Summary and general assessment

The manuscript presents a large parametric study (2772 numerical simulations with the non-hydrostatic Multilayer-HySEA model) of landslide-generated tsunamis at Stromboli. Two complementary analyses are carried out: (i) a variance-based (Sobol) global sensitivity analysis of the maximum wave height *H* with respect to six source parameters, and (ii) the derivation of volume–wave-height scaling relationships (logarithmic for subaerial sources, linear-with-exponential-depth-decay for submarine sources), which are then proposed as a rapid hazard-assessment tool and tested against the 2002 Stromboli event.

The study is well motivated, the numerical tool is established and appropriate, the parameter space is unusually broad, and the code and full dataset are openly archived (GitHub, Zenodo). The distinction between subaerial and submarine scaling regimes, and the exponential decay of the linear coefficient with submergence depth, are the most valuable and reusable contributions.

I nonetheless have some concerns, listed in the following, which I would want to see addressed before publication. I therefore lean towards **major revision**.

Specific comments

On the definition of *H*. Throughout the text, *H* is referred to in different ways. It is first introduced as the "maximum wave height", and the main structure of the paper and of the simulations involves this type of interpretation. But in other places, *H* is used as an amplitude: for instance, at page 11, row 279 I read "the peak amplitude (i.e., *H*)". These are not equivalent: in the standard IOC tsunami convention, *wave height* is the crest-to-trough vertical distance, whereas the largest positive excursion above still-water level is the (positive) *amplitude*. The waveforms in Fig. 2 indicate that the reported values correspond to the maximum positive sea-surface elevation (e.g. gauge 1: ~4 m at $V = 3\times10^6$ m^3, consistent with $\alpha\_log \cdot V = 1.33 \times 3$), i.e. the positive crest, not the crest-to-trough height, nor $\max|\eta|$ (the leading trough at gauge 1 reaches ~ −40 m at $V = 30\times10^6$ m^3, exceeding the crest). Please (i) define *H* unambiguously in the Methods (max η, $\max|\eta|$, or crest-to-trough) and adopt a single, glossary-consistent term throughout; (ii) note that the amplification factor and the scaling coefficients α are metric-dependent: at gauge 1 the factor for a ten-fold volume increase is ≈5.5 (crest), ≈6.7 (trough), or ≈6.2 (crest-to-trough); iii) at the near-source gauges the drawdown can exceed the crest; defining *H* as the positive crest alone discards the dominant excursion precisely where the signal is strongest, and the trough (sea withdrawal and return) is itself part of the hazard. Please state whether the choice of metric would change the sensitivity ranking and the near-source hazard picture. This comment is closely linked to my next comment: the amplitude metric and the non-linear regime of the receiving gauge are two facets of the same signal-interpretation issue.

Nonlinearity of the solution vs. the water depth at the sampling gauges. Multilayer-HySEA retains the non-linear terms, but whether they are actually "effective" at a given gauge depends on the local non-linearity parameter $\varepsilon = a/h$ (and, jointly with dispersion, on the Ursell number $Ur = aL^2/h^3$), not on depth alone. Using the depths in Table S1 and the amplitudes in the figures, the gauges span very different regimes: the deep offshore gauges 4 and 5 ($h \approx -1338$ to -1474 m) have $\varepsilon \lesssim 0.01$ and are effectively linear in *all* scenarios (consistent with the linear-dispersive wave trains reported there, i.e. small Ur); gauge 7 ($h = -2$ m) is strongly non-linear, essentially in the surf zone (ε of order unity for the larger events); and the intermediate gauges 0, 1 ($h \approx -50$ m), which are near-linear for small submarine waves, reach $\varepsilon \approx 0.4$–0.5 for the large subaerial events near the SdF. So, the near-source and village gauges (0, 1, 7, 8, 9, 10) can be markedly non-linear for the larger waves, while only the deep offshore pair is linear throughout. I think this has several implications worth addressing:

- Gauge 7 has the highest σ_α and σ_χ and is the only submarine case where α is not constant with volume (row 366: $\alpha_{log} \simeq$ 15 cm per 10^6 $m^3 \longrightarrow \alpha_{ref} \simeq$ 6 cm per 10^6 m^3), which the authors call "localized nonlinearity". This is very plausibly a *receiver-side* effect of its ~2 m depth. Making this connection explicit would turn an apparent anomaly into a physically understood feature.
- The logarithmic saturation of *H*(V) for subaerial sources is attributed entirely to the *source* (spreading and dissipation during descent). At shallow gauges, however, part of the saturation could be a *receiver-side* effect (non-linear shoaling and depth-limited breaking). A clean discriminating test is already contained in the dataset: the logarithmic trend also appears at the deep, linear gauges 4 and 5, where breaking is impossible. If the authors confirm and state this, it *strengthens* their source-based interpretation; I raise it as a request for clarification, not a criticism.
- A constant α is exact only where H = αV (linear superposition). At the non-linear gauges α depends on amplitude (hence volume), so a single α is a coarser approximation there, and part of the elevated σ_α is intrinsic to the non-linearity rather than to parameter spread. This qualifies the "nomogram" tool of Fig. 13 at exactly the gauges (villages) that matter for impact.
- The impact-relevant gauges (7–10, Stromboli village and Ginostra) are precisely the non-linear ones, whereas the method is most rigorous offshore, where non-linearity matters least. Moreover, the adopted hazard metrics are sea-surface elevations at fixed points, not run-up/inundation, which is even more strongly non-linear; this limitation of applicability should be stated.

Input distributions and sampling scheme for the Sobol indices. The Sobol indices are, by construction, sensitivities *conditional on the assumed input probability distributions*, not intrinsic properties of the physical system. The authors state that the interpolated response surface is sampled with 2^{15} quasi-random points "drawn from uniform distributions spanning the range of each input parameter" (Sect. 3.2). I suggest considering the following two aspects:

- Several of the chosen parameters span multiple orders of magnitude; some of them are quasi-geometrically spaced, like $m_f \in [10^{-5}, 10^{-1}]$ m^{-1} and, to a lesser extent, the volumes (3, 5, 8.7, 14, 21, 30 $\times 10^6$ m^3). If these are sampled uniformly in linear space, the samples are overwhelmingly concentrated near the upper end (for m_f, essentially between 10^{-2} and 10^{-1}), so the reported index measures sensitivity only over a small part of the nominal range. Please clarify whether m_f (and *V*) were sampled uniformly in linear or in log space and justify the choice. If linear, the authors should repeat the analysis in log space for m_f and comment on how the ranking changes.
- In Section 5.2, the manuscript claims that "The quantitative relationships between landslide volume, submergence depth, and maximum tsunami wave height derived in this study provide a potential tool for preliminary hazard assessment". I might agree, but I recommend adding an explicit caveat that the sensitivity ranking depends on the prior ranges/shapes assumed.

Physical meaning of the combined subaerial + submarine analysis. Both abstract and conclusions put the accent on the combined-dataset result (elevation $q \approx 61$ %, volume $V \approx 22$ %). My concern is that this combination involves a linear interpolation of *H* over an elevation axis *q* that runs continuously from +530 to −583 m a.s.l. Across the shoreline ($s = 0$), the physics changes (subaerial impact vs. submarine coupling) so that *H(q)* will possibly be non-continuous. This might make the origin of the very high combined-case sensitivity to *q* ambiguous. To solve this ambiguity, I ask the authors to (i) discuss explicitly the extent to which the combined-q dominance is an artifact of lumping two regimes, (ii) justify linear interpolation across $s = 0$, and (iii) consider giving the two regime-separated analyses (which are physically cleaner) at least equal importance in the abstract and conclusions.

The e-folding depth s_0 = 191 m. As far as I understand, a single, gauge-independent value $s_0 = 191$ m is imposed for all submarine exponential fits. The procedure is not completely clear to me: was 191 m obtained from a global fit and then imposed, or were per-gauge s_0 values found to be mutually consistent? If imposed, please report the per-gauge best-fit s_0 (or a goodness-of-fit metric) so the reader can judge how well a single value describes all gauges. A short statement on the uncertainty of s_0 would also strengthen the "≈ 37 % per ≈ 200 m" claim.

Application to the 2002 event (rows 553–572). The way the 2002 event is used as a demonstration looks a bit "specious". The main reasons are the following. 1) Only one tide gauge was used for the reconstruction, which would be robust if the same results were obtained also for other coastal points for which observed values are available. 2) Using the gauge-8 submarine coefficient $\alpha_0 = 0.34$ with $s = 0$ (?) gives ~29 ×10⁶ m³; the average subaerial model gives ~84 ×10⁶ m³; the maximum subaerial coefficients give ~28 ×10⁶ m³; and the maximum submarine coefficient gives ~22 ×10⁶ m³, which is then singled out as being "in good agreement" with the accepted ~20 ×10⁶ m³. Please state explicitly which α and which submergence s are used in each estimate; using α_0 (a shoreline value) for a source that was submerged should *underestimate* the required volume, and this direction of bias should be stated. 3) The fact that the 2002 event was a submarine + subaerial event (~20–30 ×10⁶ m³ total) is not only based on tsunami reconstructions, but also on seismological analyses nad on post-event bathymetric and aerophotogrammetric surveys; hence, applying either a purely subaerial or a purely submarine model is an idealization that deserves a sentence of justification. (4) Rather than selecting the single combination that best matches the accepted figure, I recommend presenting the *full spread* of estimates (~22–84 ×10⁶ m³) as the honest predictive envelope and noting that it brackets the accepted value; this is a more convincing and honest demonstration of the tool's utility.

δ and m_f treated as one-dimensional ordinal axes. The friction is varied along a single ordered family of six configurations with $\delta_1 = \delta_3$ imposed (Sect. 2), and both δ and m_f enter the interpolation/Sobol framework as 1-D axes. The reported sensitivity to "δ" is therefore sensitivity along this one-parameter path through the three-angle space, not to friction in general; likewise, the $\delta_1 = \delta_3$ constraint should be recalled when interpreting the results. Please make this scope explicit where the δ sensitivity is discussed, so readers do not over-generalize.

Specific comments

- **Page 2, rows 49-50:** The conclusions by Tinti et al. (2006) are reported approximately. Tinti et al. (2006) concluded that the sequence of events generating the 30 December 2002 tsunami could be satisfactorily modelled through a submarine landslide with a volume of about 16×10^6 m³, followed by a subaerial landslide with a volume of about 5×10^6 m³, the two landslides being separated by a 7-min lag. The ratio between the inferred submarine and the subaerial volumes is perfectly coherent with the conclusions attributed to Fornaciai et al. (2019), which are instead referred to in the text as in contrast to the Tinti et al. (2006) results. The sentence must be rephrased accordingly.
- **Page 4, rows 100--101:** selecting three references (not clear based on which criterion) for the well-known fact that landslide volume plays a crucial role in determining the characteristics

of the ensuing tsunami appears highly disputable. To partially mitigate, please at least add "e.g." at the beginning of the list. The same applies to other references lists in the paper.

- **Page 4, row 104:** "(Fornaciai et al., 2019)" → "(Fornaciai et al., 2019 and references therein)"
- **Page 6, Figure 1.** What is the criterion adopted to select offshore gauges 4 and 5 more or less at the same Easting coordinate? Nothing against this, but maybe one of the two could have been chosen more to East to enhance the role of propagation.
- **Page 4, row 92-93 and Figures 2 and 3.** The authors correctly point out that the offshore gauges 4 and 5 show dispersive wave trains and, not coincidentally, the largest δ-sensitivity. Since a 20 m grid may under-resolve the shortest dispersive components, could part of the strong offshore δ-sensitivity be influenced by grid resolution rather than being purely physical? I encourage the authors to add a sentence addressing this, maybe after a resolution check at one of the two offshore gauges.
- **Page 7, rows 155-156:** Which interpolation method was used within `RegularGridInterpolator` ?. Was it linear? If so, linear interpolation of a coarse grid (6 volumes, 11 elevations) can bias variance estimates. Are the indices stable under a smoother interpolant? Please comment.
- **Page 15, σ_α and σ_χ definitions:** σ_α is defined as a normalized min–max half-range, not as a standard deviation, and is sensitive to a single extreme simulation; σ_χ is an RMS *relative* deviation, which up-weights small-volume points where *H* is small. Please rename/clarify these as relative-spread and relative-RMS metrics respectively and note their sensitivity to extremes/small-*H* points, to avoid a statistical over-interpretation.
- **Page 23, rows 460-462 (gauge 0 anomaly).** The peculiar *D* sensitivity at gauge 0, attributed to the submerged canyon mirroring the SdF incision, is plausible but asserted rather than shown. A short supporting argument (or reference to the bathymetry in Fig. 1) would help.
- **Table 3 and Figure 12 caption.** V_{ref} is written as "$3 \cdot 10^7$ m^3" here but as "30×10^6 m^3" in the main text. Please standardize the notation for the same quantity throughout the paper (I would suggest 30×10^6 m^3 everywhere for consistency with the volume axis).

"Technical" corrections

The English is generally clear; the following are minor corrections to be implemented.

- **Abstract, final sentence:** missing full stop after "…similar islands".
- **Page 2, row 35:** "Hunga Tonga-Hunga" → "Hunga Tonga-Hunga Ha'apai"
- **Page 2, row 49:** "the last, large tsunami occurred in 2002 was triggered" → "the last large tsunami, which occurred in 2002, was triggered".
- **Page 3, row 60:** "make Stromboli the principal locations …" → "make Stromboli the principal location …".
- **Page 5, row 131:** "timeseries" → "time series"
- **Page 6, caption to Figure 1:** white dots are numbered from 2 to 10, while gauges 0 and 1 are in red. So "(1, 2, 4, 5, 6, 7, 8, 9, and 10)" should be corrected into "(2 to 10)".
- **Page 12, row 294:** "V is the **utmost** influential parameter" → "the most influential parameter"
- **Pages 12-13, end of row 296-beginning of row 300:** these sentences are duplicated. Please remove them.
- **Page 13, row 306:** "even in" → "even though"
- **Page 17, row 409:** "the dependence of H on V **in** well described" → " the dependence of H on V **is** well described".

- **Figure captions (Figs. 5–10, S15–S24):** "These fits **provides** best values" → "These fits **provide** best values ".
- **p. 23, l. ~568:** "The comparison with the 2002 event **highlight**" → "highlights".
- Throughout: "**viceversa**" → "vice versa".

Supplementary material

- **Table S1:** the number format between Easting (X) and Northing (Y) differ. Please use the same format (I suggest avoiding the thousands separator).
- **Fig. S25 caption:** "**Pouliquenne**-Forterre" → "Pouliquen–Forterre" (also check the en-dash in "Pouliquen–Forterre" throughout).