI)), 2011. ISSN 0078-3234. URL http://yadda.icm.edu.pl/yadda/element/bwmeta1.element.dl-catalog-470a75c2-d847-417b-91a4-10e0bbfcb266. Publisher: -.

WMO. *Guide to storm surge forecasting*. WMO, Geneva, 2011. ISBN 978-92-63-11076-3. OCLC: 1075529493.

T. Wolski and B. Wisniewski. Geographical diversity in the occurrence of extreme sea levels on the coasts of the Baltic Sea. *Journal of Sea Research*, 159:101890, Apr. 2020. ISSN 1385-1101. doi: 10.1016/j.seares.2020.101890. URL https://www.sciencedirect.com/science/article/pii/S1385110120300903.

T. Wolski and B. Wisniewski. Characteristics and Long-Term Variability of Occurrences of Storm Surges in the Baltic Sea. *Atmosphere*, 12(12):1679, Dec. 2021. doi: 10.3390/atmos12121679. URL https://www.mdpi.com/2073-4433/12/12/1679. Number: 12 Publisher: Multidisciplinary Digital Publishing Institute.

T. Wolski, B. Wisniewski, and S. Musielak. Baltic Sea datums and their unification as a basis for coastal and seabed studies. *Oceanological and Hydrobiological Studies*, 45, June 2016. doi: 10.1515/ohs-2016-0022.

P. Woodworth and M. Marcos. Changes in extreme sea levels. CLIVAR(16), Feb. 2018.

P. L. Woodworth, J. R. Hunter, M. Marcos, P. Caldwell, M. Menéndez, and I. Haigh. Towards a global higher-frequency sea level dataset. 3(2):50–59, 2016. ISSN 2049-6060. doi: 10.1002/gdj3.42. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/gdj3.42`. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/gdj3.42.

Acronyms

ANN Artificial Neural Network. 23–25, 30

AoI Area of Importance. 35–39, 42–47, 50, 52, 54

BANOS Baltic Sea and North Sea Oscillation. 17, 56, 58

BS Baltic Sea. 26, 27, 33, 35, 36, 39, 42, 44, 52, 53

BSH Bundesamt für Schifffahrt und Hydrographie. 21

CDO Climate Data Operators. 29

CFM Confusion Matrix. 33, 34

CNN Convolutional Neural Network. 25

ConvLSTM Convolutional Long Short-Term Memory. 25

CRPS Continous Ranked Probability Score. 24

DT Decision Tree. 30–32, 67

ECMWF European Centre for Medium-Range Weather Forecasts. 27, 66

ERA5 European Re-Analysis. 25, 27–29, 31, 33, 35–37, 39, 43, 54, 57, 66

EVRS European Vertical Reference System. 6

FI Feature Importance. 34

FNP False Negative Prediction. 34–42, 44–48, 50, 55, 57

FNR False Negative Rate. 33

GESLA Global Extreme Sea Level Analysis. 3, 24–29, 31, 33, 56

GTSR Global Tide and Surge Reanalysis. 24, 25

HP hyperparameter. 32, 33

LSTM Long Short-Term Memory. 24, 25

ML Machine Learning. 3, 25

MSL Mean Sea Level. 4, 6, 7

MSLP Mean Sea Level Pressure. 25

NaN Not-a-Number. 30

NAO North-Atlantic-Oscillation. 10, 16, 18, 19, 55, 56, 58

PCA Principal Component Analysis. 24

PF prefilling. 29–31, 33, 37, 39, 40, 42, 46, 52–55, 57

PM Predictor Map. 34–36, 38, 39, 42, 43, 46, 47, 50, 52, 54, 57

RF Random Forest. 24, 26, 30–32, 56, 67

RFC RandomForestClassifier. 32

RMSE Root Mean Squared Error. 23

RRF Random Regression Forest. 30

SP surface pressure. 2, 28–30, 33, 35–39, 41–43, 45–48, 54–58

SVM Support Vector Machine. 30

TP total precipitation. 28, 33, 37–39, 42, 52, 54, 57, 58

TPP True Positive Prediction. 34–48, 50, 52–54, 57

TPR True Positive Rate. 33–35, 39, 40

TTPR Test True Positive Rate. 35–37, 40, 42, 43, 45, 46, 49, 51–53

U10 eastward wind at 10m height. 2, 28, 30, 33, 36–50, 52, 54–58

V10 northward wind at 10m height. 28, 35–37, 42, 43, 45, 54, 55, 57, 58

VTPR Validation True Positive Rate. 35–37, 39, 42, 43, 45–52, 54, 55, 57

Appendices

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 1: Number of station as in Fig 8 and corresponding code in GESLA dataset.

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

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

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

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 5: Parameters and timelags used for experiment **B**. Parenthesis indicate that predictors are used in combination.

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)
12	(SP, SP, U10, U10)	(1, 5, 1, 7)

Table 6: Parameters and timelags used for experiment **C**. Parenthesis indicate that predictors are used in combination.

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 7: Parameters and timelags used for experiment **D**. Parenthesis indicate that predictors are used in combination.

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 8: Parameters and timelags used for experiment **E**. Parenthesis indicate that predictors are used in combination.

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 9: Parameters and timelags used for experiment **F**. Parenthesis indicate that predictors are used in combination.

B Equations

The vertically averaged and integrated Navier Stokes equations are based on two assumptions, namely, a small surge amplitude and large horizontal scale of the surge compared to the water depth (Gönnert et al. (2001)). The horizontal velocities u and v are then vertically averaged over the whole depth of the basin giving the horizontal transport components of x and y as

$$M \equiv \int_{z=-D}^{\zeta} u dz \text{ and } N \equiv \int_{z=-D}^{\zeta} v dz. \tag{12}$$

Together with the linearization of the Navier-Stokes equations, this leads to the traditional (linear) storm surge equations

$$\frac{\partial M}{\partial t} - fN = -gD\frac{\partial \zeta}{\partial x} - \frac{D}{\rho_0}\frac{\partial P_a}{\partial x} + \frac{1}{\rho_0}\left(\tau_{S_x} - \tau_{B_x}\right), \tag{13}$$

$$\frac{\partial N}{\partial t} + fM = -gD\frac{\partial \zeta}{\partial y} - \frac{D}{\rho_0}\frac{\partial P_a}{\partial y} + \frac{1}{\rho_0}\left(\tau_{S_y} - \tau_{B_y}\right), \tag{14}$$

$$\frac{\partial \zeta}{\partial t} + \frac{\partial M}{\partial x} + \frac{\partial N}{\partial y} = 0, \tag{15}$$

where

$$\begin{aligned} f &= \text{the Coriolis parameter,} \\ g &= \text{gravity,} \\ D &= \text{bottom of basin,} \\ \rho_0 &= \text{uniform density of water,} \\ P_a &= \text{atmospheric pressure above sea level,} \\ \tau_{S_x}, \tau_{S_y} &= \text{wind stress parameter,} \\ \tau_{B_x}, \tau_{B_y} &= \text{bottom friction parameter.} \end{aligned}$$

We calculated the travel time of deep water waves across the Baltic Sea as follows. Mainly deep water waves move at a speed of

$$C^2 = \frac{gL}{2\pi}, \tag{16}$$

with gravity acceleration g and wavelength L. According to Rikka (2014) wavelengths in the Baltic Sea are estimated between 20m - 70m. Taking $L = 50$m leads to a speed of approximately 32 kmh^{-1}. The maximum Length of Baltic Sea is almost 1600 km. The wave would need around 50 hours, i.e. more than 2 days, to travel this distance.

C Listings

```
#---
# Loading and preprocessing
#---
Y_tilde = Load: predictand dataset (GESLA);
Y = Preprocess: Y_tilde;
X = Load: CDO-preprocessed predictor dataset (ERA5 or PF);
X, Y = Intersect dates: X, Y;
X, Y = Timelag: X, Y;
X, Y = Convert shape according to fit model: X, Y;

#---
# Building the model
#---
X_train, X_test, Y_train, Y_test = train-test split: X, Y;
X_train, X_test, Y_train, Y_test = Scale: X_train, X_test, Y_train, Y_test;
hparams = Optimize hyperparameters: X_train, Y_train;
model = Build Random Forest: hparams;
model = Fit: model(X_train, Y_train);

#---
# Evaluate results
#---
def evaluation(model, predictor, predictand):
    Plot predictor on research area;
    Calculate & visualize: feature importance;
    Calculate & visualize: confusion matrix;
    Calculate & visualize: AUROC;

evaluate: evaluation(model, X_test, Y_test);
evaluate: evaluation(model, X_train, Y_train);

if validation_data_exists:
    X_validation, Y_validation = Load data & preprocess as above;
    evaluate: evaluation(model, X_validation, Y_validation);
```

Listing 1: Pseudocode of Software that is applied to each station for all chosen predictors.

D Versicherung an Eides statt

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit im Studiengang *Ocean and Climate Physics* selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel – insbesondere keine im Quellenverzeichnis nicht benannten Internet-Quellen – benutzt habe. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen entnommen wurden, sind als solche kenntlich gemacht. Ich versichere weiterhin, dass ich die Arbeit vorher nicht in einem anderen Prüfungsverfahren eingereicht habe und die eingereichte schriftliche Fassung der auf dem elektronischen Speichermedium entspricht.
Einer Veröffentlichung der vorliegenden Arbeit in der zuständigen Fachbibliothek des Fachbereichs stimme ich zu.

D Versicherung an Eides statt

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit im Studiengang *Ocean and Climate Physics* selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel – insbesondere keine im Quellenverzeichnis nicht benannten Internet-Quellen – benutzt habe. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen entnommen wurden, sind als solche kenntlich gemacht. Ich versichere weiterhin, dass ich die Arbeit vorher nicht in einem anderen Prüfungsverfahren eingereicht habe und die eingereichte schriftliche Fassung der auf dem elektronischen Speichermedium entspricht.
Einer Veröffentlichung der vorliegenden Arbeit in der zuständigen Fachbibliothek des Fachbereichs stimme ich zu.

Hamburg 31.10.22

Kai Bellinghausen