We would like to thank Reviewer#1 for their careful consideration of this manuscript and for their helpful and insightful comments. We have carefully considered the reviewer's comments and worked to include them in the revised version of the manuscript according to the proposed suggestions.

Please find below the responses to the reviewer's comments.

General comments

This paper is an interesting evaluation of the skill of the Global Flood Detection System to measure river discharge from satellite passive microwave signals, and is certainly worthy of publication after some correction. The correlation between the daily ground station-measured water discharge and the satellite signal is measured for a range of rivers with different widths, floodplain areas, land cover types, climatic regions and other factors. For African, Asian and North American rivers, the mean R values are less than 0.5, and the correlation is only medium. Only European and South American rivers give high correlation (>0.5). It might be argued that a judicious set of ranges of R has been employed (R < 0.3, 0.3 – 0.7, >0.7), in which many rivers lie in the middle range, but still may have R values < 0.5, so that the correlation is only medium. The authors should comment on this. The relatively low R values show the difficulty of obtaining a reasonable signal-to-noise ratio from a 10km pixel when the flood width is often substantially less than the pixel width. As a result, it is obviously a sensible idea to identify sites where the method will work because of the associated site variables, and use these for future studies, rather than trying to make the method work for all sites. The method would also appear to work best for detecting floods rather than forecasting them, since a 4-day average signal is used, partly to cope with the time lag between changes in stage at a gauging station and associated changes in flood extent.

Specific comments

7333/14: Make it clear that you are talking about river floods (or does this include deaths in the tsunami of 2001?).

Author's reply: Modified in the manuscript as suggested by the Reviewer.

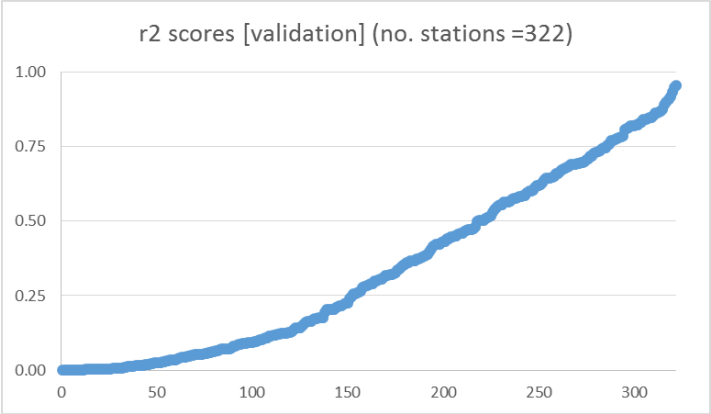
7337/6 In a flood situation, is the error on the observed discharge not higher than the 5–20% quoted?

Author's reply: Explanation added on the manuscript as suggested.

The uncertainty of river discharge is higher during floods events when the stage-discharge relationship, the so-called rating curve, is used. As evaluated by Pappenberger et al. (2006), the analysis of rating curve uncertainties leads to an uncertainty of the input of 18–25% at peak discharge. Di Baldassarre and Montanari (2009) showed that the total rating curve errors increase, when the river discharge increases and varies from 1.8% to 38.4% with a mean value of 21.2%.

7342/13 What was the spread of the R2 values for the fits?

Author's reply: answered below.



r^2 scores [validation]

[0-0.3] = 163
[0.3-0.5] = 54
[0.5-0.75] = 68
[>0.75] = 37

Figure 1. r^2 scores obtained on the validation.

In Fig. 3b, please make it clearer that different rating equations are being used for different months, not simply that in fig. 3a.
Author's reply: Clarified in the figure caption as suggested by the reviewer.
In fig. 3a, why aren't there 15 points on the graph, one for each March between 1998-2002?
Author's reply: The calibration was done for 5 years (1998-2002), therefore the five points represent the five mean values for March in this case.

7344/20 A little more description of the **Gini index** might help the reader.
Author's reply: explanation added on the manuscript.

Gini´s mean difference was first introduced by Corrado Gini in 1912 as an alternative measure of variability and the parameters derived from it, such as the Gini index, also referred to as the concentration ratio (Yitzhaki and Schechtman, 2013). The Gini index is mostly popular in economics, however it is also used in other areas, such as building decision trees in statistics to measure the purity of possible child nodes, and it has been compared with other equality measures (Gonzales,L., et al. 2010).

How does the **Random Forest** method cope if the variables are correlated (as e.g. discharge and river width probably are)? Is the correlation between variables output from the method as they would be from a principal component analysis? If so, it would be useful in the subsequent analysis to know the correlations between variables to know which were most significant.
Author's reply: explanation added on the manuscript.

The random forests algorithm, introduced by Breiman (2001), is a modification of bagging that aggregates a large collection of tree-based estimators and has better estimation performances than a single random tree: each tree estimator has low bias but high variance whereas the aggregation achieve a bias-variance trade-off. This algorithm has good predictive performances in practice, they work well for high dimensional problems and they can be used with multi-class output, categorical predictors and imbalanced problems. Moreover, the random forests provide some measures of the importance of the variables with respect to the prediction of the outcome variable.

The random forest algorithm was selected instead of the Principal Component Analysis as we had mixed data types because some of the variables to be study were categorical instead of continuous.

Although, the effect of the correlations on these measures has been studied recently (see Archer and Kimes (2008), Strobl et al. (2008), Nicodemus et al. (2010), Nicodemus (2011), Auret and Aldrich (2011), Tolosi and Lengauer (2011), Grömping, U. (2009) and Gregorutti et al. (2013)) there is no yet a consensus on the interpretation of the importance measures when the predictors are correlated and on what is the effect of this correlation on the importance measure.

In order to test the effect on the results when correlated variables were included in the analysis, an independent Random Forest analysis was carried out during the analysis (not shown in the paper) for the same variables but excluding the river width and the presence of floodplains and wetlands variables. Results also showed that the mean daily observed discharge had the highest importance and the presence of hydraulic structures (mainly dams) and of river ice had the lowest importance to classify a location as good or poor performance.

7345/25 Do you really mean that the signal may have a large natural variation, or that the noise is instrument noise?
Author's reply: answered and edited on the manuscript as follows.
We meant that the signal to noise ratio might be low for a site or have intermittent instrument noise occasionally producing intermittent positive spikes in discharge. We have edit this in the manuscript.

73446/8 R = 0.3 is chosen as a threshold in fig. 4, yet this is only a medium correlation. What happens if you chose R = 0.5 as the threshold, are there too few sites satisfying this criterion then?
Author's reply: For this study, 42 sites have R>0.5

7346/23 In fig. 5, in the eastern USA, many stations had R > 0.3 in the calibration (fig. 4), but have NSE < 0 in validation. Why is this? The rivers are presumably often wide and on floodplains near the sea at these observation points?
Author's reply: explanation added on the manuscript. Not the map below as it is complementary to figures already shown in the manuscript.
Figure 5 doesn't shows the calibration score. It shows the initial correlation between GFDS signal vs. in situ observed discharge.
The figure below (Fig. 1 of the Author's comments) shows the R score obtained during the validation for stations located in Easter USA (no. of stations=66). In addition, it shows the pixels values when the river width is higher than 1km (Yamakazi et al., 2014) and the Global Lake and Wetland Database layers (Lehner and Doll, 2014).

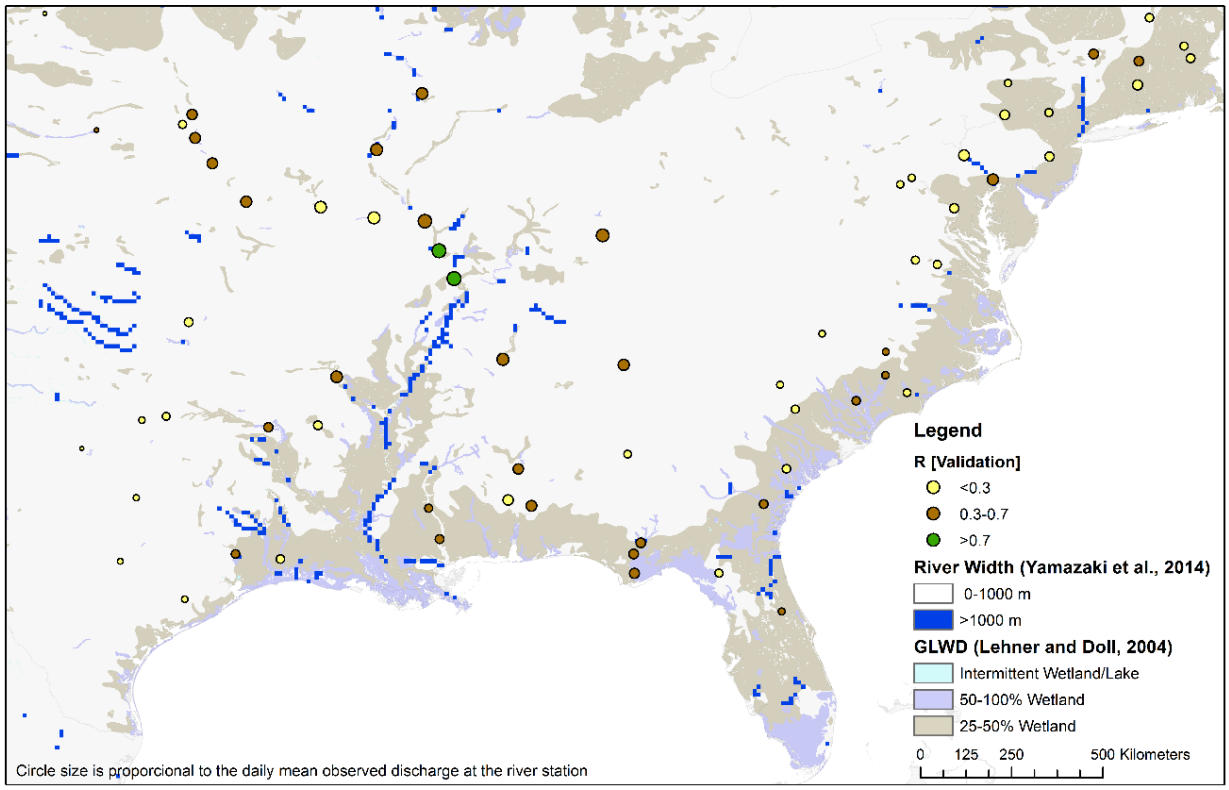


Figure 2. R score of the validation (n=66 station) for Eastern USA.

As shown in the manuscript for the whole of the stations, we conclude that most of the stations in this region obtained poor scores due to a number of factors: ~64% of these stations have a mean discharge value lower than 500 m^3s^{-1} and ~88% of the stations are located at river width lower than 1km. In addition, ~59% of the stations are located in wetlands areas. Sites with these characteristics might not provide useful outputs when aiming to measure river discharge through the use of satellite flood signal, as it is the case of some of this stations.

7347/10 It is probably true that locations with a river width higher than 1 km are more likely to score an R larger than 0.3, but it would be worth quantifying R for widths > 1km and showing that it's significantly larger than 0.3.
Author's reply: Quantification added on the manuscript.

The mean R score is 0.60. Where 26 out of 64 (~41%) have R> 0.75.

A related point is, in fig. 6a, could you explain why some rivers of 100m or less width have R values as high as the widest rivers? Intuitively you would have thought the brightness temperature for a pixel containing water would depend on the river width (perhaps I'm confusing the river width with the flood width here?).
Author's reply: explanation added to the manuscript.

The retrieval of the satellite signal also depends on the floodplain geometry. As soon as the river floods and water goes over-bank, the proportion of water in the wet pixel greatly increases. So the score should be also high for small rivers with a proportionally big floodplain.

7347/24 might not provide reliable results: : It would be better to quantify this rather than just stating it. You could use a statistical test to compare the rivers with Q < 500m3/s that have R < 0.3 with rivers with Q > 500 that have R > 0.3, and show that they were significantly different.
Author's reply: Quantification added to the manuscript.

As 77% of the stations with Q<500 m^3/s, have R< 0.3, while 91.5% of the stations with Q>500 m^3/s have R >0.3, locations with discharge of less than 500 m^3s^{-1} might not provide reliable results for a global satellite-based monitoring system.

Technical corrections
All comments were adapted according to the Reviewer's suggestion.

7333/16 Golnaraghi 2009 and Kundzewicz 2012 refs missing /28 UNOSAT 2013 ref missing.
Modified
7335/20 climate-drive -> climate-driven /27 global -> a global 7337/9 us -> as
Modified
7340/4 Example -> Examples /17 define M/C signals /22 split sentence at 'an array'
Modified
7345/8 as validated -> were validated /13 calibrate -> calibrated /14 discharge satellites -> satellite discharge
Modified
7346/16 two-years -> two years /20 shorted -> shorter
Modified
7348/8 25x 25 pixel -> 25 x 25 km pixel /28 To note -> Note
Modified
7349/22 Where highest -> The highest
Modified
7350/8 presence or not -> presence or absence /20 for - the most of – the -> for most of the
Modified
7351/12 in some -> on some
Modified
7352/2 test -> tested
Modified
7354/ fig 12 caption: was chose -> was chosen; of the stations -> or the stations; station -> stations
Modified
7353/2 replace the semicolons with commas in this long sentence /20 satellite measured -> satellite-measured
Modified
7354/10 no verb in sentence /15 a more -> more
Modified

References

Archer, K. J. and Kimes, R. V. Empirical characterization of random forest variable importance measures. Computational Statistics and Data Analysis, 52:2249–2260, 2008. doi: 10.1016/j.csda.2007.08.015

Auret, L. and Aldrich, C. Empirical comparison of tree ensemble variable importance measures. Chemometrics and Intelligent Laboratory Systems, 105:157–170, 2011. doi: 10.1016/j.chemolab.2010.12.004

Breiman, L.: Random forests, Mach. Learn., 45, 5–32, 2001.

Di Baldassarre, G. and Montanari, A.: Uncertainty in river discharge observations: a quantitative analysis, Hydrol. Earth Syst. Sci., 13, 913-921, doi: 10.5194/hess-13-913-2009, 2009.

Gonzalez, L., Velasco Morente,F. , Gavilan Ruiz, J.M., Sanchez-Reyes Fernandez, J.M. The Similarity between the Square of the Coeficient of Variation and the Gini Index of a General Random Variable. Journal of Quantitative Methods for Economics and Business Administration 10: 5–18.2010, ISSN 1886-516X.

Gregorutti,B., Michel, B., Saint-Pierre, P. Correlation and variable importance in random forests. Cornell University Library, 2013. arXiv: 1310.5726 [stat]

Grömping, U. Variable Importance Assessment in Regression: Linear Regression versus Random Forest. The American Statistician. 11/2009; 63:308-319, 2009. doi: 10.1198/tast.2009.08199

Lehner, B. and Döll, P.: Development and validation of a global database of lakes, reservoirs and wetlands, J. Hydrol., 296/1–4, 1–22, 2004. doi: 10.1016/j.jhydrol.2004.03.028

Nicodemus, K. K. Letter to the editor: On the stability and ranking of predictors from random forest variable importance measures. Briefings in Bioinformatics, 12:369–373, 2011. doi: 10.1093/bib/bbr016

Nicodemus, K. K., Malley, J. D., Strobl, C., and Ziegler, A. The behaviour of random forest permutation-based variable importance measures under predictor correlation. BMC Bioinformatics, 11:110, 2010. doi: 10.1186/1471-2105-11-110

Pappenberger, F., Matgen, P., Beven, K.J., Henry, J., Pfister, L., Fraipont de, P., Influence of uncertain boundary conditions and model structure on flood inundation predictions, Advances in Water Resources, Volume 29, Issue 10, Pages 1430-1449, 2006. doi: 10.1016/j.advwatres.2005.11.012

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. Conditional variable importance for random forests. BMC Bioinformatics, 9:307, 2008. doi: 10.1186/1471-2105-9-307

Tolosi, L. and Lengauer, T. Classification with correlated features: unreliability of feature ranking and solutions. Bioinformatics, 27:1986–1994, 2011. doi: 10.1093/bioinformatics/btr300

Yamazaki, D., O'Loughlin, F., Trigg, M. A., Miller, Z. F., Pavelsky, T. M., and Bates, P. D.: Development of the global width database for large rivers, Water Resour. Res., 50, 3467–3480, doi: 10.1002/2013WR014664, 2014.

Yitzhaki, S., Schechtman, E. The Gini Methodology. A Primer on a Statistical Methodology. 2013. Springer Series in Statistics. Volume 272, 2013, ISBN: 978-1-4614-4720-7.