We want to thank the reviewer for the constructive comments. In the following, we state the referee's comments (in blue) followed by the response and actions taken (in black). We have also highlighted the changes in the revised manuscript, where new texts are represented using blue. The line numbers given here are also from the edited version.

1 Referee 2

Comment #1

The selection of Hurricane Irene (2011) is very appropriate due to the severity and the influence on the study site. However, did the authors consider the option of using other hurricane tracks instead of perturbing Hurricane Irene or to complement the results obtained with Hurricane Irene?

Response to comment #1

We selected the Delaware Bay and River (DBR), US, as the study site to examine the role of local estuarine wind in elevating the surge-induced coastal flooding because of the extensive model calibration and validation done for both hydrology (DHSVM) and hydrodynamic (FVCOM) models in the same region previously (using a large set of available flow gauges, tide gauges, tidal current profilers, and high water marks), described in Deb et al. (2023; https://doi.org/10.1029/2022EF002947). For this region, it has been reported that Hurricane Irene (2011) caused the most damage when both fluvial and coastal flooding are considered. Two other recent extreme events, Hurricane Isabel (2003) and Hurricane Sandy (2012), also brought large coastal surges within the bay and river; however, they made landfall far away from DBR and translated through different regions, making them less suitable cases compared to Hurricane Irene (2011).

Comment #2

Regarding the generation of the ensembles, please justify the selection of 50-member ensembles and the two times selected for the initialization (separated by 12 hours). Please, discuss the sensitivity of the results/conclusions to these parameters.

Response to comment #2

Reviewer #1 posed a very similar question. We've added text to Appendix A to clarify our ensemble setup, which we'll paraphrase here: Following the methods of Reed et al. (2020; https://doi.org/10.1126/sciadv.aaw9253), we test the sensitivity of hurricane fidelity to (a) model physics parameters and (b) model initialization time. With regard to (a), a suite of model simulations (to be explained below) are conducted by perturbing model physics parameters to which hurricanes are most sensitive, according to the hurricane parameter sensitivity study of He and Posselt (2015; https://doi.org/10.1175/JCLI-D-15-0255.1) and as used in Reed et al. (2020) and Reed et al. (2021; http://dx.doi.org/10.1175/BAMS-D-20-0160.1). With regard to (b), this is accomplished by conducting "mini" 10-member ensembles, each initialized at 12-hour increments from just before Irene's first U.S. landfall in North Carolina back to 5 days previous. We analyze the (mini) ensemble mean hurricane track and intensity errors at each initialization time to identify an optimal initialization time that attempts to maximize both simulation fidelity and forecast lead time (to allow sufficient hurricane spin-up). As 10-member ensembles yield an inadequately small sample size, we expand the mini ensemble initialized at the optimal time (00Z August 26, 2011), as well as the ensemble initialized 12 hours earlier (12Z August 25, 2011), to 50 members each. This is very similar to the approach used in Reed et al. (2020), in which a 100-member ensemble was run at a single optimal initialization time. We felt that using a second initialization time might provide a greater diversity of model solutions – though admittedly the large inland track bias in the earlier initialization (E2) was unexpected.

Comment #3

The validation of the FVCOM model (Appendix C) shows good results. However, the validation was carried out by forcing the model with ERA5 instead of E3SM outputs. Why was the FVCOM model not validated with E3SM outputs? It is suggested to include the validation of the hydrodynamic model with the forcing data used in the analysis or discuss in more detail this issue in the paper. This comparison would be also useful for understanding the uncertainty that justifies the generation of the 50-member ensembles.

Response to comment #3

This work first validated hydrology and hydrodynamic models using observation-based and reanalysis forcing to show that they can accurately represent the physical processes if we provide a more realistic forcing. Then, we used the E3SM ensembles of Hurricane Irene-like events to examine the sensitivity of coastal flooding to different tracks. In response to the reviewer's question "*Why was the FVCOM model not validated with E3SM outputs?*", we want to say that the E3SM tracks used in this study represent Irene-like events which are all physically plausible; however, they do not precisely represent the observed event. In Appendix A, we have an in-depth discussion about the time evolution of errors in along- and cross-track distances, minimum central sea-level pressure, and maximum surface wind associated with the E3SM tracks compared to the observed event. See the following description:

On page 17, line 318: *Figure A1 displays the time evolution of errors in along- and cross-track distances, minimum central sea-level pressure, and maximum surface wind associated with Hurricane Irene simulated by E3SM initialized on August 26, 2011 00Z (ensemble E1). Similar time series for ensemble E2 (not shown) indicate larger distance errors—consistent with a more westward/inland track—but similar errors in minimum central pressure and maximum surface winds, by construction. Figure A1 shows that Hurricane Irene simulated for E1 generally follows the correct trajectory (cross-track errors less than 20 km) but has a forward speed that is slower than observed (along-track errors of roughly* -50 *to* -100 *km). Further, the E1 version of Irene predicts a central pressure that is too low and surface winds that are too high, indicating an overestimation of hurricane intensity. We also use the Climate Prediction Center "Unified Gauge-Based Analysis of Daily Precipitation over CONUS" product (Chen, 2008), provided at daily resolution on a 0.25°×0.25° horizontal grid, to assess space-time averaged precipitation accumulations. Figure A2 displays the evolution of mid-Atlantic watershed-averaged (left) 3-hourly precipitation amounts and (right) cumulative sum of precipitation. Watershed-averaged precipitation intensity peaks near August 28, 2011 15Z-18Z, for both the ensemble mean and selected E1 member. E3SM exhibits a slow onset of the watershed-averaged cumulative sum of precipitation through August 28, 2011 12Z (Fig. A2, right), but later overestimates cumulative precipitation through August 29, 2011 12Z. This equates to a ∼15 mm (∼20%) underestimation of cumulative precipitation during the initial impact window but a ∼30 mm (∼33%) overestimation of storm-total precipitation. Together, Figs. A1 and A2 indicate that E3SM simulates a version of Irene that is too slow and too strong, leading to a delayed onset of precipitation in the mid-Atlantic watershed but ultimately an overestimation of storm-total precipitation.*

To show that E1 members (with correct trajectory) can provide a reasonable estimate of the peak water surface elevation (WSE) when compared to observation, we have modified Figure 2c and included the observed peak WSE during Hurricane Irene (2011). This modified Figure 2c can help understand the uncertainty in FVCOM water elevation due to varying biases in meteorological forcing.

We also added the following lines to the main text:

On page 7, line 157: *E1, which produced reasonable Irene tracks (Figure A1), also shows a fair range of WSE and an ensemble mean compared to the observed along-channel peak WSE for Hurricane Irene (2011). Here, observed peak WSE means the FVCOM model WSE generated using reanalysis forcing and validated using field datasets. At the upstream end, near NB, the peak WSE range deviates from the observed due to the influence of river discharge and the biases that propagated from the E3SM precipitation field (details provided in Appendix A).*

Comment #4

In Fig.3 and several parts of the document, the authors mention 'simulations without the effect of estuarine wind fields, remote wind fill and full wind fill'. It is not clear how these types of wind fields (estuarine, remote and will) are considered in the numerical model. For example, does a simulation without estuarine wind fields refer to a simulation with astronomical tides only or with astronomical tides and pressure (but without wind)? For a simulation with remote wind fields, how are the estuarine wind fields switched off?. Please, explain in more detail.

Response to comment #4

Reviewer #1 also asked a very similar question. We multiplied the E3SM wind velocity vectors (in m/s) by 0.1 outside/within the polygon (shown in Figures 3a,b) to make estuarine/remote wind stress-only cases, respectively. The astronomical tide and pressure field were not modified. We are paraphrasing the new lines below that are added to Section 4.

On page 10, line 184: *We used this bounding polygon to select the E3SM grid cells that fully cover the FVCOM model domain. FVCOM uses a bilinear interpolation method to assign wind velocity at the unstructured grid cells from the meteorological dataset. To represent a scenario with nominal estuarine wind during hurricane landfall, we multiplied the E3SM wind velocity vectors (in m/s) within the polygon with 0.1 to uniformly dampen the wind magnitude in the selected cells, regardless of the instantaneous location of the hurricane. For the two cases, A and B, this artificial dampening reduced the peak wind speed magnitude to* $\sim 2.0\,\mathrm{m/s}$*, making the impact of the hurricane wind field negligible. In addition, the polygon and E3SM grid cells extensively covered the FVCOM model domain, going beyond the FVCOM boundary, to better interpolate the wind forcing.*

On page 11, line 215: *As described earlier, we multiplied the E3SM wind velocity vectors (in m/s) by 0.1 outside/within the polygon (shown in Figures 3a,b) to make estuarine/remote wind stress-only cases, respectively.*

Comment #5

The results are mainly based on the comparison of cases A and B. How are the remaining Ensemble 1 and 2 simulations used in the analysis?

Response to comment #5

In this study, we used all the Ensemble 1 and 2 simulations at the beginning to see the variability in water surface elevation from different events (shown in Figure 2c). Subsequently, as we focused on examining the primary driving mechanism behind the increase in along-channel water surface elevation, especially at the mid-Bay, we selected these two cases, A and B, which represented the best combination from all members to show the role of the estuarine local wind in flood amplification. We added a new supplemental material section "Hurricane case selection: Case A from Ensemble E1 and Case B from E2", based on Reviewer #1's comment, showing the step-by-step approach we took before selecting these two tracks.

Comment #6

The study focuses on convergent estuaries, can similar results and conclusions be expected for other types of estuaries? I would suggest the authors explain this in more detail in the discussion section.

Response to comment #6

The geometry can vary widely for other estuaries (e.g., sheltered tidal lagoons or river deltas), influencing the flood amplification or attenuation rate depending on the estuary's unique characteristics. From the results of this study, it would be challenging to state the response of these systems. However, based on the existing literature [e.g., Khojasteh et al. (2021); https://doi.org/10.1371/journal.pone.0257538], we can say that the along-channel flood amplification rate will always be higher for a converging and macro-tidal system like the Delaware Bay and River, USA, than for non-converging ones. We added the following lines to the main text to clarify it more.

On page 16, line 281: *In converging estuarine systems worldwide [e.g., the Delaware Bay and River (USA), Humber Estuary (UK), Hooghly Estuary (India), the Meghna River Estuary (Bangladesh), and the Pearl River Estuary (China)] that are highly vulnerable to hurricane-induced flooding, physically consistent and integrated modeling frameworks are critical to correctly resolve this nonlinear tide-wind-surge dynamics and improve the coastal hazard projections for a future climate. In other coastal systems, such as sheltered tidal lagoons or river deltas, properly resolving the estuarine local wind using an integrated framework is essential as well; however, the interacting effect of geometry and tide-wind-surge dynamics in flood amplification will be less significant than the converging ones.*