2nd Reviewer:

GENERAL COMMENTS

This study presents the results of applying the HIDRA2 deep learning model to forecast sea level along the Estonian coast of the Baltic Sea. Similar to its application in the Adriatic Sea, for which HIDRA2 was originally developed, the data-driven model generally outperforms dynamical models (3D ocean models), except for extreme events. This is expected, as extreme events are inherently rare and, therefore, challenging to accurately represent using a data-driven approach.

While the manuscript is clear and well-written, it does not significantly advance the understanding of sea surface height (SSH) forecasting or machine learning (ML)-based methods for this purpose. However, it provides a valuable report on a state-of-the-art ML-based system capable of producing fast and computationally “cheap” SSH forecasts for selected locations within the Baltic Sea.

I strongly encourage the authors to reduce the length of certain sections, particularly in the discussion and conclusion (which often reads like a summary), where some paragraphs are repetitive or restate well-known concepts—such as the efficiency and computational cost-effectiveness of ML methods compared to 3D ocean models based on primitive equations. Instead, I suggest expanding on the ensemble approach, assessing its limitations, and exploring potential strategies to improve the representativenes of the ensemble spread.

We thank the reviewer for encouraging and constructive remarks. We will follow their suggestion and significantly shortened suggested sections.

Other comments:

Line 89: Brackets should be only around the year: “Details of the encoding architecture are presented in (Rus et al. 2023).”

Thank you, corrected.

Lines 98-99: “The original meteorological data, with a domain size of 40 × 50, were subsampled to a 9 × 12 grid.”

Subsampling appears to discard valuable information. Have the authors attempted to use the full resolution? Do you anticipate any improvements by retaining the original grid size?

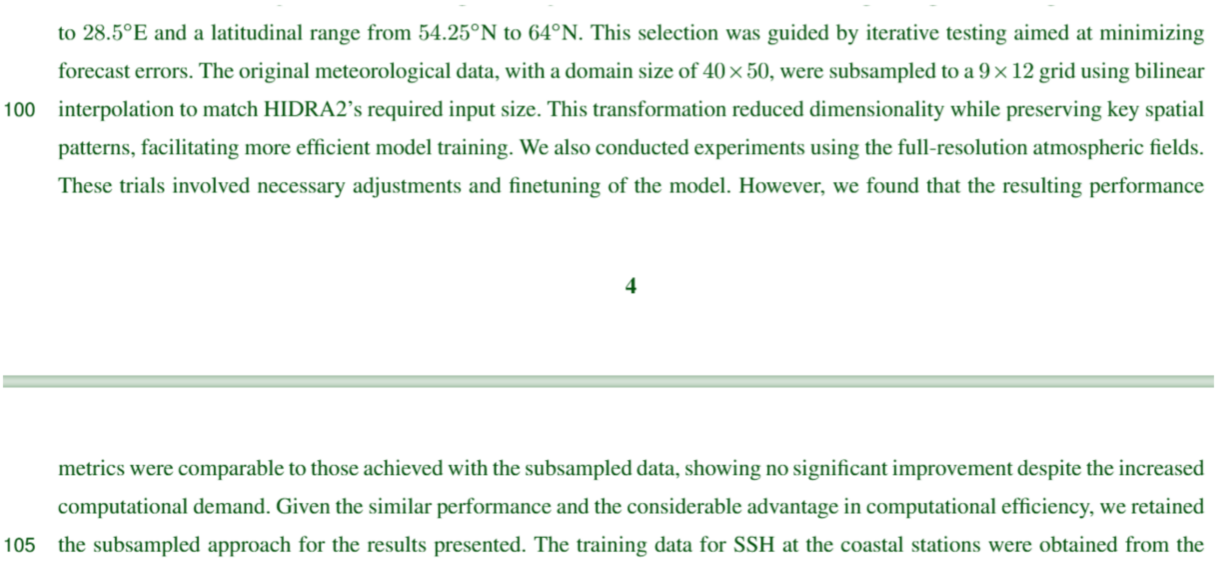
The reviewer is correct to point this out. This does merit further clarification. We did indeed conduct experiments using the full-resolution atmospheric fields. These trials involved necessary adjustments and finetuning of the model. However, we found that the resulting performance metrics were comparable to those achieved with the subsampled data, showing no significant improvement despite the increased computational demand. Based on these empirical results, while the intuition that higher resolution might hold more information is valid, our current model configuration did not seem to benefit from it for this specific task. Given the similar performance and the considerable advantage in computational efficiency, we retained the subsampled approach for the results presented.

In our implementation, the original 40 × 50 meteorological fields were not simply subsampled but downscaled to a 9 × 12 grid using **bilinear interpolation** (PyTorch’s => Resize function). This approach retains the large-scale spatial patterns while significantly reducing the input dimensionality, allowing for efficient model training and inference.

We did experiment with retaining the full-resolution input during early testing. However, the increased computational cost did not yield meaningful improvements in predictive performance, particularly in metrics such as RMSD and correlation. Given this trade-off, the interpolated grid was selected to balance model complexity and skill.

We have clarified this in the revised manuscript to indicate that bilinear interpolation was used and that the grid reduction was a design choice informed by iterative testing. Also, we acknowledge that the manuscript previously used the term **“linear interpolation”**, which has now been corrected to **“bilinear interpolation”** to accurately reflect the 2D nature of the operation applied to the spatial grid.

we have revised the manuscript as follows:



to 28.5°E and a latitudinal range from 54.25°N to 64°N. This selection was guided by iterative testing aimed at minimizing forecast errors. The original meteorological data, with a domain size of 40 × 50, were subsampled to a 9 × 12 grid using bilinear interpolation to match HIDRA2’s required input size. This transformation reduced dimensionality while preserving key spatial patterns, facilitating more efficient model training. We also conducted experiments using the full-resolution atmospheric fields. These trials involved necessary adjustments and finetuning of the model. However, we found that the resulting performance metrics were comparable to those achieved with the subsampled data, showing no significant improvement despite the increased computational demand. Given the similar performance and the considerable advantage in computational efficiency, we retained the subsampled approach for the results presented. The training data for SSH at the coastal stations were obtained from the

Figure 2: It is misleading to depict the model architecture using the Adriatic Sea instead of the Baltic Sea. If this figure is sourced from another article, please provide the appropriate citation. If it was created specifically for this study, consider replacing the Adriatic Sea with the Baltic Sea to avoid confusion.

We agree. The Figure was now changed to feature the Baltic domain.

Figure 6: This figure is not particularly informative, as most curves overlap, except for Feb-Mar 2024 in Haapsalu. Consider moving it to the supplementary material, as the key point is already well illustrated in Figure 7.

suggestion accepted. Figure 6 was moved to the supplementary material.

Line 190: “One might argue that this smoothing stems from the fact that we are working with the ensemble mean”. Please include a figure like Figure 7, but for both the best and worst ensemble members (perhaps based on RMSE). This will help illustrate the smoothing effect of the ensemble mean more effectively.

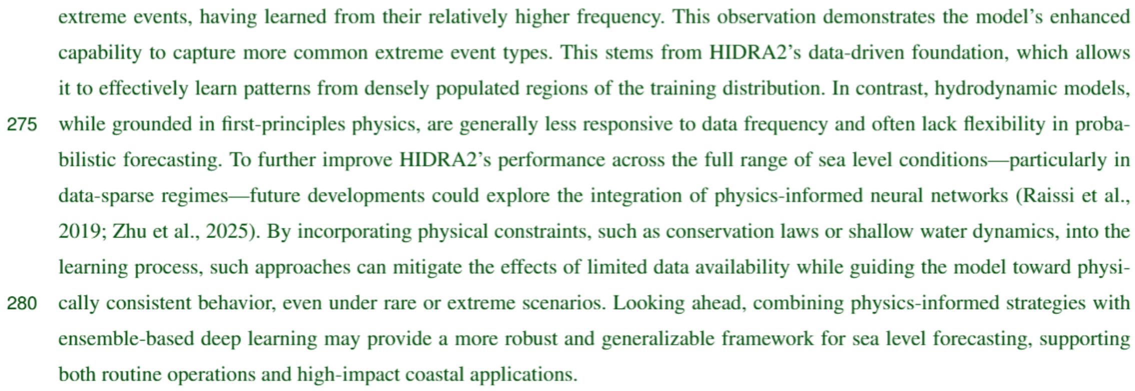
It should be noted that, in principle, no individual ensemble member performs consistently better or worse than the others. In the ECMWF Ensemble Prediction System (EPS), all members are reinitialized at each forecast cycle. This means, for example, that the 7th member in today’s forecast is unrelated to the 7th member from yesterday or tomorrow. There is no continuity between members over time—each one is statistically equivalent in terms of performance. Therefore, plotting the best and worst members across multiple cycles would result in a large number of discontinuous curves, which would reduce the clarity of the figure. We hope the reviewer agrees that such a representation would not be helpful. Nevertheless, the following clarification has been added to the Figure caption: “Grey lines in the background denote raw spectra, while black lines denote their 6-hour moving average.”

Lines 230-250: The discussion in this section largely reiterates well-established points about HYDRA2 without adding new insights. Writing a full paragraph to restate that ML-based methods are significantly more computationally efficient than traditional dynamical models seem unnecessary, as this fact is widely recognized, even by non-specialists.

This section has been significantly shortened and mentioned paragraphs were removed in the interest of brevity and non-repetitiveness.

Line 283: This observation demonstrates the model’s enhanced capability to capture more common extreme event types. This idea should be discussed also for prediction of normal vs extremes (rare) SSH conditions, where HIDRA2 generally outperforms dynamical models.

In the revised manuscript the discussion has been extended as follow:



extreme events, having learned from their relatively higher frequency. This observation demonstrates the model’s enhanced capability to capture more common extreme event types. This stems from HIDRA2’s data-driven foundation, which allows it to effectively learn patterns from densely populated regions of the training distribution. In contrast, hydrodynamic models, while grounded in first-principles physics, are generally less responsive to data frequency and often lack flexibility in probabilistic forecasting. To further improve HIDRA2’s performance across the full range of sea level conditions—particularly in data-sparse regimes—future developments could explore the integration of physics-informed neural networks (Raissi et al., 2019; Zhu et al., 2025). By incorporating physical constraints, such as conservation laws or shallow water dynamics, into the learning process, such approaches can mitigate the effects of limited data availability while guiding the model toward physically consistent behavior, even under rare or extreme scenarios. Looking ahead, combining physics-informed strategies with ensemble-based deep learning may provide a more robust and generalizable framework for sea level forecasting, supporting both routine operations and high-impact coastal applications.

Line 286-292: This section reads more like a summary rather than a conclusion and should be revised accordingly.

This section was thoroughly shortened and substantially rewritten along the lines of the reviewer's suggestions.

Line 305: I do not find this section informative. Deep-learning ensemble models, such as HIDRA2, are pertinent for advancing the development of Digital Twins and associated impact models (Li et al., 2023). These models, utilizing ensemble-based techniques, are particularly effective in capturing the complex, non-linear relationships in SSH data across diverse scales. By integrating multiple predictive models, ensemble approaches enhance the accuracy and robustness of forecasts, making them valuable for the creation of Digital Twins of the Earth systems. This, in turn, supports more precise impact assessments and decision-making processes in coastal management and risk mitigation.

We agree. This section was removed from the paper.