For Anonymous Referee #1 (RC1)

General comments:

The paper creates a new type of tsunami surrogate model. It is based on a Factorized Fourier Neural Operator (F-FNO) a good choice due for NO due to the ability of F-FNO to capture higher Fourier modes -- as shown in the comparison with regular FNOs. The application to the sea of Japan tsunami sources shows excellent timing but the amplitudes are underestimated by some margin.

The method is underperforming against more established emulation approaches cited in the context of peaks for hazard assessments. However the F-FNO benefits are more about its ability to emulate wave propagation between function spaces: inputs and outputs can be fields. The impulsive source forcing is interesting as well as physics-guided learning strategies. Nice implementation and extension of FNO, with lots of care in the accumulation of errors over hundreds of steps, and loss tailored to the problem, with efficient implementation. The design of experiments is unfortunately substandard.

Response: We sincerely thank the Reviewer #1 for very careful reading of the manuscript and for comments which helped us to identify strengths and weaknesses of the work as it was presented. Two major comments led to a completely reworked study. We adopted Latin Hypercube design for training and compared it to logic tree based training data set (see our responses to the specific comments 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:

R1M1. The main attraction of NOs as explained, is the ability to take as inputs fields v GPs and other approaches. What can we learn about the dynamics and patterns with an F-FNO emulator v GPs?

Response R1M1: The key difference lies in how the two approaches represent the propagation process. A GP emulator maps source parameters directly to scalar outputs without representing intermediate wavefield states. The F-FNO learns the sequential wavefield evolution explicitly, encoding propagation dynamics such as refraction, reflection, and basin-scale interference within the operator. The GP emulators used in this context take source parameters only, so a change in bathymetry would require a new set of simulations and a retrained emulator. We demonstrate such an analysis in our response to R1M5.

This operator also adds a feature not provided by fixed-output emulators. It outputs the total state (η, u, v) at all grid points and time steps, allowing all diagnostics to be calculated retrospectively. The arrival times, maximum elevation maps, and spectral statistics reported in the paper are all the results of running the surrogate model once, so there is no need to create a new emulator even if there are new sites, new water level thresholds, or new arrival criteria.

Changes made R1M1: The following statements were added to Sect. 4.1: “For most downstream uses, the full wavefield is what matters, not a particular statistic.” and “The arrival times, the peak-elevation maps, and the spectral diagnostics that we report here were all computed from the fields produced by a single surrogate rollout, without training separate models for individual diagnostics” The bathymetry-perturbation experiment was added as Sect. 3.6 and Supplement Sect. S4 (Figs. 16 and S4). These revisions appear at Lines 425–435 and 492–497 of the marked manuscript and Lines 67–89 of the marked Supplement (Figs. 16 and S4).

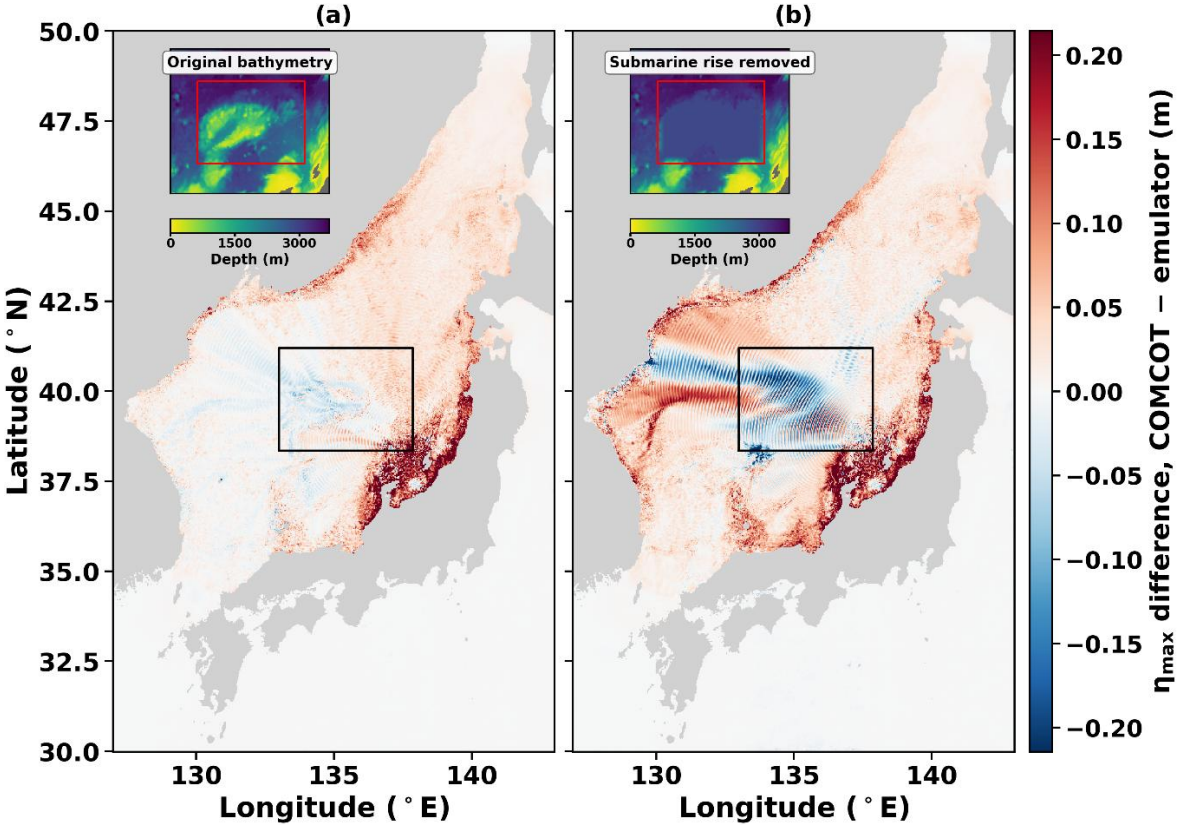

Figure 16. Difference in maximum surface elevation between COMCOT and the F-FNO for (a) the original GEBCO bathymetry and (b) the modified bathymetry after complete removal of a broad submarine rise in the central basin. The insets show the bathymetry within the modified region.

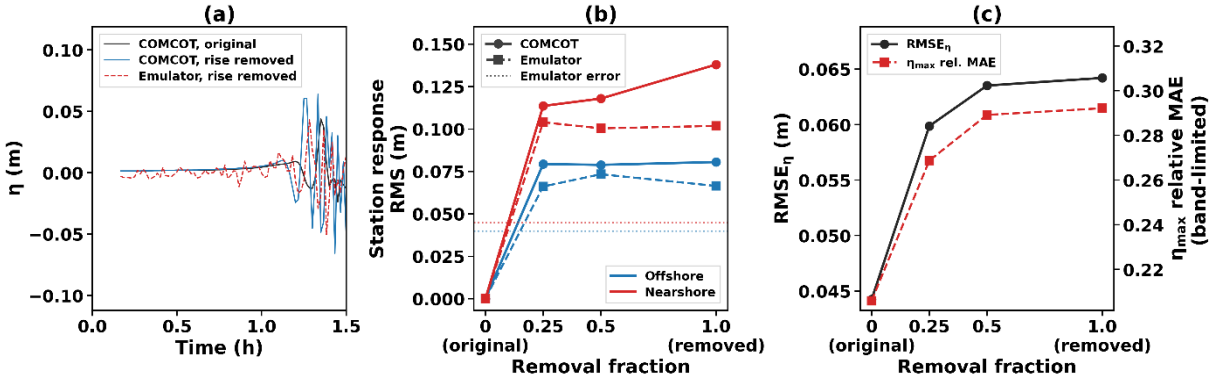

Figure S4. Response of the surrogate to the bathymetry change. (a) Free-surface elevation at gauge B400 for one test scenario, for COMCOT on the original bathymetry, for COMCOT with the rise removed, and for the surrogate with the removed-rise depth channel. (b) Gauge response RMS for the removal fraction, 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 RMSE η 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.

R1M2. The logic tree lines 202-211 is not a good design to train the F-FNO, unable to explore enough across 3-4 values only, compared to say a Latin Hypercube or a sequential design. It is also very large with 486 values despite for only 6 parameters (for each scaling law type). A GP would need less than 100 samples to be trained on 6 parameters. The training needs to be redone and discussed. I expect a big improvement in the quality of the emulation.

Response R1M2: We thank the reviewer for this suggestion and have rebuilt the training database accordingly. One clarification first, because it determines how the sample budget should be read. The six source parameters are not inputs to the surrogate. The network takes the wavefield state (η, u, v) with the static bathymetry channel and predicts the next state, while the parameters enter only through the Okada initial condition. The surrogate learns a state-transition operator on a function space rather than a six-dimensional parameter-to-scalar regression, and each scenario contributes 400 output frames on the full grid from which training windows are drawn. A sample budget comparable to that of a scalar GP emulator is thus not directly applicable.

We nonetheless agree that a factorial grid is a poor way to diversify initial conditions, and we have replaced it. We regenerated all 864 scenarios using a Latin Hypercube design. Strike, dip, rake, and moment magnitude were sampled across their ranges. Epicenter location and scaling law remained unchanged. We kept the same three test groups and the same number of test scenarios as in the original submission. The 378 test scenarios are divided into Test-E (162), Test-M (162), and Test-EM (54) (Fig. 6). To isolate the effect of the training design, we compared two models. One was trained on the new Latin Hypercube database. The other is the model of the original submission, trained on the logic-tree database. In the supplement this database is described as the gridded (factorial) counterpart with the same parameter ranges (Sect. S2). Both models were evaluated on exactly the same Latin Hypercube test sets, so any difference between the results comes from the training design and not from the test scenarios. The two models share the same architecture, loss weights, and training schedule.

The redesign did not reduce the field error, and we report this directly. The two designs are comparable on all three splits (Table S1), because the logic-tree model already reached RMSE η of 3–8 cm, and the residual error at this level is dominated by autoregressive accumulation over the 200-step rollout rather than by coverage of the source-parameter space. Further reduction would require architectural or loss-function changes. The benefit of the continuous design appears instead in the arrival diagnostics (Table S2). The median pre-arrival RMS is essentially identical for the two designs (6.4 mm and 6.5 mm), but the upper tail differs by a factor of two (90th percentile of the pre-arrival maximum, 29 mm against 61 mm). This matters because the 5 cm arrival criterion is a threshold test on the maximum rather than on the mean. The factorial model triggers it before the leading wave at 14.5 % of the Test-EM gauges against 1.9 % for the Latin Hypercube design, and its mean absolute arrival-time error is 11.5 min against 8.9 min. The same distinction explains the unchanged field errors, since RMSE η is averaged over all wet cells and is insensitive to the tail that governs the threshold-based diagnostics.

On Test-M the factorial design is marginally better, because one third of its in-range cases sit exactly at M_w = 7.8, closest to the extrapolation target, whereas the Latin Hypercube design spreads samples uniformly over [7.4, 7.8]. We regard this as a trade-off between the two designs rather than a deficiency of either, and report both.

Changes made R1M2: Sect. 2.3 and Appendix A were rewritten around the stratified LHS design, and Fig. 6 was replaced by the sampling schematic of the new design. The design-sensitivity experiment is reported in the new Supplement Sect. S2. The revised sampling design and holdout definitions appear at Lines 195–218 and 743–773 of the marked manuscript and Lines 14–36 of the marked Supplement (Fig. 6 and Tables S1–S2).

Table S1. Field errors of the identical model configuration trained on the continuous (LHS) and on the factorial database, evaluated on the common test splits (RMSEη in m, mean ± std over scenarios, 200-step rollout).

Training database	Test-E	Test-M	Test-EM
Latin Hypercube design (LHS)	0.0264 ± 0.0119	0.0650 ± 0.0220	0.0796 ± 0.0228
Logic tree (factorial)	0.0274 ± 0.0123	0.0589 ± 0.0190	0.0779 ± 0.0225

Table S2. Arrival and still-water diagnostics of the two designs on the common Test-EM cases. Gross arrival errors are gauge pairs with an absolute arrival-time error above 30 min. From Supplement Sect. S2.

Diagnostic	Continuous (LHS)	Gridded (factorial)
Median pre-arrival RMS	6.4 mm	6.5 mm
90th percentile of pre-arrival maximum	29 mm	61 mm
Gauges reaching the 5 cm criterion before the leading wave	1.9 %	14.5 %
Gross arrival errors (early detections)	8 (3)	26 (22)
Mean absolute arrival-time error	8.9 min	11.5 min
Arrival-time regression R^2	0.918	0.822

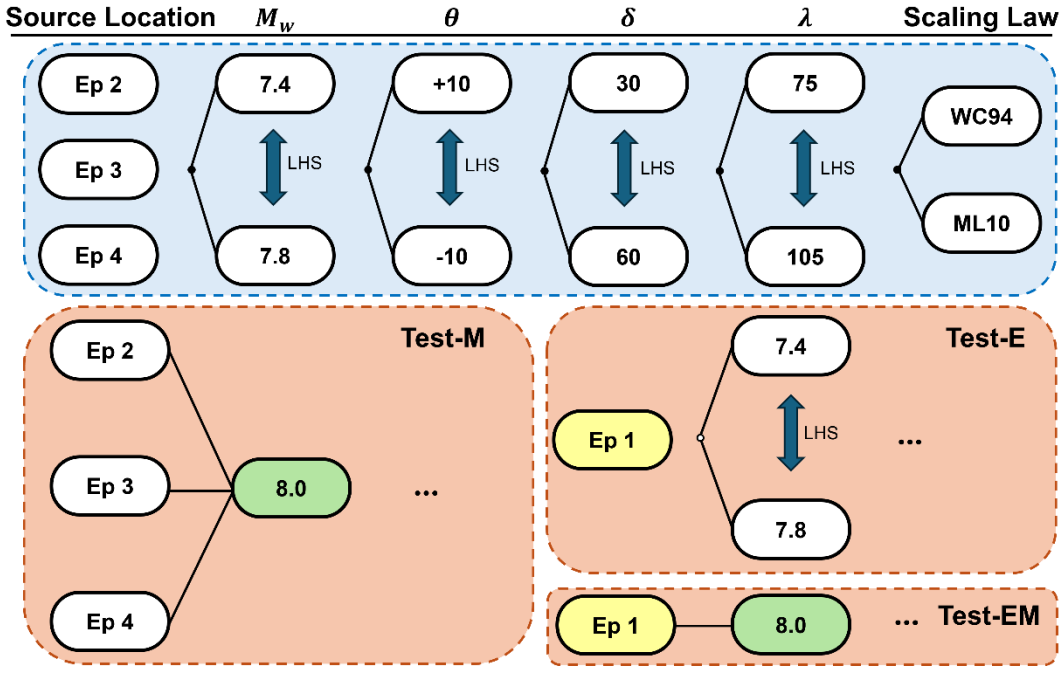

Fig 6. Schematic of the sampled source-parameter space used in this study and the resulting held-out test splits. Training data use Ep 2 through Ep 4 with strike offsets from −10 to +10, dip from 30 to 60, rake from 75 to 105, magnitudes Mw 7.4 to 7.8, and two scaling laws (WC94/ML10). Arrows indicate

that each parameter is sampled continuously between the two limits by Latin hypercube sampling. Held-out tests are defined as Test-M (Ep 2–Ep 4, Mw = 8.0 only), Test-E (Ep 1, Mw 7.4 to 7.8), and Test-EM (Ep 1, Mw = 8.0 only).

All results in the manuscript and the Supplement were recomputed on the new database. This applies to Table 3, Figs. 5 and 7 to 12, Figs. 14 and 15, and Tables S4 to S7.

R1M3. For validation, the authors hold out the largest magnitude (8.0) and one specific source location for model evaluation and to test the scalability of the neural operator.

This is too small for validation, I suggest a Leave-one-out strategy.

Response R1M3: We appreciate this comment and have tested a leave-one-source-out strategy, together with a magnitude-interpolation holdout.

We perform four tests for the generalization of the location of the four different epicenters (Ep 1–4). The Ep 1 holdout is identical to Test-E. The other three are retrainings of the model on the remaining sources, with one epicenter excluded in turn and evaluated on the in-range scenarios of the excluded epicenter. The results of these four tests are consistent. The case-specific RMSE η for the four unseen locations are of moderate size and vary between 0.024 m and 0.031 m (Table 4).

Note, however, that the largest error is produced by the northernmost location (Ep 4) since it lies at the northern edge of the sampled source region, so this fold requires spatial extrapolation and not interpolation. The magnitude-interpolation holdout, which retrained the model once with the 134 development scenarios in $M_w \in [7.55, 7.65]$ withheld, gives the smallest error of all evaluations in Table 4. Unseen magnitudes inside the sampled range are therefore easier for the model than unseen epicenters, and both are easier than the extrapolation to $M_w = 8.0$ outside the sampled range.

We keep the three original partitioning schemes for the subsequent tests in order to test the three cases of generalization. These are (1) a shift of the location of the epicenter of a single earthquake of known magnitude, (2) an extrapolation of the magnitudes of unknown earthquakes beyond the sampled range, and (3) a combination of both.

Changes made R1M3: The following was added to Sect. 2.3: “Two additional holdout protocols were used with the same database. In the leave-one-source-out protocol, we retrained the model four times with one epicenter excluded in turn and evaluated it on the in-range scenarios of the excluded epicenter. In the magnitude-interpolation holdout, we retrained the model once with the 134 development scenarios in $M_w \in [7.55, 7.65]$ withheld and evaluated it on the withheld band.” Sect. 3.1 reports the results in Table 4. The added Leave-one-source-out test and results appear at Lines 214–218 and 343–349 of the marked manuscript (Table 4).

Table 4. Leave-one-source-out and magnitude-interpolation holdout performance of the Selected configuration (mean ± std over held-out scenarios, 200-step rollout). Reproduced from Table 4 of the revised manuscript.

Holdout	RMSEη (m)	NRMSEη	N
Ep 1 excluded (= Test E)	0.0264 ± 0.0119	0.592	162
Ep 2 excluded	0.0237 ± 0.0117	0.526	162
Ep 3 excluded	0.0253 ± 0.0115	0.552	162

Ep 4 excluded	0.0307 ± 0.0134	0.662	162
$M_w \in [7.55, 7.65]$ withheld (interpolation test)	0.0169 ± 0.0050	0.389	134

R1M4. There is no comparison with other techniques on the example case. In the discussion some mention of the possibly higher accuracy of other methods due to focus on specific outputs, but some illustrations of shortcomings/strength of the method would be good.

Response R1M4: We have added a comparison with a Gaussian-process (GP) emulator, the established approach in this application area (Sarri et al., 2012; Guillas et al., 2018). A task-specific GP was trained to predict the peak free-surface elevation at each of the six virtual gauges from the source parameters, using 437 training scenarios and evaluated on the same test sets. The comparison is now reported in Supplement Sect. S9 (Table S10), and Sect. 4.1 summarizes the outcome. The epicenter enters the GP as latitude and longitude and not as a categorical label, so the emulator is free to interpolate toward the unseen source location. Here we only want to report the simple GP method and do not aim to set up an optimal GP approach in this work.

F-FNO has about one-quarter of the GP error on Test-E and about half on Test-EM. The GP uses only the source parameters and contains no information about how tsunami waves travel across the basin. The training data also include only three source locations. These limitations may reduce its accuracy at an unseen epicenter.

The ordering reverses on Test-M. At the trained epicenters, the GP error is about half the F-FNO error when predicting the response at M_w =8.0. The GP predicts a single peak value directly. F-FNO obtains the same value after 200 successive prediction steps. This difference may contribute to the lower GP error on Test-M. Note that the quantities in Table S10 are gauge-level peak elevations and are not comparable with the field-averaged RMSE η quoted in the abstract.

Figure 13 shows a limitation in magnitude extrapolation. On Test-M, the model underestimates peak elevation for 90% of the gauge-scenario pairs and has a median bias of -24%. For Test-E and Test-EM, median biases remain with less than 8%.

For the architecture, the F-FNO improved over the standard FNO for both RMSE η and the spectral band-energy error, by 23–40 % and 89–94 % respectively (Supplement Sect. S5). We did not test other deep learning architectures such as ConvLSTM, U-Net or transformers, since each would require its own separate and substantial tuning, and they would then become the main subject of the paper instead of the simple factorized architecture that we are studying here.

Changes made R1M4: Section 4.1 now states the outcome, and the two sentences that ruled out a quantitative comparison were replaced by: “We cannot quantitatively compare our error values with those reported by other works on tsunami surrogate models, because each work is characterized by different computational settings, such as the spatial domain, discretization, and validation criteria. A direct comparison is possible for the gauge peak elevation within the present database. We fitted a Gaussian process emulator (Rasmussen and Williams, 2005) to the peak elevation at the six gauges using the same 437 training scenarios. On Test-E, its mean absolute error is about four times the F-FNO error. Its error is about twice the F-FNO error on Test-EM and about half on Test-M (Supplement Sect. S9). The Gaussian process uses only the source parameters and holds no information about wave travel across the basin. Its training data include only three epicenters, so it has little support for predicting the response at a new location. At an epicenter included in the training data, the GP predicts the peak value directly. In contrast, F-FNO obtains the same value after 200 successive prediction steps.” This discussion appears at Lines 473–482 of the marked manuscript (Fig. 13).

We added Supplement Sect. S9, which reports the GP baseline in full together with Table S10. This appears at Lines 127–156 and Table S10 of the marked Supplement.

Table S10. Peak free-surface elevation at the virtual gauges for the task-specific GP baseline and for the F-FNO surrogate on the three held-out splits. Both are evaluated at the same six gauges over the same scenarios, giving 972, 972 and 324 gauge and scenario pairs. A negative median bias means the peak is placed too low. The last column gives the share of pairs below the reference.

Split	Method	MAE (m)	Median bias (%)	Below reference (%)
Test-E	GP	0.083	+38.2	15.9
Test-E	F-FNO	0.020	−7.8	72.4
Test-M	GP	0.073	−10.1	78.7
Test-M	F-FNO	0.158	−24.1	90.0
Test-EM	GP	0.126	+31.8	13.9
Test-EM	F-FNO	0.064	−5.6	62.3

We also added the peak-amplitude scatter plot as the new Fig. 13. Section 3.2 now explains how the two error measures are related: “The spectral and amplitude results show related tendencies. During the first 1.7 h of the rollout, F-FNO has a lower fraction of energy in the high-wavenumber band than COMCOT (Fig. 12b). This band represents variations over short spatial distances. Figure 13 shows a negative median bias in peak elevation for all three test sets. The largest bias occurs in Test-M. Taken together, these results are consistent with smoothing in the predicted fields. [...] These features may make small-scale patterns more difficult to preserve, but the present figures do not separate the two effects.” This is located at Lines 374-382.

Section 4.2 now reports the relative bias: “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).” This discussion appears at Lines 507–518 of the marked manuscript.

The standard FNO comparison remains in Supplement Sect. S5, and Tables S4 to S6 were recomputed on the new database. The standard-FNO comparison appears at Lines 90–106 of the marked Supplement (Tables S4–S6).

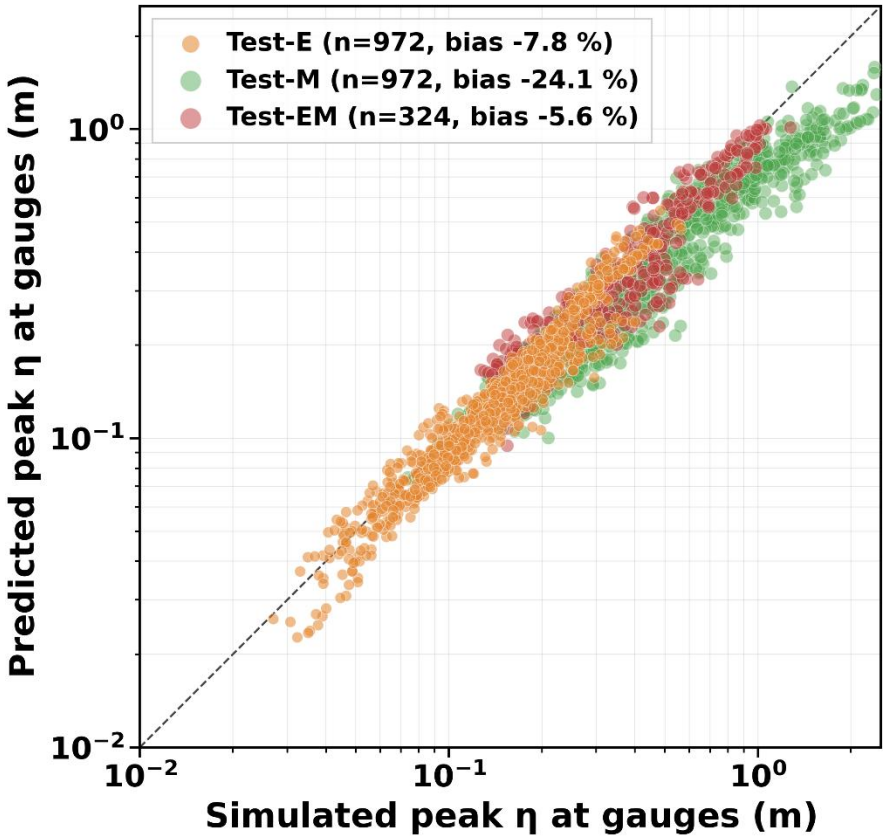

Figure 13. Predicted against simulated peak free-surface elevation at the virtual gauges for the three held-out splits. The dashed line marks the 1:1 relation. On Test-M the model underpredicts the peak at 90 % of the gauge–scenario pairs and the median bias reaches −24 %; on Test-E and Test-EM the median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

R1M5. The introduction is good, the motivation explicit with the aim to go beyond "predicting a fixed set of outputs such as maximum water level and inundation maps": explain the benefits obtained from that? Dynamics are mentioned and nice motivation but what can potentially be learned from this approach?

Response R1M5: Compared with emulators that predict selected values at fixed locations, the F-FNO provides the full wavefield. It also changes its predictions when the input bathymetry changes. We now explain these benefits in Sect. 4.1.

The F-FNO predicts η, u, and v at every grid point and output time. The same predicted fields can be used to calculate arrival times and peak-elevation maps. They can also be used to examine water velocity. A new emulator is therefore not needed for each measure.

We also tested the model after removing a broad submarine rise from the input bathymetry. The location of the wave-bending pattern shifted in the same direction as in COMCOT. The median size of the F-FNO response was 86 % of the COMCOT response. This result was obtained without retraining the model.

The model saw only the original bathymetry during training. The modified depth field was therefore new to the model. In a control test, we changed only the depth input. This test reproduced nearly the same response as the full test, with a median ratio of 0.99. The result shows that the model uses the bathymetry input when making its predictions. Supplement Sect. S4 provides the full analysis.

This experiment is limited to a controlled bathymetric change within the same basin and on the same grid. It does not establish reliable prediction for any bathymetric change or for another basin.

Changes made R1M5: The benefits are now stated in Sect. 4.1: "Access to the full wavefield allows multiple diagnostic quantities to be derived from the same surrogate output. Statistical emulators are designed to produce the very same quantity that was used for training the emulator. For a fixed-output emulator, changing the target site or diagnostic definition would require a new emulator, whereas the operator returns η, u, and v at all points on the model, and all of the above-mentioned diagnostics can be computed directly from these fields. The arrival times, the peak-elevation maps, and the spectral diagnostics that we report here were all computed from the fields produced by a single surrogate rollout, without training separate models for individual diagnostics." This revision appear at Lines 492–497. And the bathymetry experiment is reported in Sect. 3.6 and Supplement Sect. S4 (see R1M1). These revisions appear at Lines 425–435 and 538–540 of the marked manuscript and Lines 67–89 of the marked Supplement (Figs. 16 and S4).

R1M6. Fig 8 there is an obvious issue of extrapolation as outside the range of training.

Response R1M6: We agree that Fig. 8 shows a clear loss of accuracy outside the magnitude range used for training. We excluded the (M_w=8.0) scenarios from both training and model selection. This provided a strict test of prediction beyond the training range. We retained the results because they show an important limit of the current emulator.

The magnitude-interpolation test supports this interpretation. In this test, we withheld scenarios with (M_w[7.55,7.65]). This test had the lowest surface-elevation error among the evaluations in Table 4. The model therefore performed better when predicting within the trained magnitude range than when predicting beyond it.

Changes made R1M6: The magnitude-interpolation holdout is reported in Sect. 3.1 and Table 4 (see R1M3). Sect. 4.2 states the limit of magnitude extrapolation. These points are documented at Lines 205–208, 343–349, and 504–518 of the marked manuscript (Table 4).

R1M7. Fig 10 and 11 very interesting. Often underestimated of peak wave height discussed in section 3.3

Response R1M7: We thank the reviewer for highlighting this result. We agree that Figs. 10 and 11 reveal a recurring underestimation of peak surface elevation. Section 3.2, Fig.13 and Section 4.2 discusses this limitation and quantifies its dependence on the test setting. We further discuss this behavior in relation to the loss of high-frequency spectral energy.

Changes made R1M7: Fig. 13 was added, and Sects. 3.2 and 4.2 quantify the peak bias per split (see R1M4). The added analysis appears at Lines 374–381 and 507–518 of the marked manuscript (Fig. 13).

R1M8. Speed-up is interesting 10x but COMCOT is fast model, PCOMCOT is discussed as more relevant extension 50-75x but also higher resolution near the coastline (as discussed but not performed) would be more expensive and more accurate?

Response R1M8: We agree that a more detailed coastal simulation would require more computation. However, higher resolution alone does not ensure greater accuracy. It can represent smaller spatial features. Its accuracy also depends on the bathymetry, the physical equations, and validation against observations.

The measured speedup in this study is about 10 times. This value applies only to the tested COMCOT setup, which uses a single 3.2 km grid and non-dispersive shallow-water equations. The practical benefit, however, lies not in the single scenario but in the cost of the ensemble. In probabilistic hazard assessment the scenario database is evaluated repeatedly, and the marginal cost of each additional

scenario is what accumulates. On the hardware used here, COMCOT requires approximately 91 s per scenario and about 22 h for the full 864-scenario database, whereas the trained surrogate completes the same set in under 3 h. Both figures refer to sequential execution, and both codes can be run in parallel, so the ratio rather than the absolute wall-clock time is the relevant quantity. A surrogate may provide a larger speedup for a dispersive model or a nested high-resolution model. The present study does not test this possibility.

The reported PCOMCOT runtime gives a ratio of about 50–75 when compared with the current F-FNO runtime. However, this ratio is based on one PCOMCOT case and a different computing setup. The F-FNO was also trained on COMCOT results, not PCOMCOT results. We therefore report the ratio only as runtime context. It is not a measured speedup for a PCOMCOT surrogate.

We also tested the effect of grid spacing. Sixteen scenarios were repeated at 1.6 km spacing. Four of these were also repeated at 0.8 km spacing. The mean absolute difference in the basin-scale peak-elevation fields between the 3.2 and 1.6 km grids was 1.76 cm. However, the amplitudes at individual coastal gauges did not converge across the three grid spacings. The test therefore supports the basin-scale field comparisons in this study. It does not establish grid-converged coastal predictions.

A nearshore emulator in the future should be trained on a validated high-resolution coastal model. Such a model would require accurate bathymetry and suitable nearshore physics. It would also need validation against observations.

Changes made R1M8: Changes in Section 4.1 appear Line 450-459 (See R2M1). Section 4.2 describes the limits of the current grid. Supplement Sect. S3 presents the grid-sensitivity test and its results. These revisions appear at 525–540 of the marked manuscript and Lines 38–65 and 113–115 of the marked Supplement (Table S3 and Fig. S3).

Table S3. Three-level refinement of the peak-elevation field for the four scenarios run at all three grid spacings. MAE values are mean absolute differences of $|\eta|_{max}$ after restriction to the base grid, and the ratio is MAE(base, fine2) divided by MAE(fine2, fine4). From Supplement Sect. S3.

Case	MAE(base, fine2)	MAE(fine2, fine4)	Ratio
test_E_Ep01_WC94_004	0.74 cm	0.63 cm	1.17
test_E_Ep01_ML10_065	1.02 cm	0.79 cm	1.29
test_M_Ep03_ML10_002	3.09 cm	2.24 cm	1.38
test_EM_Ep01_WC94_024	1.93 cm	1.53 cm	1.26

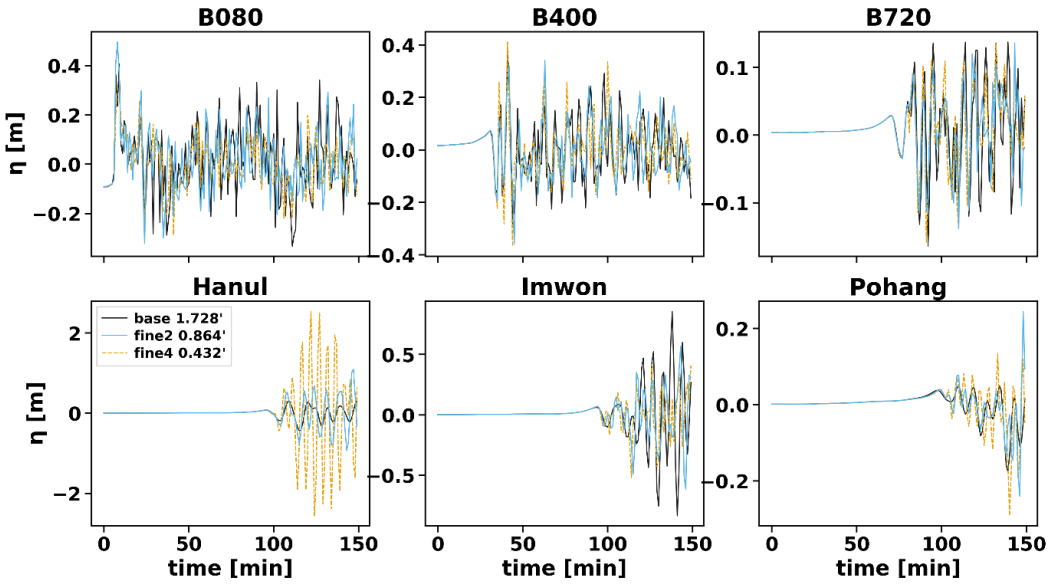

Figure S3. Free-surface elevation at the six virtual gauges for the scenario test_EM_Ep01_WC94_024 at three grid spacings: 1.728′ (base, solid), 0.864′ (fine2, dashed), and 0.432′ (fine4, dotted). The top row shows the offshore gauges and the bottom row the coastal sites. Note the different vertical scales.

R1M9. Usually surrogates allow gains of multiple orders of magnitude. Ideally a higher resolution model than a fast coarse resoution COMCOT should be the simulator that will then be emulated.

Response R1M9: We agree that a validated high-resolution coastal model would be a more relevant target for emulation. We conducted a short grid-sensitivity test to assess the adequacy of the 3.2 km grid and determine whether grid refinement materially changes the results. Sixteen held-out test scenarios were rerun at 1.6 km spacing, and four of these were also rerun at 0.8 km spacing. The results of peak-elevation fields showed little sensitivity to refinement. The mean absolute difference between the 3.2 km and 1.6 km results was 1.76 cm, while the three-level difference ratio for the 3.2 km, 1.6 km, and 0.8 km results was 1.27. However, the amplitudes at individual buoys did not converge across the three different resolutions. We therefore present the coastal-point results only as indicative values at each model resolution.

Changes made R1M9: The grid-refinement test is reported in Supplement Sect. S3 (Table S3 and Fig. S3), and Sect. 4.2 states which quantities are converged at the present spacing (see R1M8). The grid-refinement analysis appears at Lines 525–540 of the marked manuscript and Lines 38–65 of the marked Supplement (Table S3 and Fig. S3).

R1M10. The whole paragraph in the Conclusions "The surrogate functions as a rapid offshore boundary condition generator, facilitating the creation of nested grids that enhance the resolution of site-specific hazard studies. By accelerating the outer grid propagation.." is not fit for purpose. Indeed the main simulation costs are local not propagation in seconds but local high resolution coastal modelling. I suggest to remove this paragraph and ideally run the model at higher resolution -- including near the coastline -- to train te F-FNO. It should actually boost the quality of the F-FNO (using the right design) and provide a stronger statement of success in terms of impact on hazard assessements and warning and in terms of computational efficiency.

Response R1M10: We agree that the quoted paragraph overstated the nearshore application of the present model, and we have removed it from the Conclusions. The current F-FNO emulates basin-scale propagation on a fixed, coarse COMCOT grid. It does not replace the computationally dominant high-resolution nearshore simulation.

A validated high-resolution nested model would be a more relevant target for future emulation. We also ran the reference solver at 1.6 km and 0.8 km grid spacing. Higher resolution could improve the physical relevance of the training target, but it would not by itself guarantee lower surrogate errors. We now identify this extension as future work and restrict the conclusions to the scope demonstrated here.

Changes made R1M10: The Conclusions paragraph was deleted. The following sentences were also deleted from Sect. 4.1: "The surrogate operator can be adopted as a fast outer grid propagation model that provides offshore boundary conditions for nearshore simulations along the Korean coast." and "Basin-scale propagation is first simulated to provide wavefields at the boundaries of refined nearshore grids, which resolve local processes such as shoaling and inundation." From Sect. 4.3: "In cases where a basin-scale surrogate model is desired for nearshore accuracy, the results would then be combined with nested higher-resolution local models that use surrogate boundary conditions." The Conclusions now center on ensemble amortization and basin-scale screening. The deleted statements and replacement text are visible at Lines 448–450, 555–556, and 585–591 of the marked manuscript.

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.