Reviewer 1

I am pleased with the authors responses to my original review. I have a few remaining questions/corrections, which range from technical to minor.

Relating to the extended analysis of the MAPE of the accuracy ratio of CNNS trained with different fractions of flooding (Figure 9A): Did the authors test the MAPE accuracy with a proportion of more than 26% of flooding images?

Reply) For simplicity we kept the number of images constant and 26% was the maximum fraction of flooding images under that constraint.

There is also a decline in accuracy between 10 and 15%, so I'm wondering if the decline in accuracy the authors note continues past 26% or not?

Reply) We are not sure, however the decline in accuracy from 23% to 26% is much more noticeable and might indicate to an over-reliance of flooding images in the training, to the detriment to the correct identification of water fraction in non-flooding, or "normal" images. We look forward to testing this hypothesis in the future.

Following on, I believe that the authors see the lowest MAPE when 23% of the training set was composed of flooding images. It looks like the validation set was 21 flooding image and 70 normal images (line 215) making the percentage of flooded images in the validation set also 23%. Is this a coincidence? I suggest testing a different proportion of flooding images to ensure the best accuracy ratio proportion for training is independent of the percentage of flooded images of validation data used for validating/testing.

Reply) We indeed think it is a coincidence, however we can test a different proportion in future work.

What are the limitations of this approach? For example, if a camera moves, would the accuracy of the predictions be at risk?

Reply) The approach is very general. Note that we used three different cameras, with drastically different fields of view, and some of them indeed moved during the field campaign. However, the final result is quite homogeneous. See, for example, the lack of trend between the different symbol's color in Fig.5.

Line 240: Maybe I'm misunderstanding something, but isn't this statement contradictory? Accuracy of CNN predictions independent of amount of water in images, yet it was better for images with a larger fraction of water? This would tell me it might be dependent, because more water = more fraction of water?

Reply) These statements refer to two different things. The accuracy of the raw CNN predictions indeed is better for flooding images (Figs. 5 and 8A, open symbols; and Fig. 9). However, the accuracy of the filtered CNN predictions, depends little on the amount of water (Fig.5 and Fig. 8 top, filled symbols). We clarified this distinction in the revised version.

Line 252: is precise the right word to use here? Considering its just pixels from the images, I'm not sure it would be considered a precise observation of beach or back-beach overtopping. How is overtopping defined based on the imagery?

Reply) We removed the word precise to avoid confusion. By definition, the flooding of the beach or back-beach, as measured by the identification of water pixels on the beach/back-beach region, represent an overtopping event.

Line 253: I suggest editing this sentence to say, "This method has the potential to enhance…." Since it hasn't been combined with photogrammetry this is untested, so I would leave it as it has the potential, rather than it can enhance…

Reply) Done.

Figure 2: I suggest the authors label the figure to show which photos are considered flooding and which are considered normal.

Reply) Done.

Figure 3: I suggest the authors label “top row” and “bottom row” rather than just “top” and “bottom”.

Reply) Done.