Answers to RV1

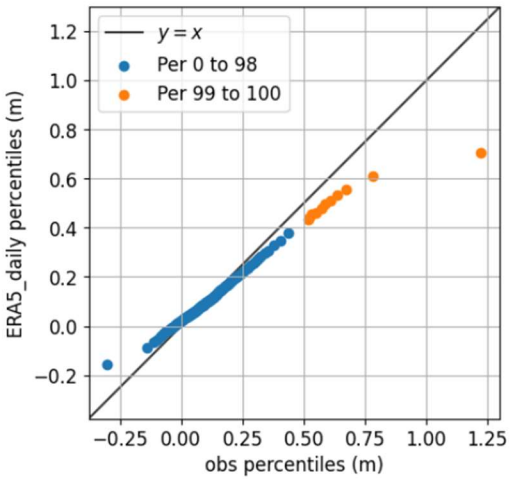
This manuscript presents a statistical reconstruction of storm surge records using 17 climate model projections along the European coastlines. It is the result of a significant computational effort, combining a (relatively) small set of dynamical numerical simulations and a data-driven model based on 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.



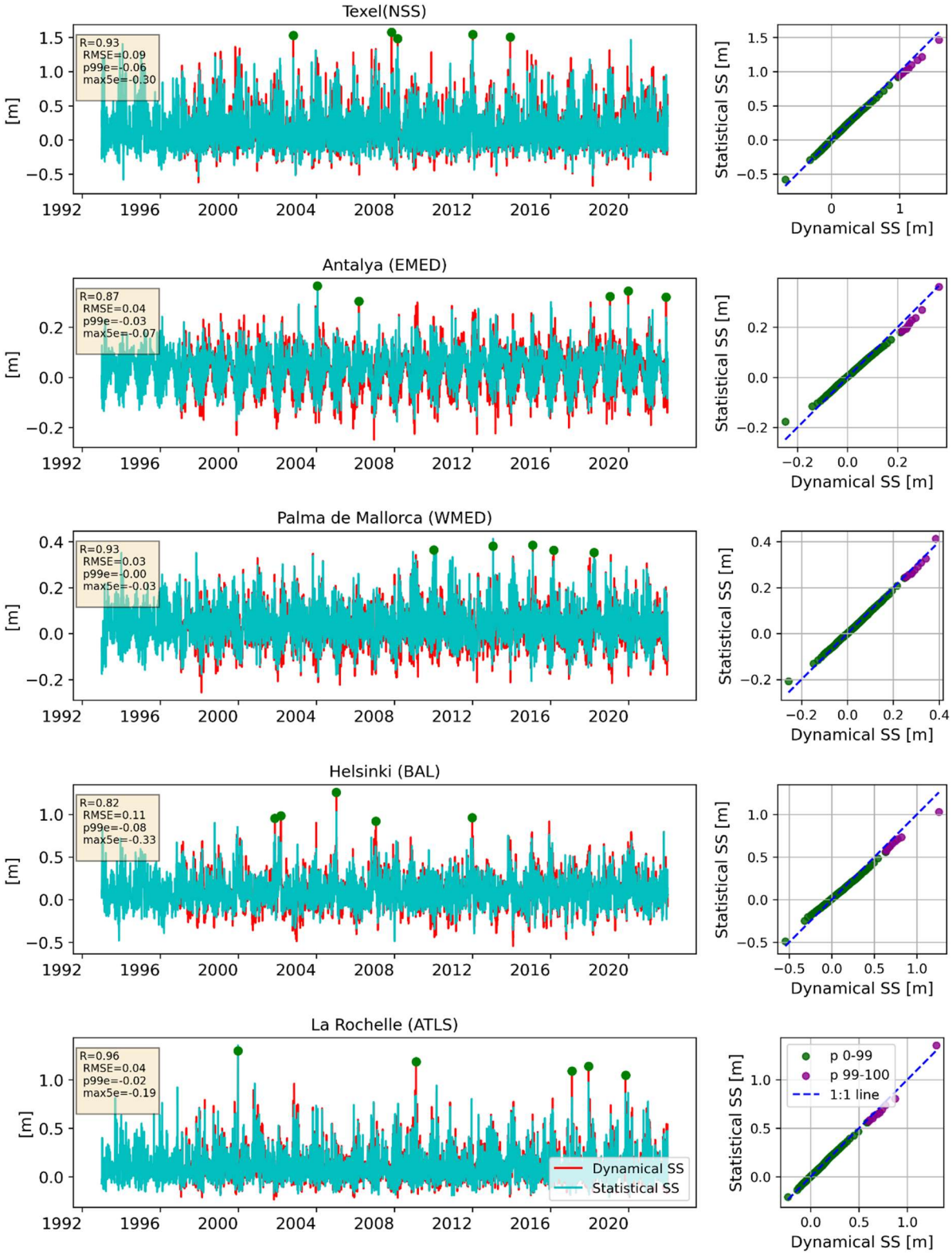
Answer to comment 1: We thank the reviewer for her useful remarks. Indeed, given that the study focuses on projections of extremes, the performance of the statistical model for extremes needs to be better demonstrated. The calibration focuses on correlation and RMSE to illustrate that the statistical 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

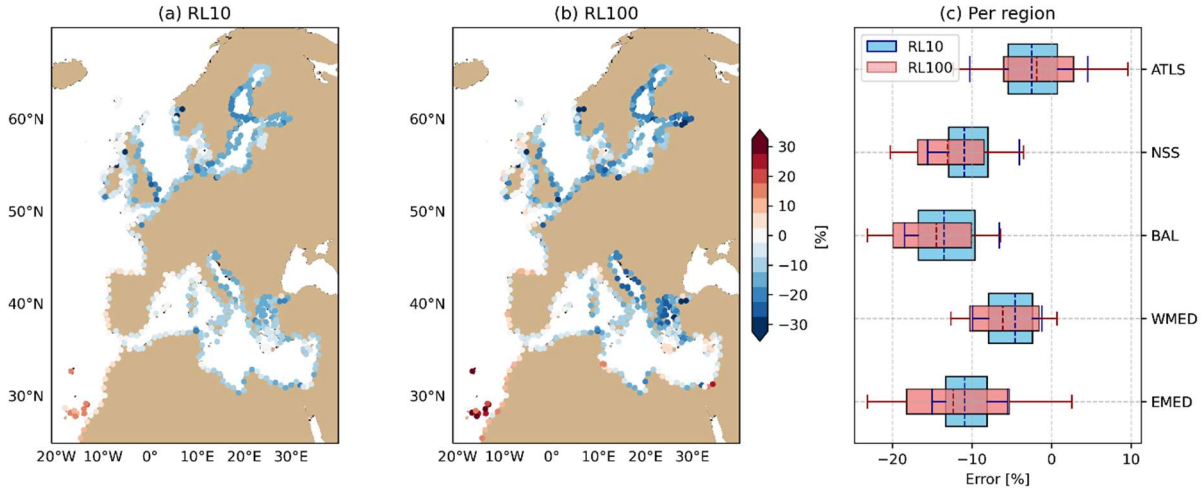
performance for high percentiles (for this subset of stations at least), other than for the station in the Baltic. For individual events marked in green, performance varies across events, some are very well captured and others are underestimated. Based on these results, we do attribute a general good 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.*

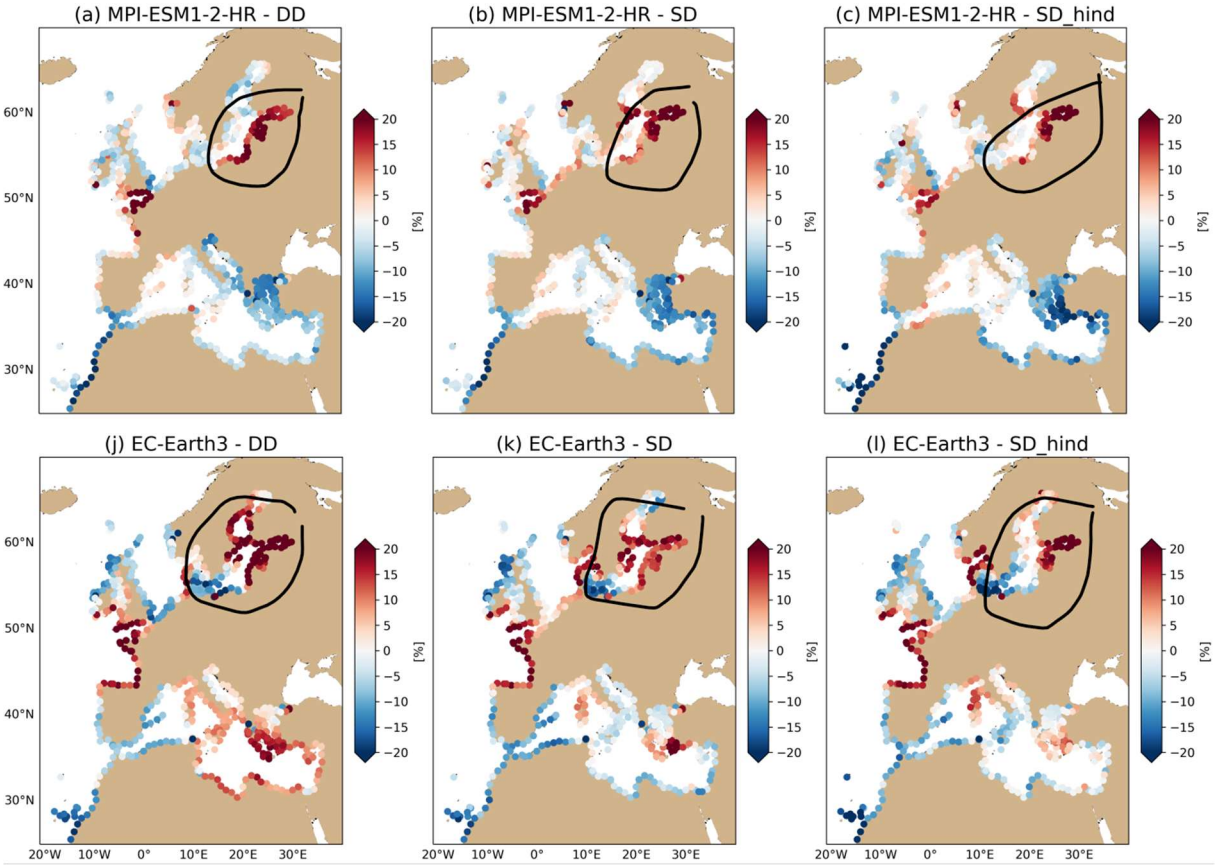
Answer to comment 2: Thank you for your insightful remark. We realize that the current description is indeed too shallow and results in Figure 6 (now Figure 8) deserve a more thorough description given the nuances observed between GCMs. Before diving into these nuances, we have better highlighted in 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*

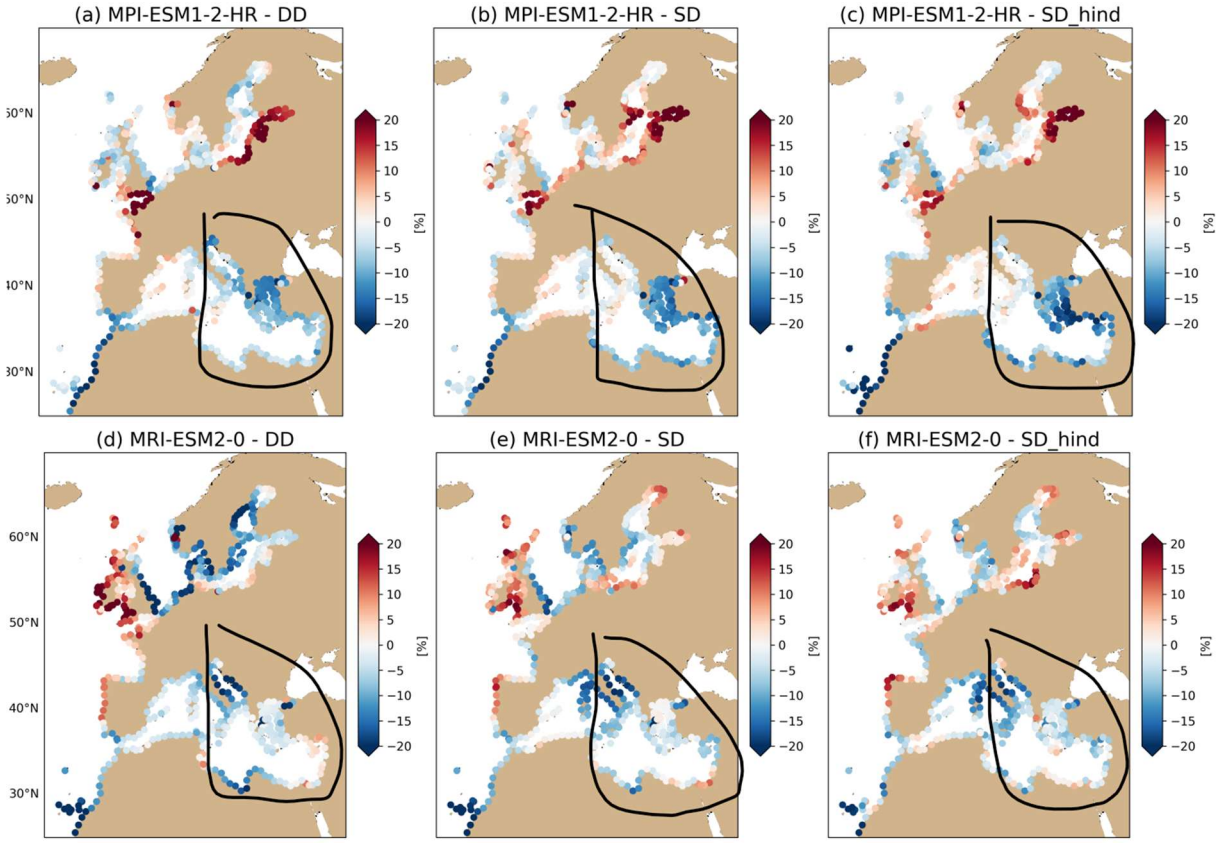
the European coastline, based on the SDM skill to reconstruct the hindcast simulation under ERA5 forcing degraded to 1°. The objective is to select a single configuration that delivers optimal performance at the European scale and can subsequently be used for European-scale ESS projections."

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.*

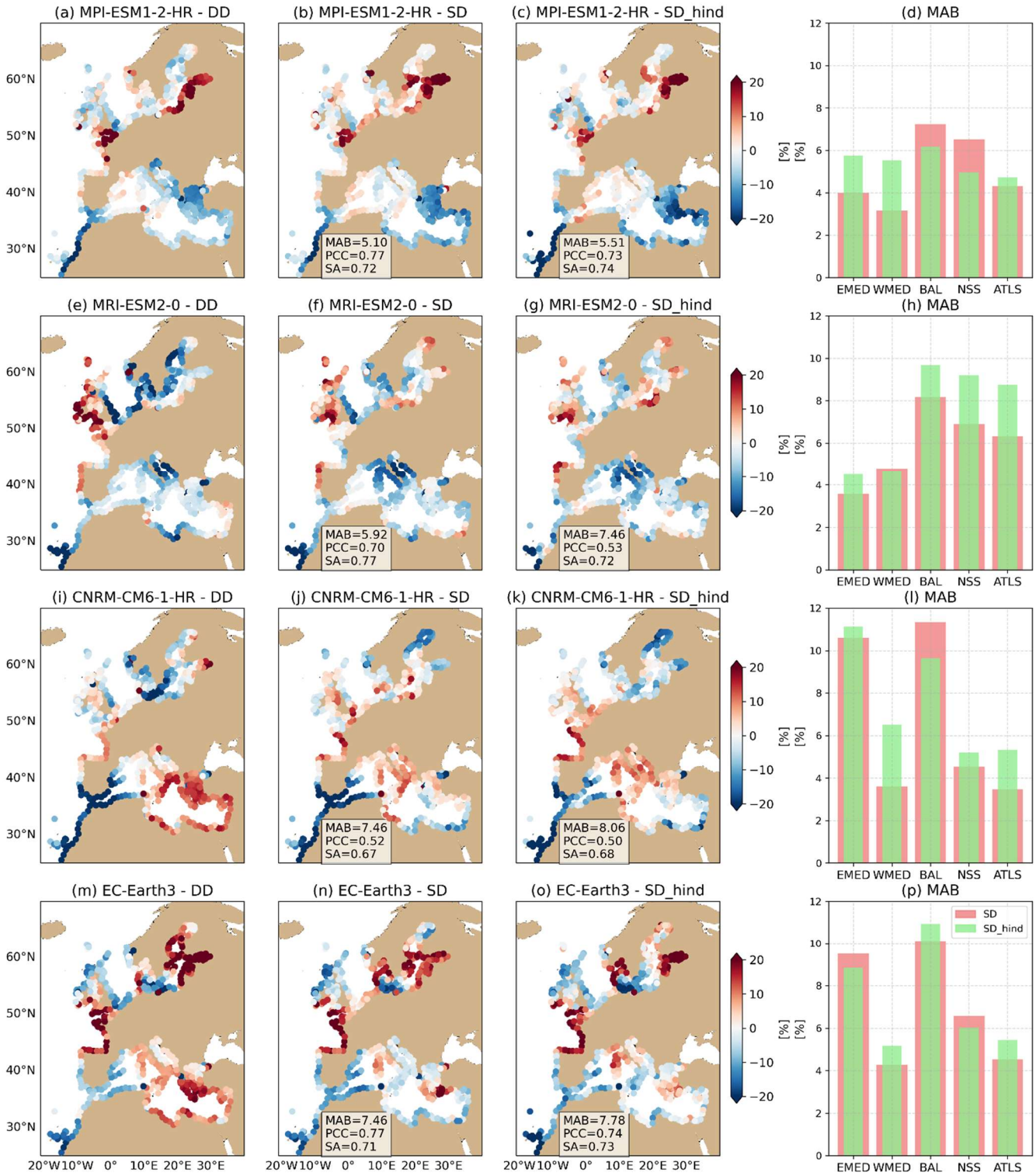


Figure 8 Projected changes (%, extreme value analysis on [2080-2099] vs [1995-2014]) in the 1 in 10 year storm surge event (RL10) for dynamical climate simulations (DD, a,e,i,m), statistical estimates trained on each historical climate simulation (SD, b,f,j,n) and statistical estimates trained on the 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.

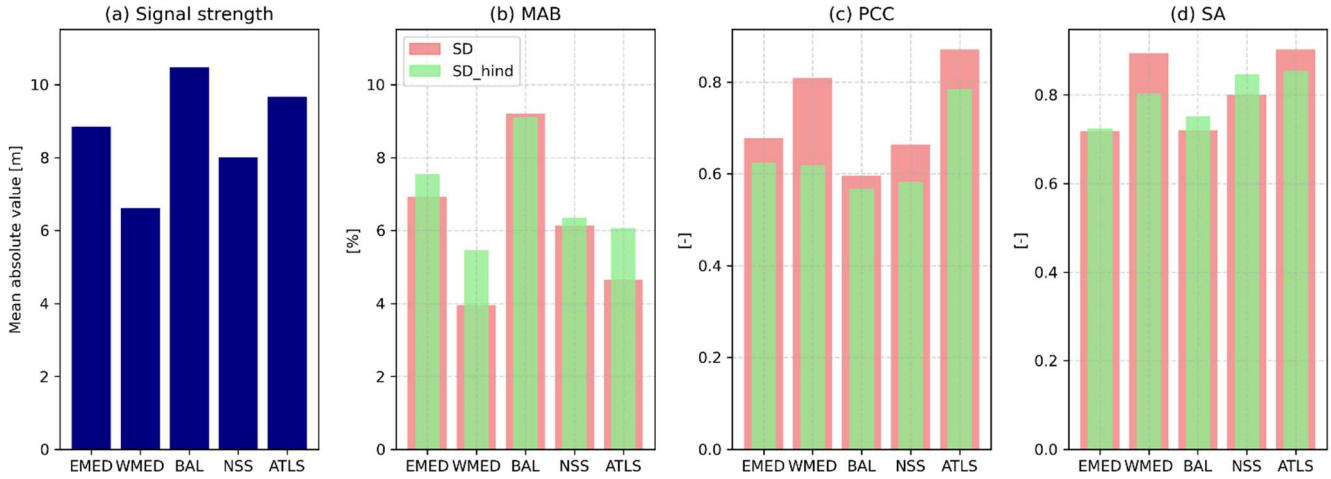


Figure 9 Regional performance metrics of statistically downscaled projections relative to dynamically downscaled projections computed by pooling projections across the 4 GCMs for each region. (a) Strength of the signal of projected changes in the 10-year storm surge return level (RL10) from dynamical 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:

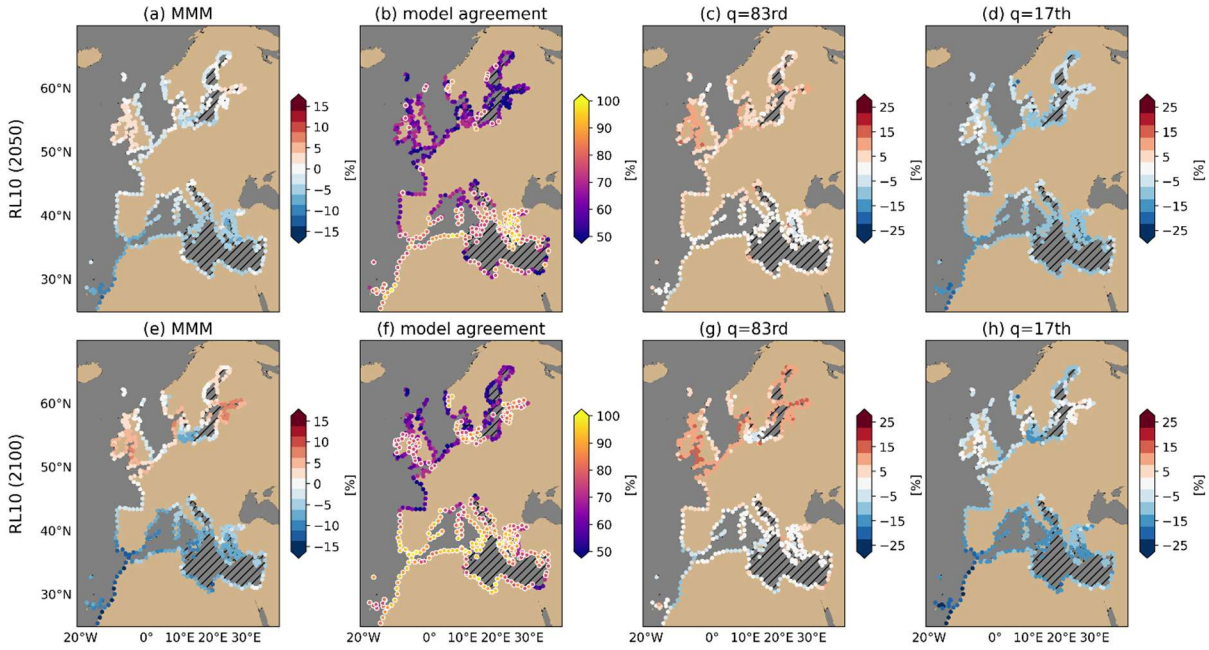


Figure 10 Muti-model mean (MMM) projected changes [%] in the 1 in 10 storm surge return level by middle (a) and end (e)-of the 21st century generated by the hindcast-trained statistical downscaling model (SDM, training period 1997-2021). Corresponding ratio of models agreeing in the sign of projected 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).