Editor Decision:

Reconsider after major revisions (24 Jun 2016) by Dr. Stacey Archfield
I have re-read the revised manuscript, the author responses to the first round of reviews, and the reviews of the current version of the manuscript.
Both reviewers note that additional revision is needed for the paper to proceed. I find that Reviewers 1 and 2 both have expressed concern that the initial reviewer comments are not fully addressed, particularly with respect to the limited sample size utilized to support the author findings.
The authors initial response to this comment indicated that there is more data to support their findings but that the authors did not want to include these additional results because it would add too much length to the manuscript. I felt quite the opposite; if there are more data to support the findings, this would increase the quality of the manuscript and strengthen the case for the use of crowd-sourced information for hydrologic modeling. Since this is, as the authors rightfully point out, one of the first contributions on this topic, additional case studies are needed to convincingly argue that the use of crowd-sourced information is viable.
Another place of particular revision that is needed is in the separation of the results and discussion. While the authors may feel the current organization of the manuscript is clear to them, Reviewer 1 has stated it was not clear in their reading. I agree that it would be a cleaner presentation of the results to have these sections separated.
This manuscript has the potential to be an important and novel contribution to the literature and I believe that the changes requested by the reviewers are constructive and will further enhance and strengthen the contribution.
I look forward to reading the next revision of the manuscript.

Dear Editor,
We greatly appreciate the time and efforts by the editor and referees in reviewing the manuscript. Following editor's suggestions we have improved the manuscript including additional flood events (1 more for the Brue catchment and 2 for the Bacchiglione) and 2 additional case studies with 3 flood events in each one, for a total number of 12 floods events instead of 3 as in the previous version of the manuscript.
Overall we found similar results in all the considered flood events and case studies. In fact, assimilation of crowdsourced observations in Brue, Sieve and Alzette catchments showed the same model behavior. The only difference can be found in the magnitude of model improvement, or μ(NSE), which is related to NSE values in case of no model updating, as reported in figure 9. In addition, the simulations required by Reviewer#2 in the previous review have been added (Figure 11). Overall, figures 2, 7, 8, 10, 11, 15, 16 and 17 have been improved included the new simulations. However, in order to reduce the total number of figures we have decided to remove three figures, 8, 9 and 13 in the previous manuscript version, which did not provide additional information to the manuscript.
As requested by both reviewers we have also separated results and discussions, using a proper color-coding to identify scenarios from 2 to 9. The evidence of the updated text has been marked using the blue color, while the red text is related to the first review.
Kind regards on behalf of all authors

Maurizio Mazzoleni

Anonymous Referee #1:

I appreciate the authors' efforts with the revision. My main concern, however, on the limited amount of test cases being used, was only addressed with some text and not with adding more events and/or catchments. One could say two catchments are sufficient for this first study (even if I would still like to see more ...), but only using 1 or 2 events just does not make any sense to me. Depending on the event details (including data quality, model performance, ...) results for single events could be largely different. I also do not understand why the authors did not include more events, as this would be rather straight-forward given that the model already is set up for the catchments.

To be honest, as long as this major limitation is not resolved, I see little use in fully reviewing the manuscript again.

We thank the reviewer for this important comment. At a first stage of this review we have decided to include 3 more flood events for the Brue catchment. The results we found are reported below.

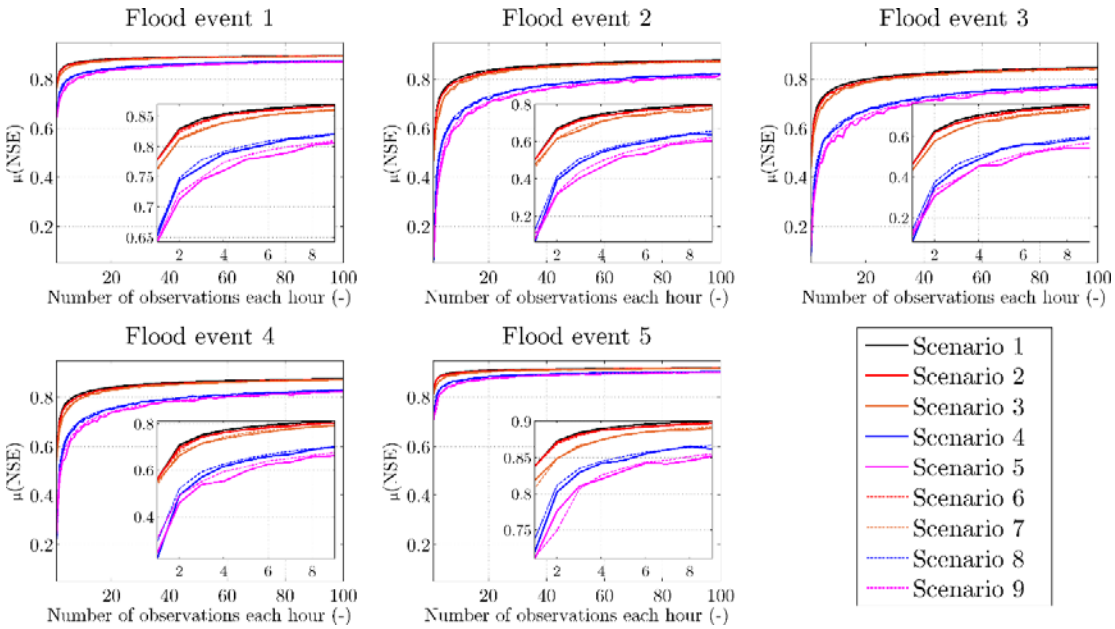

Figure 1. Dependency of μ(NSE) on the number of observations, for the scenarios 2, 3, 4, 5, 6, 7, 8 and 9 for the five considered flood events

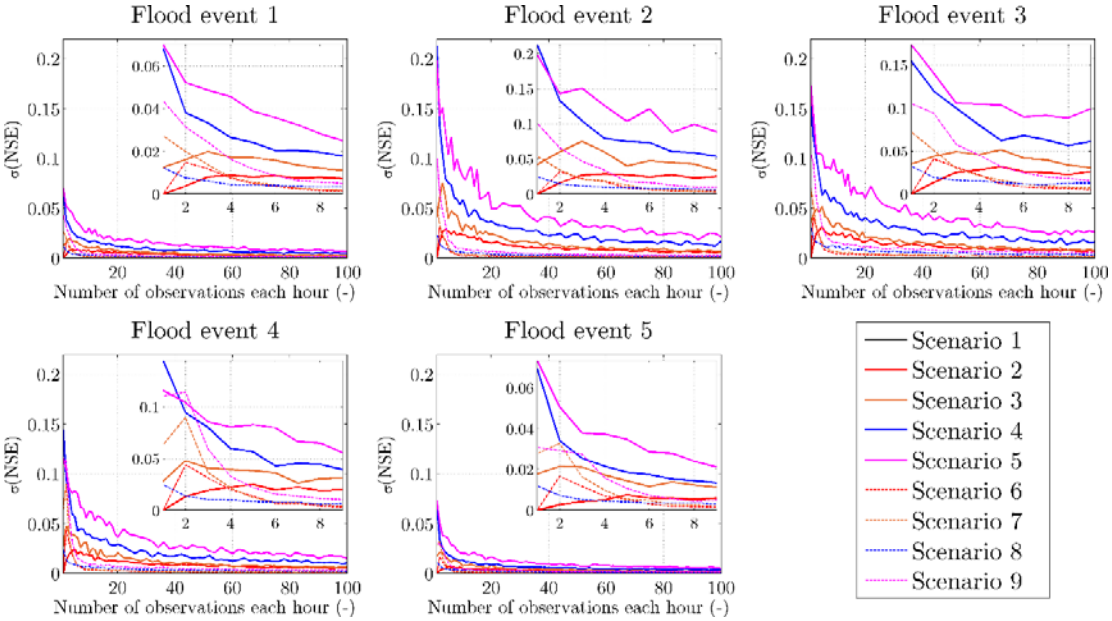

Figure 2. Dependency of σ(NSE) on the number of observations, for the scenarios 2, 3, 4, 5, 6, 7, 8 and 9 for the five considered flood events

These results are very similar between each other and we thought, as suggested by the reviewers, that adding other case studies will strengthen the manuscript. For this reason we decide to remove flood event 4 and 5 for the Brue catchment and add other 2 catchments with 6 flood events to the study. In addition, other 2 flood events are added to the Bacchiglione case study for a total of 4 case studies and 12 flood events. As it can be seen from results obtained in the Brue, Sieve and Alzette catchments, reported in new figures from 7 to 12, model behavior is very similar by changing the type of flood events. Overall, the effect of random arrival frequency of crowdsourced (CS) observations is lower than the one induces by random uncertainty in such observations. The only difference between the results of each flood events is the magnitude of the model improvement related to the model performances in case of no update. In fact, during flood events 1 in Brue or event 2 in Sieve model performances are already high (NSE about 0.82) with 1 observation. This is reflected in a rapid model improvement even for small number of CS observations as reported in Figure 9. However, in case of other flood events models does not perform as well as with the previous flood events and this brings to more CS observations needed in order to achieve good model improvement. For this reason, it is difficult to define a priori number of CS observations necessary to have a satisfactory model improvement due to the model performances without assimilation. However, in case of these case studies and during these nine flood events, it can be seen that an indicative value of 10 CS observations can be considered in average to achieve a good model improvement.

In case of the Bacchiglione catchment additional two flood events are considered. It can be seen that integrating social and physical sensors results in the best model improvement. In particular, for the two new flood events Setting D (2 social and 1 physical sensor) performs better than Setting E (3 social and 1 physical). As suggested by the reviewers and editor, we have separated results and discussions sections adding the description of the additional flood events. These and other

considerations have been added in the new version of the manuscript marked in blue. We believe manuscript has been significantly improved following the reviewers' suggestions.

Anonymous Referee #2:

Overall I find the paper which aims to develop and evaluate a data assimilation method for crowdsourced hydrological observations to be very interesting and timely given the growth of citizen science activities in environmental fields. I am reviewing the revised manuscript also in view of the open public discussion.

1. *The assimilation method and experimental development is well thought out. However, my main concern is that the evaluation is based on a very limited sample, which is something that has been highlighted by Reviewer 1 and not quite addressed in the revision. Hence I would suggest stating clearly up front (in the abstract) that the evaluations are conducted over a specific number of flood-events. This should balance the current title which is highly appealing but gives too much promise in terms of scientific insight.*

We appreciate reviewer's comment. Indeed, the previous sample of flood events was very limited. For this reason we added 2 more case studies and 9 more flood events in order to prove our previous results. Overall, as replied to Reviewer #1, results between different flood events are very consistent between each other. The only different is the magnitude of the model improvement which really depends on the initial model performances without update. Additional results and considerations can be found in the new version of the manuscript marked in blue. We believe that the quality of the manuscript has been improved and strengthened with these new results.

2. *Currently all the 11 scenarios in Experiment 1 are lumped together in one analysis. I suggest to make the results section more readable, the scenarios could be categorized better such that only one factor is studied at a given time (e.g. effect of arrival frequency is evaluated while accuracy is fixed/controlled). This could also be achieved through a better color-coding/line-typing within the figures.*

Following reviewer's suggestion we have separated results and discussions. In particular, in the results section we have described the results achieved in each single scenarios without lumping them. The same has been carried out for all case studies. In addition, different colors have been used in describing scenarios from 1 to 9 in the Brue, Sieve and Alzette catchments. In fact, in case of random arrival frequency and fixed accuracy we have used warm colors (red and orange for scenario 2 and 3) for plotting the results, while in case of random accuracy and fixed frequency we used cold color (blue for scenario 4). As stated in the manuscript, scenario 5 is a combination of random accuracy and random frequency. Therefore, a purple color was used in order to describe the combination of scenario 3 and 4 as combination of cold and warm colors (purple obtained as combination of blue and red). We have included these considerations in the "Results" section (lines 538, 578 and 584) of the updated version of the manuscript.

3. *The authors should explain why the value of 100 was chosen as the diagonal for the S matrix in the Kalman Filter. The basis for this value is unclear in the manuscript. The authors should also discuss if/in what way the results would be affected by this assumed (?) value. To my knowledge, the Kalman Gain would be very sensitive to the values used in the matrix. Also, please double check the notations used throughout this section - M_Q (line 353) is not used in the equations.*

Thanks for the comment. Indeed we realized this sentence was wrong and misleading. We apologize for the inconvenience. In fact, the proper characterization of the model covariance matrix **S** is a fundamental issue in Kalman filter. In this study, in order to evaluate the effect of assimilating crowdsourced observations, the model error is considered lower than the observations one. For this reason, a different value of the covariance matrix **S** is considered for each case study. In fact, a covariance matrix **S** with diagonal values of 1, 25 and 1 are considered for the Brue, Sieve and Alzette catchments. The bigger value of S in the Sieve catchment is due to the higher flow magnitude in such catchment if compared to the other two. A sensitivity analysis of model performances depending on the value of S is reported in the new version of the manuscript. The results show that the updated model tends to better represent flood events depending on the value of the matrix **S**. In fact, increasing the value of **S**, i.e. assuming a less accurate model, force the model towards the observations because are more accurate than the model itself. The results of the sensitivity analysis are related to the first flood event of the Brue, Sieve and Alzette catchments. Increasing the number of observations within the observation window results in the improvement of the NSE for different value of model error. However, such improvement becomes negligible for a given threshold value of streamflow observation, which is a function of the considered flood event. This means that the additional observations do not add information useful for improving the model performance. Overall, increasing the value of the model error **S** tends to increase NSE values as mentioned before. For this reason, in order to better evaluate the effect of assimilating crowdsourced observations, a small value of **S**, i.e. a model more accurate than observations, is assumed. In case of the Bacchiglione catchment, **S** is estimated, for each given flood event, as the variance between observed and simulated flow values.

In addition to that, we found a small error in the Kalman Filter code related to the definition of the matrix R. For this reason the simulations were run again and the value of σ(NSE) for scenario 4 are higher than the previous ones. These considerations have been included in the “Results” section (lines from 538 to 552) of the revised version of the manuscript.

4. *On the authors' response to Reviewer 1's comment regarding calibration period*

i) The experiments are assessed in terms of NSE. However, the calibration that authors describe for Brue catchment is based on optimising correlation. This is an inconsistency that should be explained. Furthermore, the time series length does seem to make a difference to the two parameters investigated (especially 'n' considering how it is used as the exponential term in eqn 1) as tabulated.

We thanks the reviewer for this valuable comment. In the first version of the manuscript model calibration was carried out using NSE while after the first review we ran additional experiments using the correlation R as metrics instead. As remarked by the reviewer, this is inconsistent with all the rest of the manuscript in which model performances are evaluated using NSE. However, based on Krause et al. (2005) and Price et al. (2012), these two indexes can be considered similar

because both very sensitive to peak flows, at the expense of better performance during low flow conditions. For this reason we showed results based on R. However, in order to avoid this inconsistency in the manuscript, we herein report the comparison between 4 different indexes used to calibrate the model parameter for the Alzette catchment. As it can be seen from the figure, all the matrices gave same optimal parameter values of n=1 and c=0.00064. In the new version of the manuscript, we added that the models of the three catchment were calibrated using both R and NSE, resulting in similar parameter values.

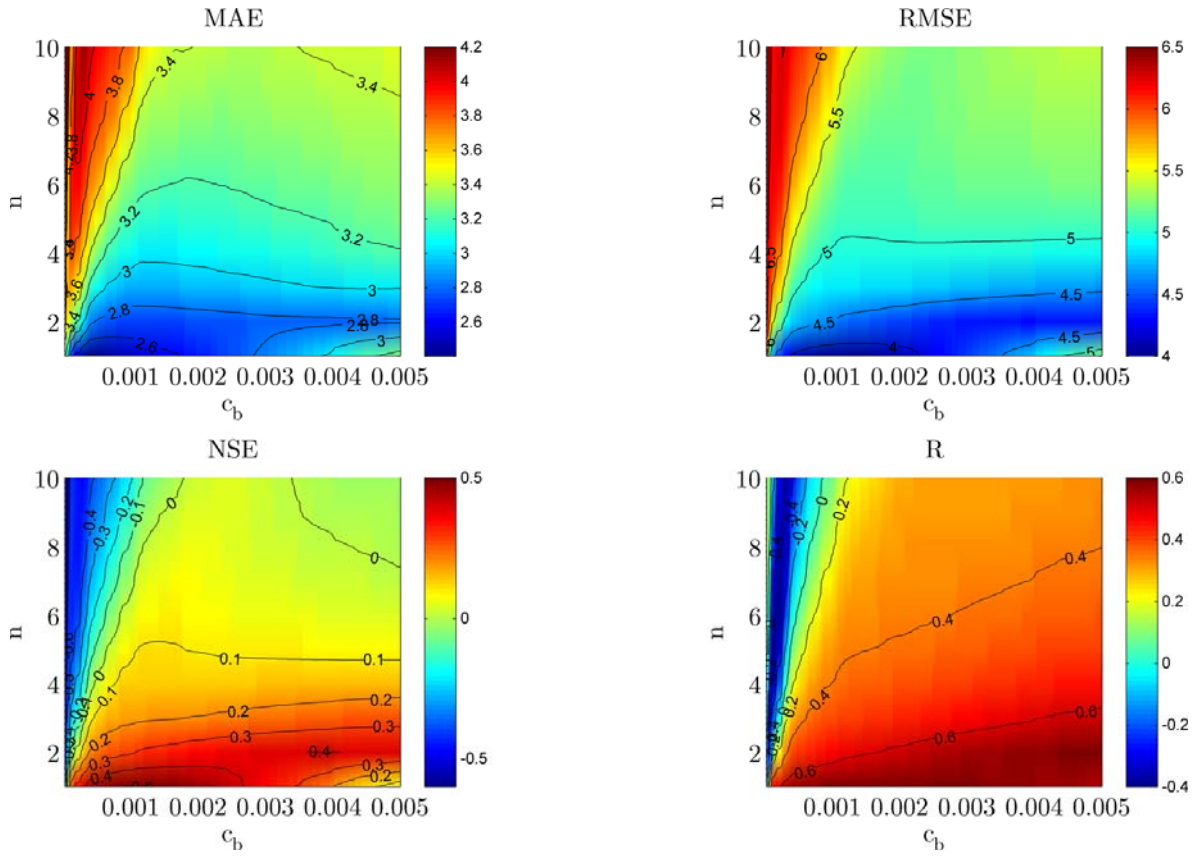

Figure 3. Model calibration using 4 different statistics, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Nash Sutcliffe Efficiency (NSE) and correlation coefficient (R) for the Alzette catchment

As the reviewer stated, in case of long time series the value of n changes. However, such value reduction of n, which tend to increase flood peak, is compensated with a reduction of c, which tend to reduce the flood peak.

ii) The reference to Buytaert et al 2014 should be accompanied by a brief extraction of their review/findings. The current statement 'detailed and interesting review' is vague and the authors' review of the other literature in the preceeding sentences could also be related to 'hydrology and water resources science'.

Thanks to the reviewer's comment we extended the description of Buytaert et al. 2014. In this review study, the potential of citizen science, based on robust, cheap, and low-maintenance sensing equipment, to complement more traditional ways of scientific data collection for hydrological sciences and water resources management is explored. In order to study the challenges and opportunities in the integration of hydrologically-oriented citizen science in water resources management, four case studies from remote mountain region (e.g. the Peruvian Andes) are considered. These text has been added in lines 73-78 of the new version of the manuscript.

Reference

- Krause, P., Boyle, D. P., and Base, F.: Comparison of different efficiency criteria for hydrological model assessment, Adv. Geosci., 5, 89-97, 2005
- Price, K., S. T. Purucker, S. R. Kraemer, and J. E. Babendreier (2012), Tradeoffs among watershed model calibration targets