Authors' reply to Antonios Parasyris's comments on the manuscript https://doi.org/10.5194/egusphere-2026-3411

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).

Dear Reviewer,

Thank you for your constructive comments and the time you dedicated to reviewing our manuscript. Our responses to your comments are provided in blue text below. The indicated locations of the added or revised text refer to the manuscript (MS) with tracked changes.

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

The suggested analysis under point (i), comparing ERA5 and CERRA against point atmospheric observations, would, as the reviewer notes, require substantial additional work and is beyond the scope of the present study. Nevertheless, in the revised manuscript we commented on the possible role of representativeness and spatial/temporal offsets, taking into account the reviewer's comment and the suggested literature (lines ~635; please see below).

630 resolutions. As a result, atmospheric forcing for meteotsunami studies is often simplified using idealis~~z~~ed pressure disturbances, highlighting the difficulty of fully representing the relevant processes in operational or research-grade atmospheric models. In this context, cases in which CERRA-based predictions perform worse than ERA5-based predictions do not necessarily mean that the higher-spatial-resolution atmospheric fields of CERRA contain less relevant information. If present, small spatial or temporal offsets in the atmospheric disturbances represented by CERRA could make this
635 information more difficult for the convolutional models to use, thereby affecting HFO-A prediction errors. Direct verification of the atmospheric fields against high-frequency observations, using methods that account for spatial and temporal displacement (e.g., Ebert, 2008; Kozyrakis et al., 2023), would be needed to determine whether this contributes to the differences between ERA5- and CERRA-based predictions.

In this light, our hypothesis was that, although ~~therefore not that~~ the input fields do not fully describe the relevant processes,

To strengthen our conclusions regarding the usefulness of CERRA, we performed the additional experiment suggested under point (ii) by interpolating CERRA to the ERA5 resolution (3 hours to 1 hour, and 5.5 km to ~27 km) and retraining the models for both stations. The results of these experiments are discussed in the revised manuscript (Results, lines 399-412; please see below). The corresponding numerical results are provided below in Tables RC1 and RC2 for completeness; they are not included as tables in the MS, but the main findings are discussed in the text.

Table RC1: Table 2 from the MS (Bakar station) updated with results from experiments using CERRA data interpolated to the ERA5 grid, resulting in an hourly temporal resolution and a spatial resolution of

~27 km (red). Underlined values denote optimal values within each category (All, Reg, and Ext)

Model	Amplitude type	Events	RMSE (cm)	MAE (cm)	Relative error (%)	Bias (cm)	Accuracy (%)
HFNet	All	730	3.15	2.22	41.87	0.90	89.95
			3.021	2.05	37.66	0.13	91.00
$HFNet_{JE}$			2.57	1.54	24.14	-0.35	93.97
			2.895	1.75	26.89	-0.54	92.37
HFNet	Reg	715	2.88	2.11	42.19	1.04	89.93
			2.60	1.90	37.76	0.31	92.35
$HFNet_{JE}$			2.07	1.39	23.96	-0.22	95.37
			2.31	1.56	26.66	0.35	93.98
HFNet	Ext	15	9.12	7.46	26.53	-5.72	42.22
			11.03	9.26	32.92	-8.28	26.67
$HFNet_{JE}$			10.87	8.95	32.64	-6.37	28.89
			12.38	10.46	37.88	-9.53	15.56

Table RC2: Table 4 from the MS (Ploče station) updated with results from experiments using CERRA data interpolated to the ERA5 grid, resulting in an hourly temporal resolution and a spatial resolution of ~27 km (red). Underlined values denote optimal values within each category (All, Reg, and Ext)

Model	Amplitude type	Events	RMSE (cm)	MAE (cm)	Relative error (%)	Bias (cm)	Accuracy (%)
HFNet	All	727	2.67	2.12	51.23	1.25	78.56
			2.50	1.94	46.13	0.85	80.56
$HFNet_{JE}$			2.22	1.45	29.31	0.02	89.45
			2.21	1.44	29.91	-0.01	89.91
HFNet	Reg	716	2.64	2.09	51.72	1.31	78.96
			2.42	1.89	46.48	0.91	81.15
$HFNet_{JE}$			2.05	1.38	29.28	0.07	90.36
			2.04	1.37	29.90	0.07	90.83
HFNet	Ext	11	4.87	3.75	18.95	-2.99	54.54
			5.55	4.51	23.27	-3.62	42.42
$HFNet_{JE}$			7.27	6.02	31.64	-3.16	30.30
			7.19	5.84	30.09	-4.85	30.30

Comparison of results obtained using ERA5 and CERRA input data, when performance is averaged across~~independent of~~ model architectures, shows relatively small differences for all and regular events, with slightly better scores obtained when CERRA input is used. In contrast, differences are larger for extreme events, for which better scores are obtained with ERA5 input across all metrics. These~~is~~ differences may be related to the different spatial and temporal ~~suggests that CERRA, with its higher spatial~~ resolutions of the two reanalyses, with CERRA having a finer spatial resolution (5.5 km versus ~27 km)~~,~~

but coarser temporal resolution ~~performs somewhat better for regular events, whereas ERA5, with its higher temporal but coarser spatial resolution (1 hour versus~~ (3 hours versus 1 hour)~~. , performs better for forecasting extremes~~.To assess whether differences in model performance obtained with ERA5 and CERRA could be related to their different spatial and temporal grids, an additional set of experiments was performed using CERRA data interpolated to the ERA5 spatial (~27 km) and temporal (1 hour) grid. Both HFNet and HFNet$_{IE}$ were retrained using these data following the same training and evaluation procedure as in the original experiments. At Bakar, when performance was averaged across model architectures, models using ERA5 and interpolated CERRA showed very similar results for all events, with neither input dataset yielding consistently better model performance across the metrics. For regular events, models using interpolated CERRA generally performed slightly better than those using ERA5. In contrast, for extreme events, models using ERA5 consistently outperformed those using interpolated CERRA across all performance metrics. This indicates that the better performance of models using ERA5 input for extreme events at Bakar cannot be explained by differences in the spatial and temporal grids alone, suggesting that other differences between the two reanalyses may also influence model performance.

Regarding CERRA bias correction, applying an extreme-preserving quantile-based correction is not straightforward in our case, as our models use multiple atmospheric variables at several vertical levels. A consistent correction would need to preserve vertical dependencies, relationships between variables, and the physical consistency of the atmospheric fields. In the literature, we found assessments of CERRA surface and near-surface variables against observations (e.g., Patra et al., 2025; Pelosi, 2023), but no comparable bias-correction approach for multivariable, multilevel CERRA atmospheric data of the type used in our models. Developing and validating such a correction would also require suitable observations throughout the atmospheric column and would constitute a substantial methodological study on its own. And last but not least, we expect our network to be learnable enough to learn any systematic bias in the input data automatically. We therefore do not consider such a bias-correction analysis feasible within the framework of the present study.

Patra, A., Oueslati, B., Chevallier, T., Renaud, P., Kervella, Y., and Dubus, L.: Evaluation of ERA5, COSMO-REA6 and CERRA in simulating wind speed along the French coastline for wind energy applications, Advances in Science and Research, 22, 69–85, https://doi.org/10.5194/asr-22-69-2025, 2025.

Pelosi, A.: Performance of the Copernicus European Regional Reanalysis (CERRA) dataset as proxy of ground-based agrometeorological data, Agricultural Water Management, 289, 108556, https://doi.org/10.1016/j.agwat.2023.108556, 2023.

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).

Yes, we agree, and we added a focused paragraph to the Introduction (lines 91–103) reviewing the existing literature on machine-learning-based predictions in meteorology and oceanography, with particular emphasis on rare and extreme events, while avoiding a broader review of AI applications.

remains computationally demanding, limiting its applicability in real-time forecasting.

In recent years, machine-learning (ML) approaches have increasingly been explored for forecasting rare, high-impact atmospheric and ocean events, motivated in part by the computational cost of high-resolution dynamical modelling and ensemble prediction. Neural-network and surrogate approaches have been developed for storm surges and extreme coastal sea levels, providing computationally efficient predictions and, in some cases, performance approaching that of hydrodynamic models (Daramola et al., 2025; Hermans et al., 2025; Ramos-Valle et al., 2021). Similar approaches have been applied to marine heatwaves (Giamalaki et al., 2022), extreme precipitation (Zhang et al., 2023; Ravuri et al., 2021), and heatwaves (Miloshevich et al., 2023), while global ML weather models have demonstrated competitive skill at very low inference cost, including in forecasting tropical cyclones and, more recently, surface extremes such as extreme temperatures and winds (Price et al., 2025; Bi et al., 2023; Lam et al., 2023). A recurring challenge in ML-based prediction of extremes is their sparse representation in training data, which can lead to reduced predictive skill and underestimation of the most extreme events (e.g., Hermans et al., 2025; Miloshevich et al., 2023). The prediction of HFO extremes considered in this study therefore represents a particularly data-limited instance of this broader effort to develop computationally efficient ML methods for forecasting rare ocean hazards.

Within the more specific context of SL forecasting, the HIDRA family of models (Rus et al., 2026; 2025; 2023; Žust et al., 2021) is particularly relevant to the present study. Developed and extensively evaluated in the Adriatic Sea, HIDRA provides a general deep-learning (DL) architecture for predicting sea-level variability. ~~In parallel, data-driven approaches have been increasingly applied to SL forecasting. Particularly relevant is the HIDRA family of models (Rus et al., 2026; 2025; 2023; Žust et al., 2021), a general deep-learning (DL) forecasting architecture thoroughly evaluated in the Adriatic Sea.~~ The architecture has since been implemented along the Estonian coast (Barzandeh et al., 2025) and tested along the Spanish,

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.

The question of reliable Early Warning Systems (EWS) is a crucial question for the interaction with Civil Rescue and other response services. One of the goals of this research was to assess whether DL methods can be used to potentially upgrade meteotsunami warning systems. However, weighting between higher recall and higher false alarm rate varies geographically and with the amplitude of the event. For example, from our experience and interactions with Civil Rescue services in the northern Adriatic, it seems that tuning for high recall at the risk of having more false alarms is not an optimal trade-off. For example, a model which predicts an HFO every day would have a 100% recall at the cost of daily false alarms. It

seems that the Civil Rescue units prefer somewhat lower recall to higher number of false alarms. On other hand, after disastrous floods in Spain, the responders in Spain are now indicating (personal communication) that at least for high amplitude events, they prefer a false alarm to not predicting the event. So there is no universal solution for when a modeling setup is ready as an EWS. Therefore at this point, we decided not to tune for higher recall. At the same time, we are of the opinion that the network is not good enough for an EWS at this moment. The EWS - on the Mediterranean scale as we are currently testing this model on Balearic Islands - however remains our priority for the near future work.

This was commented on in the revised Discussion (lines 650-653; please see the details below).

processes. This indicates that the predictability of HFOs is fundamentally constrained by the information content of the atmospheric fields, and that further improvements may require either higher-resolution inputs or additional sources of
650 information. The limited sensitivity to architectural changes and to the choice of ERA5 or CERRA input further supports this interpretation. At present, the level of performance achieved with the available inputs is not sufficient for reliable implementation of HFO early warning systems and associated downstream services. The current skill level is therefore better characterised as a research benchmark than an operationally deployable early warning system.

~~IAccordingly, i~~mproving HFO predictability likely requires ~~augmenting atmospheric inputs with additional sources of~~

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.

Thank you for pointing this out. To assess the potential influence of batch size, we performed an additional $HFNet_{JE}$ experiment using gradient accumulation to achieve an effective batch size of 128. The results, provided here in Table RC3, indicate only a minor positive influence on overall performance, but a poorer performance was found for extreme events. These findings and their implications are briefly discussed in the revised MS (Methods, lines 298-303; Results, line ~425; please see below).

Table RC3: Table 2 from the MS, updated with metrics from additional $HFNet_{JE}$ experiments (red text) using an effective batch size of 128 (achieved through gradient accumulation). Underlined values denote optimal values within each category (All, Reg, and Ext)

Model	Amplitude type	Events	RMSE (cm)	MAE (cm)	Relative error (%)	Bias (cm)	Accuracy (%)
HFNet	All	730	3.15	2.22	41.87	0.90	89.95
$HFNet_{JE}$			2.57 2.58	1.54 1.52	24.14 22.56	-0.35 -0.62	93.97 93.63
HFNet	Reg	715	2.88	2.11	42.19	1.04	89.9
$HFNet_{JE}$			2.07 2.04	1.39 1.36	23.96 22.32	-0.22 -0.50	95.37 95.04
HFNet	Ext	15	9.12	7.46	26.53	-5.72	42.22
$HFNet_{JE}$			10.87 11.18	8.95 9.25	32.64 33.72	-6.37 -6.67	28.89 26.67

The manuscript was modified as indicated below:

300 As batch size can influence the optimisation and generalisation behaviour of deep neural networks (Keskar et al., 2017), additional $HFNet_{JE}$ experiments were performed using gradient accumulation to achieve an effective batch size of 128. The results indicated that batch size had only a minor influence on overall performance, with somewhat larger differences for extreme events. The effect was not systematic, although the larger effective batch size resulted in somewhat poorer performance for extremes. Therefore, although the influence of batch size appears limited, it cannot be fully excluded when interpreting the differences between HFNet and $HFNet_{JE}$. Each experiment was repeated three times to account for variability due to random initialisations of model parameters.

425 by about 21% and 29% for regular events. In contrast, extreme events (HFO-A ≥ 20 cm) are generally better predicted by the adapted HIDRA model, HFNet, for which RMSE and MAE are lower by approximately 10% and 7%, in comparison to the ones for $HFNet_{JE}$. As noted in Sect. 3.4, however, a possible influence of the different batch sizes used for HFNet (128) and $HFNet_{JE}$ (32) on the observed performance differences cannot be excluded. It is worth noting that adding the Richardson number as an additional input variable to the $HFNet_{JE}$-ERA5 configuration (Table 2) resulted in errors for regular events that

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.

We explicitly noted this in captions of all relevant tables (Tables 2–5). As an example, we attached the revised caption of Table 2 (please see below).

380 **Table 2.** Mean performance metrics (RMSE, MAE, relative error, bias, and accuracy) for the 24-hour lead-time forecasts using ERA5 input at the Bakar station, averaged over three runs. Metrics, defined in Sect. 3.5, are reported for all HFO-As (All), regular events (HFO-A < 20 cm; Reg), and extreme events (HFO-A ≥ 20 cm; Ext). ~~Accuracy is defined as the proportion of predictions differing from observations by less than one standard deviation.~~ Underlined values denote optimal values within each category (All, Reg, and Ext). Models were selected based on the smallest maximum validation loss, thereby prioritising performance for extreme events (Sect. 3.4).

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

Thank you for your positive recommendation and constructive feedback. We appreciate your time and effort.