Reviewer A Comments

This paper is on a timely and interesting topic in seismology and has the potential to be valuable to the community upon revision. The foreshock/mainshock classification b-value scheme presented by Gulia and Wiemer 2019 is a controversial topic that has been the subject of many papers over the last four years. The b-value calculation is dependent on the Mc which is difficult to estimate accurately. This paper uses a new Mc algorithm to test the results of Gulia and Wiemer 2019 (henceforth GW19). The concept of the paper has promise and is a good start. I have added comments on additional analyses and tests that the authors should consider to make for a more convincing case and ultimately stronger conclusions.

L56-76: Consider citing sources or examples for your statements on the deficiencies/problems for each Mc method.

L77-85 Be more specific on how the Mc value is calculated.
This is my impression of the Mc calculations:

1. Godano and Petrillo (2022) calculate Mc based on curves of $\Delta = m_a - (m_{th} + \log(e))$ vs m_{th}.
2. Godano (2023) calculates Mc based on $\langle b_n(m_{th})\rangle$ vs 1/n curves.
3. This paper calculates $C_v(m_{th})$ and selects Mc = m_{th} where $C_v(m_{th})$ is closest to 1 (or the value determined to be optimal via the synthetic tests presented).

This paper primarily cites Godano (2023) for the Mc method, so please explain briefly why method 2 is the same/similar or a derivative of method 3. As Godano (2023) is not yet published I think this paper should be clear and detailed about the procedure used. In my opinion you can leave the derivation to Godano (2023) but be clear about the practical steps taken so that it is reproducible.

Section 1.2 Synthetic Tests
L115: The description of Figure 14 provided here does not match the image, which appears to be identical to Figure 13.

Can you do ETAS synthetic tests for Mc calculated via the maximum curvature method as used by GW19? This will demonstrate how this Mc method performs compared to maximum curvature. If it performs better than maximum curvature the paper becomes much stronger.

Section 2.2 Blind Tests
L134 Do you impose any depth restrictions on the seismicity?

GW19 selects seismicity within a rectangular volume defined by empirical estimates of the fault plane dimension and not using a circular radius (GW19, Methods Step #3). Gulia and Wiemer have argued that tests of their method should use the same procedure (Gulia and Wiemer 2021, https://doi.org/10.1785/0220200428, particularly deviation 1.3). Please explain why you chose to change this.

Consider doing a second blind test with the setup made to be as close as possible to GW19. This would make your conclusions even stronger if you obtain the same results.

Discuss why the different result is a product of your improved Mc algorithm as opposed to using different spatiotemporal ranges for the selected seismicity. The authors write that the GW19 catalog selection is “peculiar” and “meticulously selected”. It is unclear whether this statement is due to the Mc method used by GW19, or by the spatial selection of seismicity. The paper focuses on discussing Mc, and therefore implies that these differences are caused by the new improved Mc algorithm. If the “peculiar” nature of the catalog is due to the spatial selection then this must be discussed.

Another deviation from GW19 is the use of M5.5-5.9 earthquakes, Gulia and Wiemer (2021) argues that these require limiting seismicity closer to the fault. I don’t think it’s crucial that you follow their guidance (since I believe they used the original method when looking at a 1995 M5.8 event in Gulia et al. 2020, https://doi.org/10.1785/0220190307) but you should at least point out that the results for the M6 events carry more weight.

L137: The post-earthquake values that you report appear to be medians and averages over a year. Can you also provide values calculated within shorter time intervals for all events? Particularly a result at a time that would fall between the Amatrice and Norcia earthquakes. You could make a time series figure to display this or add a table. There are two reasons for this: large foreshocks and aftershocks are usually closer than 1 year in time and b-value changes may recover over long time periods (e.g. GW19 Figure 1a). A single b-value averaged throughout that entire time period may smooth out the initial change in b-value that is seen in GW19 Figure 1a.

How does this Mc and b-value method perform immediately after the mainshock? Short-term aftershock incompleteness is an issue for the traffic light system (e.g. van der Elst 2021 and many others, https://doi.org/10.1029/2020JB021027). How early can you obtain accurate b-values and is there a way to determine that blindly in real time? What is the earliest data point included in the reported averages after the mainshock? GW19 excludes seismicity immediately following the mainshock for this reason.

You could also include an inset or more panels in Figures 17 and 18 that show the post-mainshock results in detail. I’d like to see how the b-value evolves in between the two M6+ earthquakes for Amatrice in the blind test similar to Figure 1a in GW19. I’d also like to see this for the others: perhaps there is an increase in b after the mainshocks that recovers? Right now these interesting time frames are represented by vertical lines on the far right of these plots: zoom in.

L149: “As can be seen the b values before and after the occurrence of the mainshcks cannot be considered different at a 95% significance level.”

1. How are you calculating this threshold? Are you assuming a normal distribution (if so justify it in the text) and multiplying σ_b by two?

2. Add another column to the table showing what the 95% significant threshold is to make it easier for your readers.
3. Consider mentioning this when describing your procedure for the blind test, it should be clear what constitutes a significant b-value change before presenting the results. This can also be applied to the b-value changes reported in section 2.1.
4. GW19 uses a 10% threshold: they issue the red alert if the b-value drops over 10% of the initial value. Explain why you choose to use the 95% significance level over adopting their procedure. Consider including the 10% comparison to address if GW19 would have considered these significant changes.

L150: “This implies that the discrimination between foreshocks and aftershocks cannot be performed using a blind algorithm.“
This (and the similar sentence in the conclusion) is an overstatement. The authors designed one blind algorithm, this does not rule out all blind algorithms. Add discussion on what the failure of this specific blind algorithm means.

L159: The authors call GW19’s catalog selection peculiar here, but then say it is appropriate later on (L161). This seems to be contradictory, please expand on this before the conclusion or in relation to my previous comment about GW19’s catalog selection.

You could mention that GW19 can also classify earthquakes as mainshocks via a significant b-value increase. While less exciting, this is another result for comparison, especially if using shorter time spans for the reported b-values leads to results with higher variability.

Figures
In the text you use periods to refer to decimals (e.g. 1.5) but in the figures, you use commas (e.g. 1,5). Please make it consistent.

Figures 1-3: Consider adding a dashed vertical line at 0. Please use a consistent x-axis range for the figures.

Figures 4-8: Consider removing some of these figures or moving some to a supplement. I don’t think they are different enough from each other to warrant 5 separate figures. Perhaps you pick one that is most representative or combine two into one figure, one on the small end and one on the large end. This also applies for Figures 9-13. Consider adding a colored horizontal line indicating the true values.

You could consider adding a histogram figure to summarize the ETAS results as a whole, perhaps report histograms of b_{true} - $\langle b_e \rangle$, where <> denotes the average value calculated over the time series. You could do something similar for Mc.

Figures 9-13: I believe the y-axis label should be Mce and not b. Consider adding a colored horizontal line indicating the true values.

Figure 14: This is identical to Figure 13, I think the authors included the wrong file.

Figure 15: Please be more clear in the figure legend and use punctuation before the numbers. E.g. “b value median before Amatrice: 1.00”. Please add to the caption that the “after” window is after Amatrice and before Norcia, I know you say this in the text but I think it should be repeated in the caption for clarity.

Figure 16: Please state what the two randomly chosen time windows were for this figure. Either in the legend or caption.

Figures 17-18: Please add a vertical line to denote the time of the mainshocks. Please provide an inset or figure that zooms in on the time period following the earthquakes to show the evolution of b and Mc after the earthquake, you could include the pre-event average b as a horizontal line.

Minor Comments

There is a lack of attention to detail throughout this paper (e.g. wrong axis labels, spelling errors, a missing figure). Please read through the paper and ensure there are no other issues.

L26: Change “space” to “spatial”
L36: Change “...of the b are…” to “...of the b value are…”
L66: Define Mcv. Also missing a period after ‘b’
L91: Spelling: “quantities”
L102: The peak in Figures 1-3 is at negative 0.25 not positive 0.25
L108: “...different values of b**,** Then we thin…”. Change the comma to a period.
L141: You write that Item 3 was only done one time, but I believe you meant to say Item 4. If this is not the case, more clarification is needed.
L149: Spelling: “mainshocks”

L171: The provided codes are not well commented, which would allow for easier reproducibility.

Reviewer B Comments

For author and editor

This is a very interesting paper in which the authors verify the reliability of a method for estimating the completeness earthquake magnitude m_c, through randomly simulated catalogs. Consequently, the reliability of the estimation of the b-value parameter in the Gutenberg-Richter frequency-magnitude distribution is assessed. The method is finally employed for the real-time discrimination between foreshocks and aftershocks in the Amatrice-Norcia (2016) and other Italian main sequences. Results highlight the need of a proper selection of the earthquake catalog, which is fundamental to properly detect the b-value parameter, and therefore to possibly interpret it as a stress indicator.

The method to estimate m_c is practically proposed in the paper by Godano et al. (2023), and is based on the variability coefficient c_v and on the quantity m-m_th (m_th is a threshold value for the magnitudes). It is here just briefly summarized. I think it is fair to specify that the work by Godano et al. (2023) is still submitted and, to my knowledge, not yet published. This could prevent a proper evaluation of this paper. However, I found a preprint relative to the submission by Godano et al. (2023), which I skimmed. I referred to this preprint for doing my review.

Overall, I really appreciated the authors' effort to test the reliability of the method proposed to estimate m_c and b. A proper and reliable estimation of these key parameters in earthquake forecasting, for example based on simulations like in this work, is fundamental to correctly interpret the results of any related statistical analysis. I appreciated also the application to real-time foreshock-aftershock discrimination, thus giving a contribution to the active debate in the literature about interpreting the temporal variation of the b-value as a proxy of the stress condition in the crust. Still, before acceptance, in my opinion the paper needs some revisions mainly aimed at investigating some points that could be relevant, and making the work more easily readable and self-consistent. I hope my comments, listed below, are helpful to the authors and produce some improvements in what is already an interesting and well-written paper.

<u>Main comments</u>

- The paper is developed by using some methods and techniques illustrated in other publications. They represent integral part of the current manuscript, therefore I strongly suggest to explain them with a higher detail. This will surely help the reader to go through the text. Among the methods/techniques adopted, the main one to discuss thoroughly is the new method to estimate m_c proposed in Godano et al. (2023). If I understood correctly from the preprint corresponding to the submitted work by Godano et al. (2023), the average value of b in a finite set of magnitudes is written as a function of c_v and the quantity mu = mean(m) - m_th, for magnitudes larger than the reference value m_th. Then, m_c is found based on the b-

value stabilization. I think this should be better clarified in the manuscript, as the authors here specify that m_th = m_c when b reaches a stable value, but then they introduce the quantity c_v which appears independent of b. Some details are necessary for an easier understanding.

- The reliability of the method to estimate m_c, and consequently b, is assessed by producing synthetic catalogs, and by evaluating for these two parameters the difference between the values used for simulation and the corresponding estimated values. One thing that appears not completely clear to me concerns the thinning procedure (Ogata and Katsura, 1993). In the "Introduction" section, the *entire magnitude range method* by Ogata and Katsura (1993) is illustrated as a method to estimate completeness; in the following sections where the simulation is discussed, this method is instead addressed as "thinning". Therefore, I guess that the 10^5 simulated catalogs are created by straightforwardly applying (1) without any magnitude of completeness, i.e. no "- m_th" is used in the formulation, and the results are catalogs of only synthetic magnitudes. Then, these catalogs are "thinned", that is, the number of magnitudes is reduced by considering only those above a given threshold, where the threshold is found by means of Ogata and Katsura (1993)'s method. Finally, the authors apply Godano et al. 2023's method to get the final estimation of m_c for the thinned catalogs. If this is true, this should be better clarified in the text, where a proper explanation of the meaning of thinning is required. Besides, I was wondering how would the results change when considering a different thinning technique.

- Including a detailed discussion about the thinning method, suggested above, would also help understanding the analysis relative to the ETAS simulated catalogs. Indeed, in the relative section the authors discriminate the results according to three levels of thinning, that is, the 5 ETAS simulated catalogs are thinned at 3 different values, which is the reason why they get 15 incomplete catalogs (this is evident from the figures but, for clarity, it is better to specify it also in the text, line 109). The authors should explain in a bit more detail what means to "thin at a certain level". I would also specify in the text that the 5 different values of b used to get the 5 ETAS catalogs are b = 0.6, 0.8, 1, 1.2, 1.4.

Since here the authors discuss the b-value temporal variation, I was curious to understand if the synthetic catalogs produced are purely temporal and if, also in this case, no completeness threshold is used to get the simulations.

I suggest including some additional comments about how the trend of the b-value in Figures from 4 to 8 changes with respect to the different m_c used, and the reason why in some cases (e.g. bottom panel of Figure 7) there are some gaps.

Figures 13 and 14 appear identical to me. I guess Figure 14 would contain the distribution p(b_e) as a function of the values assumed by its random variable b_e; however it is not clear

to me, neither from the text (lines 115-116), nor from the caption of Figure 14, if the authors use here a fixed b-value (1.4?) to generate different (how many?) catalogs, then they estimate for each catalog the b, and finally analyze again the difference between the fixed b-value and the estimated b-values. The authors also write "..confirming the tendency to underestimate b independently of the m_c values..", but a few lines above, b is said to be correctly estimated. Please, clarify a bit this part.

- The parameter c_vt seems to play an interesting role in the distribution of the quantity Delta m_c. Could this parameter be interpreted as a sort of "completeness variability coefficient", being the threshold after which m_th = m_c?

Since it is observed that p(Delta m_c) is correctly picked at Delta m_c = 0 when c_vt = 0.93, I am curious to understand if the overestimation of m_c in Figures 1 and 2 (Delta m_c ~= - 0.25) would be reduced when using for this figure c_vt = 0.93 instead of c_vt = 0.97.

Small differences of the parameter c_vt (of the order of 10^(-2)) seems to introduce a shift in the Delta m_c distribution's pick. Therefore, I think it may be worth discussing a bit more about the sensitivity with respect to c_vt. It is necessary to explain how the value used for this parameter to obtain Figures 1 and 2, has been fixed (as done instead for N=10 in line 99). I also suggest to add a sentence or some reference supporting the selection of the ranges for b and mu in lines 88 and 89. Eventually, for comparison with Figure 3, I suggest considering to add the case N = 10 in Figure 1.

To numerically support the evidences found in Figures from 1 to 3, I also suggest the authors to consider performing a statistical test for the distributions of Delta m_c and Delta b to be super-gaussian with 0 mean.

- The error on the m_c and b estimates could play a relevant role and surely allows us to correctly understand and interpret the results obtained. For example, in the analysis of the temporal variations of b, it may happen that the successive estimate falls into the confidence interval of the preceding one. I believe this is an important point the authors should address and discuss in the part relative to testing the method, as they have already done when estimating the b value before and after the Italian mainshocks.

- After testing its reliability, the method proposed to estimate m_c and b is applied to assess the b-value capability to discriminate in real-time between foreshocks and aftershocks in the case of the Amatrice-Norcia sequence, as previously done by Gulia and Wiemer (2019). Interestingly, the authors find some results supporting those in the latter publication; however, this part appears to me not completely "self-consistent". The sequence analyzed

should be described a bit for an uninformed reader, and a comparison between the method proposed and the technique by Gulia and Wiemer (2019) could help. Some authors in the literature have raised doubts about the paper by Gulia and Wiemer (2019), who replied with additional analyses for example in terms of the b-positive (van der Elst, 2021). Some comments about that are necessary to contextualize the results here.

Figure 15 presents a large hole from 2015 to after 2016.5, surely due to the lack of events in this area, which did not experience consistent seismic activity before the Amatrice earthquake, but this information should be added and explained in the text. The same holds for the values of 1, 0.82, 1.05 and 0.84 appearing in the legend of this figure.

I suggest the authors to consider zooming Figure 15 (or adding a panel with the zoom) only to the 2016.25 – 2017.5 period, because, as it is now, that part of the figure seems compressed.

- In Section 2.2, “A blind algorithm”, please specify the temporal interval considered in this analysis, and the occurrence time of the mainshocks identified. What would happen when considering different input parameters (for example, a shorter preceding period, or a different threshold for the number of events)?

Others

- *Pag. 1, line 24.* The references seem not to be in the proper format.

- *Pag. 2, line 49.* Since this is the title of a Paragraph, maybe it would be better to use the bold style.

- *Pag. 2, lines 60-61.* Some punctuation would be needed, with an adverb, e.g. “When *r* reaches its maximum, or a stable value, then m_th = m_c”.

- *Pag. 3, line 66.* Please delete “v” and add the period at the end.

- *Pag. 3, line 67.* Please correct in “…multiplies relationship 1 by…”.

- *Pag. 3, line 85.* Please correct in “Let us test…”.

- *Pag. 3, line 89.* Please correct in “…[1.5,2.5]…” (there is the misprint of a period).

- *Pag. 3, line 91.* Please correct in “…quantities…”.

- *Pag. 3, line 97.* Please change in “…larger than or equal to a given value of N…”.

- *Pag. 4, line 102.* Here it should be “Delta m_c ~= - 0.25”.

- *Figures 1 to 3.* I think it would be easier to use in the top panels of these three figures the same range for the y-axis. The same for the bottom panels. As regard the fact that, for N=300 and Delta m = 3 long tails in Delta b distribution are avoided (lines 103-104), this is not perfectly clear from the figures. I suggest to make the red markers the upper layer of the figures. Eventually, in the captions of these figures the authors could add the fixed values used in all the cases (e.g., c_vt = 0.97 in the caption of Figure 1), and the catalogs used (for example, in Figure 1, the catalogs used are those with a number of events larger than or equal to the different N specified in the legend, if proper).

- *Figure 3.* Please sort in descending order the legend for the c_vt values, and use the same font dimension for all items.

- *Pag. 4, lines 108-110.* The sentence “Then we thin…one at time” is difficult to read and understand; I suggest to rephrase it in something like “Then, as we did before, each catalog is thinned by means of the same technique used is Section 1.1, thus obtaining 15 incomplete time catalogs. The temporal variations of the b value are finally obtained by considering windows of 1000 events, sliding on one event at a time”. Besides, it is not clear if the 5 catalogs are simulated from the pure-temporal ETAS or the spatiotemporal ETAS (in this latter case, only the temporal part is considered).

- *Figures 9 to 13.* The proper y-axis label is m_c_e. Maybe, to see at a glance the overestimation, the authors may add in the different panels a thin red horizontal line in correspondence of the different m_c values used.

- *Caption of Figure 15.* Please correct in “…Gutenberg-Richter…” and specify the method applied to get the two estimates of 1.1 and 1.2.

- Figures 17 and 18 could be merged.

References

van der Elst, N. J. (2021). B-positive: A robust estimator of aftershock magnitude distribution in transiently incomplete catalogs. Journal of Geophysical Research: Solid Earth, 126(2), e2020JB021027.

Answers point-to-point

November 24, 2023

1 Report of Reviewer A

This paper is on a timely and interesting topic in seismology and has the potential to be valuable to the community upon revision. The foreshock/mainshock classification b-value scheme presented by Gulia and Wiemer 2019 is a controversial topic that has been the subject of many papers over the last four years. The b-value calculation is dependent on the Mc which is difficult to estimate accurately. This paper uses a new Mc algorithm to test the results of Gulia and Wiemer 2019 (henceforth GW19). The concept of the paper has promise and is a good start. I have added comments on additional analyses and tests that the authors should consider to make for a more convincing case and ultimately stronger conclusions.

We thank the reviewer for her/his comment and, below, we provide point-by-point answers to her/his comments and questions.

1) L56-76: Consider citing sources or examples for your statements on the deficiencies/problems for each Mc method.

A1) In the new version of the manuscript we cite the article of Mignan and Woessner (2012) reporting the limitation of the quoted methods with the exception of the Godano (2017) method. To our knowledge, there are no articles quoting the method limitation.

2) L77-85 Be more specific on how the Mc value is calculated. This is my impression of the Mc calculations: 1. Godano and Petrillo (2022) calculate Mc based on curves of $\Delta = ma-(mth+log(e))vsmth$. 2. Godano (2023) calculates Mc based on $\langle b_n(mth) > vs 1/n$ curves. 3. This paper calculates $Cv(mth)$ and selects $Mc = mth$ where $Cv(mth)$ is closest to 1 (or the value determined to be optimal via the synthetic tests presented). This paper primarily cites Godano (2023) for the Mc method, so please explain briefly why method 2 is the same/similar or a derivative of method 3. As Godano (2023) is not yet published I think this paper should be clear and detailed about the procedure used. In my opinion you can leave the derivation to Godano (2023) but be clear about

the practical steps taken so that it is reproducible.

A2) We thank the referee for his comment. For the estimation of m_c we employ the method from Godano et al. (2023) which introduces second order statistics observing that the variability coefficient c_v of an exponential distribution assumes a value equal to 1. Then, evaluating c_v as a function of a lower magnitude threshold m_{th} we can obtain $m_{th} = m_c$. In the revised version of the manuscript we added more details about the technique we have used for estimating m_c.

3) Section 1.2 Synthetic Tests L115: The description of Figure 14 provided here does not match the image, which appears to be identical to Figure 13.

A3) We apologise for the error, Figures 13 and 14 were loaded identical by mistake. In the revised version of the manuscript we replace the wrong Fig.14 with the correct one (new Fig.16).

4) Can you do ETAS synthetic tests for Mc calculated via the maximum curvature method as used by GW19? This will demonstrate how this Mc method performs compared to maximum curvature. If it performs better than maximum curvature the paper becomes much stronger.

A4) We thank the referee for this useful suggestion. In the new version of the manuscript we include the b_e distribution for the case b=1 and the different m_c when the maximum curvature method is used in estimating m_c. As expected b_e appears to be significantly different expecially for larger m_c values. The figure deos not include the other b values because it results very confused and adding new figures makes the paper more heavy.

5) Section 2.2 Blind Tests L134 Do you impose any depth restrictions on the seismicity?

A5) We do not impose depth restriction, we only removed both the mainshocks ($m = 5.8$ and $m = 5.9$) occurred in the Aeolian Arc at a depth larger than 144 Km.

6) GW19 selects seismicity within a rectangular volume defined by empirical estimates of the fault plane dimension and not using a circular radius (GW19, Methods Step 3). Gulia and Wiemer have argued that tests of their method should use the same procedure (Gulia and Wiemer 2021, https://doi.org/10.1785/0220200428, particularly deviation 1.3). Please explain why you chose to change this.

A6) In the new version of the manuscript we discuss this aspect and present the selected earthquakes maps in the SI. As stated in the revised manuscript, a circular radius does not affect the selection

of the aftershocks in respect to the fault plane. Differences could be observed for the background sesimicity. However, as discussed in the new version of the manuscript, Gulia and Wiemer themselves had to enlarge the background seismicity area in order to obtain a reliable b estimation. Moreover the circular radius represents the simplest method to be implemented in a blind algorithm.

7) Consider doing a second blind test with the setup made to be as close as possible to GW19. This would make your conclusions even stronger if you obtain the same results.

A7) We think that a second test with the same setup of GW19 cannot be included in a blind algorithm. Indeed or we know the fault plane before starting the algorithm or we chose a rectangular area centered on the mainshock location. The first option is not the easiest blind algorithm and the second does not changes our results (see point A6).

8) Discuss why the different result is a product of your improved Mc algorithm as opposed to using different spatiotemporal ranges for the selected seismicity. The authors write that the GW19 catalog selection is "peculiar" and "meticulously selected". It is unclear whether this statement is due to the Mc method used by GW19, or by the spatial selection of seismicity. The paper focuses on discussing Mc, and therefore implies that these differences are caused by the new improved Mc algorithm. If the "peculiar" nature of the catalog is due to the spatial selection then this must be discussed.

A8) In the revised version of the manuscript conclusions we better remark that a more reliable method of estimating m_c and b makes stronger the confirmed Gulia and Wiemer results. Moreover, we smoothed our statements about the earthquake selection: now it is defined specific and no peculiar.

9) Another deviation from GW19 is the use of M5.5-5.9 earthquakes, Gulia and Wiemer (2021) argues that these require limiting seismicity closer to the fault. I don't think it's crucial that you follow their guidance (since I believe they used the original method when looking at a 1995 M5.8 event in Gulia et al. 2020, https://doi.org/10.1785/0220190307) but you should at least point out that the results for the M6 events carry more weight.

A9) We agree with the referee that this difference has not been highlighted well. In the revised version of the manuscript we have introduced a sentence clarifying this aspect.

10) L137: The post-earthquake values that you report appear to be medians and averages over a year. Can you also provide values calculated within

shorter time intervals for all events?Particularly a result at a time that would fall between the Amatrice and Norcia earthquakes. You could make a time series figure to display this or add a table. There are two reasons for this: large foreshocks and aftershocks are usually closer than 1 year in time and b-value changes may recover over long time periods (e.g. GW19 Figure 1a). A single b-value averaged throughout that entire time period may smooth out the initial change in b-value that is seen in GW19 Figure 1a.

A10) The suggestion of the referee is very interesting. However, when we try to apply it, the b time variation fills of holes because $N < 150$ or $\Delta m < 2$. Probably, for shorter time intervals, we need less restrictive criteria in the analysis. This could be matter of further investigation requiring too much more work and a significant enlarging of the manuscript.

11) How does this Mc and b-value method perform immediately after the mainshock? Short-term aftershock incompleteness is an issue for the traffic light system (e.g. van der Elst 2021 and many others, https://doi.org/10.1029/2020JB021027). How early can you obtain accurate b-values and is there a way to determine that blindly in real time? What is the earliest data point included in the reported averages after the mainshock? GW19 excludes seismicity immediately following the mainshock for this reason.

A11) We thank the referee for his question. We are aware that it is important to make an assessment of b and m_c as quickly as possible after the occurrence of a target event. In the explanation of the blind test algorithm we write that at least 500 events are needed to perform the b value estimation. If there are not at least 500 events, the sequence is discarded from the study. This assumption could be relaxed and accept the value estimation even with a lower number of aftershocks. With regard to the earliest data point included in the calculation, we modified, in the revised version of the manuscript, the blind test by introducing the issue of short term aftershock incompleteness (STAI) employing the functional form proposed by Helmstetter et al. (2006).

12) You could also include an inset or more panels in Figures 17 and 18 that show the post-mainshock results in detail. I'd like to see how the b-value evolves in between the two M6+ earthquakes for Amatrice in the blind test similar to Figure 1a in GW19. I'd also like to see this for the others: perhaps there is an increase in b after the mainshocks that recovers? Right now these interesting time frames are represented by vertical lines on the far right of these plots: zoom in.

A12) Done

13) L149: "As can be seen the b values before and after the occurrence of the mainshcks cannot be considered different at a 95% significance level." 1. How are you calculating this threshold? Are you assuming a normal distribution (if so justify it in the text) and multiplying σ_b by two?

A13.1) In the new version of the manuscript we clarify that we are performing a t-test.

2. Add another column to the table showing what the 95% significant threshold is to make it easier for your readers.

A13.2) The t-test is standard. We do not think that adding its threshold value is necessary.

3. Consider mentioning this when describing your procedure for the blind test, it should be clear what constitutes a significant b-value change before presenting the results. This can also be applied to the b-value changes reported in section 2.1.

A13.3) See point before.

4. GW19 uses a 10% threshold: they issue the red alert if the b-value drops over 10% of the initial value. Explain why you choose to use the 95% significance level over adopting their procedure. Consider including the 10% comparison to address if GW19 would have considered these significant changes.

A13.4) We are not following a traffic light approach. Our main goal is to show that, improving the m_c estimation, the GW19 results are confirmed.

14) L150: "This implies that the discrimination between foreshocks and aftershocks cannot be performed using a blind algorithm." This (and the similar sentence in the conclusion) is an overstatement. The authors designed one blind algorithm, this does not rule out all blind algorithms. Add discussion on what the failure of this specific blind algorithm means.

A14) In the new version of the manuscript we smoothed our observation and state that our blind algorithm doesn't work and that other could.

15) L159: The authors call GW19's catalog selection peculiar here, but then say it is appropriate later on (L161). This seems to be contradictory, please expand on this before the conclusion or in relation to my previous comment about GW19's catalog selection.

A15) In the new version of the manuscript we define the GW19

selection specific which is not contradictory with appropriate. Sorry for the mistake.

16) You could mention that GW19 can also classify earthquakes as mainshocks via a significant b-value increase. While less exciting, this is another result for comparison, especially if using shorter time spans for the reported b-values leads to results with higher variability.

A16) We agree with the referee this is less exciting. Moreover we are planning to investigate this aspect, with particular attention to STAI, in a future paper.

2 Report of Reviewer B

This is a very interesting paper in which the authors verify the reliability of a method for estimating the completeness earthquake magnitude m_c, through randomly simulated catalogs. Consequently, the reliability of the estimation of the b-value parameter in the Gutenberg-Richter frequency-magnitude distribution is assessed. The method is finally employed for the real-time discrimination between foreshocks and aftershocks in the Amatrice-Norcia (2016) and other Italian main sequences. Results highlight the need of a proper selection of the earthquake catalog, which is fundamental to properly detect the b-value parameter, and therefore to possibly interpret it as a stress indicator. The method to estimate m_c is practically proposed in the paper by Godano et al. (2023), and is based on the variability coefficient c_v and on the quantity $m - m_t h$ ($m_t h$ is a threshold value for the magnitudes). It is here just briefly summarized. I think it is fair to specify that the work by Godano et al. (2023) is still submitted and, to my knowledge, not yet published. This could prevent a proper evaluation of this paper. However, I found a preprint relative to the submission by Godano et al. (2023), which I skimmed. I referred to this preprint for doing my review. Overall, I really appreciated the authors' effort to test the reliability of the method proposed to estimate m_c and b. A proper and reliable estimation of these key parameters in earthquake forecasting, for example based on simulations like in this work, is fundamental to correctly interpret the results of any related statistical analysis. I appreciated also the application to real-time foreshock-aftershock discrimination, thus giving a contribution to the active debate in the literature about interpreting the temporal variation of the b-value as a proxy of the stress condition in the crust. Still, before acceptance, in my opinion the paper needs some revisions mainly aimed at investigating some points that could be relevant, and making the work more easily readable and self-consistent. I hope my comments, listed below, are helpful to the authors and produce some improvements in what is already an interesting and well-written paper.

We thank the reviewer for appreciating our work and defining it interesting and well-written. Below, we provide point-by-point answers to his comments and questions.

1) The paper is developed by using some methods and techniques illustrated in other publications. They represent integral part of the current manuscript, therefore I strongly suggest to explain them with a higher detail. This will surely help the reader to go through the text. Among the methods/techniques adopted, the main one to discuss thoroughly is the new method to estimate m_c proposed in Godano et al. (2023). If I understood correctly from the preprint corresponding to the submitted work by Godano et al. (2023), the average value of b in a finite set of magnitudes is written as a function of c_v and the quantity $mu = mean(m) - m_t h$, for magnitudes larger than the reference value m_{th}. Then, m_c is found based on the b-value stabilization. I think this should be better clarified in the manuscript, as the authors here specify that $m_{th} = m_c$ when b reaches a stable value, but then they introduce the quantity c_v which appears independent of b. Some details are necessary for an easier understanding.

A1) We thank the referee for his comment. We agree with the referee that some details regarding the method used should be introduced in the main text. The variability coefficient assumes values $c_v \simeq 1$ when $m_{th} = m_c$, i.e., when the distribution become a pure exponential. Therefore, evaluating c_v as a function of the threshold magnitude m_{th} allows us to identify when b reaches a stable value. In the revised version of the manuscript we have added, after Eq.(2), a more accurate explanation of the method.

2) The reliability of the method to estimate m_c, and consequently b, is assessed by producing synthetic catalogs, and by evaluating for these two parameters the difference between the values used for simulation and the corresponding estimated values. One thing that appears not completely clear to me concerns the thinning procedure (Ogata and Katsura, 1993). In the "Introduction" section, the entire magnitude range method by Ogata and Katsura (1993) is illustrated as a method to estimate completeness; in the following sections where the simulation is discussed, this method is instead addressed as "thinning". Therefore, I guess that the 10^5 simulated catalogs are created by straightforwardly applying (1) without any magnitude of completeness, i.e. no "- m_{th}" is used in the formulation, and the results are catalogs of only synthetic magnitudes. Then, these catalogs are "thinned", that is, the number of magnitudes is reduced by considering only those above a given threshold, where the threshold is found by means of Ogata and Katsura (1993)'s method. Finally, the authors apply Godano et al. 2023's method to get the final estimation of m_c for the thinned catalogs. If this is true, this should be better clarified in the text, where a proper explanation of the meaning of thinning is required. Besides, I was wondering how would the results change when considering a different thinning technique.

A2) We thank the referee for the comment and confirm the correct understanding. We numerically simulated 10^5 catalogues with different b-values. These catalogues are complete, i.e., contain all the events coherent with the distribution (1). Then, we remove events from the 10^5 simulated catalogues by means of Ogata and Katsura (1993)'s method. Next, for each catalogue, we estimate m_{c_e} and b_e using the Godano et al. (2023) method and evaluate the quantities $\Delta b = b - b_e$ and $\Delta m_c = m_c - m_{c_e}$. To better clarify the approach, we have also explained better the meaning of thinning and the procedure in the section "randomly simulated catalogues". To our knowledge, there are no other thinning techniques.

3.1) Including a detailed discussion about the thinning method, suggested above, would also help understanding the analysis relative to the ETAS simulated catalogs. Indeed, in the relative section the authors discriminate the results according to three levels of thinning, that is, the 5 ETAS simulated catalogs are thinned at 3 different values, which is the reason why they get 15 incomplete catalogs (this is evident from the figures but, for clarity, it is better to specify it also in the text, line 109). The authors should explain in a bit more detail what means to "thin at a certain level". I would also specify in the text that the 5 different values of b used to get the 5 ETAS catalogs are b = 0.6, 0.8, 1, 1.2, 1.4. Since here the authors discuss the b-value temporal variation, I was curious to understand if the synthetic catalogs produced are purely temporal and if, also in this case, no completeness threshold is used to get the simulations. I suggest including some additional comments about how the trend of the b-value in Figures from 4 to 8 changes with respect to the different m_c used, and the reason why in some cases (e.g. bottom panel of Figure 7) there are some gaps.

A3.1) We agree with the referee's suggestions regarding the need to clarify these points. In particular, the synthetic catalogues were generated using the space-time ETAS model to enhance their realism, even though we did not incorporate space information in the analysis of the ETAS simulated catalogues. In the revised version of the manuscript, we have included a comprehensive explanation of the employed ETAS model and provided details on the parameter optimization process. As the referee correctly noted, the ETAS catalogues are complete, and we did not employ a specific m_c (magnitude of completeness) value in their creation. They were generated with five different values of b. Subsequently, we applied Ogata and Katsura's method (1993) to thin each of the five catalogues, employing three different m_c values for each. This resulted in a total of 15 thinned (incomplete) catalogues. The gaps present in the figures are related to lack of data in the considered time windows. In particular, the required number of events N is greater than the number of

events recorded in the numerical catalogue. We have added the new information to the revised version of the manuscript.

3.2) Figures 13 and 14 appear identical to me. I guess Figure 14 would contain the distribution $p(b_e)$ as a function of the values assumed by its random variable b_e; however it is not clear to me, neither from the text (lines 115-116), nor from the caption of Figure 14, if the authors use here a fixed b-value (1.4?) to generate different (how many?) catalogs, then they estimate for each catalog the b, and finally analyze again the difference between the fixed b-value and the estimated b-values. The authors also write "..confirming the tendency to underestimate b independently of the m_c values..", but a few lines above, b is said to be correctly estimated. Please, clarify a bit this part.

A3.2) We apologise very much for the error, Figures 13 and 14 were loaded identical by mistake. Also, the y-label of figures 9-13 in the old manuscript are wrong. Those figures represented the time variation of m_{c_e}, the estimated completeness magnitude, for the 5 different values of b employed in the ETAS simulations (different figures) and for the 3 different values of thinning employed after the ETAS simulations (different panels for each figure). The old fig.14 of the manuscript (missing by mistake in the old version of the manuscript, and the new fig.16) represents the distribution of b_e estimated for different catalogues thinned at different m_c values for each value of b. In particular, for each numerical catalog simulated with a certain b-value, we estimate the b_e to obtain a distribution. Then, we repeated this calculus for different value of thinning m_c. We have fixed all these points in the new version of the manuscript.

4.1) The parameter c_{vt} seems to play an interesting role in the distribution of the quantity Δm_c. Could this parameter be interpreted as a sort of "completeness variability coefficient", being the threshold after which $m_{th} = m_c$?

A4.1) Since the completeness is chosen after the stabilization of the c_v value we agree that is possible to interpret c_{vt} as a "completeness variability coefficient". In the new version of the manuscript we better clarify this aspect.

4.2) Since it is observed that $p(\Delta m_c)$ is correctly picked at Delta $m_c = 0$ when $c_{vt} = 0.93$, I am curious to understand if the overestimation of m_c in Figures 1 and 2 ($Deltam_c = -0.25$) would be reduced when using for this figure $c_vt = 0.93$ instead of $c_vt = 0.97$.

A4.2) We agree with the referee, indeed the old Fig 3 was exactly that. We hope that in the new version of the manuscript this is more clear.

4.3) Small differences of the parameter c_{vt} (of the order of 10^{-2}) seems to introduce a shift in the Δm_c distribution's pick. Therefore, I think it may be worth discussing a bit more about the sensitivity with respect to $c_v t$. It is necessary to explain how the value used for this parameter to obtain Figures 1 and 2, has been fixed (as done instead for N=10 in line 99). I also suggest to add a sentence or some reference supporting the selection of the ranges for b and mu in lines 88 and 89. Eventually, for comparison with Figure 3, I suggest considering to add the case N = 10 in Figure 1.

A4.3 As the referee correctly noted, the small differences of the parameter c_{vt} of the order of 10^{-2} generate a little shift in the Δm_c distribution pick. This is clear from the $p(\Delta m_c)$ distributions for different values of c_{vt} revealing that the c_{vt}=0.93 is the correct value. In the new version of the manuscript we more clearly enlighten that this is a sensitivity analysis of c_{vt}. The case N=10 is not presented because a reliable evaluation of b using 10 events is not possible. Indeed it was a mistake in the old version of the manuscript. In the present version 10 is corrected to 100.

4.4) To numerically support the evidences found in Figures from 1 to 3, I also suggest the authors to consider performing a statistical test for the distributions of Δm_c and Δb to be super-gaussian with 0 mean.

A4.4) We thank the referee for the useful suggestion. In the new version of the manuscript, we have included a comparison with a Gaussian distribution in Figure 3. Additionally, by conducting a Kolmogorov-Smirnov test, it is possible to statistically reject the hypothesis of a normal distribution.

5) The error on the m_c and b estimates could play a relevant role and surely allows us to correctly understand and interpret the results obtained. For example, in the analysis of the temporal variations of b, it may happen that the successive estimate falls into the confidence interval of the preceding one. I believe this is an important point the authors should address and discuss in the part relative to testing the method, as they have already done when estimating the b value before and after the Italian mainshocks.

A5) We agree with the referee that error assessment is of paramount importance since it may happen that the subsequent estimate falls into the confidence interval of the preceding one. However, as the graphs are very dense, plotting error bars is complicated. In any case, we have added a small discussion of the error in the revised version of the manuscript.

6.1) After testing its reliability, the method proposed to estimate m_c and b is applied to assess the b-value capability to discriminate in real-time between foreshocks and aftershocks in the case of the Amatrice-Norcia sequence, as pre-

viously done by Gulia and Wiemer (2019). Interestingly, the authors find some results supporting those in the latter publication; however, this part appears to me not completely "self-consistent". The sequence analyzed should be described a bit for an uninformed reader, and a comparison between the method proposed and the technique by Gulia and Wiemer (2019) could help. Some authors in the literature have raised doubts about the paper by Gulia and Wiemer (2019), who replied with additional analyses for example in terms of the b-positive (van der Elst, 2021). Some comments about that are necessary to contextualize the results here.

A6.1) We agree with the referee that in order to make the section self-consistent, it is necessary to introduce the results of Gulia and Wiemer (2019) and especially the discussion on the predictive power of the b-value. In the revised version of the manuscript, we have also introduced some results from the literature that questioned Gulia and Wiemer (2019) paper. Finally, we introduced our forecasting proposal by means of a blind test. In this way, we are confident that it is easier to compare the methodologies employed.

6.2) Figure 15 presents a large hole from 2015 to after 2016.5, surely due to the lack of events in this area, which did not experience consistent seismic activity before the Amatrice earthquake, but this information should be added and explained in the text. The same holds for the values of 1, 0.82, 1.05 and 0.84 appearing in the legend of this figure. I suggest the authors to consider zooming Figure 15 (or adding a panel with the zoom) only to the 2016.25 – 2017.5 period, because, as it is now, that part of the figure seems compressed.

A6.2) We agree with the referee. Therefore in the revised version of the manuscript we added a zoom inset from the time period 2016.5 to 2017.5.

7) In Section 2.2, "A blind algorithm", please specify the temporal interval considered in this analysis, and the occurrence time of the mainshocks identified. What would happen when considering different input parameters (for example, a shorter preceding period, or a different threshold for the number of events)?

A7) In the revised version of the manuscript we added a Table including all the relevant information about the mainshocks selected in the blind test. As from the algorithm item 3 and item 5, we select 4 years before and 1 year following the occurrence of the target mainshock. Moreover, following the referee suggestion, we conducted a blind test by considering a half-period preceding and following the event target. While this adjustment did not significantly impact the results, we have incorporated this additional analysis in the revised version of the manuscript.

8) Others.

8A) We thank the referee for pointing out minor comments typos and various corrections to the English. We have implemented all the proposed changes.

Reviewer A Comments

I thank the authors for their detailed responses to my questions. They have addressed my concerns well. Going through the manuscript again I came up with some additional comments and suggestions, all of which are fairly minor.

L48: It is probably better to describe Mc as the lowest magnitude where EQs are reliably recorded, not necessarily all.

Thank you for adding the details on the Mc procedure. I think you are missing one more piece of information in the methods section: that you choose mth to be the smallest mth where cv is larger than the threshold (cvt). This makes what the authors say about Figure 1 easier to understand. By defining cvt in detail here you could delete the phrase in L111 saying "the cv threshold after which mth = mc", since the improved methodological section makes that phrase no longer necessary.

Would the b-value changes for the GW19 test in section 3.1 be a statistically significant change in b according to the t-test? Or are you just confirming that b went down? If it is the statistically significant change it would be good to mention that.

L232 You refer to epicentral maps here. In the response to reviewers, you stated that these maps were included in the Supplementary Information, however no SI was provided. If there will be a supplement for the final publication you could mention where to find these maps there.

If you are planning a supplement, I think you could consider moving some of the repetitive figures into that supplement. You probably only need one or two of the Figures 11-16 and same for 6-10.

In the response, the authors stated that reporting results at times closer to the blind test mainshocks is beyond the scope of this paper, which is primarily focused on evaluating the GW19 Amatrice-Norcia b-value changes. I see you did report results for a 6 month time frame, which is useful. I do think that reporting b-value averages/medians and assessing significance at shorter post-mainshock times is important to assess whether the b-value is predictive. However, I will defer to the authors on this, since the paper is primarily evaluating the new mc method and applying it, rather than proposing a new blind algorithm for assessing these b-value changes. That being said I think you should consider explaining in the text why you aren't doing averages/medians/significance tests at 5, 10 days, 30 days, 2 months, etc.

You did calculate the average/median b-values between Amatrice and Norcia, which were roughly 2 months apart. So I assume the reason you got a lot of holes in the blind test was because the other mainshocks had a lack of data in that time frame?

You did report results for a 6 month time window in the blind tests. What is the smallest you could make the post-event time window without getting too many gaps in the time series? Are those results any different?

Thank you for addressing the deviations from GW19, I know you added that the larger earthquakes carry more weight (L227-228) upon my review. After adding the new sentence before it in L226, you probably don't need the "carry more weight" sentence or should reword it to be more specific.

With regards to the t-test, you could make an argument as to whether or not that approach is more appropriate than the arbitrary +/- 10% used in GW19 to determine significance if you wish to.

In Table 3, why are there 6 and not 7 events? Can you add that information around L240?

Some figures use commas to denote decimals and some use periods. Please make this consistent. You use periods throughout the text so I would recommend that.

Many figures have y-labels that are partially aligned with the y-axis values. For example, Figures 2, 4, 5, etc. It would look neater if they did not overlap, similar to Figure 1, 16, 17, etc.

The multi-panel figures have a lot of space between panels that should be reduced. You could also consider labeling the panels a and b.

Specific Figure Comments

Figure 1: Consider adding a vertical line at the true Mc value or perhaps filling in or coloring the dot that would have been selected as the best mth. In the caption, mc does not have a subscript for the "c". I would also encourage you to specify what the b and mc are in the caption rather than just saying that they are reported in the figure.

Figure 3: I think the new insets on the top panel make it too complicated. I think one would be ok, perhaps one (or both) could go into a supplement.

Figures 12 and 14: The ylabel ("mce") has a white background that cuts off the y-axis label "2.5" by just a little.

Figure 16 and 17: Consider using different shapes or different colors, could make the figure more readable to those with red-green colorblindness.

Figure 18: I'm not sure what the superscript "B" in the "b_{MA}^{B}" is. I think you can remove the arrows, they cover up some data and make the figure messier.

Figure 19: I think it would be interesting to know what the two random times are.

Figures 20 and 21: Add markers to denote the time of the mainshocks. Could include the pre-event average b as a horizontal line. Please tidy up the location labels so they are placed in the same relative place. For example, the word Amatrice is really low on the plot compared to the rest and Mirandola is nearly hitting the tick marks.

Other

L19: You use the shorthand "GR" for Gutenberg and Richter later in the paper, so consider defining it here or somewhere else before then. E.g. "...Gutenberg and Richter (GR) law…"
L33: The Pino et al citation has no year
L40: Insert comma after "could be biased", so it reads "...could be biased, causing a biased estimation…"
L43: Move the comma after "and" to after "reliability". "...verify its reliability, and consequently, the reliability of…"
L58: Same as L43, move comma before the "and"
L67: Consider re-writing the sentence as "Although the method presents some advantages, in this case the instability of r can also produce biased estimation of mc and b."
L69: Say Equation 1 or Eqn 1, not just "1"
L89 Instead of "value 1" say "value of 1"
L91 Spelling: "Occurre" should be "occur"
L92: Say "let us call it" rather than "let we call it"
L103: I'd reword as "...experimental catalogues containing some, though not all, events with m < mc."
L104: Say "parameters have been selected" rather than "parameters has been selected"
L118: Delete "from", and I think it should be Figures 3 to 5 instead of 3 to 4
L120: Add space in "Fig.3" and Spelling, "hypotesis" should be "hypothesis"
L123: Consider rewording "...or to a small magnitude range." to "...or a small magnitude range" or "...or too small of a magnitude range."
L124: A couple grammatical mistakes, consider rewording as something like: "The overestimation of mc when cvt=0.97 suggests that this parameter also influences our results."
L125: Add space "Fig.5", occurs in other places too.
L128: I think you should add what ETAS stands for, so the first sentence could be "The Epidemic Type Aftershock Sequence (ETAS) model represents…"

L139: The semicolon after Zhuang et al. (2004) should be a comma, so “(2004),” instead of “(2004);”. Also add a comma in the similar place after Zhuang and Touati (2015)
L145: Spelling: “off-springs” should be “offspring”
L147: Spelling: “Ricther” -> “Richter”
L157: Add spaces between “delta m = 2” to be consistent with the other two equalities in the sentence.
L166 and other places, “Fig.s” should just be “Figs.”
L189 Here you say “b-value” rather than “b value”. I think either is ok, but should be consistent throughout
L190 Add a “of”, e.g. “...attributed to a mixture of inconsistencies…”
L206: Too many commas, could change to something like “Now we test an…” or “Let us now test an…”
L224: Add comma, “The information about the location, magnitude, and occurrence time…”
L227: make “M5.5 - M5.9 earthquake” -> “M5.5 - M5.9 earthquakes”.
L235: Add comma after b_M
L250: too many commas, perhaps “Using the cv method we then tested…”

Reviewer B Comments

-Pag. 3, line 91. Please correct in “…it can occur…”.

-Pag. 5, line 120. I would suggest to report either here or in the figure the p-value resulted from the Kolmogorov-Smirnov test, and the significance level considered.

-Caption of Figure 1. Please use for “mc” the same notation of the text.

-Pag. 9, line 142. Please correct in “Each of these elements…”.

-Pag. 9, line 145. Please correct in ““…For each off-spring…”.

Recommendation: Accept Submission

Answers point-to-point v2

January 7, 2024

1 Report of Reviewer A

We thank the reviewer for appreciating our response to previous reviewing. Please find, in the following, a point to point answer to the reviewer observations.

The referee writes
L 48: It is probably better to describe Mc as the lowest magnitude where EQs are reliably recorded, not necessarily all.

Answer: We thank the reviewer for the observation. We changed the new version of the manuscript accordingly.

The referee writes
Thank you for adding the details on the Mc procedure. I think you are missing one more piece of information in the methods section: that you choose mth to be the smallest mth where cv is larger than the threshold (cvt). This makes what the authors say about Figure 1 easier to understand. By defining cvt in detail here you could delete the phrase in L111 saying "the cv threshold after which mth = mc", since the improved methodological section makes that phrase no longer necessary.

Answer: We thank the reviewer for the observation. We changed the new version of the manuscript accordingly.

The referee writes
Would the *b*-value changes for the GW19 test in section 3.1 be a statistically significant change in b according to the t-test? Or are you just confirming that b went down? If it is the statistically significant change it would be good to mention that.

Answer: We are just confirming that b went down without performing any test. The result is already published on a very prestigious review.

The referee writes
L 232 You refer to epicentral maps here. In the response to reviewers, you stated that these maps were included in the Supplementary Information, however no SI was provided. If there will be a supplement for the final publication you could mention where to find these maps there. If you are planning a supplement, I think you could consider moving some of the repetitive figures into that supplement. You probably only need one or two of the Figures 11-16 and same for 6-10.

Answer: We thank very much the reviewer for enlightening our mistake. We forgot to upload the supplementary information. We have corrected the mistake.

The referee writes
In the response, the authors stated that reporting results at times closer to the blind test mainshocks is beyond the scope of this paper, which is primarily focused on evaluating the GW19 Amatrice-Norcia b-value changes. I see you did report results for a 6 month time frame, which is useful. I do think that reporting b-value averages/medians and assessing significance at shorter post-mainshock times is important to assess whether the b-value is predictive. However, I will defer to the authors on this, since the paper is primarily evaluating the new mc method and applying it, rather than proposing a new blind algorithm for assessing these b-value changes. That being said I think you should consider explaining in the text why you aren't doing averages/medians/significance tests at 5, 10 days, 30 days, 2 months, etc.

Answer: We agree with the reviewer, however the use of shorter time windows is too much difficult because the gaps become dominant leading to a not significant evaluation of averages and medians.

The referee writes
You did calculate the average/median b-values between Amatrice and Norcia, which were roughly 2 months apart. So I assume the reason you got a lot of holes in the blind test was because the other mainshocks had a lack of data in that time frame? ou did report results for a 6 month time window in the blind tests. What is the smallest you could make the post-event time window without getting too many gaps in the time series? Are those results any different?

Answer: See the previous point.

The referee writes Thank you for addressing the deviations from GW19, I know you added that the larger earthquakes carry more weight (L227-228) upon my review. After adding the new sentence before it in L226, you probably don't need the "carry more weight" sentence or should reword it to be more specific. With regards to the t-test, you could make an argument as to whether or not that approach is more appropriate than the arbitrary +/- 10% used in GW19 to determine significance if you wish to.

Answer: Of course the t-test is more appropriate in respect with the +-10% used by GW19. However we prefer to don't argue with them.

The referee writes In Table 3, why are there 6 and not 7 events? Can you add that information around L240?

Answer The absence of the Mirandola earthquake is due to the change of the parameters excluding this earthquake from the analysis because the halving the time, the number of events is smaller than 500. In the new version of the manuscript we added a sentence to clarify this aspect.

The referee writes Some figures use commas to denote decimals and some use periods. Please make this consistent. You use periods throughout the text so I would recommend that.

Answer: Done.

The referee writes Many figures have y-labels that are partially aligned with the y-axis values. For example, Figures 2, 4, 5, etc. It would look neater if they did not overlap, similar to Figure 1, 16, 17, etc.

Answer: Done.

The referee writes The multi-panel figures have a lot of space between panels that should be reduced. You could also consider labeling the panels a and b.

Answer: Done.

The referee writes Specific Figure Comments Figure 1: Consider adding a vertical line at the true Mc value or perhaps filling in or coloring the dot that would have been selected as the best mth. In the caption, mc does not have a subscript for the "c". I would also encourage you to specify what the b and mc are in the caption rather than just saying that they are reported in the figure.
Figure 3: I think the new insets on the top panel make it too complicated. I think one would be ok, perhaps one (or both) could go into a supplement.
Figures 12 and 14: The ylabel ("mce") has a white background that cuts off the y-axis label "2.5" by just a little.
Figure 16 and 17: Consider using different shapes or different colors, could make the figure more readable to those with red-green colorblindness.
Figure 18: I'm not sure what the superscript "B" in the "bMAB" is. I think you can remove the arrows, they cover up some data and make the figure messier.
Figure 19: I think it would be interesting to know what the two random times are.
Figures 20 and 21: Add markers to denote the time of the mainshocks. Could include the pre-event average b as a horizontal line. Please tidy up the location

labels so they are placed in the same relative place. For example, the word Amatrice is really low on the plot compared to the rest and Mirandola is nearly hitting the tick marks.

Answer: We have changed the figures accordingly to the reviewer suggestions.

The referee writes
Other
L19: You use the shorthand "GR" for Gutenberg and Richter later in the paper, so consider defining it here or somewhere else before then. E.g. "...Gutenberg and Richter (GR) law..."
L33: The Pino et al citation has no year
L40: Insert comma after "could be biased", so it reads "...could be biased, causing a biased estimation..."
L43: Move the comma after "and" to after "reliability". "...verify its reliability, and consequently, the reliability of..."
L58: Same as L43, move comma before the "and"
L67: Consider re-writing the sentence as "Although the method presents some advantages, in this case the instability of r can also produce biased estimation of mc and b."
L69: Say Equation 1 or Eqn 1, not just "1"
L89 Instead of "value 1" say "value of 1"
L91 Spelling: "Occurre" should be "occur"
L92: Say "let us call it" rather than "let we call it"
L103: I'd reword as "...experimental catalogues containing some, though not all, events with m ¡ mc."
L104: Say "parameters have been selected" rather than "parameters has been selected"
L118: Delete "from", and I think it should be Figures 3 to 5 instead of 3 to 4
L120: Add space in "Fig.3" and Spelling, "hypotesis" should be "hypothesis"
L123: Consider rewording "...or to a small magnitude range." to "...or a small magnitude range" or "...or too small of a magnitude range."
L124: A couple grammatical mistakes, consider rewording as something like: "The overestimation of mc when cvt=0.97 suggests that this parameter also influences our results."
L125: Add space "Fig.5", occurs in other places too.
L128: I think you should add what ETAS stands for, so the first sentence could be "The Epidemic Type Aftershock Sequence (ETAS) model represents..." 139: The semicolon after Zhuang et al. (2004) should be a comma, so "(2004)," instead of "(2004);". Also add a comma in the similar place after Zhuang and Touati (2015)
L145: Spelling: "off-springs" should be "offspring"
L147: Spelling: "Ricther" -¿ "Richter"
L157: Add spaces between "delta m = 2" to be consistent with the other two equalities in the sentence.

L166 and other places, “Fig.s” should just be “Figs.”
L189 Here you say “b-value” rather than “b value”. I think either is ok, but should be consistent throughout
L190 Add a “of”, e.g. “...attributed to a mixture of inconsistencies...”
L206: Too many commas, could change to something like “Now we test an...” or “Let us now test an...”
L224: Add comma, “The information about the location, magnitude, and occurrence time...”
L227: make “M5.5 - M5.9 earthquake” -¿ “M5.5 - M5.9 earthquakes”.
L235: Add comma after bM
L250: too many commas, perhaps “Using the cv method we then tested...”

Answer: We thank the reviewer for the accurate revision of all these details. We changed the text accordingly to her/his suggestions.

2 Report of Reviewer B

We thank the reviewer for appreciating our work. We accepted all her/his suggestions and changed the manuscript accordingly.

The referee writes
-Pag. 3, line 91. Please correct in “...it can occur...”.
-Pag. 5, line 120. I would suggest to report either here or in the figure the p-value resulted from the Kolmogorov-Smirnov test, and the significance level considered.
-Caption of Figure 1. Please use for “mc” the same notation of the text.
-Pag. 9, line 142. Please correct in “Each of these elements...”.
-Pag. 9, line 145. Please correct in ““...For each off-spring...”.

Answer: We changed the text accordingly to reviewer suggestions.