Primary Drivers of Sea Level Variability in the North – Baltic Sea Transition Using Machine Learning

David Ek

Degree of Master of Science (120 credits)
with a major in Earth Sciences
60 hec

Department of Earth Sciences
University of Gothenburg
2022 B1199

Faculty of Science

UNIVERSITY OF GOTHENBURG

UNIVERSITY OF GOTHENBURG
Department of Earth Sciences
Geovetarcentrum/Earth Science Centre

Primary Drivers of Sea Level Variability in the North – Baltic Sea Transition Using Machine Learning

David Ek

ISSN 1400-3821

B1199
Master of Science (120 credits) thesis
Göteborg 2022

Mailing address
Geovetarcentrum
S 405 30 Göteborg

Address
Geovetarcentrum
Guldhedsgatan 5A

Telephone
031-786 19 56

Geovetarcentrum
Göteborg University
S-405 30 Göteborg
SWEDEN

ABSTRACT

Global mean sea level is rising, however not uniformly. Regional deviations of sea surface height (SSH) are common due to local drivers, including surface winds, ocean density stratifications, vertical land- & crustal movements and more. The contribution of each background driver needs to be better understood to create reliable sea level rise projections, enable effective local policymaking and aid in urban planning decisions.

In this study, we assess region-specific historic sea levels along the western Swedish coastline (Kattegat, Skagerrak & South Baltic Sea). We use monthly satellite altimetry observations spanning 26 years and daily observations spanning 6 years, as well as *in situ* tide gauge measurements to identify SSH covariance between sub-regions. We employed a number of manual statistical methods and found that the North – Baltic Sea transition can be effectively split up into four separate subbasins of sea level covariance. We found that SSH variability in the Skagerrak and Kattegat Seas is different from that of the Belts and south of the Danish Straits. While typically the correlation between SSH time series from different locations declines with distance, this is not seen at the entrance to the Baltic Sea due to the complexity of the region. To investigate this further and identify underlying primary forcings, we quantified the correlation between climatic drivers derived from the ERA5 reanalysis such as 10m-winds, sea surface temperature and sea level pressure, and principle components of the SSH variability signal within these regions. Zonal winds are most important for determining short-term sea level variability throughout the study area. As freshwater discharge from rivers and tributaries to the Baltic Sea is large, pressure- & density gradients may be more important as SSH regulators in this area.

Additionally, we used neural networks to try to capture non-linear dependencies between the sea level drivers and sea level that are not apparent from statistical analyses. By predicting sea level at selected locations from different combination of drivers, we can determine which drivers have the highest influence. Since it is important to capture long-term dependencies between variables, we employed a recurrent neural network with a long short-term memory architecture and found that it is possible to predict daily sea level variability within a few cm of error with only a handful of background drivers. We found that excluding the zonal wind component was the most detrimental for model accuracy, which agrees with the statistical analysis.

Table of Contents

1 INTRODUCTION

Global sea level rise (SLR) is an area of intense study in the global scientific community (Church et al., 2013). It is both a good indicator of climate change and also a large cause of concern for the impact it may have on coastal ecosystems as well as on human societies and settlements (Oppenheimer et al., 2019). Not only is the mean sea level rising, the rate of SLR is accelerating. Indications from tide gauge observations show that the global mean sea level (GMSL; the spatial average global height of the sea surface) rose by 1.7 (±0.2) mm yr^{-1} between 1901 and 2010 (Church et al., 2013). Since the start of the satellite altimetry data record 1993-2018, the GMSL rose at an accelerated rate of 3.1 (±0.3) mm yr^{-1} (WCRP, 2018). Projections presented in the IPCC Special Report on the Ocean and Cryosphere in a Changing Climate (SROCC) estimate GMSL's 0.43 to 0.84 m higher by 2100 relative to 1986-2005 under the "best" emission scenario RCP 2.6 and worst-case scenario 8.5 respectively (Oppenheimer et al., 2019).

Most global SLR is attributed to two dominant factors, the thermal expansion of the ocean and the influx of freshwater from ice sheets and glaciers. Thermal expansion is caused by rising ocean temperatures. Generally, higher temperatures lead to lower densities, hence, as the sea water gets warmer, it expands and occupies more space. Thermal expansion accounted for 1.32 mm yr^{-1} of SLR recorded between 1993-2015 (WCRP, 2018). Higher global temperatures in both the oceans and the atmosphere cause the net decrease of ice mass in glaciers worldwide. The Greenland and Antarctic ice sheets were together responsible for 0.75 mm yr^{-1} of SLR while other glaciers contributed 0.56 mm yr^{-1} during the same period (Church et al., 2013). Global SLR is not spatially equal or uniform however, and local or regional deviations from the global mean are common (Oppenheimer et al., 2019; Slangen et al., 2014a). Slangen et al. (2014a) found regional coastal sea levels to range from 30% above to 50% below the global mean. Local controlling drivers include surface winds, air pressure systems, ocean density stratifications, changes in the Earth's gravity field as well as basin-wide deformation, vertical land- & crustal movements, and more (Cazenave & Llovel, 2010; Church et al., 2013; Mitrovica et al., 2018; Stammer et al., 2013; Woodworth et al., 2019). The sea surface height (SSH) at any given moment can be considered the superimposition of these various drivers. However, the contribution of each background driver on local sea levels needs to be better understood to enable effective local policymaking and urban planning

decisions. As regional rates of SLR have significantly deviated from the global mean in the past, it should be expected that future regional sea levels will vary as well. Increased sea levels, coupled with storm surges and tidal effects are expected to lead to more severe flooding events (H.-O. Pörtner et al., 2022). This creates a need to assess region-specific historic sea levels and its drivers accurately, in order to create reliable SLR projections for local governments and decision-making bodies.

1.1 AIM

This project is part of a larger, FORMAS funded, multi-year project called NEEDS. The aim of NEEDS is to determine the complete dynamics of sea level and provide sea level and flood projections for the next 30 years over Northern Europe using machine learning techniques to help determine the main drivers of sea level variability (SLV). Ultimately, the aim of the project is to verify if the proposed northern European enclosure dam would be a relevant option to protect Scandinavian coastlines.

This Master's project in particular aims to complete parts of the first of three objectives of NEEDS, which is to identify spatial coherence in the wider Northern European seas and identify/map regions that covary on daily to decadal timescales. The focus in this project is on high frequency sea level variability at areas along Swedish coastlines, mainly along the North Sea – Baltic Sea transition. I also refine these maps with historical tide-gauge measurements. After identifying regions that covary, I explain how certain subregions covary, using both conventional statistical methods as well as machine learning techniques through Recurrent Neural Networks (RNN). The results from this Master's project will be directly compared to those produced for NEEDS using other machine learning methods, and allow the project to move to the next phase.

I aim to compare the results from a classic statistical approach to analyze sea level variability and forcing components, with results obtained from a machine learning approach. In my case, I intend to predict sea levels based on different combinations of background forcing variables.

1.2 STUDY AREA

The study area encompasses the wider northern European seas, including the North Sea, Baltic Sea, Norwegian Sea and North-East Atlantic Ocean. After preliminary work (not shown) and out of personal interest I choose to focus primarily on a limited area

consisting of the Skagerrak Sea, Kattegat Sea, Danish Straits, and the south-westernmost areas of the Baltic Sea, with effort dedicated to understanding these region's primary SSH influencing drivers.

Skagerrak is the deepest of the three basins, encompassing the Norwegian trench (>700 m depth – **Figure 1**). The salty Jutland Current (black arrows on **Figure 1**) enters the basin from the west, bringing water from the North Sea (Christensen et al., 2018). Once past the tip of Denmark, it converges with less salty water originating from the Baltic (red arrows), turns north and later follows the Norwegian coast west eventually forming the Norwegian Coastal Current (yellow arrows) (Christensen et al., 2018). The Kattegat is in comparison shallow (average 25 m) and connects the Baltic Sea to the rest of the open ocean through the Danish Straits (namely, from west to east, the Little Belt, Big Belt and Oresund i.e., the Belt Sea). Inside the Baltic Sea itself there are no permanent currents. However, to compensate for the large freshwater inflow and precipitation events into the Baltic there is usually an outflow through the Belts. The Baltic Outflow Current flows through these passages, before following the Swedish coast north, undergoes mixing, and eventually joining the Norwegian Coastal Current (SMHI, 2014b). Like most currents, they can be highly variable and certainly wind dependent (Christensen et al., 2018; Hordoir et al., 2013).

There are however other forcings acting on the currents that we must also consider, for instance pressure and density gradients (Gustafsson & Andersson, 2001; Hordoir et al., 2013). Large river inputs into the Baltic creates a large contrast in salinity content between the basins, while saltwater inflow is provided by the North Sea through the Skagerrak (Lass & Mohrholz, 2003). As the Baltic Sea drains through the Danish Straits, the surface layer of water is brackish with an inflow of salty water travelling in the opposite direction beneath the top layer (Lass & Mohrholz, 2003). The outflow from the Baltic is driven by a barotropic pressure gradient due to the differences in sea level between the Kattegat and Baltic seas. Sea level difference is maintained mainly by winds that cause a pileup of water at the coastlines but also due to the many rivers and tributaries draining into the Baltic Sea (Lass & Mohrholz, 2003).

In conclusion, the sub-basins are quite different, with the deeper northern parts of the study area open to the wider North Sea, resulting in saltier water, larger tidal variability and increased vulnerability to Atlantic storm surges. The Baltic Sea on the other hand is

semi-enclosed with fresher water, where short frequency SLV signals do not propagate efficiently through the narrow straits, leading to small tidal amplitudes (Hieronymus et al., 2017). Thus, there exists differences between these basins in terms of sea level variability.

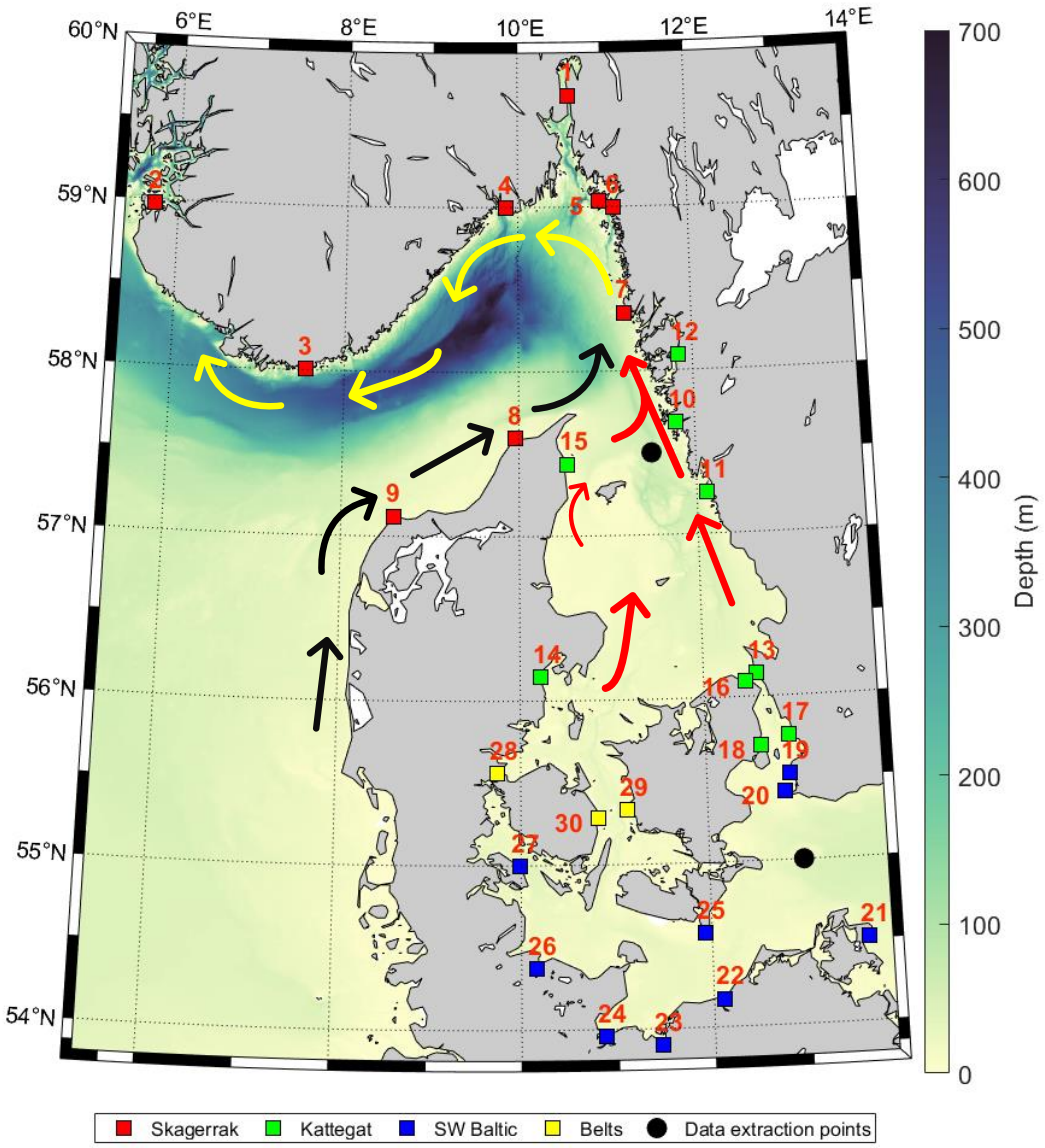


Figure 1: Bathymetry over study area. The colored squares represent the tide gauge locations, the color indicates which basin they are located within and the number label their station ID. The arrows indicate the general flow of surface currents, the Baltic Outflow Current in red, the Jutland Current in black and the Norwegian Coastal Current in yellow. The two black points show the location of the areas analyzed using neural networks, the northernmost point being the Kattegat location and the other being the SW Baltic location.

1.3 SEA LEVEL DRIVERS

While much of global SLR is attributed to the addition of freshwater from ice sheets and glaciers, when considering regional sea surface variability (SSV; sea surface height variations over time), the redistribution of existing water mass becomes more significant, especially in shallow shelf seas and at high latitudes (>60°N) (Meyssignac et al., 2017; Oppenheimer et al., 2019). For instance, as wind-driven currents shift, sea levels may rise in one location and fall at another (Woodworth et al., 2019). Additionally, melt water released from ice sheets does not cause a uniform rise of global sea levels. Rather, local changes to SLR caused by ice sheets yield distinct patterns to regional SLR. Changes to the mass distribution of the ice sheets cause the gravitational attraction between the ocean and the ice sheet to decrease. Combined with the effect of glacial isostatic adjustment (GIA; the response of solid Earth to ice mass loads), relative sea levels surrounding the ice sheet will fall, while areas far from the melting ice sheet experience enhanced SLR compared to the global mean (Mitrovica et al., 2018). In Northern European seas, SLR is expected to be mostly dominated by ice mass loss from the Antarctic ice sheet, while ice mass loss from the Greenland ice sheet may induce sea level fall. These changes are often referred to as sea level fingerprints (Mitrovica et al., 2018). Over time however, long-term trends of SLR accumulate and are expected to dominate over the 21st century (Church et al., 2013).

In situ and satellite observations have shown regional SSH trend variability on decadal to interannual timescales (Cazenave & Llovel, 2010; Church et al., 2013), which is also the case for northern European Seas. In the North Sea basin rates of SLR have ranged from 1.3 to 3.9 mm yr^{-1} between 1993-2014, with higher rates found off the Danish-German coast and at isolated regions surrounding NE Great Britain (Sterlini et al., 2017). Low levels of SLR were found midway between NE Scotland and SW Norway (Sterlini et al., 2017).

Regional deviations of SSH are caused by a combination of various forcings such as changes in ocean dynamics, the atmospheric circulation and in the Earth's gravity field, as well as basin-wide deformation and vertical land- & crustal movements (Stammer et al., 2013). SSH controlling drivers at any location can be local, remote, dynamic and/or static in nature, and operate on a great varying degree of spatial and temporal scales. Some examples include storm surges, which would be considered local and short-lived,

while changes in atmospheric modes of variability can be considered to be remote and long-lived. Slangen et al. (2014b) explained the contributing processes associated to regional sea level change across a number of regions including the North Sea. Their results indicated that steric/dynamic changes and GIA are the largest factors on multidecadal timescales. Dangendorf et al. (2014a) conducted an extensive study where they examined sea level variations driven by different forcing factors across a range of timescales in the North Sea and found that subannual variability is largely controlled by meteorological forcings such as winds and surface air pressure, and that the relative importance of the background forcings varies throughout the region.

SSV is a dynamical system, meaning that it is inherently chaotic and difficult to predict. SSV in coastal areas proves to be even more challenging to explain, since these areas inherently possess shallow waters, complex coastlines, and river runoff, as well as struggle with difficulties associated with satellite altimetry products (Woodworth et al., 2019).

Nonuniform thermosteric expansion originating from uneven ocean warming has been found to be largely responsible for observed spatial trend patterns in regional sea levels (Cazenave & Llovel, 2010). Thermosteric expansion is larger out on the open ocean where the water column is deep. It does however effect coastal areas due to dynamic equilibrium seeking to be maintained in order to obey mass conservation laws, resulting in water mass moving from the open ocean towards the coast (Stammer et al., 2013). The second part of steric expansion in the ocean consists of halosteric, i.e., salinity changes. While globally much smaller than the thermosteric component (Meyssignac et al., 2017) they can still be large locally and should still be considered (Llovel & Lee, 2015), especially in the North-Baltic Sea transition area where large salinity contrasts are present.

Previous studies have shown the importance of surface winds in influencing water mass transport and SSV in the North – Baltic Sea transition zone (e.g. Hieronymus et al., 2017; Hordoir et al., 2013; Passaro et al., 2015). For instance, Baltic outflow of freshwater has been found to be highly restricted by wind driven SSV in the Kattegat (Hieronymus et al., 2017; Hordoir & Meier, 2010). The predominant wind pattern are the westerlies, i.e. the prevailing western winds in the middle latitudes determined by the general atmospheric circulation, usually strongest during the winter months (Passaro et al., 2015). The mean

transport of wind-driven surface currents in the northern hemisphere is offset clockwise to the wind direction by 45°, as a result of the balance between the Coriolis force and drag created between underlying water layers. This is known as Ekman transport. This means that strong westerlies typically cause higher than usual SSH as they drag water from the North Sea into the Skagerrak and Kattegat Seas (Passaro et al., 2015). The coastal boundary blocks the wind-driven water transport leading to sea level elevation (Woodworth et al., 2019). This in turn causes a reduced slope of the ocean surface between the Kattegat and the Baltic that decreases the pressure gradient and consequently water flow (Gustafsson & Andersson, 2001; Hordoir et al., 2013). During spring and summer, when the westerlies lose strength, the pressure gradient fails to sustain itself, which drives seasonal freshwater pulses (Gustafsson & Andersson, 2001; Hordoir et al., 2013). Passaro et al. (2015) have previously found that all sea level maxima that occurred over an 8-year period in these regions coincided with strong westerlies, and that the lowest levels of SLA occurred during easterlies.

During calm wind conditions, the in- and outflow between the Kattegat and the Baltic Sea is bi-directional and is separated by a steep halocline, with salty water flowing into the Baltic underneath a much fresher top layer flowing out of it (Sayin & Krauss, 1996). Besides the barotropic pressure gradient that exists due to sea level differences, this flow is also density-driven by the large contrast in bottom and surface salinity between the basins (from 35 to 8 PSU). During strong winds, the flow between the basins is rather unidirectional in either direction across the entire water column (Sayin & Krauss, 1996; Weisse et al., 2021).

Through these processes, it is clear that I must consider both surface and bottom salinity fields in addition to surface winds when examining water movement in my study area. Surface salinity may act as a proxy for the top layer, while the bottom salinity may act as a proxy for the bottom layer of water flow. To give additional insight, assuming the bi-directional flow, I can use mixed layer depth (MLD) as an indication of the state of the baroclinic currents. Over the entire study area, the MLD is on average 13 meters deep. In the southwest (SW) Baltic, the MLD goes even deeper, reaching more than 20 meters on average (not displayed). The relationship between surface salinity, bottom salinity and MLD is displayed in **Appendix A**, together with the depth of the MLD in regard to the depth of the water column.

Besides wind, one other important meteorological forcing are the changes in atmospheric pressure loading known as the Inverse Barometer (IB) effect. As increased air pressure exerts a force on its surroundings, it coerces water movement. This is a well-known process that states that, approximately, for every 1 mbar increase in surface air pressure, the sea level decreases by 1 cm (Roden & Rossby, 1999). Since normal air pressure ranges between 950-1050 hPa across the length of a year, air pressure induced sea level variability can be expected to reach -37 to +63 cm around the mean sea level annually from this effect alone (SMHI, 2014a).

Just as local sea levels can be varied, so can vertical land motion (**Figure 2**). Since the last Ice Age, the continental crust has been isostatically rebounding after being depressed by the ice sheets. In Sweden, GIA is responsible for inducing a land rise that varies from less than 1 mm yr^{-1} in the southernmost parts to 10 mm yr^{-1} in the northernmost parts of the country (Vestøl et al., 2019).

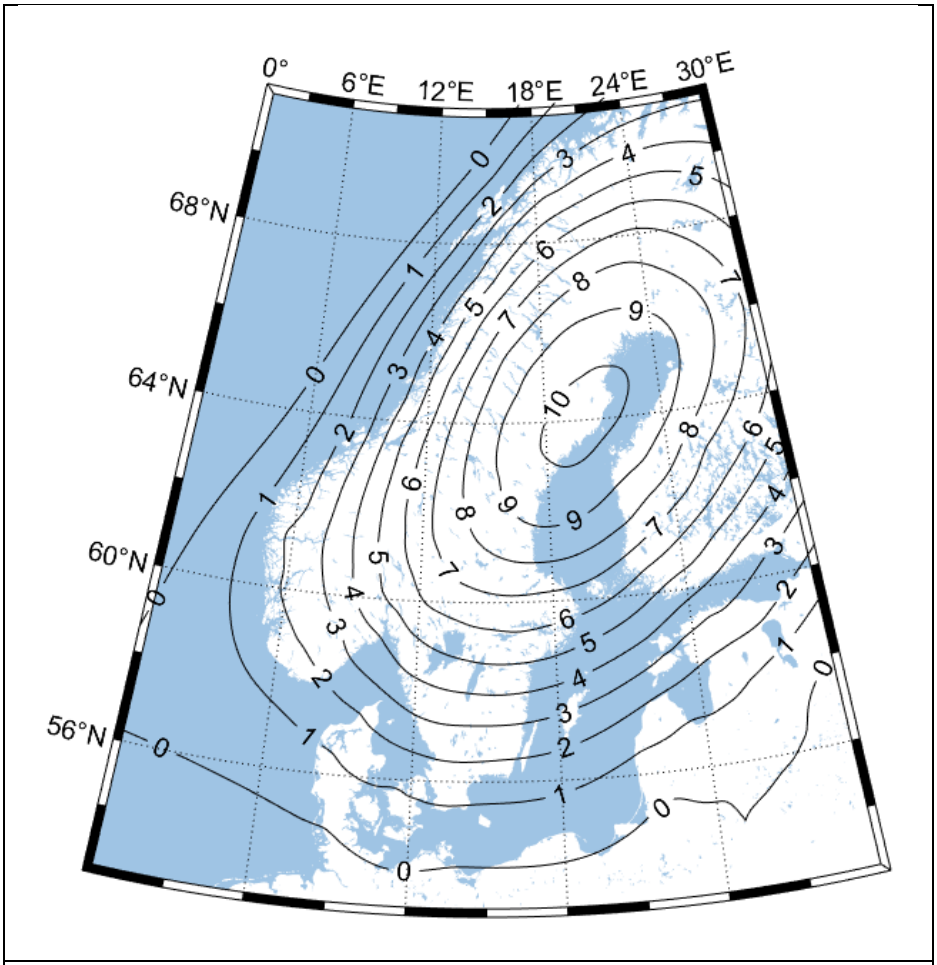


Figure 2: The vertical GIA [mm/yr] over Fennoscandia derived from the land uplift model NKG2016LU (Vestøl et al., 2019).

The diurnal tidal pattern is the most dominant tidal component in the Skagerrak and Kattegat Sea, it is however typically not larger than 5-10 cm offshore (Christensen et al., 2018). In the Baltic, the tides are virtually non-existent since the tidal signal undergoes significant filtering through the Danish Straits (Carlsson, 1998; Samuelsson & Stigebrandt, 1996). Such high-frequency variability is anyhow not captured in this study considering the temporal resolution of the datasets being used (daily to monthly). Therefore, I do not make any tidal corrections myself, and only use the tidal corrections already applied in the satellite altimetry dataset.

2 METHODS

2.1 DATA ACQUISITION

2.1.1 Satellite Altimetry

I use satellite altimeter gridded sea surface height data downloaded from the Copernicus Marine Environment Monitoring Services (CMEMS) database. The sea surface height timeseries is estimated through optimal interpolation techniques, merging all of the altimeter missions available since the first recordings in 1992. This includes Jason-3, Sentinel-3A/B, HY-2A, Saral/AltiKa, Cryosat-2, Jason-2, Jason-1, Topex/Poseidon, ENVISAT, GFO and ERS-1/2 (Pujol & Mertz, 2020). The dataset presents sea surface height data as sea level anomaly (SLA), in reference to a 20-year 1993-2012 average. To study long-term sea surface trends monthly SLA data spanning 1993-2019 was downloaded over the study area in 0.25° × 0.25° spatial resolution. To study short-term SSV, daily data spanning 2014-2019 was downloaded over the study area in 0.25° × 0.25° spatial resolution. The daily dataset is also utilized in the neural network made for SSV prediction.

Satellite altimetry works by means of transmitting a nadir-viewing radar pulse from the satellite down towards the Earth (Robinson, 2010). The distance that this pulse travels is referred to as the altimetric range (European Space Agency, 2022). The signal is reflected off the surface of the Earth and is received again by the satellite. The returned signal, or waveform, has a standard shape over most of the ocean, with a sharp leading edge followed by a gradually diminishing trailing edge (Cipollini et al., 2017). Given that the velocity of the pulse propagation is known, the time it takes the signal to be reflected is used to calculate the distance between the satellite transmitter/receiver and the Earth's surface. By knowing the satellites precise orbit and the distance between the orbit and an arbitrary reference ellipsoid, the height of the Earth's surface can be determined (European Space Agency, 2022; Robinson, 2010). This geocentric or absolute form of sea level observation has been measured over the last three decades, and has provided accurate nearly global observations on a near real-time basis (Church et al., 2013).

The authors of the satellite altimetry dataset have applied a number of corrections prior to distribution. These are explained in the following section.

As is presented in Fernandes et al. (2014), the height of the water surface (h) above a reference ellipsoid can be expressed as:

$$h = H - R_{obs} - \Delta R$$

where H is the height of the satellites center off mass above a reference ellipsoid, R_{obs} is the altimetric range corrected for all instrument effects, and ΔR is the combined corrections applied for all range and geophysical effects. The corrections, or ΔR, can independently be expressed as:

$$\Delta R = R_{ion} + R_{wet} + R_{dry} + R_{SSB} + R_{DAC} + R_{tides}$$

R_{ion}, R_{wet} and R_{dry} are corrections that have to be made to account for different speeds of light through the Earth's ionosphere and troposphere, which cause slowdown of the electromagnetic signal. This entails ionospheric, wet tropospheric and dry tropospheric correction. In the ionosphere, the delay is caused by signal refraction by free electrons, the correction for which can accurately be applied by employing dual-frequency altimeters. In the wet troposphere the delay is caused by water vapor and in the dry troposphere the delay is caused by other dry gasses, mainly nitrogen and oxygen (Fernandes et al., 2014). To accurately apply these corrections, a three-channel microwave radiometer to determine atmospheric water vapor content can be used (Robinson, 2010). However, since not all satellite systems are equipped with the necessary sensors, different correction methods may be used for different satellites. Since the dataset used in this study merges many different satellite systems together, in cases when the satellite is not equipped with all necessary sensors, the creators instead opt for model-based estimates of atmospheric water vapor content, which for coastal areas may even be preferred as we discuss in a later chapter. Corrections for R_{SSB} (Sea State Bias), R_{DAC} (Dynamical Atmospheric Correction) and R_{tides} (tidal components) also come included in the dataset. SSB correction is based on wind and wave height estimates and is caused by the influence of ocean waves on the returned waveform. The pulse is better reflected from the smoother throughs than the peaks or crests, which results in estimates

of sea level being too low (Cipollini et al., 2017). DAC corresponds to the removal of the barotropic ocean response to atmospheric forcings, necessary to isolate the response in terms of sea level. The dataset uses barotropic models forced by pressure and wind simulations. While it does filter out low frequency variability caused by the IB-effect, high frequency variability remains. Finally, ocean tidal effects are removed using the FES 2014b tidal model, which removes 34 ocean tidal components, in addition to the correction for pole tide and solid earth tide.

The limitations of satellite altimetry is its rather short span of available data, obviously too short to derive GMSL estimates on century length time scales.

2.1.2 Tide Gauges

The other form of sea level observation is instead measured in respect to the solid earth, and is thus referred to as relative sea level (RSL). RSL has in some locations been measured for centuries, in the form of in situ tide gauge stations. In this report I use monthly averaged tide gauge data from 30 stations from the `Revised Local Reference (RLR)´ dataset acquired from the Permanent Service for Mean Sea Level database (https://www.psmsl.org/) (Holgate et al., 2012; PSMSL, 2022). Tide gauge records were chosen on the basis of foremost being inside my study area and secondly, between the 1993-2012 reference period not having more than 3 years with less than 75% data completion (meaning 9 out of 12 months per year). The locations of the stations are shown in **Figure 1** while **Table 1** gives an overview of the name and data series length and data completeness of tide gauge stations used. Fortunately, most gaps in the data are concentrated to the earlier stages of the tide gauge station timeseries. Since 1993, only occasional years are incomplete.

Table 1: Station number and name of the monthly tide gauge dataset. Years available indicates how many years the tide gauge is available. The value in parenthesis shows how many years have at least 75% data completion.

Station number	Station name	Total years available	Availability 1993-2012
1	OSCARSBORG	148 (64)	20 (20)
2	STAVANGER	102 (91)	20 (20)
3	TREGDE	93 (89)	20 (20)
4	HELGEROA	55 (41)	20 (20)
5	VIKER	30 (30)	20 (20)
6	KUNGSVIK	47 (47)	20 (20)
7	SMOGEN	110 (110)	20 (20)
8	HIRTSHALS	126 (118)	20 (20)
9	HANSTHOLM	65 (49)	20 (20)
10	GOTEBORG - TORSHAMNEN	52 (52)	20 (20)
11	RINGHALS	53 (50)	20 (20)
12	STENUNGSUND	58 (54)	20 (20)
13	VIKEN	44 (44)	20 (20)
14	AARHUS	129 (126)	20 (19)
15	FREDERIKSHAVN	124 (118)	20 (20)
16	HORNBAEK	127 (124)	20 (20)
17	BARSEBACK	84 (60)	20 (20)
18	KOBENHAVN	129 (124)	20 (18)
19	KLAGSHAMN	91 (91)	20 (20)
20	SKANOR	28 (28)	20 (20)
21	SASSNITZ	84 (74)	20 (20)
22	WARNEMUNDE 2	164 (164)	20 (20)
23	WISMAR 2	170 (170)	20 (20)
24	TRAVEMUNDE	163 (155)	20 (20)
25	GEDSER	126 (126)	20 (20)
26	KIEL-HOLTENAU	63 (50)	20 (20)
27	FYNSHAV	50 (48)	20 (18)
28	FREDERICIA	128 (127)	20 (20)
29	KORSOR	121 (116)	20 (20)
30	SLIPSHAVN	122 (119)	20 (19)

Modern tide gauges function in a somewhat similar fashion to satellite radar altimetry. The main instrument is often a microwave radar sensor placed within a sounding tube that is connected to the ocean. The time it takes for the radar pulse to travel back from the water surface is recorded and used to calculate the sea level (NOAA, 2021). Other instruments include pressure sensors, which instead are submerged in the water. The sea level is then determined by measuring the pressure exerted by the water column (SMHI, 2021). Before the digital age, the sea level was measured using floats connected to an analog recorder (NOAA, 2021).

While sea level timeseries from tide gauge stations exist that started long before the first satellite altimeter missions, the data is rather limited in its spatial distribution. Additionally, the sea level datasets from tide gauges are largely restricted to the

coastlines, particularly those with the longest data records (Woodworth & Player, 2003). This gives limited to no information about open ocean processes. Furthermore, since tide gauges record sea level in reference to the solid earth, it always includes vertical land & crustal movements of the ground itself. Throughout Fennoscandia, considerable vertical land movement occurs every year, often larger than the ocean movement itself (Vestøl et al., 2019). In the Baltic-North Sea transition zone, rates of GIA range from +4 mm yr^{-1} in northern Skagerrak to 0 mm yr^{-1} in the southern regions of the Baltic (**Figure 2**). This entails that the relative SLR trend changes throughout the region. In order to make the tide gauge dataset comparable to both other tide gauges and the satellite altimetry dataset, the difference in local GIA must be accounted for. Previous studies make corrections for this by applying land-uplift models such as the NKG2016LU presented in Vestøl et al. (2019) (**Figure 2**). I instead opt for a linear least square detrending of both datasets. This removes the uneven vertical land movement over the area as well as any trend of sea level rise/fall. Since I am interested in studying the short-term variability of sea level and not the long-term changes, this is not an issue. Additionally, since the satellite altimetry dataset is presented as SLA over a 20-year 1993-2012 average, the same 20-year average is computed and subtracted for each tide gauge station, obtaining the sea level anomaly over this reference period.

2.1.3 Accuracy between sea level datasets

A comparison and correlation analysis between all tide gauge stations and nearest cell of the gridded satellite altimetry timeseries is visualized in **Figure 3**. The tide gauge timeseries are here restricted to January 1993-December 2019 to match the satellite altimetry dataset. If the tide gauge timeseries are missing data for any month, the month in question is also excluded from the altimetry dataset.

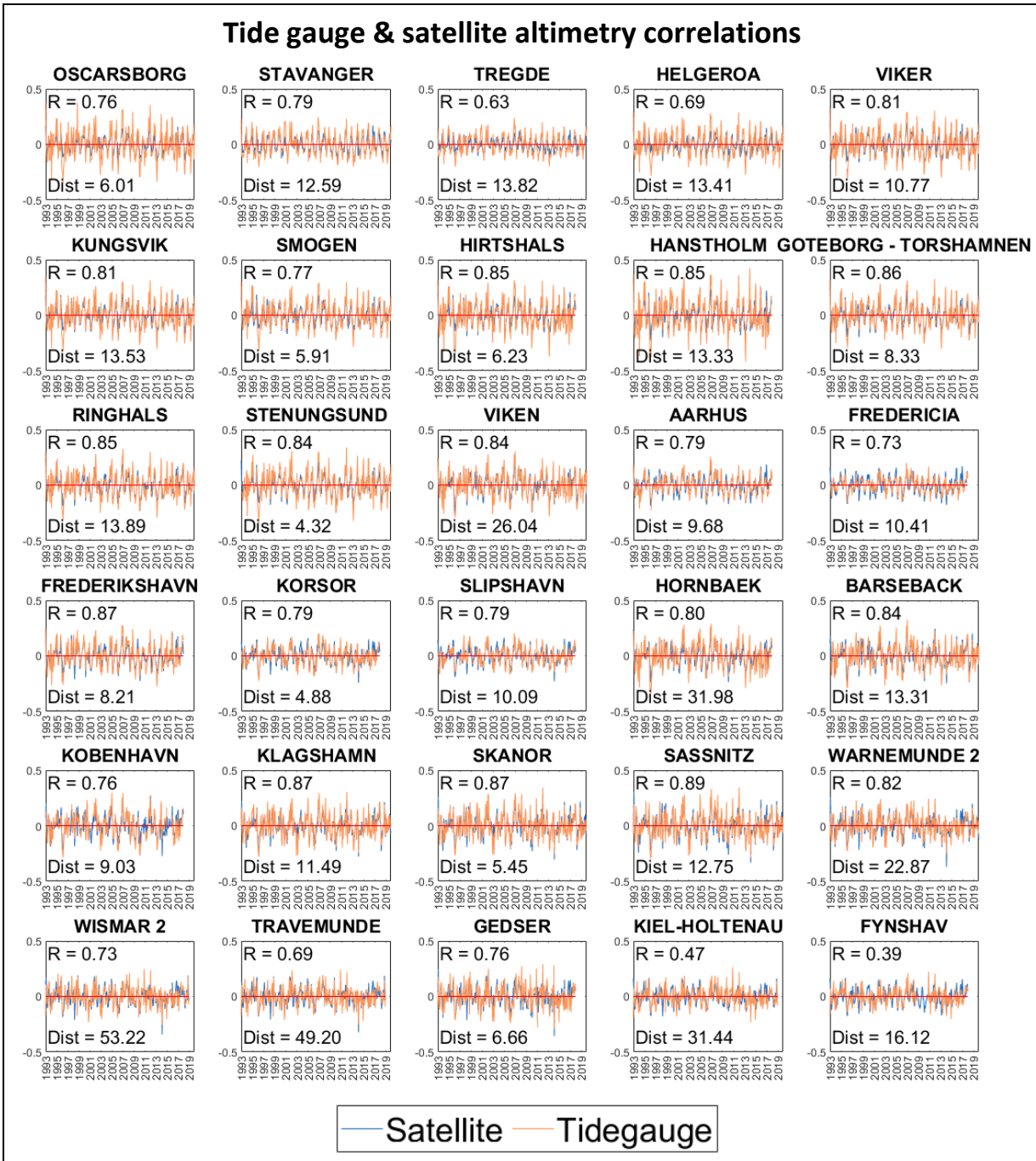


Figure 3: For all 30 tide gauge stations, correlation between the detrended tide gauge data (orange) and that of the closest point of the gridded altimetry dataset (blue). Correlation coefficient R and distance to the center of closest cell indicated for each station.

Generally, there is a high agreement between tide gauge records and the satellite altimetry dataset. 25 out of the 30 tide gauge–satellite altimetry pairs exhibit correlation coefficients higher than 0.7, and only 2 have coefficients lower than 0.6. Lowest correlations were found at Fynshav, Denmark (R = 0.39) and Kiel, Germany (R = 0.47) (**Figure 3**). Both these stations are also located south of the Little Belt (station no. 27 &

26, **Figure 1**), an area that exbibits inconsistent sea level behavior. This is further discussed in Chapter 3.1.

Non-perfect correlations can be expected due to the spatial resolution of the altimetry dataset. At 0.25° × 0.25° resolution, each gridded datapoint represents the average sea level over approximately 420 km^2 of ocean and leads to many coastal areas to be unrepresented. In contrast, tide gauges report the sea level at a distinct point, the placement of which could be inside a protected bay or harbor, and may be far away from the closest cell of the satellite altimetry dataset it is being compared to. This leads to tide gauges not being well representative of offshore processes.

Furthermore, there are known issues that contribute to the decrease of confidence in satellite altimetry data near the coasts. Traditionally, satellite altimetry is designed for the open ocean and within 10-15 km of the coast, it is often deemed unreliable (Madsen et al., 2007). The work to develop new methods, retracking algorithms and satellite sensors to increase data precision at the coasts is an active area of study for the satellite altimetry scientific community (Cipollini et al., 2017). For instance, the received waveform can be distorted by surface inhomogeneities, as is the case at the ocean-land transition areas where the presence of land corrupts the echo (Passaro et al., 2015). It is then important to implement retracking algorithms that are able to analyze the distorted waveforms (Cipollini et al., 2017). Near the coast, the corrections that need to be applied become unreliable as well. For instance, the water-vapor correction in the troposphere can be distorted by intruding land in the radiometer footprint, which can cause several centimeters of error (Cipollini et al., 2017). Such errors can often be remedied by instead implementing model-based corrections. The Baltic-North Sea transition zone possess countless islands and jagged coastlines, all sources for satellite altimetry error that leads this area to be particularly prone to data inaccuracies.

Additionally, significant differences in corrections and filtering exists between the two data products. Since the altimetry dataset is corrected for the tidal components and DAC while the tide gauge data is not, discrepancy between the two datasets should be expected for this reason as well.

2.1.4 Sea level drivers

A total of nine possible sea level drivers are selected to be included in the analysis. These are drivers that past studies have found to be the most important in determining SLV in the study area as I have demonstrated in Chapter 1. More variables could be included, such as precipitation, evaporation, surface run-off, solar irradiance, NAO index etc. but considering the time-scope of this project, I have decided to not include more than the ones listed in **Table 2**. The variables are collected from both the "*ERA5 hourly data on single levels from 1979 to present*" (ERA5-h) dataset from the Copernicus Climate Change Service (C3S) Climate Data Store and the "*Baltic Sea Physics Reanalysis*" (BSPR) dataset from the Copernicus Marine Environmental Monitoring Services (CMEMS) database. A brief description of the sea level drivers is presented below in **Table 2**.

Table 2: The nine possible sea level drivers that I include in the analysis against the SLA datasets.

Driver	Description	Unit	Dataset
U-component of 10 m wind (Zonal)	The horizontal speed of air moving towards the east at 10 meters above the Earth surface.	$m\ s^{-1}$	ERA5-h
V-component of 10 m wind (Meridional)	The horizontal speed of air moving towards the north at 10 meters above the Earth surface.	$m\ s^{-1}$	ERA5-h
Sea surface temperature	The temperature of the sea water at the surface.	K	ERA5-h
Sea level pressure	The pressure exerted at the Earth's surface by the weight of a vertical column of air.	Pa	ERA5-h
U-component of surface currents	The horizontal velocity of eastward surface currents.	$m\ s^{-1}$	BSPR
V-component of surface currents	The horizontal velocity of northward surface currents.	$m\ s^{-1}$	BSPR
Surface salinity	The amount of salt dissolved at the ocean surface	PSU	BSPR
Bottom salinity	The amount of salt dissolved at the sea floor	PSU	BSPR
Mixed layer depth	The depth from the sea surface of the homogenous mixed layer	m	BSPR

2.1.4.1 ERA5 Atmospheric Reanalysis

Four drivers come from the "*ERA5 hourly data on single levels from 1979 to present*" dataset from the C3S Climate Data Store: eastward and northward surface winds, sea surface temperature (SST) and sea level pressure (SLP) (Hersbach et al., 2018). The data are provided as hourly estimates on a 0.25° × 0.25° spatial grid and are downloaded over the 2014-2019 period covering the entire study area. The variables are derived from the ECMWF re-analysis that follows data assimilation principles of combining model data with observations, consisting of both satellite and in-situ observations of temperature,

humidity, 10 m winds and more. Since the data is only available in hourly estimates, I computed the 24-hour daily averages for these variables before they could be used in the analysis.

2.1.4.2 Baltic Sea Physics Reanalysis

The remaining five drivers come from the "*Baltic Sea Physics Reanalysis*" dataset from the CMEMS database: eastward and northward surface currents, surface and bottom salinity, and mixed layer depth (https://doi.org/10.48670/moi-00013). The data are provided as daily estimates on a 4 × 4 km spatial grid; I downloaded them over the 2014-2019 period. The dataset unfortunately does not cover the entire study area, and is limited by its western longitudinal boundary at 9° East. It does however include the whole Baltic Sea and Kattegat as well as most of the Skagerrak. The variables are derived from the ice-ocean model NEMO-Nordic (based on NEMO-3.6) together with assimilated sea surface temperature profiles and salinity profiles. The ocean model is developed and used by the Swedish Meteorological and Hydrological Institute (SMHI) (Hordoir et al., 2019). Since this dataset is on a much finer grid compared to the rest, for analysis with the satellite altimetry dataset I interpolated it onto a matching 0.25° × 0.25° spatial grid using a 2-D linear interpolation technique.

The wind is split into two variables of zonal (u10) and meridional (v10) winds. These range from negative to positive vector values that are projected onto the x or y axis. Negative zonal winds simply mean that the wind is positive in the westward wind direction. To obtain the true wind direction and wind speed, the zonal and meridional wind components must be combined. However, by leaving them as separate variables, one can get a better understanding of how the true North-South-East-West wind bearings affect sea level variability. For this reason, I chose to not combine the wind variables, and leave them as their separate components.

2.2 IDENTIFYING BASINS OF COVARIANCE

Regions of covariance are determined by plotting the correlations between the tide gauge timeseries against the "straight-line" distance between them. I only include correlations that are significant on a 95% confidence level. Generally, the correlation between the timeseries will decrease with distance between the tide gauge stations. I initially placed the tide gauge stations within three geographical sub-basins: the Skagerrak, Kattegat, and

the SW Baltic. I took a line-of-best-fit approach, where the goal was to minimize the root mean squared error (RMSE) and the slope of the line with as few sub-basins as possible. I also considered other statistical metrics such as the coefficient of determination (R^2), which represents the proportion of the variability seen in the response variable Y (the correlation) that is explained by the distance variable X. I did this manually by testing different configurations of tide gauge groupings. The final result includes a fourth sub-basin, which I call the Belts, as these stations did not fit well into any other sub-basin (**Figure 1**).

The cross-basin analysis was done by similar means, except I instead calculated the correlation coefficients between every possible tide gauge station pair featuring stations from the now-defined separate sub-basins. For instance, the three tide gauge stations located in the Belts sub-basin are individually paired with each of the nine tide gauge stations located in the Kattegat sub-basin. Between them, they create 27 pairs of tide gauge timeseries combinations. Both the Pearson correlation coefficient and the "straight-line" distance is calculated between each tide gauge combination, and plotted against each other.

2.3 MAIN DRIVERS AS DETECTED BY STATISTICAL METHODS

To determine the main drivers of sea level variability by statistical methods, I use MATLAB (v.2021B) and the Climate Data Toolbox – a set of functions written for the analysis of climatic data (Greene et al., 2021). I also use M-Map, a mapping package for MATLAB, to create the various maps seen in the report (Pawlowicz, 2020).

2.3.1 Pre-processing of datasets

Primarily, in most sea level drivers there exists a strong seasonal signal that dominates the annual short-term variability. It is particularly apparent in variables such as SLA, SST, surface salinity and meridional wind (**Figure 4**).

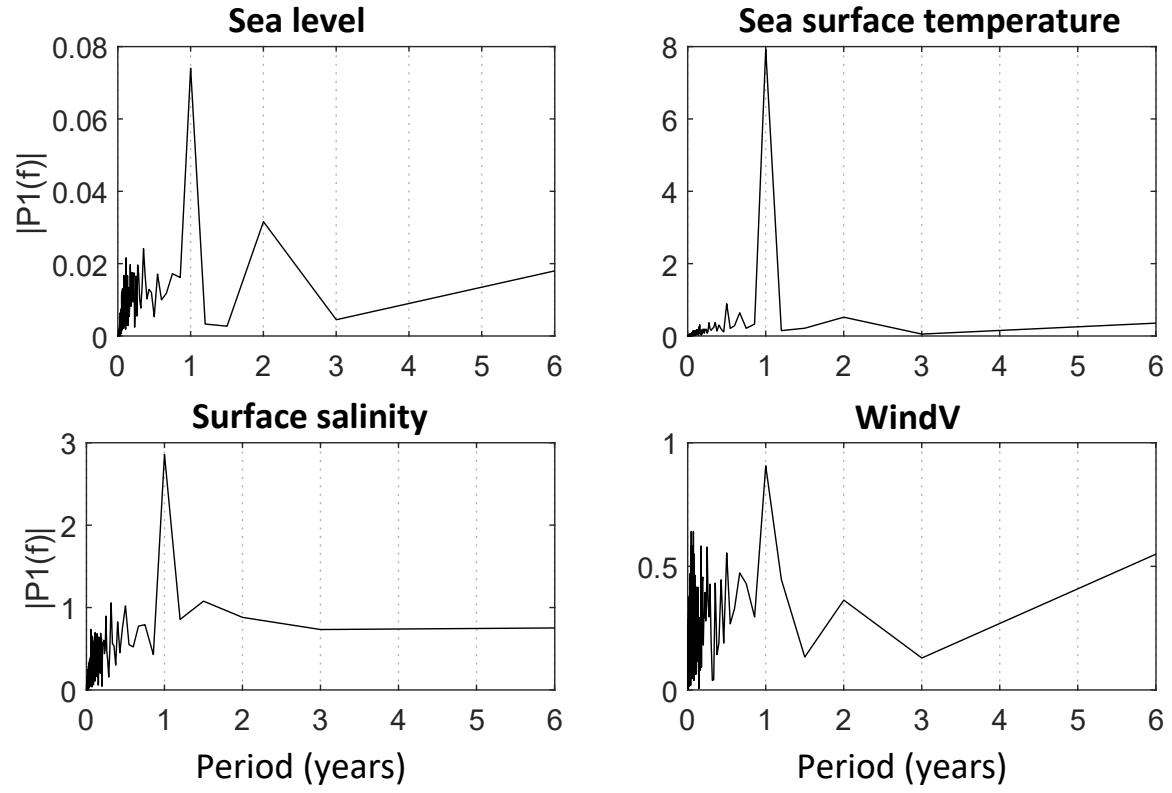


Figure 4: Fast Fourier Transforms of some select sea level variable time-series. The large peak at the 1-year period indicates a strong recurring signal that returns every year. This is the seasonal cycle of the variables that is to be removed.

There are multiple reasons to remove the seasonal cycle when computing multi-variate analysis. First, as is apparent in **Figure 5**, each timeseries has its own seasonality. For practical reasons, it is better to remove the seasonal cycle altogether than to deal with them individually. Second, many climatic variables are inherently seasonal, for instance Nordic sea levels being higher during winter months. When sea levels decrease in the fall it does not always signal an important change to any background driver, but could simply be the seasonal decrease, which in my case holds no important information. By removing the seasonal cycle, I limit any spurious correlation between drivers and sea level. Within a machine learning approach, it is also important to remove any seasonality when forecasting from timeseries, as it provides a clearer relationship between input and output variables. As we will see, it is important to ensure the input variables are independent from one another. If I do not remove the seasonal cycle, there is a large risk of the variables not behaving independently. To remove the seasonal cycle, I estimate the climatology by fitting multi-year daily averages of the data. The approximated climatological cycle is then removed from the annual sequence. This produces a seasonal

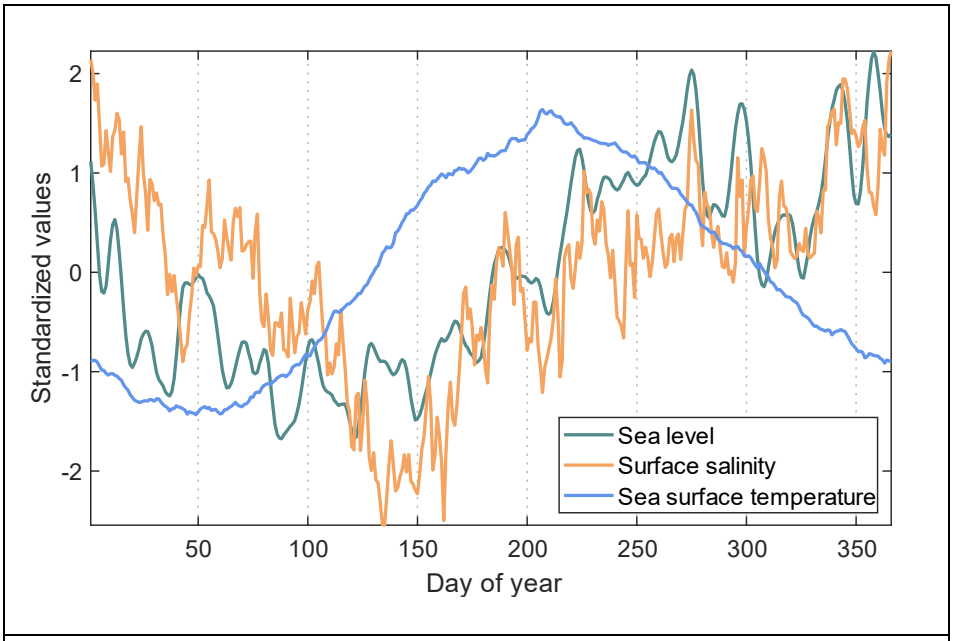


Figure 5: The seasonal cycle of sea level, surface salinity, and sea surface temperature at the Kattegat research point.

stationary time-series suitable for my analysis. The timeseries datasets are also detrended for the same purpose, using the aforementioned linear least-squares regression techniques. The estimated trend is then removed from the signal.

2.3.2 Statistical analysis

First, I calculate the Pearson Correlation Coefficient for each gridded point between each sea level driver and sea level anomaly to get an overview of the relationships between sea level variability and its background drivers. Only correlation coefficients within a 95% confidence interval are included in these figures. While it is important to distinguish between correlation and causation, correlation analysis gives a good sense of how well the full variability exhibited in the sea surface height signal is accounted for, or represented by, the full variance of the sea level drivers.

I follow statistical decomposition methods of Principal Component Analysis (PCA) and Empiric Orthogonal Function (EOF) analysis for timeseries signal breakdown of the daily SLA data. PCA/EOF are common and useful multivariate statistical techniques to analyze climatic data because they can reduce many variables in a dataset to much fewer new variables, which provide insight into both spatial and temporal variations (Wilks, 2006). It is common that a large number of principal components (PCs) are needed to explain all

of the variance within a signal. Luckily, the first few PCs usually capture sufficient variance. In my case, the first four PCs explain almost 90% of observed variance within the daily SSH dataset, and the first two explain almost 80%. Each PC functions as an orthogonal vector in time, which means that they are independent of each other. In other words, the variance explained by one does not overlap with the variance explained by another. The first principal component will capture the most dominant part of the variance, the second will capture the second largest part of the variance that is not explained by the first, and so on. This is in part why PCA is such an efficient and useful tool. In reality, PCs correspond to eigenvectors, the magnitude of which are determined by their accompanying eigenvalues. These are vector and scalar properties that have been calculated from the dataset's covariance matrix. It is the eigenvalues of the covariance matrix that describe the fraction of variance explained by each PC. Why and how this works can be explained in mathematical detail, but it is beyond the scope of this work and will not be covered in this report. While the PCs show the temporal variance of the signal, the EOF show the spatial structures of them. They are however both calculated simultaneously. Likewise, as the PCs are orthogonal in time, so are the EOFs in space, meaning there does not exist a spatial correlation between two EOFs. It is then possible to visualize the pattern of variability for each mode, and possible to identify underlying causes for each one. The EOF maps show which areas co-vary in the same opposition of phase, which areas co-vary in the opposite opposition of phase and which areas that are not affected by the mode in question at all.

As I will show when I describe my EOF results in Section 3.2, I find that the explained variability drops dramatically after the first two EOFs, so I choose to focus on characterizing these two modes that together explain 80% of the sea level variability in the area. I compare the first and second PCs independently with the sea level drivers, as seen in for instance Passaro et al. (2021) where they compare the two most prominent PCs with both zonal and meridional winds. I do this by correlating the PCs with each of the deseasoned and detrended sea level drivers.

2.4 MAIN DRIVERS AS DETECTED BY MACHINE LEARNING

Machine Learning is the science of utilizing computer algorithms that are capable to continually improve its accuracy by self-learning. It has been used in many vastly

different fields of science to solve clustering, classification, regression problems and more. In this study, I use a type of Recurrent Neural Network (RNN) with a Long Short Term Memory (LSTM) architecture to predict sea levels, and to determine the primary drivers of sea level variability at two distinct points in the North – Baltic Sea transition zone (shown on **Figure 1**).

Neural networks are a subsection within Deep Learning and are built out of nodes connected through layers. The general idea is that the network mimics the architecture of the human brain, with thousands of interconnected nodes/neurons (Goodfellow et al., 2016). There are input layers, one or more hidden layers and an output layer. The model is fed with training data, which in my case consist of the nine possible sea level drivers listed in **Table 2**. All nodes in adjacent layers are connected to each other while nodes in the same layer are not, however the connection between particular nodes can sometimes be weak, effectively cutting the node off from some other nodes. This happens through the training process. At the start of the training session, each node within the network is assigned random weights and bias values. As each node receives an incoming value from each of its connections, the values are multiplied with the associated weight and added together. Only if the resulting product exceeds the bias value will it send the signal forward to the next layer. During the training process, the weights and bias values are continually adjusted to minimize the loss, which in my case is the root mean squared error between the predicted sea level and the true values. The process is then repeated over as many iterations as necessary to get the most accurate result, which is called gradient descent. This type of basic network is called feed-forward because information only flows in one direction, from input x, through the intermediary layers that define $f(x)$, which lead to an output $\hat{y}$. No feedback function, where for instance the output of the model feeds back into itself, exists in this type of network (Goodfellow et al., 2016).

Early testing using feed-forward neural networks did not perform particularly well for my problem. While it did capture some variability, it was limited in its execution. One likely reason is the fact that change within a sea level driver does not always result in an immediate response in sea level. More likely, there exists a delay between signals, as I will show in the Results. Unfortunately, standard feed-forward neural networks treat each point in the sequence independently, unable to remember what happened before. RNNs are a type of modulation of the feed-forward neural network, that also possess the ability

for the nodes to use the previous output and store it within its memory for a short time. What this means in practice, assuming daily values, is that it can store a certain number of past days for each forcing, which it will use to predict the current sea level. However, there is an issue with RNNs related to backpropagation between the layers that occurs when updating the weights of each node: the vanishing gradient problem. As the algorithm moves backwards to update the weights related to earlier and earlier time steps, the gradient may start to get smaller in size until the weights are no longer updated, which prevents training (Goodfellow et al., 2016; Yu et al., 2019).

The LSTM architecture consists of a chain of repeating modules or cells within the hidden layers. Within each cell there are three gates, called the forget-gate, input-gate and output-gate that simply put, are filters for the ingoing and outgoing data (Yu et al., 2019). The gates filter the data through sigmoid activation functions, meaning that it assigns a weight between 0 and 1 depending on the importance. The LSTM networks solve the vanishing gradient descent problem by using these gates that ensure that previous information is retained.

Although I will not focus on the mathematical proof behind RNNs, I will present the fundamental equations which describe the models below. For a full and detailed explanation of RNNs and LSTMs, I refer the reader to Sherstinsky (2020).

Mathematically, the first layers of a simple RNN can be described as:

$$h_t^{(1)} = tanh\,(W_h^{(1)} x_t + b_h^{(1)} + W_h^{(1)} h_{t-1}^{(1)}) \qquad (1)$$

where h_t is the output of the current layer at time *t*, *tanh* is the activation function, W_h is the weight, x_t is the input data, b_h is the bias and $W_h h_{t-1}$ is the output of the past hidden layer. The following layers *n* can be described as:

$$h_t^{(n)} = tanh\,(W_h^{(n)} h_t^{(n-1)} + b_h^{(n)} + W_h^{(n)} h_{t-1}^{(n)}) \qquad (2)$$

And the output layer $\hat{y}_t$ as:

$$\hat{y}_t = W_o^{(L)} h_t^{(L-1)} + b_o^{(L)} \qquad (3)$$

The final layer is a weighted linear combination of the input plus a bias.

The architecture of a LSTM cell is more complex than the RNN, since it includes multiple gates. Following Yu et al. (2019) it can be described as:

$$f_t = \sigma\left(W_{xf}\, x_t + W_{hf}\, h_{t-1} + b_f\right) \quad (4)$$

$$i_t = \sigma\left(W_{xi}\, x_t + W_{hi}\, h_{t-1} + b_i\right) \quad (5)$$

$$o_t = \sigma\left(W_{xo}\, x_t + W_{ho}\, h_{t-1} + b_o\right) \quad (6)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \phi\left(W_{xc}\, x_t + W_{hc}\, h_{t-1} + b_c\right) \quad (7)$$

$$h_t = o_t \odot \phi(c_t) \quad (8)$$

where f_t, i_t and o_t are the forget gate, input gate and output gate respectively at time t. The gates all have the same format, and are calculated using the previous hidden state and current input data. They use a sigmoid activation function which assigns a value from 0 to 1. x_t is the input data, h_t is the current hidden state and c_t is the cell state. The cell state is a weighted sum of the previous cell state controlled by the forget gate and the simple RNN equation controlled by the input gate. W_i, W_c, and W_o are the weights and b_f, b_i, b_o, and b_c are the bias. ϕ is the activation function, and the $\odot$ operator denotes elementwise multiplication of two vectors.

After computing the output value, the root mean square error (RMSE) between the predicted and true value can be calculated. This is expressed as the loss, which the goal is to minimize as much as possible.

To be able to test the predicted results against independent sea level values, the data are first split into 3 categories. 70% of the data is used as training data. The remaining 30% are evenly split into validation data, which are used for model evaluation during training, and test data, which are used to evaluate the model after the training is completed. Knowing the current training set loss, backpropagation algorithms calculate the error gradient, which is used to update neuron weights and biases. This process is then repeated for several iterations/epochs until the loss no longer decreases. If the loss continues to decrease, while the validation set loss does not there might be a case of overfitting, which means that the neural network memorizes the training dataset while showing worse results on the validation dataset. I implement an early stopping algorithm

to preemptively combat overfitting by ending model training when the validation loss stops improving. The algorithm has a patience of 3, meaning that should the validation loss not improve over 3 epochs, I stop the training and revert the model back to its best state, i.e., the smallest validation loss.

The LSTM network is built and run on a local machine on Jupyter Lab, using the Tensorflow library. Two locations, one in the Kattegat Sea and one in the SW Baltic, are chosen for the ML analysis. These were chosen based on the differences between the spatial coherence regarding sea level in these regions, which I will describe in Section 3.1. Since the model is sensitive to the magnitude of the data values, all datasets except the sea level time series are normalized before use. There is no need to normalize the target of the neural network.

I first conduct an experiment to determine the optimal number of past days to include in the prediction, or rather, the optimal sequence length. I let the number of layers (2) and number of neurons within the layers (200 & 400) stay the same between model runs, and only change the number of past values the neural network should use to predict. The amount of layers and an approximation of a good amount of neurons within the layers were determined by early testing and experimentation. I start with running the model with a sequence length of 1, 2, 3, 4, 5, 6, 8, 10, 15, 20, 30 and 90 days. If the minimum loss is possibly between two of the values, then the model is run again with sequence lengths in between these values that previously have been untested. Since initial weights and biases are random, the model will produce different results despite using the same settings. For this reason, multiple runs should be computed for each setting. I train the network 30 times for each run. The results showing the RMSE between predicted and true values and the optimal sequence length for each location are discussed in Section 3.3.

The second experiment is conducted in order to describe each location's SSH primary drivers. Before I train the models however, it is important to optimize the model in order to get the best results. I run the Keras Tuner, a deep learning optimization framework that tests for the optimal hyperparameter values (O'Malley et al., 2019). Knowing the best hyperparameter values to use before you start training a model is challenging and finding the optimal configuration manually is extremely time consuming and virtually

impossible. Keras Tuner is built to automate this process. The hyperparameters I determined this way are the number of neurons in the two hidden layers. I set the tuner to test for values between 50 – 400 neurons in the first hidden layer and 200 – 600 neurons in the second hidden layer, with increments of 50 in-between. These values were found to give good model result based on early testing. The tuner tests different configurations of neurons between the layers and returns the best one. Knowing how many past days each location should use from the previous experiment, the model is first run 30 times at both locations using the tuned hyperparameter values with all possible drivers to produce "base" runs. Thereafter, I exclude one of the drivers and I rerun the tuner and run the model 30 more times. This process is replicated until each driver has been excluded from a model run. Based on the median RMSE between predicted and true values across the 30 model runs for each excluded driver, a ranking can be produced where an increased RMSE compared to the base run indicates that the excluded driver is significant in controlling local sea level.

In order to visualize how well the model performed in predicting sea levels, I produce graphs showing all 30 model runs from one experiment, highlighting the best performing one. The graphs have been smoothed using a low-pass Butterworth filter using a 4-day running mean. This removes noise created by the prediction, a result of the model predicting sea levels for each day separately, independent of past predictions.

3 RESULTS AND DISCUSSION

In the following chapter, the results are presented concurrently with the discussion. The first section presents the results from the identification of basins of co-variance estimated from in-situ tide gauge station data. The second section introduces the results from the statistical analysis between sea level variability and background drivers. The third section presents the results from the LSTM neural network experiments.

3.1 BASINS OF COVARIANCE

The Baltic – North Sea transition zone can effectively be organized into four separate regions or sub-basins of sea level variability. These are Skagerrak, Kattegat, SW Baltic and the Belts region. One common attribute of all regions presented in **Figure 6**, is that the correlation between tide gauge stations generally decreases with distance.

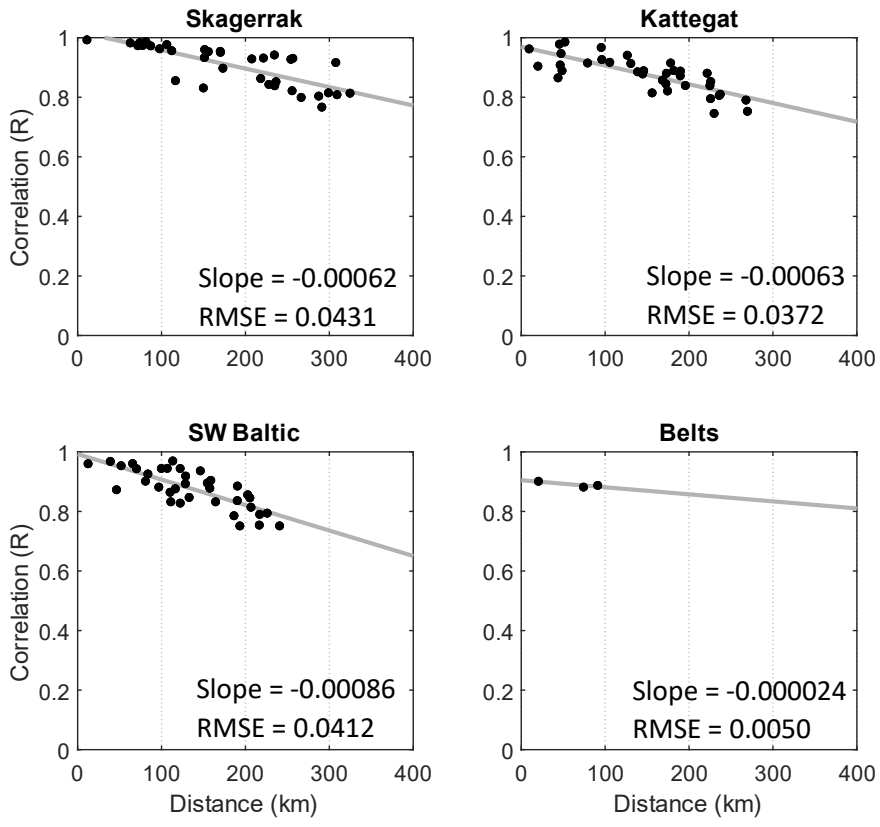


Figure 6: Dependence of correlation between pairs of time series on the distance between the tide gauges calculated separately for each basin. The displayed values are the slope of the line and the RMSE of the regression.

The Belts region contains only three tide gauge stations however, leading to this region being poorly represented and results not being entirely conclusive. Nevertheless, these three stations did not fit well into any other of the sub-basins. When included in the two neighboring sub-basins of Kattegat and SW Baltic, it became clear that they do not correlate well with any of the other tide gauge stations, significantly worsening the correlation – distance relationship. In the Kattegat, the slope of the line steepened from -0.00063 to -0.0010, the RMSE increased from 0.037 to 0.071 and the R^2 decreased from 0.61 to 0.56. In the SW Baltic, the RMSE increased from 0.041 to 0.086 and the R^2 decreased from 0.6 to 0.1. They do however correlate well with each other, leading to them being separated into their own sub-basin. The three tide gauge stations cover two of the three main belts that connects the Baltic Sea to the Kattegat. The third passage, the Oresund, is instead best sorted into the Kattegat sub-basin. This passage is represented by tide gauge station 13, 16, 17 and 18 (**Figure 1**). While this divide between the sounds may seem surprising, one should first consider the method for calculating the distance between the tide gauge stations. I here used a "straight-line" method of measuring distance, but this is not always well representative of the true distance the water would have to travel between two individual stations. The true distance between tide gauge stations can hence be substantially longer. For instance, as can be seen in **Figure 1**, while the "straight-line" distance between station 29 in the Belts and station 18 in the Kattegat may appear relatively short (100 km), the stations are situated on opposite sides of and separated by the large Danish island of Zealand. This leads to the timeseries of tide gauge station 18 not being well represented by the timeseries at tide gauge station 29, leading the two tide gauges to have a seemingly short "straight-line" distance but poor correlation.

Additionally, it has been shown that the water transport between the Baltic and the Kattegat occurs faster through the Oresund in comparison to the other two straits. This is thought to be mostly caused by the shorter length but also a smoother local bathymetry. This leads it to be much more straightforward for the water to travel through the Oresund in comparison to the others (Lass & Mohrholz, 2003). This most likely has a large contribution to the Belts sub-region experiencing different sea level variability compared to the Oresund.

Figure 7 presents the results from the cross-basin tide gauge correlation analysis, i.e. the correlation of pairs of stations from different basins. These results confirm the general trend of correlation decreasing with distance that I just discussed. This result holds true for all cross-basin analysis combinations except for the Belts – SW Baltic duo (**Figure 7f**). Between these sub-basins, there is a general reversed trend of correlation instead increasing with distance. However, the datapoints are indeed quite scattered (RMSE = 0.08) making the linear relationship more unreliable. The Skagerrak – SW Baltic cross-basin analysis (**Figure 7d**) also possesses more scattered datapoints (RMSE = 0.14). In fact, all cross-basin analyses that include the SW Baltic sub-region appear to contain more scattered datapoints and higher RMSEs (0.12, 0.14, and 0.08) compared to the other combinations (0.06, 0.04, and 0.05). This may indicate that the SW Baltic sub-basin's sea level variability is more complex and different from the rest of the sub-basins.

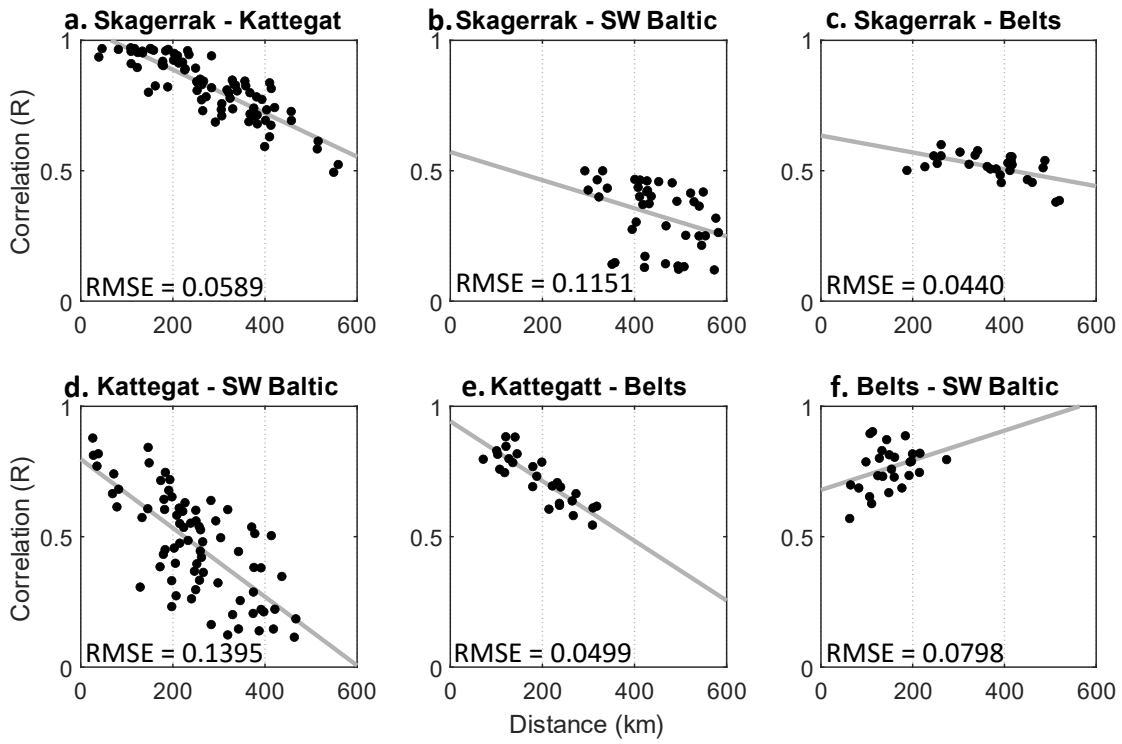


Figure 7: Cross-basin correlation analysis: Dependance of correlation between pairs of time series from different basins on the distance between the tide gauges

The need to separate the basins becomes even more clear when considering **Figure 7*b*** and ***d***. Here, many tide gauge correlation pairs are weak. This proves that it would be inappropriate to consider the whole North Sea – Baltic Sea transition zone as one singular area of covariance. This will be further demonstrated in the EOF analysis of Section 3.2

It is also important to consider that, since the tide gauge station timeseries have not undergone any tidal corrections, this has most likely influenced the results. In shallow seas, tides can vary over short distances (Robinson, 2010), especially in areas situated closer to the open North Sea such as the Skagerrak and Kattegat. Additionally, previous studies have also found that a filtering effect of SLV signals such as tides occurs through the Belts (Hieronymus et al., 2017), leading this area to be naturally complex. Since areas south of the Belts experience very little tidal variability while areas north of them do, this most likely increases the correlation discrepancies between tide gauges between and even within sub-basins.

It should also be mentioned that it is difficult to establish clear borders between basins of covariance using this method. For instance, the analysis between Skagerrak – Kattegat (**Figure 7a**) have multiple tide gauges between them with very strong positive correlations. This means that it is difficult to sort these tide gauge stations into one basin or the other.

To summarize, the North – Baltic Sea transition zone can be split into 4 separate regions of sea level covariance following the tide gauge correlation analysis. In the following section, I will discuss how these results compare to the satellite altimetry dataset, particularly how this relates to the EOF analysis.

3.2 MAIN DRIVERS AS DETECTED BY STATISTICAL METHODS

The correlation analysis between altimetry-based sea level anomaly and sea level drivers presented in **Figure 8**, shows that zonal winds have by far the strongest connection to sea level anomaly variability in most of the area. This reaffirms what has been found in past research (Hieronymus et al., 2017; Passaro et al., 2021; Sterlini et al., 2016). This is especially true in Kattegat and most of Skagerrak. Slightly lower correlations are present at the deeper waters of the Skagerrak and the onset of the deeper Norwegian Trench. Whereas strong positive correlation exists between zonal winds and sea level anomaly for all areas north of the Danish Straits, only negative correlations are found south of them. In fact, such a divide can be found in many of the correlation analyses made between sea level and background drivers. This includes drivers such as meridional winds, sea level pressure and zonal currents (**Figure 8**). Only correlation coefficients within a 95% confidence interval are included in these figures.

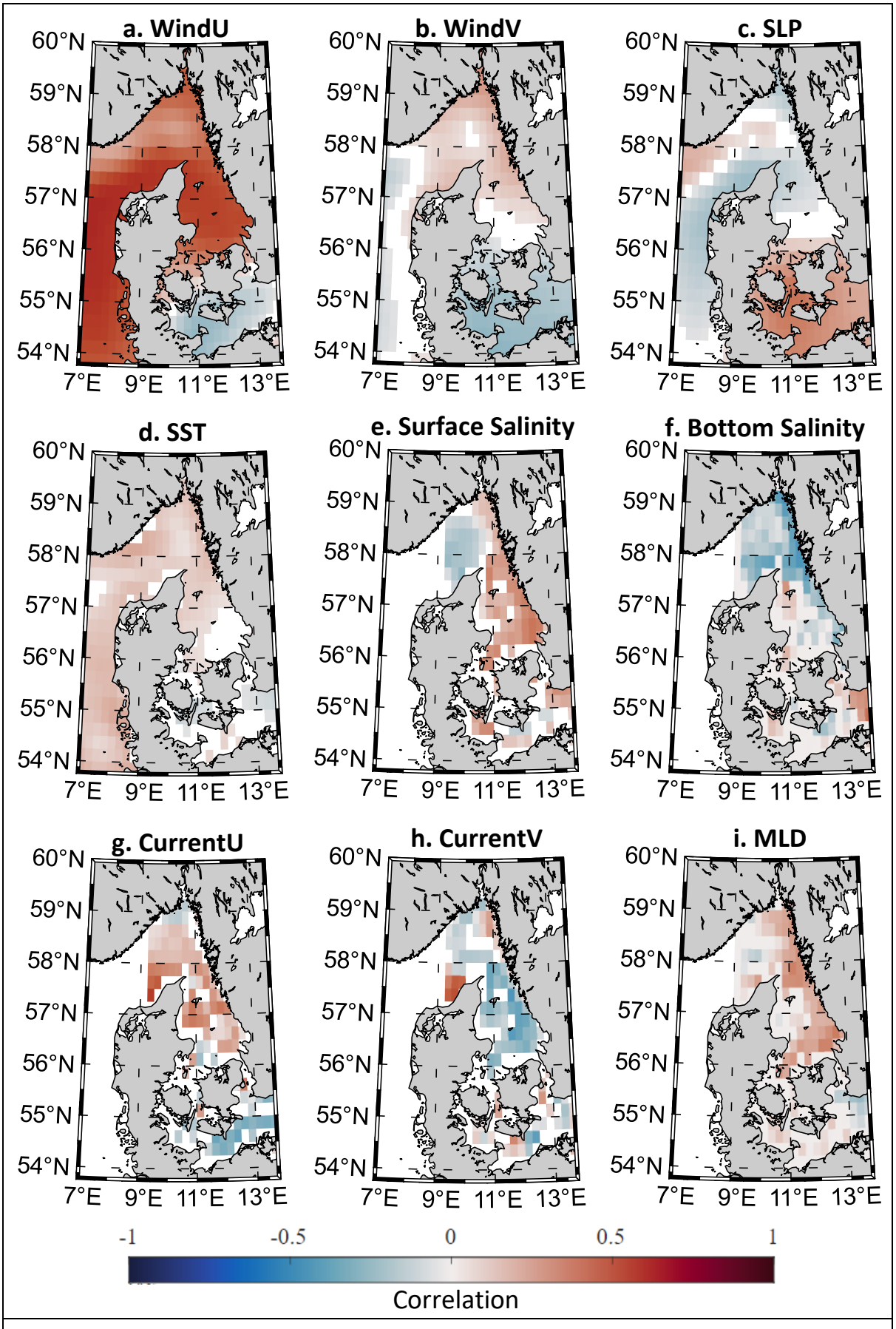


Figure 8: Correlation between SLA and background drivers.

Which of the two wind components is the most significant in controlling local sea levels is largely conditioned by local coastal geometry. As wind blows into and over the coasts, it drags water along with it. As described in Dangendorf et al. (2014a), who investigated the North Sea barotropic response to atmospheric forcings and in Sterlini et al. (2016), who investigated the major drivers of SSV in the Danish North Sea, Ekman transport in the North Sea is generated by eastward positive winds that drive a net movement of water towards the south. Ekman transport from northward positive winds drive a net water transport towards the east. Southward winds drive water transport towards the west, and eastward winds drive water transport towards the north. As the water nears the coastlines, it converges, which leads to an SLA response in the form of higher sea levels. This coastal zone effect between surface winds and sea level anomaly can be seen throughout the world but is especially apparent in the Nordic seas, including the North – Baltic Sea transition zone. In the Skagerrak and Kattegat Seas, higher than usual SLA coincide with eastward wind and to a lesser extent northward wind. Higher than usual SLA in the SW Baltic is mostly associated with westward and southward wind. Major Baltic outflow events are generally caused by weakened or reversed westerly winds that fail to sustain the sea level gradient between the basins. Outflow events can rapidly decrease the mean sea level in the Baltic (Carlsson, 1998). In the Belts, there is little to no correlation between SLA and eastward/westward wind, while there is a larger connection to southward wind, meaning that high SLA is caused by northern wind in this region. This difference between the basins might be the reason for why the tide gauge stations in the Belts did not correlate well with those in either the Kattegat nor SW Baltic.

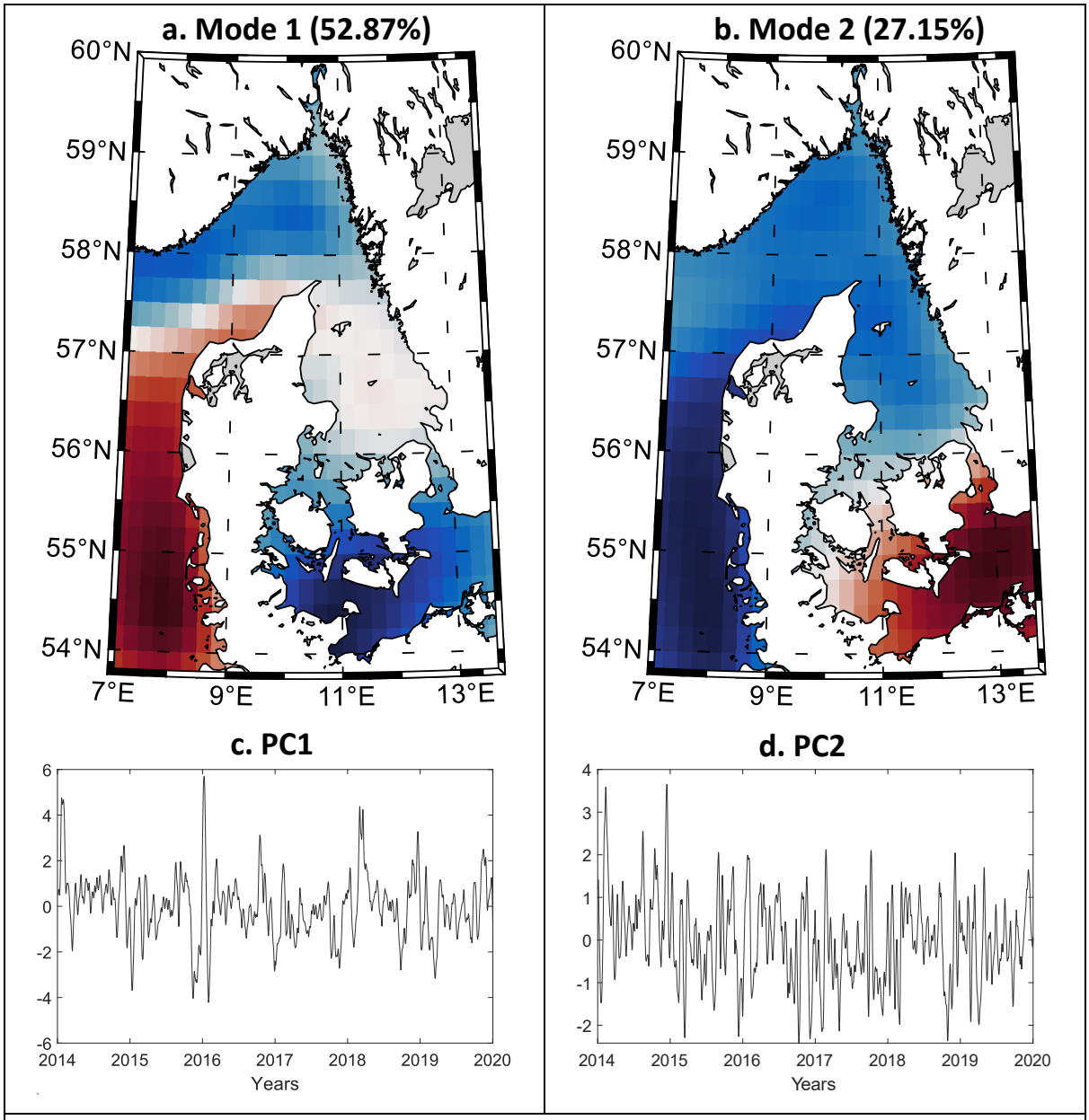


Figure 9: The first two EOF modes of SLA variability (top) and the corresponding PC:s (bottom), with their respective explained variance. Together they explain almost 80% of the variance.

SLA variability in the Belts and SW Baltic have a clear positive correlation to variability in sea level pressure (**Figure 8c**). There also exists a weak positive correlation (0.2) following most of the Norwegian coastline, with weak negative to non-significant correlations in the Kattegat. This result is surprising, as the inverse barometer (IB) effect on daily timescales has not been filtered out of the altimetry dataset. If the SLA response to changes in SLP was only confined through the IB effect, this known linear relationship would lead to negative correlations between SLP and SLA throughout the region. Instead, we see areas that experience higher sea levels in tandem with higher SLP values. This

further suggests that there is another physical process, driven by spatial changes in SLP, and in effect wind, that exerts changes to SLA. It may be that external SLA surges caused by pressure systems elsewhere lead to these results. Such an effect is for instance studied in Wolski et al. (2014), where it was shown that a deep low pressure system passing over central Scandinavia and the north Baltic Sea caused decreased sea levels in the SW Baltic. It is important to consider that the correlation analysis has only been considered between SLA and background forcings at the same location, while it has been proven that remote changes in atmospheric pressure fields and wind movement also influence SLA remotely (Dangendorf et al., 2014a; Sterlini et al., 2016; Wolski et al., 2014). Remote forcings on local sea levels are however not easily distinguishable in this study.

In the Kattegat, there is a salinity gradient of surface waters going north to south, going from 20-25 PSU in the north to 10-12 PSU in the south (Christensen et al., 2018; Stigebrandt, 1983). In most of Kattegat I find a positive correlation between surface salinity and sea surface height (**Figure 8e**). This is most likely also related to the variability of surface winds. Since strong zonal winds in the Kattegat control outflow of freshwater from the Baltic, it should be expected that strong winds also bring saltier water from the North Sea while simultaneously inhibiting freshwater outflow from the Baltic. There is general agreement that during weak or moderate zonal wind conditions, a two-layer halocline system of in-and-outflow occurs through the Danish Straits (Sayin & Krauss, 1996; Weisse et al., 2021). As the westerlies pick up strength, the system shifts to in-or-outflow across the entire water column, the strength of which is determined by the sea level gradient between Skagerrak and SW Baltic (Weisse et al., 2021). Assessing the ocean mixed layer depth can give us a better understanding of these processes. The ocean mixed layer is defined as the layer at the surface of the ocean where most intense mixing takes place, and therefore temperature and salinity (and hence density) are fairly uniform. In this regard, the ocean mixed layer depth is heavily related to the SST, and surface salinity fields. In the Danish Straits for instance, the mixed layer depth is also that of the halocline. The correlation between the mixed layer depth and SLA fields follows a similar spatial structure as to that of the surface salinity. While not as apparent, a similar spatial structure with inversed correlation values can also be seen between mixed layer depth and SLA analysis and the bottom salinity and SLA correlation analysis.

There is a weak positive correlation between local SLA and sea surface temperature across nearly the entire study area. This can be expected as thermal expansion of the water column has been found to be a leading cause of rising sea levels. This result confirms our expectation, however, since the mean depth throughout most of our study area rarely exceeds 30 meters, the effect that the local thermosteric signal has on SLA variability is expected to be weak (Dangendorf et al., 2014a; Woodworth et al., 2007). An exception to this is the Norwegian Trench, where the water reaches depths of >700 meters. Coincidentally, this area also has the strongest correlation between SLA and SST variability. Both Bingham and Hughes (2012) and Chen et al. (2014) suggest that, while shallow North Sea waters are unable to produce a steric SLA signal, coastal SLA response to SST originate from deeper oceanic waters, which then propagate along the coastlines.

It should also be noted that most analyses between SLA variability and background drivers presented in **Figure 8** contain many areas that do not have a significant correlation on a 95% confidence level. Note that missing datapoints west of 9° East in **Figure 8e, f, g, h** and **i** is due to spatial limitations of the dataset and not insignificant correlations.

The spatial divides previously discussed are also visible in the EOF analysis presented in **Figure 9**, most notable in the second EOF map. Here the phase opposition of sea level variability is strong on opposite sides of the Danish Straits, with little to no variability associated with this second principal component in the Belts sub-region. This is also visible in the first EOF albeit somewhat weaker. While the Kattegat is unaffected by the first EOF, there is a strong signal around and south of the Danish Straits.

It is not surprising that there is a common spatial structure. Wind patterns are largely driven by air pressure fields, that further drive the circulation of surface currents through wind stress. As such, zonal and meridional winds are coupled and follow pressure distributions in the atmosphere. Many researchers therefore argue that large-scale atmospheric modes, in this region predominantly the North Atlantic Oscillation (NAO), can be effectively used as an atmospheric proxy, which combines coupled processes such as the IB effect, wind stress, air pressure variations and precipitation (Dangendorf et al., 2014b; Hurrell, 1995; Suursaar & Sooäär, 2007). While the changes to the NAO are mostly

related to long-term decadal variability that is not entirely captured in this study, the connection still exists.

To examine which of the forcings best correspond to the PC variability, I also show results from the correlation analysis between the PCs and sea level drivers in **Figure 10a** and **b**. I find that the first PC is strongly linked to zonal winds throughout the study area (average R = 0.56), as it is well described by the local zonal wind component, while the second PC is best described by the SLP (average R = 0.38). In comparison, PC1 and SLP have an average correlation of 0.15, PC2 and zonal winds had an average correlation of 0.16. Meridional winds have average correlations to PC1 and PC2 of 0.11 and 0.21 respectively. Other results can be viewed in **Appendix B**.

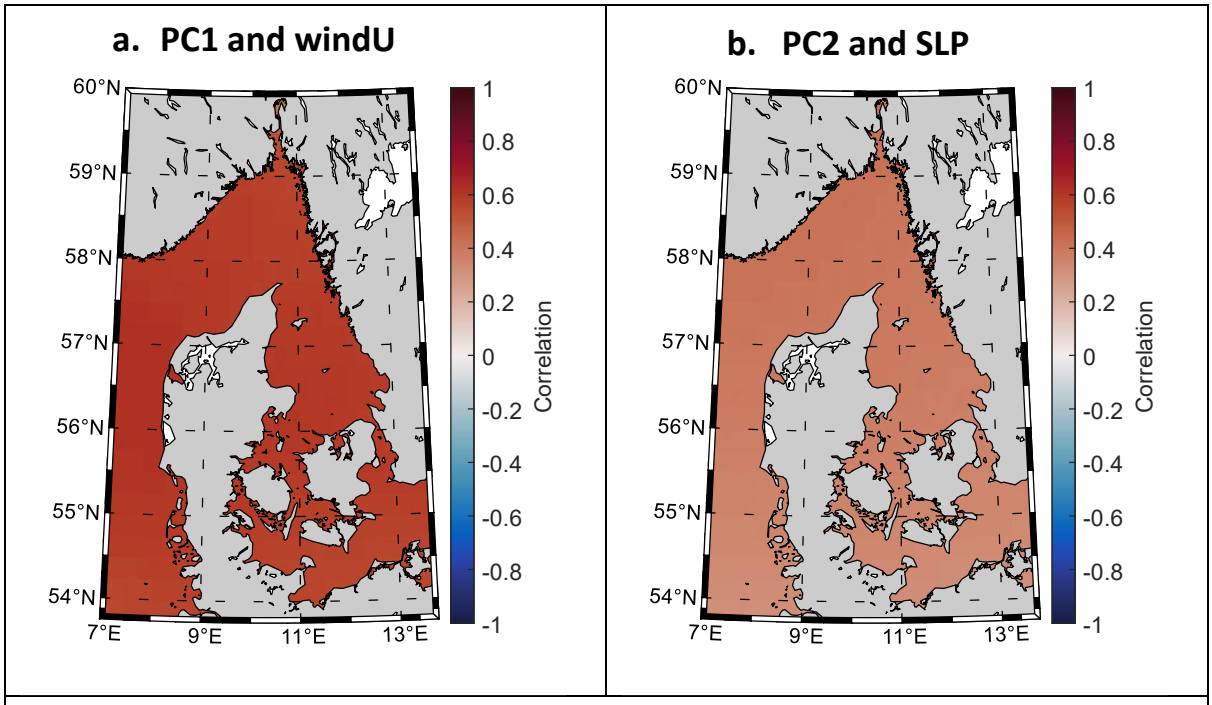


Figure 10: Correlation analysis between (a) PC1 and zonal (u) wind component, and (b) PC2 and sea level pressure. Both sea level drivers were the highest correlated ones to each PC.

To conclude this section, it is clear through our results and what has been discussed by past research (e.g. Slangen et al., 2014b), that the short-term variability of sea level anomaly is strongly dominated by changes in wind stress, particularly to changes of the zonal wind component, but also to changes in the atmospheric pressure field.

3.3 MAIN DRIVERS AS DETECTED BY MACHINE LEARNING

The similar spatial structures between SLA and background variables described in the previous section suggests the existence of significant cross-correlations between forcings, also known as collinearity. Within a machine learning approach, collinearity between input variables, which stems from the variables not being entirely independent from one another, can muddle the results. **Table 3** shows the complete cross-correlation matrix between background drivers, as well as the correlations between the SLA and the drivers for the two locations defined in **Figure 1**. In the Kattegat location, SLA shows highest correlation with the zonal wind. Zonal wind is also moderately well correlated with the zonal current and surface salinity, and have a strong negative correlation with meridional currents. In the SW Baltic location, the relationship between 10 m winds and

Table 3: To the left is displayed the correlation coefficients between the full SLA timeseries and the background drivers at the two locations. These values are not considered when evaluating collinearity between the background variables. To the right, cross-correlation between SLA and background drivers at a Kattegat and SW Baltic location (for exact locations, see Figure 1). Where the correlation between two variables exceed ±0.4, they are shaded red. Values are marked as NS if the correlation is non-significant at a 95% confidence level.

Kattegat

	SLA			u10	v10	SLP	SST	SS	BS	CurrU	CurrV
u10	0.53		u10	1.00							
v10	0.20		v10	0.08	1.00						
SLP	-0.11		SLP	-0.27	-0.21	1.00					
SST	0.12		SST	NS	NS	NS	1.00				
SS	0.28		SS	0.37	NS	-0.24	NS	1.00			
BS	-0.35		BS	-0.18	NS	NS	NS	-0.34	1.00		
CurrU	0.18		CurrU	0.33	0.12	-0.16	NS	0.24	0.04	1.00	
CurrV	-0.20		CurrV	-0.49	0.28	0.20	NS	-0.19	-0.11	-0.28	1.00
MLD	0.26		MLD	0.28	0.08	-0.18	0.06	0.42	-0.19	0.05	-0.19

SW Baltic

	SLA			u10	v10	SLP	SST	SS	BS	CurrU	CurrV
u10	-0.11		u10	1.00							
v10	-0.18		v10	NS	1.00						
SLP	0.25		SLP	-0.27	-0.25	1.00					
SST	-0.08		SST	-0.08	NS	0.10	1.00				
SS	0.16		SS	0.13	0.05	-0.13	-0.07	1.00			
BS	0.23		BS	0.22	NS	-0.14	NS	0.52	1.00		
CurrU	NS		CurrU	0.43	0.28	-0.20	NS	0.09	NS	1.00	
CurrV	NS		CurrV	-0.46	0.42	NS	0.08	-0.08	-0.12	NS	1.00
MLD	-0.05		MLD	-0.16	-0.16	NS	-0.12	NS	-0.37	-0.07	-0.10

surface currents is even stronger. Moderate negative correlations are also found between sea level pressure and zonal winds. In terms of SLA, the strongest correlations are found with SLP.

Sterlini et al. (2016), who undertook a multiple linear regression approach to model sea levels, had used a threshold limit of ±0.35 when comparing and excluding sea level drivers in their analysis. I however choose to be slightly more lenient, and incorporate a threshold limit of ±0.4. Such high values are for instance found between the wind and surface current, meaning that much of the variability displayed in the surface current signal is captured by the wind signal. Since currents are primarily wind driven, and not the other way around, I choose to remove both the zonal and meridional current from both locations for the machine learning analysis. Correlations above my threshold are present between surface and bottom salinity in the SW Baltic and surface salinity and MLD in the Kattegat. Since surface salinity also acts as a proxy for precipitation and evaporation, I choose to keep this forcing and instead remove bottom salinity and MLD. The two variables are removed from both locations for consistency.

As seen in **Figure 11**, there exists a large signal delay between the zonal wind component and SLA in the SW Baltic sub-basin. The signal delay is calculated using cross-correlation techniques to determine the optimum data shift between the two variables. In the SW Baltic, the optimum correlation can be found by shifting the SLA variable back, in some regions up to 15 days. This is related to the results presented in **Figure 12**, which shows the differences in RMSE between different model runs using variable sequence length of sea level drivers for the two locations. There is a stark contrast between the two locations. At the Kattegat location, the optimal sequence length is 6 days, while in the SW Baltic location, it is 18 days. While the results in **Figure 11** are the cause for the large contrast between the areas in regard to the optimal sequence length for the RNN, there exist no clear explanation as to why the signal delay between zonal wind forcing and the SLA response is so large. It is possible that the natural complexity in this region plays a large role, but also that there might be other variables that I have not considered. Samuelsson and Stigebrandt (1996) mentioned that SW Baltic variability is influenced by sea levels in the Kattegat, so by including more possible background forcings, in particular remote ones, the sequence length may be shortened.

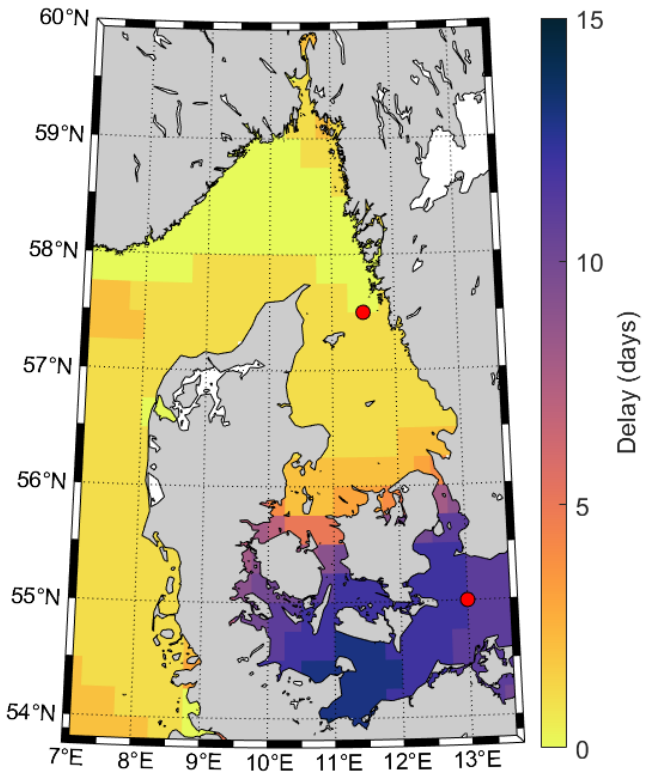


Figure 11: Lag (days) yielding the maximum correlation in absolute values between the zonal (u) wind signal and SLA. The values show by how many days the SLA signal is delayed in reference to the u wind. The two red dots are the location of the Kattegat and SW Baltic data extraction points.

Generally, the neural network is better at predicting the sea level at the Kattegat location (best RMSE = 2 cm, **Figure 12a**) compared to the SW Baltic location (best RMSE = 4 cm, **Figure 12b**). It also requires a shorter sequence length to produce the better results. The model does not seem to perform better after a sequence length of circa 18. Since longer sequence lengths make the model run slower, I choose this value for the forcing experiment displayed in **Figure 13**, rather than a higher value of e.g., 30. This indicates that SW Baltic SLV is not as easily described as SLV in the Kattegat Sea.

Figure 13 present the results of the second machine learning experiment. The sea level drivers are ordered by median RMSE values and follow as such a ranking system of Any sea level drivers below the "All forcings" experiment are those whose exclusion improved

the prediction. In the SW Baltic, the model did not perform better when excluding any of the background drivers. In the Kattegat, only when excluding SST did the model run better on average.

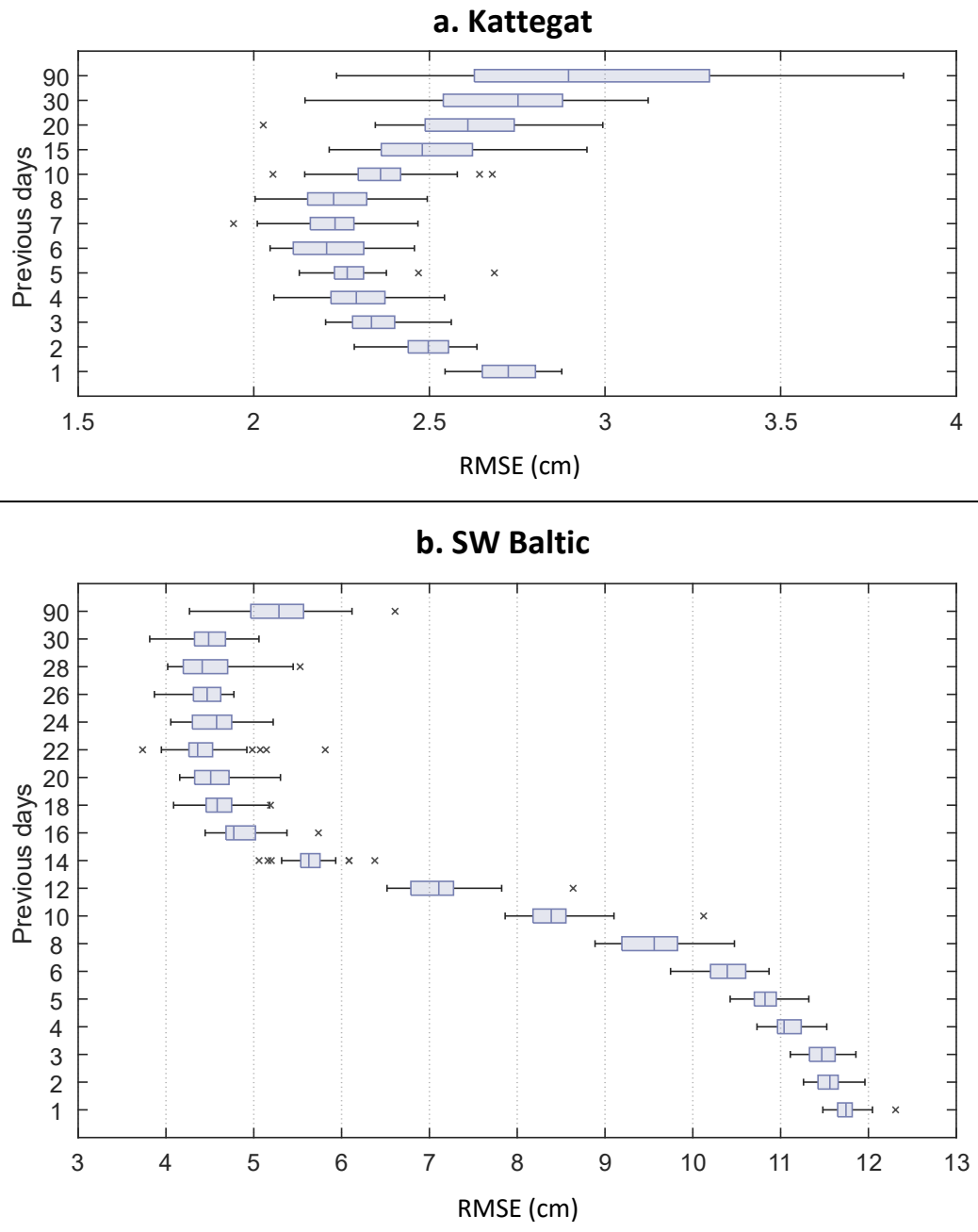


Figure 12: Boxplots visualizing the results from 30 model runs featuring all forcings at the (a) Kattegat location and (b) SW Baltic location. Root-mean-square error depends on the number of past days the model uses to predict.

It is clear that the zonal wind component serves as the most dominant background driver in both locations. In the Kattegat, the RMSE increases by almost 1.5 cm compared to the base run when excluding zonal winds from the predictions. In the SW Baltic, the RMSE

increases by nearly 8 cm when excluding zonal winds from the predictions, which is a significantly worse performance. The second most important driver in the SW Baltic is meridional winds, where the RMSE increases by roughly 1 cm compared to the base run. In the Kattegat, sea level pressure is found to be the second most important driver, where the RMSE increases by 0.5 cm when excluded. This confirms our results from the statistical analysis and agrees well with other research that find that atmospheric drivers are those most responsible for driving local high frequency sea level variability in the North – Baltic Sea transition zone.

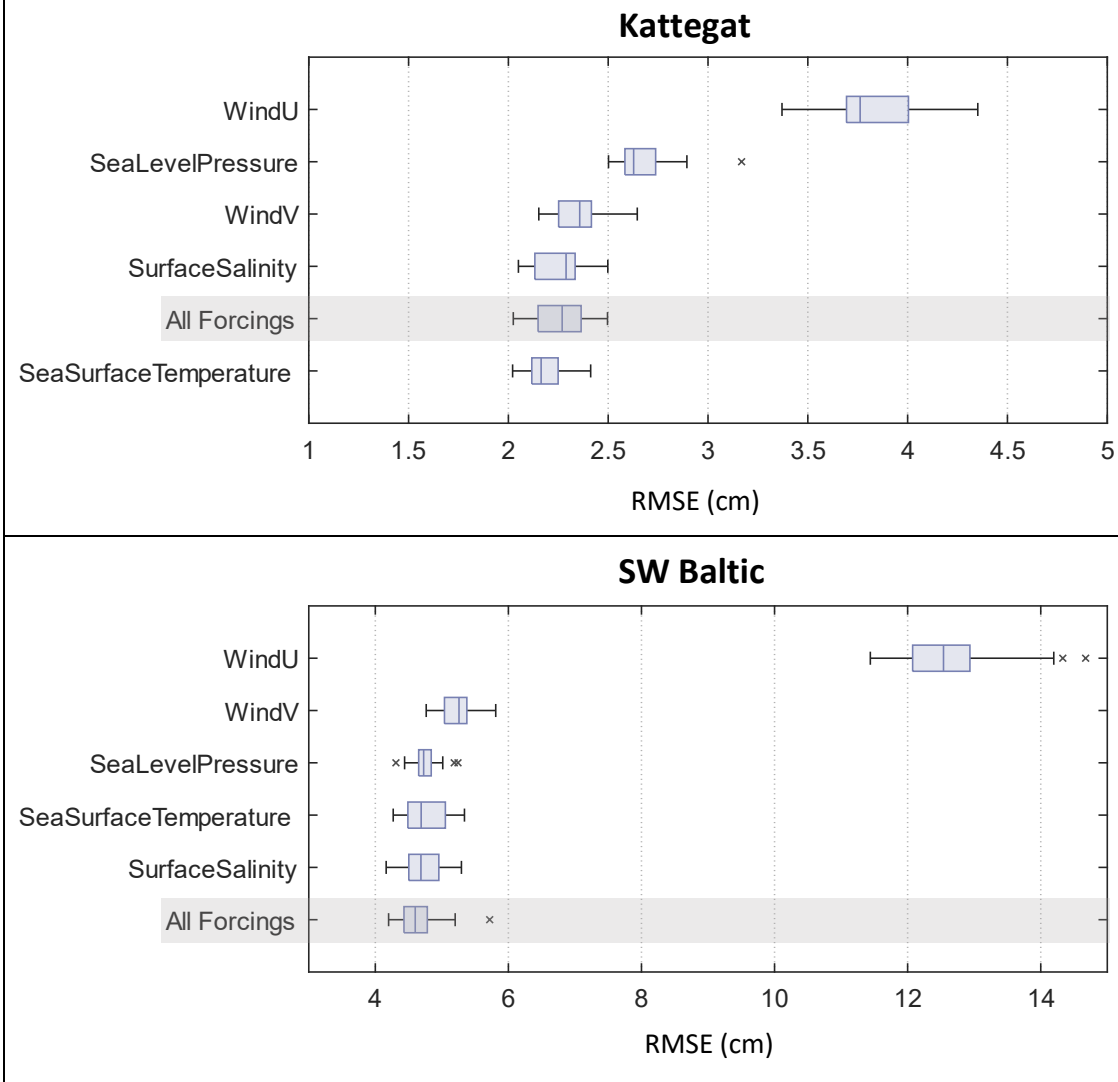


Figure 13: Forcing experiment at the Kattegat location. Each boxplot visualizes the results from 30 model runs featuring all forcings excluding the one named. The highlighted boxplot is the base run that includes all forcings.

The difference in RMSE between the other sea level drivers are small, and it is difficult to say that one is more important than the other since the difference in RMSE is minimal. Such a small change may well be due to the random initial weights of the neural network at the start of the training phase. However, some points can still be made. For instance, sea surface temperature is the least important driver for the Kattegat location, and its exclusion even slightly improves predictions. While a known relationship exists between thermal expansion and sea level rise, the effect that SST has on sea level variability on daily timescales is small. Surface salinity has a negative effect on predictions when excluded from the model at both locations. Surface salinity does not only give an approximation of precipitation and river run-off, both which would increase local sea levels, but may also act as a secondary proxy to the wind stress signal since salinity is influenced by wind-driven North Sea inflow of surface waters (Hordoir et al., 2013). In the Baltic especially, salinity intrusions from the Kattegat are forced by strong westerlies, and in the Kattegat, freshwater inflow from the Baltic is influenced by weak westerlies.

Regarding the neural network performing consistently much worse at the SW Baltic location, this may suggest that other sea level drivers ought to be considered that I have currently left out. It is possible that this location in the SW Baltic is more strongly influenced by remote drivers, say winds in the Kattegat or freshwater discharge from streams and tributaries throughout the entire Baltic proper.

Finally, **Figure 14** displays the results from the base "All forcings" predictions from the two locations. In both locations, the models often overestimate sea levels, yet both still fail to accurately predict the observed sea level peak at both locations at the end of September and in early October 2018, suggesting that this extreme was caused by a mechanism not represented in the training data. In the context of protecting coastlines and planning preemptive measures in urban environments against rising sea levels, a slight overestimation is better than an underestimation. It is however much more vital to also be able to capture and predict the extreme sea level events, as these are the ones that will be the costliest. In Kattegat (**Figure 14a**) only during a period between May 2018 – June 2018 do most prediction seem to instead underestimate sea levels. In SW Baltic (**Figure 14b**), large discrepancies can be seen between September 2018 – October 2018 and around January 2019.

The range of observed values is also larger in the SW Baltic (-0.3 to +0.28 m) compared to the range of observed values in Kattegat (-0.2 to +0.15 m) during this period. It is possible that the larger range in values contributed to the model prediction having larger RMSE at the SW Baltic location.

Despite these issues, the model performs well in predicting sea levels. The neural network analysis of the sea level drivers also performs well and confirms the conclusions about the sea levels primary drivers reached by the traditional statistical approach.

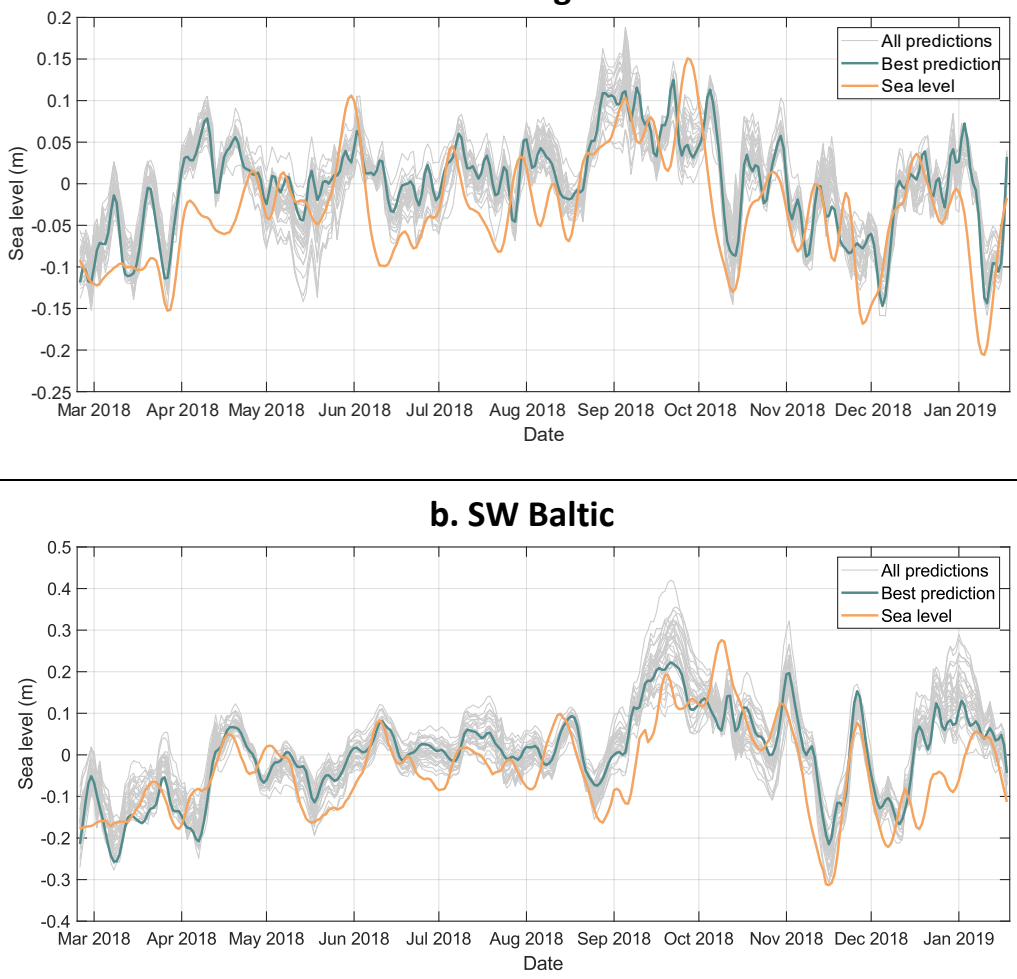


Figure 14: Results from the base run including all forcings model prediction at the (a) Kattegat location and (b) SW Baltic location. The green line shows the best prediction with the lowest RMSE, the light grey show all other predictions from the same model and the yellow line show the observed SLA values. Each model is run 30 times.

4 CONCLUSION

The results from this thesis directly contribute to the project NEEDS. Mapping of the regions of coherent sea level variability is used in the project to conduct separate experiments in each region to determine region-specific drivers of sea level variability.

Through the work conducted in this thesis, I have successfully defined and mapped regions of coherent sea level variability in the North – Baltic Sea transition zone. By analyzing the tide gauge timeseries, the area can be effectively divided into four separate subbasins of covariance: Skagerrak, Kattegat, South-West Baltic and the Belts. The correlations between the pairs of stations in each basin are very strong, ranging from the minimum of 0.75 in the Kattegat to the maximum of 0.99 in Skagerrak. The division between areas of covariance is a necessary procedure that enabled me to better understand each subbasins background sea level drivers on high frequency sea level variability.

Results show that the zonal wind component is the main driver of sea level variability in most of the study area, be it through strong predominant westerlies or through their weakened or even reversed state. It has a large positive correlation to sea level anomaly north of the Belts and a negative correlation south of them. It is closely correlated to the first principal component that explains more than 50% of the sea level variability. The meridional wind component also has some influence, though much smaller. Which wind is most responsible at each location is mostly controlled by local coastal geometry and changes significantly in the relatively small area due to the complexity of the coast. Sea level pressure is also found to be an important driver in sea level anomaly variability and is most correlated to the second principal component.

When predicting sea level using a neural network, I achieve good model results using only a handful of forcings. The information about the dominant drivers of sea level variability determined here will be used to build new machine learning models that will quantify the flood risk in Sweden and Denmark for the next 30 years based on different climate model predictions. The forcings used were zonal wind, meridional wind, sea level pressure, sea surface temperature and surface salinity. The main sea level drivers, i.e., the drivers whose exclusion resulted in significantly worse prediction are zonal winds, meridional winds (for both Kattegat and the South-West Baltic) and sea level pressure (only in the

Kattegat). On the other hand, the inclusion of surface salinity or sea surface temperature did not improve model results.

I also showed that the model performed consistently worse at predicting sea levels in the South-West Baltic than it did in the Kattegat. This may be a result of not including enough forcings as input variables. As has been suggested by other research, Baltic sea level variability is most likely also influenced by remote atmospheric forcings outside my study area.

I will finally make some remarks about potential improvements to the thesis and further research opportunities. First, to fully utilize the benefits of satellite data coverage, I would have liked to calculate spatial averages of both SLA and drivers across all four sub-basins delimited by the tide gauge analysis. This might give a more definitive answer to each sub-basins primary forcings, as well as be more representative of the sub-basin's variability compared to a single location. I showed that using a machine learning approach to determine primary sea level drivers works well, and that it is comparable to that of traditional statistical approach. However, to first compute PCA/EOF analysis for each sub-basin, and then attempt to predict and find primary forcings for each PC separately may give even more substantial results. Lastly, it would be most interesting to also include remote drivers in both the statistical and machine learning analysis (e.g. to use zonal winds in the Kattegat Sea to determine South-West Baltic sea level variability, or river run-off from nearby land catchments), a subject which this report is admittedly lacking.

5 REFERENCES

Bingham, R. J., & Hughes, C. W. (2012). Local diagnostics to estimate density-induced sea level variations over topography and along coastlines. *Journal of Geophysical Research: Oceans*, *117*(C1). https://doi.org/https://doi.org/10.1029/2011JC007276

Carlsson, M. (1998). A coupled three-basin sea level model for the Baltic Sea. *Continental shelf research*, *18*(9), 1015-1038. https://doi.org/https://doi.org/10.1016/S0278-4343(98)00025-9

Cazenave, A., & Llovel, W. (2010). Contemporary Sea Level Rise. *Annual Review of Marine Science*, *2*(1), 145-173. https://doi.org/10.1146/annurev-marine-120308-081105

Chen, X., Dangendorf, S., Narayan, N., O'Driscoll, K., Tsimplis, M. N., Su, J., Mayer, B., & Pohlmann, T. (2014). On sea level change in the North Sea influenced by the North Atlantic Oscillation: Local and remote steric effects. *Estuarine, coastal and shelf science*, *151*(C), 186-195. https://doi.org/10.1016/j.ecss.2014.10.009

Christensen, K. H., Sperrevik, A. K., & Broström, G. (2018). On the Variability in the Onset of the Norwegian Coastal Current. *Journal Of Physical Oceanography*, *48*(3), 723-738. https://doi.org/10.1175/jpo-d-17-0117.1

Church, J. A., P.U. Clark, A. Cazenave, J.M. Gregory, S. Jevrejeva, A. Levermann, M.A. Merrifield, G.A. Milne, R.S. Nerem, P.D. Nunn, A.J. Payne, W.T. Pfeffer, D. Stammer, & Unnikrishnan, A. S. (2013). Sea Level Change. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.*

Cipollini, P., Benveniste, J., Birol, F., Fernandes, J., Passaro, M., & Vignudelli, S. (2017). Satellite Altimetry in Coastal Regions. In. https://doi.org/10.1201/9781315151779-11

Dangendorf, S., Calafat, F. M., Arns, A., Wahl, T., Haigh, I. D., & Jensen, J. (2014a). Mean sea level variability in the North Sea: Processes and implications. *Journal of Geophysical Research: Oceans*, *119*(10). https://agupubs.onlinelibrary.wiley.com/doi/pdfdirect/10.1002/2014JC009901?download=true

Dangendorf, S., Wahl, T., Nilson, E., Klein, B., & Jensen, J. (2014b). A new atmospheric proxy for sea level variability in the southeastern North Sea: observations and future ensemble projections. *Climate Dynamics*, *43*(1), 447-467. https://doi.org/10.1007/s00382-013-1932-4

European Space Agency. (2022). *Sentinel User Guide*. https://sentinels.copernicus.eu/web/sentinel/user-guides

Fernandes, M. J., Lázaro, C., Nunes, A. L., & Scharroo, R. (2014). Atmospheric Corrections for Altimetry Studies over Inland Water. *Remote Sensing*, *6*(6), 4952-4997. https://www.mdpi.com/2072-4292/6/6/4952

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. https://www.deeplearningbook.org/

Greene, C. A., Kaustubh Thirumalai, Kelly A. Kearney, José Miguel Delgado, Wolfgang Schwanghart, Natalie S. Wolfenbarger, Kristen M. Thyng, David E. Gwyther, Alex S. Gardner, & Blankenship., D. D. (2021). *The Climate Data Toolbox for MATLAB.* In Geochemistry, Geophysics, Geosystems 2019.

Gustafsson, B. G., & Andersson, H. C. (2001). Modeling the exchange of the Baltic Sea from the meridional atmospheric pressure difference across the North Sea. *Journal of Geophysical Research: Oceans*, *106*(C9), 19731-19744. https://doi.org/https://doi.org/10.1029/2000JC000593

H.-O. Pörtner, D.C. Roberts, M. Tignor, E.S. Poloczanska, K. Mintenbeck, A. Alegría, M. Craig, S. Langsdorf, S. Löschke, V. Möller, A. Okem, & Rama, B. (2022). IPCC, 2022: Climate Change 2022: Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Sixth

Assessment Report of the Intergovernmental Panel on Climate Change. *Cambridge University Press.*

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., & Thépaut, J.-N. (2018). *ERA5 hourly data on single levels from 1979 to present* Copernicus Climate Change Service (C3S) Climate Data Store (CDS). . https://doi.org/10.24381/cds.adbb2d47

Hieronymus, M., Hieronymus, J., & Arneborg, L. (2017). Sea level modelling in the Baltic and the North Sea: The respective role of different parts of the forcing. *Ocean modelling (Oxford)*, *118*, 59-72. https://doi.org/10.1016/j.ocemod.2017.08.007

Holgate, S. J., Matthews, A., Woodworth, P. L., Rickards, L. J., Tamisiea, M. E., Bradshaw, E., Foden, P. R., Gordon, K. M., Jevrejeva, S., & Pugh, J. (2012). New Data Systems and Products at the Permanent Service for Mean Sea Level. *Journal of Coastal Research*, *29*(3), 493-504. https://doi.org/10.2112/jcoastres-d-12-00175.1

Hordoir, R., Axell, L., Höglund, A., Dieterich, C., Fransner, F., Gröger, M., Liu, Y., Pemberton, P., Schimanke, S., Andersson, H., Ljungemyr, P., Nygren, P., Falahat, S., Nord, A., Jönsson, A., Lake, I., Döös, K., Hieronymus, M., Dietze, H., . . . Haapala, J. (2019). Nemo-Nordic 1.0: a NEMO-based ocean model for the Baltic and North seas – research and operational applications. *Geosci. Model Dev.*, *12*(1), 363-386. https://doi.org/10.5194/gmd-12-363-2019

Hordoir, R., Dieterich, C., Basu, C., Dietze, H., & Meier, H. E. M. (2013). Freshwater outflow of the Baltic Sea and transport in the Norwegian current: A statistical correlation analysis based on a numerical experiment. *Continental shelf research*, *64*, 1-9. https://doi.org/https://doi.org/10.1016/j.csr.2013.05.006

Hordoir, R., & Meier, H. E. M. (2010). Freshwater fluxes in the Baltic Sea: A model study. *Journal of Geophysical Research: Oceans*, *115*(C8). https://doi.org/https://doi.org/10.1029/2009JC005604

Hurrell, J. W. (1995). Decadal Trends in the North Atlantic Oscillation: Regional Temperatures and Precipitation. *Science*, *269*(5224), 676-679. https://doi.org/doi:10.1126/science.269.5224.676

Lass, H. U., & Mohrholz, V. (2003). On dynamics and mixing of inflowing saltwater in the Arkona Sea. *Journal of Geophysical Research: Oceans*, *108*(C2). https://doi.org/https://doi.org/10.1029/2002JC001465

Llovel, W., & Lee, T. (2015). Importance and origin of halosteric contribution to sea level change in the southeast Indian Ocean during 2005–2013. *Geophysical Research Letters*, *42*(4), 1148-1157. https://doi.org/https://doi.org/10.1002/2014GL062611

Madsen, K. S., Høyer, J. L., & Tscherning, C. C. (2007). Near-coastal satellite altimetry: Sea surface height variability in the North Sea–Baltic Sea area. *Geophysical Research Letters*, *34*(14). https://doi.org/https://doi.org/10.1029/2007GL029965

Meyssignac, B., Piecuch, C. G., Merchant, C. J., Racault, M. F., Palanisamy, H., MacIntosh, C., Sathyendranath, S., & Brewin, R. (2017). Causes of the Regional Variability in Observed Sea Level, Sea Surface Temperature and Ocean Colour Over the Period 1993–2011. *Surveys in Geophysics*, *38*(1), 187-215. https://doi.org/10.1007/s10712-016-9383-1

Mitrovica, J. X., Hay, C. C., Kopp, R. E., Harig, C., & Latychev, K. (2018). Quantifying the Sensitivity of Sea Level Change in Coastal Localities to the Geometry of Polar Ice Mass Flux. *Journal of Climate*, *31*(9), 3701-3709. https://doi.org/10.1175/jcli-d-17-0465.1

NOAA. (2021). What is a tide gauge? https://oceanservice.noaa.gov/facts/tide-gauge.html

O'Malley, T., Bursztein, E., Long, J., Chollet, F., Jin, H., & Invernizzi, L. (2019). *KerasTuner*. In https://github.com/keras-team/keras-tuner

Oppenheimer, M., B.C. Glavovic , J. Hinkel, R. van de Wal, A.K. Magnan, A. Abd-Elgawad, R. Cai, M. Cifuentes-Jara, R.M. DeConto, T. Ghosh, J. Hay, F. Isla, B. Marzeion, B. Meyssignac, & Sebesvar, Z. (2019). Sea Level Rise and Implications for Low-Lying Islands, Coasts and Communities. *IPCC Special Report on the Ocean and Cryosphere in a Changing Climate [H.-O.*

Pörtner, D.C. Roberts, V. Masson-Delmotte, P. Zhai, M. Tignor, E. Poloczanska, K. Mintenbeck, A. Alegría, M. Nicolai, A. Okem, J. Petzold, B. Rama, N.M. Weyer (eds.)]. In press.

Passaro, M., Cipollini, P., & Benveniste, J. (2015). Annual sea level variability of the coastal ocean: The Baltic Sea-North Sea transition zone. *Journal of Geophysical Research: Oceans*, *120*(4), 3061-3078. https://doi.org/10.1002/2014jc010510

Passaro, M., Müller, F. L., Oelsmann, J., Rautiainen, L., Dettmering, D., Hart-Davis, M. G., Abulaitijiang, A., Andersen, O. B., Høyer, J. L., Madsen, K. S., Ringgaard, I. M., Särkkä, J., Scarrott, R., Schwatke, C., Seitz, F., Tuomi, L., Restano, M., & Benveniste, J. (2021). Absolute Baltic Sea Level Trends in the Satellite Altimetry Era: A Revisit [Original Research]. *Frontiers in Marine Science*, *8*(546). https://doi.org/10.3389/fmars.2021.647607

Pawlowicz, R. (2020). *M_Map: A mapping package for MATLAB*. In (Version 1.4m) www.eoas.ubc.ca/~rich/map.html

PSMSL. (2022). *Tide Gauge Data*. http://www.psmsl.org/data/obtaining/

Pujol, M. I., & Mertz, F. (2020). Product User Manual For Sea Level SLA products SEALEVEL_*_PHY_L[3/4]_[NRT/REP]_OBSERVATIONS_008_0*. (3.1).

Robinson, I. S. (2010). *Discovering the Ocean from Space: The Unique Applications of Satellite Oceanography* (P. Blondel, Ed.). Springer Heidelberg Dordrecht London New York. https://doi.org/10.1007/978-3-540-68322-3

Roden, G. I., & Rossby, H. T. (1999). Early Swedish Contribution to Oceanography: Nils Gissler (1715–71) and the Inverted Barometer Effect. *Bulletin of the American Meteorological Society*, *80*(4), 675-682. https://doi.org/10.1175/1520-0477(1999)080<0675:Escton>2.0.Co;2

Samuelsson, M., & Stigebrandt, A. (1996). Main characteristics of the long-term sea level variability in the Baltic sea. *Tellus A: Dynamic Meteorology and Oceanography*, *48*(5), 672-683. https://doi.org/10.3402/tellusa.v48i5.12165

Sayin, E., & Krauss, W. (1996). A numerical study of the water exchange through the Danish Straits. *Tellus A: Dynamic Meteorology and Oceanography*, *48*(2), 324-341. https://doi.org/10.3402/tellusa.v48i2.12063

Sherstinsky, A. (2020). Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. *Physica D: Nonlinear Phenomena*, *404*, 132306. https://doi.org/https://doi.org/10.1016/j.physd.2019.132306

Slangen, A. B. A., Carson, M., Katsman, C. A., van de Wal, R. S. W., Köhl, A., Vermeersen, L. L. A., & Stammer, D. (2014a). Projecting twenty-first century regional sea-level changes. *Climatic Change*, *124*(1), 317-332. https://doi.org/10.1007/s10584-014-1080-9

Slangen, A. B. A., van de Wal, R. S. W., Wada, Y., & Vermeersen, L. L. A. (2014b). Comparing tide gauge observations to regional patterns of sea-level change (1961–2003). *Earth Syst. Dynam.*, *5*(1), 243-255. https://doi.org/10.5194/esd-5-243-2014

SMHI. (2014a). *Air pressure and sea level*. https://www.smhi.se/en/theme/air-pressure-and-sea-level-1.12266

SMHI. (2014b). *Surface currents*. https://www.smhi.se/en/theme/surface-currents-1.12286

SMHI. (2021). *Mätning av vattenstånd*. SMHI. https://www.smhi.se/kunskapsbanken/oceanografi/oceanografiska-matningar/vattenstandsstationer-1.132796

Stammer, D., Cazenave, A., Ponte, R. M., & Tamisiea, M. E. (2013). Causes for Contemporary Regional Sea Level Changes. *Annual Review of Marine Science*, *5*(1), 21-46. https://doi.org/10.1146/annurev-marine-121211-172406

Sterlini, P., de Vries, H., & Katsman, C. (2016). Sea surface height variability in the North East Atlantic from satellite altimetry. *Climate Dynamics*, *47*(3), 1285-1302. https://doi.org/10.1007/s00382-015-2901-x

Sterlini, P., Le Bars, D., de Vries, H., & Ridder, N. (2017). Understanding the spatial variation of sea level rise in the North Sea using satellite altimetry. *Journal of Geophysical Research: Oceans*, *122*(8), 6498-6511. https://doi.org/https://doi.org/10.1002/2017JC012907

Stigebrandt, A. (1983). A Model for the Exchange of Water and Salt Between the Baltic and the Skagerrak. *Journal Of Physical Oceanography*, *13*(3), 411-427. https://doi.org/10.1175/1520-0485(1983)013<0411:Amfteo>2.0.Co;2

Suursaar, Ü., & Sooäär, J. (2007). Decadal variations in mean and extreme sea level values along the Estonian coast of the Baltic Sea. *Tellus A: Dynamic Meteorology and Oceanography*, *59*(2), 249-260. https://doi.org/10.1111/j.1600-0870.2006.00220.x

Vestøl, O., Ågren, J., Steffen, H., Kierulf, H., & Tarasov, L. (2019). NKG2016LU: a new land uplift model for Fennoscandia and the Baltic Region. *Journal of Geodesy*, *93*(9), 1759-1779. https://doi.org/10.1007/s00190-019-01280-8

WCRP, G. S. L. B. G. (2018). Global sea-level budget 1993–present. *Earth Syst. Sci. Data*, *10*(3), 1551-1590. https://doi.org/10.5194/essd-10-1551-2018

Weisse, R., Dailidienė, I., Hünicke, B., Kahma, K., Madsen, K., Omstedt, A., Parnell, K., Schöne, T., Soomere, T., Zhang, W., & Zorita, E. (2021). Sea level dynamics and coastal erosion in the Baltic Sea region. *Earth Syst. Dynam.*, *12*(3), 871-898. https://doi.org/10.5194/esd-12-871-2021

Wilks, D. S. (2006). *Statistical Methods in the Atmospheric Sciences* (Vol. 2). Elsevier Inc.

Wolski, T., Wiśniewski, B., Giza, A., Kowalewska-Kalkowska, H., Boman, H., Grabbi-Kaiv, S., Hammarklint, T., Holfort, J., & Lydeikaitė, Ž. (2014). Extreme sea levels at selected stations on the Baltic Sea coast**This work was financed by the Polish National Centre for Science research project No. 2011/01/B/ST10/06470. *Oceanologia*, *56*(2), 259-290. https://doi.org/https://doi.org/10.5697/oc.56-2.259

Woodworth, P. L., Flather, R. A., Williams, J. A., Wakelin, S. L., & Jevrejeva, S. (2007). The dependence of UK extreme sea levels and storm surges on the North Atlantic Oscillation. *Continental shelf research*, *27*(7), 935-946. https://doi.org/https://doi.org/10.1016/j.csr.2006.12.007

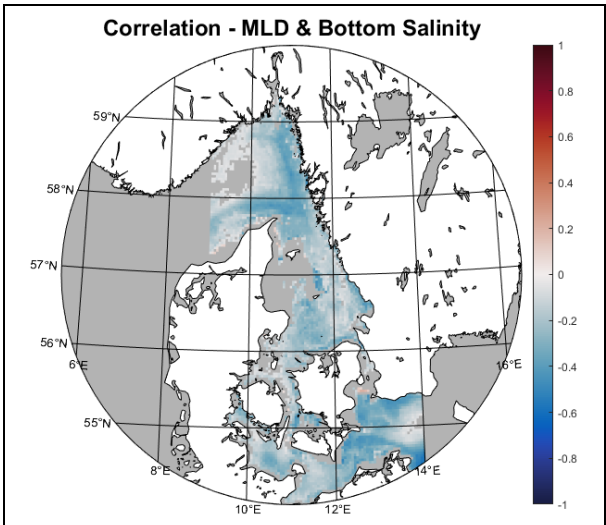
Woodworth, P. L., Melet, A., Marcos, M., Ray, R. D., Wöppelmann, G., Sasaki, Y. N., Cirano, M., Hibbert, A., Huthnance, J. M., Monserrat, S., & Merrifield, M. A. (2019). Forcing Factors Affecting Sea Level Changes at the Coast. *Surveys in Geophysics*, *40*(6), 1351-1397. https://doi.org/10.1007/s10712-019-09531-1

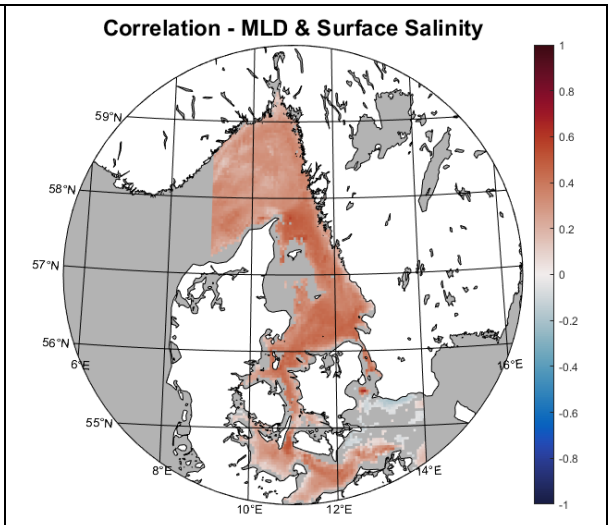
Woodworth, P. L., & Player, R. (2003). The Permanent Service for Mean Sea Level: An Update to the 21st Century. *Journal of Coastal Research*, *19*(2), 287-295. http://www.jstor.org.ezproxy.ub.gu.se/stable/4299170

Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. *Neural Computation*, *31*(7), 1235-1270. https://doi.org/10.1162/neco_a_01199

6 APPENDIX

6.1 A





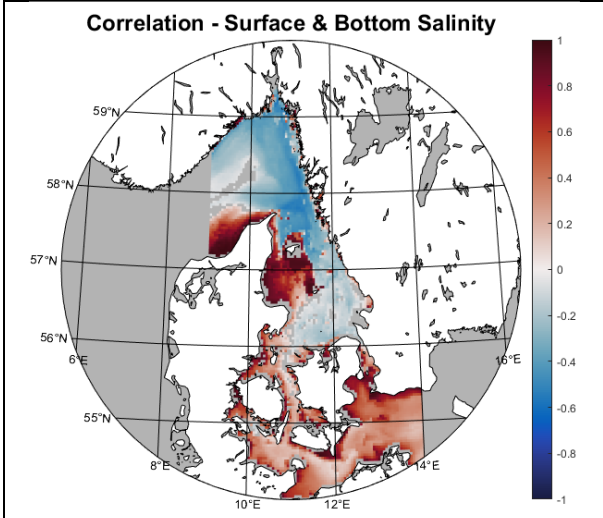


Figure 1: Correlation maps between mixed layer depth (MLD) and bottom salinity (top-left), MLD and surface salinity (top-right) and surface and bottom salinity (bottom-left).

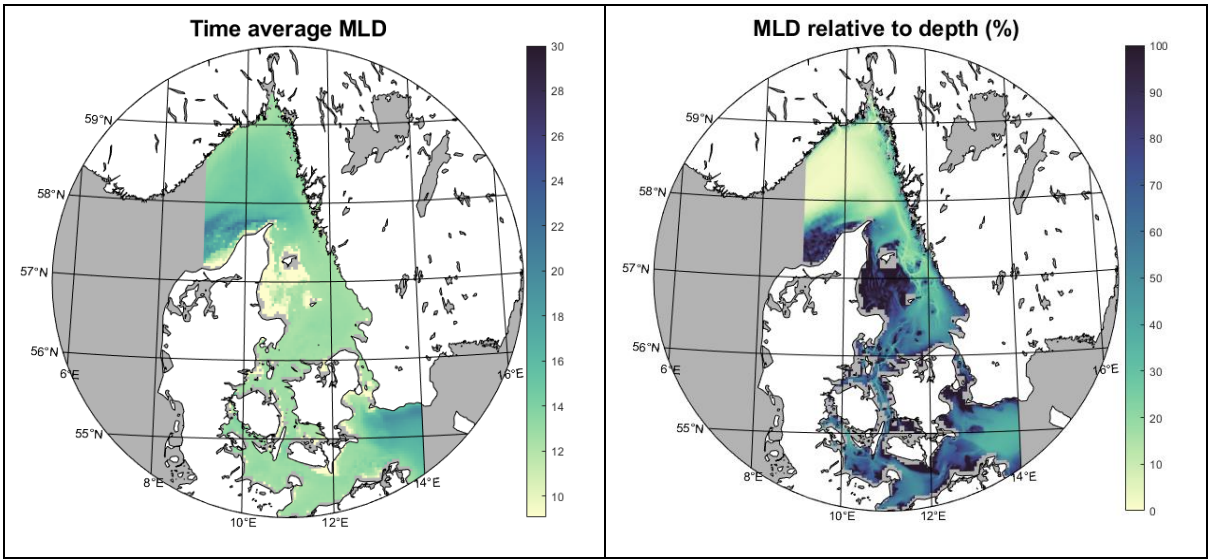


Figure 2: Figures showing the time averaged MLD over 2014-2019 (depth in meters, left) and the relationship between time averaged MLD and depth (MLD given as a percentage of the total depth, right)

6.2 B

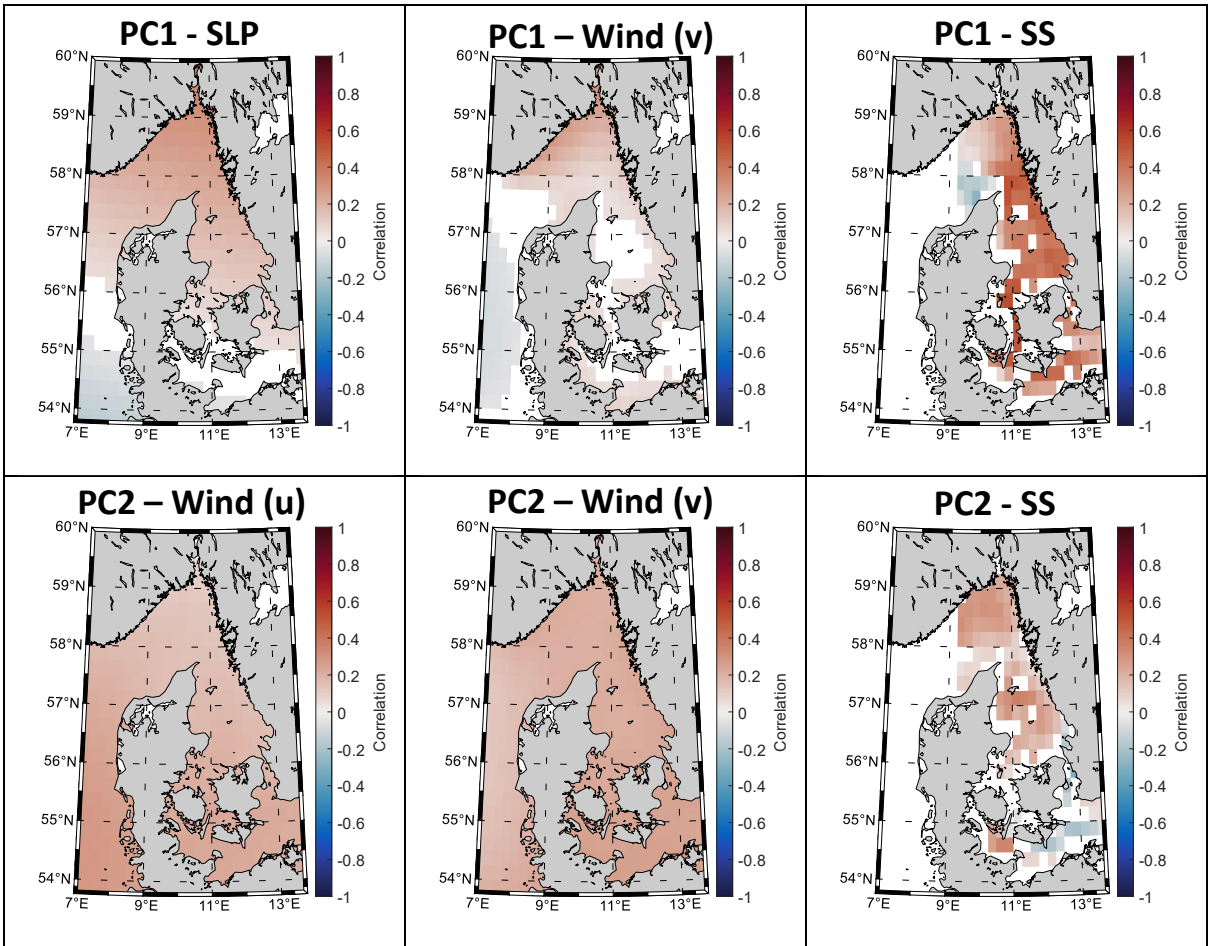


Figure 1: Correlation analysis between sea level variability PC1, PC2 and background sea level drivers.