Anonymous Referee #2:

Thank you for the opportunity to review the article "Can assimilation of crowdsourced streamflow observations in hydrological modelling improve flood prediction?" (hess- 2015-415). This article presents an evaluation of methods for improving the accuracy of hydrologic models by incorporating crowdsource (social sensors) data. This is an interesting idea and the first paper on the topic that I have read. The opportunity to get the public to engage in extreme-events using technology they are already familiar with is exciting and will likely be a great success. I think the paper is generally written well and accurately presents the methods and results and that the discussion and conclusions are reasonable. That said, I have included a few comments/suggestions/questions for the authors to consider. I have not provided an editorial review, though I do believe the paper should have a thorough editorial review prior acceptance. There are several instances with subject / verb agreement, some words are unnecessarily plural, and acronyms that do not appear to have definition (DA for example). Additionally, figures need to be checked to make sure they include relevant information included in the text (for example, include "setting A" on figure 15 or describing (a) and (b) on figure 13).

***RC:** Are there any methods currently in use to quantify the accuracy of crowdsource (CS) data? This is particularly important given that the methods you for including crowdsourced data are workable. I think you mention briefly about assessing accuracy of actual social sensors. Please expand on this in terms of current ideas, particularly ideas that would assess accuracy in an objective manner*

AC: We thank the reviewer for this valuable comment. Following his suggestion, we included the following additional information about methods used to assess quality of CS data, in the introduction:

> "According to Bordogna et al. (2014) and Tulloch and Szabo (2012), quality control mechanisms should consider contextual conditions to deduce indicators about reliability (expertise level), credibility (volunteer group) and performance of volunteers such as accuracy, completeness and precision level. Bird et al. (2014) addressed the issue of data quality in conservation ecology by means of new statistical tools to assess random error and bias in such observations. Cortes et al. (2014) evaluated data quality by distinguishing the in-situ data collected between a volunteer and a technician and comparing the most frequent value reported at a given location. They also gave some range of precision according to the rating scales. With in-situ exercises, it might be possible to have an indication of the reliability of data collected (expertise level). However, this indication does not necessarily lead to a conclusion of high, medium or low accuracy every time a streamflow observation

of a contributor is received. In addition, such approach is not enough at operational level to define accuracy in data quality. In fact, every time a crowdsourced observation is received in real-time, the reliability and accuracy of observations should be identified. To do so, one possible approach could be to filter out the measurements following a geographic approach which defines semantic rules governing what can occur at a given location (e.g. Vandecasteele and Devillers, 2013). Another approach could be to compare measurements collected within a pre-defined time-window in order to calculate the most frequent value, the mean and the standard deviation."

RC: *Please consider restructuring the Introduction. While the Introduction is very informative, it is quite long and digresses into a discussion of sensor technologies, issues of quality control, other CS networks, oceanographic models, and assimilation of asynchronous observations among other things. The paper is supposed to be about assimilation of CS data assimilation. The Introduction should go directly to this point. As written, the introduction of the topic and explanation of the objectives are separated by a considerable amount of material. Please shorten the Introduction to clearly present the topic, current understanding of how to include CS data, gaps in that understanding, and what you propose to do to fill that knowledge gap. The other information should be retained, but put into a different sections ("Background", "Existing CS Networks"). I personally find the material on existing CS networks very interesting and would like to see that information discussed a bit more.*

AC: Following reviewer's suggestion we shortened the introduction and focused on assimilation of CS observation, providing also some details about the past and ongoing projects in which CS are used to improve models predictions. In particular, we firstly defined the necessity of improving the model introducing the concept of model updating and DA. Secondly, we described the need of CS observations and some CS projects are illustrated. We focused on two main characteristics of the crowdsourced observations: a) data quality and b) variable life span (asynchronous observations). Thirdly, data quality issues and method used to deal with this problem are described (following previous reviewer's comment). Finally, we described existing methods used to assimilated asynchronous observations in hydrology and other water related models. The text related to the assimilation of distributed hydrological observations has been removed. We believe that in the present form the introduction is more readable and objectives of this paper are clearer. We also provided additional details about methods to assess observational uncertainty as proposed by reviewer in a previous comment. The new version of the introduction included in the revised manuscript is reported below:

"Observations of hydrological variables measured by physical sensors have been increasingly integrated into mathematical models by means of model updating methods. The use of these techniques allows for the reduction of intrinsic model uncertainty and improves the flood forecasting accuracy (Todini et al., 2005). The main idea behind model updating techniques is to either update model input, states, parameters or outputs as new

observations become available (Refsgaard, 1997; WMO, 1992). Input update is the classical method used in operational forecasting as uncertainties of the input data can be considered as the main source of uncertainty (Bergström, 1991; Canizares et al., 1998; Todini et al., 2005). Regarding the state updating, Kalman filtering approaches such as Kalman filter (Kalman, 1960), extended Kalman filter (Aubert et al., 2003; Kalman, 1960; Madsen and Cañizares, 1999; Verlaan, 1998) or Ensemble Kalman filter (EnKF, Evensen, 2006) are ones of the most used when new observations are available.

Due to the complex nature of the hydrological processes, spatially and temporally distributed measurements are needed in the model updating procedures to ensure a proper flood prediction (Clark et al., 2008; Mazzoleni et al., 2015; Rakovec et al., 2012). However, traditional physical sensors require proper maintenance and personnel which can be very expensive in case of a vast network. For this reason, the technological improvement led to the spread of low-cost sensors used to measure hydrological variables such as water level or precipitation in a distributed way. An example of such sensors, defined in the following as "social sensor", is a smart-phone camera used to measure the water level at a staff gauge with an associate QR code used to infer the spatial location of the measurement (see **Error! Reference source not found.**). The main advance of using these type of sensors is that they can be used not only by technicians but also by regular citizens, and that due to their reduced cost a more spatially distributed coverage can be achieved. The idea of designing such alternative networks of low-cost social sensors and using the obtained crowdsourced observations is the base of the EU-FP7 WeSenseIt project (2012-2016), which also sponsors this research. Various other projects have also been initiated in order to assess the usefulness of crowdsourced observations inferred by low-cost sensors owned by citizens. For instance, in the project CrowdHydrology (Lowry and Fienen, 2013), a method to monitor stream stage at designated gauging staffs using crowd source-based text messages of water levels is developed using untrained observers. Cifelli et al. (2005) described a community-based network of volunteers (CoCoRaHS), engaged in collecting precipitation measurements of rain, hail and snow. An example of hydrological monitoring, established in 2009, of rainfall and streamflow values within the Andean ecosystems of Piura, Peru, based on citizen observations is reported in Célleri et al. (2009). Degrossi et al. (2013) used a network of wireless sensors in order to map the water level in two rivers passing by Sao Carlos, Brazil. Recently, the iSPUW Project is aims to integrate data from advanced weather radar systems, innovative wireless sensors and crowdsourcing of data via mobile applications in order to better predict flood events in the urban water systems of the Dallas-Fort Worth Metroplex (ISPUW, 2015; Seo et al., 2014). Other examples of crowdsourced the water-related information include the so-called Crowdmap platform for collecting and communicating the information about the floods in Australia in 2011 (ABC, 2011), and informing citizens about the proper time to drink water in an intermittent water system (Alfonso, 2006; Au et al., 2000; Roy et al., 2012). A detailed and interesting review of the

examples of citizen science applications in hydrology and water resources science is provided by Buytaert et al. (2014)

The traditional hydrological observations from physical sensors have a well defined structure in terms of frequency and accuracy. On the other hand, crowdsourced observations are provided by citizens with varying experience of measuring environmental data and little connections between each other, and the consequence is that the low correlation between the measurements might be observed. So far, in operational hydrology practice, the added value of crowdsourced data it is not integrated into the forecasting models but just used to compare the model results with the observations in a post-event analysis. One reason can be related to the intrinsic variable accuracy, due to the lack of confidence in the data quality from such heterogeneous sensors, and the variable life-span of the observations.

Regarding data quality, Bordogna et al. (2014) and Tulloch and Szabo (2012) stated that quality control mechanisms should consider contextual conditions to deduce indicators about reliability (expertise level), credibility (volunteer group) and performance of volunteers such as accuracy, completeness and precision level. Bird et al. (2014) addressed the issue of data quality in conservation ecology by means of new statistical tools to assess random error and bias in such observations. Cortes et al. (2014) evaluated data quality by distinguishing the in-situ data collected between a volunteer and a technician and comparing the most frequent value reported at a given location. They also gave some range of precision according to the rating scales. With in-situ exercises, it might be possible to have an indication of the reliability of data collected (expertise level). However, this indication does not necessarily lead to a conclusion of high, medium or low accuracy every time a streamflow observation of a contributor is received. In addition, such approach is not enough at operational level to define accuracy in data quality. In fact, every time a crowdsourced observation is received in real-time, the reliability and accuracy of observations should be identified. To do so, one possible approach could be to filter out the measurements following a geographic approach which defines semantic rules governing what can occur at a given location (e.g. Vandecasteele and Devillers, 2013). Another approach could be to compare measurements collected within a pre-defined time-window in order to calculate the most frequent value, the mean and the standard deviation.

Regarding the variable life-span, crowdsourced observations can be defined as asynchronous because do not have predefined rules about the arrival frequency (the observation might be sent just once, occasionally or at irregular time steps which can be smaller than the model time step) and accuracy. In a recent paper, Mazzoleni et al. (2015) we have presented results of the study of the effects of distributed synthetic streamflow observations having synchronous intermittent temporal behaviour and variable accuracy in a semi-distributed hydrological model. It has been shown that the integration of distributed uncertain intermittent observations with single measurements coming from physical sensors would allow for the further improvements in model accuracy. However, we have

not considered the possibility that the asynchronous observations might be coming at the moments not coordinated with the model time steps. A possible solution to handle asynchronous observations in time with EnKF is to assimilate them at the moments coinciding with the model time steps (Sakov et al., 2010). However, as these authors mention, this approach requires the disruption of the ensemble integration, the ensemble update and a restart, which may not feasible for large-scale forecasting applications. Continuous approaches, such as 3D-Var or 4D-Var methods, are usually implemented in oceanographic modeling in order to integrate asynchronous observations at their corresponding arrival moments (Derber and Rosati, 1989; Huang et al., 2002; Macpherson, 1991; Ragnoli et al., 2012). In fact, oceanographic observations are commonly collected at not pre-determined, or asynchronous, times. For this reason, in variational data assimilation, the past asynchronous observations are simultaneously used to minimize the cost function that measures the weighted difference between background states and observations over the time interval, and identify the best estimate of the initial state condition (Drecourt, 2004; Ide et al., 1997; Li and Navon, 2001).In addition to the 3D-Var and 4D-Var methods, Hunt et al. (2004) proposed a Four Dimensional Ensemble Kalman Filter (4DEnKF) which adapts EnKF to handle observations that have occurred at non-assimilation times. In this method the linear combinations of the ensemble trajectories are used to quantify how well a model state at the assimilation time fits the observations at the appropriate time. Furthermore, in case of linear dynamics 4DEnKF is equivalent to instantaneous assimilation of the measured data (Hunt et al., 2004). Similarly to 4DEnKF, Sakov et al. (2010) proposed the Asynchronous Ensemble Kalman Filter (AEnKF), a modification of the EnKF, mainly equivalent to 4DEnKF, used to assimilate asynchronous observations (Rakovec et al., 2015). Contrary to the EnKF, in the AEnKF current and past observations are simultaneously assimilated at a single analysis step without the use of adjoint model. Yet another approach to assimilate asynchronous observations in models is the so-called First-Guess at the Appropriate Time (FGAT) method. Like in 4D-Var, the FGAT compares the observations with the model at the observation time. However, in FGAT the innovations are assumed constant in time and remain the same within the assimilation window (Massart et al., 2010). Having reviewed all the described approaches, in this study we have decided to use a straightforward and pragmatic method, due to the linearity of the hydrological models implemented in this study, similar to the AEnKF to assimilate the asynchronous crowdsourced observations.

The main objective of this novel study is to assess the potential use of crowdsourced observations within hydrological modelling. In particular, the specific objectives of this study are to a) assess the influence of different arrival frequency of the crowdsourced observations and their related accuracy on the assimilation performances in case of a single social sensor; b) to integrate the distributed low-cost social sensors with a single physical sensor to assess the improvement in the flood prediction performances in an early warning system. The methodology is applied in the Brue (UK) and Bacchiglione (Italy) catchments,

considering lumped and semi-distributed hydrological models respectively. The Brue catchment is considered because of the availability of precipitation and streamflow data, while the Bacchiglione river is one of the official case studies of the WeSenseIt Profect (Huwald et al., 2013), which is funding this research. Due to the fact that streamflow observations from social sensors are not available in the Brue catchment while in the Bacchiglione catchment the sensors are being recently installed, the synthetic time series, asynchronous in time and with random accuracy, that imitate the crowdsourced observations, are generated and used.

The study is organized as follows. Firstly, the case studies and the datasets used are presented. Secondly, the hydrological models used are described. Then, the procedure used to integrate the crowdsourced observations is reported. Finally, the results, discussion and conclusions are presented."

RC: *Is the discussion about oceanographic studies / models needed? It was not clear to me what that material added to the paper. If it is needed, please make it more clear what the connection is? Is it technology of oceanographic models that can be used in your process of including CS data into models?*

AC: DA in oceanographic models is in the paper since oceanographic observations are commonly collected at not pre-determined, or asynchronous, times – and this has relevance for the paper. Indeed, the DA technologies used in oceanography (continuous (variational DA)) could have been used also for hydrology, but they require building adjoint models and this limits their use in case of using real complex hydrological models. (The reasons of using KF instead are given in Introduction.)

RC: *Why is the MIKE11 model presented as the model for representing flood propagation on the main channel in the Bacchiglione basin? Immediately after stating that the MIKE11 model was used, it appears that it was replaced by the Muskingum - Cunge model. Maybe they were used to represent two different processes in this basin? Obvioulsy, this was not clear. If you used the M-C model, then why even bother with the MIKE11 part of the discussion? Please reconsider your wording to make it clear. If both were used, please explain the role of each.*

AC: We thank the reviewer for pointing out this aspect of our study, which perhaps was not clearly explained. MIKE11 model was originally used by AAWA, the water authority, within their early warning system on the Bacchiglione basin. However, in this study, in order to reduce the computational time of the simulations and since main part of the uncertainty sources come from the hydrological model, MIKE11 was replaced with a Muskingum-Cunge model. We mentioned this point in section 3.2:

"In the early warning system implemented by AAWA in the Bacchiglione catchment, the flood propagation along the main river channel is represented by one-dimensional hydrodynamic model, MIKE 11 (DHI, 2005). This model solves the Saint-Venant equations in case of unsteady flow based on an implicit finite difference scheme proposed by Abbott and Ionescu (1967). However, in order to reduce the computational time required by the analysis performed in this study MIKE11 is replaced by a hydrological routing Muskingum-Cunge model (see, e.g. Todini 2007), considering river cross-sections as rectangular for the estimation of hydraulic radius, wave celerity and the other hydraulic variables"

Due to limitation in the number of figures we do not present a comparison between MIKE11 and MC routing. However, we are currently working on a study in which we demonstrate that these two methods show similar results in terms of estimated discharge at Ponte degli Angeli.

RC: *Increases in model accuracy due to assimilating CS observations needs to be presented in different ways. I understand the value in evaluating model accuracy and improvements in accuracy in terms of NSE. Several times in the paper, the value of including these CS observations is couched in terms of increased accuracy of flood peak magnitudes and timing. Discussing this increased accuracy in terms of NSE only is not all that informative. Statistics such as NSE only speak to overall model accuracy, not to real increases/decreases in prediction error. Please include discussion about percent change in flood peak prediction (in text and/or table) for a few of the peaks in your evaluation period.*

AC: Indeed, the averages statistics like NSE may not correctly present the model performance gains during floods, so the other error metrics which reflect flood-time performance more explicitly can be used. As suggested by the reviewer we have carried out additional analyses to assess the change in flood peak prediction considering 3 peaks occurred during flood event 2 (see Figure 3), in the Brue catchment.

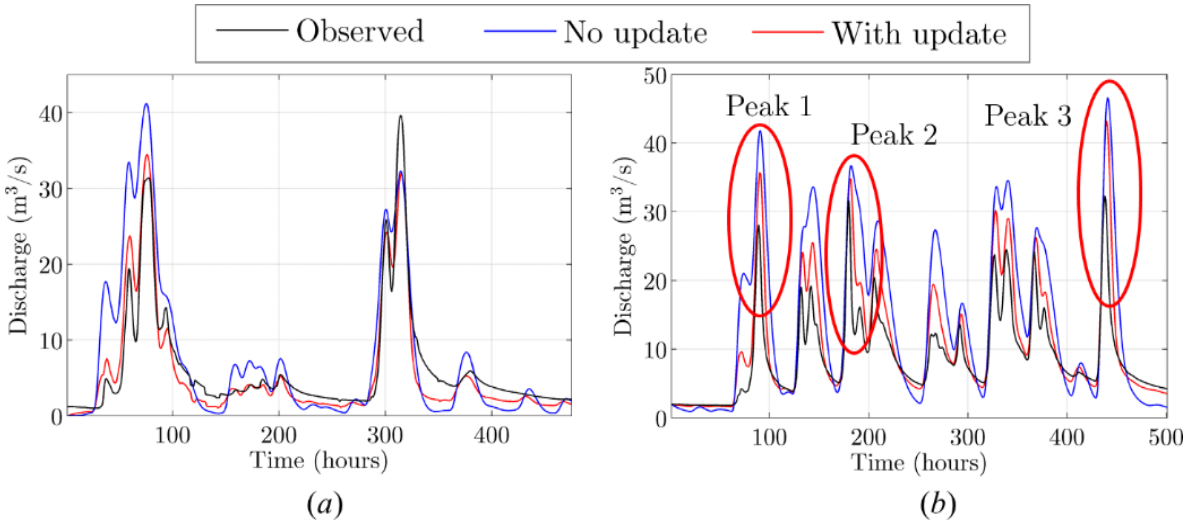

Figure 1. Indication of the 3 flood peak occurred during flood event 2 in Brue catchment

Error in the flood peak timing and intensity is estimated using Err_t and Err_I equal to:

$$Err_t = t_P^o - t_P^S . \quad (1)$$

$$Err_I = \frac{Q_P^o - Q_P^S}{Q_P^o} . \quad (2)$$

Where t_p^o and t_p^s are the observed and simulated peak time (hours), while Q_p^o and Q_p^s are the observed and simulated peak intensity (m^3/s). From the results in Figure 2 and 3, considering 12-hours lead time, it can be observed that, overall, errors reduction is achieved for increasing number of observations within 1 hour. In particular, assimilation of CS observations has more influence in the reduction of the peak intensity rather than peak timing. In fact, as Figure 4 shows, e.g. in case of peak 1, a small reduction of Err_t is obtained even increasing the number of CS observations. In fact, in all the 3 considered peaks, maximum reduction in Err_t is around 1 hour. On the other hand, higher error reduction is achieved if we considered the peak intensity rather than its timing. In particular. Smaller Err_I error values are obtained in case of scenario 1, while scenario 5 is the one that shows the lowest improvement in terms of peak prediction. This can be related to the random moment and accuracy of the CS observation in such scenario. Similar results are obtained in case of scenario 6 and 9.

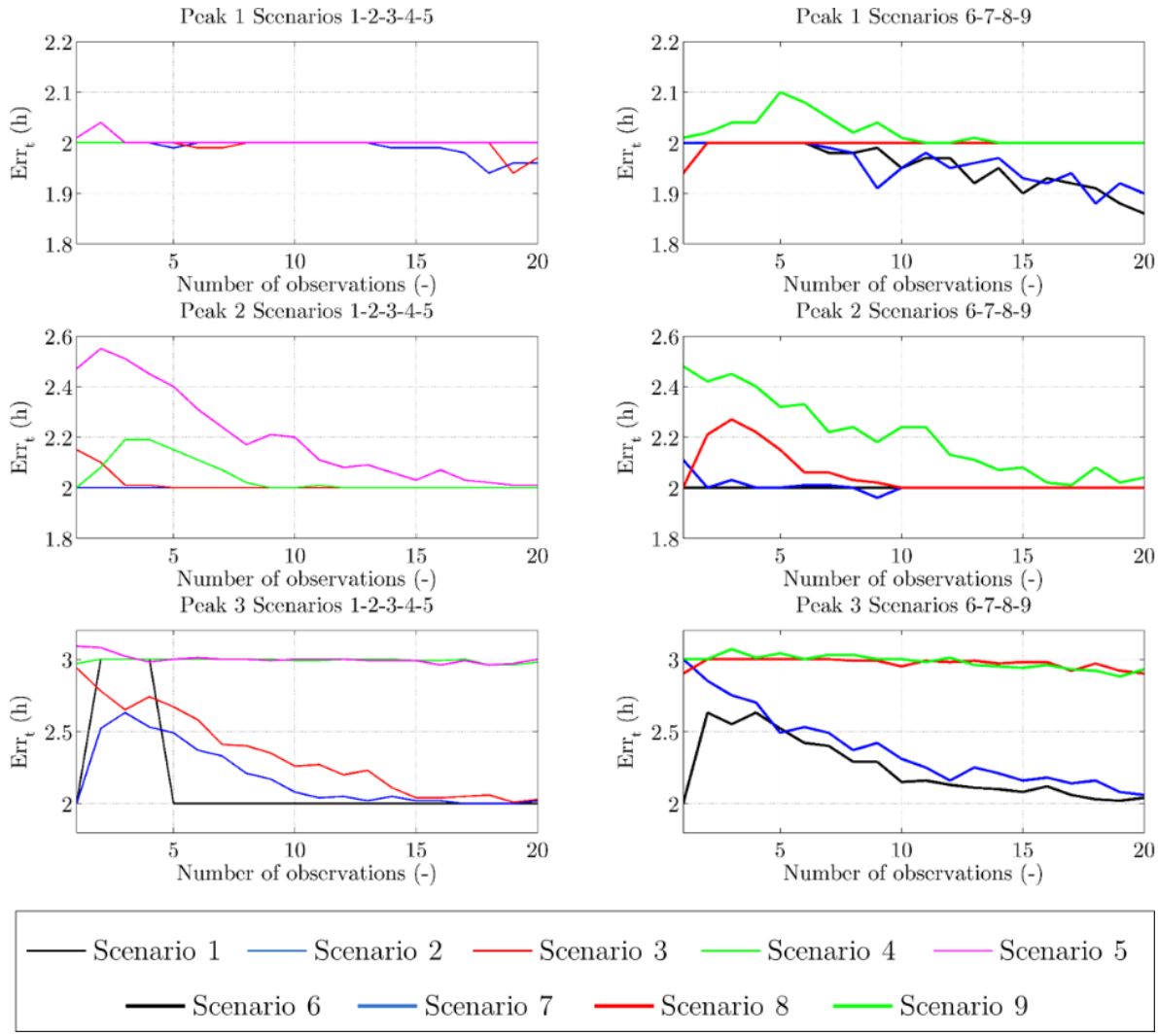

Figure 2. Representation of Err_t as function of the number of CS observations for 3 different peaks in case of scenarios from 1 to 9

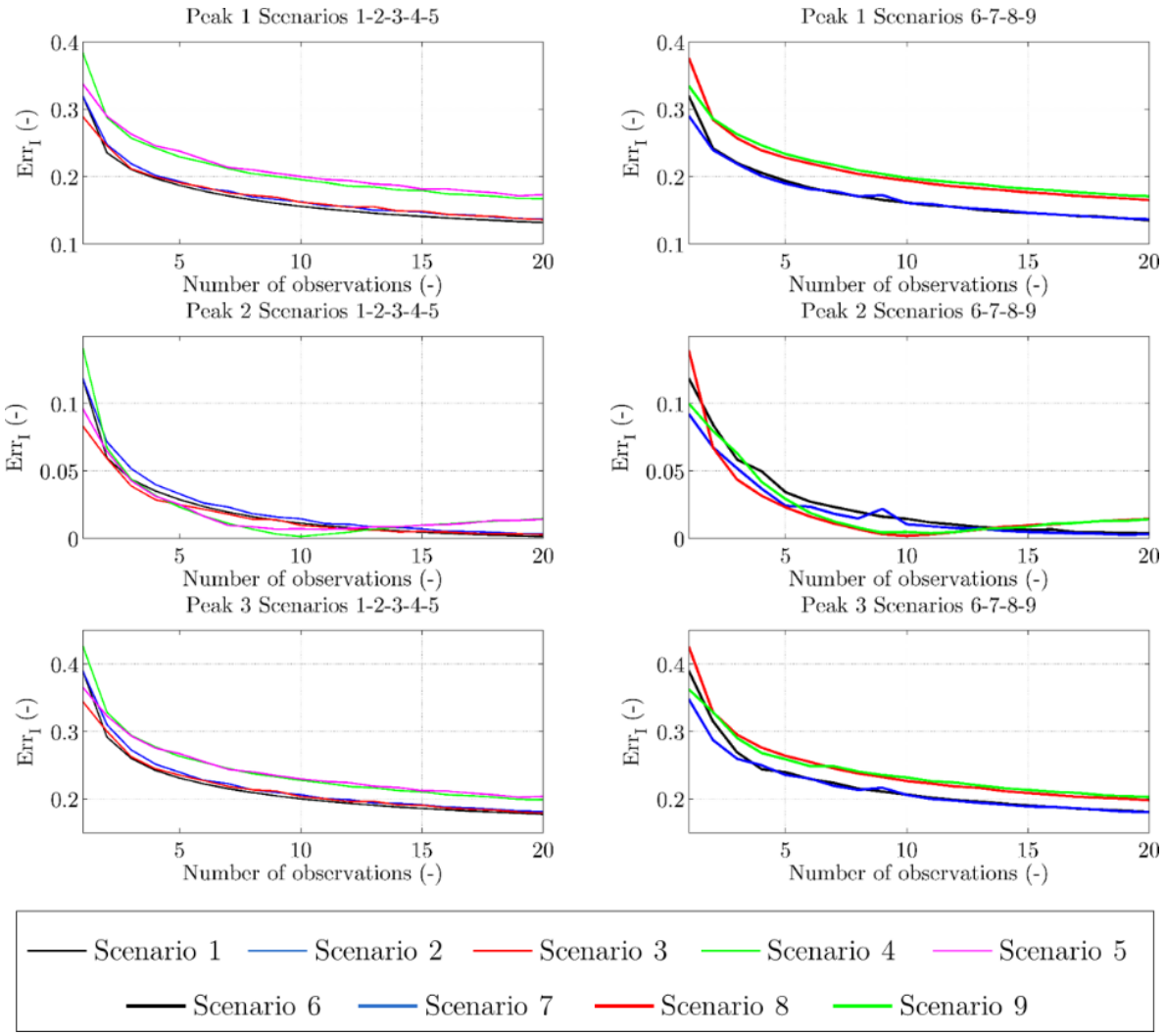

Figure 3. Representation of Err_I as function of the number of CS observations for 3 different peaks in case of scenarios from 1 to 9

These conclusions are very similar to the ones obtained analysing only NSE as model performance measures. This can be related to the linear nature of the model and the consequent DA approach used in this work. Due to the already high number of figures included in the revised version of manuscript, we have prepared an additional table (see below) indicating the percentage of error Err_t and Err_I reduction for each scenario changing from the assimilation of 1 to 20 observations. We leave the decision to the Editor whether to add or not these latest results and Figures/Table.

Table 1. Percentage of error Err_t and Err_I reduction

	Err_t			Err_I		
Scenario	**peak1**	**peak2**	**peak3**	**peak1**	**peak2**	**peak3**
1	0	0	0	0.588007	0.990132	0.545782
2	0.02	0	-0.01	0.571863	0.97161	0.535791
3	0.015	0.069767	0.309524	0.529608	0.967312	0.480587
4	0	0	-0.00337	0.564742	0.897512	0.535304
5	0.004975	0.186235	0.029126	0.486836	0.855197	0.443373
6	0.07	0	-0.02	0.578038	0.969569	0.537394
7	0.05	0.052133	0.313333	0.528956	0.96833	0.481274
8	-0.03093	0	0	0.560649	0.896826	0.535814
9	0.004975	0.177419	0.023333	0.489236	0.858807	0.441452