Anonymous Referee #2

Received and published: 24 June 2017

The authors use Cryosat-2 SAR data for inferring water levels over the Mekong river system. As with the earlier pulse-limited radars, applying SAR for inland water bodies requires a (static or time-variable) land-water mask for separating water echoes from land returns, that is usually derived from optical and/or radar remote sensing. In this article, the authors suggest to derive the mask from the SAR Delay-Doppler stacks itself, using a standard classification method. It must be noted that Cryosat-2 levels, due to the satellite's unusual 'repeat' orbit, are difficult to validate. SAR promises to enable water level measurements for rivers of width down to few 100 m or even less, and the availability of river masks poses a challenge, in particular when seasonal inundation is present. The work is timely and touches upon a relevant topic.

However, the scientific hypothesis and the paper's objective are not well-described and it is difficult to read. In large parts, the article is written in an explorative style: the authors apply a certain sequence of approaches and report about success, but it is not explained why these particular approaches are chosen nor are systematic tests provided. This applies to the classification approach (k-means), and the same goes for the features chosen to be used for classification.

Thank you for this concern. More than the presented features were tested. But some were not sensitive for the water classification or redundant with one of the used features (tested with correlation between features). We added in the text at the end of the feature presentation:

All of these features were chosen due to their sensitivity for the posed problem of water classification and independent from each other. More features were tested but discarded because they were either not sensitive for the classification or highly correlated to one of the used features.

As for the k-mean algorithm: We needed an unsupervised classification algorithm as we do not have reliable training datasets for a supervised classification. The k-means algorithm is widely used

and was already used for Cryosat-2 waveform classification (Göttl et al.). We also tested the k-mediods algorithm but found no real difference between the results.

We added in the text to the k-mean algorithm in section 4:

An unsupervised clustering algorithm is used as no reliable training data is available. The unsupervised k-means clustering algorithm is widely used and was already tested for waveform classification in Göttl et al. 2016.

The authors jump back and forth with approach and validation. I would suggest a more systematic writing: data, method, results, validation, interpretation.

This is a point we discussed among the co-authors before submitting the paper. We found that because of the diverse results it is easier to incorporate the discussion directly into the result and validation section.

Also, it would be helpful to have 1) a flowchart for the approach including region subdivision, classification, retracking, and outlier removal, and 2) a flowchart for the validations.

Thank you for this very good suggestion. We added a flowchart of the approach at the beginning of manuscript. As for the validations, we think that such a flowchart would not be as helpful. The steps of the validation are in parallel and not like the steps of the approach in a hierarchical order. But we added at the beginning of section 6 a table summarizing the different validations done.

Other remarks

Page 2 lines 24-30: Here, the authors somehow suggest the range-integrated power, RIP, provides a kind of independent observation that is not available for other altimeters. Although it is true in a literal sense I find this slightly misleading: In fact, both the RIP and the SAR waveforms are derived from the stacks (Fig. 2) which contains the primary observable. In fact all the features they used

can be interpreted as properties of the stack matrix. This leads to the question whether the authors actually use the multi-look stacks for focusing on the rivers when deriving water levels? This should be answered in section 3.

We changed the text to:

For the land-water classification a set of features derived from the Cryosat-2 stack data over the intermediate step of the waveform and the RIP is used. The features are summarized in Table 1.

We are not sure if we understood the second question correctly. Should we have used not the full stack matrix for the waveform which in turn is used for the height determination? In fact we tested using not all single looks of the stack but only those singe looks with high power to calculate the waveform. Another test we did was retracking all single looks and using only those with equal height. Nonetheless, both ways were not improving the results as far as we could validate. Due to the higher computational load this idea was discarded. We state already clearly in section 3 that we use all single look waveforms for the SAR waveform.

In my printout of figure 1, the upstream / mid-stream / downstream mask hachures are not shown as indicated in the inset.

This seems to be a general problem (see review #1). We hope that we solved it now for all platforms and printers.

The authors argue that for the smaller rivers of the Mekong basin no reliable land-water mask is available. But there are definitely regions of the world where very precise land-water masks are available (e.g. all EU) – why not testing the classification approach properly for such a region? If this paper is meant to provide a new method, it would be perfectly ok to add comparisons in a test region outside the Mekong.

It is true, that all rivers inside the EU are very good mapped. But transferring the problem to Europe poses major difficulties. Europa has no nearly as large and diverse river systems as the Mekong which is measured in SAR by Cryosat-2. The Danube River is in SARin and the Po River is according to the mode mask the only larger river measured in SAR. The Po River is not comparable in terms of size and complexity of its surrounding topography to the Mekong River.

As the other reviewer pointed out, the mask available for the Mekong with 30m accuracy is very precise for rivers. Therefore, the comparison between the classification and the land water mask shown in section 6.3.2 is what you ask for.

Eq. (1): the terminology is awkward. Peakiness pwf is not a vector, why is the symbol bold?

Max(wf) should read max-over-I wfi

We changed the symbols

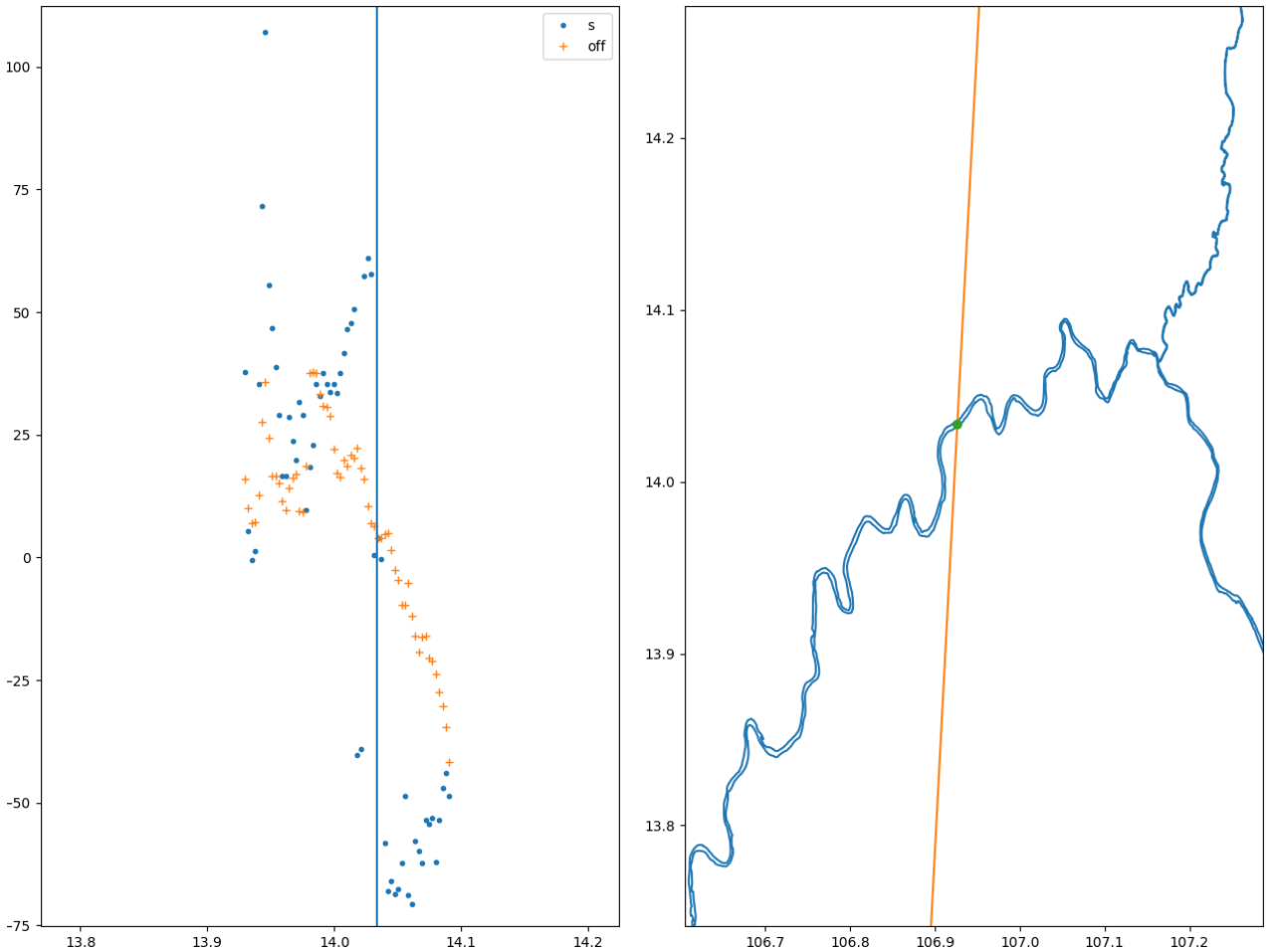

Figure 1: Symmetry (blue) and off centre (yellow) features for one river crossing (left). The y-axis are the values of the features, x-axis is latitude.

The pass (yellow) is ascending and the river (blue) flows towards South-West (right).

Can the authors please provide an example for the RIP asymmetry parameter caused by a realistic river sloped, assuming the satellite flies along the river (from 246 single looks)?

Please see figure 1 in this document. While approaching the river the surface is sloped towards the satellite and both s and off are positive. s is even decreasing towards the river and then 'jumping' down after the river is passed. How can this be explained? While approaching a water target the overall returned energy increases but the side of the stack towards the water receives relatively more energy which causes the bigger unbalance between the two sides.

The slope of the line of s before and after the crossing are different, after, where the river is sloped away from the satellite, it is flatter than before the crossing.

The slope of the river in this part is between 30 and 40cm/km. We do not think that this example is worth including in the paper.

We also added in the text:

A positive s indicates a water surface sloped towards the approaching satellite.

From the subdivision of the overall basin in three regions, the authors find that in the overlap regions the classification (and thus water levels) does not always agree. That's worrysome but may be expected given the approach. But what is the implication for applications?

First of all, the result shows that the classification approach proposed in this study is not a method one to fits it all. The characteristics of water returns differ too much between small and narrow rivers and wide open rivers to be classified with the same parameters. This is something a user has to keep in mind for such a classification; the classification is most successful with a homogeneous target. Still, most heights agree in the overlap considering the accuracy of satellite altimetry over inland waters.

After classification, the water surface is identified in retracked levels through searching for a horizontal (or, I guess, sloped) line. Outliers at the margins should tell about the misclassification in the first step. Is it possible to quantify this, e.g. in % of identified water surface from k-means vs. the straight-line in water levels?

Yes, we search for a horizontal line, but only if enough data points are available (more than 5), which is not that often. We now quantified the number of heights discarded and also looked into the spatial distribution of those discarded. Around 90% of the classified measurements are taken for the height determination. Many of those discarded are apart of the cluster of heights taken for the water level estimation. It seems a reasonable assumption that those measurements are not river reflections but could come from ponds or paddy fields. As we explained in the text another

reason for the outliers could be off nadir effects which occur in SAR data across track. In both cases the classification is not wrong because it classifies correctly water. But we do not want all water but just river water. We do not think that this investigation is interesting enough for the paper.

k-means can be seen as a special case of maximum likelihood classification under assuming Gaussian distribution and spherical clusters – the distance measure is not weighted in the original method. But weighted k-means may be appropriate whenever features have different geometric meaning and/or units. Are all features equally weighted, does this really makes sense?

It does not make sense to use the unweighted k-mean clustering if the units and orders of magnitudes of the features are different or the variance is different. Therefore, we normalized all features before the clustering which also reduced the differences in variance. Still we assume the equal Gaussian distribution for the features. We tested the features for this; all except the maximum power feature passed this test.

We added in the text:

The k-means algorithm assumes normally distributed features with equal variance, which we ensured by the normalization of the features.

7 Conclusions: ‘We demonstrate in this study the possibilities of classifying CryoSat-2 SAR data in the Mekong river basin and using this classification for water level extraction’. This statement carries no information at all – please be concise and provide real conclusions (the above is not even a summary).

We changed the sentence (also according to suggestions of other reviewer) to:

We demonstrate in this study the advantage of CryoSat-2 SAR altimetry data for measuring rivers which are identified by a classification, which is independent of a precise land-water-mask.

Eq. (3) While (3) and (4) may have been 'derived' from the OCOG retracker, they represent standard statistical measures. I find it misleading to refer to the OCOG in this respect, which simply makes use of the same statistical moments.

We removed the reference to the OCOG.

There are some thresholds chosen for the outlier selection based on the near-annual repeat of C-2, height difference of 7 m, 10km / 30day spacing in the second step. It is said these are based on a conservative approach, but more should be provided on how robust the overall results are with respect to these thresholds.

The time of 30 days has only a small impact, if we take anything larger than 30 days it does not change the results. For 10 days only very few data points are additionally kept but these do not have any influence on the final validation approach. Changing the distance of 10 km has similar small effects as the time spacing. But it should be kept in mind, that if the spatial and temporal spacing is too small no comparison can be done and the measurement is not considered an outlier. And the mean height to which the height is compared to is weighted by distance in space and time.

As for changing the 7m the effect is a bit larger. Unsurprisingly, a value less than 7m results in more outliers and vice versa. Looking at Figure 9, the threshold determines where the tail of the distribution is cut off. As one can see, the number of differences in this region of the histogram is very small; therefore a shift of the threshold is not affecting most of the observations. If one looks at the two regions separately the upstream region has a heavier tail of the histogram and thus the threshold of 7 m has a higher influence.

We added in the text:

Of the three thresholds used for the outlier detection the difference of 7\,m in between years is the most sensitive for the later result. The time and distant weighted mean in the second part of the outlier detection limits the sensitivity of the other two thresholds.

Appendix: All four figures have the label A1?

This was an error of the Latex Template we were now able to fix.