Dear Editor and reviewers, thank you very much for revising our manuscript. We appreciate the time taken by each reviewer to revise our manuscript and the suggestions that helped improve our work. In the following, we reply to each comment. We made all changes accordingly in our revised version. Lines here refer to the revised manuscript with tracked changes.

Janbert Aarnink on behalf of the co-authors

CC1 Comments by Prof. Andres Iroumé

Is a very well written and interesting manuscript.

I have a few suggestions intended to complete/improve some aspects.

Response: Thank you for your comments and help in increasing this manuscript's quality. The suggestions were well appreciated.

They are:

Introduction:

- Page 1, L19-20. Natural mortality wind, snow loads, wildfires and beaver activities can also be recruitment sources.

Response: thank you. The reviewer is correct; we added the abovementioned processes in Line 18.

- Page 1, L20. “Wood plays a crucial role by trapping sediment, creating pools, and generating spatially varying flow patterns”, not only as it distributes along the riverbanks but also when stored within the active or bankfull channel.

Response: we edited the sentence to clarify this aspect in Line 19 in the revised ms.

- Page 2, L34. The number of observations of instream wood is scarce? I do not fully agree. Perhaps the amount of observations of instream wood dynamics is scarce, so please clarify.

Response: The other reviewer also raised this point, so we edited this paragraph to clarify what we meant. Lines 35-37.

- Page 2, L43, about the best methods to quantify wood transport. Not only video-based methods, but also the installation of a GPS in each wood is a very good method, but extremely expensive.

Response: yes, agreed. We added this and other approaches in the revised text. Lines 37-45.

Methods:

- Page 3, L86. Figure 1 does not give an overview of the data collection and processing. It gives an overview of the process to follow to collect and process data. Please also correct the title of Fig. 1 below the figure.

Response: we corrected the text accordingly. New caption of Figure 1: Overview of the methodology used for data collection and processing. .

- Page 4, L107 and 115. Figure or figure? Please decide.

Response: we corrected the text accordingly across the manuscript.

Discussion and conclusion:

- I do not find comments related to the limitations of the use of low-cost cameras, and how to avoid these limitations, may be by using high resolution cameras, installations, others. Please discuss and conclude.

Response: yes, this is an important point. We added some discussion about the use of high-resolution cameras. Section 4.4.

RC1 Comments by Prof. Diego Panici

The manuscript is about the automatic detection of instream large wood in video recording using deep learning tools. The results are really intriguing, but I believe that a substantial revision will be needed before considering this paper for publication. Here are some major comments:

First, there is limited to no comparison with other existing models. CNNs are widely used for image recognition (and, indeed, the authors acknowledged YOLO being the most widespread algorithm), yet, there is no comparative analysis with other studies or algorithms.

Second, the overall aim and output of this manuscript is really unclear. It is necessary to explicitate this further and emphasise what the study has revealed and what increase in scientific knowledge it has brought. As things stand, it is hard to discern what is the new scientific knowledge that this paper has produced.

Third, the paper structure needs substantial changes. The results and discussion sections merged together makes difficult to discern between the actual observations and the authors' analysis. It is essential that the two sections are kept separate. The language used is also not appropriate for a scientific paper: this was mostly informal and colloquial and needs thorough revision.

Fourth, the method was unclear and lacked explanation (at times it was not even easy to understand what cameras have been used, where and how, whilst a schematic would have helped). Overall, this limits the generalisation of the method proposed.

An annotated version is also provided with in-line comments.

Response: Thank you very much for the comments; we appreciate the time taken to revise our manuscript and the suggestions that contributed to a significant increase in the quality of the paper.

Line comments:

Line 27: DOI

Response: replaced

Line 38: In the field, though. Recently, there has been a lot more work on experimental work to try and define transport dynamics:

Innocenti et al., 2023 https://doi.org/10.1029/2022WR034363

Innocenti et al., 2022 https://doi.org/10.1002/esp.5516

Panici, 2021 https://doi.org/10.1029/2021WR029860

just to cite some that focused almost exclusively on LW transport in flumes

Response: the reviewer is right; we added more information regarding flume experiments and previous studies; thanks for the suggested references. Lines 32-24.

Line 56 : typo

Response: corrected. Lines 65-66.

Lines 71-76 : This probably needs to be more detailed, as to evaluate what different algorithms do and how they have been adapted to tracking LW and other objects in rivers

Response: we have extended the section regarding the existing methodolofies for river monitoring using machine learning. Lines 74-89.

Line 89 : No problem on this, but was there a reason why iPhones were not included? Just because they represent a significant portion of the phone market

Response: This is something to discuss, but we did not consider iPhones as low-cost mobile phones and we did not have those phones lying around to use.

Lines 91-92 : Can you add a few more details about this?

Response: we added more details about the previous studies in the Ain and Allier Rivers in France and the datasets from these previous works. Lines 124-134.

Lines 102-103 : This is unclear: what are the "rest of the labels"? If 10% is manually labelled, and then the remaining 90% is labelled by means of a CNN, what is the remaining amount of labels?

Response: We clarified this aspect, that wasn't very clear in the original text. Lines 139-144.

Line 103 : Would be worth stating the accuracy and how it was checked that

Response: We added the accuracy. Line 146.

Line 112 : I noticed that there's a mix of British and American spelling, e.g., 'greyscale' (British), 'labeling' (American). I would recommend to stick to one spelling

Response: We carefully revised the text and homogenized the style.

Line 113 : Does it mean this is the total number of LW observed for the whole database?

Response: We clarified this value and more clearly stated we have 15,228 images with a totalk of 33,160 detection in the database. Line 149.

Camption figure 2 : Perhaps it may help to have a sub-figure with maps where the images have been captured, rather than just coordinates.

Response: We added a map as suggested. New Figure 2.

Line 116 : It is unclear why the PCA is being used here. What is its purpose?

Response: We explained why this PCA was applied. Lines 149-170.

Line 125 : This is a rather blunt statement. It can be split like this (and is a fairly common practice), but it is not a necessity.

Response: We smoothed the sentence. Lines 172.

Line 141 : By whom?

Response: We added this missing information. Lines 191-193.

Line 240 : This needs more detailed explaining what was effectively done: how does the algorithm work?

Response: we clarified this. Lines 321-330.

Line 245 : It is difficult to disentangle results from discussion here. Could you not split this into two sections where results are commented separately from any analysis or discussion from the authors?

Response: We understand the concern, it was challenging to discuss and interpret our results, as this is a methodological paper mostly, and each step and result needed to be justified and explained. However, we improved the structure and split the results from the discussion.

Line 248 : I wouldn't necessarily call this in scientific terms

Response: we rephrased it throughout the manuscript.

Line 250 : This is true for stationary LW, but for waterborne LW?

Response: we added a clearer explanation. Lines 374-376.

Line 258 : This needs definition

Response: we defined this term. Lines 337-339.

Lines 263-264 : This was already said

Response: We removed the sentence.

Line 265 : Surely these are typos?

Response: Yes, sorry. We don't know how those got through. Corrected. Lines 345.

Line 271 : This needs to be said in the methods

Response: we moved this part to the methods as suggested. Lines 379.

Line 276 : I struggle to follow this section, and why this was needed. Consider re-structuring this paragraph to outline objectives and values displayed in the table

Response: we restructured the section (Lines 350-354) and moved the explanation to the methods seciton. Lines 308-320.

Line 285 : Were some cameras attached to bridges? This was not really clear in the methodology. There really needs to be a schematic and an addition to Table 1 with details of the type of camera used (fixed, non-fixed, etc.). Currently it is very hard to understand the setup, as it is quite confusing.

Response: We elaborated and added some more details in the methods section and to table 1. Section 2.2.1.

Line 290 : largely smaller?

Response: we rephrased this sentence. Lines 364.

Line 295 : Which are...?

Response: we added more details in the methods. Lines 389-391.

Line 304 : This is not proper scientific writing

Response: we changed this term. Lines 398.

Line 311 : Is this the right word here?

Response: we removed this term. Lines 415.

Line 320 : What is 'smaller' here?

Response: we clarified this aspect. Lines 442.

Line 339 : How does georectifying come into play here?

Response: we removed this from here and clarified the sentence. Lines 432-437.

Line 342 : Such as?

Response: we added more details. Lines 451-452.

Lines 342-345 : I don't think this is a proper argument: Neural Networks use statistical relationships that take into account specific characteristics. There are no 'shortcuts' in this

Response: thank you for the clarification, we rephrased the sentence to make it clearer. Lines 453-459.

Lines 348-349 : Where online? What was the reason to use this image and not another one? If you found it online, why is it not properly credited?

Response: We expanded this part of the results and credited the source properly in the methods section. Lines 460-462.

Line 355 : 'a lot' is very much colloquial

Response: we replaced this term. Lines 469.

RC2 Comments by Dr. Chris Tomsett

Lines 19-20: There are quite a few more important sources of large wood, such as windthrow and natural mortality, or influence from fauna.

Response: we have added more sources. Line 18.

Line 25: I think localised rather larger is more appropriate here when talking about inundation.

Response: we have adjusted the term larger to localised. Lines 23.

Line 26: Could likely do with some more up to date references here to show advances in this area. E.g. https://www.mdpi.com/2076-3417/13/18/10454#B65-applsci-13-10454 and https://www.mdpi.com/2077-1312/10/7/911

Response: we have added one of the suggested references and added another one. Lines 24.

Line 36: Although this is introductory, I think some reference to those existing algorithms, and why they are location specific would be beneficial, especially as one of the goals of the paper appears to be to reduce the site-specific nature of current wood detection algorithms.

Response: we have added references to current monitoring sites. Later on in the introduction we talk about why they are site specific. Lines 54-64.

Line 40: Comma needed after high-resolution aerial surveys. This occurs in other places where you are listing.

Response: we hadded a comma. And we went through the document adding comma's to listings of 3 or more items.

Line 41: Give some examples of how RFID is being used, such ashttps://onlinelibrary.wiley.com/doi/full/10.1002/esp.3463?casa_token=G-p1V7DmbDEAAAAA%3AcGP5b3hqKzIPyE8YEHHSPK78ppWrXGMST7iPG-JZUsnuplmrtM2Vs6gkX-LIQYRsjPiCq-bqfgCXEA, https://onlinelibrary.wiley.com/doi/abs/10.1002/esp.1888?casa_token=w1NFWrbZJ7gAAAAA:0YvqxuFyU7vHaDZ2FQ3hHxlDP474jAXCKdoHvR_oKZbKkLphb0btepE7Yw0yjn9ZpJW3KPwQb6tyYQ, and the authors prior work using RFID to improve CNNs.

Response: we added an additional use of the RFID tags and GPS loggers. Lines 37-42.

Line 54: Introducing a sentence along the lines of 'they are also limited by their spatial locations, and rely on specific setups being installed prior to an event' This would lead nicely into the use of citizen science.

Response: we added a similar sentence. Lines 63-64.

Line 61: Similar to the above, another link sentence here would help the flow, think along the lines of 'Advances in machine learning methods may help to overcome this and allow for widespread wood detection'.

Response: we have adjusted the last sentence and added a similar sentence to improve the flow. Lines 72.

Line 62: I am not sure this first sentence is needed, feels informal and unnecessary.

Response: we deleted the sentence as its contribution is indeed limited.

Line 67: Starting 'The CNN has multiple…', move this to next paragraph in methods and needs expanding (see comments relating to this below). Then add in some objectives as to how you plan to run, test, and evaluate your algorithm development.

Response: we have moved this part of the section to the methods section. Lines 92-99.

Line 78: The first sentence explaining your choice of YOLO algorithm should be proceeded by a small review (one paragraph) on how CNNs work, and why they may be more suitable for detection than other algorithms, building on the information from the introduction.

Following this, I feel that just saying YOLO was chosen for speed and accuracy is limited, especially as this is tested on generic imagery in their paper not large wood. Is there any reason to suggest it would be better for large wood? If not, have other studies trying to detect wood compared between algorithms? Think this is quite a crucial area to justify.

Response: we have added a section comparing different algorithms and explaining more clearly why we chose the YOLO algorithm. Lines 92-111.

Line 80: With above changes, a new paragraph could likely be started with 'Training a CNN…'

Response: we have started a new paragraph there. Lines 99.

Line 86: Combine these two sentences for better flow.

Response: we have combined the two sentences. Lines 118-119.

Line 89-95: I think some more details about the quality of the cameras here would be useful, such as resolution. Do you have a record of how much wood was added to each stream? How long have the monitoring programmes been underway in France and is any of that manual input to the channel? Where are the online videos from and what helps to make the images and wood a more diverse setting? These are all questions that need addressing. Does each image represent a single piece of wood, or are there more pieces of wood in each image, relating to the 15,228 number here. I think you should refer to table 1 here, and also adjust figure 2 as outlined below in figures and tables section.

Response: we have addressed all questions, clarified these aspects, and referred to table 1. Section 2.2.1.

Line 98: Figure 2 also shows bounding boxes, maybe reference this instead.

Response: we have changed the figure we refer to. Lines 138.

Lines 99-105: This section could do with some additional clarity, especially as this is an additional CCN alogirthm being deployed I assume? Were you checking that the automated bounding boxes for these were detecting wood, as this is not clear which dataset you are referring to by saying the labels were checked manually. There is also no explanation of why this worked better for 11 of the 15 datasets, or what your tolerance for acceptable mAP was, especially as not good enough was below 20 percent. Moreover, when images were checked manually, were incorrectly labelled frames eliminated or adjusted, or left as incorrect?

Response: We have given a clearer explanation of the labelling process. Section 2.2.2.

Line 112: Why 80*80, is this purely incidental that no wood was larger than this, I also assume this is in pixel size not other units?

Response: we have expanded on the explanation. Lines 155.

Line 116: There is no statement of why this PCA was undertaken, and it only becomes clearer when reading the results. You need to add some context as to why this is undertaken. Furthermore, if the results of the t-SNE test are stochastic, could you not run the test numerous times to assess the diversity, akin to monte-carlo scenarios?

Response: we have explained why we are using PCA and added an explanation on the application of the t-SNE, which is only used for visualization purposes. Lines 149-170.

Line 125: Swap must for ‘is typically’, if smaller datasets don’t allow it, there is not always a split in this fashion or a separate test dataset.

Response: we have replaced ‘must’ with ‘is typically’. Lines 172.

Line 125 –142: This section is trying to explain a somewhat complex training and validation procedure, whereby computational trade-offs mean omitting some of your data as validation. However, it feels as though how these 6 examples were selected is not overly clear, besides not

being at the same place and time. It may have made sense to use dataset 14 also, purely as that would give you validation samples at a range of sizes.

The section took a while to become clear as to what the process was, and that datasets weren't being dropped from training, just the number of validation sets dropped. Perhaps trying to simplify the wording in places and go through the order. For example, 6 validation cycles were run, for each one a single dataset was dropped for validating and the model trained on the remaining 19. These 6 were chosen to represent a range of conditions, and reduced computational overhead by not undertaking 20 validation cycles.

Response: we have rephrased most of the paragraph to better explain the training and validation procedure. Section 2.2.4.

Line 139: Where has this extra dataset come from, and why was it not introduced with the other datasets? Who has been studying this, a research group(s) or monitoring agencies?

This is a useful case example that in essence the paper could have been framed around. I.e. instead of can we implement a cool algorithm, can we reduce human labour of monitoring wood?

Response: we have added an explanation of the dataset and information on the research in the introduction. Lines 189-193.

Line 145-152: This seems to be an odd way to do your sensitivity test, as although you are trying to identify the effect of number of inputs on output quality, if these inputs are multiplied for smaller datasets, then they are not adding any extra information, only overtraining the model? Would it not have been better to undertake this at a smaller number of images to assess performance, on possibly a limited number of datasets. I.e. for 10 of your datasets over 300, or 8 over the 1000, sequentially go from including 100, 200, 300, etc, and then quantify at what point there is no improvement in the model? This in itself could be one of the scenarios.

Response: We agree that the suggested way would also have been an excellent sensitivity analysis. We did it in the described way because we wanted to keep most of the data in the large datasets, without overrewarding the large datasets. This is because the model is biased towards large datasets as it is rewarded equally on each image on the total database. We have added this explanation to the text. We made a separation between first testing how much data was needed in the training process before testing how to improve the performance of the model in the next section. We have explained this in a better way. Section 2.3.1.

Line 154: Although it is clear there are no river based large wood studies to learn from, it seems that casting the net a little wider shows these studies have been used in similar ways on living trees and perhaps other wood related scenarios. E.g. https://ieeexplore.ieee.org/abstract/document/9643113?casa_token=Vm749u_aLtQAAAAA:lQ8hGqEscqOOTf4M5Co8uVAJ1qsiJtDGUoMrQDFj-oSM14tTKiVBbKzIUl1G0OTwZ5AGgRy_qw

Response: we clarified this and added that a CNN had not yet been trained for our specific purpose. And we added a reference to the referred article in the introduction. Lines 76-89.

Line 158: Although in principle I can understand how all these parameters effect wood detection, but has any worked actually been conducted on this? If so, reference it.

Response: we have added an explanation regarding augmentation strategies transferring poorly between datasets, and that it's why we explored different strategies. We have added a reference on how to augment datasets, but we haven't found a source that specifically explores training scenarios for wood detection purposes. Lines 209-220.

Lines 158-160: It says 14 were trained and compared to a baseline, but there are 13 outlined. Either say 13 models trained, or 14 including the baseline for which the other 13 are compared to.

Response: we have changed it accordingly to 14 including the baseline. Line 209.

Lines 162-163: Why were the values of 4% and 30% chosen, is there a rationale for this? The logic behind this makes sense, but just need to clarify reasoning for thresholds, even if they were just decided as no previous study to base upon.

Response: we have added more explanation on how we got those numbers. Lines 221-228.

Line 170: Can the dataset size be given for total number of images, i.e. how similar is 'approximately the same'. This also feeds into informal language comments.

Response: we have added the dataset size. Line 234.

Line 171: Why such a high number, when lowering it slightly could reduce the need for oversampling from some datasets? This appears similar to the sensitivity analysis you performed.

Response: we have clarified why we chose this number. Lines 236-237.

Lines 176-179: Can you specify the number that were rotated vs mirrored, as the and/or makes it unclear if this was randomly done and randomly distributed. Were any both mirrored and rotated?

I agree, that only partial rotation is necessary, this seems like a sensible decision to have made.

Response: we have explained the total number of mirroring and rotations performed. Lines 241-247

Line 180: Again, how many were altered, and what proportion were mirrored or rotated. I think this needs more detail so the user knows what was done. How much extra data did this result in?

Response: we have added an explanation of the total number of images in the dataset for this scenario, and also for the two scenarios after. Lines 248-251.

Line 188: I feel that the inclusion of the phrase 'non-living wood' implies you are adding living wood, as opposed to wood that is not floating. Could be removed.

Response: we meant that the there may be a difference between detecting living trees and wood, but we understand the confusion and removed it. Lines 258-262.

Line 189: Change example to wood sample.

Response: we have changed example to wood sample. Lines 260.

Lines 192 –196: This is an interesting scenario, primarily as these are open source datasets, with lower quality, but greater geographical diversion. In essence, I am not sure this is just testing data quality. Again, with the addition of other datasets, I think they should be mentioned in the original introduction of data, and their locations (even if approximate) included on a figure map. They can be highlighted/commented that they are only used for testing or specific scenarios, but curious as to why they were not included from the outset?

Response: We agree that we are also testing the geographic diversity of the data in this scenario. We did not include this data because we labelled the data only in a later stage. The tests before were already performed by that time. Therefore, this is not part of the main dataset and is merely a test to check whether adding any data of even low quality would help in training the model. As this part of the test is small, we do not feel like further elaboration on the data is justified.

Regarding data quality, YouTube and Twitter generally highly compress videos, and therefore, the data quality is generally lower. We have added this aspect to the discussion. Lines 263-269.

Line 199: Which datasets were removed?

Response: we have added the datasets that were removed, and clarified this aspect in the revised text. Lines 274.

Lines 201 –206: This is really well explained and justified here, so should therefore be a model for your other scenarios where the justification is weaker.

Response: we have added more detailed explanations to the other scenarios. Lines 276-281.

Lines 207-211: This is an interesting scenario to assess, as many secondary data sources may be of lower quality. However, are the double- precision images used either a) the down sampled images at 416*416 resampled again to a higher resolution, or b) the original images resampled to 832*832? The text make it seems like you double the resolution of the down sampled image (scenario a), as opposed to changing the original resampling (scenario b). Make this clear either way.

Response: we have changed the explanation and made clear it is scenario b. Lines 282-286.

Lines 213-226: This is a really well written section on the statistics being used, what they mean, and how a reader should interpret them.

Response: thank you.

Lines 232–239: How is this sensitivity test different to the one introduced earlier, and why has this one got more dataset sizes to test the sensitivity? This is also not referred to in the results as far as I can tell, so what is the purpose of this section?

Response: This is indeed not different and we have removed it from the manuscript.

The variance method also adds some confusion, is each of these models run several times, and the best results taken? If so, why the best results, does that not overestimate model performance? This could be explained better.

Response: We have elaborated on the explanation of the process in lines 310-320.

You then talk about comparing between models, which again is fine but is very brief as to why, needs more explanation. You then mention a final model, is this not just your optimal model from all your testing?

Response: We have changed the section to explain why our method actually stops the overestimation of the models' performance. Lines 311-313.

Line 238: Which dataset is this, is it the same as the one introduced previously on the river Inn? Again, this needs to be stated.

Response: we have elaborated on the Inn dataset further in section 2.1.3 and in this part referred to that section.

Lines 240–244: This seems like a really sensible addition and is good to see some unpicking of what is happening behind the scenes. Maybe a brief idea of how this works, and what you hope to find and why you picked certain images (one river, across rivers, different angles?) would be sensible? In this case you might hope to hypothesise why some images/datasets are less well classified? This would be nice to see expanded on in the discussion.

Response: We have extended this part to better explain what we are trying to understand from the method. Lines 321-330.

Line 248: I am not sure 'blob' is appropriate here, and if they are so small how can you be certain these are pieces of wood? You mention wood remaining stationary, does that mean moving wood was not included in the study?

Response: We have moved this part to the discussion section, and adjusted the terminology throughout the manuscript. Also, we removed the word stationary as it was confusing.

Lines 254–264: Unfortunately, no supplementary could be found on the online interface for comparison. However, I do wonder if whether double panelling a figure to include one of these plots for clustering with figure 5 could help to show the variation.

I would also argue that the relative sizes of the bounding boxes compared to images were not that different, with many similar distributions and a few outliers, primarily from external datasets which is to be expected.

You also state that for 12, 18, and 19, the drop in relative size could be due to low camera resolution or distance from stream, but 12 is one of the model setup cameras so surely you know this, and could tell for the others by looking at the original images?

Response: we have added supplementary material, and included a figure to the revised manuscript. Also, we have added a part in which we compare images from different datasets in the

supplementary material and explain what we mean by different in quality. In fact, including some of this data posed additional challenges, as it was hard to identify and label the wood even manually.

Line 265: Assume this is meant to be Database Configuration.

Response: Yes, correct. Something went wrong here. Lines 345.

Lines 266-270: As this is both a results and discussion section currently, there is a lack of discussion here about why this may be, and that by oversampling images you may not see an improvement in model performance purely due to the model become more tuned to those specific examples.

Response: we have decoupled the two sections, expanded the results part and added a separate part in the discussion. Lines 379-383.

Line 270: This is very important, if you do not now oversample, in your scenarios where you mentioned oversampling smaller datasets, did you now not do this? This seems like quite a big change. If so, I think the sensitivity results need to come within the methods inclusive so that you do not explain changes in your methods during the results.

Response: The text was unclear. We have not used these results to adjust the methods, and now we clarified this and deleted.

Line 273: What were these results, and are they really comparable considering the differences in the object types?

Response: The results have a comparable mean average precision. We have added this to the section. Lines 377.

Lines 274-276: This section is not overly clear, I think it needs better wording to explain what is being done here, especially regarding the multiple training rounds. This feeds back into above comments at the end of the methods.

Response: This line summarized a larger part of the methods section in 1 sentence, and was indeed unclear. Therefore, we removed this sentence and indicated that the table shows the results from the training scenarios as explained it in the methods section. Lines 310-320.

Lines 281–284: I can see what is trying to be said here, about training for specific or general wood detection, but feel it could have been said better. This is also the first mention of how cameras were mounted, perhaps this should be mentioned in the data section also.

Response: we have added information on the mounting points of the cameras in the methods section. Also, we have adjusted the explanation to be more clear. Lines 355-359.

Lines 285 -293: There is a focus here on the high-definition wood images in this analysis, and yet there are only 9 images in the dataset. As such, are larger changes in mAP not more likely due simply to the lower number of objects to compare against? This is somewhat shown by the weighted average, and so overstating the importance of a vast performance decrease or increase here may be unjustified. The narrative however, that good wood images lead to

better training than poor wood images, is justified by the average and weighted average outputs.

Response: It is correct that if there are fewer samples in a validation dataset, the model missing 1 more piece of wood already drastically decreases its performance. However, all samples in the dataset are clearly pieces of wood with explicit characteristics of wood. So the model missing one or two more pieces does indicate that is it not as good at understanding the characteristics of wood. That being said, the fact that we state 'vastly' and show a large percentage is indeed unfair. Therefore we have revomed the statements of the amount of decrease. Lines 360-368.

Lines 294– 295: Has a significance test been undertaken here?

Are these broadly speaking not the only two factors, apart from manual labelling to begin with for training.

What are the worst performing models?

Response: we have now explicitly mentioned which models performed best and worst. Also, as the data quality seems to be the larger limiting factor, we have removed the algorithm from the statement, as it performs better with different datasets. Also, we changed the word significant. Lines 386-389.

Lines 296 –297: This sounds like you have added in an extra scenario, rather than describing one of your scenarios.

Change 'where the datasets with lower performance than 30% mAP were excluded' to 'where the datasets with a mAP of lower than 30% were excluded'.

Response: we are talking about scenario 11 here. We have added that information to the section. We also changed the wording following the reviewer's advice. Lines 390.

Lines 300– 301: Which scenario is this, can't find a reference to 19% in the table that is positive? If this is just assuming the inverse, then the addition of these images back wouldn't be the same 19% as the base conditions would be a different value.

Response: we are trying to describe the opposite, so the -19 percentage points we see with scenario 11 can be interpreted as a +19 percentage points when we add data of the same scene but from another day. We have made this explanation clearer in the text. Lines 392-396.

Lines 301 –306: This is a really important and useful point, and should be one of the key take home messages that adding to existing databases with some data from a site improves the algorithms performance. Check some wording here though, especially when speculating performance benefits.

Response: we altered this part. Lines 394-401.

Lines 307–313: This is an interesting section about whether the time component is critical. However, I fell it is overplayed in its significance. Of the two worst performing datasets (11 and 18) only one shows an increase of 6%, the other a decrease. Therefore, to say improvements of nearly 10% are made is an

exaggeration. Arguably, this is somewhat upstaged by the large decrease in one of the better performing datasets (3).

Response: we agree, and have rephrased it, although we believe it is still very interesting for future research. Lines 411-417.

Lines 313 –316: Make this a separate paragraph as it feels separate from the temporal component.

Compared to the emphasis placed on scenario 12, scenario 13 appears to show much greater performance gains, and the importance of image resolution in tracking wood. As this has implications for how wood should be monitored, both from a hardware and software perspective, it likely needs more attention and discussion around the trade-offs between image resolution, computational efficiency, and expected wood size.

Response: We have elaborated the disussion on scenario 13 in its own paragraph. Lines 418-423.

Line 315 references image 5, is this from figure 2 as these seem to be larger wood size, if not, please be clearer as to what this refers to.

Response: we meant to say that from the 6 reference datasets, the 3 that show the greatest improvement are the one that have the smallest relative Bbox sizes. We have made this more clear in the text. We have also elaborated on the implications for practitioners. Lines 421-423.

Lines 317 – Onwards: This almost feels like a different section or subsection, as it is a change from training and validating to assessing the model used. It seems as though this section itself however is limited in just comparing two models, moreover, these results have differences greater than many of the scenarios provided above, which indicates that model choice may be more important than datasets, something that is not discussed in great detail. As a result, the take home would switch from the importance of data, to the importance of model selection in getting the best outputs...

Response: we have created a new section for this part. The differences between the models are indeed larger than between many of the scenarios. We have adjusted the text as well. Section 4.2.

Line 329: Perhaps, if a new subsection is introduced for the above, this should be moved prior to it.

Response: we have moved this part up. Lines 426-437.

Line 334: Reference figure 7 here, as it is not referenced anywhere in the text

Response: we have referred to the methods section where we have elaborated on this dataset. And we have referred to figure 7. Lines 425.

Lines 335 –337: You identify that the model is better at identifying large wood, and then state how large wood components compromise the greatest proportion of transport, but this needs to be referenced to support this. Furthermore, small wood components also play a role in increasing the total volume of log jams etc and so important to monitor. Commenting on how this is missed in the dataset is probably needed.

If possible, it would be great to look at those that are missed and estimate the size of these to identify a limit of detection. However, that may be beyond the scope of this investigation and potential for future research.

Response: We have added references. Lines 432. It would indeed be interesting to find that out. And we are working on that, but is indeed beyond the scope of this manuscript.

We have adjusted the section to also indicate the importance of small wood and measured that can be taken to have a better change of detecting small wood also.

Line 338: Have these images been georectified in the processing? If so this needs to be explained for reproducibility. Moreover, if they have then they could be used to identify the limits of detection for wood as per above?

Response: no, they have not been georectified in this study. We are elaborating on the potential of the method. However, as this was confusingly phrased, we changed the wording. Lines 432-437.

Line 342: Give examples here please, and comment on how they may differ or align to wood detection (e.g. shape and background).

Response: we have added an example (humans walking through the frame). Lines 455-459.

Line 342–346: think this needs to be reworded, at times this sounds speculative and also non-scientific. The theory of not being able to detect outside of the training sample is sound, just the transmission of this information is not clear enough.

Response: we have adjusted this section and added examples. Lines 451-452.

Line 347: Where was this from and why not use one of your current data? Again, this points to questions going back to your initial data introduction, and consistently adding new bits of information.

Response: we have elaborated on the source of the image and added it to the methods section and better explained why this image was used. Lines 460-462.

Lines 350– 360: Does this not come back to simple survey and image design. If most of your images are from roads and bridges overlooking rivers, and you provide an image much closer to the channel, it will struggle, until as you say you include images of large bits of wood close up. Therefore, to use the word remarkably again seems a little overstated.

Response: with the word remarkable we do not necessarily mean something positive. More like something we might not have expected. Lines 464-475.

Lines 357-358: Can you expand on how you know it is using the wood texture, is this hypothesised from the location of the pixels used, or can this be proven?

Response: you are right, we have no way of knowing that the model actually understands the texture of wood, and therefore the wording was too enthusiastic. We have changed the wording to be more careful. Line 472.

Lines 363–371: This is a nice start to your conclusion, summarising your results well to give an overview of the paper. However, there is no comment on how increasing data sizes or changing their angles/mirroring had no effect.

Response: we have added this to the conclusion. Line 493.

Lines 372 –382: I feel that to say your model struggles on the definition of wood, unless its given high-quality images of wood not in rivers, is overly harsh on your model. The purpose of this paper and method is to detect wood in rivers, likely from monitoring stations above the rivers surface (on bridges etc). So the model works if it detects these well, and shouldn't necessarily be able to detect wood such as in Figure 8. Therefore, the model CAN generalise the concept of wood 'in rivers', which is the main purpose is it not?

I think the word blob should be removed throughout, perhaps in this instance they are best referred to as fragments or segments, i.e. not all the wood is on show? Make sure this distinction is first explained when replacing the initial occurrence of the word 'blob'.

This may be clarified by an earlier point, is this 19% increase simply the opposite of the 19% reduction when the Allier dataset (18) is removed? If so. This is not 19% (e.g. 20% decrease from 100 is 80, a 20% increase from 80 is not 100). If this is a separate analysis, make sure this is clear during the methods and results. It could even be viewed as an additional scenario (e.g. adding same site from different date).

Response: We have adjusted the explanation about the model understanding the concept of wood.

-we have removed the word blob from the paper.

-you are right that we made a mistake in explaining the percentages. We have adjusted the explanations to the words 'percentage points' where we made this mistake.

Lines 383– 387: This could likely be grouped into areas of future research. 1) real time monitoring 2) algorithm development and miniaturisation 3) temporal imagery for object detection. These could also form some structure for a separated discussion, allowing room to discuss the impacts of the research.

Response: we have adjusted the last part to more clearly touch upon those points. Lines 507-511.

Figure 1: This figure could benefit from labelling the boxes with the sections of the method that they refer to. This will allow readers to quickly understand which bit of the process they are referring to. Make sure the naming matches to, it will help the reader.

This could also be improved by creating this as an overall schematic of the methods, which would better describe the whole process as mentioned prior.

Response: we have adjusted the titles in the text and in the figure to correspond. We have also added the section numbers to the figure.

Figure 2: It is great to see some visual examples of what these images look like, and how they differ, especially in regard to the additional imagery. However, I think it would be good to possibly

remove one or two images, and add an inset location map showing where in the world these were taken from, rather than coordinates in the caption. This would give a better idea to the readers of where your data is coming from. You could colour or size location dots based on the number of images from a location as well.

Response: we have added the locations to the images instead of coordinates.

Table 1: Could this table also have a column or some stars which denote the datasets used in validation, these are mentioned later on but will help the reader when scanning back and forth. Consider making either camera lowercase, or the unknown and differing upper case.

Response: we have indicated the 6 representatative datasets and added more information about the cameras.

Figure 3: Why is this figure not further up in the manuscript? It is referenced first several pages earlier and causes confusion in the current section. Appreciate this may just be a current formatting error for the preprint.

Response: it is now further up in the manuscript.

Figure 4: No changes required for this figure, it is clearly laid out, shows the size of datasets, and helps to explain what is happening in terms of the number of training vs validation datasets.

Response: thank you, we kept it as is.

Figure 5: Again, another clear figure which adds to the manuscript and is broadly easy to interpret. The inclusion of a double headed arrow along the x axis, pointing to larger wood and smaller wood may help with interpretation, so readers know if the value is indicating a lot of the image is the woods bounding box, or little.

Response: thank you, we kept it as is.

Figure 6: This figure is good, however it could do with stretching along the x axis, as this will help to show the variation in IOU training loss which show subtle differences.

Response: thank you for the suggestion, however the detail in the image does not allow for stretching. The bandwidth between the epochs is already clear in our eyes.

Table 2: The table layout is fine, but the text is a little hard to read in places. For those reading in non-colour or with colour-confusion, perhaps as well as colours a marker could be used to quickly attribute greater than 3% increases or decreases.

Response: colours are actually not allowed in esurf tables, so we indicated them with stars.

Figure 7: A useful figure, make sure it is referenced in the text. Are these bounding boxes ones predicted by the model or drawn manually for users. It could be better to include boxes created by the model as well to show the types of

wood it is missing (perhaps detected and missed wood as two separate colours?).

Response: we want to show an example of the dataset here and not qualitatively go into specific pieces that are missed by the model. For the analysis, we only use mean Average Precision in the text. We have referenced it in the manuscript.

Figure 8: Are the bounding boxes in this figure manually drawn? If so, they should probably better align with the extent of the wood. Likewise, as the percentage is referring to overlap in bounding box size, perhaps indicating the bounding box of the detected wood would help to illustrate these differences? Otherwise, this is a very helpful and useful figure.

Response: The bounding boxes are created by the model and are, therefore, not perfect. Here we go into the detections qualitatively. We have adjusted the explanation of the image to stress these are model generated boxes.