

Dear Editor,

we have responded to all the points and comments raised by the two reviewers. Below we highlight, point by point, the changes made to the manuscript. Where we provide line numbers, they refer to the NOT TRACKED version of the revised manuscript.

Kind regards,
Laura Sandri on behalf of the authors.

REV #1

This manuscript develops a new database for eruptions at Mount Etna, separates it into flank and summit eruptions, and analyses the data to determine the best fitting model in the two cases. There are many examples of good practice in the analysis, including the testing of multiple hypotheses, the use of an appropriate information criterion to choose the best one, and the simulation of model variability and averaging by running an ensemble of models, not just using the best fit. The results show a clear difference between the two types of eruption, with summit eruptions showing clear evidence of short-term clustering, even in the primary data, and the best fit model for flank eruptions consistent with long-term anti-clustering indicating a renewal process with some memory. The results could in principle be used in managing the future risk of eruptions of the different types in operational forecasting mode for a time varying hazard rate, and some examples of this are given to illustrate this.

The paper is of a suitable standard for publication, and I recommend publication after addressing the following mostly minor points.

The abstract could be more specific about the example application to hazard forecasting.

Thank you for this comment. We added a sentence to be more specific on what the temporal model proposed allows in terms of PVHA.

Section 2.1. Can you say something about the uncertainty in dating events in such a long catalogue, and how this varies in time?

The catalogue we start from is publicly available on Zenodo at <https://doi.org/10.5281/zenodo.14284932>. There, the authors provide uncertainty bounds on the radio-dated events, but this is beyond the topic of the present paper. Furthermore, the age of the eruptions considered in the analysis (for completeness reasons, see below) are historically dated, so there is no uncertainty on the year of the eruptions.

Line 50. Can you also demonstrate and justify the statement of completeness and how this was determined? (In any case you change the completeness starting date later and explain why, so this is an initial guess, not the final version). In fact, the completeness is determined effectively by trial and error with stationarity as a criterion.

We thank the reviewer for this point, as the first version on the manuscript was unclear on this. At first, talking with volcanologists at Osservatorio Etneo, they told us to consider the catalogue since 1600 CE as complete (the so-called “historical” completeness in seismic hazard analysis). However, as our previous analysis in Table 1 (first version) showed, the catalogue since 1600CE was not stationary.

Thanks to this reviewer’s and the other one’s comments, we revisited the whole procedure. First, now we inspect the whole catalogue from Proietti and Branca (2023) since the year 1000 CE (as one can see from Sandri et al 2024, figure 1b, there is an eye-visible abrupt change in slope around that year). Then we run a change-point analysis to look for significant changes in the temporal trend of eruptions since 1000 CE. We find that the first-order change-point is exactly before 1763 eruption, so we take 1763 as the date of completeness under the assumption of stationarity.

We changed the manuscript accordingly, in particular we delete the sentence previously at line 50 and we explain the new completeness analysis in Section 3.1 and in the consequent results in Section 4.1. We also deleted the column corresponding to the flank eruptions since 1600 from Table 1.

Line 58. How was the catalogue merging done? Was this simply by combining these databases, or did it involve some recalibration and/or decisions on which to use in the case of overlapping events? Would you say this resulted in a homogeneous catalogue after the merger, and why?

We have added the following sentence in Section “2.2” to explain better how the catalogue was put together: “Specifically, for the few overlapping events among the different sources, our first choice-data are the results derived directly from radar measurements (Mereu et al., 2023). When this is not possible, we use data from the literature (in Calvari et al., 2018-2022 or, if not available there, in Andronico et al., 2021). This approach was adopted to standardize the dataset and ensure rigorous quality control over the information included in the catalogue.”

Line 70. ‘does not show evident clustering’. Where is the evidence for this statement? This assertion (in fact all unsupported assertions) should be justified by a logical narrative and cross-reference to the evidence. To identify clustering, we rely on the value of the Coefficient of Variation (CV), shown in Table 1. We have now added a reference to the table now at line 75.

Line 82. ‘due to their tendency to cluster’ - see previous comment. This assertion should be justified. See point above. We added here too a reference to CV and Table 1.

Line 93. Can you explain how the compound model works? Is this a sum of two different processes or is this a temporary change where one process switches off and another switches on?

The way we have modelled the data do not provide a direct switch from one regime to the other, as for example is done in Selva et al (2022) Science Advances. In our study, we use the two models to build synthetic catalogues by assuming that, while within a cluster, the cluster is over when a generated intra-cluster IET is “sufficiently” long to be very unlikely. In practice, as explained, we switch to the next cluster when the generated intra-cluster IET is longer than a high percentile of its best model’s pdf (we tried several values for this, and we obtain stable results if considering a cluster over when a generated intra-cluster IET is larger than 90 to 95th percentiles of its pdf; larger values make the procedure “diverge” as intra-cluster IETs become very often much longer than Inter-cluster IET, which should not be).

Line 121. Can you compare these annual rates with what you would get from a standard Poisson model? What is the probability gain on the forecasts compared to this null hypothesis?

For comparison, the standard Poisson model for flank eruption would provide an annual rate of 46 eruptions out of 260 years approximately (the completeness period), so a rate of circa 0.18. We have added for comparison the Poissonian hazard function in Figure 2 for the “rare” processes, i.e., for flank eruptions in panel a and for cluster onsets on panel b.

However, since the best model (BPT) has a time-dependent hazard function, this comment made us change our mind and we now prefer to provide the probability of at least one event in a given time interval (1, 3, 5, 10, 30 and 50 years), rather than the annual rate in Table 3.

We also provide a new panel (new panel b) in figure 2 showing the increase in probability as a function of the time window of exposure. Similarly, in the new panel d of that figure we show the expected number of summit paroxysms as a function of the exposure time, according to our model. Finally, for flank eruptions we also computed the probability gain of the preferred model over the Poissonian occurrence, for each event and for the whole dataset, as in Passarelli et al (2010) and Kagan and Knopoff (1987). The results are now shown in a new Figure and show that the intermediate IETs are those that make the Brownian Passage Time model outperform the Poisson model, leading to a global positive probability gain. This is now added in the Discussion section.

Lines 169-171. The evidence and chain of logic behind this inference on the hazard function is not presented, but it should be. Do you mean a steady increase, or a temporary increase and transient decay to

the background level after the cluster stops? I think you should cross-refer to the relevant figure for the hazard functions here for both types of eruption from the two variants of the BPT model, and compare and contrast these with each other rather than simply assert the inference. It would also be useful to add a time-independent rate to the relevant figure and discussion for comparison.

We thank the reviewer for this comment, who is similar to one from the other reviewer. Now we have expanded the description of the Brownian Passage Time model, we have added the reference to Figure 2 where we show the hazard functions of the preferred models explaining our data, and we have added in Figure 2 the hazard function for the Poisson model for comparison.

Figure 1. It would be good to see the incremental probability function here as well as the cumulative version. **We have added also the incremental probability function and the incremental distribution of the real data in figure 1, adding a right-side axis.**

Figure 3. There is clear evidence of clustering here, but you don't refer to this in the main text as evidence when you introduce the notion of clustering.

We added it.

Figure 4. I think it would be useful to also show the variability or uncertainty in the best fit curves (red) from the simulations, similar to Figure 1.

We have modified Figure 4 by adding 3 more panels, one for each ΔT previously shown. In each panel we show all the individual synthetic ECDFs, and the real one. The individual synthetic ECDFs provide a visual idea of the dispersion in the expected number of summit paroxysms, accounting for our temporal model's intrinsic uncertainty. We also run 4000 individual KS2 tests, of which very very few are rejected (but this was somehow expected, given the low number of data points in each individual ECDF).

Figure 8. Can you explain why the lower percentile curve is shorter than the upper one?

Yes, the values are too low and they are numerically zero. Since the y axis is logarithmic, these are treated as NaNs.

Ian Main, University of Edinburgh, 24 February 2026

REV #2

The paper aims to provide reliable statistical distributions of different styles of eruptive activity at Mt. Etna, an important topic for volcanic hazard analysis.

Overall I found the manuscript interesting and readable. However, several issues need to be addressed in a revised version.

- Completeness of the flank-eruption catalog. The manuscript is inconsistent about the catalog's completeness: line 50 states eruptions since 1600 are considered "for reasons of completeness," while line 110 claims completeness only since 1763. Beyond correcting this contradiction, the authors should clearly explain how they assessed completeness. Section 4.1 gives the impression that a cutoff date was chosen to produce a "coherent" distribution (e.g., starting from 1660 would yield a different distribution), which is not an appropriate justification. The long quiescence after the 1669 eruption might reflect a time-predictable behavior suggested in previous work (e.g., Marzocchi and Zaccarelli, JGR, 2006). That possibility is not necessarily inconsistent with the authors' results, but the paper needs a more thorough discussion of the completeness issue and its implications for the analysis.

We thank the reviewer for this point, as the first version on the manuscript was unclear on this. At first, talking with volcanologists at Osservatorio Etneo, they told us to consider the catalogue since 1600 CE as complete (the so-called "historical" completeness in seismic hazard analysis).

However, as our previous analysis in Table 1 (first version) showed, the catalogue since 1600CE was not stationary.

Thanks to this reviewer's and the other one's comments, we revisited the whole procedure. First, now we inspect the whole catalogue from Proietti and Branca (2023) since the year 1000 CE (as one can see from Sandri et al 2024, figure 1b, there is an eye-visible abrupt change in slope around that year). Then we run a change-point analysis to look for significant changes in the temporal trend of eruptions since 1000 CE. We find that the first-order change-point is exactly before 1763 eruption, so we take 1763 as the date of completeness, specifying that we determine completeness under the assumption of stationarity.

We changed the manuscript accordingly, in particular we delete the sentence previously at line 50 and we explain the new completeness analysis in Section 3.1 and in the consequent results in Section 4.1. We also deleted the column corresponding to the flank eruptions since 1600 from Table 1.

Also, we acknowledge in the Discussion in line 212-214 that cutting out the "unusually" long IET after the largest eruption occurred in 1669 CE (which is out of the stationary portion of the catalogue) might be real and may reflect the "generalized" (not linear) time-predictable behaviour suggested by Marzocchi and Zaccarelli, JGR, 2006 and Passarelli et al, 2010 JVGR, although those studies used a different and less revised catalogue. Based on the most recent available catalogue for flank eruptions at Etna (Proietti and Branca, published on Zenodo and used in Sandri et al, 2024 NHESS), the unusually long IET following that large eruption breaks the stationarity assumption and thus we removed it, suspecting a possible incompleteness in the aftermath of such a large eruption which destroyed Catania layout and may have disrupted the recording of Etna eruptions.

- AIC is known to favor models with more parameters. Here all distributions have two parameters except the exponential, which has one. Would the results change if you used AICc (the small-sample corrected AIC) to account for this bias?

We thank the reviewer for this important point. For both the cases with small sample sizes (flank eruptions, N=45; inter-cluster IETS, N=11) we also compute the AICc. Both AIC and AICc consistently identify the BPT models in both cases. We have now added this in text and tables.

- The use of the KS1 test may be inappropriate here. KS1 assumes the model parameters are known a priori, not estimated from the data; ignoring parameter estimation can substantially bias goodness-of-fit results. Use an alternative that accounts for fitted parameters (e.g., the Lilliefors modification for KS1 with normal or exponential distributions) or adjust KS via a parametric bootstrap to obtain correct p-values. Other robust options include the Anderson–Darling test, which IS LESS sensitive to this problem.

We thank the reviewer for this comment, which is actually a very important point. We have implemented the bootstrapped KS method proposed by Zeimbekakis, A., Schifano, E. D., & Yan, J. (2024). On Misuses of the Kolmogorov–Smirnov Test for One-Sample Goodness-of-Fit. *The American Statistician*, 78(4), 481–487. <https://doi.org/10.1080/00031305.2024.2356095>.

We find that the p-value for the bootstrapped KS method, based on 5000 bootstrap samples from the fitted Inverse Gaussian, is 0.154 so we do not reject the null hypothesis as the 90% confidence interval on p-values among the bootstrapped samples is 0.068-0.218. We also tested the same null hypothesis but considering the uncertainty on the values of the preferred model's parameters, with the same test but sampling only 350 pairs of the model parameter values, and for each pair we sample 1000 bootstrap. In such case, more than 50% of the sampled pairs achieve bootstrap KS test p-values larger than 1%.

We have added this modification in the text.

- The authors use the term "hazard function." While statistically correct, this can be confused with the "hazard curve" (often a survival or exceedance function). I suggest defining both terms explicitly when first used.

Ok, we add this sentence "Note that the term "hazard function" (also termed "hazard rate") defines the instantaneous probability of a flank eruption as a function of the time elapsed since the last one,

and should not be confused with the "hazard curve" (e.g., Tonini2015)" the first time we mention hazard function.

- Many volcanologists may be unfamiliar with the BPT distribution. I suggest citing Matthews et al. (BSSA, 2002) and referencing their Figure 2 to illustrate the distinction between a "periodic" BPT and a "clustered" BPT. This will help readers visualize and understand the different behaviors.

Thank you for this suggestion, we have added the reference to that Figure. Also, to follow the suggestion by the other reviewer, we have added a description of the Brownian Passage Time model in the beginning of the Discussion and added also the hazard function for the Poisson model for comparison.