For Referee #2 (RC2, Dr. Erlend Storrøsten)

General comments:

The manuscript presents an emulator for offshore tsunami wave propagation, which is a useful and relatively less explored direction compared with the larger body of emulator studies focused on inundation. The authors approximate the numerical propagation model using a Fourier Neural Operator in an iterative prediction framework, with the aim of reproducing the full simulation output rather than only selected quantities. This is a meaningful contribution, and the manuscript is overall well written. Although the dataset is somewhat limited, the evaluation is systematic and gives a useful first assessment of the method's potential. That said, some important aspects of the method's applicability and limitations should be clarified more explicitly.

Response: We sincerely thank the Reviewer #2 for the encouraging assessment and for comments that helped us clarify the applicability and limitations of the method. In the revised manuscript, we added a gauge-level scatter analysis of bias in maximum amplitude. We also conducted a bathymetry experiment to examine how the trained operator responds to changes in bathymetry. In addition, we investigated the sensitivity of the reference solver to grid resolution. The specific revisions made in response to each comment are described below. The responses are in blue text and the changes made are shown in red text. The revisions will be traceable in the revised manuscript.

Specific comments:

R2M1. The efficiency comparison is currently made against a relatively traditional CPU-based solver. This is informative, but it does not fully establish the computational advantage of the emulator relative to modern high-performance tsunami codes. I would therefore encourage the authors to discuss how the reported efficiency compares with highly optimized GPU-accelerated implementations, even if only qualitatively (see https://link.springer.com/chapter/10.1007/978-3-319-55480-8_16). In this respect, the inclusion of PCOMCOT in Table 4 is interesting, since it highlights that dispersive models are substantially more computationally demanding; this broader context may deserve slightly more attention.

Response R2M1: We agree that comparison with the CPU-based COMCOT configuration alone does not establish a general computational advantage over modern GPU-accelerated tsunami solvers. We revised the manuscript to clarify the comparison and the scope of our efficiency claim. As the reviewer noted, the PCOMCOT entry in Table 5 in the revised manuscript highlights the substantially higher cost of dispersive solvers. We now discuss this contrast explicitly.

Change made R2M1: Section 4.1 now discusses "This acceleration lowers the cost of evaluating each additional scenario. It therefore makes larger scenario ensembles more practical within the validated parameter range. The appropriate computational comparison also depends on the reference solver. Highly optimized GPU solvers for the shallow-water equations reach much shorter runtimes than CPU codes. Tsunami-HySEA, for example, simulated 4 h of basin-scale propagation at 60 arc-second resolution in less than 4 min on a single GPU (Macias et al., 2016). Because its domain, grid, hardware, and numerical formulation differ from those used here, that runtime is not a direct benchmark for our surrogate. We therefore do not claim a general single-scenario speed advantage over GPU-optimized numerical solvers. The main computational benefit of the surrogate is its low marginal evaluation cost once the training cost has been amortized over repeated ensembles. Thus, the computational tradeoff is between a one-time training cost followed by 8.5--12 s per additional scenario and a COMCOT cost of approximately 91 s per scenario that increases linearly with the number of scenarios." These revisions appear at Lines 450-459 of the marked manuscript.

R2M2. The manuscript leaves some uncertainty regarding how site specific the model is. While bathymetry is included as an input (L61), the model appears to be trained on a fixed grid. If so, does

the model need to be retrained for a different bathymetry or site? More generally, the extent to which the method generalizes with respect to bathymetry and resolution should be clarified. This is particularly relevant because the iterative prediction strategy may involve both error accumulation and higher cost than direct prediction.

Response R2M2: We thank the reviewer for raising the important point. Because bathymetry is provided as an input, we were also interested in examining whether the trained model could respond to controlled bathymetric changes in the same basin without retraining. We therefore conducted an additional bathymetry perturbation test. We added a mention that application to a different basin and resolution would require retraining and validation.

Change made R2M2: The experiment is summarized in the new Sect. 3.6 and reported in full in Supplement Sect. S4 (Fig. S4). Supplement Sect. S4 now evaluates the model's response to controlled bathymetric changes in the existing basin. These revisions appear at Lines 425–435 and 538–540 of the marked manuscript and Lines 67–89 of the marked Supplement (Fig. 16 and Fig. S4).

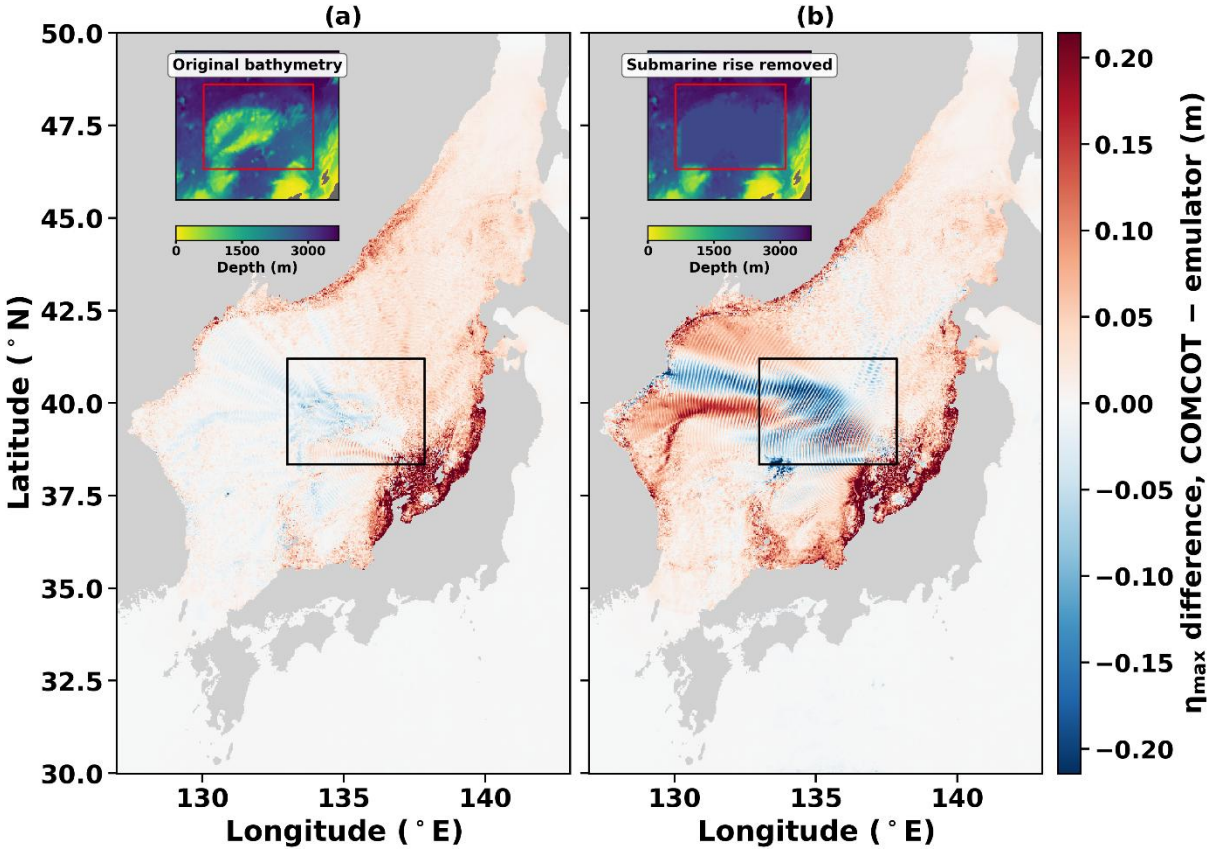

Figure 16. Difference in maximum surface elevation between COMCOT and the F-FNO for (a) the original GEBCO bathymetry and (b) the modified bathymetry after complete removal of a broad submarine rise in the central basin. The insets show the bathymetry within the modified region.

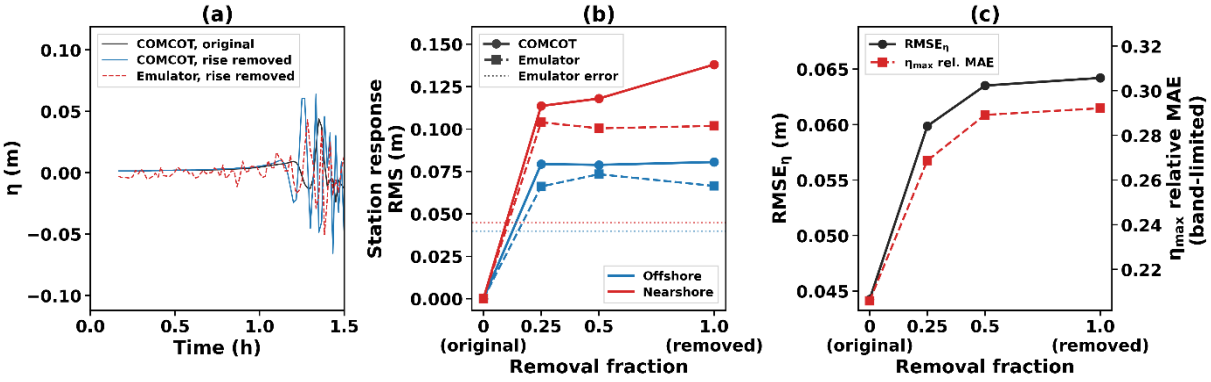

Figure S4. Response of the surrogate to the bathymetry change. (a) Free-surface elevation at gauge B400 for one test scenario, for COMCOT on the original bathymetry, for COMCOT with the rise removed, and for the surrogate with the removed-rise depth channel. (b) Gauge response RMS for the removal fraction, averaged over the three offshore and the three nearshore gauges. Circles show COMCOT and squares show the surrogate. The dotted lines mark the surrogate error on the original bathymetry, one for each gauge group. Panel (b) reports group means, while the transfer-ratio statistics mentioned in the text are computed for each gauge and level. (c) Domain RMSEη on the left axis and the band-limited relative MAE of the peak-elevation field on the right axis, averaged over the four scenarios. The two axes are scaled independently.

R2M3. Figures 11 and 13 suggest a possible bias toward underprediction, especially for sharp peaks. If such a bias is present, it is not well captured by the current metrics. A scatterplot of predicted versus simulated maximum amplitude could help clarify this point. It may also be worth discussing whether the use of RMSE contributes to smoother, slightly damped predictions. In a related study, similar issues were mitigated using an asymmetric loss function

(https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2025JH001100).

Response R2M3: We thank the reviewer for this suggestion. The requested scatter analysis confirms an underprediction bias, most clearly for magnitude extrapolation. A spatially averaged RMSE term and the finite Fourier-mode representation may contribute to peak damping, although the present analysis does not isolate their individual effects. We therefore discuss both as possible explanations and cite the suggested asymmetric-loss method as a potential remedy.

Changes made R2M3: We added Fig. 13, which compares F-FNO and COMCOT maximum surface elevations at the virtual gauges. The per-split median biases (−7.8 % for Test-E, −24.1 % for Test-M, −5.6 % for Test-EM) are reported in the figure legend. Section 4.2 now reads: “The model underestimates the peak elevation at 90 % of the Test-M gauges, and the median bias reaches −24 %. On Test-E and Test-EM the median biases stay within 8 % (Fig. 13).” Section 3.2 relates this bias to the reduced fraction of high-wavenumber spectral energy and notes that the present figures do not separate the effects of the finite mode count and the spatially averaged loss. Section 4.2 also cites the loss function proposed by Ragu Ramalingam et al. (2026), which assigns a larger penalty to underprediction. These revisions appear at Lines 374–381 and 507–518 of the marked manuscript (Fig. 13).

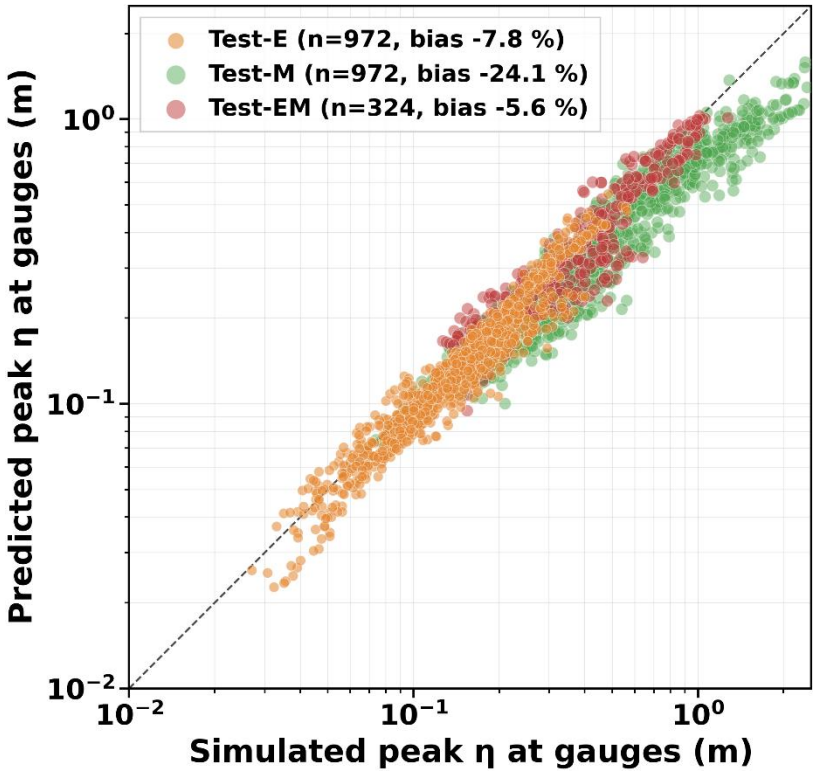

Figure 13. Predicted against simulated peak free-surface elevation at the virtual gauges for the three held-out splits. The dashed line marks the 1:1 relation. On Test-M the model underpredicts the peak at 90 % of the gauge–scenario pairs and the median bias reaches −24 %; on Test-E and Test-EM the median bias stays within 8 %.

R2M4. In Figure 11, the time series appear to show some spurious predictions prior to wave arrival that are not present in the simulations. Since preserving the sea at rest is an important property in tsunami modeling, it would be useful to comment on whether the F-FNO approach introduces spurious oscillations, and whether this could also help explain the outliers in Figure 12a.

Response R2M4: We thank the reviewer for pointing this out. The training procedure penalizes water level changes during calm periods. Small fluctuations still remain before wave arrival. We measured their size at the six virtual gauges. As shown in Table S9, the median pre-arrival RMS was about twice the COMCOT value. A predicted amplitude of 50 mm before the leading edge was uncommon. The highest rate was 1.87% in Test-EM.

We also examined the large arrival-time errors in Fig. 12a. Only three of the eight errors above 30 min were early predictions. Pre-arrival fluctuations may explain some early errors. They do not explain most of the large errors.

Changes made R2M4: Section 4.1 summarizes methods and results of the pre-arrival fluctuations evaluation tests. Table S9 now reports the pre-arrival free-surface fluctuations metrics for F-FNO and COMCOT. Sect. S8 and Table S9 provide the full method and results. Section 4.1 appear at Lines 460–467 of the marked manuscript. The related description in Sect. S8 is at Lines 117–125 of the marked Supplement (Table S9). Table S2 in Sect. S2 also tested arrival diagnostics for the new and old designs. Lines 29–31 reports early error results in Sect. S2 (See Table S2's Gross arrival errors).

Table S9. Pre-arrival free-surface diagnostics for the three held-out tests. The final column reports windows in which the predicted η reached 50 mm at least once before the COMCOT leading edge. This diagnostic does not apply the consecutive-sample condition used for the formal arrival-time calculation.

Split	Windows	F-FNO RMS median (mm)	F-FNO RMS p90 (mm)	COMCOT RMS median (mm)	Windows with predicted $\lvert \eta \rvert \geq 50$ mm
Test-E	737	3.56	8.77	1.87	0 (0.00 %)
Test-M	788	3.50	6.05	1.99	1 (0.13 %)
Test-EM	214	6.40	11.08	3.45	4 (1.87 %)

Minor comments:

L9: COMCOT has not been introduced.
Response: COMCOT is now spelled out at first use in the abstract (Cornell Multigrid Coupled Tsunami Model). (marked manuscript, Lines 9–10).

L12: "From the logic tree, we hold out...". It sounds like you construct a single test set, however in reality you explore multiple different train-test splits.
Response: Abstract now states that "evaluate the surrogate on three test splits that isolate the two factors and their combination". (marked manuscript, Lines 14–16).

L35: Due to the large number of potential sources PTHAs are often based on several thousand simulations. See "Selva et al. Quantification of source uncertainties in Seismic Probabilistic Tsunami Hazard Analysis (SPTHA)".
Response: The introduction now states that "hundreds to thousands of possible scenarios are usually evaluated" and cites Selva et al. (2016). (marked manuscript, Lines 27–29 and Lines 32).

L48: "...learn wavefields as function spaces rather than vectors.". It is not entirely clear to me what is meant here.

Response: The sentence has been rewritten: "The Fourier Neural Operator (FNO) and its factorized variant (F-FNO) learn an operator that maps wavefield states at successive times in the simulation to the subsequent wavefield states. In contrast to a fixed-dimensional vector of values at a set of stations, the input and output here are functions on the model domain." (marked manuscript, Lines 50–54).

L75: "The simulation inputs are normalized with wet-cell masking..". What does it mean?
Response: The sentence now reads: "The simulation outputs are restricted to ocean cells with a land–sea mask and standardized channel-wise with training-set statistics, and the scenarios are partitioned into training, validation, and test sets (Step 3)." (marked manuscript, Lines 82–84).

Figure 1: I like the figure, however it contains several notions that has not yet been defined such as Y_t, ATE, Context length,.. This is valuable when looking back at the figure after reading a bit more, but may bee confusing at first.
Response: The caption now ends with: "Notation used in the figure (y_t, context length S, rollout horizon H, LHS, ATE, BEE) is defined in Sects. 2.1, 2.2 and 2.4." Fig. 1 was also updated to show the stratified LHS design in Step 01. (revised Fig. 1 and its caption).

L122: "All dynamic variables are normalized channel-wise...". Normalization is often based on some

kind of invariance, is this the case here?
Response: Physical invariance is not intended. Normalization is a channel-by-channel standardization for optimization, and since statistics are calculated based on wet cells, land values do not affect moments. As the same fixed transformation is applied to both the validation and test data, this is merely a fixed transformation of units and does not imply dynamic symmetry. The clause added to Section 2.2 is as follows: "This normalization is a fixed linear scaling for each channel. It helps the training converge and does not represent any physical invariance." (marked manuscript, Lines 127–132).

L136: Are the variables H, W and d_v defined? What are the dimensions?
Response: They are now defined in Sect. 2.2: "where H and W are the grid dimensions and d_v is the number of latent channels." (marked manuscript, Lines 144–146).

Equation 6: The circle with a dot (\odot) operator has not been defined.
Response: The operator is now defined below Eq. (5) in the revised manuscript: "where $\odot$ denotes element-wise multiplication in the spectral domain". (marked manuscript, Lines 151).

L179: The training parametere are stated both in the text and in the table caption below.
Response: The body text now reads: "All variants are implemented in PyTorch and share the same training schedule and loss weights, summarized in the caption of Table 2." The numerical values no longer appear in the body text. They are given once in Table 2 for the four model configurations, and the full set of loss hyperparameters is listed in Table B1 of the appendix. (marked manuscript, Lines 188–193 and the caption of Table 2).

Figure 7: p-values are reported here and before (L284). Could you briefly mention the underlying hypothesis associated with this p-value?
Response: The following sentence was added to Sect. 3.1: "The p-values in these panels come from a two-sided Mann–Whitney U test of the null hypothesis that the scenario-level errors of the two sets are drawn from the same distribution." (marked manuscript, Lines 300–302).

Figure 11: Are the y-axes aligned in both columns of the plots? Markers seems to be placed differently.
Response: Thanks for pointing this out. Figure 11 has been regenerated as buoy signals for representative cases. The y-axis limits are now set separated for each panel. The same correction has been applied to Figs S1 and S2. (revised Fig. 11 and Supplement Figs. S1–S2).

Figure 14: Looks like allot is happening in the interval 0-10. Perhaps a log transform could make the details of the plot somewhat more visible?

Response: We agree and thank the reviewer for this suggestion. The horizontal axis of the revised figure (now Fig. 15) uses a logarithmic scale. The early error growth is now clearly visible. The log axis also shows that the growth slows down much earlier than step 100 as stated in the original submission. (revised Fig. 15, See also Lines 395–403 of marked manuscript).

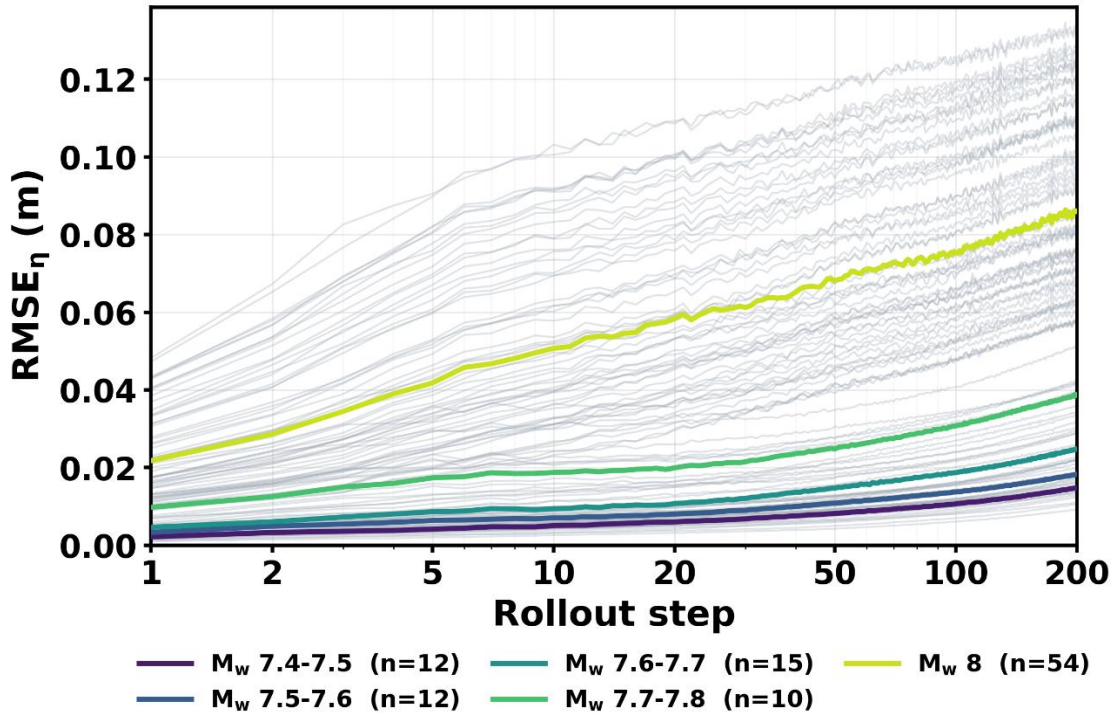

Figure 15. RMSE η evolution over the 200-step prediction horizon for the Selected model. Each gray line represents a single scenario from the validation set and Test-EM split (103 scenarios total). The colored curves give the median of each magnitude bin, with the number of scenarios in parentheses.

Additional Revisions: We revised Table 1's description related to the bottom friction as we use the frictionless linear shallow-water equations in COMCOT. We also modified the Eqs. (2) and (3) accordingly and deleted bottom friction source terms in the previous manuscript. We also corrected minor grammatical and wording errors found throughout the manuscript.