Uncertainties and discrepancies in the representation of recent storm surges in a non-tidal semi-enclosed basin: a hind-cast ensemble for the Baltic Sea

Marvin Lorenz[1] and Ulf Gräwe[1]

[1]Leibniz Institute for Baltic Sea Research Warnemünde, Rostock, Germany

Correspondence: Marvin Lorenz (marvin.lorenz@io-warnemuende.de)

Abstract. Extreme sea level events, such as storm surges, pose a threat to coastlines around the globe. Many tide gauges have been measuring sea level and recording these extreme events for decades, some for over a century. The data from these gauges often serve as the basis for evaluating the extreme sea level statistics, which are used to extrapolate sea levels that serve as design values for coastal protection. Hydrodynamic models often have difficulty in correctly reproducing extreme sea levels and, consequently, extreme sea level statistics and trends. In this study, we generate a 13-member hind-cast ensemble for the non-tidal Baltic Sea from 1979 to 2018 using the coastal ocean model GETM (General Estuarine Transport Model). In order to cope with mean biases in maximum water levels in the simulations, we include both simulations with and without wind speed adjustments in the ensemble. We evaluate the uncertainties in the extreme value statistics and recent trends of annual maximum sea levels. Although the ensemble mean shows good agreement with observations regarding return levels and trends, we still find large variability and uncertainty within the ensemble (95% confidence levels up to 60 cm for the 30-year return level). We argue that biases and uncertainties in the atmospheric reanalyses, e.g. variability in the representation of storms, translate directly into uncertainty within the ensemble. The translation of the variability of the 99th percentile wind speeds into the sea level elevation is in the order of the variability of the ensemble spread of the modelled maximum sea levels. Our results emphasise that 13 members are insufficient and that regionally large ensembles should be created to minimise uncertainties. This should improve the ability of the models to correctly reproduce the underlying extreme value statistics and thus provide robust estimates of climate change-induced changes in the future.

1 Introduction

Extreme sea levels (ESLs) result from complex interactions between different processes that lead to very high sea levels that can cause coastal flooding, often damaging coastal communities. It is therefore important to understand and assess how sea-level dynamics and ESLs have changed in the past and how they will change in the future.

For the non-tidal Baltic Sea, the Baltic Sea Assessment Reports (BEARs, https://esd.copernicus.org/articles/special_issue1088. html) provide an overview of the current state of knowledge and knowledge gaps regarding sea-level dynamics and ESLs in Weisse et al. (2021) and Rutgersson et al. (2022), as well as projected changes of mean sea level and ESLs in the future in Meier et al. (2022a, b). Weisse et al. (2021) argue that improved numerical models with higher resolution are needed since

25 the occurrence of ESLs is not always well described by classical statistical distributions (Männikus et al., 2019). However, the representation of ESLs in numerical models ~~adds~~ is depending on the representation of storms in the atmospheric dataset, adding another source of uncertainty. In this study, we will assess the uncertainty and variability that can be expected from numerical models when investigating simulations of ESLs, focusing here on hind-cast simulations.

The most prominent ESLs in the Baltic Sea are storm surges. However, even during storm surges, other effects such as wave
30 setup and seiches can simultaneously modulate the sea level. For the Baltic Sea, *preconditioning*, i.e., the mean filling of the Baltic Sea, plays an important role. It describes the process of mean sea level changes due to water mass exchange with the North Sea caused by persistent westerlies or easterlies on a time scale of weeks (e.g. Madsen et al., 2015; Soomere and Pindsoo, 2016; Weisse and Weidemann, 2017). This significantly increases or decreases the background sea level upon which the storm surges act and thus the overall height of the ESL event. One of the highest sea levels observed in the Baltic Sea resulted from
35 the coincidence of preconditioning that filled the Baltic Sea to very high levels and the simultaneous occurrence of a storm surge. This was in 1872 when the maximum water levels in the western Baltic Sea along the German coast exceeded 3 m above the mean sea level (Rosenhagen and Bork, 2009; Feuchter et al., 2013; Hofstede and Hamann, 2022). The return period of this event is estimated to be ~ 3400 years (Bork and Müller-Navarra, 2009). Recently it has been shown that this ESL could have been even higher if the preconditioning and the weather systems had been slightly different ~~(?)~~(Andrée et al., 2023).
40 Sea-level rise is projected to increase the global mean sea level by 28 to 102 cm (median) by 2100, depending on anthropogenic greenhouse gas emissions as well as the response of the ocean and the cryosphere in the upcoming decades ~~(IPCC, Chapter 9, Fox-Kemper et al., 2021)~~(IPCC 2021, Chapter 9, Fox-Kemper et al., 2021). Considering ice sheet dynamics in more detail, Bamber et al. (2019) ~~found~~ list in their Tab. 2 possible ranges from 69 to 111 cm (median), 49-98 to 79-174 cm (17-83%), and 36-126 to 62-238 cm (5-95%)~~, their Tab. 2. With this rise, the exchange of water masses with the North~~
45 ~~Sea is expected to increase, which could affect preconditioning~~. The sea-level rise is expected to be additive to the ESLs in the open Baltic Sea (Gräwe and Burchard, 2012; Hieronymus et al., 2018). Recent scenario studies for the Baltic Sea have shown no significant changes in maximum wind speeds (Christensen et al., 2022) from which Meier et al. (2022a) conclude that ESLs relative to mean sea level do not change beyond statistical significance. ~~This means that singular historical events~~ ~~such as the 1872 storm surge can be expected to exceed the estimated 200-year return values under additional mean sea level~~
50 ~~rise scenarios.~~

Past mean sea-level rise for the Baltic Sea has been shown to align with the global mean ~~(Weisse et al., 2021, and references therin)~~(Weis However, sea-level rise has been spatially heterogeneous due to the glacial isostatic adjustment (GIA, Peltier, 2004) and will continue to be heterogeneous in the future (Grinsted, 2015). In some areas, the GIA exceeds the sea-level rise, e.g. in the Gulf of Bothnia, leading to decreased water depth. ~~For~~ Mean sea-level rise in the Baltic Sea ~~, it has also been shown that mean sea-level~~
55 ~~rise~~ also depends on atmospheric changes~~(Gräwe et al., 2019)~~, such as atmospheric pressure, wind speed, and wind direction (Gräwe et al., 2019). In general, sea-level rise shifts the ESLs (relative to present-day mean sea level) to higher base levels, and ~~return periods of~~ thus high ESLs become more frequent (Wahl et al., 2017). ~~This inhomogeneity in~~ Present sea-level rise and atmospheric changes ~~has~~ have already led to more frequent and longer ESL events (Wolski and Wiśniewski, 2021). Furthermore, ESLs have also become significantly higher in the eastern Baltic Sea, e.g., up to 5 to 10 mm yr^{-1} along the Estonian coast and

60 the Gulf of Finland (Pindsoo and Soomere, 2020). Because of these changes, non-stationary statistical tools have been used to infer further these trends, e.g. non-stationary distributions of the Generalised Extreme Value (GEV) (Coles et al., 2001). For the ~~Baltic Proper~~Gulf of Riga, Kudryavtseva et al. (2021) show that the shape parameter has changed significantly and often abruptly in the past. A sudden decrease of the shape parameter between 1986 and 1990 may be related to the North Atlantic Oscillation (NAO). High correlation values indicate a direct connection between the NAO and the mean sea-level of the Baltic Sea

65 (e.g. Andersson (2002); Hünicke and Zorita (2006); Karabil et al. (2018)), winds (e.g. Lehmann et al. (2011)), and hence ESLs ~~(e.g. Suursaar and Sooäär (2007); Wolski and Wiśniewski (2022)~~(e.g. Suursaar and Sooäär, 2007; Wolski and Wiśniewski, 2022). However, extreme winds show little correlation with the NAO (Bierstedt et al., 2015).

Due to all these different components leading to the observed ESLs, not only do numerical models have difficulty correctly reproducing each ESL in hind-cast simulations, but these uncertainties and biases are also present in scenario simulations for

70 the future. Since most atmospheric processes cause sea level changes, high-resolution (both in time and space) atmospheric forcing datasets are required to reproduce the observed ESLs. In modern atmospheric reanalyses, data assimilation is used to force the model to reproduce observed weather systems such as passing storms. Nevertheless, data assimilation does not make the atmospheric datasets perfect. Depending on how the models are configured and the focus of the respective simulation, each reanalysis has its strengths and weaknesses. Global reanalyses have a coarse resolution, and regional reanalyses depend on

75 the boundary conditions of the global reanalyses. To cope with these issues, ensembles are created from as many members as possible to compensate for the strengths and weaknesses of each member by examining the ensemble mean and the ensemble variability. However, ESL hind-cast simulations are sparse, and hind-cast ensembles focusing on ELSs are basically nonexistent or include only a few members. Global ESL reanalyses (e.g. Muis et al., 2016) have the disadvantage that they cannot focus and adjust for local biases. ~~On the other hand, regional~~ Regional hind-casts can be adjusted and benefit from regionally

80 down-scaled atmospheric datasets. However, these datasets are rare. Alternatively, data-driven models can be used to generate an ensemble (e.g. Tadesse and Wahl, 2021; Tadesse et al., 2022). However, these data-driven models are only applicable to gauge stations. Numerical models are needed to fill the spatial gaps and to untangle underlying physical processes.

In this study, we generate a hind-cast ensemble for the Baltic Sea to investigate the ability of the ensemble to reproduce observed ESLs and trends. We will focus solely on the variability of the representation of Baltic Sea ESLs within the ensemble

85 and trends in annual maximum sea levels. We will examine only ESLs caused by atmospheric processes. We use different extreme value distribution methods, the Generalised Extreme Value Distribution (GEV) and the Generalised Pareto Distribution (GPD), to evaluate the variability of expected return levels. We show that the variability within state-of-the-art atmospheric hind-cast products is large for extremes. Therefore, reproducing observed ESLs, estimating return levels, and recent trends is challenging for numerical simulations. We use six atmospheric reanalyses to compile an ensemble of ESL simulations in the

90 Baltic Sea from 1979 to 2018.

2 Methods

2.1 Numerical Simulations

We use the General Estuarine Transport Model ~~(GETM, Burchard and Bolding, 2002)~~(GETM (version 2.5), Burchard and Bolding, 2002),
a structured coastal ocean model (Klingbeil et al., 2018), to simulate the surface elevation in the Baltic Sea. The model has
been widely used for many Baltic Sea applications, including storm surges: Gräwe and Burchard (2012). Here, we use GETM
to solve the vertically integrated shallow water equations, considering only barotropic effects and ignoring baroclinic processes
(except for one simulation where we include static baroclinic circulation). We employ the superbee advection scheme (Pietrzak,
1998)~~and a bottom roughness length of $z_0 = 0.001$ m for all simulations~~. The wind stress is calculated from the 10 m wind
fields with

$$\boldsymbol{\tau} = C_D \rho_{\mathrm{air}} |\boldsymbol{u}_{10}| \boldsymbol{u}_{10}, \tag{1}$$

with $\rho_{\mathrm{air}} = 1.25$ kg m^{-3} and the 10 m wind speed vector $\boldsymbol{u}_{10}$. The drag coefficient C_D is computed by the formulation of
Kara et al. (2000). In our barotropic simulations we do not consider temperature differences between the ocean and atmosphere.
Therefore, the drag coefficient reads as

$$C_D = 10^{-3} \left(0.862 + 0.088w - 0.00089w^2\right),$$

where the wind speed is limited: $w = \max(2.5, \min(32.5, |\boldsymbol{u}_{10}|))$. The numerical setup of the Baltic Sea has a resolution
of one nautical mile, which is nested into a coarse (5 n.m.) model of the Northwest Atlantic with a bottom roughness of
$z_0 = 0.005$ m, see Fig. 1, similar to the setups of e.g. Gräwe et al. (2015). The Northwest Atlantic model is forced with
ERA5 (Hersbach et al., 2020) with unchanged wind speeds. Along its boundary, air pressure-induced water level changes are
imposed using the atmospheric pressure from ERA5 to include large pressure systems from the Atlantic into the model chain
(inverted barometric effect). This simulation prescribes boundary conditions for the one nautical mile North Sea / Baltic Sea
domain (all simulations). This may introduce some inconsistencies in the sea level when using other atmospheric forcings
for the North Sea / Baltic Sea nesting stage. However, we assume these errors to have little effect on Baltic Sea ESLs since
the Danish Straits act as a low-pass filter. Thus, ESLs are generated inside the Baltic Sea as a superposition of storm surges,
filling states, and local seiches which are all included in the model setup. Note that the boundary conditions for this nesting
do not include tides to avoid introducing tide-surge interactions (Arns et al., 2020). ~~However, tides~~Tides are negligible for
the Baltic Sea, as explicitly shown by Gräwe and Burchard (2012). However, standing-wave-surge (seiche-surge) interactions
(Arns et al., 2020) are included. Furthermore, the mean sea level is kept constant in order to study only the atmospheric-induced
extreme sea levels themselves. The North Sea / Baltic Sea domain is set up with a bottom roughness of $z_0 = 0.001$ m, a value
that is also used by other coastal ocean models than GETM for the Baltic Sea (e.g. Kärnä et al., 2021).

High-resolution atmospheric forcing in both time and space is required to model extreme sea levels. For the Baltic Sea,
we consider the regional atmospheric reanalyses listed in Tab. 1, which provide the needed high-resolution fields for winds
speeds (10 m height) and sea level pressure. For this study, we investigate the period from 1979 to 2018, the overlap between
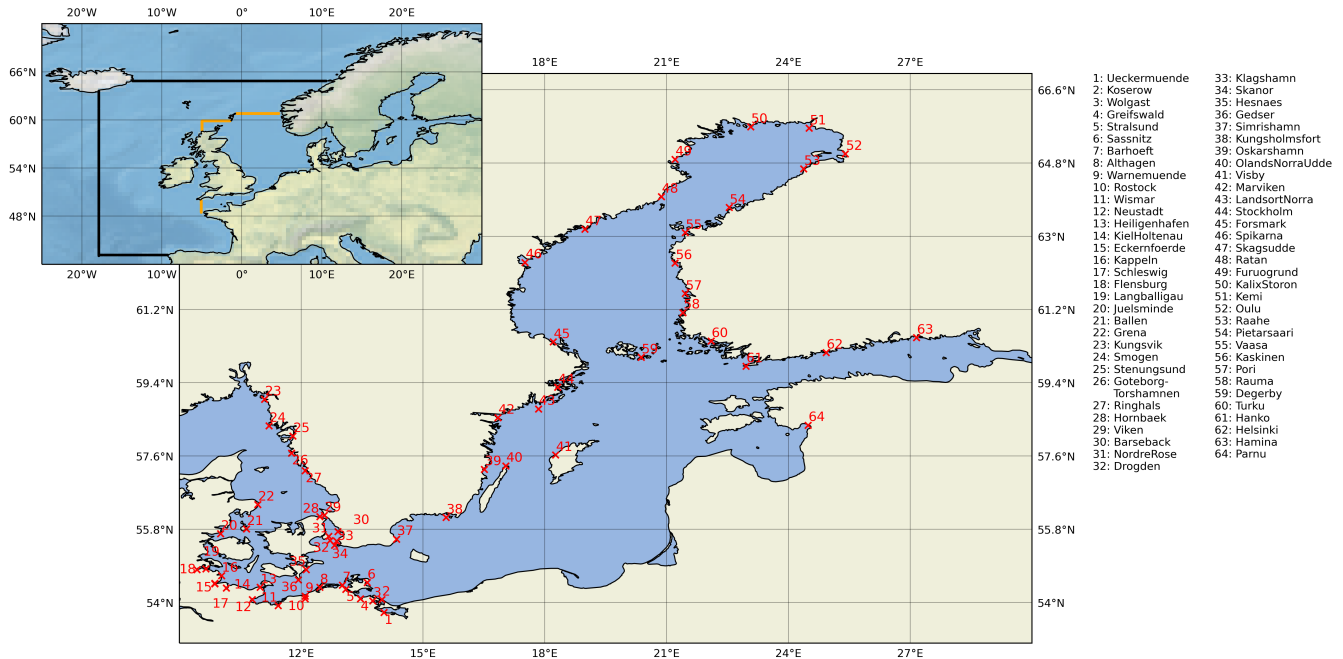
Figure 1. Model domain and locations of the stations (red crosses and numbers) used for validation. ~~Data is obtained from the European Marine Observation and Data Network (EMODnet,).~~ The black line indicates the boundary of the coarse Northwest Atlantic Ocean domain. The orange lines mark the boundaries of the one nautical mile setup of the North Sea & Baltic Sea.

all atmospheric datasets, focusing on the variability between these ensemble members. We included one run ('UERRA baroclinic') where we included monthly mean static density fields and static ice cover (taken from Gräwe et al. (2019)). For each atmospheric forcing, we ran one simulation with default wind speeds and one with increased wind speeds, see Tab. 1, for a total of 13 ensemble members. Before all analyses, we subtracted the linear trends from each time series to explicitly consider only ESLs relative to the respective mean sea level. This excludes changes in the mean sea level due to persistent changes in wind speed, wind direction, sea level pressure changes, ~~external loading, and land uplift~~ and external loading (Gräwe et al., 2019).

2.2 Observational Gauge Data

We use several gauges to validate the model results along the Baltic Sea coastline. The data is obtained from the European Marine Observation and Data Network (EMODnet, https://emodnet.ec.europa.eu) and the Global Extreme Sea Level Analysis (GESLA, Woodworth et al., 2016; Haigh et al., 2022). The station record lengths and data gaps are summarised in Fig. 2 (a table of the gauges is provided in Tab. S1 in the Supplement). Their locations are shown in Fig. 1 and also listed in Tab. S1 in the Supplement. Note that we only included stations with time series starting in 2004 and earlier, and that we use the data from 1979 onwards to match the simulation periods. For each gauge, the mean sea level rise is subtracted (linear fit over the entire

Table 1. Overview of the different atmospheric datasets used in this study. Each atmospheric forcing is included one time with default wind speeds and one time with increased wind speeds except the UERRA (baroclinic) run which is only computed with increased wind speeds. ~~This run includes monthly mean temperature and salinity fields, thus baroclinic circulation effects, as well as sea ice cover. This totals the number of ensemble members to 13.~~

Atmospheric forcing / run	time period	temporal resolution	spatial resolution	wind factor increase	reference
UERRA	1961 - 2018	hourly	11 km	+ 7 %	Ridal et al. (2017)
UERRA (baroclinic)	1961 - 2018	hourly	11 km	+ 7 %	Ridal et al. (2017)
ERA5	1950 - 2018	hourly	32 km	+ 11 %	Hersbach et al. (2020)
CFSR	1979 - 2018	hourly	32 km	+ 3 %	Saha et al. (2010, 2014)
coastDat1	1950 - 2018	hourly	0.22°	+ 7 %	Geyer (2014)
coastDat2	1979 - 2018	hourly	11 km	+ 4 %	Geyer (2014)
coastDat3	1948 - 2020	hourly	11 km	+ 3 %	Geyer et al. (2021)

time series) before analysis and comparison with model results. This also excludes sea level changes due to GIA from the tide gauges.

2.3 Extreme Value Statistics

140 We emphasise that we assume that the statistical properties of the extreme value distributions are constant in time within the considered time frame. Therefore, we do not consider temporal changes in the extreme value distributions. This assumption may not be valid ~~(Kudryavtseva et al., 2021)~~(Kudryavtseva et al., 2018, 2021). However, this study focuses on the variability within the ensemble and not the comparison between static and temporally varying extreme value distributions.

2.3.1 Generalised Extreme Value Distribution (GEV)

145 To describe the distribution of storm surge heights and return periods, we use the general extreme value (GEV) distribution (Coles et al., 2001) using the time series of annual storm season (July to June) ~~maxima, such as block maxima~~ block maxima (Männikus et al., 2020). The GEV (cumulative distribution function) is defined by

$$F(z, \mu, \sigma, \xi) = \exp\left\{ -\left[1 + \xi\left(\frac{z - \mu}{\sigma} \right) \right]^{-1/\xi} \right\}, \qquad (2)$$

where z is the sea level, μ is the location parameter, σ is the scale parameter, and ξ is the shape parameter. The shape parameter
150 ξ governs the tail of the distribution and depending on its value ~~subfamilies are defined~~the distribution reduces to:

1. Gumbel distribution for $\xi \to 0$:

$$F(z, \mu, \sigma) = \exp\left\{ -\exp\left\{ -\frac{z - \mu}{\sigma} \right\} \right\}. \qquad (3)$$
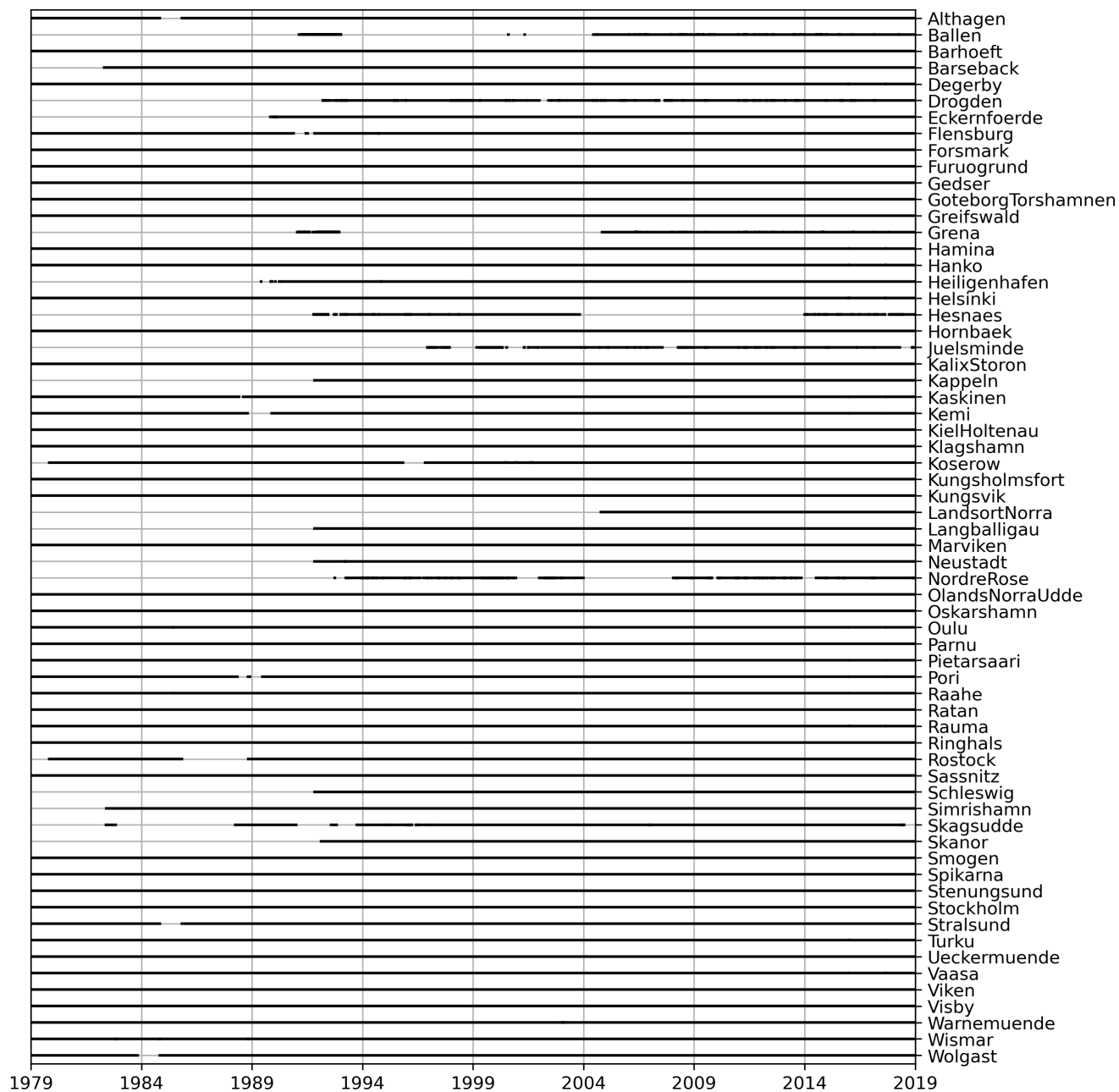
Figure 2. Overview of the record lengths and data gaps of the gauge stations. A table with exact locations and more details of the exact time spans and gaps can be found in the Supplement in Tab. 1.

2. Fréchet distribution for $\xi > 0$:

$$F(z, \mu, \sigma, \underline{\alpha}\xi) = \begin{cases} 0 & \text{for} \quad z \leq \mu, \\ \exp\left\{-\left(\frac{z-\mu}{\sigma}\right)^{-\xi}\right\} & \text{for} \quad z > \mu. \end{cases} \tag{4}$$

3. Weibull distribution for $\xi < 0$:

$$F(z, \mu, \sigma, \underline{\alpha}\xi) = \begin{cases} \exp\left\{-\left(\frac{z-\mu}{\sigma}\right)^{\xi}\right\} & \text{for} \quad z < \mu, \\ 1 & \text{for} \quad z \geq \mu. \end{cases} \tag{5}$$

For each gauge used in this study, the GEV is fitted to the annual storm season maxima. We use the Python code from Reinert et al. (2021) for the fitting, which uses the maximum likelihood estimation method.

2.3.2 Generalised Pareto Distribution (GPD)

Another function to describe the statistical distribution of extreme values is the Generalised Pareto Distribution (GPD) (cumulative distribution function),

$$F(z, \mu, \sigma, \xi) = \begin{cases} 1 - \left(1 + \frac{\xi z}{\tilde{\sigma}}\right)^{-1/\xi} & \text{for} \quad \xi \neq 0, \\ 1 - \exp\left(-\frac{z}{\tilde{\sigma}}\right) & \text{for} \quad \xi = 0, \end{cases} \tag{6}$$

where $\tilde{\sigma}$ is the scale parameter, $\tilde{\sigma} = \sigma + \xi(u - \mu)$, and u is the threshold (Coles et al., 2001). This method uses a peak-over-threshold to sample the time series (minus the mean sea-level rise). In this study, we use the 99.7% percentile of the sea level distribution as the threshold suggested by Arns et al. (2013) and only consider events separated by more than 48 hours. We use the maximum likelihood estimation method and a python script for the fitting.

3 Results

3.1 Comparison of ~~extreme~~ sea levels~~exceeding one meter above mean sea level~~

Since we are interested in the ~~extreme sea levels~~ESLs, we compare the observed ESLs for each station with the modelled ESLs ~~, see~~ (Fig. 3), where the stations are sorted in a clockwise direction along the Baltic Sea coast, starting from the Western Baltic Sea ~~. We use one meter as a threshold which aligns with the~~ (Fig. 1). The time series comparison for each tide gauge station shows a good agreement with low Root Mean Square Errors (RMSE ≤ 0.1 m) and high Pearson correlation coefficients $R \approx 0.9$, see Fig. S1. For the ESL comparison, we use the German Federal Maritime and Hydrographic Agency's definition of a storm surge in the Baltic Sea~~by the German Federal Maritime and Hydrographic Agency~~, which defines a storm surge as a sea level of more than 1 metre above mean sea level. Furthermore, we only consider events that are separated by more than 48 hours. We search for events that fulfil these criteria in the observed tide gauges and compare them with the modelled

sea level for each station. This comparison shows that the default wind speed simulations have a clear negative mean bias, red dots and values in 3. On average, the simulations underestimate the maximum water levels. The bias is different for each atmospheric forcing. To improve this comparison of ESLs in the model, we increased the wind speeds by a constant factor for

180 each atmospheric dataset, see Tab. 1. With this increase, the mean bias of the model results ~~improves~~ reduces significantly; see black dots and values in Fig. 3. However, it should be kept in mind that due to the homogeneous wind speed increase, some areas, i.e. stations, in the model domain show a larger bias than before, e.g. the central and Northern Baltic Sea. However, the wind speed increase is necessary to capture the correct ESLs in the Western Baltic Sea. Therefore, in the following analysis of the ESLs, we consider both the default simulations (larger mean bias) and the simulations with increased/adjusted wind

185 speeds (smaller mean bias) as an ensemble of 13 members. Note that the ESLs for stations in shallow coastal lagoons such as 'Althagen' and 'Schleswig' are overestimated because the model does not resolve the local topography and inlets.

3.2 Return levels and variability within the ensemble

3.2.1 Return levels obtained from the GEV method

Since the period considered is only 40 years, we compare the 30-year return periods, which should be well estimated by the

190 underlying 40 years of data, see also Gräwe and Burchard (2012). The comparison of modelled and observed 30-year return levels for the stations ~~,~~ (Fig. 4~~,~~) shows a large variability between the atmospheric reanalyses. A more detailed view of all return level fits and the different simulations' confidence intervals is shown in Fig. ~~S1~~ S2 for the station 'Warnemuende' in the Supplement. We have again sorted the stations clockwise around the Baltic Sea coast, starting in the Western Baltic Sea. This depiction shows that the default wind speed simulations have a clear negative bias in the Western Baltic Sea but show

195 good agreement in the central and Northern Baltic Sea regions. On the other hand, the adjusted wind speed simulations show good agreement in the Western Baltic Sea and an overestimation in the central and Northern regions. This pattern is similar to the general comparison of ESLs in Section 3.1. This shows that the biases from the previous section are directly reflected in the return level estimates. Each atmospheric forcing has its strengths and weaknesses regarding how well storm systems are represented in terms of intensity and tracks in different regions. Therefore, the spatially homogeneous adjustment of wind

200 speeds improves the representation in some areas but deteriorates it in others. However, these problems seem to cancel out for the ensemble ~~,~~ (Fig. 4h~~. The ensemble mean and spread (95% confidence interval)show that the ensemble is close to the observed return levels~~). For most of the stations, the ensemble mean return level is close to the mean return level based on the observations and well within the respective 95% confidence interval.

  Using the model to fill the area between the observed gauges, we show the ensemble mean 30-year GEV return level for

205 every fourth data point (every second point in latitude and longitude) in the Baltic Sea, see Fig. ~~6a. This depiction shows that the general pattern of the high and low return levels is similar in each simulation, but the estimated return levels vary considerably, as already seen previously.~~ 5a. Since westerlies are the dominant winds in the Baltic Sea region, the eastern coasts show higher return levels as expected and well documented (e.g. Wolski et al., 2014). In the South-Western Baltic Sea, west-facing coasts have higher return levels than those facing north or south but lower expected return levels than the Northern and Eastern Baltic
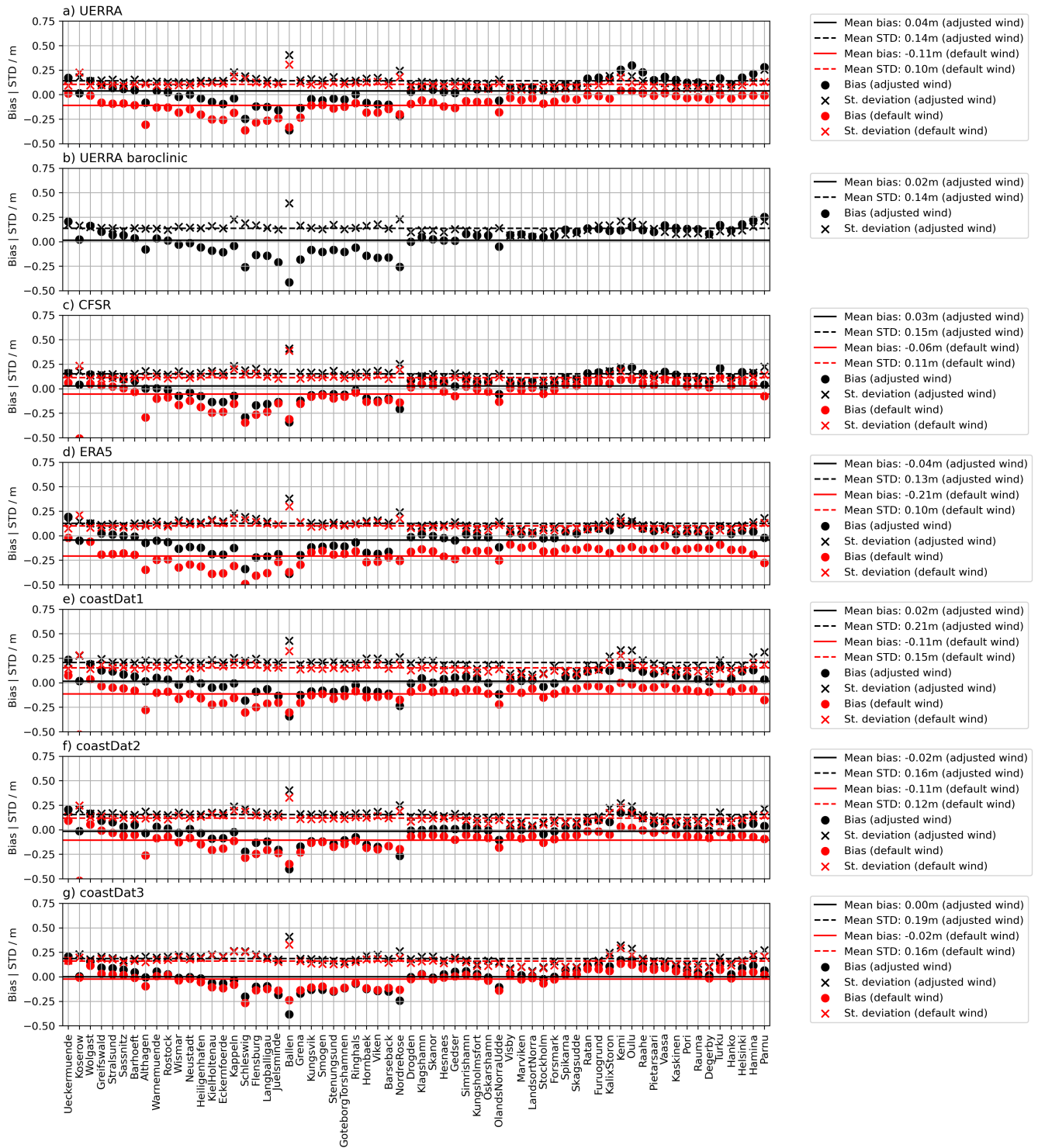
Figure 3. Mean bias of the model's ESLs and standard deviation summarised (STD) for each station of Fig. 2 for simulations ~~with and without the increased wind speed. The correction of the wind factor, see~~ listed in Tab. 1 ~~for the values, improves the mean bias of the model significantly for all atmospheric datasets. Note that for 'UERRA baroclinic' the model was only integrated for the adjusted wind factor. However, the improvement of the mean biases for the stations varies regionally. The default wind simulations show better agreement / smaller biases in the central and Northern Baltic Sea. In contrast, the Western Baltic Sea has a better agreement in the increased wind speed~~

Figure 4. Summary of the 30-year return levels using the GEV method for each gauge station and each ensemble member: a)-g) return levels and 95% confidence intervals for each atmospheric forcing and each simulation. h) ensemble mean and ensemble 95% confidence interval. The large 95% confidence interval of the station 'Hesnaes' is explained by a large uncertainty of the fitted shape parameter ξ.

210 Sea. Although the ensemble mean shows good agreement with the estimated return levels based on the observations, the spread, i.e. the 95% confidence interval, shows large values up to 50 cm and more (Fig. 5b).

Comparing the different ensemble members to the ensemble mean ,(Fig. 6,) we see more clearly how each member has its regions where it is close to the ensemble mean and regions where it deviates strongly. This depiction shows that the general pattern of the high and low return levels is similar in each simulation, but the estimated return levels vary considerably, as

215 already seen previously. Without again describing the spatial pattern of each ensemble member, we want to emphasise the discrepancies between the individual simulations. Generally, the adjusted wind speed simulations are at the high end of the distribution, whereas the default wind speed simulations are at the low end. The simulation with the lowest return levels is ERA5. The simulation with the highest return levels is the UERRA simulation, which includes the increased wind speeds and incorporates static density effects. Already when comparing the default simulations, the variability differences between the

220 30-year return levels of these six simulations can be as high as 60 cm, even though they are all reanalyses.

3.2.2 Return levels obtained from the GPD method

Similar to the return levels estimated by the GEV method, we estimated the return levels based on the GPD method for the same period. In the Supplement, we provide plots comparing the estimated return levels of the simulations with the return levels based on the observations: for the station 'Warnemuende' in Fig. S2 S3 and for the 30-year return levels in Fig. S3 S4. In

225 short, comparing the 30-year return levels for each station, the patterns are similar to the GEV results except for the generally lower estimates for the 30-year return levels. Again, the adjusted wind speed simulations are in better agreement with the observations for the Western Baltic Sea than for the Northern and Eastern Baltic Sea, where the default wind speed simulations are in better agreement with the observations. Overall, based on the observations, the ensemble mean 30-year return levels lie within the 95% confidence intervals of the 30-year return level estimates.

230 The general pattern of high and low GPD 30-year return levels ,(Fig. 5c,) is very similar to the pattern using the GEV method of high and low GEV 30-year return levels in Fig. 5a. Also, the spatial GPD 95% confidence interval of the return levels from the GPD method is pattern and the values are similar to the results obtained from the GEV method, GEV 95% confidence interval (Fig. 5b and d). This suggests that the uncertainty within the ensemble is independent of the method used to estimate the return levels.

235 Again comparing the ensemble members to This claim is supported by the deviations of the 30-year GPD return levels from the ensemble mean ,(Fig. S4 S5 in the Supplement, shows similar results to the same comparison for the). Each ensemble member shows similar deviations (both in pattern and values) to the GEV return levels with slightly different values, but the general patterns are the same in Fig. 6.

3.3 Trends

240 We have seen in the previous sections that the comparison of ESLs with observations and modelled ESLs differ from observations to a different extents for each atmospheric forcing. Therefore, the estimated 30-year return levels differ significantly within the ensemble, leading to significant uncertainty. Previous studies have shown some trends in storm surge occurrence and heights.
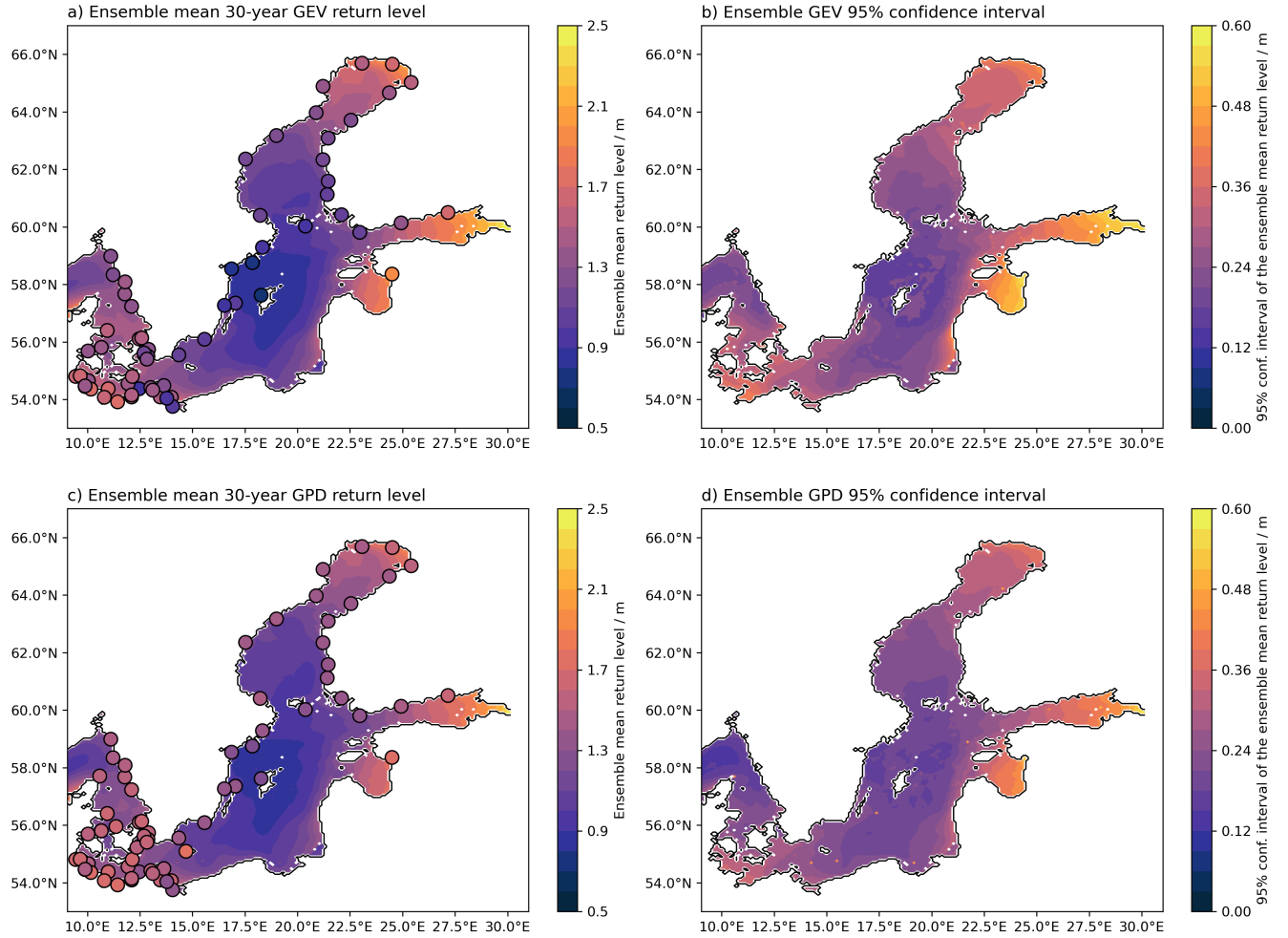
Figure 5. ~~Spatial maps~~ Maps of the return levels and the 95% confidence intervals: a) ensemble mean of the 30-year GEV return levels. b) the 95% confidence interval of the GEV return levels. c) ensemble mean of the 30-year GPD return levels. d) the 95% confidence interval of the GPD return levels. In a) and c), the coloured dots indicate the 30-year GEV/GPD return levels of the observations.
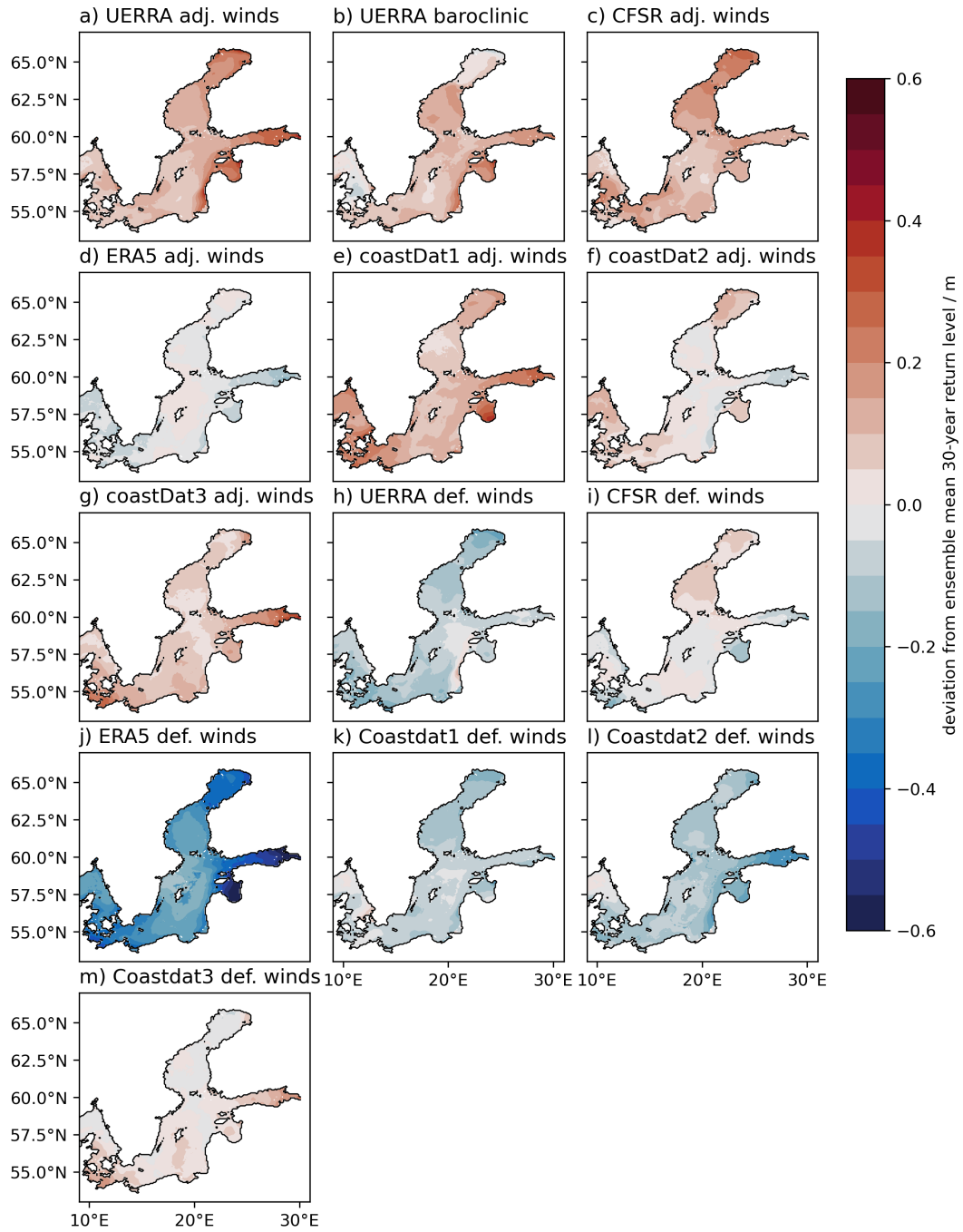
Figure 6. Spatial distribution of the 30-year GEV return level deviation from the ensemble mean (see Fig. 5a) for each ensemble member.
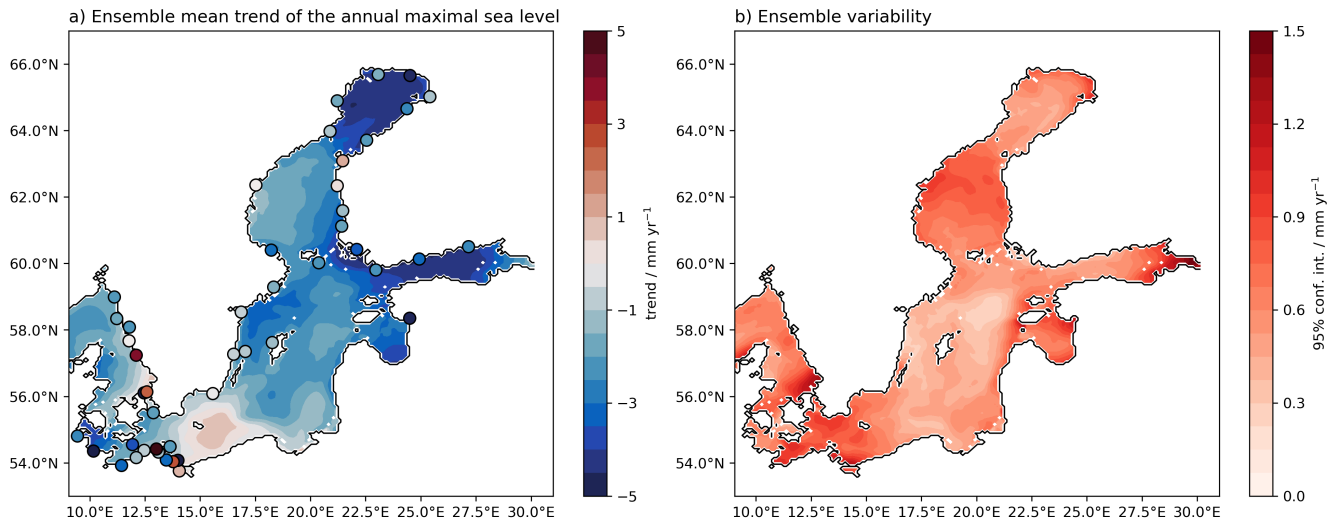
Figure 7. The ensemble mean trend in a) and the 95% confidence interval of the mean trend of the annual storm season maxima in b) for the time period from 1979 to 2018. The coloured dots show the trends of the tide gauges covering the same 40 years. Note that the mean sea level trends have been subtracted beforehand.

Therefore, we look at linear annual storm season maxima trends within the ensemble period, 1979-2018. For the ensemble mean , (Fig. 7,) we find a negative mean trend in the Baltic Sea except for the eastern Kattegat and the waters north of Poland.

245   The ensemble mean shows the largest negative trends in the Northern Baltic Sea and the Gulf of Finland. However, the uncertainty within the ensemble is large up to 1.5 mm yr^{-1} , (Fig. 7b). Comparing the trends with the observed trends, coloured dots in 7a, we see a general agreement in the trends and in the spatial distribution. However, some gauge stations show deviations in the sign of the trend, e.g. stations in the Western Baltic Sea. This indicates that details in the trends are not captured by either the ocean model resolution (one nautical mile) or the atmospheric reanalysis data.

250   Comparing the trend of each member , (Fig. 8), we see an agreement in the spatial pattern of trends as represented by the ensemble mean. However, how the size and shape of the different areas of positive and negative trends differ, e.g. the extent of the area in the Western Baltic Sea that shows a positive trend varies for each ensemble atmospheric forcing. Furthermore, the actual values of the trends differ greatly within the different simulations, not only between the adjusted and default wind speeds but even within the sub-sample of default simulations.

4   Discussion

From the results of our study, we find that the choice of atmospheric reanalysis significantly impacts how well the ocean model simulates past sea level extremes. Not only are the estimated return levels significantly different depending on the choice of the atmospheric reanalysis, but the modelled trends in sea level maxima are also affected. One of the reasons is that each simulation forced by a different reanalysis product underestimates or overestimates different ESL events since each real storm system is
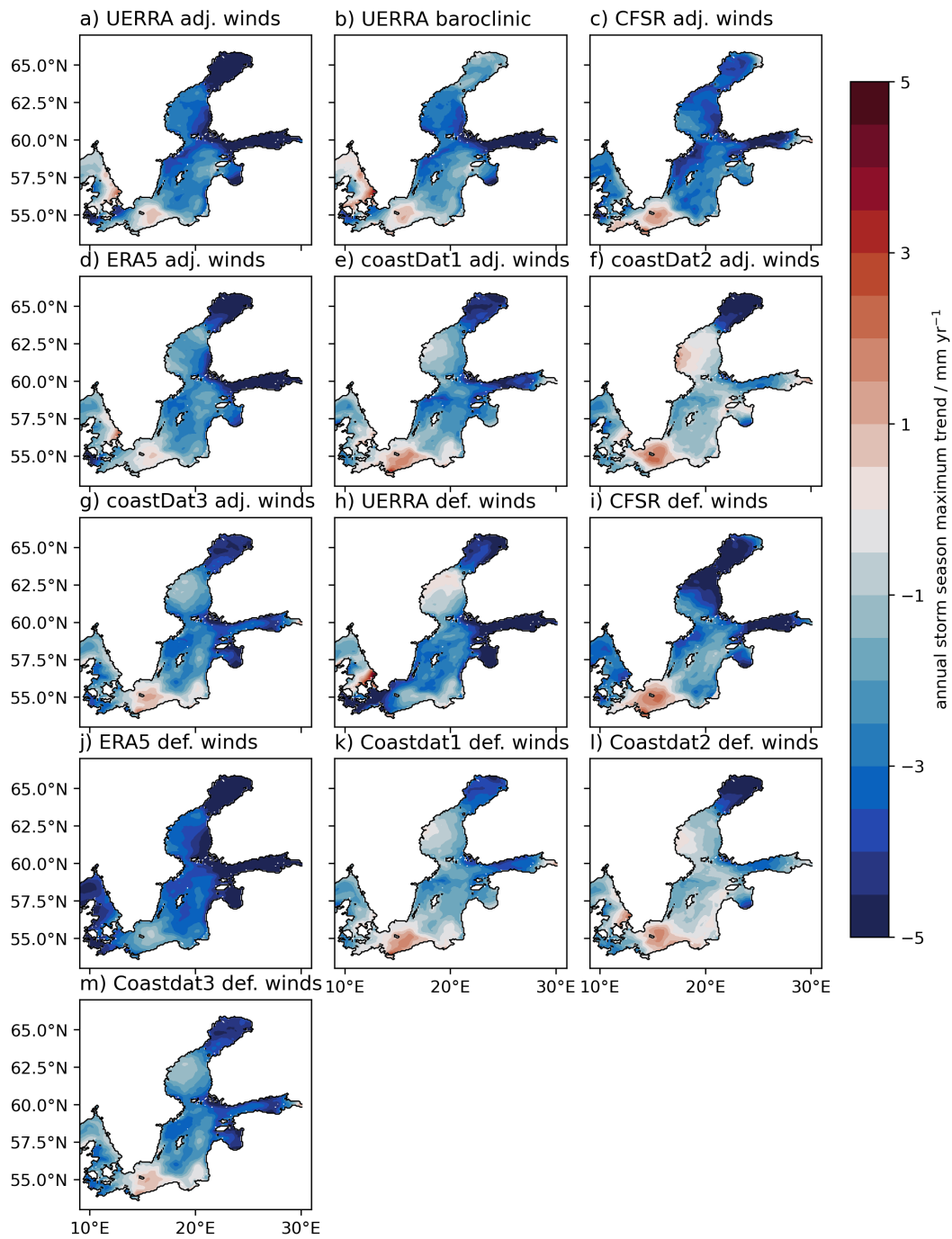
Figure 8. Trend in each ensemble members' annual storm season maxima. Note that the mean sea level trends have been subtracted beforehand.

260 represented slightly different. Furthermore, each reanalysis has different biases, both spatially and in wind speed distribution and direction. Therefore, the variability within the generated ensemble can be quite large, both for the return levels and trends. Nevertheless, the ensemble ~~agrees well with~~ mean can reproduce observed return levels and observed trends.

The variability within the ensemble for the 30-year return levels can be as high as 60 cm in the Gulf of Finland or the eastern part of the Baltic Proper. Along the western coasts of the Baltic Sea, the 95% confidence values are around 30-40 cm. The
265 ~~variability of the atmospheric high wind speeds, i.e. storms, in the different atmospheric datasets forward into the ESLs. The~~ 99th percentiles for each default 10 m wind speed (u_{10}) dataset from Tab. 1 ~~,~~ (Fig. 9~~,~~) show the variability between the different datasets. ERA5 has the lowest ESLs and has the lowest 99th percentile wind speeds over the Baltic Sea. CFSR shows the highest 99th percentile wind speeds over the Baltic Sea of the ensemble, which translates into the smallest adjustment for wind speeds of 3%, Tab. 1. The 95% confidence intervals due to the variability within the ensemble are $\sim 2.5\,\mathrm{m\,s^{-1}}$ between the default
270 atmospheric datasets over the entire Baltic Sea. This can result in a difference in the maximal sea levels of approximately

$$|\Delta\eta| = \left| \frac{L C_D \rho_{\mathrm{air}} \left((u_{10}^2 - (u_{10} \pm 2.5)^2) \right) \mathrm{m^2\,s^{-2}}}{g \rho_0 H} \right| \approx 10 - 60\,\mathrm{cm}, \tag{7}$$

where L is a distance scale (fetch length) ($\mathcal{O}(L) = 100\,\mathrm{km}$), the drag coefficient $C_D \approx 0.001$ C_D, which is approximated with 0.001 for simplicity, the density of air $\rho_{\mathrm{air}} \approx 1.25\,\mathrm{kg\,m^{-3}}$, the density of seawater $\rho_0 \approx 1000.0\,\mathrm{kg\,m^{-3}}$, the water depth H, and g the gravitational constant. The sea level difference depends strongly on the water depth and the distance L. For
275 $u_{10} = 18.0\,\mathrm{m\,s^{-1}}$, $L = 300\,\mathrm{km}$, and $H = 50\,\mathrm{m}$, the resulting elevation change would be $\sim 43\,\mathrm{cm}$. The resulting sea level change would be even higher for shallower waters or greater distances. This approximation of the effect of the atmospheric variability is in the order of magnitude of the sea level confidence interval of the ensemble. This highlights the importance of correctly representing of high percentile wind speeds, i.e. storms, in the atmospheric reanalysis, to capture the ESLs in the simulations. The variability within the ensemble of our study is in agreement with other studies in this regard: Dieterich et al. (2019) discuss
280 the historical period (1961-2005) of a regional climate model for the North Sea and the Baltic Sea to study the variability within an ensemble of six members (six different global model boundary conditions for their regional general circulation model). They find uncertainties between 20-40 cm added by the choice of the global circulation model boundary conditions as the largest uncertainty for the 100-year return levels. Eelsalu et al. (2014) show for their small three-member ensemble for the Estonian coast that uncertainties for 15-year return levels are in the range of about 20 cm. However, for longer return periods, their
285 spread grows significantly to 60 cm and more. In general, the errors grow with longer return periods, as does the variability of our ensemble. To reduce the uncertainty, we believe a large regional ensemble is needed ~~to minimise the errors~~.

The variability of the return levels within the ensemble is independent of the choice of chosen extreme value distribution. Our results show similar confidence intervals within the ensemble for both the GEV and GPD methods: both show similar spatial patterns and similar values. However, their absolute values for the return levels differ, especially for the short
290 and very long return periods. Based on the annual maxima of the storm season, the GEV method yields higher 30-year return levels than those obtained from the GPD method. We do not want to discuss the pros and cons of the methods. We refer to Arns et al. (2013) for detailed discussion of this. However, the difference in the return levels of the two methods for the 30-year return levels shown in this study is smaller or in the range of variability within the ensemble. This indi-
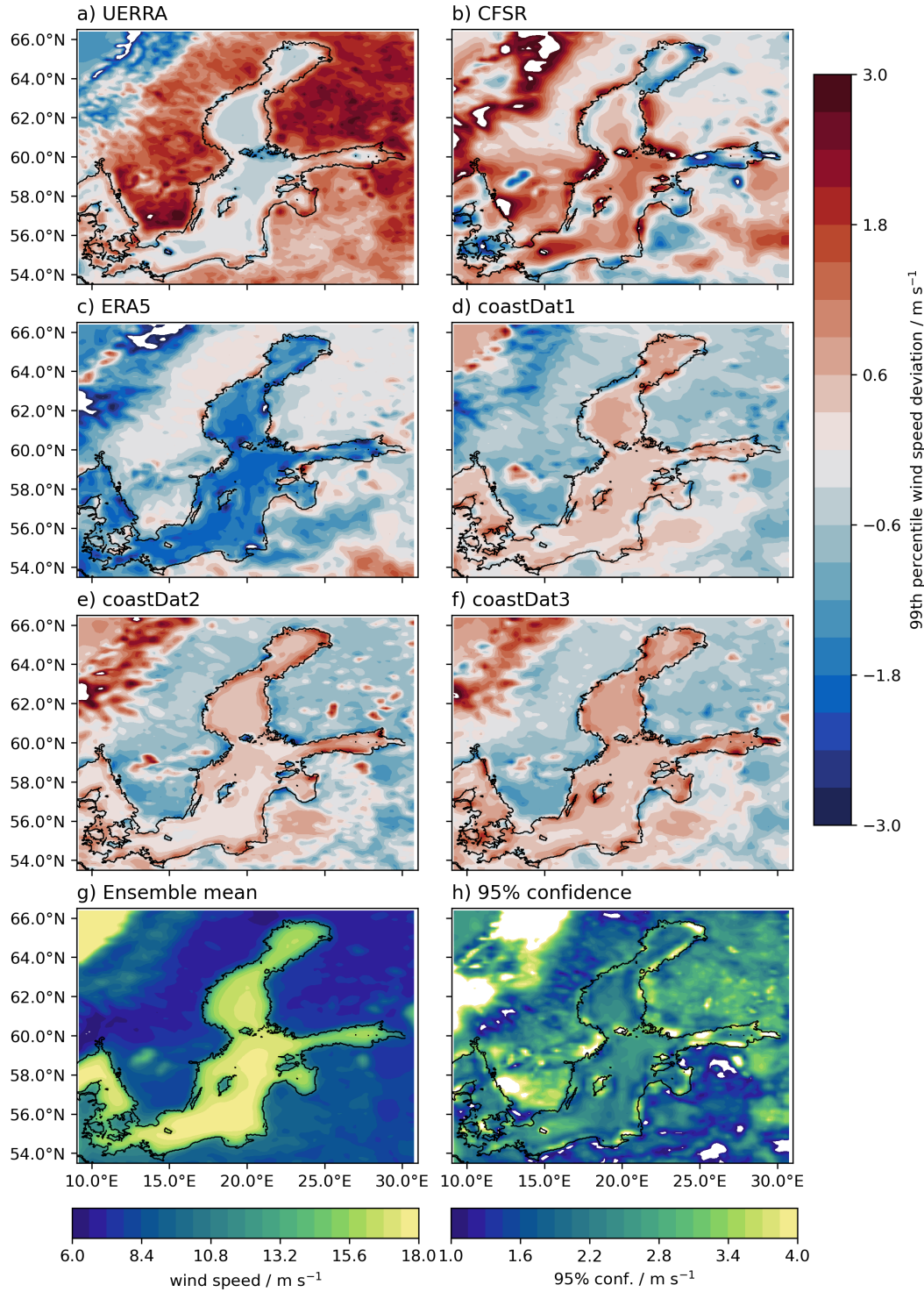
Figure 9. Deviations of the 99th percentiles of the 10 m wind speeds from their mean for the meteorological forcings used to generate the ensemble. The mean in g) and the standard deviation in h) are computed by interpolating the individual percentiles of each forcing onto the grid of the UERRA dataset.

cates that future efforts should be made to minimise the uncertainty. ESLs are the superposition of many sea level processes: preconditioning, storm surges, and standing waves (seiches). For example, since preconditioning can increase the mean sea level in the Baltic Sea on weekly time scales, the total sea level during a storm surge can be increased. These two (and more) compounding mechanisms follow different statistical distributions (e.g. Suursaar and Sooäär, 2007). Our analysis neglects this point as the GEV and GPD methods assume a single statistical distribution which still led to good fits. Furthermore, the mentioned mechanisms are interacting non-linearly (Arns et al., 2020). Disentangling these two (and more) compounding processes (Soomere and Pindsoo, 2016; Pindsoo and Soomere, 2020), of this ensemble remains a matter for a future study.

We have shown with the generated ensemble that the trend of the annual storm season maxima is different for each ensemble member, but still shows some agreement in the general spatial pattern. Since each storm is captured differently in each atmospheric dataset, this influences the trends of the ESLs. Fig. 10 shows the linear trends of the annual 99th percentile wind speeds for each dataset and the ensemble mean (interpolated to the UERRA grid). The datasets agree on negative trends for the southern, western, and central Baltic Sea for the hind-cast period. We found a positive trend for most datasets for the northern and eastern parts. CoastDat2 deviates from this general pattern with very small trends in the Western Baltic Sea and negative trends in the north. Since the CFSR-member is forced with both CFSR and CFSv2 no consistent model is used and the trends shown are expected to be mostly due to the change in the model configuration. An estimate of how the wind trends can be translated into sea level trends can be made using the derivative of equation (7):

$$|\partial_t \eta| = \left| \frac{2 L C_D \rho_{\text{air}} u_{10} \partial_t u_{10}}{g \rho_0 H} \right| \approx 0.1 - 1.0 \, \text{mm/yr}, \tag{8}$$

where the value $\partial_t X$ ~~should denote the trend~~ denotes the slope of the trend line. For $u_{10} = 18.0 \, \text{m s}^{-1}$, $L = 300 \, \text{km}$, and $H = 50 \, \text{m}$, and $\partial_t u_{10} = 0.03 \, \text{m s}^{-1} \, \text{yr}^{-1}$ the resulting change in elevation would be $\sim 0.7 \, \text{mm yr}^{-1}$. However, the ensemble trends of the sea level maxima are higher, indicating that directional changes and preconditioning must play a critical role. However, this is beyond the scope of this study. Furthermore, if we compare the trends of the ensemble mean to the results of Soomere and Pindsoo (2016) ~~for the Estonian coast, for which we did not have tide gauges of the desired lengths available~~ and Pindsoo and Soomere (2020), we find an apparent discrepancy. Whereas Soomere and Pindsoo (2016) and Pindsoo and Soomere (2020) find a clear positive trend of up to $10 \, \text{mm yr}^{-1}$, the ensemble of this study shows a negative trend. They further show that preconditioning is responsible for up to $4 - 5 \, \text{mm yr}^{-1}$ of the total trend, showing that preconditioning indeed plays a crucial role. It should also be noted that these trends are sensitive to the considered period since the natural variability of the Baltic Sea region includes oscillations in the order of decades due to the North Atlantic Oscillation (NAO) and the Atlantic Multidecadal Oscillation (AMO) (also named Atlantic Multidecadal Variability, AMV) (Meier et al., 2022b, and references therein). The considered time period of Soomere and Pindsoo (2016) and Pindsoo and Soomere (2020) is from 1961 - 2005 with 6-hourly model output. We compared the trends for tide gauge station which go back to 1961 for both time periods, 1979-2018 and 1961-2005, Tab. S2. The results show that both time periods lead to different trends, even changing the sign of the trend. Many of these stations show a positive trend for 1961-2005 and negative trends for 1979-2018. For example, stations in the Western Baltic Sea show a positive trend up to $4 \, \text{mm yr}^{-1}$ which is in agreement with the previous studies. However, this

is not the case for 'Parnu' which still shows a negative trend. Therefore, the discrepancy between this study and previous studies can be only partly attributed to the different time periods. This hints that the variability of ESLs to the NAO and AMO/AMV should be revisited. Furthermore, the 6-hourly output of Soomere and Pindsoo (2016) and Pindsoo and Soomere (2020) could introduce some aliasing since peak water levels may not be captured, compare to e.g. to Kiesel et al. (2023) for mean time series of ESLs in the Western Baltic Sea.

Our ensemble has a record length of 40 years. ~~FThe~~ The 100-year or 200-year return levels are often used as protection targets for coastal protection planning. Therefore, an extrapolation of the data leads to significant uncertainty ranges, see e.g. Fig. ~~S1 and~~ S2 and S3 in the Supplement, where the confidence ~~interval grows~~ intervals grow significantly when extrapolating, especially for the GEV method. To reduce the uncertainties, longer time series would be necessary to estimate the long return periods reliably. For the Baltic Sea, very long (> 100 years) time series exist and have been used to study the large return levels (e.g. Suursaar and Sooäär, 2007). However, our results show that climate change makes reliable estimates of the highest return levels for the current and future climate states based on small ensembles or single simulations complicated and uncertain, see also Muis et al. (2022). It is expected that the underlying statistical distribution will change with climate change. Therefore, the non-stationary extreme value statistics have been used in other studies, e.g., Kudryavtseva et al. (2021) and temporal trends in ESLs have been studied, e.g., with quantile regression (Barbosa, 2008). Disentangling these changes in statistical extreme value distributions is only possible with numerical models. However, much larger ensembles are needed to separate ensemble variability from changes due to climate change. Currently, the ensemble variability for climate scenarios for the European coasts is in the order of magnitude of the expected differences between low and high RCP scenarios (e.g. Vousdoukas et al., 2017). In addition, the large natural variability in the Baltic Sea regions makes this disentanglement even more complicated in this region.

To reduce model bias (Section 3.1), we increased the wind speeds by a constant factor which reduced the mean bias of ESLs. However, the results suggest that the wind factor should be varied spatially to account for regional biases in ESLs in the simulations. The ~~biases may~~ main negative model bias in our, but also previous studies, is found in the ESLs of the Western Baltic Sea. In early model studies the bias was because of coarse resolution which allowed water to exit through the Danish Straits during ESL events (e.g. Meier et al., 2004). Furthermore, local convergence effects need high model resolution to be correctly reproduced. But also in high-resolution hydrodynamic models negative biases in ESLs are found: 1 km resolution in Gräwe and Burchard (2012), 200 m resolution in Kiesel et al. (2023). Therefore, biases can be attributed to wind speed biases (e.g. Gräwe and Burchard, 2012). In the present study, we found the ESL biases to be caused by biases in the atmospheric datasets or how the ocean model translates the forcing into sea level changes. Applying better corrections is not a trivial undertaking and would certainly further disturb the physical consistency. Adjusting wind speeds may help with the biases in the ESLs, but it also has implications for general hydrodynamics, which is not the focus of this study. Because we use vertically integrated Reynolds averaged Navier-Stokes equations, we do not consider changes in the surface mixed layer depths, or upwelling and downwelling. Nevertheless, these would be strongly modified by the ~~default increase in~~ adjusted wind speeds and may lead to over- or underestimation of these~~with~~. This would have consequences for the general dynamics, e.g. stratification or the total salt import by Major Baltic inflows. One possible way of tackling this problem could be adjoint/inverse models
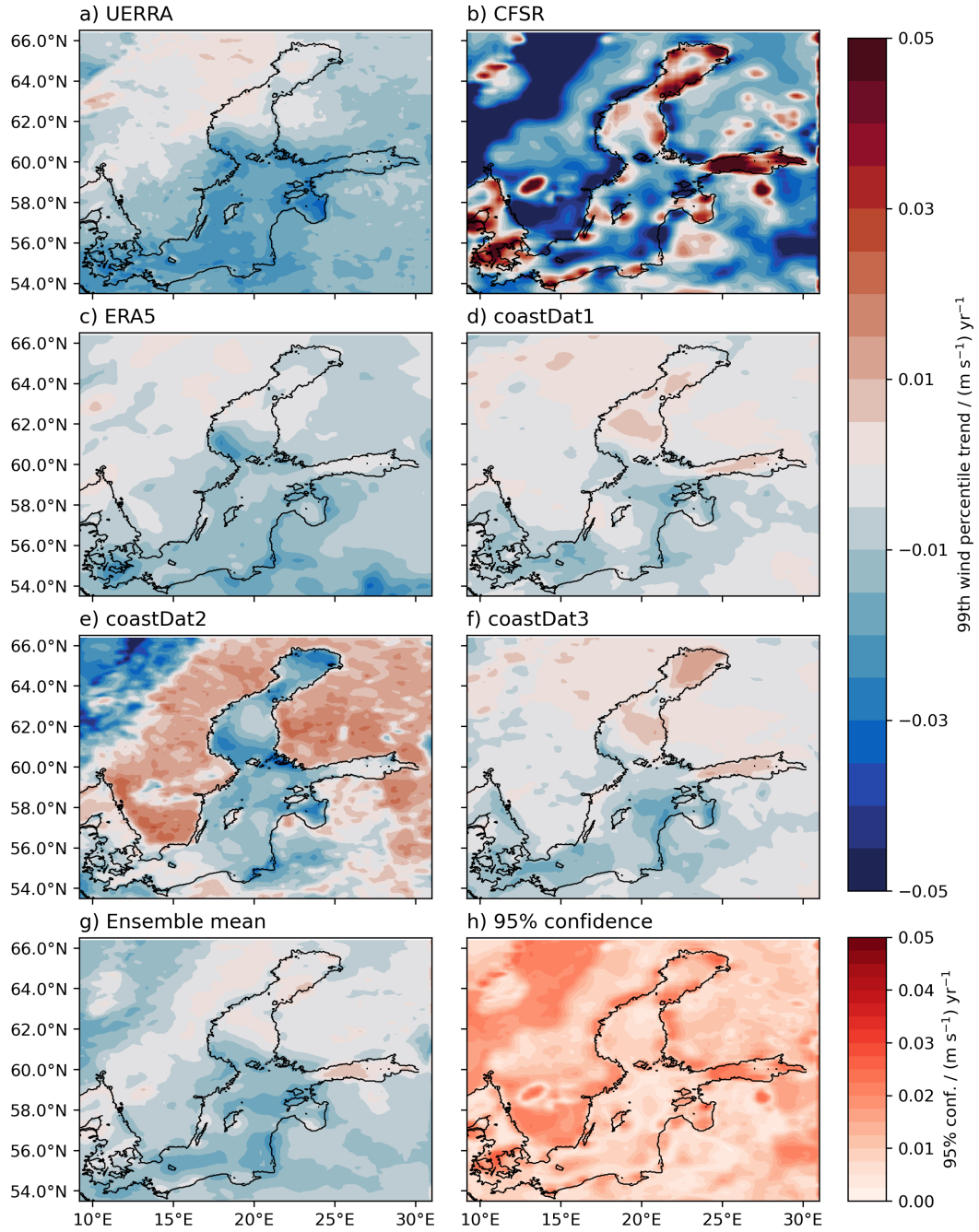
Figure 10. Trend of the annual 99th percentiles of 10 m wind speeds for the meteorological forcings used to generate the ensemble.

(Errico, 1997) which can be used to calibrate coastal ocean input fields like bottom friction (Kärnä et al., 2023) or in this case wind speed factors. However, this does not solve the underlying issue which we believe is found in the resolution of the atmospheric models over the Western Baltic Sea. Compared to the size of the Western Baltic Sea, the number of grid cells in the atmosphere is probably not high enough to resolve the winds properly.

There exist many formulations, i.e. polynomial fits, of the drag coefficient which translate wind speeds into winds stresses (e.g. Kara et al., 2000; Large and Yeager, 2009). Here we have used the formulation of Kara et al. (2000). Using a different formulation, e.g. one that yields a larger drag coefficient for high wind speeds, could have improved the representation of ESLs and potentially decreased the wind factors we applied. Since there exist many formulations, each formulation would lead to different ESLs, thus introducing a source for variability. The different formulations depend on the geographic locations of the underlying data which is fitted, e.g. offshore versus onshore (Smith et al., 1992). The formulation we have used (Kara et al., 2000) is based on observations made in the Arabian Sea. One could argue that this formulation may not suitable for the Baltic Sea. However, we expect for this study that the variability of the atmospheric datasets, Figs. 9 and 10, to be much larger than the differences in ESLs due to the drag coefficient formulations. Still, the exploration of the effect of different drag coefficients formulations on ESLs in the Baltic Sea should be carried out in a future study. A calibrated drag coefficient formulation using an adjoint model could be a possible solution (Peng et al., 2013). However, depending on the wind input fields, different best fits have to be expected. Therefore, this inverse method cannot be expected to provide a one-size-fits-all solution.

We have increased the ensemble spread by including the default and adjusted wind speed simulations. We could have either used only the default wind speed simulations and accepted the negative biases in the Western Baltic Sea, thus reducing the ensemble spread, or we could have included them to improve the ensemble mean for that region, with the drawback of larger uncertainties. Ultimately, uncertainties are shifted from one problem to another and would not be minimised.

5 Summary and Conclusions

We show that modelling extreme sea levels is difficult and relies heavily on correctly representing storms in the atmospheric forcing data. We derive and provide a 13-member hind-cast ensemble focusing on extreme sea levels, covering the period from 1979 to 2018. We constructed the ensemble using present state-of-the-art reanalysis datasets for the Baltic Sea region. We have explicitly not considered long-term changes in mean sea level. The ensemble mean can reproduce similar extreme value distributions compared to tide gauges around the Baltic Sea, both for the Generalised Extreme Value ~~Distribution~~ distribution (GEV) and the Generalised Pareto ~~Distribution~~ distribution (GPD). However, spatially heterogeneous biases can be expected depending on the choice of atmospheric forcing. We ~~, therefore,~~ find a large variability within the ensemble: up to 60 cm (95% confidence interval) within the ensemble for the 30-year return levels and trend variability up to $1.5\,\mathrm{mm\,yr^{-1}}$ for the annual maximum sea level. This variability is entirely due to the different atmospheric representations of storms and how these translate into the sea level computation in the ocean model. ~~This~~ One approach trying to minimise this uncertainty can be ~~reduced by using~~ much larger ensembles. Hence, we invite other scientists to incorporate our numerical simulations and results into future ensembles to increase the number of ensemble members. Initiatives such as the Baltic Sea Model Intercomparison

Project (BMIP, Gröger et al. (2022), https://www.baltic.earth/working_groups/model_intercomparison/index.php.en) should therefore be expanded in the future to build a large regional ensemble with multiple forcing combinations.

Since the trends of annual maximum sea levels are sensitive to the time period considered and the trends of this study differ from previously published studies (e.g. Soomere and Pindsoo, 2016; Pindsoo and Soomere, 2020), we believe that the dependence of ESLs on the NAO and AMO/AMV should be reconsidered. Furthermore, the contributions of preconditioning, seiches, and surges to ESLs in the context of climate modes need to be disentangled.

Code and data availability. The results, a frozen GETM version, and the analysis code can be found here: https://doi.org/10.5281/zenodo.8340649. The model output can be accessed here: https://thredds-iow.io-warnemuende.de/thredds/catalogs/publications/lorenz/catalog_Lorenz_Graewe_2023.html. The atmospheric forcings can be accessed from following links: UERRA https://doi.org/10.24381/cds.44ec8078, ERA5 https://doi.org/10.24381/cds.adbb2d47, CFSR https://www.ncei.noaa.gov/products/weather-climate-models/climate-forecast-system, coastDat1, coastDat2, coastDat3: personal communication from HEREON, general website https://www.coastdat.de/.

Author contributions. ML and UG designed the study together. UG created the model setups. ML performed the simulations, did the analyses, and wrote the first draft of the manuscript with input from UG.

Competing interests. The authors indicate no competing interests.

Acknowledgements. The authors thank Tarmo Soomere and one anonymous reviewer for their constructive feedback. This research is part of the ECAS-BALTIC project: Strategies of ecosystem-friendly coastal protection and ecosystem-supporting coastal adaptation for the German Baltic Sea Coast. The project is funded by the German Federal Ministry of Education and Research (BMBF, funding code 03F0860H). All simulations were performed on clusters of the German National High Performance Computing Alliance (NHR). Figures in this paper use the colourmaps of the cmocean package (Thyng et al., 2016).

References

Andersson, H. C.: Influence of long-term regional and large-scale atmospheric circulation on the Baltic Sea level, Tellus A: Dynamic Meteorology and Oceanography, 54, 76–88, https://doi.org/10.3402/tellusa.v54i1.12125, 2002.

Andrée, E., Su, J., Dahl Larsen, M. A., Drews, M., Stendel, M., and Skovgaard Madsen, K.: The role of preconditioning for extreme storm surges in the western Baltic Sea, Natural Hazards and Earth System Sciences, 23, 1817–1834, https://doi.org/10.5194/nhess-23-1817-2023, 2023.

Arns, A., Wahl, T., Haigh, I., Jensen, J., and Pattiaratchi, C.: Estimating extreme water level probabilities: A comparison of the direct methods and recommendations for best practise, Coastal Engineering, 81, 51–66, https://doi.org/10.1016/j.coastaleng.2013.07.003, 2013.

Arns, A., Wahl, T., Wolff, C., Vafeidis, A. T., Haigh, I. D., Woodworth, P., Niehüser, S., and Jensen, J.: Non-linear interaction modulates global extreme sea levels, coastal flood exposure, and impacts, Nature communications, 11, 1–9, https://doi.org/10.1038/s41467-020-15752-5, 2020.

Bamber, J. L., Oppenheimer, M., Kopp, R. E., Aspinall, W. P., and Cooke, R. M.: Ice sheet contributions to future sea-level rise from structured expert judgment, Proceedings of the National Academy of Sciences, 116, 11 195–11 200, https://doi.org/10.1073/pnas.1817205116, 2019.

Barbosa, S. M.: Quantile trends in Baltic sea level, Geophysical Research Letters, 35, https://doi.org/10.1029/2008gl035182, 2008.

Bierstedt, S. E., Hünicke, B., and Zorita, E.: Variability of wind direction statistics of mean and extreme wind events over the Baltic Sea region, Tellus A: Dynamic Meteorology and Oceanography, 67, 29 073, https://doi.org/10.3402/tellusa.v67.29073, 2015.

Bork, I. and Müller-Navarra, S. H.: Modellierung von extremen Sturmhochwassern an der deutschen Ostseeküste, Die Küste, 75 MUSTOK, pp. 71–130, 2009.

Burchard, H. and Bolding, K.: GETM: A General Estuarine Transport Model; Scientific Documentation, European Commission, Joint Research Centre, Institute for Environment and Sustainability, 2002.

Christensen, O. B., Kjellström, E., Dieterich, C., Gröger, M., and Meier, H. E. M.: Atmospheric regional climate projections for the Baltic Sea region until 2100, Earth System Dynamics, 13, 133–157, https://doi.org/10.5194/esd-13-133-2022, 2022.

Coles, S., Bawa, J., Trenner, L., and Dorazio, P.: An introduction to statistical modeling of extreme values, vol. 208, Springer, https://doi.org/10.1007/978-1-4471-3675-0, 2001.

Dieterich, C., Gröger, M., Arneborg, L., and Andersson, H. C.: Extreme sea levels in the Baltic Sea under climate change scenarios–Part 1: Model validation and sensitivity, Ocean Science, 15, 1399–1418, https://doi.org/10.5194/os-15-1399-2019, 2019.

Eelsalu, M., Soomere, T., Pindsoo, K., and Lagemaa, P.: Ensemble approach for projections of return periods of extreme water levels in Estonian waters, Continental Shelf Research, 91, 201–210, https://doi.org/10.1016/j.csr.2014.09.012, 2014.

Errico, R. M.: What is an adjoint model?, Bulletin of the American Meteorological Society, 78, 2577–2592, https://doi.org/10.1175/1520-0477(1997)078<2577:wiaam>2.0.co;2, 1997.

Feuchter, D., Jörg, C., Rosenhagen, G., Auchmann, R., Romppainen-Martius, O., and Brönnimann, S.: The 1872 Baltic Sea storm surge, https://doi.org/10.4480/GB2013.G89.10, 2013.

Fox-Kemper, B., Hewitt, H., Xiao, C., Aðalgeirsdóttir, G., Drijfhout, S., Edwards, T., Golledge, N., Hemer, M., Kopp, R., Krinner, G., Mix, A., Notz, D., Nowicki, S., Nurhati, I., Ruiz, L., Sallée, J.-B., Slangen, A., and Yu, Y.: Ocean, Cryosphere and Sea Level Change. In Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, https://doi.org/10.1017/9781009157896.011, 2021.

Geyer, B.: High-resolution atmospheric reconstruction for Europe 1948–2012: coastDat2, Earth System Science Data, 6, 147–164, https://doi.org/10.5194/essd-6-147-2014, 2014.

Geyer, B., Ludwig, T., and von Storch, H.: Limits of reproducibility and hydrodynamic noise in atmospheric regional modelling, Communications Earth & Environment, 2, 17, https://doi.org/10.1038/s43247-020-00085-4, 2021.

455 Gräwe, U. and Burchard, H.: Storm surges in the Western Baltic Sea: the present and a possible future, Climate Dynamics, 39, 165–183, https://doi.org/10.1007/s00382-011-1185-z, 2012.

Gräwe, U., Naumann, M., Mohrholz, V., and Burchard, H.: Anatomizing one of the largest saltwater inflows into the Baltic Sea in December 2014, Journal of Geophysical Research: Oceans, 120, 7676–7697, https://doi.org/10.1002/2015JC011269, 2015.

Gräwe, U., Klingbeil, K., Kelln, J., and Dangendorf, S.: Decomposing mean sea level rise in a semi-enclosed basin, the Baltic Sea, Journal
460 of Climate, 32, 3089–3108, https://doi.org/10.1175/jcli-d-18-0174.1, 2019.

Grinsted, A.: Projected Change—Sea Level, pp. 253–263, Springer International Publishing, Cham, https://doi.org/10.1007/978-3-319-16006-1_14, 2015.

Gröger, M., Placke, M., Meier, H., Börgel, F., Brunnabend, S.-E., Dutheil, C., Gräwe, U., Hieronymus, M., Neumann, T., Radtke, H., et al.: The Baltic Sea Model Intercomparison Project (BMIP)–a platform for model development, evaluation, and uncertainty assessment,
465 Geoscientific Model Development, 15, 8613–8638, https://doi.org/10.5194/gmd-15-8613-2022, 2022.

Haigh, I. D., Marcos, M., Talke, S. A., Woodworth, P. L., Hunter, J. R., Hague, B. S., Arns, A., Bradshaw, E., and Thompson, P.: GESLA Version 3: A major update to the global higher-frequency sea-level dataset, Geoscience Data Journal, https://doi.org/10.31223/x5mp65, 2022.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al.: The
470 ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Hieronymus, M., Dieterich, C., Andersson, H., and Hordoir, R.: The effects of mean sea level rise and strengthened winds on extreme sea levels in the Baltic Sea, Theoretical and Applied Mechanics Letters, 8, 366–371, https://doi.org/10.1016/j.taml.2018.06.008, 2018.

Hofstede, J. and Hamann, M.: The 1872 catastrophic storm surge at the Baltic Sea coast of Schleswig-Holstein; lessons learned?, Die Küste, https://doi.org/10.18171/1.092101, 2022.

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

Kara, A. B., Rochford, P. A., and Hurlburt, H. E.: Efficient and accurate bulk parameterizations of air–sea fluxes for use in general circulation models, Journal of Atmospheric and Oceanic Technology, 17, 1421–1438, https://doi.org/10.1175/1520-0426(2000)017<1421:eaabpo>2.0.co;2, 2000.

480 Karabil, S., Zorita, E., and Hünicke, B.: Contribution of atmospheric circulation to recent off-shore sea-level variations in the Baltic Sea and the North Sea, Earth System Dynamics, 9, 69–90, https://doi.org/10.5194/esd-9-69-2018, 2018.

Kärnä, T., Ljungemyr, P., Falahat, S., Ringgaard, I., Axell, L., Korabel, V., Murawski, J., Maljutenko, I., Lindenthal, A., Jandt-Scheelke, S., Verjovkina, S., Lorkowski, I., Lagemaa, P., She, J., Tuomi, L., Nord, A., and Huess, V.: Nemo-Nordic 2.0: operational marine forecast model for the Baltic Sea, Geoscientific Model Development, 14, 5731–5749, https://doi.org/10.5194/gmd-14-5731-2021, 2021.

485 Kiesel, J., Lorenz, M., König, M., Gräwe, U., and Vafeidis, A. T.: Regional assessment of extreme sea levels and associated coastal flooding along the German Baltic Sea coast, Natural Hazards and Earth System Sciences, 23, 2961–2985, https://doi.org/10.5194/nhess-23-2961-2023, 2023.

Klingbeil, K., Lemarié, F., Debreu, L., and Burchard, H.: The numerics of hydrostatic structured-grid coastal ocean models: state of the art and future perspectives, Ocean Modelling, 125, 80–105, https://doi.org/10.1016/j.ocemod.2018.01.007, 2018.

490 Kudryavtseva, N., Pindsoo, K., and Soomere, T.: Non-stationary Modeling of Trends in Extreme Water Level Changes Along the Baltic Sea Coast, Journal of Coastal Research, pp. 586–590, https://doi.org/10.2112/SI85-118.1, 2018.

Kudryavtseva, N., Soomere, T., and Männikus, R.: Non-stationary analysis of water level extremes in Latvian waters, Baltic Sea, during 1961–2018, Natural Hazards and Earth System Sciences, 21, 1279–1296, https://doi.org/10.5194/nhess-21-1279-2021, 2021.

Kärnä, T., Wallwork, J. G., and Kramer, S. C.: Efficient optimization of a regional water elevation model with an automatically generated 495 adjoint, 2023.

Large, W. and Yeager, S.: The global climatology of an interannually varying air–sea flux data set, Climate dynamics, 33, 341–364, https://doi.org/10.1007/s00382-008-0441-3, 2009.

Lehmann, A., Getzlaff, K., and Harlaß, J.: Detailed assessment of climate variability in the Baltic Sea area for the period 1958 to 2009, Climate Research, 46, 185–196, https://doi.org/10.3354/cr00876, 2011.

500 Madsen, K. S., Høyer, J. L., Fu, W., and Donlon, C.: Blending of satellite and tide gauge sea level observations and its assimilation in a storm surge model of the North Sea and Baltic Sea, Journal of Geophysical Research: Oceans, 120, 6405–6418, https://doi.org/10.1002/2015JC011070, 2015.

Männikus, R., Soomere, T., and Kudryavtseva, N.: Identification of mechanisms that drive water level extremes from in situ measurements in the Gulf of Riga during 1961-2017, Continental Shelf Research, 182, 22–36, https://doi.org/10.1016/j.csr.2019.05.014, 2019.

505 Männikus, R., Soomere, T., and Viška, M.: Variations in the mean, seasonal and extreme water level on the Latvian coast, the eastern Baltic Sea, during 1961–2018, Estuarine, Coastal and Shelf Science, 245, 106 827, https://doi.org/https://doi.org/10.1016/j.ecss.2020.106827, 2020.

Meier, H. E. M., Dieterich, C., Gröger, M., Dutheil, C., Börgel, F., Safonova, K., Christensen, O. B., and Kjellström, E.: Oceanographic regional climate projections for the Baltic Sea until 2100, Earth System Dynamics, 13, 159–199, https://doi.org/10.5194/esd-13-159-2022, 510 2022a.

Meier, H. E. M., Kniebusch, M., Dieterich, C., Gröger, M., Zorita, E., Elmgren, R., Myrberg, K., Ahola, M. P., Bartosova, A., Bonsdorff, E., Börgel, F., Capell, R., Carlén, I., Carlund, T., Carstensen, J., Christensen, O. B., Dierschke, V., Frauen, C., Frederiksen, M., Gaget, E., Galatius, A., Haapala, J. J., Halkka, A., Hugelius, G., Hünicke, B., Jaagus, J., Jüssi, M., Käyhkö, J., Kirchner, N., Kjellström, E., Kulinski, K., Lehmann, A., Lindström, G., May, W., Miller, P. A., Mohrholz, V., Müller-Karulis, B., Pavón-Jordán, D., Quante, M., Reckermann, 515 M., Rutgersson, A., Savchuk, O. P., Stendel, M., Tuomi, L., Viitasalo, M., Weisse, R., and Zhang, W.: Climate change in the Baltic Sea region: a summary, Earth System Dynamics, 13, 457–593, https://doi.org/10.5194/esd-13-457-2022, 2022b.

Meier, H. M., Broman, B., and Kjellström, E.: Simulated sea level in past and future climates of the Baltic Sea, Climate research, 27, 59–75, https://doi.org/10.3354/cr027059, 2004.

Muis, S., Verlaan, M., Winsemius, H. C., Aerts, J. C., and Ward, P. J.: A global reanalysis of storm surges and extreme sea levels, Nature 520 communications, 7, 1–12, https://doi.org/10.1038/ncomms11969, 2016.

Muis, S., Aerts, J., Antolínez, J. A. Á., Dullaart, J., Duong, T. M., Erikson, L., Haarmsa, R., Apecechea, M. I., Mengel, M., Le Bars, D., et al.: Global projections of storm surges using high-resolution CMIP6 climate models: validation, projected changes, and methodological challenges, Authorea Preprints, https://doi.org/10.1002/essoar.10511919.1, 2022.

Peltier, W. R.: Global glacial isostasy and the surface of the ice-age Earth: the ICE-5G (VM2) model and GRACE, 525 https://doi.org/10.1146/annurev.earth.32.082503.144359, 2004.

Peng, S., Li, Y., and Xie, L.: Adjusting the wind stress drag coefficient in storm surge forecasting using an adjoint technique, Journal of Atmospheric and Oceanic Technology, 30, 590–608, https://doi.org/10.1175/jtech-d-12-00034.1, 2013.

Pietrzak, J.: The use of TVD limiters for forward-in-time upstream-biased advection schemes in ocean modeling, Monthly Weather Review, 126, 812–830, https://doi.org/10.1175/1520-0493(1998)126<0812:tuotlf>2.0.co;2, 1998.

530 Pindsoo, K. and Soomere, T.: Basin-wide variations in trends in water level maxima in the Baltic Sea, Continental Shelf Research, 193, 104 029, https://doi.org/10.1016/j.csr.2019.104029, 2020.

Reinert, M., Pineau-Guillou, L., Raillard, N., and Chapron, B.: Seasonal shift in storm surges at Brest revealed by extreme value analysis, Journal of Geophysical Research: Oceans, p. e2021JC017794, https://doi.org/10.1029/2021jc017794, 2021.

Ridal, M., Olsson, E., Unden, P., Zimmermann, K., and Ohlsson, A.: Uncertainties in Ensembles of Regional Re-Analyses–Deliverable D2.
535 7 HARMONIE reanalysis report of results and dataset, 2017.

Rosenhagen, G. and Bork, I.: Rekonstruktion der Sturmwetterlage vom 13. November 1872, Die Küste, 75 MUSTOK, pp. 51–70, 2009.

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

540 Saha, S., Moorthi, S., Pan, H.-L., Wu, X., Wang, J., Nadiga, S., Tripp, P., Kistler, R., Woollen, J., Behringer, D., et al.: The NCEP climate forecast system reanalysis, Bulletin of the American Meteorological Society, 91, 1015–1058, https://doi.org/10.1175/2010bams3001.1, 2010.

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., Behringer, D., Hou, Y.-T., Chuang, H.-y., Iredell, M., et al.: The NCEP climate forecast system version 2, Journal of climate, 27, 2185–2208, https://doi.org/10.1175/jcli-d-12-00823.1, 2014.

545 Smith, S. D., Anderson, R. J., Oost, W. A., Kraan, C., Maat, N., De Cosmo, J., Katsaros, K. B., Davidson, K. L., Bumke, K., Hasse, L., et al.: Sea surface wind stress and drag coefficients: The HEXOS results, Boundary-layer meteorology, 60, 109–142, https://doi.org/10.1007/bf00122064, 1992.

Soomere, T. and Pindsoo, K.: Spatial variability in the trends in extreme storm surges and weekly-scale high water levels in the eastern Baltic Sea, Continental Shelf Research, 115, 53–64, https://doi.org/10.1016/j.csr.2015.12.016, 2016.

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

Tadesse, M. G. and Wahl, T.: A database of global storm surge reconstructions, Scientific Data, 8, 1–10, https://doi.org/10.1038/s41597-021-00906-x, 2021.

Tadesse, M. G., Wahl, T., Rashid, M. M., Dangendorf, S., Rodríguez-Enríquez, A., and Talke, S. A.: Long-term trends in storm surge climate
555 derived from an ensemble of global surge reconstructions, Scientific reports, 12, 1–18, https://doi.org/10.1002/essoar.10511038.1, 2022.

Thyng, K. M., Greene, C. A., Hetland, R. D., Zimmerle, H. M., and DiMarco, S. F.: True colors of oceanography: Guidelines for effective and accurate colormap selection, Oceanography, 29, 9–13, https://doi.org/10.5670/oceanog.2016.66, 2016.

Vousdoukas, M. I., Mentaschi, L., Voukouvalas, E., Verlaan, M., and Feyen, L.: Extreme sea levels on the rise along Europe's coasts, Earth's Future, 5, 304–323, https://doi.org/10.1002/2016ef000505, 2017.

560 Wahl, T., Haigh, I. D., Nicholls, R. J., Arns, A., Dangendorf, S., Hinkel, J., and Slangen, A. B.: Understanding extreme sea levels for broad-scale coastal impact and adaptation analysis, Nature communications, 8, 1–12, https://doi.org/10.1038/ncomms16075, 2017.

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

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

Wolski, T. and Wiśniewski, B.: Characteristics and Long-Term Variability of Occurrences of Storm Surges in the Baltic Sea, Atmosphere,
12, 1679, https://doi.org/10.3390/atmos12121679, 2021.

Wolski, T. and Wiśniewski, B.: Characteristics of seasonal changes of the Baltic Sea extreme sea levels, Oceanologia,
570   https://doi.org/10.1016/j.oceano.2022.02.006, 2022.

Wolski, T., Wiśniewski, B., Giza, A., Kowalewska-Kalkowska, H., Boman, H., Grabbi-Kaiv, S., Hammarklint, T., Holfort, J., and Lydeikaitė,
Ž.: Extreme sea levels at selected stations on the Baltic Sea coast, Oceanologia, 56, 259–290, https://doi.org/10.5697/oc.56-2.259, 2014.

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