RESPONSE TO REVIEWER #2

Paper: Towards assimilation of crowdsourced observations for different levels of citizen engagement: the flood event of 2013 in the Bacchiglione catchment

M. Mazzoleni, V.J. Cortes Arevalo, U. Wehn, L. Alfonso, D. Norbiato, M. Monego, M. Ferri and D.P. Solomatine

General comment

This paper on the potential use of citizen science data for flood forecasting is interesting to the readers of HESS but I have several major and some minor comments and concerns.

We appreciate the critical reviewer's comments and suggestions on our paper. We have addressed them all, in some cases by adding new experiments and in others re-structuring, clarifying and/or removing text and acknowledging the theoretical approach of our study. We therefore believe that the manuscript has improved substantially. Below, you can find a point-to-point discussions of all the comments.

Major comments

RC1: *The paper describes the results for multiple experiments, for different river stretches, lead times and stations but the multitude of results are never integrated or discussed. In fact, there is almost no discussion of the results at all. The lack of integration of the different results leaves the reader at loss about what the main take home message or contribution of this work is. This seriously harms the impact of this study and paper. Often the results for different stream sections or sub-catchments are described in detail and while these specific results (and the differences for the sections or subcatchments) may be of interest for people working in this basin, it is unclear why the results are different and what was learned from these differences that can be used outside this particular basin. Due to this lack of synthesis and discussion, the paper reads a lot more like a report of a modelling study for an agency (or thesis) than a paper for an international journal. Overall, much more integration and discussion of the results are needed and to clearly state what new thing was learned from this study. The paper has 17 figures, many of which contain multiple subplots and look similar. It is hard for the reader to pin-point what the main or most important "take home" figures are. Is there not a way to merge some of the figures or to summarize the results in a more clever way so that it is clear what figure (and thus what result) the reader should remember from this manuscript? The different figures don't integrate and compare the results from the different experiments and therefore it is hard to compare the model simulation results for the different data types and thus to appreciate the value of the different data types.*

AC 1: We thank the reviewer for the comprehensive comment, and indeed noting the deficiencies of way material is presented in the manuscript. In the complex process of assimilation of crowdsourced observations in water system models many factors play an important role in the correct flood estimation. Those are for example the type of a social sensor, citizen engagement, decay of the engagement over time, type of hydrological and hydraulic model, and quality control of CS observations. We agree with the reviewer and accept that results were presented without giving the reader a clear "take home" message. In this study we focused on the type of social sensors, on the citizen engagement level and its variability in time and space. To better integrate the results and to give a stronger message on the effect of CEL on flood prediction, we removed figures 6, 9, 10, 12, 16 and 17, which contains many multiple sub-plots.

We have included additional analyses on the effect of diminishing citizen engagement in time and variable spatial smartphone penetration on the model performances. In addition, we have reduced the number of scenarios in the experiment 3. In the revised manuscript version we will reduce the number of presented scenarios from 6 to 3 (see next figure). Instead, we will include new analyses on the effect of temporal and spatial variability of citizen engagement levels. Finally, we will divide results and discussions into two separate sections to clearly present the findings of this study.

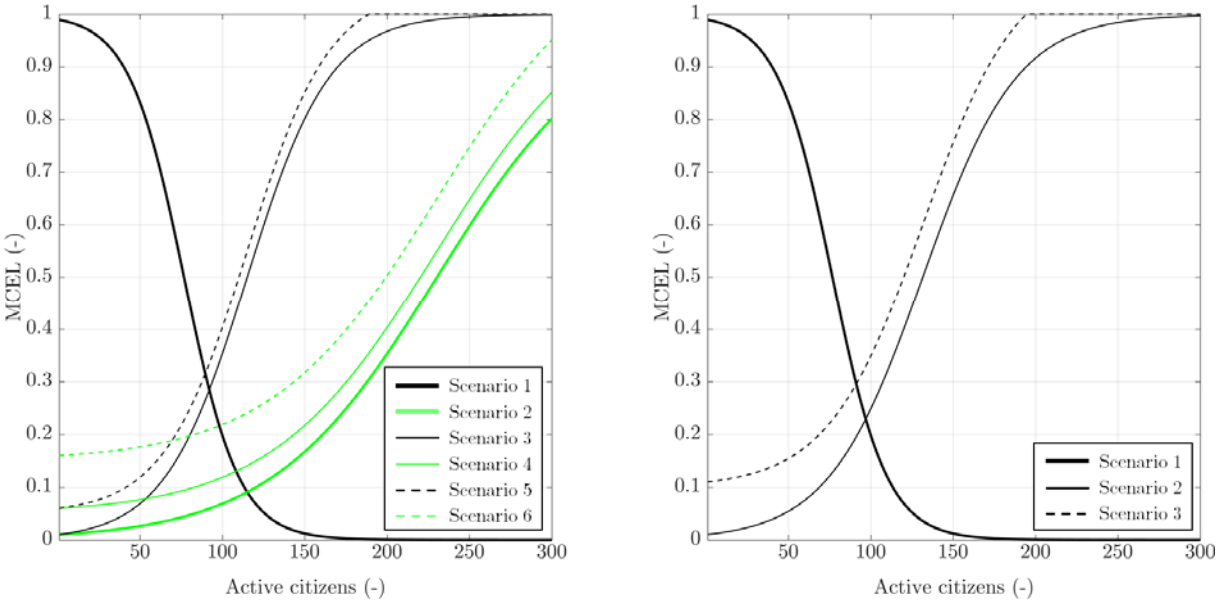

Figure 1. Original 6 scenarios (left side) and new 3 scenarios (right side) theoretical engagement curves

The main "take home" message we wanted to convey is: assimilation of CS observations provided by citizens improves model performance, and we can show how much, and how this improvement depends on the level of engagement. In particular, assimilation of CS observations in hydrological model tends to lead to a lower improvement than the assimilation in hydraulic models. Bias in water level observations plays an important role. Finally, the effect of spatial variability of active citizens and the decay of citizen engagement in time it is of vital importance to keep adequate model results. This study demonstrates that high model performance can still be achieved even for decreasing engagement in time. We will make these statements clearer in the revised version.

RC2: *P4L34: It is unclear how this paper is different from the four other papers by the authors on this topic. It would be good to specifically state here (or elsewhere) what is different between this paper and these previous papers and how this paper builds on the work of (and goes further than) these previous papers.*

AC 2: We thank the reviewer for this valuable comment. Indeed, we should have been clearer. In the revised version, we are clarifying this aspect in the discussion section. In the four previous studies we investigated the effects of assimilating real-time (synthetic) crowdsourced (CS) observations into *hydrological* models. However, in Mazzoleni et al. (2015; 2017a and 2017b), we have not investigated the effect of assimilating CS observations into hydraulic models. Furthermore, we have not considered neither (theoretical) scenarios of citizen engagement, nor the simultaneous assimilation of CS observations from static and dynamic social sensors. For this reason, the main objective of this study is to assess the usefulness of assimilating CS observations in both hydraulic and hydrological models-based predictions of flood events. To that end, we analyse the flood event occurred in May 2013 in the Bacchiglione basin with the data which would come from a distributed network of static physical (StPh),

static social (StSc) and dynamic social (DySc) sensors (however this data was simulated). These (synthetic) CS observations for water level are assimilated in a cascade of hydrological and hydraulic models. The experiments are analysed as CS observations and the assessment of citizen engagement levels are yet not operational nor available in the case study. CEL is further defined as the probability of receiving a CS observation based on the citizen's own interest or intention in collecting water levels. We assume that CEL mainly limit the intermittency or timely availability of observations. All mentioned above will be reflected in the revised manuscript and we hope will improve its clarity.

RC 3: *Methods: It is not fully clear what data that could come from citizen science observations was used. On P5L11, both water level and precipitation data are mentioned. From the methods (P5L25) it appears that only the water level data are used and the precipitation data are not used (except the precipitation from the standard measurement stations - not the simulated citizen science data) but then on P18L20-21, P29L4, 11 and P33L1 it is suggested that amateur weather observers will take more measurements. Why would amateur weather observers be particularly interested in water levels? Weather stations don't regularly measure water levels. Or was the weather data used as well? Also, it is not clear when the water level data was converted to streamflow and when it was just used as water level. On P15L21-25 it is suggested that for this experiment the water levels were not converted. Were they only used in the hydraulic model or also in the hydrologic model? Or did that depend on the situation/experiment? If so, then it should be explained much better when water level and when streamflow (converted from the water level observations) was used. If sometimes water level and other times streamflow was used, it should be discussed how this hinders comparison of the results for the different experiments.*

AC 3: We fully agree with this comment, and regret the confusion. On P5L11 we were referring to the mobile app developed within the WeSenseIt project, used to observe both water levels and precipitation. However, in this study we only used water level, since precipitation data are provided by static physical sensors within the catchment. We clarify this aspect in the revised version of the manuscript.

We referred to amateur weather observers to point at the active citizens which will (regularly) provide water level data driven by moral norms and the wish to create knowledge about the hydrological status of the river. We agree with reviewer that the term amateur weather observers might be confusing. For this reason, in the revised version we have modified the term "weather enthusiast individuals" into "enthusiast individuals" which do not use any weather stations for providing water level observations. No weather data was used from these hypothetical citizens, just water level.

In both hydrological and hydraulic models, water level observations are assimilated. Ideally, water level values can be directly assimilated into the hydraulic model. However, in this study we use a Muskingum-Cunge model which requires flow information rather than water level. For this reason, water level value at a certain random location is converted into flow value using Manning equation. Similarly, for the hydrological model the water level observations at the outlet of each sub-catchment has to be converted into streamflow values to be assimilated. This can be done by means of the rating curves available at the sub-catchment outlet cross-sections. As suggested by the reviewer, in the revised manuscript we are providing more details and explanation on the assimilation of water level and streamflow within the hydraulic and hydrological models.

RC 4: *P11L27: In this study, the modelled streamflow was used to obtain the water level and streamflow data. However, when real citizen science water level data are used, a rating curve is needed for every potential measurement location to obtain information on the streamflow. How would you do that? This is crucial information that is needed when this approach would be used with actual citizen science data (rather than this hypothetical or virtual study). However, almost no guidance is given on how this rating*

curve information would be obtained for the real citizen science case or how the huge uncertainty in any assumed rating curve will affect the model results. This really needs to be addressed to make the proposed approach useful for real cases with citizen science data. On P15L19 it is suggested that cross sections can be derived from natural cross sections elsewhere but cross sections vary hugely. So this will significantly impact the results.

AC 4: As the reviewer correctly mentioned, it is quite unlike to have the information of the rating curve at a random location of the CS observation provided by dynamic sensors in real world applications. In this case, a properly calibrated Manning equation can be used along the river to convert water level into streamflow. Roughness parameter in the Manning equation is calibrated by comparing the observed and simulated rating curve at the outlet section of the catchment. Such a calibrated value is optimal for the cross section of Vicenza but may not be optimal for other upstream sections. In addition, Manning equation uses the cross-section information to estimate hydraulic variables like wetted area and perimeter. When there are no data regarding the cross-section, assumptions should be made about a rectangular cross-section with a given width and depth. Obviously, in case of a more complex hydraulic model, the estimation of streamflow from water level is not required since it can directly assimilate water level.

On the other hand, in case of static sensors the water level can be converted into streamflow using a rating curve. During the installation of the sensor/staff gauge it would be possible to derive the rating curve and cross section in order to convert water level into discharge. However, both Manning equation and rating curve introduce a significant degree of uncertainty in the streamflow estimation. For this reason, CS observations from social sensors are assumed to have lower (and variable) accuracies. These considerations have been included in the conclusion section of the updated manuscript.

RC 5: *Table 2: How were these values chosen? What are they based on? No references or information is given and therefore it cannot be determined if these values are reasonable at all!! Give references to back up these values or describe how they were chosen and why they are considered reasonable.*

AC 5: In this study we assumed that observations from DySc sensors are randomly biased adding a white noise with standard deviation proportional to a coefficient γ. This coefficient has the same absolute value of the error value of α in case of StSc sensors. No reference to the choice of γ was provided in the manuscript since it was subjectively assumed. Obviously, we do not want to conclude that 0.3 should be considered as default value to estimate bias in real-life crowdsourced observations. Such bias has to be defined based on field experiments with volunteers proving water level observations during real flood conditions. The main point of this analysis was to provide a sensitivity analysis of model results based on a subjective value of γ.

This study demonstrated that the effect of biased observations on flood prediction strongly depends on the model performance without any assimilation. In the flood event we considered, model without update tends to underestimate the observed water level and streamflow. For this reason, assimilation of overestimated water level data provides higher model performances (N_{SE}). On the other hand, underestimation of water level data will give lower NSE. These results can be seen in figure 7 of the paper. These considerations have been included in the conclusion section of the updated manuscript.

RC 6: *Table 3: What are these alpha values based on? A reference should be given or the choice of these values should be discussed in detail! P13L7: do these values really suggest that if water levels can be measured from a staff gauge at 1 cm increments that citizen scientists can estimate the distance between the bank and the water level without a staff gauge with a 2-5 cm accuracy? This latter value*

seems not reasonable to me at all (since already the surface level of the bank probably differs by a few cm).

AC 6: We agree with the reviewer that the assumption of a citizen scientists estimating the distance between the bank and the water level without a staff gauge with a low error is unrealistic. A more appropriate method for measuring flow at a random location using a dynamic sensor can be the one proposed by Beat et al. (2014). The authors proposed an approach to measure water level with good accuracy, surface velocity and runoff in open channels. However, this approach requires a-priori knowledge on the channel geometry at the random location of the measurement, which is one of the main sources of uncertainty. For this reason, it is assumed that DySc sensors have lower accuracy than StSc. We modified the section 3.1 in the manuscript accordingly.

One of the main and obvious issues in citizen-based observations is to maintain the quality control of the water observations. In this study, the coefficient α is assumed to be a random variable, uniformly distributed between 0.1 and 0.5 based on the type of the social sensors. The high values of α for the StSc and DySc sensors are due to the different sources of uncertainty introduced in the water level estimation and the consequent conversion to discharge, since both hydrological and hydraulic models assimilate flow data.

In case of StSc sensors, water level can easily be measured by citizens using a staff gauge as reference. The main source of uncertainty is introduced in the streamflow estimation from water level by means of the Manning equation or available rating curve. The value of α equal to 0.3 used for StSc is based on our previous studies. In case of DySc sensors, besides the uncertainty in the flow estimation, the assessment of the water level is affected by the uncertainty in the proper knowledge of the cross-section geometry at the random location. For this reason, an error value of 0.5, almost double than for case of StSc, is assumed for DySc sensors.

Unfortunately, we did not have any real crowdsourced observation to test validity of these coefficients. In this study, no sensitivity analysis was performed on the maximum value of α using dynamic sensors. However, for the revised manuscript, we have already run an additional simulation in which the maximum value of α is set equal to 0.8 (during Experiment 3) in order to assess the change in model performance due to data assimilation. In this case, the coefficient K on the logistic curve is set equal to 1 (see the μ(NSE) values reported in the figure below).

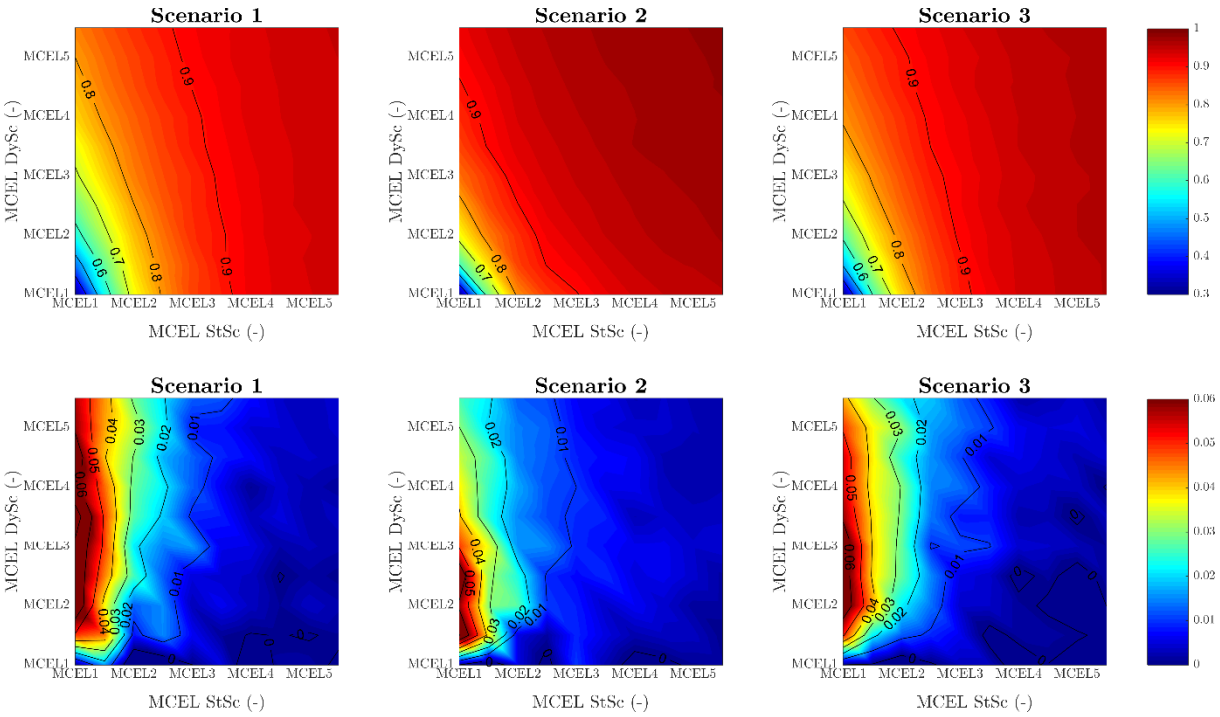

Figure 2. First row: μ(NSE) values in case of high observation error (α_{max}=0.8) in the DySc sensors; Second row: difference between μ(NSE) with α_{max}=0.5 and α_{max}=0.8 for different engagement levels from StSc and DySc (experiment 3)

From the figure you can see higher gradient in the contour lines if compared to the results obtained with α_{max} of 0.5, meaning a higher dependence of μ(NSE) towards StSc sensors since the uncertainty in the DySc sensors is increasing. In addition, Figure 2 shows small differences between μ(NSE) with α_{max}=0.5 and α_{max}=0.8. Due to the already high number of figures and results we have not included this analysis in the updated version of the manuscript. We leave the editor to the decision on whether add or not these results in updated version of the manuscript.

Reference:

Beat L., Philippe, T., and Peña-Haro, S.: Mobile device app for small open-channel flow measurement, (June 18, 2014), International Congress on Environmental Modelling and Software, Paper 36, 2014.

RC 7: *P14L27: Does this indeed mean that for any given time step there is a 40% chance of getting 1 measurement? Even at night? That does not seem realistic. In the figure CEL values of >80% are used. This is certainly not likely. It would be better to at least also zoom in to the much lower and more realistic CEL values. On P25L20 it is mentioned that the results are highly sensitive to the CEL values. This makes it even more important to show only (or mainly) the results for reasonable CEL values!*

AC 1: Indeed, we agree with this comment. As the reviewer mentioned, we did not distinguished between observations provided during day time or night time (as addressed in Mazzoleni et al. 2015). This is one of the limitations of this study and it is mentioned in the conclusion section of the revised manuscript.

The reviewer also suggested to focus more on the lower part of the theoretical engagement curve, assuming more realistic CEL value than the ones assumed in our study (CEL>80%). For this reason, we have carried out an additional simulation where the maximum carrying capacity of the logistic curve (K) is considered variable from 0.01 up to 1.

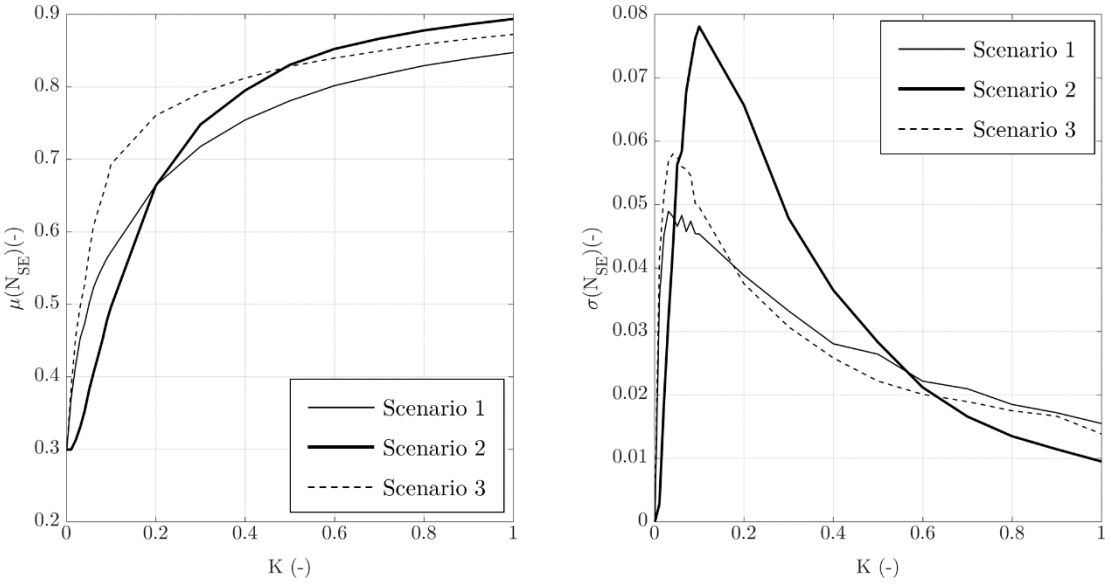

Figure 3. Difference between $\mu(N_{SE})$ and $\sigma(N_{SE})$ values obtained considering varying values of K for different engagement levels from StSc and DySc during experiment 3

Finally, in the next figure the $\mu(N_{SE})$ obtained with varying values of K are summarized. For a given scenario and value of K, the single value of $\mu(N_{SE})$ is estimated as the mean average of the different $\mu(N_{SE})$ values corresponding to the MCEL for StSc and DySc.

The results of this analysis showed an expected reduction in the model performances for low values of parameter K (which indicates the maximum possible level of engagement). It can be noted that if K is equal to 0.5, although the engagement values are halved, assimilation of crowdsourced observations still provide significant model improvement for all the different scenarios. As expected, $\sigma(N_{SE})$ values tend to increase for low engagement of citizens. From **Error! Reference source not found.**, it can been that $\mu(N_{SE})$ values do not follow a linear trend as somehow expected. On the other hand, tends to drop for values of K between 0 and 0.2 (for example in scenario 3), while for higher K values the $\mu(N_{SE})$ do not grow significantly. In particular, for K values higher than 0.5, scenario 2 provides the highest $\mu(N_{SE})$ values. On the other hand, for lower K values than 0.5 scenario 3 is the one leading to better model performances. This is because the presence of enthusiast individuals keeps high engagement values even for low values of K. Regarding the variability of NSE, i.e. $\sigma(N_{SE})$, for values of K lower than 0.4, high $\sigma(N_{SE})$ can be observed in scenario 1. These considerations and other details are reported in the updated version of the manuscript. Figure are included in the revised paper.

RC 8: *P18L28-29: This is unclear and not logical. In the case that actual citizen science data are used, you don't know which measurement is most accurate and so you can't use this criteria. You would most likely use both measurements. Why was that not done here?*

AC 1: Indeed, this issue needs explanation. As mentioned in the conclusions of this study, one of the main problems in citizen science data is the proper definition of the observation error which changes according to the citizen and sensor type. However, in data assimilation methods it is necessary to define both model and observation errors in order to optimally update model states. Based on the assumptions of observation errors for multiple observations, we decided to consider the one with the lowest error. Nonetheless, the approach proposed by the reviewer of using all measurements instead of only the most accurate one is also valid. In that case, each observation will be used in the assimilation scheme based on certain assumptions of observational errors. Less weight is given to the more uncertain observations

while more weight is given to the more reliable. We are including this consideration in the updated version of the manuscript.

RC 9: *The results (e.g fig 5-6) show that the NSE values are low when the lead time is more than one hour. I miss a discussion on how useful these model simulations are for operational flood management. Is a model prediction with an NSE of 0.4 still useful? It seems unlikely to me that roads can be closed and people evacuated with a lead time that is much less than an hour. Currently there is no discussion about this at all – this really should be included. Also why was NSE used as a criteria and not peak water level or peak flow as well, as in the end that is what is most important in flood management.*

AC 9: We agree with the reviewer that for more than 1 hour lead time the results coincide with the ones achieved in open-loop condition (i.e. without data assimilation) using forecasted precipitation as input - which is the current practice for flood forecasting in the catchment used by Alto Adriatico Water Authority. Any improvement of model performance with respect to this situation provide additional useful information for flood risk management. However, this is valid only to the case in which observations are assimilated within hydraulic model at cross sections close to Vicenza. For operational flood management it is advisable to consider model results in which observations at upstream location of the catchment are assimilated in both hydrological and hydraulic model. This can be observed for example in figure 8 of the manuscript, where for high lead time values (4 hours) assimilation of observations in the hydrological model allows for a better model prediction.

In this study, N_{SE} is used as the only performance measure without considering improvement in the prediction in the peak and rising limb of the hydrograph, which are extremely important in case of operational flood management. Based on reviewer's comment, we have included additional performance measures, i.e. the relative error between observed water level peak and simulated peak during Experiment 3 (theoretical engagement level scenarios), to better assess the assimilation of crowdsourced observations from an operational point of view.

$$E_{RR} = \frac{\left(\mathrm{WL}_P^O - \mathrm{WL}_P^S\right)}{\mathrm{WL}_P^O}$$

Where WL_P^O and WL_P^S are the observed and simulated peak water levels correspondingly.

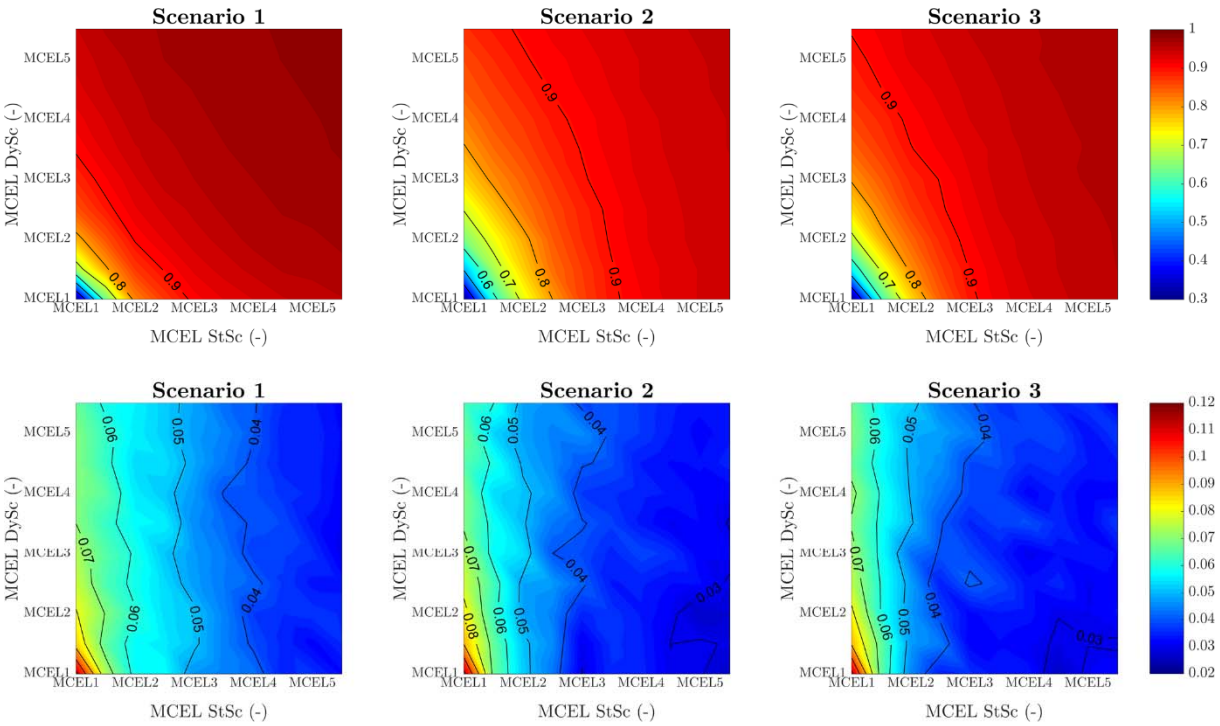

Figure 4. $\mu(N_{SE})$ (first row) and $\mu(E_{RR})$ (second row) values obtained considering different engagement levels from StSc and DySc during experiment 2

Analysis using E_{RR} leads to conclusions similar to those drawn when using N_{SE}. However, smaller error values are obtained in scenarios 3 rather than scenario 1. In addition, it can be observed that E_{RR} values are more sensitive to engagement levels from StSc sensors than from the DySc ones (vertical gradient). More details and discussions are provided in the revised version of the manuscript. Figure 4 is added to the revised paper.

RC 10: *Overall, the paper is not particularly well written. For many sentences, a more direct or less complicated sentence structure could be used. This would make the paper much easier to read. Some of the information is given twice (e.g. P3l25-28 = p4L14-15), other information is not really necessary (e.g. P2L25-28). Elsewhere lists with other studies are given without any information about them, thus also not the important aspects that are relevant for this study (e.g. P3L28-32). In other places, there are sentences that may be remnants from moving text around or previous versions that don't fit with the content of the remainder of the paragraph at all (e.g. P4L15- 19). I suggest that the authors critically read through the manuscript, include missing information but also remove sentences that do not fit (i.e. break the flow) or don't provide any information that is pertinent to this study.*

AC 10: Indeed we should have been more careful in writing and structuring. We appreciate reviewer's comment and suggestions. Based on them, we have re-phrased complicated sentences, removed double text and polished the manuscript. We believe the readability of the paper is improved.

Other specific comments

RC11: *Title: The title doesn't really tell what the paper is about or what the main findings are. I suggest that you consider changing the title to make it much clearer that this is a hypothetical study that assumes that crowd-sourced data is available (using model results as observations)"*

AC11: Thank you for the comment. We have considered your suggestion and adjusted the title which will now read "Exploring the influence of citizen engagement on assimilation of crowdsourced observations: a model study based on the flood event of 2013 in the Bacchiglione catchment".

RC12: *P2L13-15: Add references for each of these attempts.*

AC12: We have included the reference in the text below L13-15 of the revised version of the manuscript. In particular, (1) Data assimilation techniques (see a detailed review provided by Liu et al. 2012); (2) assimilation of multiple physical sensors; and more recently (Aubert et al., 2003; McCabe et al., 2008; Pan et al., 2008; Lee et al., 2011; Montzka et al., 2012; Pipunic et al., 2013; Andreadis et al., 2015; Lopez Lopez et al., 2015; Rasmussen et al., 2015); 3) assimilation of crowdsourced (CS) observations from static social and dynamic social sensors (Shanley et al., 2013; Buytaert et al., 2014; Lahoz and Schneider, 2014; Fava et al. 2014; Smith et al. 2015; Fohringer et al., 2015; Gaitan et al., 2016; Giuliani et al., 2016; de Vos et al., 2017; Rosser et al., 2017; Starkey et al., 2017; Yu et al., 2017).

RC13: *P2L21-29: Remove this text. This may be useful in a report but is not really necessary in a scientific publication.*

AC13: We have removed the text as suggested

RC14: *P3L7: Are 'heat flux sensor' data really that widely available and are they really that useful for flood prediction?*

AC14: We removed heat flux sensor, as not really useful for flood prediction

RC15: *P3L8: Add references!*

AC15: We have include Liu et al. (2012) as main reference

Liu, Y., Weerts, A. H., Clark, M., Hendricks Franssen, H.-J., Kumar, S., Moradkhani, H., Seo, D.-J., Schwanenberg, D., Smith, P., van Dijk, A. I. J. M., van Velzen, N., He, M., Lee, H., Noh, S. J., Rakovec, O., and Restrepo, P.: Advancing data assimilation in operational hydrologic forecasting: progresses, challenges, and emerging opportunities, Hydrol. Earth Syst. Sci., 16, 3863-3887, doi:10.5194/hess-16-3863-2012, 2012.

RC16: *P3L25-29: I don't think that it is necessary to include this information. The paper is already very long.*

AC16: We have included this information in the second section where a background of citizen engagement theories and methods are briefly described. We believe this section will help the reader to understand the main motivations that drive citizen for sharing data and be engaged within the observatories of water.

RC17: *P3L29-32: Either take this list of references out or tell what these studies have looked at and how this is relevant for this study.*

AC17: Based on the length of the manuscript, we have removed this text.

RC18: *P5L8: I thought that this was done by the civil protection. Make it clear that this is not an "average citizen". * P5L12: Isn't the system already operational?*

AC18: One of the goals of the WeSenseIt project is to involve any kind of citizens in collecting and sharing hydrological measurement for improving model performances and prediction. That is why in P5L8 we referred to "citizens", or more in particular "active citizens" if they are already involved in the observatory of water. However, the different case studies of the project have different types of involved (active) stakeholders. In the Bacchiglione catchment, for example, both Civil Protection and active citizens are involved, albeit with different level of engagement. Currently, the Alto Adriatico Water Authority is testing the possibility to make the assimilation of crowdsourced observations in the Bacchiglione catchment operational.

RC19: *P5L17: What are typical response times (and/or travel times of the flood wave) for this catchment? Without any information on this, it is not possible to interpret the results for the different lead times.*

AC19: The following table shows typical response times for the sub-catchment and the reaches:

Location	Time (hours)
Sub-catchment A	1.5
Sub-catchment B	3.5
Sub-catchment C	6.0
Reach 1	2.2
Reach 2	2.0
Reach 3	7.2
Reach 4	9.5
Reach 5	3.4
Reach 6	5.2

We are considering including this information in the revised manuscript.

RC20: *P5L24: Were these traffic disruptions indeed due to flooded streets (or due to landslides, etc)?*

AC20: Yes, they were mainly due to flooded streets and river closed to overtopping the levee system. We will mention this in the revised manuscript.

RC21: *P6L17: I would not use the word 'sensor' in this context. The text will be much clearer when 'observation' is used as no sensors are used in the DySc. This is particularly the case for wording such as on P26L9, 10, 13, where the number of observations is mentioned and not a particular sensor.*

AC21: Thank you for the comment. We used the term "sensor" to define a device that responds to a physical input (e.g. heat, light, sound, pressure, etc.) and transmits a resulting output representing the status of a physical system. That is why, a citizen using a mobile app to measure flow characteristic can be considered to be sensors as well (called dynamic social sensors). However, an "observation" is not considered to be dynamic since it represents the physical system at a particular location and time, which is a consequence of the dynamic behaviour of the sensor in time and space. For this reason, we would like to keep the term "sensor" when referring to a dynamic device used by a moving/static citizen to take a measurement of the system.

RC22: *P13L2: A reference is needed here. I don't think that technicians or hydrologists are necessarily better at estimating depths, volumes or flows than other people. In my experience when multiple hydrologists estimate the depth or flow in a river, their estimates still vary widely.*

AC22: Thank you for the comment. The uncertainty of a neophyte or interested volunteer can be larger than that of an experienced volunteer or expert technician under the assumption that the experience of a large number of observations, enough training or sufficient background may increase the quality of observations. We agreed with the reviewer that both volunteers and experts are subjected to biases in either qualitative or crowdsourced observations. In a data collection exercise carried out with only technicians, Cortes Arevalo et al. (2016) highlighted the variability of expert collected observations. CS observations can be subjected to (random) bias regardless of their expertise level and pointed out that quality control procedures should done on a regular basis and upon submission regardless of the expertise level. Based on experience of online mapping, Kerle & Hoffman, (2013) pointed out that both volunteers and technicians have difficulties to provide observations, as all possibilities cannot and should probably be covered into the training procedures. Kerle & Hoffman, (2013) suggest that corrective feedback about the consistency of collected observations can improve the quality of results. That is for example against CS observations carried out by the same observer or suggested reference for the observations. In the discussion of results we have made some suggestions. Further research is needed on how corrective feedback can best be provided via mobile and web-based easy-to-use sensors and low-cost monitoring technologies.

RC23: *Figure 4: Make it clearer in the caption that these are hypothetical curves and not based on previous studies. If not, please include the reference.*

AC23: Thank you for the comment. We have modified the caption and the description of the Experiment 2 section to underline the theoretical validity of our approach.

RC24: *Figures 4-5: Use different line styles so that the figure is also clear when printed in black and white.*

AC24: Based on reviewer's comment we have modified line styles and colour to make the figure visible also in black and white (see figure 4 above).

RC25: *Figure 5a: For which lead time is this result? This is not clear from the caption.*

AC25: Figure 5a is referred to lead time 1 hour. However, we have removed this figure as explained in one of the previous replies.

RC26: *P21-L1-7: This should be part of the methods (not the results).*

AC26: We thank the reviewer of his/her comment. Based on that, we moved the description in P21-L1-7 in the method section Experiment 2.

RC27: *P21: Only the mean simulation results are shown and discussed. The variability in the results should at least be mentioned (or shown with an error band in fig 6).*

AC27: As previously discussed, we removed figure 6 to give more space to additional analyses on the effect of citizen engagement.

RC28: *P23L1-2: Why? This is an interesting result but not discussed. Just saying that results for A are better than for another catchment may be interesting for people working in this basin but not for the readers of HESS. For them it is much more interesting why these results are so different or what can be learned from these differences. Similar on L21-24 (and many other locations throughout the results) what is interesting about this result for people outside this basin/what can be learned from this?*

AC28: We completely agree with the reviewer. In the updated version of the manuscript we have divided the section "Results and discussion" in "Results" and "Discussions" respectively. In the discussion section, a more general description of the findings of this research (for example of figure 7) is provided.

RC29: *P25L1-4: So here only the water level and not the derived streamflow data are used? But doesn't that make that the comparison between the static and dynamic sensor network results more difficult? This is unclear and needs to be discussed in more detail! * P28L13-15: Don't overstretch your results. This study shows the model results for different chosen engagement levels but does not provide any information about the actual motivation. * P29L10: Why is no bias assumed? Isn't it likely that when people estimate the distance between the water level and the stream bank, there is a bias in the resulting water depth information? * P32L16-18: Add why this was the case.*

AC29: Synthetic water level values are used to derive streamflow values in both static and dynamic sensors. We have included more details in the experimental setup description. As stated by Gharesifard and Wehn (2016), and now to be included in section 2 of the revised manuscript, we acknowledge that stronger motivations or intentions are not only driven by a combination of more positive and favourable attitudes. The motivations also rely on stronger positive social pressure and greater perceived control or self-sufficiency about the means to provide CS observations. Gharesifard and Wehn (2016) further recognized that such rational choices may not apply in case of emergency situations. A simplified model has been formulated under the consideration that: i) only volunteers and/or trained volunteers will participate in providing water level observations. ii) the mobile application available for the project is easy-to-use and accessible for all participants. As mentioned in the paper (experimental setup), we assumed that engagement level is driven by citizen's own interest, which may be i) own personal purposes, ii) shared or community interests and iii) societal benefits. Each one of this different citizen's own interests corresponds to a curve in out theoretical (and simplified) approach. That is why, we stated that sharing CS observations driven by feeling of belonging to a community of friends (scenario 2) can help improve flood prediction. Obviously, this is a theoretical result that should be validated with real CS observation and a proper social analysis to better define the engagement curves based on particular motivation of the citizens.

We agree with the reviewer that bias in the CS observations should be included in the simulations. For this reason, the simulations in Experiment 3 are now performed considering CS observations in case of Bias 2 instead than Bias 1. In addition, we have performed an additional analysis considering negative and positive bias (Bias 3 and 4 in table 2) in the crowdsourced observations assimilated in the experiment 3. The difference between $\mu(N_{SE})$ values obtained using observations with Bias 2, Bias 3 and Bias 4 are displayed in the next figures. As expected, it can be observed that Bias 4 provides higher N_{SE} values than Bias 2 since model without update underestimate observed streamflow/water level. Moreover, results obtained using observations with Bias 3 have lower N_{SE} than results with Bias 2. However, in both Bias3 and 4, such changes in N_{SE} are very small, leading to the conclusion that assimilation of biased (observations) water level observations during the May 2013 flood event in the Bacchiglione

River do not significantly improve or reduce model performances. We have included these analyses and figure 5 in the updated version of the manuscript.

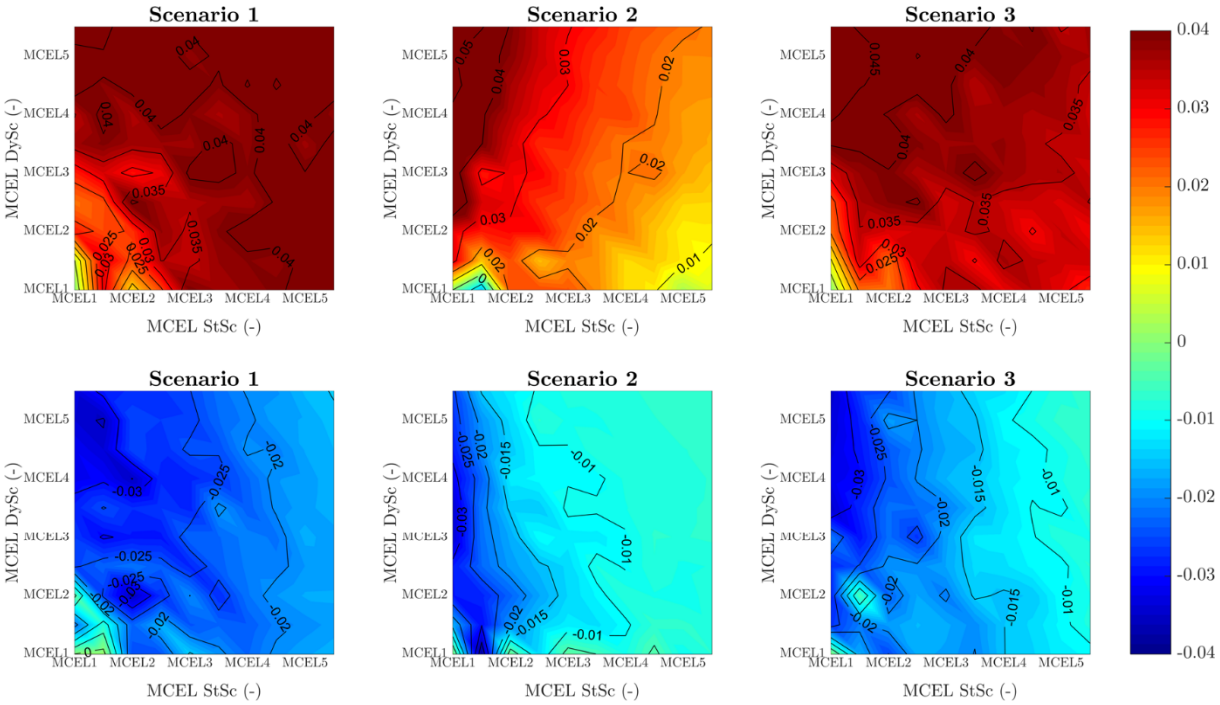

Figure 5. *Difference between $\mu(N_{SE})$ values obtained considering Bias 2 with Bias 3 (first row) and Bias 2 with Bias 4 (second row) different engagement levels from StSc and DySc during experiment 3*

RC30: *Conclusion: This is not a conclusion of the results or a summary of the main take home messages but rather a list of things that were done. That is much less useful than an actual conclusion.*

AC30: We thank the reviewer for her/his valuable comment. In the revised manuscript we have improved the conclusion providing a short summary of the main findings, a critical analysis of the novelty and main take home messages, limitations and recommendations for future studies.

RC31: *P33L14-16: Yes this is true but not a part of this study so don't include it in the conclusion.*

AC31: We have removed this sentence

Minor editorial suggestions:

- *P1L18: remove 'for model performance' and insert 'for improving model performance' at the end of the sentence.*
- *P1L19: insert 'of inclusion of social sensor data'*
- *P1L29-30: try to rewrite this sentence to make is clearer and easier to understand.*
- *P2L2: do you mean 'maximum' engagement instead of 'minimum engagement'?*
- *P2L13: remove 'over'*
- *P2L17-18: replace 'the benefits' by 'how citizen science data could have benefitted' to make it much clearer that this is a hypothetical situation and actual citizen science data were not available for this event.*

- *P4L2: Rather than 'minimize low' you could say 'maximize accuracies'*
- *P4L3-14: This part is about engagement and would fit much better at P5L4 (but this requires a sentence to link it to the previous sentence)*
- *P4L14-15: Double and not necessary – take out*
- *P4L16-18: Move to P4L2 where it fits much better.*
- *P5L29 (and elsewhere): replace 'arrival time' by 'measurement interval'*
- *Table 1: replace 'lecture' by 'reading'*
- *P21L4: "random uniform' – this is confusing is it random and variable or uniform?*
- *P21L15: The caption needs to be improved because it doesn't explain the figure (the figure is not clear for someone who only reads the caption).*
- *P32L28: Rewrite this sentence- it is unclear*

We thank the reviewer for all these valuable suggestions and comments. We have addressed all of them and include in the updated version of the manuscript.