Reviewer 1

Dear authors,

I have read your manuscript entitled "Water depth estimate and flood extent enhancement for satellite-based inundation maps" with great interest.The manuscript’s focus on the development of the FLEXTH algorithm to address the limitations of existing flood mapping methodologies is commendable. The algorithm's utilization of topographic information for enhancing flood delineation and providing estimates of water level and depth across entire flood extents represents an advancement in the field. However, the discussion on the algorithm's key features, such as accuracy, limited supervision requirements, and computational efficiency, lacks credibility due to the weakness or absence of supporting evidence. As a result, its potential applicability in large-scale flood assessments is called into question.

The introduction section offers a comprehensive review of methodologies for estimating flooded area and water depth, highlighting both the limitations and advancements in current approaches. This provides valuable insights into the current state of the field. However, apparent lack of awareness regarding some of the latest developments in the field casts doubt on the claims of novelty surrounding the method presented in the manuscript.

Overall, the paper is well-written, logically structured, and most of the figures are appropriate. The introduction of the FLEXTH algorithm represents a notable contribution to the field, with the potential to enhance flood assessment and disaster response strategies. The open access Python script is also a welcome addition to facilitate further research and collaboration within the scientific community.

Detailed below are some specific comments. I strongly suggest a significant revision of this manuscript to address these issues thoroughly.

We express our gratitude to the Reviewer for the interest in our study and for investing time in conducting a thorough assessment. We value the recognition of our methods as a significant contribution to the field, as well as the positive assessment of the manuscript's clarity and logical structure. Additionally, we are pleased that our commitment to open science, demonstrated by making the code freely accessible, has been acknowledged.

In revising the manuscript, we are specifically addressing the issues concerning the “accuracy, limited supervision requirements, and computational efficiency, lacks credibility due to the weakness or absence of supporting evidence”. In particular, in order to address the remark about the limited validation of the method, and the effect of the no-data areas on the flood propagation routine, we will include two new section to the manuscript. The first providing a comprehensive comparison of the methodology against what is probably the state of the art in the field. The section will include ad-hoc hydrodynamic simulations (which will be made openly available to the scientific public given the notorious lack of openly available dataset for validation purposes). The second new section will systematically assess the effect of the no-data areas on the performances of the algorithm, in particular on the flood propagation routine.

Finally we will revise the references as suggested and modified the two figures where improvements were necessary.

We hope the revised manuscript will meet the expectation of the Reviewer, as we are planning to incorporate most of his/her request.

Below each comment is addressed in details.

Detailed comments:

1. Line 4's mention of “billions of euros” seems Europe-centric, overlooking the global nature of flooding, which disproportionately impacts lower socio-economic regions over developed countries. It would be beneficial to provide a more inclusive perspective on the economic impacts of flooding, considering the varying economic contexts and vulnerabilities worldwide. Additionally, emphasizing the socioeconomic disparities exacerbated by flooding in vulnerable regions can underscore the urgency of addressing this global challenge.

   We acknowledge the remark. We will remove the reference to euro and added a sentence to stress that floods have disproportionate impacts in less developed areas.

2. The term "Exclusion Mask" introduced in Line 51 and used throughout the manuscript pertains specifically to the Global Flood Monitoring (GFM) product. However, a more suitable term might be "No Data Areas," as this is commonly encountered in satellite-derived flood extents regardless of the satellite or product used. While it's assumed that no data areas can be treated similarly to the exclusion mask mentioned here, this assumption should be explicitly addressed.

   Agreed, we will exchange the “Exclusion Mask” notation.

3. The assertion in line 75 regarding the unproven applicability of water depth estimation approaches for large-scale assessments is inaccurate. For instance, Teng et al. (2022) (https://doi.org/10.1029/2022WR032031) have conducted a comprehensive comparison of various methods and conclusively demonstrated their effectiveness for large-scale assessments. Peter et al. (2022) (https://doi.org/1 0.1109/LGRS.2020.3031190) have also implemented FwDET into Google Earth Engine for rapid and large-scale flood analysis. It

is imperative to acknowledge and incorporate these findings to ensure the accuracy and completeness of discussion on this matter.

Thanks for pointing that out. These studies will be acknowledged. Furthermore, the revised manuscript will features new sections including a comparison between FLEXTH and the Google earth engine implementation of FwDET (the one of Peter et al., 2022 mentioned by the reviewer in the comment).

4. In lines 87-89, the claim regarding computational efficiency uses the term "for areas of up to tens of thousands of square kilometres", which lacks rigor and specificity, particularly in terms of resolution. It is recommended to provide the number of grid cells or include the specific resolution considered to accurately assess its significance in the context of flood mapping. In addition, you are making a claim about the computational advantages of the proposed method without providing solid evidence or comparisons with other existing approaches in terms of run speed. It's essential to provide empirical data or benchmarks to support this claim and accurately assess the computational efficiency of the proposed method relative to other methods in the field.

   We agree with the comment (scale without resolution is not very informative). We will specify the issue. The first new section will includes new simulation and comparison metrics with the widely used FWDet V2.0 (Cohen et al.,2019; Peter et al.,2022) specifically addressing these aspects.

5. I find it difficult to follow Figure 1 in its current horizontal layout. Consider changing it to a vertical layout and using standard flow chart shapes to improve clarity and ease of understanding.

   We will try to improve the figure following the suggestion.

6. Section 2.1 is titled "Input and Output Products" but does not mention output at all. Consider revising the title to accurately reflect the content or include information about output products in the section.

   Agreed, thanks for pointing that out.

7. In the paragraph starting from Line 156, Method A is effectively using Inverse Distance Weighting (IDW), while Method B seems to be a crude percentile-based averaging algorithm. It would be beneficial to consider other interpolation methods such as Spline,

Kriging, or more advanced machine learning methods. These alternative approaches may offer advantages in terms of accuracy, robustness, and flexibility, especially in handling complex spatial relationships and varying data distributions. Therefore, exploring and comparing these different interpolation techniques could provide a more comprehensive understanding of the flood water level along the dry-wet borders and potentially improve the accuracy of the results.

In method B the distribution depends on the distance of each border pixel from each target location inside the flooded area where water levels are computed. In this way the impact of each reference elevation in the distribution is weighted based on its distance and the ensuing distribution is substantially different from the mere un-weighted version.

About the interpolation methods we appreciate the remark. In fact, other interpolating methods have been considered (including some of those mentioned by the Reviewer). However, they require more tweaking, and, in our tests, they did not evidence any substantial advantage, particularly considering that they are not so robust for systematic, unsupervised and large-scale applications (as they are more sophisticated). We also would like to point out that standard Kriging methods (e.g. simple and ordinary Kriging) won't be ideal for applications where a regional topographic gradient is present (as it would be the case for large-scale applications). In these settings, more refined versions of Kriging would be necessary, where the regional trends can be properly accounted for (e.g. universal Kriging or Kriging with drift). Such methods would increase the degree of complexity and might cast doubts on the overall robustness of the methodology.

8. Figure 2（especially C） requires additional clarification to enhance its interpretability in its current form. Providing detailed annotations, labels, and a clear legend could help elucidate the information presented and make the figure more intuitive for readers to understand. Additionally, including a brief description of the data represented in Figure 2C within the main text could provide context and aid in interpretation.

   Thanks for rising this issue. We will revise the figure following the suggestion.

9. The approach introduced in Section 2.3 is novel. It delineates new dry-wet borders informed by DTM in excluded or no data areas. This method, while simple, represents a step forward and deserves emphasis as the main novelty of this manuscript. However, the manuscript does not sufficiently demonstrate the effectiveness of the flood propagation routine. To address this, it is recommended to block out areas of flood extent and propagate flood into those areas as if they were excluded, then compare the results with the actual border. This step is critical to substantiate the effectiveness of this novel method.

   Thanks for appreciating the novelty of the methodology. This aspect will be highlighted.

In order to address the doubt of the Reviewer concerning the propagation routine (also shared with the second Reviewer), we will include a new section to the revised manuscript, which will specifically address the effect of masking on the performances of the algorithm. We are confident the reviewers will appreciate the new systematic analysis.

10. One significant critique of this study is its reliance on a single case study: the Pakistan 2022 case study. This limited scope is insufficient, particularly considering the lack of easily accessible validation data for this specific case study. Moreover, it hinders the ability to compare the method's accuracy and computational efficiency with other existing approaches. To address this limitation, it is recommended to include additional case studies using published datasets. This would allow for a more convincing demonstration of the advantages of the proposed method.

    Unfortunately freely available data sources, especially when it comes to water depth, are usually not openly available. See for example the review paper mentioned in comment #3 (Teng et al. (2022)). That study does not provide the modelled water depth as the access to such data is restricted.

    Acknowledging the lack of validation data as a critical aspect, the revised manuscript will include a new section dedicated to assess FLEXTH against hydrodynamic simulations in 2 geographically different locations. Although hydrodynamic simulations do not necessary reproduce real-world cases, they provide realistic physically-based scenarios useful for validation purposes and they readily provide flood extents and water depths, circumventing the limitation of remote-sensing methodologies.

11. Lines 234-235, how does this run speed compare to other existing approaches? You are claiming computational advantages without solid evidence.

    We are planning to provide further evidences to support our claims in the revised version.

12. Line 243: Please clarify the meaning of CEMS. Please spell out acronyms before their first reference. Similarly, for EMSR629, FABDAM, and other acronyms, provide their full expansion before their initial mention.

    Thanks for the remark. We will fix the issue.

13. You are using one satellite product to validate another satellite product of flood extent, which warrants clarification. Please explicitly state the advantage of CEMS over GFM flood extents for validation purpose.

    This is true but it is the only possible way unless ground-based flood delineations are available (unrealistic) or if aerial-based flood delineation are provided (rare and with difficult data access). An alternative would be to use numerical models (which we will employ in the revised manuscript, see the response to comment #10). In the current text it is specified that the CEMS flood delineation are semiautomatic with expert supervision/refinement. GFM on the other hand is a fully automatic and unsupervised system (with the consequent limitations). Following the Reviewer's comment we will try to stress the aspect better in the revised manuscript.

14. In addition to the metrics listed in Table 2, it would be recommended to include F-stat as an additional accuracy metric.

    Agreed, we will included the metric suggested by the Reviewer.

15. Using ICESat-2 altimetry data as truth to validate water depth estimates could be problematic due to mismatching of footprint and timing, as you have discussed in Section 4. To mitigate this concern, it would be advisable to incorporate additional case studies with more suitable validation data, as mentioned in the comments above.

    The use of ICESat-2 to validate flood depth estimates is per se a novelty in the field which was probably not stressed sufficiently (as noted by the second Reviewer). In fact, we are not aware of any published study that uses similar methods. Errors due to the spatial mismatch are minimal (as discussed in the manuscript). Errors due to the temporal mismatch between GFM flood maps and ICESat-2 acquisitions are also mentioned and, as correctly pointed out by the Reviewer. However, we would like to underline that similar problem would occur regardless the source of the satellite products used for validation. This happens, for example, when validating flood extent products with other products having higher mapping capabilities.

    To address this issue the revised manuscript will include a new section dedicated to validation which will employ hydrodynamic simulations in 2 additional case studies.

16. Line 322 acknowledges the critical importance of DTM accuracy and resolution. However, it raises the question of why the study does not utilize high-resolution and high-accuracy DEM data, which are available in many regions globally. If data accessibility is an issue,

it would be more beneficial to include additional case studies that utilize such data to enhance the robustness and applicability of the findings.

We believe that showing the performances of the methodology even without using site-specific high resolution topographical data is an additional proof for the suitability of the methods for large scale applications. As suggested for similar methods (see Cohen et al., Remote sensing, 2022) higher quality DTM will systematically improve the performances of the methodology.
The new test cases that will be included in the revised manuscript will use DTM with different resolution, thus helping clarifying this aspect.

17. Thank you for providing the source code. After reviewing the code, I was unable to identify memory control or chunking algorithms that would support your claims of computational efficiency for large-scale studies. Could you please clarify this aspect?

    The revised manuscript will address this aspect. Thanks for pointing that out.

18. Validation data have not been provided along with the source code.

    All the validation data are freely accessible online on GFM and ICESat-2 portals(https://portal.gfm.eodc.eu/login?redirect=%5Bobject%20Object%5D ; https://openaltimetry.earthdatacloud.nasa.gov/data/index.html). FABDEM is also publicly accessible (https://data.bris.ac.uk/data/dataset/s5hqmjcdj8yo2ibzi9b4ew3sn). We don't see the point to include in a dedicated repository a duplicate of these bulky data. However, we will specify in the revised manuscript where the data can be retrieved. What we are planning to make available instead is the result the ad-hoc hydrodynamic simulation that we performed to assess the performances of the algorithm. We believe that the latter data is more critical given the notorious lack of validation/testing datasets for flood-related studies of this type.