Answers to Reviewer 1 (Marta Marcos)

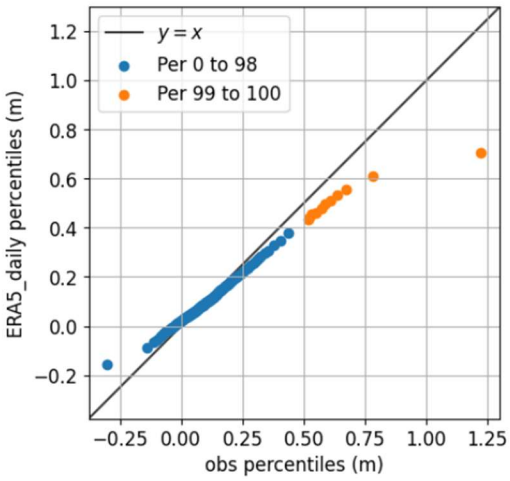
This manuscript presents a statistical reconstruction of storm surge records using 17 climate model projections along the European coastlines. It is the result of a significant computational effort, combining a (relatively) small set of dynamical numerical simulations and a data-driven model based on multiple linear regression. The methods are well described and sound. All the details are provided for the calibration and choices of the statistical model parameters. However, I have some concerns on the application of the model, and the presentation and interpretation of the results. I think the manuscript requires a major revision before being suitable for publication.

We thank the reviewer for their comprehensive review, which has helped to greatly improve the manuscript. Before answering to the comments, we'd like to highlight that the other reviewer requested a major reorganization of the paper structure, better separating between Methods and Results. We add here the new structure to guide the reviewer through the answers below, although we have tried to point out in each answer the previous and new section numbering when necessary.

1. Introduction
2. Methods
    2.1. General workflow (new)
    2.2. Training and benchmark datasets
    2.3. Statistical downscaling model
    2.4. Experimental design (new)
        2.4.1. Validation under climate forcing
        2.4.2. Multi-model ensemble projections
3. Results
    3.1. Statistical hindcast reconstructions (new)
    3.2. Validation of statistical projections
        3.2.1. Stationarity assumption
        3.2.2. Extrapolation to climate forcing
    3.3. Statistical ensemble projections

Comment1*: One major concern is the focus on extreme storm surges. The results of the statistical model applied to ERA5 forcing fields are daily records of storm surges that, when compared with the benchmark results of the dynamic simulations, provide satisfactory performance in terms of averaged storm surges. This is shown in Figures 3 and 4 that show the capabilities of the statistical method in terms of correlation and RMSE. The only metric focusing partly on extremes is the bias of the 99th percentile, but this is not representative of the ability of the statistical records to capture extreme events and reproduce return levels, which are the results that are analysed later on. In fact, the comparisons in figure 4 show that discrepancies can be quite large for individual events. This is something well known for the statistical model based on Tadesse et al (2020). The approach is good in simulating the mean storm surge climate but displays limited accuracy with the extremes, and this is a major shortcoming that must be reflected in the manuscript. I attach below an example for a tide gauge in Brest that I produced some time ago for an assessment of data-driven models. Although the differences with the dynamic simulation are expected to be smaller than in this example, as the extremes will be also underestimated (shown e.g. in Figure 1), it is clear that the statistical model is not particularly*

well suited for extreme storm surges. I am not suggesting that the authors should change the statistical approach, I believe it has its value. I think, though, that the ability of the method needs to be better described, particularly concerning extremes. To do so, I suggest using qq-plots instead of time series to evaluate model performance. Likewise, mapping the differences in maxima or yearly maxima and/or return levels between statistical and dynamical approaches forced by ERA5 would provide the required information to the reader.



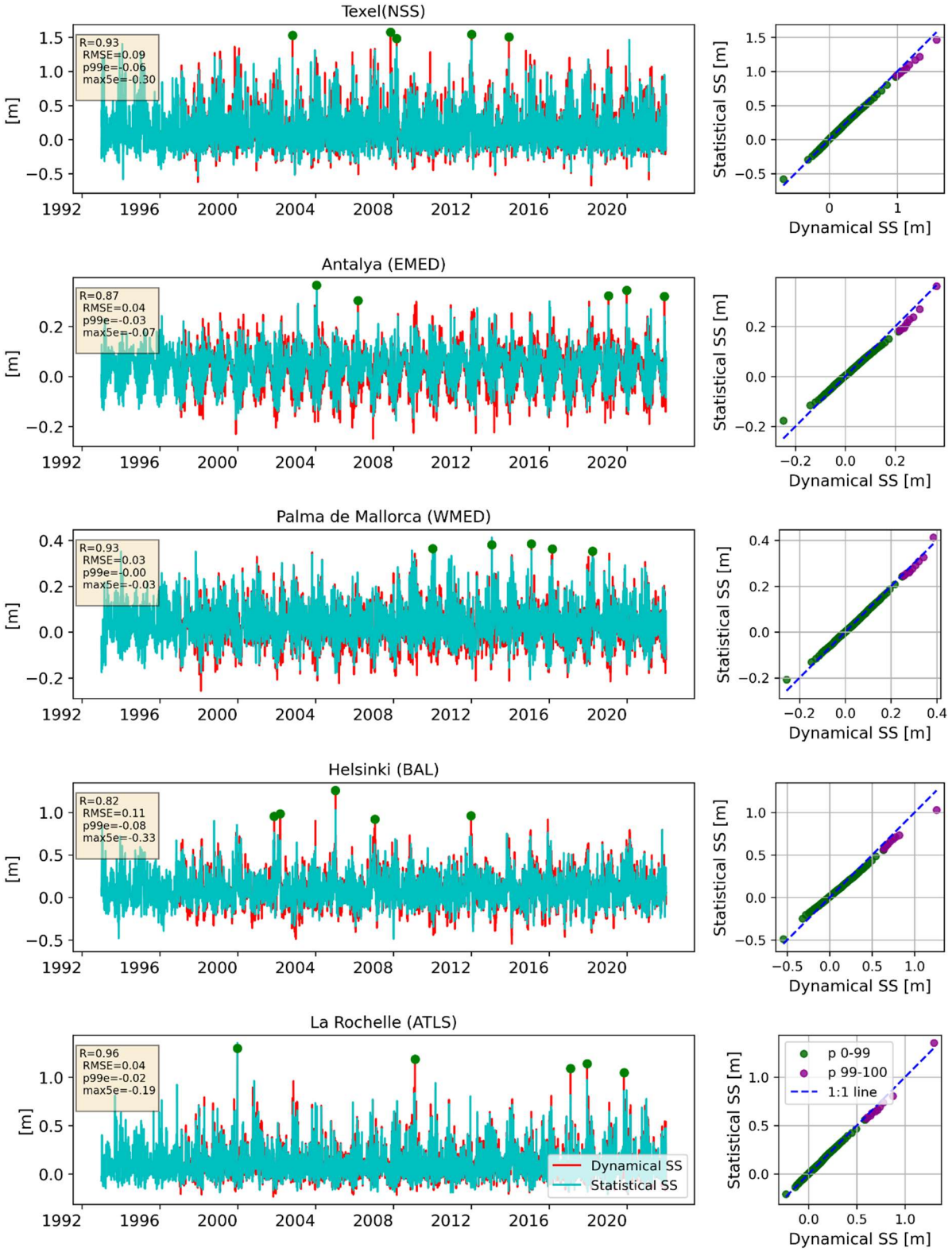
Answer to comment 1: We thank the reviewer for her useful remarks. Indeed, given that the study focuses on projections of extremes, the performance of the statistical model for extremes needs to be better demonstrated. The calibration focuses on correlation and RMSE to illustrate that the statistical model produces a storm surge signal that is coherent with the dynamical model, showing that the predictors and regression encompass sufficiently realistic physical relationships to replicate storm surge dynamics, which is important for the credibility of the statistical model. Notably, our study did not intend to elaborate a new statistical downscaling method (i.e. a methodological paper), but to evaluate the capability of existing methods for projections, which hasn't been proved in literature so far. The adopted dynamical-statistical downscaling framework enables this by providing a benchmark for future storm surges (given by the numerical simulations). This has been now highlighted in the Introduction:

Lines 80-84 (of revised manuscript, similarly for all answers below unless specified otherwise): "*Rather than developing a new statistical approach, we adopt an existing method used for broad-scale storm-surge reconstructions — multi- linear regression — and assess its capability for projecting ESSs, which has not yet been demonstrated. The adopted framework enables this evaluation by using dynamically downscaled projections as a benchmark, which is not possible for observations-based statistical downscaling*".

In our study, we have shown the mean bias for the top 1% of the data. We note that this represents more extreme events than the 99th percentile. In fact, we see that such bias is very similar to the error in the 10-year return level. Nevertheless, for completeness, we have modified the manuscript with the following items to better illustrate performance for extremes:

- We have adapted Figure4 (now Figure 5) for better visualization of the timeseries and individual events, and including the Q-Q plots on the right-hand side, in a similar fashion as proposed by the reviewer, with percentiles>99 (99-100, every 0.1 pct) in a different color to illustrate performance for extreme events in the data. The description of these new results have been added in lines 380-389. These plots show a satisfactory

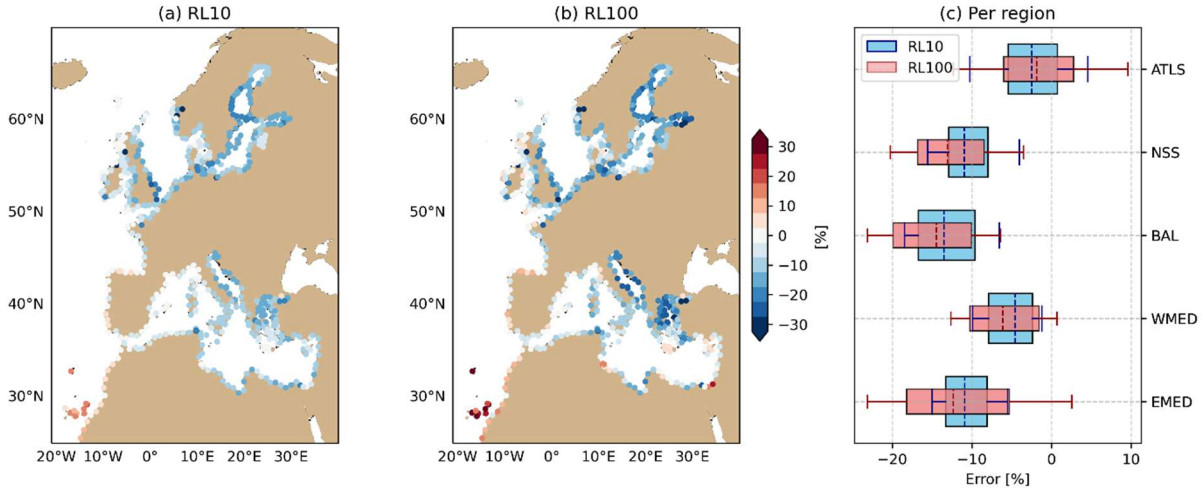
performance for high percentiles (for this subset of stations at least), other than for the station in the Baltic. For individual events marked in green, performance varies across events, some are very well captured and others are underestimated. Based on these results, we do attribute a general good skill to the statistical model for the representation of sampled extremes, although knowing that for some events it will not perform well. Since we don't evaluate future changes based on individual events, but on statistics (return levels), we complement these results by performing a dedicated validation for point estimates (see next point).



Lines 381-390:" *Time series and quantile-quantile plots for the statistical hindcast reconstruction in selected example locations across different European seas (Figure 5) illustrate the skill of the SDM to accurately reproduce the storm surge signal relative to dynamical simulations. Poorer performance in the representation of general storm surge variability is observed in Antalya and Helsinki (correlations of 0.87 and 0.82 vs. > 0.93 in the others), in line with the cross-validation analysis which showed poorer performance in the eastern*

Mediterranean and the Baltic Sea, respectively (section 2.3). The overall good agreement between statistical and dynamical estimates extends into the extreme tail – represented by the 99–100th percentiles at 0.1-percentile resolution. The Baltic station is the main exception, reflecting a lower SDM skill for extreme conditions in this region. For the largest 5 events in the series (green circles), performance strongly depends on the specific extreme event at hand and is largest for Texel (mean error of -30cm) and Helsinki (-33cm). For a thorough evaluation of extreme events across Europe, a dedicated evaluation is carried out next using extreme-value theory."

- We have added a figure (Figure 6) illustrating the statistical hindcast performance for the point estimates at 10 and 100-year return levels. The figures show the relative error in %, to evaluate whether the performance for higher vs lower return levels decreases (as absolute errors are expected to be larger for larger-magnitude events). The plot does show a moderately decreasing performance for RL100 vs RL10. 90% of the coastal locations show errors smaller than 16 and 21% for RL10 and RL100 respectively. The drop in performance for RL100 vs RL10 is most pronounced for EMED and BAL. This analysis is described in lines 400-420 and is now used as justification to focus on 10-year changes in projections.



Spatial plots of the relative error (%) in the 10-year (RL10, a) and 100-year (RL100, b) return levels between the statistical and the dynamical models, calculated using stationary extreme value analysis over 1997-2021. (c) Corresponding regional box plots, with boxes covering the interquartile range (25th-75th percentiles), whiskers extending between the 10th-90th percentiles, and dashed lines indicating the median in each region. EMED: Eastern Mediterranean Sea; WMED: Western Mediterranean Sea; BAL: Baltic Sea; NSS: North Sea; ATLS: Atlantic Shelf. See Figure 1 for the geographical coverage of each region.

Lines 400-420: "Extremes are evaluated focusing on the SDM skill for the 10-year (RL10) and 100-year (RL100) return levels relative to the dynamical hindcast (Figure 6) using the chosen EVA method (section 2.4). As suggested by previous results, the SDM systematically underpredicts extremes, except around the Canary Islands and Moroccan Atlantic coast. For the 10-year return level, relative errors average -9% across Europe, with reduced skill in the Baltic (-13%; down to -18% in the southern Gulf of Bothnia), the Adriatic and Aegean Seas (-14%), and locally along southeastern UK (-20%). Errors for the 100-year return level show a similar pattern, with a moderate amplification overall (average absolute error change of +2%) but more pronounced (+7-10%) in the Gulf of Finland, the southern Adriatic sea, the Aegean sea, and around the Canary Islands. As a result, regional boxplots (Figure 6-c) highlight the Baltic and eastern

Mediterranean as regions where negative biases reach markedly higher values (<-20%) for the 100 vs. 10-year return level, reflecting extensive coastal areas with amplified errors for more extreme events. Although errors in statistical estimates of ESSs across Europe remain overall modest (with 90% of all coastal points exhibiting errors with absolute values smaller than 16% and 21% for the 10- and 100-year return levels, respectively), they indicate lower confidence in SDM-based estimates of ESSs for regions such as the Baltic and eastern Mediterranean, and extending to the North Sea for high return periods, which should be considered when interpreting corresponding statistical climate projections. The increase of error for increasing storm surge magnitudes suggests that the storm-surge–predictor relationship departs from linearity between average and extreme conditions. This reflects the ordinary least squares formulation, which optimizes the mean response and leads to heteroscedastic errors for rare extremes; this behaviour is inherent to the methodology and is not expected to change qualitatively when the model is driven by bias-adjusted CMIP6 predictors.

Based on these results, we decide to focus statistical projections in the following to the 10-year storm-surge event to limit the impact of the decreasing SDM performance for high return periods on the confidence of the target statistical ensemble projections."

Comment2: *A second major concern is related to the discrepancies between dynamical and statistical simulations in climate models for some particular regions. As shown in Figure 5, regions as the Mediterranean (CNRM-CM6-1-HR) and the Baltic (MPIESM1-2-HR) indicate opposite changes in projected storm surges using dynamical and statistical models. To a lesser extent, also the western of the British Isles and the southern North Sea display different patterns. This needs to be clearly described. I do not think that the patterns are similar and only the magnitudes change, as claimed the lines 289-290. I agree, though, that using the hindcast instead of historical simulations for the training is mostly fine. These discrepancies hinder the interpretability of the projected storm surges presented in figure 7. The regions where the statistical and dynamical models clearly differ should not be discussed or even mapped. This includes the Baltic and the Eastern Mediterranean Seas (the western Mediterranean seems consistent in the validation). In addition, the uncertainty range is very high for the 100-year return levels, ranging from negative to positive values. This indicates that there is no confidence in the multimodel ensemble means (panels 7i and 7m). I suggest focussing only in the 10-year return level. This is also consistent with the fact that the statistical model is less reliable for the most extreme events.*

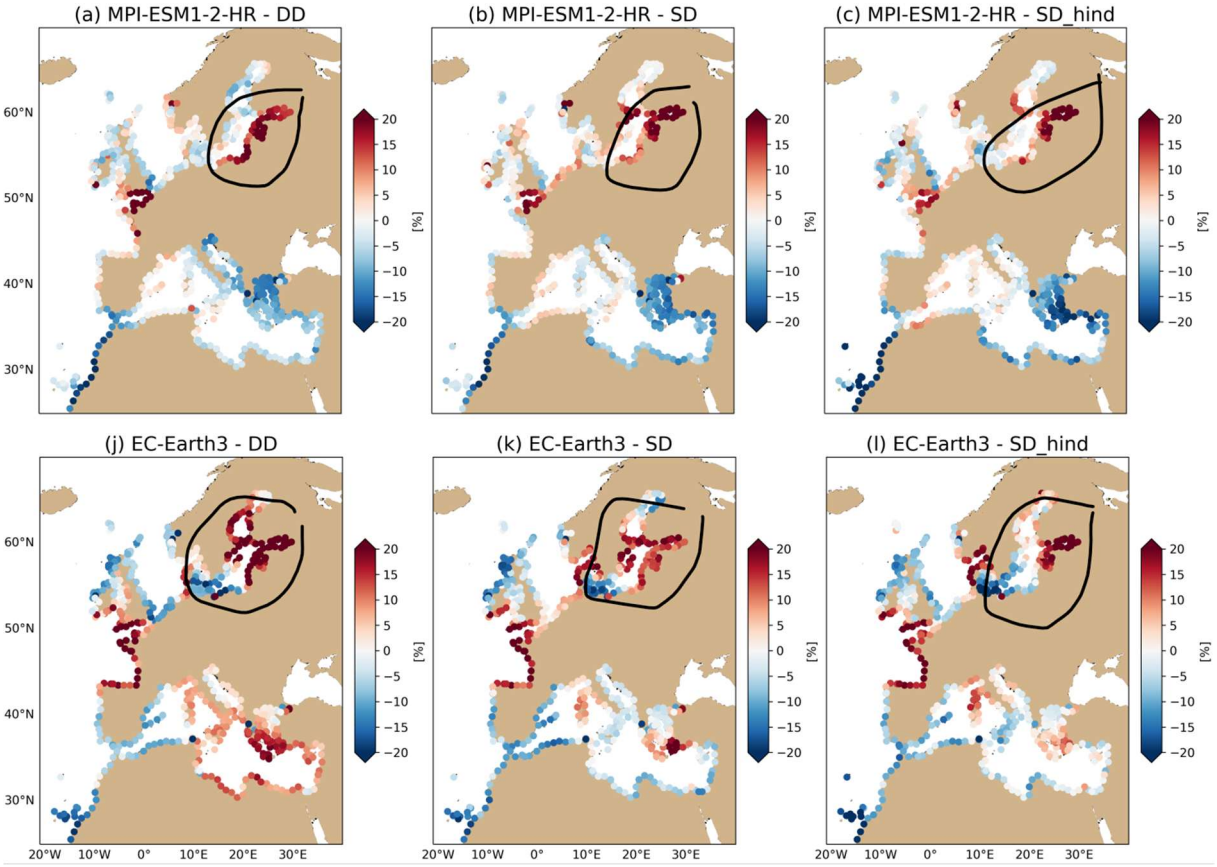
Answer to comment 2: Thank you for your insightful remark. We realize that the current description is indeed too shallow and results in Figure 6 (now Figure 8) deserve a more thorough description given the nuances observed between GCMs. Before diving into these nuances, we have better highlighted in the manuscript the objective of the developed SDM – finding a single configuration that delivers optimal performance at European scale (that is, without a dedicated site or basin- specific configuration), such that the model can be subsequently applied for European-scale ESS surges. This is important because there are probably ways of optimizing the configuration for a given site or region (in terms of the configuration choices of predictors, domain size and lag, and probably others not considered in our broad-scale application), but it is not our objective to do so here. The focus is rather on exploring the usability of the PCA+MLR-based SDM for projections when optimized at broad scale such as done in other studies (e.g. global scale for Tadese et al. 2020):

Lines 183-186 (beginning section calibration 2.3): "*First, an SDM selection phase is conducted to identify the optimal SDM configuration for the representation of daily maxima storm-surge along*

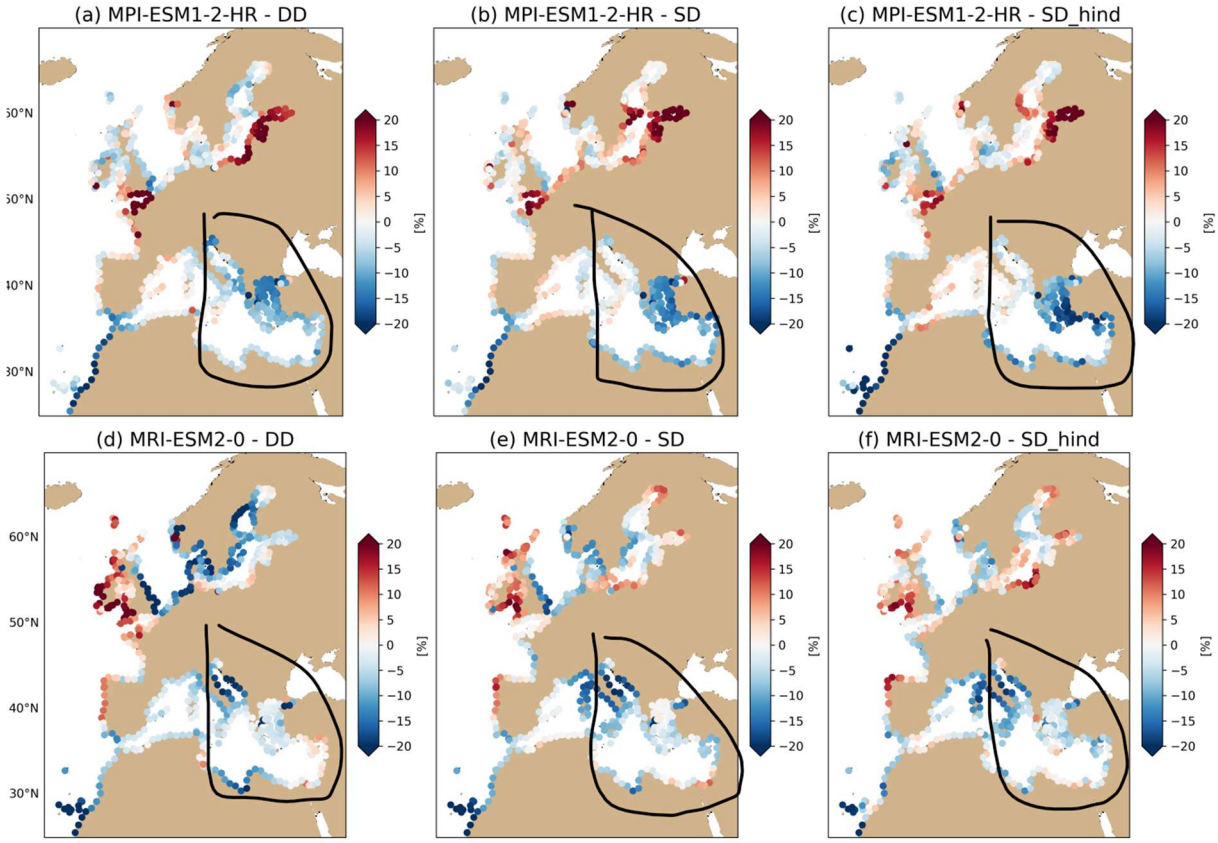
the European coastline, based on the SDM skill to reconstruct the hindcast simulation under ERA5 forcing degraded to 1°. The objective is to select a single configuration that delivers optimal performance at the European scale and can subsequently be used for European-scale ESS projections."

It was highlighted in lines 291-294 of the submitted manuscript that statistical vs. dynamical projections for the Baltic and eastern Mediterranean notably differed for 2/4 models, as you point out, and that this could be associated with the lower performance of the statistical model in these regions identified during calibration/validation for the hindcast. However, it is also true that in the other 2 models (50% of the models) the match between statistical and dynamical estimates is quite good in these regions:

- **Baltic**: for MPI-ESM1-2-HR both DD and SD project positive changes along the eastern Baltic; for EC-Earth3 a more widespread positive pattern is projected by DD across the Baltic, with negative changes in the Danish straits, and these features are largely kept (though attenuated) when moving to SD except for the norther Gulf of Bothnia. For EC-Earth3, its actually when moving onto the hindcast trained statistical estimate (SD_hind) that we lose much of that positive signal (e.g. north-western Baltic).



- **Eastern Mediterranean**: for MPI-ESM1-2-HR, for DD and SD project negative future changes of comparable magnitude and with larger values around the Aegean Sea. For MRI-ESM2-0-HR , patterns are also quite similar, with small changes in the Levantine basin but substantial negative changes in the Adriatic sea across datasets (though negative changes in the Gulf of Sidra fade away as we move through DD-SD-SD_hind).



Given that for 50% of the models estimates don't match, but for the other 50% they broadly do, we cannot categorically say that performance in projecting future changes is poor in these regions and invalidate the statistical projections based on the 17 GCMs for these regions. In fact, results in Figure 6 (now Figure 8) show that despite the lower performance of the statistical model identified in the calibration/validation for the Baltic and eastern Mediterranean, projected changes may still be reasonably well captured for some models. Our assessment is limited by the use of 4 GCMs only, having dynamical projections for the full 17-GCM ensemble would probably shed light on the capability of the statistical model to project future changes in these more challenging regions. From our 4 GCMs, it seems as if the statistical model is skillful when future changes are of a (general) given sign in these regions: changes are well reproduced by the statistical model when they are negative in the eastern Mediterranean (MPI-ESM1-2-HR and MRI-ESM2-0) and positive in the Baltic (MPI-ESM1-2-HR and MRI-ESM2-0), while they are not well reproduced when they are positive in the easter Mediterranean (CNRM-CM6-1-HR and EC-Earth3) and negative in the Baltic (MRI-ESM2-0 and CNRM-CM6-1-HR). These might indicate that future ESS changes are only well reproduced by the statistical model when associated with specific drivers or mechanism. These subjects warrant further investigation in future works. We have substantially extended the discussion in section 3.1.2 (now section 3.2.2, after reorganizing the paper based on the other reviewer's remarks) to describe all of these features and nuances.

Additionally, to add a quantitative dimension to the results, we have computed performance metrics to evaluate the skill of the SDM in reproducing projected ESS changes. We have included in each panel in Figure 6 (now Figure 8) the mean absolute bias and the pattern correlation coefficient across all coastal points, to illustrate errors in both the amplitude and the spatial structure of the climate change signal. We have also included the proportion of points for which the sign of the changes is correctly reproduced, as this will determine the inter-model

agreement in the subsequent ensemble projections. Furthermore, to illustrate the heterogeneous skill across regions, we have added an extra column showing the mean absolute bias for each region. The mean absolute bias is chosen as it reflects both magnitude and spatial structure errors. These new results illustrate differences in regional performance across GCMs, and the impact of switching to a hindcast trained SDM. They highlight the pronounced difficulties for the Baltic and eastern Mediterranean for 2 of the GCMs. Finally, we have added a Figure (9) showing the 3 computed metrics for each region when pooling data across GCMs to summarize overall performance. Given the substantial additions, we copy here the whole section (3.2.2, previously 3.1.2), lines 454-537:

==========section 3.2.2: Extrapolation to climate forcing==========

Once the stationarity assumption validated, we next evaluate the extrapolation capability of the hindcast-trained SDM to climate forcing by assessing its skill to reproduce dynamically downscaled changes in ESSs (section 2.5.1), and which hence constitutes the ultimate test to justify its application for multi-model ensemble projections of changes in ESS (section 2.5.2) focusing on the 10-year storm surge level.

*Dynamical projections reveal considerable inter-model spread, with regional changes typically reaching ±20% (Figure 8-a,e,i,m, 5^{th}-95^{th} percentiles of results pooled across GCMs), exceptionally higher (-25%/+32%, 1^{st}/99^{th} percentiles respectively). Statistical projections trained independently on each GCM (SD, Figure 8*Figure *-b,f,j,n) replicate the main European-scale spatial features of the dynamically downscaled projections, demonstrating the SDM's skill to replicate GCM-specific climate responses: mean absolute biases (MAB) remain modest (5–7.5%) and pattern correlation coefficients (PCC) for three of the four GCMs are >=0.7 (CNRM-CM6-1-HR being the exception with PCC=0.52), which is often deemed satisfactory in climate model performance evaluations (Back et al., 2024; Berhanu et al., 2025; Zebaze et al., 2025). Additionally, the sign of the projected change (sign agreement, SA) is correctly reproduced at ≥70% of grid points (67% for CNRM-CM6-1-HR). These indicators show that, despite some amplitude biases, the spatial imprint of the projected changes in the 10-year return level based on dynamical simulations is reasonably well reproduced by the SDM.*

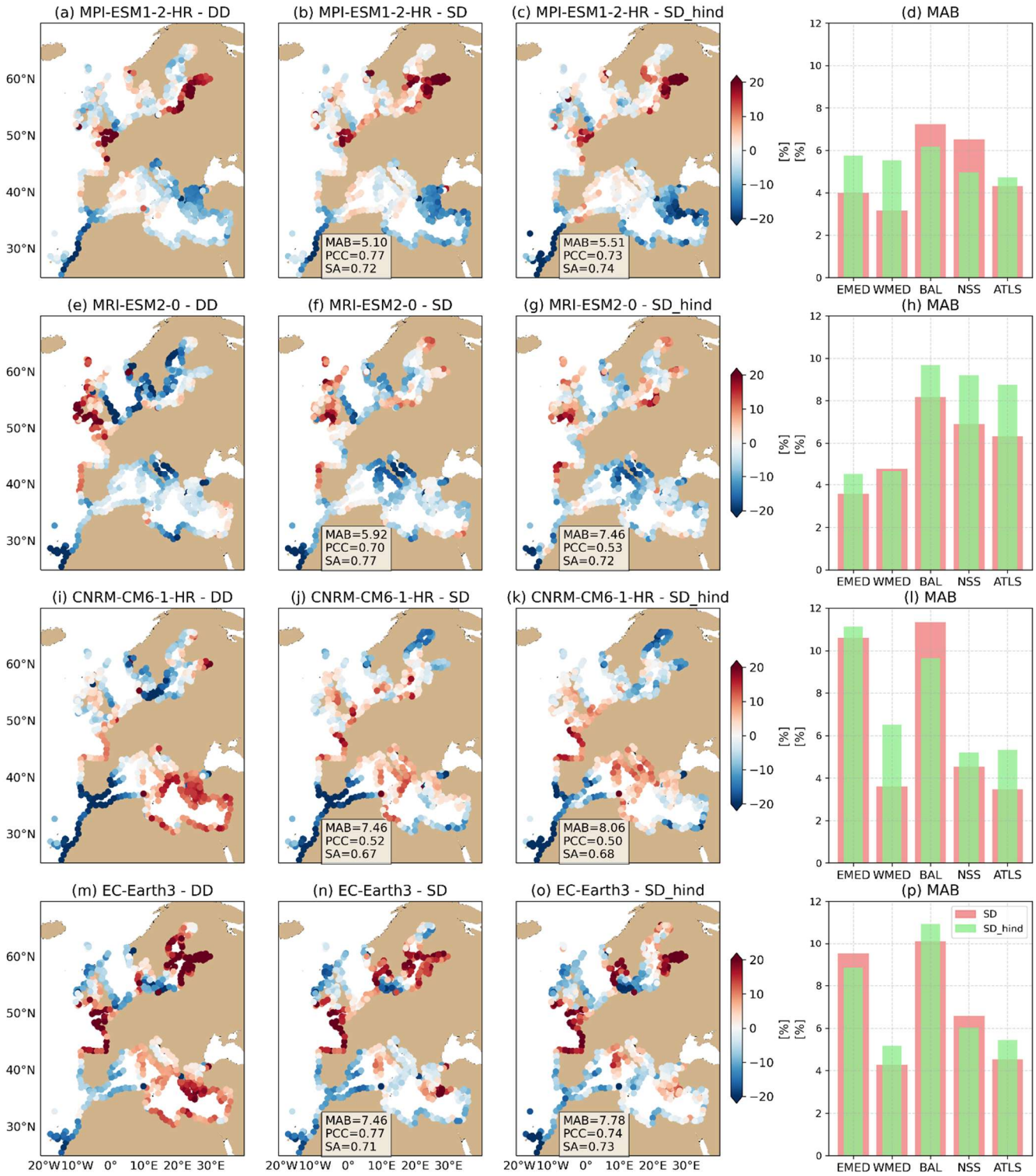


Figure 8 Projected changes (%, extreme value analysis on [2080-2099] vs [1995-2014]) in the 1 in 10 year storm surge event (RL10) for dynamical climate simulations (DD, a,e,i,m), statistical estimates trained on each historical climate simulation (SD, b,f,j,n) and statistical estimates trained on the hindcast forced by ERA5 (SD_hind, c,g,k,o) for the 4 GCMs downscaled. For each GCM (each row), the regionally averaged mean absolute bias (MAB) of SD and SD_hind estimates relative to DD estimates are given on the right-most column (d,h,l,p). EMED: Eastern Mediterranean Sea; WMED: Western Mediterranean Sea; BAL: Baltic Sea; NSS: North Sea; ATLS: Atlantic Shelf. See Figure 2 for the definition of the geographical coverage of each region. For a fair comparison between SD and SD_hind, both are trained on 20-yr periods (1995-2014 and 1997-2016 respectively).

Performance, however, varies considerably across regions (Figure 8-d,h,l,q). The Baltic Sea exhibits the largest amplitude errors for most GCMs, and both the Baltic and eastern Mediterranean perform notably worse for CNRM-CM6-1-HR and EC-Earth. While this could be attributed to the lower hindcast skill identified in these regions (Figure 3, Figure 6), performance in reproducing projected changes is not systematically lower for these regions across GCMs. In the eastern Mediterranean, poor SDM performance appears only for the two models projecting positive ESS changes (CNRM-CM6-1-HR and EC-Earth3), while for the other two GCMs (MPI-ESM1-2-HR and MRI-ESM2-0), the projected negative ESS changes with differing spatial patterns are well reproduced by the statistical model. In the Baltic Sea, dynamical ESS changes for MRI-ESM2-0 and CNRM-CM6-1-HR—which are broadly negative—are very poorly reproduced by the statistical model, whereas for MPI-ESM1-2-HR and EC-Earth3 the positive signal seen in the dynamical simulations is broadly retained in the statistical projections, albeit with reduced amplitude.

These results indicate that a limited hindcast skill of the statistical model does not necessarily imply a corresponding poor performance in projecting ESS changes. This likely depends on how well the SDM captures the effect of the dominant atmospheric predictors and associated variability modes driving future ESS changes: if these are well represented, the main climate-change signal can still be recovered even when other predictors or specific modes are less accurately represented. While out of scope for the current study, a more detailed analysis of the predictors and variability modes dominating projected ESS changes across GCMs should be pursued in future works to help clarify and better interpret the SDM's skill for climate projections, particularly in challenging regions such as the Baltic Sea and the eastern Mediterranean Sea.

When using the SDM trained solely on the hindcast (SD_hind, Figure 8-c,g,k,o), spatial patterns and relative amplitudes of changes in the 10-year return level are generally well preserved across GCMs, as reflected by the performance metrics (MAB, PCC and SA) which remain broadly comparable to those for GCM-specific statistical projections (SD). The exception is MRI-ESM2-0, for which performance decays substantially between SD and SD_hind in both the spatial pattern (PCC) and the amplitude (MAB) of the signal. This decay is largely owed to a pronounced reduction of the ESS change signal across the northwest Shelf (UK coasts, North Sea) (see regional metrics in Figure 8-h). Across GCMs, the transition to a hindcast-trained SDM tends to only moderately amplify regional amplitude biases (Figure 8-d,h,l,q). Overall, these results support the applicability of the SD_hind setup for climate projections, with the added advantage of requiring a single simulation for training (the hindcast). However, differences with GCM-specific statistical projections (SD) can be notable for some coastal sections, which might be explained by the fact that ERA5-based EOFs do not always fully explain GCM predictor variability for specific models and regions. As such, SD_hind estimates can only account for future storm surge changes linked to the identified ERA5 principal components, and not to novel atmospheric conditions or different modes of variability/covariance structures that may be present in GCMs. An analysis of the retained explained variance after projecting GCM fields onto hindcast-based principal components for the target 17-GCM ensemble (Fig S3, historical climate) reveals generally strong representativity across Europe, though with notable reductions for certain models and regions (e.g. the Mediterranean Sea for HadGEM3-GC31-MM). The retained variance also reveals stable between historical and end-of-century climates (not shown), supporting the stationarity of the predictands. Further analyses are needed to understand the observed differences between ERA5 and GCM variability and their impact on ESS change projections.

We finally compute performance metrics by pooling projections across GCMs for each region (Figure 9). Results confirm the eastern Mediterranean and the Baltic to be the worst performing regions, and notably so for the Baltic, with highest MAB (9.2%) and lowest PCC and SA (0.59 and 0.72, respectively, for SD). Given that dynamical simulations highlight these regions to display relatively strong future ESS changes (Figure 9-a), these results highlight the need to improve statistical projections in these regions for reliable future storm surge hazard assessments. The ensemble statistical projections presented hereafter should therefore be interpreted with caution in the Baltic and eastern Mediterranean regions. The best performing regions are the Atlantic façade and the western Mediterranean with lowest MAB (<5% for SD) and highest PCC and SA (>0.8 and >0.85, respectively, for SD). For the North Sea, results are somewhat mixed, as amplitude errors are moderate, the sign of ESS changes is well captured but the spatial pattern is less well resolved. The switch to a hindcast-trained SDM incurs a general but moderate decay in the SDM performance across regions, through with a notably larger impact on the western Mediterranean Sea.

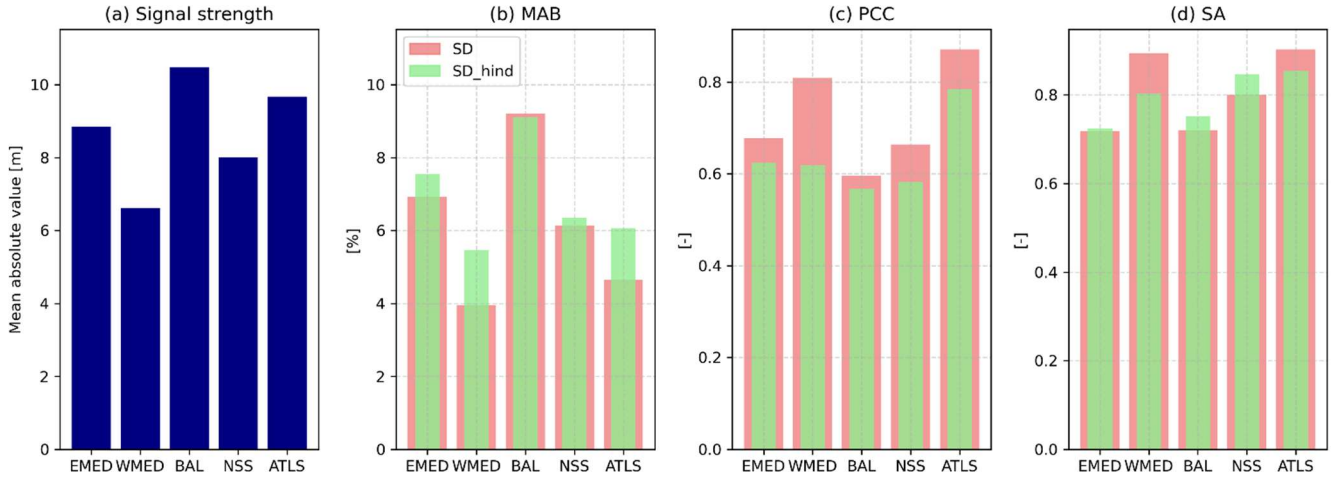


Figure 9 Regional performance metrics of statistically downscaled projections relative to dynamically downscaled projections computed by pooling projections across the 4 GCMs for each region. (a) Strength of the signal of projected changes in the 10-year storm surge return level (RL10) from dynamical simulations (regional mean of absolute projected changes). (b) Mean absolute bias (%). (c) Pattern correlation coefficient (PCC). (d) Fraction of coastal points per region for which the sign of the RL10 change is reproduced (-), considering coastal points where the absolute amplitude is >=5%.

Overall, the hindcast-trained SDM shows a sufficiently satisfactory skill in reproducing GCM-specific responses of European-scale ESSs simulated by dynamical simulations. While several limitations apply which should be considered when interpreting associated projections — including notably reduced skill in some regions (e.g., the Baltic Sea and eastern Mediterranean Sea), potential inconsistencies where GCM atmospheric variability departs from ERA5, and a systematic underestimation of ESS change magnitudes—our results support the broader use of the hindcast-trained SDM for cost-efficient multi-model projections of European-scale ESS changes. The SDM enables projections for a substantially larger ensemble of GCMs than previously reported, hence allowing a more rigorous identification of main regional trends, and importantly, a more comprehensive evaluation of inter-model variability in ESS projections, which has been poorly constrained in studies to date.

==========

Note that the lower performance for the Baltic Sea and the eastern Mediterranean is now evident in the quantitative metrics, and it's explicitly highlighted in the text:

Lines 514-516: "*Given that dynamical simulations highlight these regions to display relatively strong future ESS changes (Figure 9-a), these results highlight the need to improve statistical projections in these regions for reliable future storm surge hazard assessments. The ensemble statistical projections presented hereafter should therefore be interpreted with caution in the Baltic and eastern Mediterranean regions.*"

We have also added in the introduction to the main results – the 17-GCM projections – a disclaimer that results for these regions are subject to lower confidence. Instead of masking results in these regions as proposed by the reviewer – as we have shown that the SDM skill is lower but not completely null – we have added in the maps in Figure 10 (formerly 7) a hatching (//) in these regions, highlighting lower confidence due to the SDM skill limitations:

Lines 540-542: "*We highlight that based on the validation results, statistical ensemble projections in the Baltic and eastern Mediterranean seas are subject to lower confidence given limited skill of the statistical model, illustrated in Figure 10 through hatching in these regions.*"

Below the modified Figure 10, which besides the hatching, it now focusing on the 10-year event. Nnote we have also adjusted the colorbar limits for the likely range, as they were saturated when kept equal to those on the MMM:

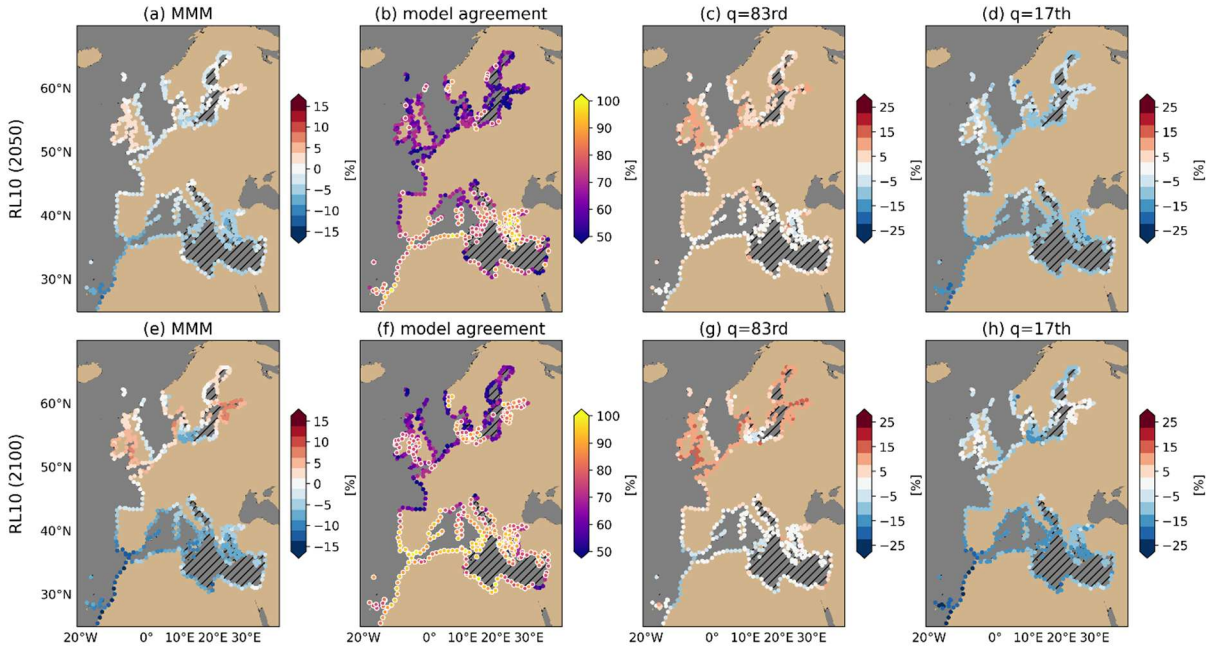


Figure 10 Muti-model mean (MMM) projected changes [%] in the 1 in 10 storm surge return level by middle (a) and end (e)-of the 21st century generated by the hindcast-trained statistical downscaling model (SDM, training period 1997-2021). Corresponding ratio of models agreeing in the sign of projected changes (b, f), 83rd (c, g) and17th (d, h) percentiles, indicating the likely range as per IPCC definitions. Hatching over the eastern Mediterranean Sea and the Baltic Sea indicate lower confidence in statistical projections in these regions given limited skill of the SDM for both past and future extreme storm surges (sections 3.1, 3.2.2). Extreme value analysis is computed for 30-year periods: baseline [1985-2014], middle of the century [2035-2064], and end of the century [2070-2099]. For reference, in a 17-model ensemble, the 17th and 83rd quantiles correspond to the lower/higher ~3 models. The % model agreement represents confidence in the sign of projected changes. Those with ratio >80% (here, >=13/17 models) are marked with white circle edges in panels (b,f).

The lower confidence in these regions is now explicit throughout the discussion of the ensemble results, including in comparisons to previous studies. We have now also added some physical interpretation of the overall pattern of projected ESS changes. Together, these elements provide

another dimension for increasing/reducing confidence in the results beyond the SDM skill evaluated from 4 GCMs:

Lines 568-582: *"The regions where robust changes have been identified broadly agree in sign with previous literature on dynamically downscaled projections of changes in the 10-year storm surge return level, despite the different GCMs being employed (Makris et al., 2023; Muis et al., 2022; Vousdoukas et al., 2016). Across studies, positive and negative ESS changes are concentrated in northern and southern Europe, respectively. However, the exact extents and magnitudes may differ substantially. For example, Muis et al. (2022) and Vousdoukas et al. (2016) identify positive future ESS changes across the Baltic Sea, while in our results substantial positive changes are limited to the eastern Baltic Sea (nothing that, in our case, the SDM is subject to lower confidence here). These studies also identify regions with substantial signals which are not emerging in our ensemble (e.g. the south-eastern North Sea in Vousdoukas et al., 2016), which may result from the use of smaller ensembles which underrepresent inter-model variance in storm-surge projections. In the North Sea, mismatches may be influenced by moderate skill of the SDM for future ESS changes (section 3.2.2). In contrast, the widespread reduction in future ESSs throughout the Mediterranean Sea identified in our ensemble is consistent across studies, even considering the lower confidence of our SDM in the eastern Mediterranean Sea. The latitudinal dipole in projected ESS changes in Europe may reflect a poleward shift of the mid-latitude jet stream and corresponding northward displacement of storm tracks (Köhler et al., 2025; Yu et al., 2023, 2024), though further analyses would be needed to attribute ESS changes to this phenomenon. "*

We have also better highlighted the lower performing regions in the discussion, and provided also potential future avenues for improving performance in the Baltic:

Lines 608-622: " *Regarding the statistical downscaling approach chosen, we have shown that multiple linear regression leads to a systematic underprediction of the target (predictand) extremes, despite achieving very satisfactory performance for normal conditions. Our results have shown that this negative bias has a limited impact on reproducing regional-scale projections of 10-year storm surge level changes for most of Europe, but for specific regions such as the Baltic Sea and the eastern Mediterranean Sea, the statistical model shows substantially lower skill, and hence our statistical projections are subject to lower confidence in these regions. Future works should explore ways of improving the SDM skill in these regions. Spectral analyses (not shown) indicate that the SDM still struggles to capture lower-frequency (>monthly) Baltic Sea variability, likely because much of it is driven by non-local processes (Weisse et al., 2021). Key remote influences include Baltic Sea volume changes driven by barotropic exchanges with the North Sea and low-pass-filtered storm surges entering through the Danish Straits (Andrée et al., 2023; Hieronymus et al., 2017). These could be better represented, respectively, by adding to the predictor set pressure-gradient indices spanning the North Sea–Baltic region (Karabil et al., 2018) or by including as regressor a low-pass-filtered surge proxy on the North Sea side. Beyond these regional challenges, and despite broadly agreeing regional patterns of changes in the 1 in 10 year storm surge level, differences between statistical and dynamical ESS changes can be substantial locally, and are expected to amplify for higher return periods given the associated decreasing capability of the presented statistical model."*

We have also highlighted regionally varying skill in the **conclusions**:

Lines 691-696: *"The model demonstrated stable skill across both historical and future climates, and showed overall satisfactory skill in reproducing the European-scale patterns of future changes in the 10-year return level given by dynamical simulations, although with a tendency for reduced amplitudes, and a notably lower skill in the Baltic Sea and the eastern Mediterranean regions. In these regions, statistical projections are therefore subject to lower confidence, highlighting the need to improve the statistical method for accurate assessments. In contrast, the model showed excellent performance along the European Atlantic façade and the western Mediterranean Sea, and moderate skill in the North Sea."*

Regarding results for the 100-year event, as previously highlighted, we have now focused our results on the 10-year event given the increased biases shown in the hindcast for higher return periods. This is also now more in line with the validation which also focuses on this return level. Ensemble projections for the 100 year are now shown in the Supplementary Materials, and briefly mentioned in the text, highlighting the associated lower confidence and the different (compounding) possible sources of the very large inter-model spread for the ensemble projection:

Lines 583-592*: "Projections of the 100-year return level (Fig. S9), which are subject to overall lower confidence due to reduced SDM skill, show generally amplified multi-model mean changes but substantially reduced inter-model agreement and larger spread than for the 10-year return level. This results in very wide likely ranges across Europe (see Figs. S10 and S11 for individual ensemble members), supporting the conclusion that the SDM should be used with caution for high return periods. In addition to degraded SDM performance, the reduced inter-model agreement likely reflects limitations in GCMs' ability to resolve the most severe extratropical storms (Priestley et al., 2020) and uncertainties in estimating the GPD shape parameter, which controls tail behaviour but is poorly constrained over 30-year periods. For this reason, several studies assume a time-invariant shape parameter when analysing changes in extremes (Cheynel, Pineau-Guillou, Lazure, Marcos, & Raillard, 2025; Lobeto & Menendez, 2024a; Marcos & Woodworth, 2017; Roustan et al., 2022). Further research is needed to assess how these different aspects affect the robustness of projections of ESSs at high return periods."*

We still argue that sampling uncertainty can play a role in the low inter-model agreement for RL100. This can be the case because inter-model agreement is evaluated based on the agreement on the *sign* of the projected *changes*, and slightly different values of the shape parameter (within its uncertainty band) might arbitrarily push projected changes to be positive/negative when these are small. We attempted to demonstrate this by evaluating RL100 projections when keeping the shape parameter constant (equal to that of the historical period). Fixing the shape parameter is often done precisely because (temporal changes in) the shape parameter are difficult to constrain and errors in the shape parameter fitting can manifest as temporal changes for large return period events. We showed that fixing the shape parameter leads to a much more robust RL100 change signal (that is, a much stronger agreement in the sign of the RL100 changes across GCMs), and less noisy when looking GCM by GCM (not shown). However, based on the other reviewer's comments, we have performed an Anderson-Darling test to evaluate the validity of different fits (free shape as originally done, fixed shape to the historical value, and also exponential distribution), and results showed the latter 2 to be rejected in widespread coastal regions, while the free shape worked well almost everywhere. Therefore we have now removed the fixed shape figure in Supplementary Materials and any discussion around such results.

Comment 3: *Another issue that I think requires some attention is the discussion about longterm variability in the atmospheric patterns and its impact on the performacen of the statistical model (lines 430-449). I do not think that the long-term modulation of low-frequency climate modes affects the results of the statistical model. Storm surges are caused by synoptic systems. These can be altered in frequency and magnitude by large-scale climate modes. However, the synoptic systems are still the same. In other words, changes in large-scale atmospheric conditions, like more blocking patterns, shifts of NAO, etc, will modulate the frequency and the intensity of the systems that generate the storm surges, but will not change the process and the type of system, nor the response of the storm surge. There are some statements in this paragraph in this line that I do not think are correct (lines 434-435, 437-438, 443-444). The only exception I can think of is the arrival of tropical-like cyclones in the future climates to the European coasts. However, these would not be well captured by the coarse resolution models anyway. I think this part needs to be reconsidered.*

Answer to comment 3: We thank the reviewer for her critical comment. The intention in this paragraph was to highlight that the principal components derived from the specific 25-year time slice used for the SDM training are not necessarily time-invariant. Despite the stationarity tests comparing specific future and past 20-year slices show rather stable results, we still wonder about whether internal variability can impact the EOFs and the SDM skill. But since our tests suggest otherwise and there's no specific literature demonstrating the contrary, we agree this discussion should be avoided. We have now removed the paragraph in lines 430-436 of the original manuscript (first block of lines you pointed out). For consistency, lines 443-444 (referring to future changes in the principal components) have been removed too. Lines 437-438 of the original manuscript don't talk about a potential temporal variability in the ERA5-based PCs, but rather on differences between the ERA5 vs GCM variability, which is illustrated by Fig S9 (now Fig. S3). We see that the variance retained after projecting the GCM fields onto the ERA5-based PCs is very high overall (close to the designed 99%) but noticeably lower in some regions as pointed out in the text, highlighting that GCM variability lies partly outside the ERA5 EOF basis on certain regions. These results require further analysis to understand the underlying causes, as now pointed out. We have now moved these lines to section 3.2.2 (Extrapolation to climate forcing, see whole section copied in previous answers):

Lines 504-511:*" As such, SD_hind estimates can only account for future storm surge changes linked to the identified ERA5 principal components, and not to novel atmospheric conditions or different modes of variability/covariance structures that may be present in GCMs. An analysis of the retained explained variance after projecting GCM fields onto hindcast-based principal components for the target 17-GCM ensemble (Fig S3, historical climate) reveals generally strong representativity across Europe, though with notable reductions for certain models and regions (e.g. the Mediterranean Sea for HadGEM3-GC31-MM). The retained variance also reveals stable between historical and end-of-century climates (not shown), supporting the stationarity of the predictands. Further analyses are needed to understand the observed differences between ERA5 and GCM variability and their impact on ESS change projections."*

Comment 4: *On a personal note, I find the reading more difficult with the use of so many acronyms. My preference would be to avoid the use of at least some of them. For example: SS as storm surges, or even SDM and DDM could be referred to as, simply, statistical model and dynamical model.*

Answer to comment 4: We thank the reviewer for her suggestion. We have removed the following acronyms throughout the manuscript: REVISE!!

-SD (we only kept it to refer to the specific experiments in previous section 3.1 'Validation of statistical projections', now section 3.2)

-SS

-DDM

-MLR

-MMM

-RL10

-RL100

For SDM, we have chosen to keep it, to refer to our specific statistical model setup (based on PCA + MLR, with the given predictor options, with coarsened predictors, etc), in a similar fashion to other papers where one defines the name of the model (e.g. GTSM instead of simply numerical model). When referring to statistical models in general, we mention explicitly 'statistical models'.

Other minor comments:

- Line 92: I am unsure what this means. Perhaps that one in every N(?) coastal points are analysed? If so, what is the averaged distance among coastal points? Please, clarify.

-Changed to '*For computational speedup, we analyze storm-surge outputs at one in every 10 coastal points (every 50-100 km, ~600 coastal points in total).*'

- Figure 1 has a wrong caption. Reference to Fig 1c in line 105 is unclear.

-Thank you for spotting that. We have corrected the label of panel (c) and the caption accordingly: '*Performance of the dynamically downscaled storm-surge hindcast against non-tidal residuals from GESLA3 tide-gauge observations (Haigh et al., 2023) for 1997– 2015. (a) Pearson correlation coefficient (R). (b) Root mean square error (RMSE). (c) 99th percentile error. GESLA3 storm surge has been extracted after yearly tidal analysis (considering a minimum of 80% coverage for each year) and is computed relative to the annual mean sea level (detrended). Stations with at least 4 years at the assessed period are retained, and statistics are evaluated only at valid observation timestamps. EMED: Eastern Mediterranean Sea; WMED: Western Mediterranean Sea; BAL: Baltic Sea; NSS: North Sea; ATLS: Atlantic Shelf*'

-Lines 99-104: please, provide references here. This pattern is shown multiple times in the literature.

-Added multiple references: Lines 139-146 now "*Correlations are lower than average for the southern part of the domain, which probably reflects the contribution of baroclinic processes to the non-tidal residual in tide-gauges (García et al., 2006; Mohamed & Skliris, 2022), which is not captured in the 2D barotropic model. RMSEs are higher than average in the North Sea (15cm), which is expected given the larger storm surge amplitudes in the region (Calafat & Marcos, 2020; Pineau-Guillou et al., 2023). The correlation and RMSE spatial patterns and values reported for the dynamical model are comparable to those reported for other European barotropic*

hydrodynamic models (Agulles et al., 2024; Cheynel, Pineau-Guillou, Lazure, Marcos, Lyard, et al., 2025; Fernández-Montblanc et al., 2020) . "

-Table S1: homogenise units.

-Done, thank you for spotting it.

- Figure S2: Please, increase the size of the figure and the font size. It is not readable.

- We have increased the size of each subpanel to the maximum while keeping the figure page-size (including caption) by putting the colorbars on the bottom. Regarding the statistics shown, these are by default plotted in our plotting routine for us to check the values, but they were not intended to be included in the publication version plots (as for the other plots, where they have been removed). We have now removed them in S2 as well. We have also adjusted the colorbar for the relative differences against the hindcast (%) to show discrete values, as for the other SM plots, to better identify values visually in the maps.

- Line 150: SS (I guess storm surge) has not been defined. Please, limit the use of acronyms.

-This acronym is now not used .

- Line 156: I do not see the reason to include both the gradient of SLP and the winds. At 1deg resolution, they are likely the same fields, and this would be overfitting the model. Please, discuss.

- Based on Pyykkö and Svensson (2023) (https://doi.org/10.1175/JCLI-D-22-0705.1), CMIP6 surface winds are shown to substantially deviate from geostrophic balance even at their typical >1 degree resolutions. We have now included this reference to justify the addition of winds. Additionally, our calibration results show clearly improved performance when including winds as predictors on top of the SLP-based predictors (highlighted in lines 196-197 of the original manuscript), and since the performance metrics are derived on data unseen by the model (k-fold cross validation) this improvement cannot be linked to overfitting due to redundancy of the newly added wind fields. On the contrary, they highlight the added value of wind fields as additional predictors. We have now discussed these aspects explicitly:

Lines 189-192: "*we then add the daily maximum squared atmospheric pressure gradient-SLPG- as a proxy for geostrophic winds (Rueda et al., 2016) (T2); finally, as surface winds may substantially deviate from geostrophic balance even at 1° resolutions (Pyykkö & Svensson, 2023), we also account for the influence of zonal and meridional near-surface winds (U10, V10),"*

Lines 234-236:" *Predictor-wise, the addition of wind variables (T > 2) markedly improves correlation and RMSE, demonstrating their added value relative to the purely geostrophic information contained in SLP gradients*"

- Figure 4: units are missing in the legends.

-Units are given in the plot y axis ([m]). We have now added them also in the caption for clarity ***"****Comparison between target coastal storm-surge (meters)…****"***

- Lines 253-254: by errors, do you mean the uncertainties in the maximum likelihood adjustment of GEV? You also show 100-year return levels in the projections.

- This comment is now removed, the justification of focusing on the 10-year event (for both validation in this section, and later ensemble projections) is based now on the lower SDM skill for higher return periods.

- Line 271: the underprediction of high storm surges is larger when trained with the hindcast. This can be due to the hindcast having smaller storm surges than the historical simulations. It would be worth checking if this is the case. That would mean that the model extrapolation is biased low. In addition, the climate models have been bias-corrected, adjusting means and variances to those in ERA5. It would also be good to check how the extremes are affected by this bias correction (probably less than the mean storm surges and this would explain these differences).

-Indeed, the hindcast simulation has lower extremes than the historical GCM simulations, as shown in Figure S1 (Fig S1-c,f,I,l, showing positive GCM RL10 errors relative to the hindcast simulation ,Fig S1-a). This was raised in the discussion in Lines 301-305 of the original manuscript (stating the GCMs *overpredict* the historical RL10). We note than in the highlighted line 271, the underprediction is relative to the dynamical simulation benchmark – that is, each GCM dynamical simulation – not the reference hindcast simulation. This *underprediction* points out that when training on the hindcast, the high storm surges are further reduced relative to the dynamic simulation than for the GCM-trained SDM. Indeed, independently of the SDM (composed by the EOFs and regression coefficients), the bias correction applied on the hindcast-trained statistical reconstructions might be playing a strong role here. When compared to the reference hindcast simulation (Figure S1), we actually see that statistical reconstructions better match the dynamical hindcast than dynamical historical simulations (smaller RL10 errors), so while we are getting further away from the dynamical simulation results, we are getting closer to the hindcast. Unfortunately, we don't investigate the relative importance of the different aspects of the extrapolation (bias correction vs filtering to ERA5 principal components vs inherent negative bias for extremes), and how extremes might be affected differently than average conditions by the correction. Nevertheless, it is not the intention in this section to judge the *value* of the metrics, but to assess their *stationarity* in time (future vs historical). This is also why performance is further based on the reproduction of the climate change signal (relative change future – historical), and not the historical performance for example, as we can expect that the bias correction performed for *SD_hind* and not present in the other estimates will bring the *SD_hind* estimates closer to the hindcast (hence 'better' in principle), but doesn't reflect a better skill of the hindcast-trained statistical model over the others. We have highlighted this aspect now in the description of the stationarity test results:

Lines 443-448: "*Particularly, the underprediction of high storm surges (>99th percentile, Figure 7-i-l) is systematically more pronounced for SD_hind. This could be the result of the bias correction if predictors in GCMs were systematically biased relative to ERA5 —for instance due to overestimated average wind speeds— producing larger storm surges for GCMs. This is suggested by results in Fig S1, which show systematically overpredicted ESSs in historical dynamical simulations across GCMs relative to the hindcast simulation, while hindcast-trained SDM estimates show much reduced errors*".

While reflecting on this issue, we have noticed that it was not clear in the text that the bias correction of GCM fields only applied to the projections based on the hindcast-trained SDM (*SD_hind*) (as, again, it was implemented because it was required for consistent projections of CMIP6 fields onto ERA5 EOFs, not as a target methodological aspect for projections in general). We have now explicitly explained this at the beginning of the results section, explaining that it

does however not affect conclusions from our validation tests (stationarity, reproduction of projected changes):

Lines 535-537: "*We emphasize that the bias correction needed for SD_hind estimates but absent in the other sets (DD, SD) does not impact the validation of the SDM for climate projections, as it has no impact on the two target validation tests (stationarity and reproduction of projected changes).*"

- Lines 279-280: Is this delta method necessary when the climate models are bias-corrected? I guess no for the mean characteristics of the storm surges, but extremes could still behave differently (also relates to my previous point above).

- Indeed, we are correcting for mean amplitude and variance, not explicitly for the extremes. Note that for the dynamical simulations, and statistical projections trained on each GCM (*SD* estimates), a bias correction has not been performed on the forcings. The correction was required to safely project the GCM fields onto the ERA5-based EOFs (*SD_hind* estimates). This is why we validate projected *changes* in this section. To clarify this aspect, we have added the following lines:

Lines 361-366: "*As previously highlighted, biases have been corrected for SD_hind (through simple mean bias and variance corrections, for safe projection onto ERA5-based EOFs), but they haven't been corrected for SD and DD estimates. Additionally, even when corrected, biases in extremes haven't been explicitly addressed, which may differ from those related to mean conditions. Together, these factors justify our focus on evaluating projected changes of ESSs (future vs. past) instead of directly assessing their future projections in the DD/SD/SD_hind intercomparison.*"

- Figure 6: some scales seem saturated. If this is the case, it should be explained in the text (line 289 states that changes are +/-20%).

- Yes, some scales are saturated. The colorbar range was chosen based on the range of 90% of the data across GCMs (5-95th percentiles). But some points show larger changes. We have made this explicit now in the text. Lines 458-460: "*Dynamical projections reveal considerable inter-model spread, with regional changes typically reaching ±20% (Figure -a,e,i,m, 5th-95th percentiles of results pooled across GCMs), exceptionally higher (-25%/+32%, 1st/99th percentiles respectively).*"

- Line 302: what does overprediction mean here?

- Relative to the hindcast, as highlighted at the beginning of the phrase. However, we have now included a discussion on historical performance on section 3.2.1 ('stationarity assumption',previously section 3.1.1) following a previous comment of the reviewer, to better explain possible nuances between performance metrics for different statistical estimates, so we have decided to remove it here (lines 302-305 of original manuscript removed)

- Lines 326-326: In most of the Mediterranean Sea the statistical approach does not provide reliable results (see my second major comment above), which means that even if the models are consistent, the result is not robust.

-As explained in comment 2, we have now better described the skill of the SDM in reproducing projected changes (for the 4 GCMs at hand) in section 3.2.2 (previously 3.1.2), highlighting that skill in the eastern Mediterranean and the Baltic is lower and hence projections are of lower

confidence for these regions (i.e. less reliable). Regarding terminology, we limit the term *robustness* to inter-model consistency (in this case in terms of the sign of projected changes, IPCC 'simple' method), while reliability is referred to as *confidence*. As highlighted in previous answers (**comment 2**), we have now added a hatching in the Figure 10(previously Figure 7) depicting the lower confidence in these regions, with corresponding explanation on the figure caption, and related comments in the description of results and conclusions.

-Lines 355-361: The fitting of a GEV using maximum likelihood comes with its uncertainties, that are related to the sample size and its empirical statistical distribution. The increase of uncertainties in the return levels for low-probability events is inherent to the approach, so it cannot be blamed for the decrease in the confidence of the results. The high uncertainties come from the use of a relatively short record (20 years, i.e. 20 maxima). Even with a high goodness of fit of the shape parameter, the uncertainties would increase. Therefore, please, reconsider this text.

Thank you for your remark. The aspects raised here have been elaborated on answer to comment 2. The highlighted lines are no longer in the text, and the RL100 results are not part of the main body. All in all, we will reiterate here why we think sampling uncertainty can play a role in the inter-model agreement as per our definitions.

We would first like to emphasize that we are not using GEV with annual maxima, but GPD + POT with a threshold leading to an average of 5 events per year (see section 2.4). Secondly, for multi-model projections in section 3.2, we derive extremes for periods of 30 years, not 20 (2070-2099 vs 1985-2014). These aspects are specified in Figure7's (now 10) caption. What we argue is that even for 30-year time series, sampling uncertainty for high return periods such as RL100 – affecting our confidence on the RL100 *change* quantification - could partly explain the reduced inter-model agreement (=robustness of changes per our definition). This is because robustness in our projections is determined by the sign of the projected *changes*, and given that RL100 changes are relatively small (order of centimeters – decimeters), slight variations on the shape parameter (given for example by a slightly different magnitude of the most extreme events in the data) in the future 30 years can push changes to be either positive or negative relative to the baseline, impacting the agreement between models. In other words, the sensitivity of projected RL100 changes to the error in the estimation of the shape (because of too few samples) can be in the same order of magnitude as the 'mean' changes given by point estimates.

However, we agree that we cannot strictly associate the reduced inter-model agreement (robustness) for RL100 to the effects of sampling uncertainty, as we cannot differentiate this aspect from others that might be affecting the skill to determine RL100 changes (e.g. SDM skill for high return periods, GCM skill in resolving these events, etc), or even from real shape parameter changes driven by climate change when these are small. Therefore, we now simply list the possible sources of the lower inter-model agreement for the RL100 (see reply to comment 2 for the associated paragraph).

Answers to Reviewer 2 (Anonymous)

COMMENTARY

I congratulate the authors for the substantial effort invested in this study. The manuscript addresses an important and timely topic and provides a valuable contribution by exploring future projections of daily maximum storm surges along European coasts using statistical downscaling applied to a large CMIP6 ensemble. The numerical and statistical workload behind this study is impressive, and the authors make a huge effort to assess the validity of their statistical downscaling approach under climate change conditions. The pan-European scope and ensemble-based perspective are clear strengths.

That said, I believe the manuscript would benefit significantly from improvements in structure, clarity, and methodological rigor, particularly regarding hypothesis formulation, inference, and the separation between methods and results. In its current form, methods and results are often interwoven, making the paper difficult to follow. In addition, while many validation steps are presented, the confidence in the final projections remains limited, partly due to the lack of formal hypothesis testing and inference in several key analyses.

My comments below are intended to be constructive. Some are necessarily subjective or based on my interpretation; please feel free to disregard them where they are not useful or where I may have misunderstood aspects of the work.

Answer to general commentary:

We sincerely thank the reviewer for their thorough, careful, and constructive review, as well as for their positive assessment of the scope, ambition, and relevance of this work. We particularly appreciate the recognition of the substantial numerical and statistical effort involved, the pan-European perspective, and the ensemble-based approach adopted in this study.

Regarding the structure of the manuscript, the original organization—where some "intermediate" results appeared within the Methods section—was intentional. Our aim was for the main Results section to focus clearly on the projections themselves, rather than on intermediate analytical steps. We believed this improved readability and maintained a coherent narrative flow. Presenting all validation and model-selection analyses as standalone results would have required a very extensive Methods section and a correspondingly large Results section, forcing readers to move repeatedly between the two to interpret the results, as these included definitions and experimental protocols specific to our study. We hence adopted a more integrated, storyline-oriented approach, introducing methods alongside the corresponding analyses to provide context as the reader advances.

The intermediate analyses in question (e.g., training data assessment, model configuration selection, and reconstruction of past extremes) were not intended as standalone findings, but as necessary steps to establish the credibility and fitness-for-purpose of the statistical downscaling model prior to generating the main results—the projected changes in extreme storm surges.

That said, we acknowledge the reviewer's concern that the interweaving of methods and results reduced clarity. Following this advice, **we have reorganized the manuscript substantially** to

collect all analyses of the SDM outputs—both for historical reconstructions and future projections—in the Results section. The Methods section still contains the intermediate results considered part of the methodology, namely (i) validation of the numerical hindcast –here, the *training dataset* - which is intended as a quality control step of the data used for training (e.g. are storm surge (extremes) properly represented in the data) rather than a full, thorough model evaluation (as the latter is not an objective of this paper), and (ii) the SDM calibration, predictor selection, and cross-validation procedure, which describe how the final model configuration was chosen. **We have also moved all methodological explanations** previously presented in the Results section **to the Methods section.** Notably, we have **created a new section in the Methods to clarify procedures followed for validation of the SDM under climate projections and its application for the final 17-model ensemble (section 2.5, 'Experimental design')**, with the experiments performed and specifications for their analysis (Table 3). For clarity, we add here the new structure

1. Introduction
2. Methods
   - 2.1. General workflow (new)
   - 2.2. Training and benchmark datasets (corresponding to previous section 'Dynamical downscaling model', but with more explicit explanation of the role of such model in this study – a dataset)
   - 2.3. Statistical downscaling model (largely unchanged, but hindcast reconstruction plots/metrics not in 3.1)
   - 2.4. Experimental design (new)
     - 2.4.1. Validation under climate forcing
     - 2.4.2. Multi-model ensemble projections
3. Results
   - 3.1. Statistical hindcast reconstructions (new, with focus on extremes now)
   - 3.2. Validation of statistical projections (largely unchanged in structure)
     - 3.2.1. Stationarity assumption
     - 3.2.2. Extrapolation to climate forcing (now including quantitative skill metrics)
   - 3.3. Statistical ensemble projections (largely unchanged but focusing on the 10-year event, and highlighting regions of low confidence)

More broadly, we would like to clarify that the objective of this paper is not to introduce a novel statistical downscaling methodology, but to adopt a PCA–multilinear regression–based SDM that has been used in several studies for the reconstruction of storm surges (as highlighted in the Introduction, e.g., Tadese et al., 2020), tailor it to a pan-European coastal domain, and evaluate its suitability for projecting future changes in extreme storm surges under climate change. The choice of predictors, dimensionality reduction, and SDM configuration follows approaches that have been extensively evaluated and applied in the previous studies, and appropriate references have been added to clarify this point. The methodological emphasis is therefore on careful implementation and validation rather than methodological innovation. **We have now clarified the objective in the Introduction section**, lines 79-84 , and recalled it throughout the paper.

Lines 79-84: *"In this study, we address these gaps by producing the first expanded multi-model ensemble (17 models) of pan-European extreme storm surge projections using a dynamical-statistical downscaling approach. Rather than developing a new statistical approach, we adopt an existing method used for broad-scale storm-surge reconstructions — multiple linear*

regression —, tailor it to Europe, and assess its capability for projecting ESS changes, which has not yet been demonstrated. The adopted framework enables this evaluation by using dynamically downscaled projections as a benchmark, which is not possible for observations-based statistical downscaling."

Finally, we appreciate the reviewer's broader comments regarding inference and confidence in the projections, and the suggestion for formal hypothesis testing. Following standard practice in the storm-surge validations, we have formulated the validation in terms of diagnostic skill metrics that quantify agreement between statistical and dynamical projections (our benchmark here), rather than formal hypothesis testing. In this regard, **we have added skill metrics to the comparison between statistical and dynamical estimates of future ESS changes** (section 3.2.2, formerly 3.1.2) to better justify claims regarding the SDM skill. **We have reinforced and clarified the interpretation of the validation results,** and strengthened the discussion of uncertainties and limitations **to better contextualize the robustness of the projected changes** (see answers to specific points below).

We believe these revisions significantly improve the clarity and structure of the manuscript while preserving its original objectives, and we thank the reviewer again for their insightful suggestions, which have helped strengthen the paper.

Detailed commentary

Section 1 Introduction

***Line 29**: The reference provided supports the statement on hazards well. I recommend adding a second reference explicitly supporting the role of extreme storm surges in flood risk across Europe.*

***Line 32**: The statement "Even without changes in storm characteristics, rising mean sea levels…" should be nuanced. Its validity depends on coastal context, such as the availability of accommodation space, sediment supply, and the presence of coastal squeeze*

Thank you for the suggestion. We now explicitly refer to the IPCC projections referenced at the end of the phrase (which consider the effect of long-term SLR only), and noting that "are expect to" instead of "will". Line 33: "*Even without changes in storm characteristics, rising mean sea levels alone are expected to dramatically increase the frequency of today's high-impact events (e.g., Fox-Kemper et al., 2021).*"

***Lines 31–35**: I suggest softening the claims and acknowledging contrasting findings in the literature. For example: Sterl et al. (2009) find no statistically significant changes in the 10,000-year return value of surge heights along the Dutch coast during the 21st century and show that higher mean sea level does not necessarily a[ect surge height. Land and Mikolajewicz (2019) show that extreme sea levels in the German Bight are dominated by strong internal variability and multidecadal fluctuations rather than clear climate-change signals. Sterl et al. (2009) https://doi.org/10.5194/os-5-369-2009 Land and Mikolajewicz (2019) https://doi.org/10.5194/os-15-651-2019*

Thank you for the suggested relevant literature, we have modified the phase accordingly, including a reference to Calafat et al. (2022) (https://doi.org/10.1038/s41586-022-04426-5). Lines 35-39 in reviewed manuscript: "*In addition, stormy conditions over the ocean and the*

induced storm surge behaviour may also change under a warmer climate and may further contribute to changes in future coastal hazards, although current evidence for Europe remains inconclusive and strongly region-dependent. While Calafat et al. (2022) report trends in European storm surge extremes matching the rate of sea level rise, other studies emphasize the lack of significant future storm surge changes and dominance of internal variability over forced trends (e.g., Lang & Mikolajewicz, 2019; Sterl et al., 2009)."

***Line 40**: The statement on low confidence in surge projection ensembles aligns with findings for GCM and RCM wind projections, which often show larger inter-model variability than model-mean changes (e.g. wind energy studies https://doi.org/10.5194/wes-7-2373-2022).*

Thanks for the reference, we have added this reference in Results section for the ensemble projections to support the large inter-model spread found for ESS changes in the North Sea:

Lines 566-567: *In the North Sea, the large inter-model spread in ESS changes is in line with studies reporting similar findings for CMIP6-based wind projections (Hahmann et al., 2022).*

***Lines 36–41**: In my experience, robust regional projections are also limited by hydrodynamic model uncertainty related to GCM resolution, particularly due to the sensitivity of surge models to forcing resolution in semi-enclosed basins such as the North Sea and Mediterranean Sea.*

Thank you for your comment. We agree that this is the case, however here we want to specifically highlight the reduced confidence in projections due to the poor inter-model uncertainty characterization in storm surge projections due to limited ensemble sizes, stemming from computationally expensive dynamical downscaling. We have therefore rephrased to streamline towards this message (now lines 40-45):

Lines 40-45:"*Despite the relevance of storm surges for European extreme sea levels, regional projections of future changes in storm surges remain limited, including for Europe. Most existing studies rely on hydrodynamic simulations to dynamically downscale climate model outputs. While physically detailed, these methods are computationally expensive, restricting ensemble sizes to a small number of global climate models (GCMs) (e.g., Chaigneau et al. 2024; Muis et al. 2020, 2022; Vousdoukas et al. 2016). As a result, inter-model uncertainty is poorly characterized, reducing the confidence in projected ESS changes, and inter-model spread across small ensembles is reported to be high, emphasizing the need for larger ensembles.*"

The impact of GCM resolution was already briefly discussed in the Discussion (lines 450-463 of submitted manuscript). Furthermore, uncertainty stemming from poor representation of extreme storm events in GCMs has been now highlighted as one of the uncertainties affecting the projection of changes for high storm surge return periods (lines 586-588):

"In addition to degraded SDM performance, the reduced inter-model agreement likely reflects limitations in GCMs' ability to resolve the most severe extratropical storms (Priestley et al., 2020) and uncertainties in estimating the GPD shape parameter, which controls tail behaviour but is poorly constrained over 30-year periods."

***Paragraph starting line 42:** The discussion of statistical downscaling under climate change conditions could be strengthened by referencing earlier work demonstrating its applicability and limitations. https://doi.org/10.1175/JCLI-D-11-00687.1 https://doi.org/10.5194/gmd-13-2109-2020*

We thank the reviewer for this helpful suggestion and for pointing us to an interesting and relevant study. We have carefully considered whether to include this reference, but we have decided not to add it for two main reasons. First, the improved capability of neural-network-based methods in representing extremes is already discussed in the Introduction, with references that specifically address storm surges (our objective**)**, which are the focus of this study.

Second, the extrapolability assessed in the proposed study is not directly comparable to the extrapolation problem addressed here. That work evaluates extrapolation within reanalysis-based frameworks and relatively modest climate shifts, whereas our study explicitly focuses on extrapolation to GCM-driven climate forcing and substantially warmer end-of-century conditions. Notably, extrapolation issues may arise between reanalysis and GCM forcing already in the overlapping historical period, with minimal effect of climate change. As such, while the study is valuable for contextualizing statistical downscaling in general, it does not directly address the specific extrapolation challenge examined in this manuscript.

We nonetheless appreciate the suggestion and agree that it provides useful broader context for the ongoing discussion on statistical downscaling under climate change

Line 51*: The phrase "less continuous framework" is unclear and should be clarified.*

We made it more explicit:

Lines 52-55: "More complex methods based on Weather Types cluster synoptic atmospheric conditions into circulation regimes and link them probabilistically to local surge responses (Anderson et al., 2019; Costa et al., 2020; Zhong et al., 2025), offering a physically interpretable but regime-based (discrete) representation of surge–atmosphere relationships."

Line 60*: Please clarify whether you mean that SDMs can be trained directly using reanalysis data.*

Clarified:

Lines 66-70: "Alternatively, the use of outputs from physically based numerical simulations (e.g., reanalysis products) as the predictand enables the training of statistical models on spatially and temporally continuous storm surge fields, effectively replicating the behaviour of dynamical models at a fraction of the computational cost"

Lines 60–66*: The discussion of hybrid downscaling is confusing. The manuscript applies a state-of-the-art statistical downscaling method. If hybrid downscaling is mentioned, it would help to clearly distinguish between purely data-driven statistical downscaling (this work), and hybrid approaches combining targeted numerical simulations with interpolation techniques.*

We have now modified the manuscript to remove instances of *hybrid* downscaling to avoid confusions and rather referring to *dynamical-statistical downscaling* when needed. E.g. lines 66-70:

"Alternatively, the use of outputs from physically based numerical simulations (e.g., reanalysis products) as the predictand enables the training of statistical models on spatially and temporally continuous storm surge fields, effectively replicating the behaviour of dynamical models at a fraction of the computational cost. Such dynamical-statistical approaches have been successfully applied to reconstruct historical storm surge fields at regional (Tausía et al., 2023) and global scales (Cid et al., 2017)."

Section 2 Methods

***About astronomical tides**, are astronomical tides explicitly modelled or removed? This should be clarified early.*

They are not modelled. This is now clarified in lines 123-124:.

"Astronomic tides are not included in the configuration, and hence non-linear tide-surge interactions (e.g., Jenkins et al., 2025) are not resolved"

***Workflow schematic**: A schematic of the full workflow (numerical model calibration/evaluation, SDM calibration/evaluation, application, and analysis) would greatly improve clarity. Figure 2 appears to serve this purpose and should be introduced earlier. Figure 1 currently presents results and could be moved later*

Thank you for your suggestion. We have added the following figure, under a new first section in the Methods called 'General workflow', where this workflow is described (high-level) and further detailed in subsequent sections in 'Methods':

Lines 94-106: "The general workflow of the study – including the training, validation and application of the envisioned statistical downscaling model for climate projections of future ESS changes – is presented in Figure 1. First, a dynamical downscaling model is used to generate both the dataset of past storm-surges (a hindcast forced by the ERA5 reanalysis, Hersbach et al., 2020) to train the statistical model - also used to evaluate the skill of the statistical model to reproduce past ESSs - and the benchmark datasets for future changes in storm surge extremes (CMIP6-forced simulations for historical and future climates, 4 GCMs) to validate the statistical projections. The validation for statistical projections focusses on evaluating two aspects: whether the statistical model trained in past conditions remains valid in future climates (stationarity) and the extrapolation capability of the hindcast-trained statistical model to climate forcing (extrapolability), with a focus on reproducing projected ESS changes (our ultimate goal). To our knowledge, this constitutes the first study to explicitly assess these two aspects, addressing a key limitation in previous approaches that apply hindcast-trained models without prior validation for storm surge projections (Cagigal et al., 2020; Dubois et al., 2025). Finally, the hindcast-trained statistical model is used to produce an ensemble of projections of ESS changes based on 17 GCMs from CMIP6. The different components and steps involved in the workflow are further elaborated in the following sections."

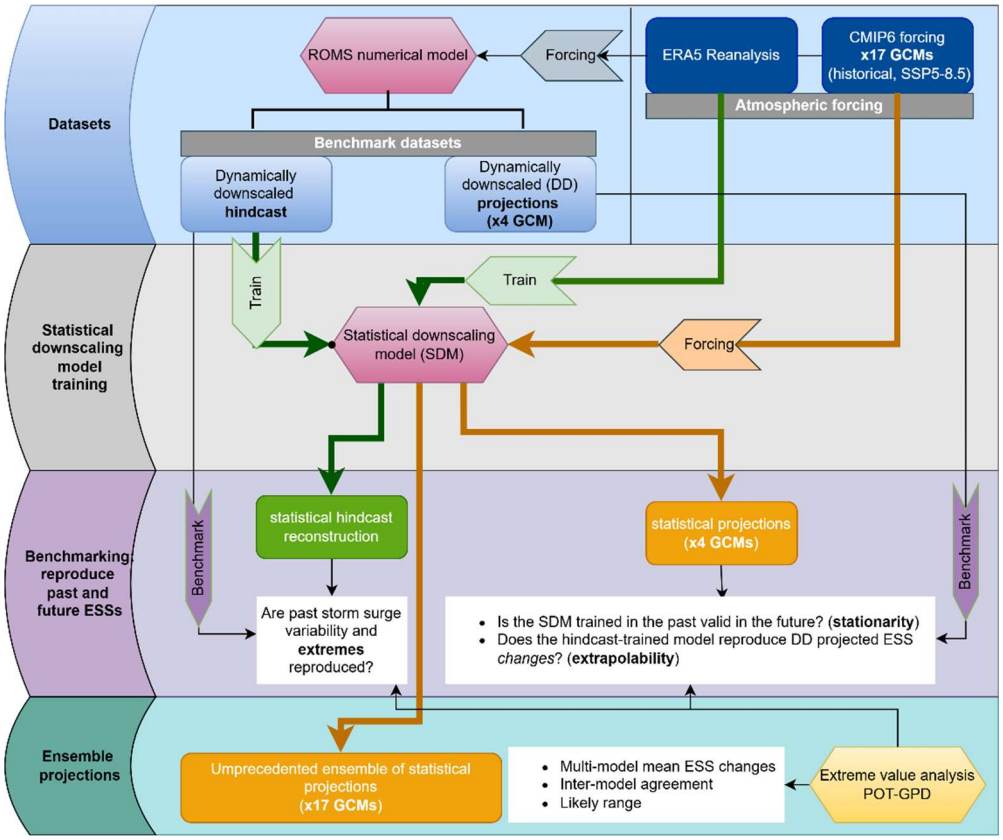


Figure 1 General workflow of the study, including the atmospheric forcings and benchmark datasets, training of the statistical downscaling model (SDM), benchmarking of the statistical model for projections of extreme storm surge (ESS) changes, and application of the model for ensemble ESS change projections based on 17 Global Climate Models (GCMs).

Original Figure 2 (now Figure 4) concerns the workflow for the statistical projections of daily maxima storm surges, not of the full study workflow which also includes the benchmarking against numerical projections and the analysis of extremes. This figure is now moved to section 2.5 (see answer to general commentary).

As clarified in the answer to the main commentary, the validation of the numerical hindcast is retained in the Methods section, as it is intended as a general validation of the training dataset (see next point)

***Line 108**: I disagree with the statement as written. The authors should demonstrate that the hydrodynamic model robustly translates wind stress and inverse barometer effects into storm surge estimates. Otherwise, the paper should be framed more explicitly as a methodological contribution. The hydrodynamic model is a key component of the hybrid SD framework. I recommend: introducing the hydrodynamic methodology in the Methods section, and moving its evaluation against observations to the Results section. Authors are encouraged to compare hindcast surge results with other available storm surge reanalysis products.*

We thank the reviewer for this comment and agree that the statement as originally written was inaccurate and required clarification. In this study, the hydrodynamic storm-surge hindcast is not presented as a fully validated or newly calibrated numerical model, but rather as a benchmark dataset used to train and evaluate the statistical downscaling model. Its ability to reasonably represent storm-surge variability and extremes is therefore essential, as the SDM is designed to reproduce both the general variability and the extreme behavior of the training data.

However, a comprehensive calibration and validation of the hydrodynamic model—including detailed sensitivity analyses (e.g. tuning of surface drag or Charnock coefficients) or event-based validation of specific historical storms, as performed in studies primarily focused on model development and skill presentation —is beyond the scope of this paper. Instead, we assess whether the hindcast reproduces the main contemporary storm-surge statistics and extremes at the continental scale, and whether its performance is consistent with other state-of-the-art European storm-surge datasets, to validate the use of its outputs for training our SDM (and also to use as benchmark for projections). We have added a phrase at the beginning of section 2.2 to clarify the role of the hydrodynamic model in this study:

Lines 113-117:"*In this study, a hydrodynamic storm-surge model is used to generate the benchmark dataset for training the statistical model – the hindcast - and to validate the statistical downscaling model under climate projections. Accordingly, the hydrodynamic model description and validation are intentionally limited in scope, focusing on documenting the hydrodynamic framework and assessing its ability to provide a suitable benchmark representation of storm-surge variability and extremes, rather than at presenting or exhaustively calibrating the numerical storm-surge model.*"

We have expanded the description of the hydrodynamic modelling framework (e.g., wind stress parameterization used, lack of astronomic tides, bottom stress) in the Methods section to provide greater methodological clarity. The evaluation of the numerical hindcast remains presented as a quality assessment of the training data, rather than as a standalone model calibration and validation exercise and is therefore retained in the Methods section. We believe this framing more accurately reflects the role of the hydrodynamic model within the overall statistical downscaling framework and scope of the study.

Finally, we have expanded the description of the hydrodynamic model skill figures, including comparing with state-of-the-art similar models, and we have modified the highlighted Line 108 to convey the right message (corresponding to the last line in the paragraph below):

Lines 137-153*: "The dynamically downscaled hindcast demonstrates satisfactory agreement with the storm surges observed in the GESLA3 tide-gauges (Figure 2; see caption for processing details), yielding mean correlation (Figure 2-a). and RMSE (Figure 2-b) values of 0.76 and 10cm, respectively. Correlations are lower than average for the southern part of the domain, which probably reflects the contribution of baroclinic processes to the non-tidal residual in tide-gauges (García et al., 2006; Mohamed & Skliris, 2022), which is not captured in the 2D barotropic model. RMSEs are higher than average in the North Sea (15cm), which is expected given the larger storm surge amplitudes in the region (Calafat & Marcos, 2020; Pineau-Guillou et al., 2023). The correlation and RMSE spatial patterns and values reported for the dynamical model are comparable to those reported for other European barotropic hydrodynamic models (Agulles et al., 2024; Cheynel, Pineau-Guillou, Lazure, Marcos, Lyard, et al., 2025; Fernández-Montblanc et al., 2020). The hindcast simulation shows a general tendency for underpredicting high storm surges (mean -3 cm), again most pronounced around the North Sea (-10cm, Figure 2-c). The reported underestimation of high storm surges is also a well-documented limitation in similar hydrodynamic model simulations (Chaigneau et al., 2024; Cheynel, Pineau-Guillou, Lazure, Marcos, Lyard, et al., 2025; Fernández-Montblanc et al., 2020), which is likely attributable to inaccuracies in the representation of storm events in the atmospheric forcing data (Irazoqui et al. 2022). Based on these findings, we conclude that the simulated hindcast shows a sufficiently high skill in reproducing storm-surge variability and extremes to serve as training for the*

statistical downscaling model, though the systematic underprediction relative to observations should be considered when interpreting projections of both dynamical and statistical models."

Data description: *A clearer and more complete description of all datasets used is missing.*

In our case, datasets include the atmospheric forcings (ERA5, CMIP6 models) and the training/benchmark datasets produced with the numerical model. Table 1 describes training and benchmark datasets (simulation periods, forcing dataset and frequency). ERA5 is not strictly described but properly referenced. The CMIP6 models are listed in the Supplementary Materials (Table S1), highlighting the horizontal resolutions and now referenced in the section 2.5 'Experimental design', subsection 2.5.2. We have added the temporal resolution of winds and atmospheric pressure of the 17 GCMs in the Table S1 caption '*Zonal and meridional winds at 10-meters are available at 3-hourly frequency, while atmospheric pressure is available at 6-hourly frequency, except for the MPI models which provide 3-hourly frequency*'.

We believe the datasets are thus sufficiently described for the purpose of the study.

Line 92: *"We thin coastal points by a factor of 10" is unclear. Does this refer to reduced output resolution relative to Cid et al. (2014), or to subsampling of coastal grid points?*

It refers to subsampling. We have clarified the phrase, lines 126-127 "*For computational speedup, we analyze storm-surge outputs at one in every 10 coastal points (every 50-100 km, ~600 coastal points in total).*"

Table 1*: In my experience, bias correction is required to reconcile differences between hydrodynamic models forced at 3-hourly versus 1-hourly resolution.*

We agree that differences in the temporal resolution of atmospheric forcing can affect storm-surge estimates and, in some contexts, may require bias correction. We acknowledge that the accuracy of both the dynamical and statistical projections is inherently limited by the temporal resolution of the available climate forcing, and this limitation is now explicitly stated in the manuscript.

Lines 158-160: *It is noteworthy that the temporal resolution of the climate forcing (3–6 hourly) is coarser than that used for the hindcast, which may affect the representation of extremes in the projections.*

In this study, however, both the dynamically downscaled projections and the statistical downscaling model are constrained by the same 3-hourly atmospheric forcing available for the climate simulations. The SDM is trained and applied using predictor–predictand pairs derived under identical forcing frequency (1 hourly) and is therefore designed to reproduce the response of the hydrodynamic model under 3-hourly forcing, rather than to correct for temporal resolution-related biases. The impact on the EOFs is expected to be limited as they are designed to represent synoptic variability.

Lines 290-293: *While the SDM is trained on daily-aggregated data at original 1 hourly resolution (hindcast), the comparison between dynamical and statistically downscaled projections remains internally consistent under the coarser 3-hourly climate forcing, as the statistical model is trained and applied using predictor–predictand datasets at the same temporal resolution, and the impact on the EOFs is expected to be negligible.*

Line 120*: Please explicitly state that the SDM is based on multiple linear regression.*

Corrected, Line 165 now: "*Following our goal to study storm-surge extremes, we build a statistical downscaling model based on multiple linear regression to establish…"*

***Line 121**: Why are daily maxima chosen rather than another temporal aggregation? Clarify the distinction between predictor and predictand. As written, this section is confusing. If daily maxima are the target variable, it would be useful to evaluate how daily maxima from the DDM compare with GESLA observations.*

We have clarified the motivation behind daily maxima:

Lines 168-170*: We target daily maxima to represent event-scale storm-surge extremes, retaining synoptic variability while reducing temporal dimensionality, as commonly done in previous statistical storm-surge downscaling studies employing multiple linear regression (e.g., Cid et al., 2017; Tadesse et al., 2020; Tausía et al., 2023).*

We slightly re-worded the section for predictor-predictand clarity, we hope it's clear enough now.

Lines 165-168: *"Following our goal to study storm-surge extremes, we build a statistical downscaling model based on multiple linear regression to establish a relationship between the dynamically downscaled daily maxima storm-surge at each target coastal location (predictand) and daily aggregated forcing atmospheric fields (SLP, U10, V10) in a region of influence around each coastal point (predictors)."*

As for the suggestion to compare daily maxima against GESLA, daily maxima is not the target variable per see, but the temporal aggregation that allows us to retain extremes at the event scale, so that we can analyze extremes afterwards. **Extremes are our real target**. This has been clarified in the objective of the paper (see Lines79-84, modified in response to the general commentary, see above) and throughout paper.

***PCA details:** How are atmospheric fields standardized prior to PCA? Is the PCA weighted? Wind components may dominate variance relative to pressure variables— this should be discussed.*

We had forgotten to add these details, thanks for raising them.

Lines 174-176 *"To avoid dominance by variables with larger variance, principal componenta analysis is applied to atmospheric predictor fields standardized by removing their mean and scaling them by the temporal standard deviation, with no additional weighting."*

***Lines 130–134:** Several steps here require more detailed description for reproducibility, and some claims (e.g. that the SDM "effectively" captures certain behavior) need supporting evidence.*

A more elaborated explanation of the setup for the SDM under climate forcing is provided now in section 2.5 (new), lines 285-293. The interpolation to a common 1-degree grid is kept in the same location (section 2.3) because it needs to be highlighted before the SDM calibration, which is done on coarsened ERA5 fields to 1 degree.

Lines 285-293 :"*For statistical projections for validation of the hindcast-trained SDM (SD_hind in Table 3) under climate forcing (experiment 2) and subsequent ensemble projections (experiment 3), CMIP6 forcings are bias-corrected relative to the ERA5 reanalysis (both previously remapped to 1° resolutions, see section 2.3**Error! Reference source not found.**) before applied to the SDM by adjusting their mean and standard deviation at each grid cell over the reference period 1995-2014. This step crucially ensures compatibility in magnitude and variance between CMIP6 and*

ERA5 predictors, and hence a consistent projection of CMIP6 predictors onto the ERA5-based principal components. While the SDM is trained on daily-aggregated data at original 1-hourly resolution (hindcast), the comparison between dynamical and statistically downscaled projections remains internally consistent under the coarser 3-hourly climate forcing, as the statistical model is trained and applied using predictor–predictand datasets at the same temporal resolution, and the impact on the EOFs is expected to be negligible."

Table 3 Experiments involved in the validation of the statistical model (SDM) for past and future extreme storm surges (ESSs), and SDM application for ensemble projections of ESS changes. *SD* and *SD_hind* refer to statistical projections using an SDM trained in each GCMs historical simulation and the hindcast simulation, respectively. See section 2.5.1 for further details.

	Experiment	SDM training dataset	SDM training period	Timespan for stationary EVA		Forcing pre-processing
				Historical	Future	
1	Statistical hindcast reconstruction	Hindcast simulation	1997-2021	1997-2021	x	Degraded to 1° resolution
2	Validation of SDM under climate forcing	*SD*: historical climate simulations	1995-2014	1995-2014	2080-2099	Degraded to 1° resolution
		SD_hind: hindcast simulation				Degraded to 1° resolution + bias corrected relative to ERA5
3	Ensemble projections	*SD_hind*: hindcast simulation	1997-2021	1985-2014	2070-2099	Degraded to 1° resolution + bias corrected relative to ERA5

We have softened the claim around the fact that the SDM may incorporate some interpolation effect between storm surges forced by low and high resolution forcing:

Lines 181-182: "*Consequently, when applied to 1° CMIP6 predictors, the SDM may partly reflect differences between storm-surge responses linked to coarse and higher-resolution atmospheric forcing through the learned regression relationships*"

Bias correction and EOFs *(Line 133): Bias correction assumes stationarity. Did the authors verify that leading EOFs from ERA5 are spatially coherent with those from bias corrected CMIP6 fields? Similarly, was the internal variability of the principal components examined?*

We thank the reviewer for raising this important point. In this study, stationarity is assessed at the level of the SDM **output**, rather than by separately examining individual components such as EOF spatial patterns. This choice reflects the overarching philosophy of the methodology, which focuses on the stability of the surge-relevant atmospheric–oceanic relationship rather than on the stationarity of intermediate statistical constructs.

The philosophy of assessing aspects such as stationarity in time (section 3.1.1) and extrapolability to climate models (section 3.1.2) through the analysis of the SDM storm-surge outputs allows to propagate the impact of these two aspects on all the components of the SDM where they may play a role (bias correction, EOFs, regression coefficients). We have clarified this methodological choice and its rationale in section 2.5.1 'statistical storm surge projections : Validation under climate forcing':

Lines 320-325: *" The purpose of these experiments is to evaluate the stationarity of predictor-predictand relationships between periods and the extrapolability of the hindcast-based statistical relationships to climate models, ultimately validating the use of the hindcast-trained SDM for ESS projections. These crucial aspects are evaluated at the level of the storm surge outputs rather than through separate analyses of individual SDM components such as the EOFs or the regression coefficients. This integrated approach allows to propagate stationarity and extrapolability issues across the different parts that compose the SDM (bias correction, EOFs, regression coefficients) and assess their combined effect on the target storm-surges. "*

While we did not perform a separate, explicit comparison of EOF spatial patterns between ERA5 and bias-corrected CMIP6 fields, nor a standalone analysis of PC internal variability, **we have evaluated the retained variance** after projecting the bias-corrected CMIP6 fields onto the ERA5-based EOFs. This was shown in Figure S8 in the submitted manuscript for the historical period. We have generated a similar figure for the future period. These figures provide:

1. Information on the similarity between ERA5 and each GCM's predictor variability, through the retained explained variance and
2. Information on the stationarity of the predictors, by comparing the retained variance between future and historical periods for each GCM.

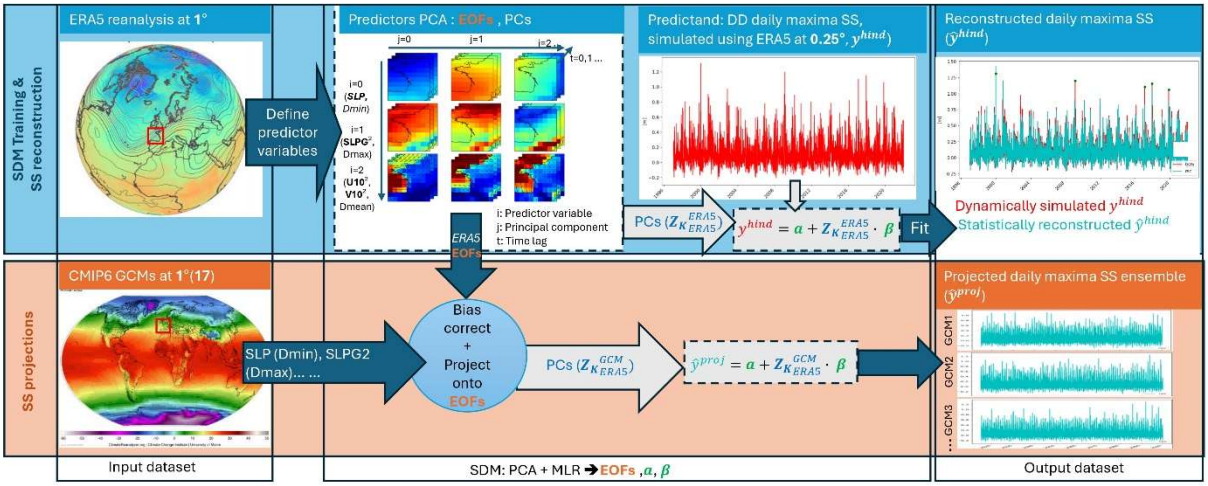
For (1), we observe that the retained variance is close to that designed for the SDM (99%) overall, but can be lower for certain regions in certain GCMs. This feature was highlighted in the discussion section in the submitted manuscript (lines 439-444, noting there was a typo on the figure referencing -it was S8 not S9). For (2), we observe very minimal differences between periods, confirming stationarity, and therefore refrain from showing the plot for the future. We have now extended the discussion on these two aspects, and moved it to the section 3.2.2 ('Extrapolation to climate forcing') in the revised manuscript (lines 502-511), to serve as basis for the legitimacy of the hindcast-trained SDM for climate projections, though ultimately the decisive tests are those reflecting the stationarity and extrapolability <u>for storm surges</u> presented in this section. We also highlight that this section includes now a much more elaborated description of the results (as pointed out in the reply to main commentary), including performance metrics added based on the general need to better *quantify* and hence justify the validity of the SDM for ESS change projections, raised by both reviewers.

Lines 502-511: "*However, differences with GCM-specific statistical projections (SD) can be notable for some coastal sections, which might be explained by the fact that ERA5-based EOFs do not always fully explain GCM predictor variability for specific models and regions. As such, SD_hind reconstructions can only account for future storm surge changes linked to the identified ERA5 principal components, and not to novel atmospheric conditions or different modes of variability/covariance structures that may be present in GCMs. An analysis of the retained explained variance after projecting GCM fields onto hindcast-based principal components for the target 17-GCM ensemble (Fig S8, historical climate) reveals generally strong representativity across Europe, though with notable reductions for certain models and regions (e.g. the*

Mediterranean Sea for HadGEM3-GC31-MM). The retained variance also reveals stable between historical and end-of-century climates (not shown), supporting the stationarity of the predictands. Further analyses are needed to understand the observed differences between ERA5 and GCM variability and their impact on ESS change projections."

Figure 2: Currently unreadable; resolution and clarity should be improved.

We have improved the resolution of the Figure 2 (now Figure 4) and added details for clarity, and increased the font overall. We have also changed to algebraic notation for simplicity, and expanded the terms in the figure caption.



Workflow followed for the training and application of the statistical downscaling model (SDM) projections of daily maxima storm surges (SS, represented by y), illustrated for La Rochelle, France. Predictor atmospheric variables are those corresponding to the chosen configuration T5 defined in Error! Reference source not found.**. The SDM is trained on the dynamical downscaling model hindcast simulation outputs (y^{hind}), forced by the ERA5 reanalysis. The SDM is defined by the empirical orthogonal functions (EOF) – extracted through principal component analysis (PCA) and representing the dominant modes of variability in the atmospheric fields around the target coastal point – and the linear regression coefficients (a, β) derived from multiple linear regression (MLR) between the principal component series (Z^{ERA5}_{ERA5}) and y^{hind}. SDMbased reconstructions ($\hat{y}^{hind}$) successfully reproduce the hindcast outputs (y^{hind}). Once these SDM elements are defined, global climate model (GCM) atmospheric fields from CMIP6 interpolated at 1 degree are projected onto the EOFs (Z^{GCM}_{ERA5}) and combined through the regression coefficients to produce storm-surge projections ($\hat{y}^{proj}$).**

Section 2.3

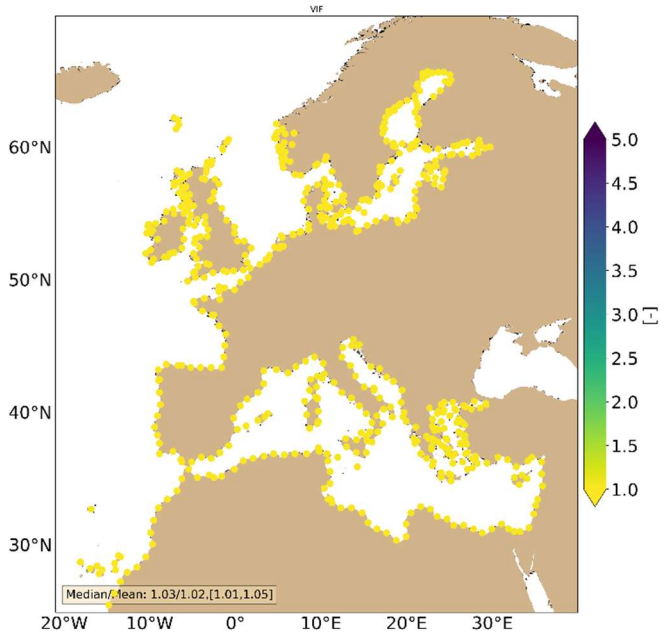
I suggest renaming this from "Calibration" to "SDM selection".

The section has been now absorbed into section 2.3 (following the general reorganization of the paper), and *calibration* has been replaced for *SDM selection* throughout the paper.

***Multicollinearity and significance:** After applying MLR, how are predictor variables assessed for multicollinearity and statistical significance?*

We thank the reviewer for this question. In the present framework. The predictors used in the multiple linear regression are principal components derived from a PCA, which are by construction orthogonal and therefore uncorrelated. As a result, multicollinearity among predictors is inherently minimized.

A plot of the median variance inflation factor (VIF) for the leading 10 principal components for the hindcast simulation shows VF~1, confirming very low collinearity:



We have added a phrase to highlight this feature of PCA in lines 172-174: *"Principal component analysis (targeting a 99% of explained variance) is employed to reduce the high dimensionality of the predictors to lighten and stabilize the subsequent regression, as orthogonal principal components inherently minimize multicollinearity."*

We did not perform significance-based filtering or regressors. Cross-validation was employed to find the configuration that best generalizes and hence present the less amount of unstable modes and non-significant predictors, but selected model configuration may still include predictors with limited individual relevance for daily maxima storm-surge. This is now acknowledges in lines 212-214:

"Coefficient-level significance filtering is not applied; robustness is instead sought through the assessment of a large set of configurations, acknowledging that some residual unnecessary complexity may remain. ***"***

Figure 3: *This figure is very insightful. Please refer to sub-panel labels rather than "top" and "bottom". Are the metrics shown in Figure 3 averaged over the five cross-validation folds? How is over- or under-fitting avoided? Overall, I recommend moving results currently embedded in the Methods section to the Results section.*

We have modified the figure caption to refer to sub-pannel labels. Metrics are computed over the concatenated test folds (5-yr each) in each cross-validation fold, leading to a reconstruction of the 25-yr timeseries based on test folds only , as highlighted in lines 173-176 of the submitted manuscript, and on the figure caption *"Metrics are derived over the 1997-2021 statistically downscaled daily maxima storm surge, reconstructed piece-wise for data independent from training following a k-fold approach".*

As highlighted in previous answers, the SDM selection is considered part of the methodology section, with the Results section focusing on storm surge reconstructions/projections.

Finally, cross-validation is employed to select the configuration with the best out-of-sample performance among those tested. While this provides a clear assessment of generalization error, it does not diagnose its origin, and some residual over- or under-fitting may remain,

especially where performance is modest. This reflects a necessary compromise given the finite number of configurations that could be explored. Our choice of possible configurations and selection approach is strongly inspired in previous works (e.g. Tadese et al. 2020, Tausia et al. 2023); as highlighted in previous answers, our objective is not develop a novel SDM (or a novel methodology for SDM model selection, relative to previous studies) but to select an SDM largely based on previous works but optimized for Europe.

Selection of T5-D9-L2: *The selection appears somewhat arbitrary as differences between T5, T6, and T7 are small in several regions (I acknowledge the large effort to generate a large set of model configurations). Expanding the colorbar range or using standardized error metrics could help. Figure 4 aids interpretation, but scatter plots comparing DDM and SDM time series would further strengthen evaluation.*

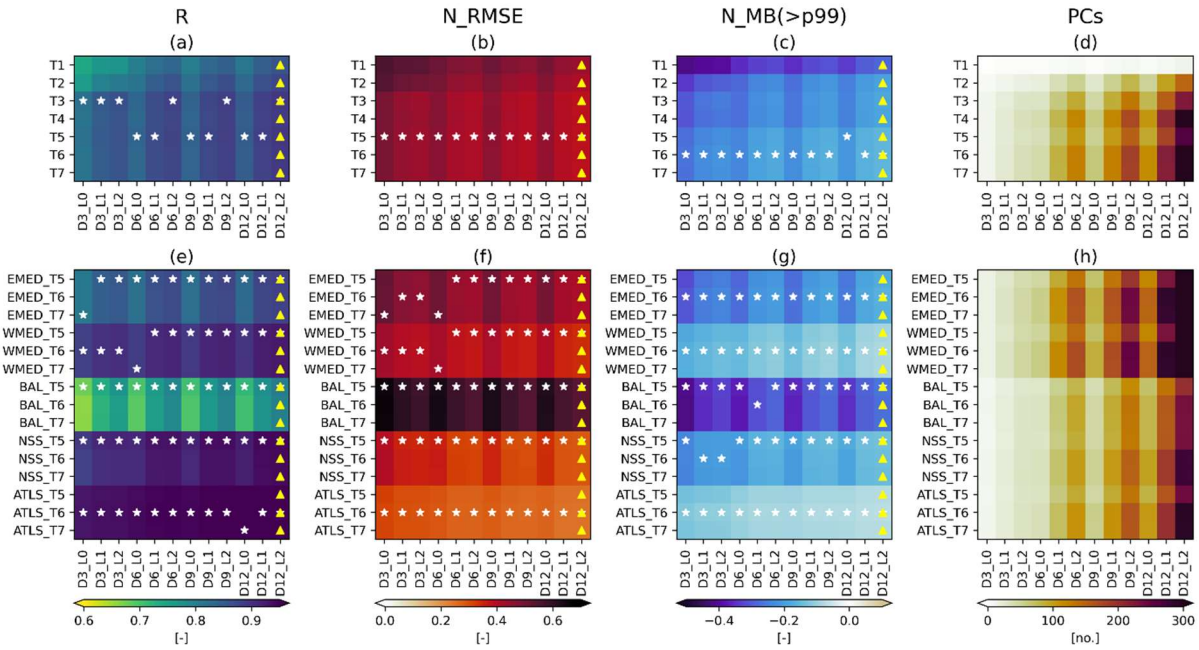
We thank the reviewer for this comment and agree that the skill differences between configurations T5, T6, and T7 are relatively small in several regions. However, as shown in Fig. 3-h, the impact of the temporal aggregation choice on the number of retained principal components is substantial: configuration T5 consistently requires fewer PCs than T6 and T7 across domains and lag combinations, as indicated by the lighter colours.

In view of this, we prioritize configuration T5 as the least complex model that achieves comparable skill while providing a stronger dimensionality reduction. This suggests that T5 extracts more stable and significant modes of variability and limits the contribution of higher-order, noise-dominated components. We have clarified this rationale in the revised manuscript (lines 245–247).

Lines 245–247: *"Performance differences between T5, T6 and T7 are negligible; T5 is therefore favoured due to its lower complexity, which allows comparable skill to be achieved with fewer retained principal components, suggesting a more stable and less noise-dominated representation of variability."*

Regarding normalized metrics, we had produce both absolute (in paper) and normalized metrics (shown below) in our analyses. We decided to retain the former for the paper to make the errors more explicit, and be able to compare to other studies. In the normalized metrics (RMSE, MB(>99th percentile)), the poorer performance for the Baltic and eastern Mediterranean is more obvious (see below). We had chosen to refer to these normalized performance only in the text (Line 202 in the original manuscript, now line 241), and we stand by this choice, but we now add the comment that these are not shown:

Lines 241-243: *Considering characteristic regional storm-surge variance (i.e., normalizing RMSE and MB, <u>not shown</u>), the SDM performs best along the Atlantic façade (ATLS), North Sea (NNS), and western Mediterranean (WMED), while performance is weaker in the eastern Mediterranean (EMED) and especially in the Baltic Sea (BAL).*



Finally, Figure 4 was not intended for comparison between SDM configurations but to illustrate the performance of the selected model for past storm surge reconstructions. As such, and given the reorganization of the paper, we have moved this plot to section 3.1, and added quantile-quantile plots side by side to the timeseries plots. This section now also contains the explicit evaluation of storm-surge extremes derived using EVA.

Figure 4 caption: Please explicitly refer to daily maxima values.

We have corrected this, thanks for spotting it.

Section 2.4 Extreme Value Analysis

***Line 221**: Is the period 20 or 30 years? Please clarify.*

The exact period depends on the specific analysis at hand. This ranges between 20 years (e.g. in the validation of projected changes against dynamical simulations, which are only available for 20-years in the historical period) and 30-years (used in the multi-model ensemble projections, as we statistically downscaled a longer portion of the historical climates to perform more robust EVA on projected changes). This is now explicitly pointed out in Table 3 for the different experiments performed, and is referenced in the EVA section, lines 263-265.

Lines 263-265: "*The periods used for the stationary-on-slice EVA range from 20 to 30 years, depending on the experiments at hand, which are further elaborated on the following section 2.5 (Table 3).* "

Is POT applied to daily maxima? What percentile corresponds to an average of three exceedances per year, and over which reference period? Please provide more detail on the GPD fitting procedure, parameter inference, and uncertainty estimation

We have added all of the details requested by the reviewer in this section. We have also elaborated on tests to evaluate the suitability of the adopted fit across experiments, which have been performed in this revision to add to the methodological rigor and support the discussion of whether a fixed vs. a free parameter should be used for evaluation of future changes in extremes.

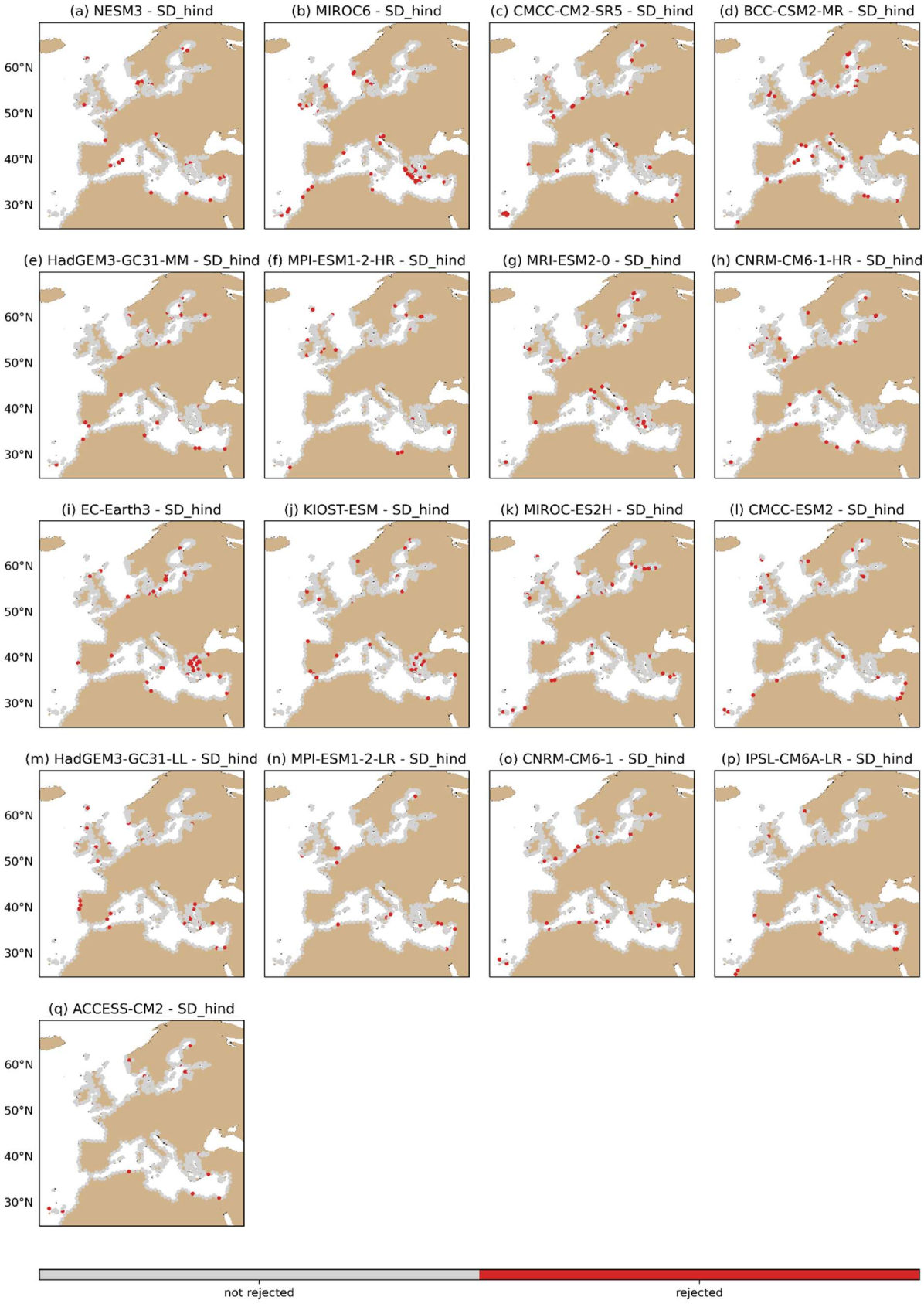
Lines 254-276: "*ESSs in past and future climates are analyzed through stationary-on-slice extreme value analysis (EVA). We fit a Generalized Pareto Distribution (GPD) on the daily-maxima storm-surges that exceed a threshold (known as the Peak Over Threshold method, POT). The selected threshold was based on an average rate of 5 extreme events per year in the considered time-slice, which corresponds to the 99.14th percentile in average across Europe for the hindcast simulation (1997-2021). Events are considered independent when separated by at least 3 days, considered the approximate time most storm events influence water levels at the coast (Wahl et al., 2017) and typically employed in extreme value analysis in Europe (Chaigneau et al., 2024; Haigh et al., 2016; Vousdoukas et al., 2016). GPD parameters are estimated using maximum likelihood estimation. Confidence intervals are derived using a mean-adjusted bootstrap (Efron & Tibshirani, 1994), based on 600 resamples, and reported at the 5th–95th percentile levels.*

The periods used for the stationary-on-slice EVA range from 20 to 30 years, depending on the experiments at hand (e.g. validation against dynamical projections, ensemble projections), which are further elaborated on the following section 2.5 (Table 3). Across all experiments, the suitability of the extreme-value model is assessed using the Anderson–Darling test at the 0.05 significance level. Both GPD and exponential distributions were evaluated. Results (not shown) indicate that the GPD provides a more robust fit across European coastlines, whereas the exponential distribution is frequently rejected over large coastal stretches for multiple climate models, in both historical and future periods, and for both dynamically and statistically downscaled datasets.

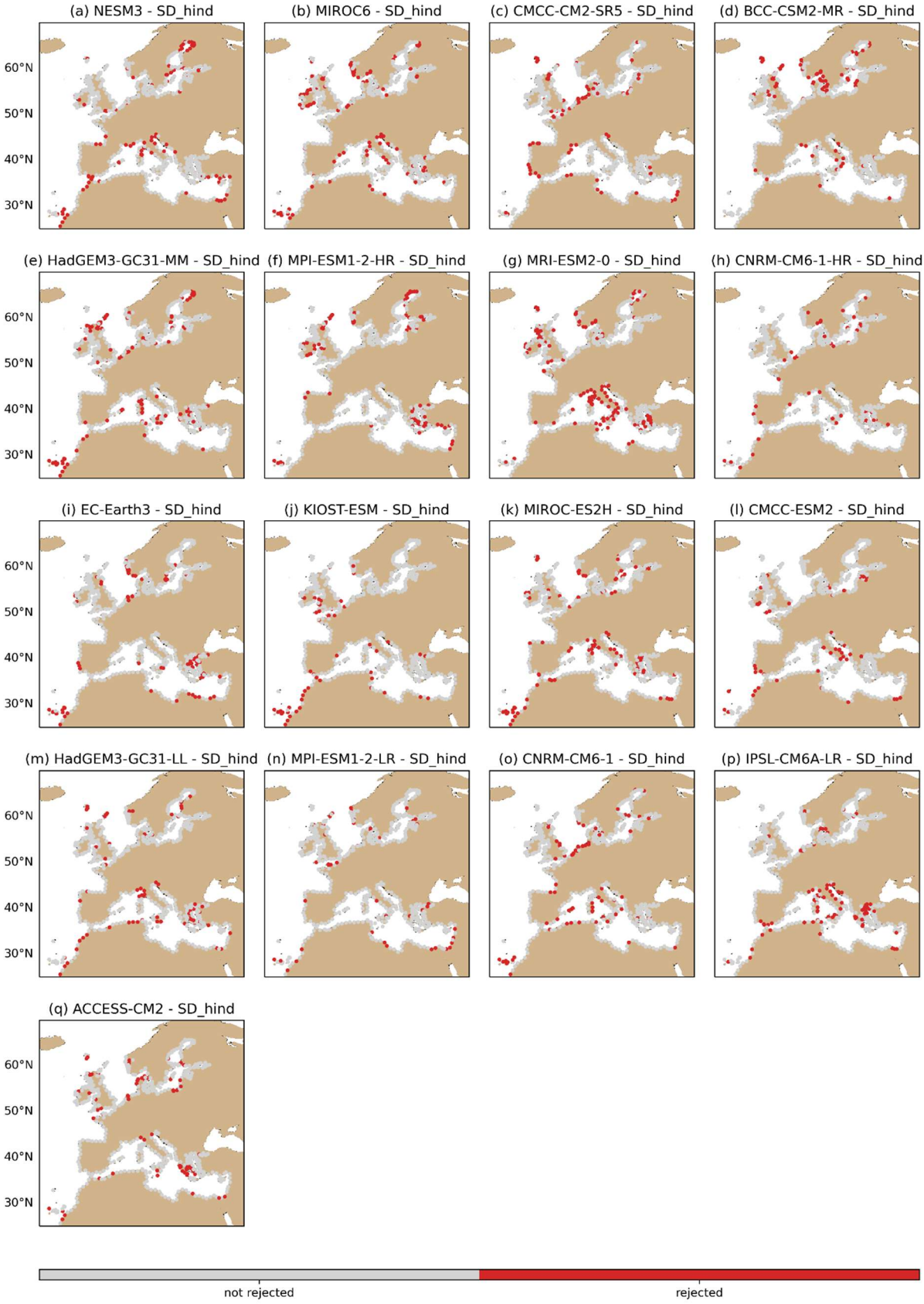
Accordingly, the GPD is adopted consistently across datasets and climates, with the shape parameter allowed to vary between periods. Sensitivity tests in which the future-period GPD shape parameter was fixed to its historical value resulted in widespread rejection by the Anderson–Darling test and were therefore not retained. While allowing the shape parameter to vary increases uncertainty in high return-level estimates compared to an exponential fit—potentially reducing the detectability of future changes—our tests indicate that this flexibility is required for an adequate representation of extremes. Under the retained distribution (GPD with freely varying shape), shape parameter estimates have been inspected and show to be stable across experiments, GCMs, and epochs, with predominantly negative values (except in the Canary Islands) and absolute magnitudes below 0.5."

We attach here the plots with the Anderson Darling test performed across the ensemble projections for 2070-2100 (similar to those based on 20-year EVA for validation of SD vs DD, and to the hindcast) :

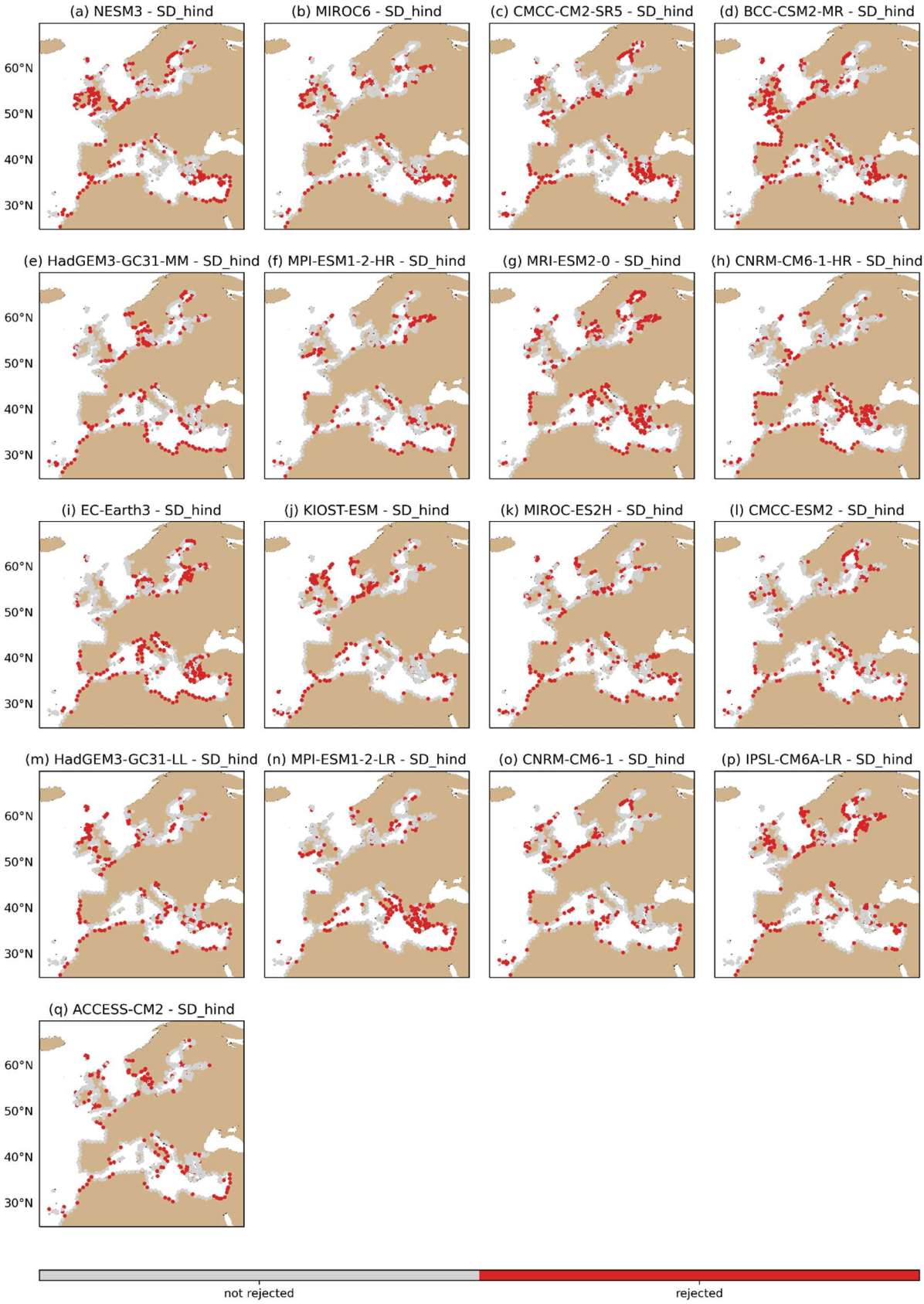
- Free shape (chosen distribution):



-
- Exponential


(a) NESM3 - SD_hind
(b) MIROC6 - SD_hind
(c) CMCC-CM2-SR5 - SD_hind
(d) BCC-CSM2-MR - SD_hind
(e) HadGEM3-GC31-MM - SD_hind
(f) MPI-ESM1-2-HR - SD_hind
(g) MRI-ESM2-0 - SD_hind
(h) CNRM-CM6-1-HR - SD_hind
(i) EC-Earth3 - SD_hind
(j) KIOST-ESM - SD_hind
(k) MIROC-ES2H - SD_hind
(l) CMCC-ESM2 - SD_hind
(m) HadGEM3-GC31-LL - SD_hind
(n) MPI-ESM1-2-LR - SD_hind
(o) CNRM-CM6-1 - SD_hind
(p) IPSL-CM6A-LR - SD_hind
(q) ACCESS-CM2 - SD_hind
60°N
50°N
40°N
30°N
not rejected
rejected

- Fixed shape to the historical period:



Section 3 Results

The opening paragraph of Section 3 describes methodology and should be moved to the Methods section.

We have moved all methodological descriptions in the results to the corresponding sections in Methods, in this case to section 2.5.

***Lines 233–235**: Please clarify the experimental design used to test the stationarity of the predictor–predictand relationship. Can MLR capture changes in variability and trends, or is something else implied? Given the bias correction and projection in ERA5 EOF space, it is important to test whether SD-MLR errors using CMIP6 inputs are homoscedastic*

The experimental design for stationarity (and extrapolability) is now described in section 2.5.1. The MLR can capture changes in variability and trends associated with the identified modes of variability, this is now clarified in lines 352-354.

Lines 352-354*: "We highlight that although the EOFs and regression coefficients are assumed stationary, SDM-based storm-surge estimates can reflect temporal changes in variability and trends through changes in the atmospheric predictors associated with the identified modes of variability."*

Regarding homoscedascity, our results show that the errors are indeed heteroscedastic as they increase for larger storm surges (as illustrated in the validation for extreme events added in this round of reviews, Figure 6 , showing larger errors for the 100-yr than the 10-yr event), suggesting a non-linear predictor-predictand relationship across average and extreme conditions. Our regression is calibrated on the full range of storm-surge values and therefore optimizes overall fit rather than tail-specific performance. This is a limitation of the adopted linear regression and ordinary least squares approach. We don't expect the error structure to qualitatively change under bias-corrected CMIP6 forcings. We have now highlighted this in the results, lines 413-417:

Lines 413-417: *"The increase of error for increasing storm surge magnitudes suggest that the storm-surge–predictor relationship departs from linearity between average and extreme conditions. This reflects the ordinary least squares formulation, which optimizes the mean response and leads to heteroscedastic errors for rare extremes; this behaviour is inherent to the methodology and is not expected to change qualitatively when the model is driven by bias-adjusted CMIP6 predictors."*

We note that a direct evaluation of the error structure under bias-corrected CMIP6 inputs would require a reference ("ground-truth") surge simulation driven by the same (bias-adjusted) CMIP6 atmospheric fields, which is not available in our framework. We therefore cannot compute statistical projection residuals in a strict sense. Nevertheless, our evaluation of SDM extrapolability under CMIP6 forcing is based primarily on the ability to reproduce **projected changes** in extreme storm surges (future vs. historical), rather than absolute surge levels. This choice is motivated by the presence of systematic GCM biases in atmospheric forcing (e.g., wind speed biases) that can induce biased absolute storm surges in the CMIP6-driven numerical simulations, which have not been bias corrected; focusing on changes largely mitigates this effect and allows for benchmarking against dynamical simulations, and its standard practice in climate projections (*delta* method). In this context, strict homoscedasticity of absolute SD-MLR residuals is not a central requirement; instead, the relevant question is whether the error structure is stable across periods (stationarity) and does not introduce

spurious trends in the extremes. We therefore assess the skill of the statistical projections by comparing SDM-derived and dynamically simulated **changes in return levels**. This rationale is now clarified in section 2.5.1:

Lines 356-366*: "Extrapolability to climate forcing: The combined effect of stationarity and extrapolability are finally assessed through evaluating the ability of the hindcast-trained SDM to capture projected changes in ESSs given by dynamical simulations. Projecting changes in ESSs constitutes the ultimate objective of our study. Projecting relative changes between historical and future periods and then adjusting observational or reanalysis baselines accordingly —called the delta method—is commonly done in climate science to minimize the influence of biases in the GCMs in future projections. As previously highlighted, biases have been corrected for SD_hind (through simple mean bias and variance corrections, for safe projection onto ERA5-based EOFs), but they haven't been corrected for SD and DD estimates. Additionally, even when corrected, biases in extremes haven't been explicitly addressed, which may differ from those related to mean conditions. Together, these factors justify our focus on evaluating projected changes of ESSs (future vs. past) instead of directly assessing their future projections in the DD/SD/SD_hind intercomparison."*

For completeness, we have slightly extended the section of the discussion that highlights alternative methodological aspects of the statistical method to improve the performance for extremes:

Lines 625-627: "*When targeting extreme events, the multiple linear regression can be modified to optimize for extremes, for example through generalized linear formulations or weighted regression approaches*."

Line 235: The claim "To our knowledge…" should be moved to the Introduction and supported with relevant literature

This was already highlighted in the Introduction, referencing the relevant literature. We have now added the study of Dubois et al. (2025) (10.16993/tellusa.4101) which was published during the review process:

Lines 71-78: *"While widely used for past storm surge reconstructions, the application of statistical downscaling to future storm surge projections remains limited. Only a few studies have explored this approach for regional projections, such as Cagigal et al. (2020) for New Zealand and Boumis et al. (2025) for Japan. In Europe, regional-scale statistical ESS projections remain scarce, having been developed only for the Baltic Sea (Dubois et al., 2025) and based on a limited set of 4 GCMs and discrete coastal locations. Furthermore, such studies have not evaluated whether the statistical relationships established under past conditions (whether from observations of reanalyses) remain valid under forcing from climate models, nor have they explicitly assessed the assumption of stationarity in the predictor–predictand relationship between past and future periods. As such, the reliability of statistically downscaled storm surge projections under climate change remains under-explored."*

The highlighted claim was reiterated here to emphasize this novelty aspect of the study. The phrase has now been moved to the Method's section describing the procedure for validation of statistical storm-surge projections (section 2.5.1).

Section 3.1 – SDM Evaluation

The experimental design belongs in Methods.

It has been moved to section 2.5.1('Experimental design') in Methods.

If 10-year return levels are used, why not also use empirical return values?

Our evaluation focuses on projected *changes* in the 10-year return level - rather than on absolute values - as already justified before. Such changes cannot be strictly isolated from the sampled extremes. Moreover, absolute return levels may differ between dynamically and statistically downscaled datasets due to systematic biases in GCMs, whereas changes between periods may be well resolved. This rationale is now clarified in Section 2.5.1 (Lines 356-366, see previous answers for the corresponding text).

The terminology should be consistent: use reconstruction for past climate and projection for future climate.

We have corrected the text to consistently refer to *reconstruction* for past climate and *projection* for future climate, and simply (DD/SD/SD_hind) *estimates* when referred to either or.

Training appears to occur over the same period—this is benchmarking rather than validation.

What we are validating here is not the statistical estimates relative to the dynamical ones (i.e., the value in each of the matrices presented, for which indeed the *SD* estimate for the historical period would be benchmarking not validation) but the stationarity of such error (compare future vs historical matrices). This is highlighted at the beginning of this results section:

Lines 431-432:*"Following the procedure described in section 2.5.1, stationarity in the predictor–predictand relationship is assessed based on the stability of the SDM skill between historical and future climates. "*

The hypotheses should be stated clearly: Both SDMs (ERA5-trained and CMIP6-trained) reproduce the distribution of daily maxima from the DDM. SDM performance is stationary across climate periods

As explained in the main reply, we have based our validation on skill metrics to diagnose agreement between statistical and dynamically downscaled storm surge projections, rather than formal hypothesis testing.

To clarify the purpose in this section, the hypothesis here is not comparing SDM to DDM daily maxima distributions. It's about evaluating the stationarity of the differences between the two (the error). This is now explicitly explained in section 2.5.1 (method for evaluating stationarity):

- Lines 347-354: "*Stationarity assumption: We compare the skill of the SDM to reproduce storm surge climates in both historical and future periods when the model is trained exclusively on historical outputs of the dynamical model, either from the hindcast (SD_hind) or from each GCM-specific historical climate simulations (SD). When the skill*

(Pearson correlation coefficient, RMSE, MB above the 99th percentile) is stable between future and historical periods, the stationarity assumption is validated. We highlight that although the EOFs and regression coefficients are assumed stationary, SDM-based storm-surge estimates can reflect temporal changes in variability and trends through changes in the atmospheric predictors associated with the identified modes of variability."

Figure 5- Showing bias and variance ratios would be informative for the first hypothesis above.

We think that the chosen metrics reflect the performance in each period well (R for temporal coherence of the signal, RMSE for general amplitude and MB(p>99) performance for extremes)

Line 272: Please support this statement with quantitative estimates.

Lines 448-450:"*Nevertheless, when comparing performance metrics between historical and future periods, a remarkably strong stationarity is found for both SD and SD_hind and across performance metrics, with maximum regional absolute differences between periods of 0.02, 0.01m and 0.06m for correlation, RMSE and MB above the 99th percentile, respectively, across SD and SD_hind estimates*"

Section 3.1.2 – Future ESS Changes

I am not convinced the 10-year return level metric robustly tests the hypothesis that SDM projections capture future change. I recommend formal hypothesis testing: Test whether historical and future distributions differ (e.g. KS test). Test whether GPD shape parameters differ (e.g. Wald test or bootstrap).

We thank the reviewer for this comment and hope we have interpreted it correctly. We clarify that this section is not intended to assess the detectability or statistical significance of future changes in extreme storm surges (as suggested by the proposed statistical tests, again hoping we have interpreted the comment correctly), but rather to evaluate whether the SDM **reproduces the future ESS changes** simulated by the dynamically downscaled benchmark. In particular, we focus on changes in the 10-year return level, which was originally motivated by the wish to validate the SDM on a robust metric compared to higher return levels which may be strongly affected by sampling uncertainty. However upon reflection, we have realized that sampling uncertainty (associated with a limited sampling of very rare events in our dataset) should play a minimal impact in the SD vs DD comparison because the two estimates are targeting the same extreme events, so sampling uncertainty should be largely similar, and any differences for high return periods should predominantly stem from SDM errors in representing those extreme events relative to DD. This has now been highlighted in the Methods section 2.5.1.

Lines 342-344: *" As all downscaled estimates (DD/SD/SD_hind) represent the same storm surges under identical forcing and periods, sampling uncertainty is expected to be largely shared across estimates, and model evaluation focuses on point estimates."*

All in all, we have reformulated the paper results to focus on projections for the 10-year event based on the poorer SDM performance identified for higher (e.g. 100 year) return periods in the hindcast reconstruction (Figure 6 in revised manuscript), as suggested by the first reviewer.

Lines 418-420: "*Based on these results, we decide to focus statistical projections in the following to the 10-year storm-surge event to limit the impact of the decreasing SDM performance for high return periods on the confidence of the produced multi-mode statistical projections.*"

Therefore, this validation section still focuses on the 10-year event in this reviewed version:

Lines 454-457: *"Once the stationarity assumption validated, we next evaluate the extrapolation capability of the hindcast-trained SDM to climate forcing by assessing its skill to reproduce dynamically downscaled changes in ESSs (section 2.5.1), and which hence constitutes the ultimate test to justify its application for multi-model ensemble projections of changes in ESS (section2.5.2), focusing on the 10-year storm surge level. "*

Finally, to improve the quantification of the skill of the SDM in reproducing dynamically downscaled 10-year return level changes (point estimates), we have substantially expanded this section with a more detailed description of the results and with skill metrics to support the different claims, given both reviewer's comments. The skill metrics are the pattern correlation coefficient, the mean absolute bias and the % of points where the sign of the projected change is correct, in order to evaluate the SD vs DD match in the magnitude and spatial pattern of future changes. We copy here the whole section 3.2.2, with extended discussion of results, including the changes to the figures:

======Section 3.2.2==========

Once the stationarity assumption validated, we next evaluate the extrapolation capability of the hindcast-trained SDM to climate forcing by assessing its skill to reproduce dynamically downscaled changes in ESSs (section 2.5.1), and which hence constitutes the ultimate test to justify its application for multi-model ensemble projections of changes in ESS (section 2.5.2) focusing on the 10-year storm surge level.

*Dynamical projections reveal considerable inter-model spread, with regional changes typically reaching ±20% (Figure 8-a,e,i,m, 5th-95th percentiles of results pooled across GCMs), exceptionally higher (-25%/+32%, 1st/99th percentiles respectively). Statistical projections trained independently on each GCM (SD, Figure 8*Figure *-b,f,j,n) replicate the main European-scale spatial features of the dynamically downscaled projections, demonstrating the SDM's skill to replicate GCM-specific climate responses: mean absolute biases (MAB) remain modest (5–7.5%) and pattern correlation coefficients (PCC) for three of the four GCMs are >=0.7 (CNRM-CM6-1-HR being the exception with PCC=0.52), which is often deemed satisfactory in climate model performance evaluations (Back et al., 2024; Berhanu et al., 2025; Zebaze et al., 2025). Additionally, the sign of the projected change (sign agreement, SA) is correctly reproduced at ≥70% of grid points (67% for CNRM-CM6-1-HR). These indicators show that, despite some amplitude biases, the spatial imprint of the projected changes in the 10-year return level based on dynamical simulations is reasonably well reproduced by the SDM.*

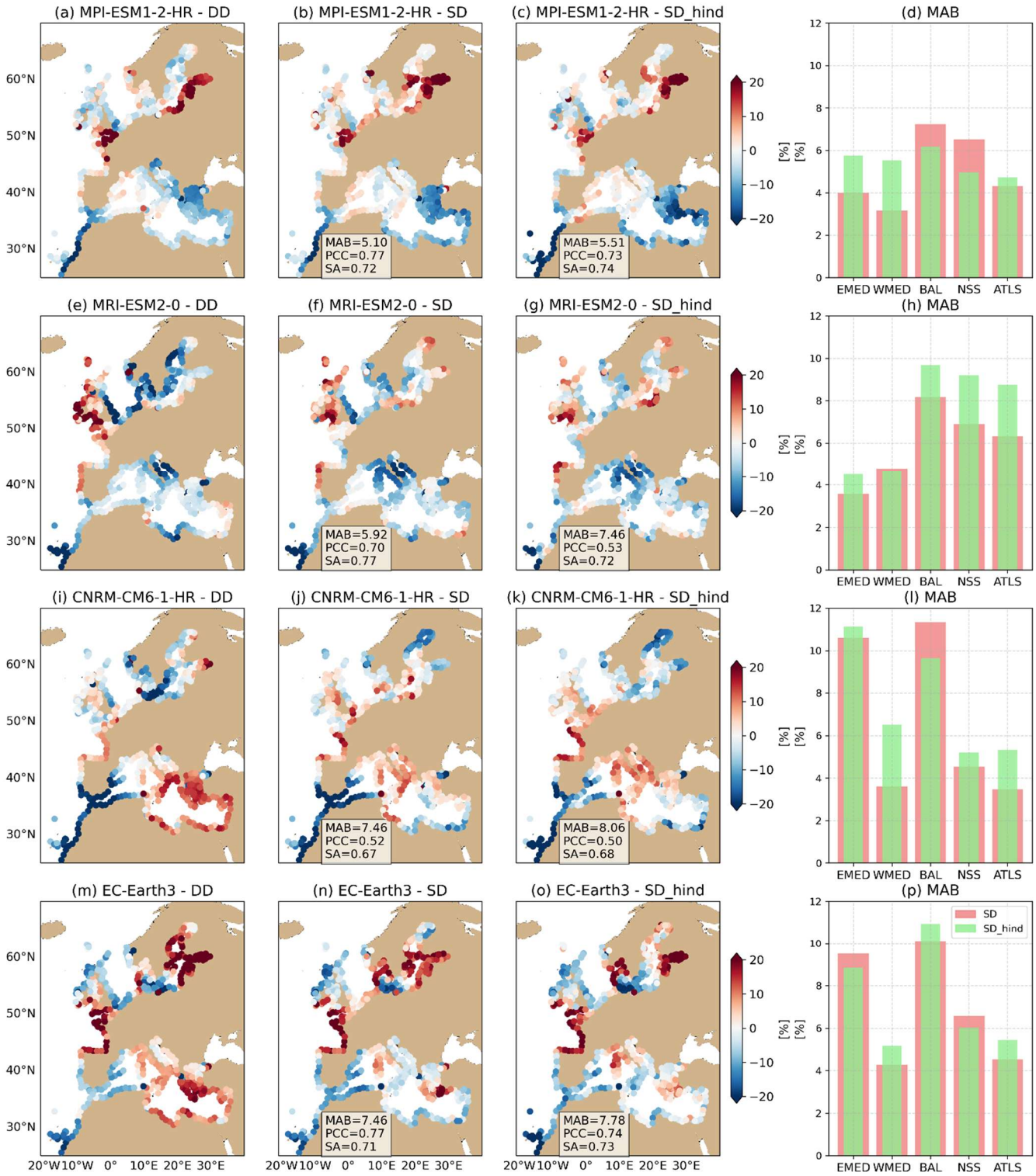


Figure 8 Projected changes (%, extreme value analysis on [2080-2099] vs [1995-2014]) in the 1 in 10 year storm surge event (RL10) for dynamical climate simulations (DD, a,e,i,m), statistical estimates trained on each historical climate simulation (SD, b,f,j,n) and statistical estimates trained on the hindcast forced by ERA5 (SD_hind, c,g,k,o) for the 4 GCMs downscaled. For each GCM (each row), the regionally averaged mean absolute bias (MAB) of SD and SD_hind estimates relative to DD estimates are given on the right-most column (d,h,l,p). EMED: Eastern Mediterranean Sea; WMED: Western Mediterranean Sea; BAL: Baltic Sea; NSS: North Sea; ATLS: Atlantic Shelf. See Figure 2 for the definition of the geographical coverage of each region. For a fair comparison between SD and SD_hind, both are trained on 20-yr periods (1995-2014 and 1997-2016 respectively).

Performance, however, varies considerably across regions (Figure 8-d,h,l,q). The Baltic Sea exhibits the largest amplitude errors for most GCMs, and both the Baltic and eastern Mediterranean perform notably worse for CNRM-CM6-1-HR and EC-Earth. While this could be attributed to the lower hindcast skill identified in these regions (Figure 3, Figure 6), performance in reproducing projected changes is not systematically lower for these regions across GCMs. In the eastern Mediterranean, poor SDM performance appears only for the two models projecting positive ESS changes (CNRM-CM6-1-HR and EC-Earth3), while for the other two GCMs (MPI-ESM1-2-HR and MRI-ESM2-0), the projected negative ESS changes with differing spatial patterns are well reproduced by the statistical model. In the Baltic Sea, dynamical ESS changes for MRI-ESM2-0 and CNRM-CM6-1-HR—which are broadly negative—are very poorly reproduced by the statistical model, whereas for MPI-ESM1-2-HR and EC-Earth3 the positive signal seen in the dynamical simulations is broadly retained in the statistical projections, albeit with reduced amplitude.

These results indicate that a limited hindcast skill of the statistical model does not necessarily imply a corresponding poor performance in projecting ESS changes. This likely depends on how well the SDM captures the effect of the dominant atmospheric predictors and associated variability modes driving future ESS changes: if these are well represented, the main climate-change signal can still be recovered even when other predictors or specific modes are less accurately represented. While out of scope for the current study, a more detailed analysis of the predictors and variability modes dominating projected ESS changes across GCMs should be pursued in future works to help clarify and better interpret the SDM's skill for climate projections, particularly in challenging regions such as the Baltic Sea and the eastern Mediterranean Sea.

When using the SDM trained solely on the hindcast (SD_hind, Figure 8-c,g,k,o), spatial patterns and relative amplitudes of changes in the 10-year return level are generally well preserved across GCMs, as reflected by the performance metrics (MAB, PCC and SA) which remain broadly comparable to those for GCM-specific statistical projections (SD). The exception is MRI-ESM2-0, for which performance decays substantially between SD and SD_hind in both the spatial pattern (PCC) and the amplitude (MAB) of the signal. This decay is largely owed to a pronounced reduction of the ESS change signal across the northwest Shelf (UK coasts, North Sea) (see regional metrics in Figure 8-h). Across GCMs, the transition to a hindcast-trained SDM tends to only moderately amplify regional amplitude biases (Figure 8-d,h,l,q). Overall, these results support the applicability of the SD_hind setup for climate projections, with the added advantage of requiring a single simulation for training (the hindcast). However, differences with GCM-specific statistical projections (SD) can be notable for some coastal sections, which might be explained by the fact that ERA5-based EOFs do not always fully explain GCM predictor variability for specific models and regions. As such, SD_hind estimates can only account for future storm surge changes linked to the identified ERA5 principal components, and not to novel atmospheric conditions or different modes of variability/covariance structures that may be present in GCMs. An analysis of the retained explained variance after projecting GCM fields onto hindcast-based principal components for the target 17-GCM ensemble (Fig S3, historical climate) reveals generally strong representativity across Europe, though with notable reductions for certain models and regions (e.g. the Mediterranean Sea for HadGEM3-GC31-MM). The retained variance also reveals stable between historical and end-of-century climates (not shown), supporting the stationarity of the predictands. Further analyses are needed to understand the observed differences between ERA5 and GCM variability and their impact on ESS change projections.

We finally compute performance metrics by pooling projections across GCMs for each region (Figure 9). Results confirm the eastern Mediterranean and the Baltic to be the worst performing regions, and notably so for the Baltic, with highest MAB (9.2%) and lowest PCC and SA (0.59 and 0.72, respectively, for SD). Given that dynamical simulations highlight these regions to display relatively strong future ESS changes (Figure 9-a), these results highlight the need to improve statistical projections in these regions for reliable future storm surge hazard assessments. The ensemble statistical projections presented hereafter should therefore be interpreted with caution in the Baltic and eastern Mediterranean regions. The best performing regions are the Atlantic façade and the western Mediterranean with lowest MAB (<5% for SD) and highest PCC and SA (>0.8 and >0.85, respectively, for SD). For the North Sea, results are somewhat mixed, as amplitude errors are moderate, the sign of ESS changes is well captured but the spatial pattern is less well resolved. The switch to a hindcast-trained SDM incurs a general but moderate decay in the SDM performance across regions, through with a notably larger impact on the western Mediterranean Sea.

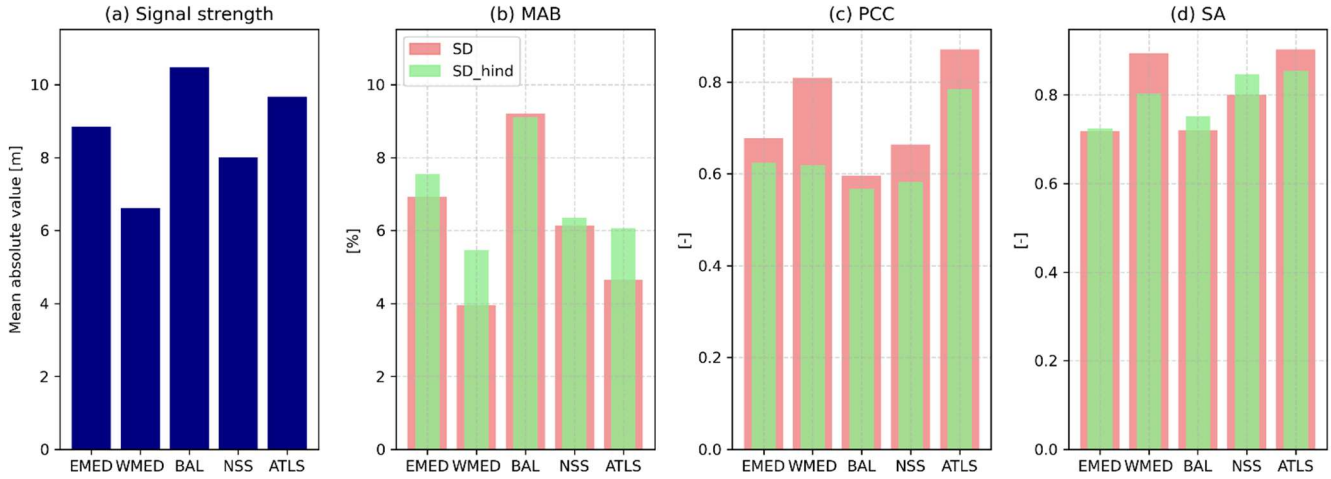


Figure 9 Regional performance metrics of statistically downscaled projections relative to dynamically downscaled projections computed by pooling projections across the 4 GCMs for each region. (a) Strength of the signal of projected changes in the 10-year storm surge return level (RL10) from dynamical simulations (regional mean of absolute projected changes). (b) Mean absolute bias (%). (c) Pattern correlation coefficient (PCC). (d) Fraction of coastal points per region for which the sign of the RL10 change is reproduced (-), considering coastal points where the absolute amplitude is >=5%.

Overall, the hindcast-trained SDM shows a sufficiently satisfactory skill in reproducing GCM-specific responses of European-scale ESSs simulated by dynamical simulations. While several limitations apply which should be considered when interpreting associated projections — including notably reduced skill in some regions (e.g., the Baltic Sea and eastern Mediterranean Sea), potential inconsistencies where GCM atmospheric variability departs from ERA5, and a systematic underestimation of ESS change magnitudes—our results support the broader use of the hindcast-trained SDM for cost-efficient multi-model projections of European-scale ESS changes. The SDM enables projections for a substantially larger ensemble of GCMs than previously reported, hence allowing a more rigorous identification of main regional trends, and importantly, a more comprehensive evaluation of inter-model variability in ESS projections, which has been poorly constrained in studies to date.

==========

Section 3.2 – Ensemble Projections

Line 344: Check parentheses and referencing style.

Thanks, corrected

Lines 341 onward: Agreement between SD and DD appears limited in regions such as the Baltic and North Sea. This weakens the stated confidence. This section would benefit from uncertainty and inference on return-level estimates.

Based on the 1st reviewer's comments, we have modified the description of the ensemble projection results to highlight the lower confidence in the Baltic and eastern Mediterranean, and we have added a hatching in the plots in these regions denoting lower confidence. Regarding the North Sea, based on the newly available skill metrics in section 3.2.2, we conclude that SDM skill in this region is lower than along the Atlantic façade/western Mediterranean but notably better than in the aforementioned low confidence regions. So the confidence in the North Sea, in relation to the SDM skill, is probably medium. We have added this now in the lines 574-577:

Lines 574-577: "*These studies also identify regions with substantial signals which are not emerging in our ensemble (e.g. the south-eastern North Sea in Vousdoukas et al., 2016), which may result from the use of smaller ensembles which underrepresent inter-model variance in storm-surge projections. In the North Sea, mismatches may be influenced by moderate skill of the SDM for future ESS changes (section 3.2.2).*"

Regarding uncertainty and inference, we have added the confidence interval range [95th - 5th percentile confidence limits] for each GCM for historical and end-of-century periods in Supplementary Materials(Figures S6-S8), calculated through adjusted bootstrap (see section 2.4). Finally, we refrain from evaluating statistical significance of the projected changes, as it should reflect both uncertainty stemming from the applied methods (e.g. EVA) but also significance relative to internal variability in extremes, which would typically require large multi-member simulations for each GCM which are out of scope for this study. However, the potential of using a cost-efficient statistical downscaling method for such a quantification was already highlighted in the discussion (lines 472-487 in the original manuscript) and remains a topic for future research.

Lines 542-544: "*For reference, 10-year event ESS changes for individual GCMs are provided in Figures S4,S5, and corresponding confidence interval (given by the 5th and 95th percentile levels) are provided in Figures S6-8 for each of the epochs assessed.*"

Figure 7: Please check the baseline period stated.

Thanks for spotting this typo, it has been corrected

Figure 8: Parts of the caption describe methods and should be moved accordingly. The random sampling of models is unclear: for 17 models taken 3 at a time there are 680 combinations—why 2000 random iterations?

Given the substantially increased length of the manuscript after the revisions, we have decided to remove this Figure and corresponding description in the final version.