Response to reviewer comments on manuscript "Evaluating the effects of preprocessing, method selection, and hyperparameter tuning on SAR-based flood mapping and water depth estimation"

Dear reviewers,

First and foremost, we are grateful for your constructive feedback. The manuscript was revised in accordance with the recommendations of the two reviewers.

Please find hereafter our answers to each of your comments separately:

- In red: Replies to comments raised only by Reviewer 1 (Wolfgang Wagner)
- In blue: Replies to comments raised only by Reviewer 2 (Anonymous Reviewer)
- In purple: Replies to comments raised by both Editor and Reviewer 2

Sincerely,
Jean-Paul Travert, on behalf of the authors
Cédric Goeury, Sébastien Boyaval, Vito Bacchi, and Fabrice Zaoui.

Response to the reviewers

Reviewer 1: Wolfgang Wagner

Reviewer Comment 1.1 — This is a very detailed and interesting study that compares different SAR based flood extent and water depth estimation techniques. The study is carried out for two flood events (December 2019 and February 2021) in the floodplains of the Garonne river in France. An impressive number of simulations were carried out using different SAR preprocessing approaches, models and model parameterizations and assessed using high quality hydraulic model outputs and observed watermarks. In my view this is much needed and highly valuable study investigating the strength and weaknesses of different algorithms in SAR data processing + flood mapping + water depth estimation. However, I have several major and minor comments.

Reply: We thank the reviewer (Wolfgang Wagner) for the positive feedback regarding the scope, design, and relevance of our study. We are grateful for the constructive comments provided in the review, and we address each major and minor point in detail below to improve its clarity.

Reviewer Comment 1.2 — My main comment is that the authors do not properly account for the effect that Sentinel-1 cannot map flooding in non-sensitive areas (forests, urban areas) and water-look-alike areas (streets, smooth fields, etc.). While they do mention these effects, it has no impact on their methodology of how the SAR derived flood and water depths maps are assessed. The point is that for Sentinel-1 the "ground truth" is not the actual flood area as simulated by the hydrological model or observed by the watermarks. Instead, for Sentinel-1 the "ground truth" is the real flood area minus the exclusion areas (e.g. all pixels where Sentinel-1 cannot detect flooding due to physical reasons). This has major implications for how the results are interpreted.

Reply: Thank you for this comment. Indeed, we pointed out that Sentinel-1 cannot map flooded areas in non-sensitive areas such as urban areas and water-look-alike areas, but it was not adequately accounted for in the evaluation methodology.

In the evaluation process, we have changed the methodology to consider the "ground truth" of the Sentinel-1 extracted flood map by removing the exclusion areas from the evaluation when comparing simulated to observed flood maps. The computation of the F1-score, accuracy, and flooded surface area is now performed over the floodplain, excluding both the minor river bed and the exclusion zones. These exclusion zones were derived from the Global Flood Monitoring Service at 10 m resolution (`https://global-flood.emergency.copernicus.eu/`). This revised methodology was described in the updated manuscript, and all flood extent evaluations were recomputed accordingly. All figures presenting evaluation results (on flood maps) have been updated to ensure methodological consistency.

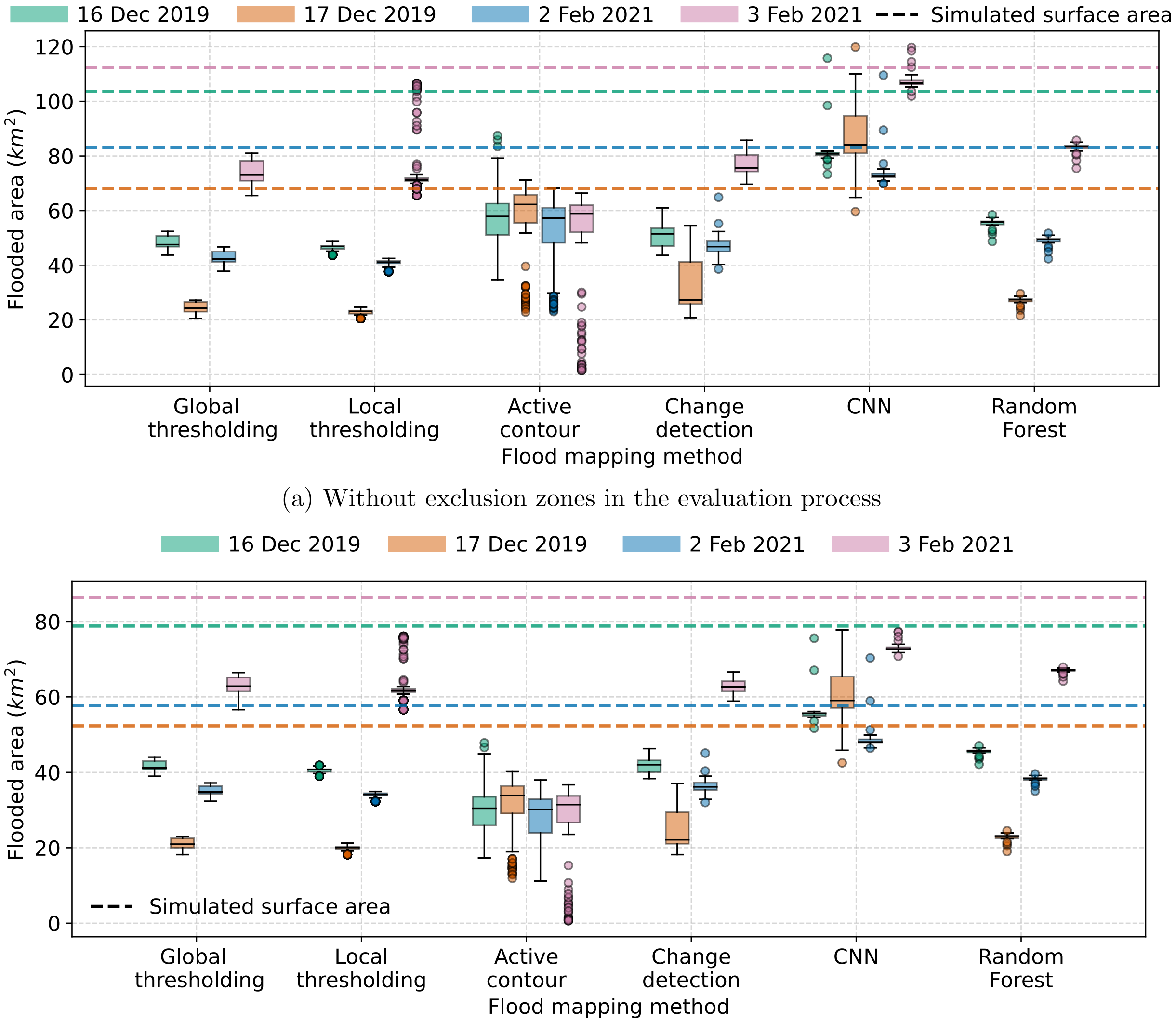
Incorporating the exclusion zones did not change the main conclusions of the study. While numerical values change, the relative performance of the different methods, the observed trends, and the influence of the hyperparameters remain consistent, as visualized in Figure 1 below for the global analysis of the generated flood maps, both with and without exclusion zones in the evaluation process.

In the revised Figure 21 (comparison of flooded surface area for simulations with varying Strickler coefficients), intersections between simulated and observed flooded areas are now observed for methods other than the CNN as visualized in Figure 2. Regarding the water depth evaluation process, the methodology remains unchanged. Approaches such as FLEXTH already account for exclusion zones by propagating water levels from observed flooded areas into excluded regions. Similarly, the cross-section–based method relies on flood extent boundaries to infer water depths, including within non-detectable areas. Therefore, the existing water depth evaluation framework remains valid.

Reviewer Comment 1.3 — At the moment the authors conclude that the CNN flood map outperforms the other results as the total flood area estimated by the CNN comes closest to the real flood extent. However, comparing the SAR backscatter images and the CNN maps shows that the CNN has a tendency to overestimate flood extent as seen by Sentinel-1. This problem then becomes apparent when estimating water depth from the CNN derived flood maps: The large RMSE values for the CNN shown in Figure 21 are evidence that some parts of the CCN maps are physically wrong.

Reply: Although the CNN-derived flood maps had a total flooded surface area closer to the simulated values, the CNN also tended to overestimate the flood extent locally compared with Sentinel-1 backscatter imagery. We have therefore moderated our conclusions throughout the manuscript to better reflect this limitation. The apparent agreement in flooded surface area hides local overestimation of inundation in the CNN maps, which directly propagates into the water depth estimation step, resulting in physically inconsistent water depths and larger RMSE values. This limitation is discussed in the revised manuscript, and the conclusions on CNN performance are mitigating, emphasising the variability between the methods and their hyperparameters.

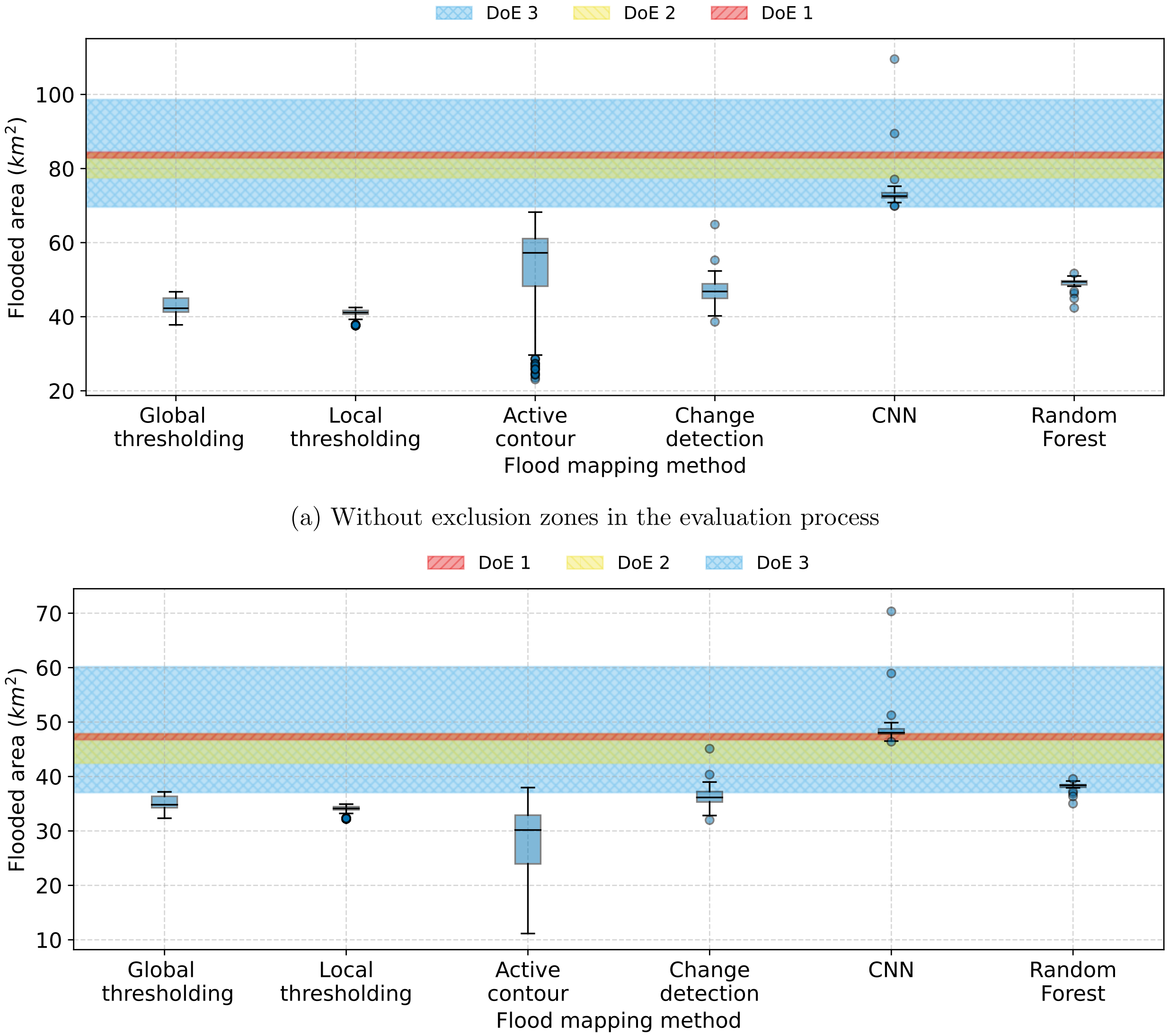
Reviewer Comment 1.4 — My second major concern is that the comparisons are unfair. While the four non-ML based approaches (global threshold, local threshold, active contour models, change detection) only use VH data as input, both ML models use both VH and VV data as input. Apparently, this is an advantage for the two ML models, and should be considered in the discussion of the results or – even better – lead to a modification of the experimental setup!



(a) Without exclusion zones in the evaluation process

(b) Including exclusion zones in the evaluation process

Figure 1: Flooded area for the generated flood maps for four satellite images and six flood mapping methods compared against their respective simulated flood maps.



(a) Without exclusion zones in the evaluation process

(b) Including exclusion zones in the evaluation process

Figure 2: Comparison of flooded area variability from hydraulic simulations using three Design of Experiment (DoE 1–3) against SAR-based flood maps. The colored surface for the three DoE indicates the 95% confidence interval of simulated flooded areas.

Reply: Using both VH and VV polarizations for the ML-based methods may provide additional information and therefore represent an advantage over the other four approaches, which rely only on VH polarization in our intercomparison.

Thus, we have modified the flood mapping methodology and retrained both ML-based models (Random Forest and CNN) using only the VH polarization, ensuring consistency with the input data used for non-ML approaches. All corresponding figures were updated accordingly, in parallel with the revisions related to the exclusion zones in the evaluation process.

The use of both VV and VH polarizations does not improve the performance of the ML-based methods compared to VH alone in the revised manuscript for this test case (the variation of the flooded area induced by using one polarization instead of two is limited to less than 0.1 km^2).

Reviewer Comment 1.5 — Morphological filters may help to reduce underdetection in flood situations, but may also lead to overdetections in non-flood situations or to wrong flood pixels altogether. So while these filters may have improved the statistics with respect to the model simulations and watermarks, it is not always clear that the physical representation of observed flood areas is improved.

Reply: We acknowledge that while these filters can improve statistical metrics such as the F1-score, they may not always improve the physical accuracy of flood maps. In the morphological filter section, we clarified that despite the statistical improvements, the physical representation of observed flood areas may not always improve the physical representation of observed floods ("While the morphological operations improved the statistics for the F1-score, the physical representation of the flood may not be improved, with for instance flooded regions on small hills when using filling operations.").

Reviewer Comment 1.6 — Most flood mapping algorithms are optimized for flood scenes, but may lead to wrong results during dry conditions. To judge the robustness of algorithms, also their performance for non-flood scenes should be tested.

Reply: For non-flooded scenes, the tested algorithms are not suitable. Thus, we added in the manuscript the limitation that the algorithms used in this article were implemented for flood scenes. Indeed, for instance, the basic thresholding method for extracting flood maps is not adapted for non-flooded events, as it is not possible to detect two Gaussian distributions, which lead to a non-representative threshold and hence flood map. Similarly, for the change detection method, it was implemented to be used automatically with a log ratio between the images and then detecting the flooded pixels using the global thresholding technique. Thus, we acknowledged in the discussion the limitations of these methods for extracting flood maps. However, some methods were more robust (given our implementation of the algorithms) to detecting water in non-flooded scenes, such as supervised classification or active contour models (but difficult to parameterize).

The other methods, such as change detection methods, could be implemented differently to work also for non-flooded scenes, without using a thresholding technique (Tupas et al., 2023). We acknowledged this limitation and proposed alternatives to the reader in the discussion section.

Reviewer Comment 1.7 — Section 1: What is the definition of a "hyperparameter"? What makes it different from a "normal" model parameter?

Reply: We clarified the definition of a hyperparameter and its distinction from model parameters in Section 1. In the revised manuscript, we explain that hyperparameters are parameters set before the model optimization or processing begins (e.g., filter window size, classification threshold), whereas model parameters are connected to the simulated physics learned or estimated thanks to observations. A sentence has been added in Section 1 to clearly state this distinction "In this work, we refer to hyperparameters as user-defined parameters (e.g., window size for speckle filtering, threshold level for flood classification) that govern the behavior of an algorithm, as opposed to model parameters that can be measured or estimated from physical.".

Reviewer Comment 1.8 — Section 1: There are some recent studies that investigated the effect of different model parameters on flood mapping accuracy (e.g. recent studies investigating different change detection algorithms). Please review the literature and relate this work to the existing publications (also come back to this point in the discussion section).

Reply: We have reviewed the literature to provide a more exhaustive overview of recent studies investigating the effects of model parameters on flood-mapping accuracy. Indeed, we found "new" publications discussing model parameters in change detection methods (Tupas et al., 2023), thresholding methods (Chini et al., 2017), and ML-based methods (Ghosh et al., 2024). These references were added to the introduction and used in the discussion section to offer alternative perspectives to the readers.

Reviewer Comment 1.9 — End of section 1 / beginning of section 2: Check for repetitions

Reply: The first sentences of Section 2 ("The study presents a workflow for processing Synthetic Aperture Radar (SAR) images to support hydraulic information extraction, specifically for generating flood maps and water depth fields. The main objective was to investigate the influence on the final outputs (flood maps and water depth fields) of different method choices, including their associated hyperparameters.") were repetitive with the end of Section 1. These sentences were removed, and the beginning of Section 2 was adapted to avoid repetitions. Furthermore, throughout Section 2 and the manuscript, we improved clarity by removing redundancies.

Reviewer Comment 1.10 — Line 86: Why are so many configurations tested for the local threshold approach (36 versus 2 / 2 / 6 configurations for the other three methods). Does this mean that the threshold approach has advantages?

Reply: We tested the local threshold approach with additional configurations, since 4 hyperparameters need to be set. To systematically assess their individual and combined influence on the results, we explored multiple configurations for each hyperparameter (typically 2–3 options per variable). Consequently, the total number of hyperparameter combinations is higher for the local threshold method than for the other approaches, which require fewer hyperparameters. The larger configuration set does not imply an inherent advantage of the threshold-based method, but rather reflects the need to evaluate the sensitivity of this approach to more hyperparameters.

Reviewer Comment 1.11 — Figure 3: Show location of in situ sites

Reply: As suggested by Reviewer 2, the zoom on the study area was adjusted in Figure 2 for better coherence. Following your recommendation, we have also added the locations of the in-situ sites to the study area figure.

Reviewer Comment 1.12 — Line 163: Only in a narrow sense I would agree to this statement: "The main source of error in SAR imagery is speckle noise . . . ". In practice, there are many physical reasons for uncertainties in the SAR derived flood maps.

Reply: We agree that the sentence was too affirmative. We have rephrased it to underline that it is a usual step in flood mapping studies to remove noise in the image. The sentence was changed from "The main source of error in SAR imagery is speckle noise, which arises from the coherent summation of scattered electromagnetic waves." to "In SAR imagery, speckle noise, which arises from the coherent summation of scattered electromagnetic waves, is a well known source of error."

Reviewer Comment 1.13 — Line 248: Visually, the SAR2SAR filter looks indeed nicer than the other filters. But are there any quantitative indicators that can substantiate that SAR2SAR "outperforms the traditional methods"? How much of this filtered image is invented, how much of it is true?

Reply: Indeed, in homogeneous areas, the SAR2SAR filter looks nicer with more homogeneous statistics. To provide a quantitative assessment, we evaluated the Equivalent Number of Looks (ENL). The original sentence, "Visually, for these hyperparameter configurations and on this zone, the SAR2SAR approach seemed to outperform the traditional methods." was removed because we cannot guarantee that the SAR2SAR approach outperforms the traditional methods solely on visual inspection. This statement was premature in the manuscript. Instead, we now present ENL-based quantification after this section, offering more objective information on filter performance, which has been analyzed in detail.

Reviewer Comment 1.14 — Figure 6: These are the VH images, right?

Reply: The images in Figure 6 are in VH polarization. We added in the caption the polarization of the images.

Reviewer Comment 1.15 — Figure 7: Use the same y-axis for a direct comparison

Reply: Corrected.

Reviewer Comment 1.16 — Line 349: How many flood cases of the Sen1Flood11 cases show similar conditions as for the Garonne river flood.

Reply: It is difficult to determine how many flood cases in the Sen1Floods11 dataset are truly comparable to the Garonne River event. The dataset consists of numerous image chips extracted over large flooded areas that exhibit heterogeneous conditions. Flood characteristics depend on local topography, hydrology, land cover, and river morphology, which vary significantly across regions. Events that appear visually similar may correspond to very different hydrological processes and flooding dynamics. This makes it challenging to objectively define and quantify flood cases that are truly comparable to the Garonne River flood. However, in the dataset related to the Sen1Flood11, they tried to add more geographic dispersion and heterogeneity to the data ("including 11 flood events with various geographic conditions"). This comment has been added to the manuscript.

Reviewer Comment 1.17 — Figure 9: I find the spread of the results for different algorithm / flood case combinations surprisingly low. What is the reason for this? Does this also reflect different pre-processing options?

Reply: The spread of results may appear low when considering F1-score and accuracy because even small variations in these metrics can correspond to a large number of pixel-level differences, given the overall image size. The limited range observed in Figure 9 is primarily due to the choice of metrics (accuracy and F1-score), which tend to compress variability. To provide a more tangible comparison, we also analyzed the difference in flooded area between the simulated and observed flood maps (excluding masked zones), as shown in Figure 9. In this figure, differences between methods and flood cases are more pronounced, with variations of several square kilometers. To improve clarity, Figure 10 was merged with Figure 9, resulting in a single figure for the global flood evaluation assessment.

Reviewer Comment 1.18 — Line 535: Some grasslands may cause "water-look-alike conditions", but normally vegetation causes a loss of sensitivity of backscatter to flooding.

Reply: Indeed, you are right: the vegetation may hinder the satellite's ability to detect flooding. We meant vegetation such as wetlands. The word "vegetation" was replaced with "wetlands".

Reviewer 2: Anonymous Reviewer

Reviewer Comment 2.1 — This study investigates the impacts of various preprocessing, mapping, and depth estimation methods on flood mapping and water depth estimation using Synthetic Aperture Radar (SAR) images. The results suggest that an ensemble approach, accounting for various uncertainty sources during the modeling process, should be preferred. Overall, this study is well-designed and comprehensive, and the findings are meaningful for flood risk management. However, I still have several comments and suggestions for improving the current work.

Reply: We thank the reviewer for the positive evaluation of our work and for acknowledging the contribution of our study to flood mapping and risk management. We appreciate the comments and suggestions provided. We modified the manuscript accordingly to improve its clarity. Detailed responses to each point are provided below.

Reviewer Comment 2.2 — This paper is a little lengthy, especially for the method sections. If some of those methods are not new, it is suggested to make the description more concise, or put some of those in the appendix.

Reply: The beginning of the second section was rewritten for more concision. While adding details to address Reviewer 1 and Reviewer 2's comments, we maintained the paper's length. We even reduced the number of lines in the core of the manuscript by moving some technical methodological explanations to the Appendix. Furthermore, the manuscript was improved and some repetitive parts were removed. We also removed 2 figures (one was redundant and one was merged with another figure) and 1 table from the manuscript.

Reviewer Comment 2.3 — The overall study is like presenting the details of a calibration process for SAR-based flood modeling, which involves rigorous evaluation as well as subjective adjustment based on the model's experience. Could you provide some general guidelines on how to make the model calibration more effective and efficient?

Reply: In Section 6, in the calibration process, we added information to improve the model calibration, given the uncertain observations and uncertain simulations. Specifically, we mentioned the difficulty of obtaining a reliable data source to calibrate the flood model, given high uncertainties in the flood maps and water depth outputs. However, we added guidelines to improve the model's calibration. For instance, the practitioner can use data assimilation for estimated the Strickler values. Furthermore, considering the high variability of the generated flood maps, the calibration process should take into account this variability, using, for instance, the GLUE methodology (to estimate the uncertainty of the model predictions) or consider the worst-case scenario to have conservative simulations regarding this natural hazard.

Reviewer Comment 2.4 — Figure 2: Does the longitude label "5-degree O" represent "W"? If yes, and to avoid confusion, it suggested to change it to "W". Also, it would be better to add a zoom-in view for the study area.

Reply: The "O" stands for "Ouest" in French. We forgot to modify the language in the generation of this Figure. The Figure was modified by changing O to W. Furthermore, we added a zoom-in view of the study area, in this Figure, including the different roughness subdomains and removed the zoom on the study area in Figure 3. Thus, Figure 2 focuses on the study area and Figure 3 on the measured discharge only, which should be more coherent.

Reviewer Comment 2.5 — Lines 118-120: It is acknowledged that the simulated flood maps just served as a reference in this study. However, it should be noted that the uncertainty in the physics-based flood modeling process is not negligible. Furthermore, the limited flood events/areas selected in the study and the uncertainty in the "ground-truth" observations may play a role in the model evaluation process. Please refer to the two papers below.

References:

"Uncertainty analysis and quantification in flood insurance rate maps using Bayesian model averaging and hierarchical BMA" (https://doi.org/10.1061/JHYEFF.HEENG-58)

"Beyond a fixed number: Investigating uncertainty in popular evaluation metrics of ensemble flood modeling using bootstrapping analysis" (https://doi.org/10.1111/jfr3.12982)

Reply: The uncertainty in the "ground-truth" observations has been discussed in the paper and serves only as an intercomparison tool against the observed flood maps or water depth fields. However, we acknowledge that uncertainty in the ground-truth observations can be accounted for. This point has been clarified in the text, and we added the references you proposed, which highlight the need to use the GLUE methodology, for instance, and to assess the uncertainty in the physics-based flood modelling process. We insisted on this limitation and added in the discussion a reference "In future works, we could also consider the uncertainty of the outputs of the physics-based model in the model evaluation process, based on GLUE analysis as proposed in (Huang and Merwade, 2024)".

Reviewer Comment 2.6 — Figure 7 and Line 259: Why was the performance of SAR2SAR for the flooded area not as good as that for the dry area?

Reply: First, Figure 7 has been rescaled to have the same y scaling. The performance of SAR2SAR for the flooded and dry areas cannot be directly compared, as the ENL differs between the unprocessed images. The statistics of the image in the dry area are more homogeneous, hence leading to higher

ENL values. However, we can see the same tendency in the dry and flooded areas, with an increase in ENL for the SAR2SAR method. Furthermore, instead of numerically comparing ENL values, we changed the Figure to compare the improvement in the ENL ratio between processed and non-processed images, thus comparing the non-dimensional numbers for the flood and dry zones. In the sentence "The SAR2SAR consistently achieved the highest ENL values across all dates and regions, particularly for the dry region (see Fig. 7a).", we removed particularly for the dry region which made the sentence confusing. Moreover, we rephrased the analysis on the ENL values to make it clearer.

Reviewer Comment 2.7 — Lines 370 and 464: No hyperparameters were considered in the random forest and CNN approaches, but they are also critical in determining the model performance in both the training, testing, and application processes.

Reply: This limitation was included in the discussion of the manuscript, as the supervised classification model's hyperparameters could influence the outputs. However, in this work, for supervised classification methods, we chose not to consider any hyperparameters, as the hyperparameters were optimized during the training procedure using Bayesian inference.

Reviewer Comment 2.8 — Line 382: There are 1222 generated flood maps, but 26*48=1248?

Reply: This is right indeed, it should be 1248 in the text, and it was corrected.

Reviewer Comment 2.9 — Lines 91 and 502: The DEM data was used in water depth estimation. Does it include bathymetry for the river channel?

Reply: In the water depth estimation framework, we used the same data as that used in the numerical model. Thus, the DEM data include the bathymetry for the river channel that was measured through cross-section measurements and then interpolated. We changed the text at these two specific lines to make it clearer. In particular, we added in the numerical model section information on the bathymetry measurements in the river channel. We added at line 502 "The DEM is a combination of topographic data at 1 m spatial resolution with vertical accuracy between 0.2 and 0.5 m, and the bathymetry of the channel (based on 70 cross-section measurements)."

Reviewer Comment 2.10 — Line 646: The word "First" does not make sense here since there is no "Second" in the following text.

Reply: The word "First" was removed.

Reviewer Comment 2.11 — Table 3: Please add units for the Strickler parameters.

Reply: The units have been added to the column names.

Reviewer Comment 2.12 — Line 652: Is it "94%" or "95%"?

Reply: Indeed, it is 95%, we corrected it in the manuscript.

References

Chini, M., Hostache, R., Giustarini, L., and Matgen, P. (2017). A hierarchical split-based approach for parametric thresholding of sar images: Flood inundation as a test case. *IEEE Transactions on Geoscience and Remote Sensing*, 55(12):6975–6988.

Ghosh, B., Garg, S., Motagh, M., and Martinis, S. (2024). Automatic flood detection from sentinel-1 data using a nested unet model and a nasa benchmark dataset. *PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, 92(1):1–18.

Tupas, M. E., Roth, F., Bauer-Marschallinger, B., and Wagner, W. (2023). An intercomparison of sentinel-1 based change detection algorithms for flood mapping. *Remote Sensing*, 15(5):1200.