Review of "Deep-learning prediction of high-frequency sea-level oscillations in the Adriatic Sea" by Međugorac et al.

General assessment

This manuscript presents a rigorous and carefully constructed evaluation of deep-learning architectures for predicting high-frequency (T < 1 h) sea-level oscillations at two Adriatic tide-gauge stations. The topic is of considerable practical importance: meteotsunamis and related high-frequency oscillations are locally destructive, poorly served by existing operational infrastructure in the Adriatic (as the authors note, no hydrodynamic HFO forecasting system is currently running in real time), and have historically resisted skillful forecasting because the atmospheric disturbances that generate them operate at spatial and temporal scales below the resolution of both routine numerical weather prediction and standard reanalysis products. Given this context, the manuscript's central contribution is a systematic test of whether a data-driven architecture, without explicit physical parameterisation of resonance mechanisms, can extract usable predictive signal from coarser large-scale atmospheric fields. This is a meaningful and well-motivated scientific question. The authors treat both the successes and the failures of their approach with appropriate rigor, and I found the manuscript's honesty about the limits of extreme-event predictability to be one of its most valuable features. I recommend acceptance subject to minor revisions.

Assessment of contribution and significance

The comparative design (HFNet vs. HFNet_JE) is a strength: rather than optimising toward a single headline metric, the authors demonstrate a real architectural trade-off between skill on typical events and skill on rare, high-amplitude events, and they support this with three independent lines of evidence (station-level results, ablation experiments, and forecast-mode ECMWF ensemble evaluation). This is a more scientifically defensible contribution than a single "best model" claim would have been, and it will be of direct use to groups considering similar architectures for other coastal hazard applications, including storm surge and compound flooding forecasting, marine heatwaves etc. where the same data-scarcity problem for extremes recurs.

The finding that higher-resolution CERRA input does not translate into better extreme-event forecasts, while ERA5's higher temporal resolution does, is a substantive and somewhat counterintuitive result. It is well supported by the data presented and is discussed with appropriate physical reasoning (Proudman resonance sensitivity to the temporal evolution of the pressure disturbance rather than purely its spatial structure).

Substantive comments

1. The apparent lack of added value from CERRA's finer spatial resolution for extreme events may partly reflect a double-penalty/representativeness effect rather than a genuine absence of useful information. CERRA’s sharper small-scale structures could be correctly capturing the disturbance but scoring worse under pointwise verification if positional/timing offsets exist, an effect well documented in high-resolution NWP verification (e.g., Ebert, 2008; Roberts and Lean, 2008 , Kozyrakis, 2023: “timing errors in the synchronization between the model and the actual flow are likely to be magnified and subsequently penalized in terms of RMSE”). It would

strengthen the paper's conclusion on CERRA's usefulness to adapt the tests with a neighborhood-based or displacement-tolerant metric (e.g., FSS) alongside the current pointwise RMSE/MAE. Since the DL model's output is a pointwise scalar (daily max HFO-A) rather than a spatial field, standard neighborhood verification metric (e.g., FSS) cannot be applied directly to the model output. However, the underlying representativeness concern may still be relevant at the input stage: CERRA's finer spatial/temporal grid could represent the same atmospheric disturbance with small positional or timing offsets relative to ERA5, which given the network's convolutional architecture and the scarcity of extreme-event training examples may make the extra information harder to exploit rather than simply unhelpful. Two experiments could help disentangle this: (i) if available data: independently verifying ERA5 and CERRA surface fields against point atmospheric observations (e.g., coastal microbarograph or wind stations) using feature-based methods, to establish whether CERRA more faithfully represents the disturbance itself; and (ii) spatially/temporally coarsening CERRA to ERA5's resolution and retraining, to separate the effect of resolution from other differences between the two reanalysis products (assimilation system, temporal frequency). This is left up to the authors to decide if they'll add at this stage as it is substantial work, or left for a future work, but could significantly enhance and complement this work, answering more questions in the downscaling experiment.

Separately, applying extreme-preserving quantile-based bias correction (e.g., QDM, Cannon et al., 2015) to CERRA prior to model training could help disentangle whether CERRA's underperformance for extremes reflects a correctable systematic bias or an intrinsic representativeness limitation.

Ebert, E.E. (2008), *"Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework,"* Meteorol. Appl.

Roberts, N.M. and Lean, H.W. (2008), *"Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events,"* Mon. Wea. Rev. — introduces the Fractions Skill Score (FSS), designed specifically to reward high-res fields for getting the right structure near the right place without the double penalty.

Cannon, A.J., Sobie, S.R., and Murdock, T.Q. (2015), *"Bias correction of GCM precipitation by quantile mapping: how well do methods preserve changes in quantiles and extremes?"* J. Climate

Kozyrakis, G.V.; Condaxakis, C.; Parasyris, A.; Kampanis, N.A. Wind Resource Assessment over the Hellenic Seas Using Dynamical Downscaling Techniques and Meteorological Station Observations. *Energies* 2023, *16*, 5965. https://doi.org/10.3390/en16165965

2. The manuscript would benefit from situating its motivation more explicitly within the broader shift in the ocean/atmosphere forecasting community toward machine learning for the prediction of rare, high-impact extremes. Including some literature on AI predicted storm surges, marine heatwaves, and other events for which traditional deterministic and statistical

approaches have proven both difficult to make skillful and computationally expensive to run operationally, will increase the impact of the current study. The authors touch on this implicitly (e.g., the comparison of HFNet's training cost to AdriSC's operational cost, and the general framing of HIDRA as a low-cost alternative to deterministic ocean modelling), but the paper does not explicitly connect its own contribution to this wider literature-review context. Given that the paper's own conclusions (Section 6, points 2 and 7) center on the same tension between typical-case skill and extreme-case predictability that motivates much of the ML-for-extremes literature more broadly, an explicit paragraph in the Introduction acknowledging this trend and framing the present study as a contribution to it would strengthen the paper's positioning and give readers outside the immediate sea-level community a clearer sense of why this approach is being pursued, both for the sake of raw predictive skill and for the practical benefit of fast, low-cost inference relative to dynamical ensemble approaches such as AdriSC or the stochastic surrogate of Denamiel et al. (2019b).

3. Given the operational stakes implied by this work (an eventual early-warning capability for the eastern Adriatic coast), I would encourage the authors to comment more directly, even if briefly, on what the reported skill levels for extreme events (Tables 2–5) would mean in a genuine early-warning context e.g., whether the observed negative bias and underestimation of extremes would be acceptable for a warning system tuned toward high recall at the cost of false alarms, or whether the current skill level is better characterised as a research benchmark than an operationally deployable capability. This distinction is implicit in the Discussion but would benefit from being stated explicitly, since it affects how readers outside the immediate modelling community should interpret the practical readiness of the approach.

4. The claim that the HFNet_JE batch-size constraint (32 vs. 128 for HFNet) does not confound the architectural comparison is asserted rather than demonstrated. Given that batch size is known to interact with generalisation behaviour in deep networks, and given that the paper's central finding rests on a comparison between these two architectures, I would ask the authors to either provide evidence that the comparison is robust to this asymmetry (e.g., a supplementary run of HFNet_JE at a larger effective batch size via gradient accumulation) or to state more explicitly, as a limitation, that this asymmetry cannot be fully disentangled from the architectural differences being tested.

5. Related to point 3: the "best validation loss" selection criterion (smallest maximum validation loss, i.e. tuned toward extreme-event performance) is a reasonable and well-justified choice, but it does mean that the reported metrics for typical/regular events are not the models' best achievable typical-event performance. This is worth restating plainly near Table 2, since a reader who does not carefully track Section 3.4 could otherwise interpret the "Reg" columns as representing each model's optimum for that regime.

Recommendation: Accept, subject to minor revisions and added discussion and check with fuzzy metrics, as outlined above.