## Reviewer Comments

The manuscript investigates whether weather-pattern-informed ensemble subselection can improve seasonal storm surge predictions at Cuxhaven in the German Bight. It combines MPI-ESM-HR seasonal forecasts with neural-network-based surge estimation and atmospheric circulation classification. The approach offers a physically interpretable basis for examining seasonal coastal hazards, and the results show modest improvements in anomaly correlations, particularly for surge-event counts. However, these improvements are not consistent across verification metrics, and probabilistic forecast performance deteriorates after subselection. The manuscript requires further clarification of the training and validation procedures, the ensemble subselection method, and the storm surge metrics. The robustness of the reported improvements and the distinction between idealized predictive potential and practical forecast skill also need closer examination. My specific comments are provided below.

**Comment #1:** The Introduction and Discussion cover circulation-based storm surge prediction and ensemble subselection, but would benefit from a stronger connection to the literature on signal-to-noise relationships, probabilistic forecast reliability, and ensemble-size effects on verification scores. These issues are directly relevant because the proposed subselection increases anomaly correlations while worsening CRPS and reducing ensemble spread. The authors should discuss these contrasting outcomes in light of prior research and more clearly explain how their approach advances existing ensemble forecasting methods. The following studies may be considered where appropriate:

- Scaife, A. A., & Smith, D. (2018). A signal-to-noise paradox in climate science. npj Climate and Atmospheric Science, 1, 28. https://doi.org/10.1038/s41612-018-0038-4.
- Ferro, C. A. T. (2014). Fair scores for ensemble forecasts. Quarterly Journal of the Royal Meteorological Society, 140, 1917–1923. https://doi.org/10.1002/qj.2270.
- Smith, L. A., Du, H., Suckling, E. B., & Niehörster, F. (2015). Probabilistic skill in ensemble seasonal forecasts. Quarterly Journal of the Royal Meteorological Society. https://doi.org/10.1002/qj.2403.
- Zengin, E. (2023). A Combined Assessment of Sea Level Rise (SLR) Effect on Antalya Gulf (Türkiye) and Future Predictions on Land Loss. Journal of the Indian Society of Remote Sensing, 51, 1121–1133. https://doi.org/10.1007/s12524-023-01694-0.
- Zengin, E. (2023). Inundation risk assessment of Eastern Mediterranean Coastal archaeological and historical sites of Türkiye and Greece. Environmental Monitoring and Assessment, 195, 968. https://doi.org/10.1007/s10661-023-11549-3.

**Comment #2:** The procedure for training and validating the surge neural network lacks clarity. Please detail the training, validation, and testing timeframes, and clarify if the forecast skill was assessed on independent data. Also, verify whether standardization and weather-pattern classification methods utilized information from the evaluation period.

**Comment #3:** The ensemble subselection procedure needs further clarification. Please explain how days with no eligible members are handled and how daily subsets with different members are combined to produce seasonal predictions. The choice of the majority threshold (>15 members) and the ±0.5 percentage-point classification threshold should also be justified.

**Comment #4:** The manuscript links the observed q98 threshold of approximately 1.49 m to the operational definition of storm surges. However, the analyzed variable is a filtered residual after removing tides and annual mean water levels, whereas the operational definition refers to water levels above mean high water. Please clarify this distinction and avoid treating these thresholds as equivalent.

**Comment #5:** The seasonal cumulative q98 surge-height metric requires a clearer physical interpretation. Summing daily q98 values across the season does not represent the height of an individual extreme event. Please explain the advantage of this metric and clarify whether ordinary days, including days without threshold exceedances, influence its seasonal variability.

**Comment #6:** The reported increases in correlation are modest. A significant correlation for the subselected ensemble does not necessarily establish a significant improvement over the full ensemble. Please assess the uncertainty in the difference between their skill scores and adjust statements about improvement accordingly.

**Comment #7:** Table 1 shows higher CRPS values after subselection for both variables, indicating poorer probabilistic performance. This should be stated consistently in the Discussion and Conclusions. Please also consider whether the reduced and varying ensemble size contributes to these differences, alongside changes in ensemble spread.

**Comment #8:** The “Best Member” is selected separately for each year using the observed outcome. Its high correlation, therefore, demonstrates that the ensemble contains values close to the observations, but does not, by itself, establish usable predictive information. Please moderate this interpretation and maintain a clear distinction between diagnostic selection and achievable forecast skill.

**Comment #9:** Subselection shows better performance from 1992 to 2023 but does not reliably enhance correlations from 1960 to 1991. This limitation should be highlighted more strongly. While improved observations and initialization might explain this, these should remain hypotheses unless further evidence supports them.

**Comment #10:** Please check consistency in dataset labels, analysis periods, and sample sizes throughout the manuscript. Tide-gauge observations and ERA5-derived data are often grouped under *“ERA5 observations”* making the verification reference unclear. The use of 1960–2023 versus 1961–2023, and references to 63 observations versus results apparently based on 64 seasons, should also be reconciled.