Assimilation of probabilistic flood maps from SAR data into a coupled hydrologic-hydraulic forecasting model: a proof of concept.

Concetta Di Mauro[1,2], Renaud Hostache[1], Patrick Matgen[1], Ramona Pelich[1], Marco Chini[1], Peter Jan van Leeuwen[2,4], Nancy Nichols[2], and Günter Blöschl[3]

[1]Luxembourg Insitute of Science and technology
[2] University of Reading,UK
[3]Vienna University of Technology
[4]Department of Atmospheric Science, Colorado State University, USA

Correspondence: Concetta Di Mauro (concetta.dimauro@list.lu)

Abstract. Coupled hydrologic and hydraulic models represent powerful tools for simulating streamflow and water levels along the riverbed and in the floodplain. However, input data, model parameters, initial conditions and model structure represent sources of uncertainty that affect the reliability and accuracy of flood forecasts. Assimilation of satellite-based Synthetic Aperture Radar (SAR) observations into a flood forecasting model is generally used to reduce such uncertainties. In this context, we have evaluated how sequential assimilation of flood extent derived from SAR data can help improve flood forecasts. In particular, we carried out twin experiments based on a synthetically generated data-set with controlled uncertainty. To this end, two assimilation methods are explored and compared: the Sequential Importance Sampling (standard method) and its enhanced method where a tempering coefficient is used to inflate the posterior probability (adapted method) and to reduce degeneracy. The experimental results show that the assimilation of SAR probabilistic flood maps significantly improves the predictions of streamflow and water elevation, thereby confirming the effectiveness of the data assimilation framework. In addition, the assimilation method significantly reduces the spatially averaged root mean square error of water levels with respect to the case without assimilation. The critical success index of predicted flood extent maps is significantly increased by the assimilation. While the standard method proves to be more accurate in estimating the water levels and streamflow at the assimilation time step, the adapted method enables a more persistent improvement of the forecasts. However, although the use of a tempering coefficient reduces the degeneracy problem, the accuracy of model simulation is lower than the one of the standard method at the assimilation time step.

1 Introduction

Floods represent one of the major natural disasters with a global annual average loss of US $104 billion (UNISDR, 2015). Extent of flood damages have risen during the last years due to climate-driven changes and an increase in the asset values

of floodplains (Blöschl et al., 2019). This emphasizes the need for reliable and cost-effective flood forecasting models to predict flood inundations in near real-time. Hydrologic and hydraulic models represent useful tools for simulating flood extent, discharge and water levels in the river bed and on the floodplain. However, both the models and the observations used as inputs for running, calibrating and evaluating the models are affected by uncertainty.

Data assimilation (DA) aims at improving model predictions by updating model states and/or parameters based on observations (Moradkhani et al., 2005). It optimally combines observations with the system state derived from a numerical model accounting for both model and observation errors. Ideally, *in situ* data are systematically assimilated into flood forecasting models, but these observations are not always available (e.g. in *ungauged* catchments) and only provide space-limited information (Grimaldi et al., 2016). Therefore, satellite Earth Observation (EO) data, and in particular Synthetic Aperture Radar (SAR) images, represent a valuable complementary dataset to *in situ* observations due to their capacities to provide frequent updates of flooded areas at a large scale. In addition, as the corresponding EO data archives are growing fast, historical observational data spanning an extended period of time can be assimilated into large scale hydrodynamic models. SAR sensors are able to acquire images of flooded areas and permanent water bodies during day and night almost regardless of weather conditions. The backscattered signal depends on the dielectrical properties of the imaged objects. Smooth surfaces, such as open water bodies, interact with the transmitted pulse so that a very limited part of the signal is backscattered to the satellite resulting in dark areas in the acquired image. Different information about water extent can be extracted from a SAR image and used to improve the forecasts using DA techniques.

Directly assimilating flood extent maps is not straightforward because these do not correspond to a state variable of the model. Therefore, some studies suggested to transform the SAR backscatter information into state variable prior to the assimilation. For instance, several studies have used EO-derived water levels to improve flood forecasts [e.g. Andreadis et al. (2007), García-Pintado et al. (2015), Matgen et al. (2010), Revilla-Romero et al. (2016), Giustarini et al. (2011), Hostache et al. (2010)]. The water levels are estimated by merging pre-selected flood extent limits extracted from the SAR satellite imagery with a digital elevation model (DEM). This step requires precise flood contour maps and high resolution DEMs which are not always available (Hostache et al., 2018).

In the existing literature only a few studies have used DA for directly assimilating flood extent maps into flood forecasting models [e.g. Lai et al. (2014), Revilla-Romero et al. (2016), Cooper et al. (2018b), Cooper et al. (2018a), Hostache et al. (2018)]. Among the advantages of a direct use of the SAR backscatter values is that it reduces the data processing time that is a key-element in near-operational applications. ~~The main difficulty of assimilating flood extent is due to the fact that the latter is not a state variable of the model since it only represents the set of water pixels of the satellite image So far, in the existing literature, only a few studies have used used Kalman Filter (KF), Four-Dimensional Variational (4DVar) and Particle Filter (PF) techniques for assimilating flood maps into flood forecasting models [Lai et al. (2014), Revilla-Romero et al. (2016), Cooper et al. (2018), Cooper et al. (2019), Hostache et al. (2018)]. Lai et al. (2014) have assimilated the surface water extent extracted from 250 m resolution MODIS data via the 4DVar. Revilla-Romero et al. (2016) have used the ensemble Kalman filter (EnKF), which is a variant of the KF where an ensemble of state vectors are drawn from the distribution of the state to repres,~~

~~to Revilla-Romero et al. (2016) assimilate the information of surface water extent from the GFDS (Global Flood Detection System) data with a resolution of 10 km~~

~~Cooper et al. (2019) have used an EnKF to update a 2D hydrodynamic model. In this case, the backscatter values are not transformed into state variables of the system. The simulated dry and wet pixels are converted into equivalent SAR backscatter values corresponding to the spatial mean of the SAR backscatter observations at a given time. Cooper et al. (2019) compared the SAR backscatter-based assimilation method with the flood edge assimilation method and showed that the new observation operator performs well compared to the assimilation of flood-edge water elevation observations.~~ Cooper et al. (2018a) have used an Ensemble Kalman Filter to update a 2D hydrodynamic model. In this case, the backscatter values are directly assimilated into the model without being transformed into state variables of the flood forecasting system. The dry and wet pixels of the simulated binary flood map are converted into equivalent SAR backscatter values corresponding to the spatial mean of the SAR observations. Cooper et al. (2018a) showed that the SAR backscatter-based assimilation method performs well compared to the assimilation method where the SAR backscatter is transformed into water levels.

~~Even though 4DVar and KFs may account for the non-linearity of the system evolution, the probability density funtions (PDFs) of the observations and of the model errors are still considered Gaussian and the analysis step is built-up on iterative linearizations of model equations and observation operators (van Leeuwen, 2010). PFs have the advantage of relaxing the assumption that PDFs of both the observation and model errors are Gaussian (Moradkhani et al., 2005).~~

Hostache et al. (2018) used a variant of the Particle Filter (PF) with Sequential Importance Sampling (SIS), to assimilate probabilistic flood maps (PFMs) derived from SAR data into a coupled hydrologic-hydraulic model with the assumption that the rainfall is the main source of uncertainty together with SAR observations. Their study showed that the assimilation of PFMs is beneficial: the number of correctly predicted flooded pixels increases as compared to the case without any assimilation, hereafter called *Open Loop* (OL). Forecast errors are reduced by a factor of 2 at the assimilation time and improvements persist for subsequent time steps up to 2 days. However, the improvements are not systematic: for some cases the updated hydraulic output deviates from the observations.

One of the reasons for such outliers could be the assumption that rainfall represents the dominating source of uncertainty together with satellite observation errors, thereby excluding other possible sources of uncertainty in the model system such as input data, model parameters, initial conditions and model structure. Even though the assumption seems to be rather realistic and suitable in operational cases, given that rainfall uncertainty has been generally identified as one of the major causes of poorly performing models [Koussis et al. (2003), Pappenberger et al. (2005)], coupled models may have additional sources of uncertainty affecting the results.~~In this context,~~

The present study is a follow up of the real world experiment by Hostache et al. (2018) ~~; we carry out a similar experiment but this time in a controlled environment so that rainfall is effectively the only source of uncertainty~~ and carries out a similar experiment in a controlled environment that considers the estimated rainfall together with SAR observations as the only sources of uncertainty. ~~Therefore, the objective of this study is to further assess and validate the proposed framework when the underlying assumptions are respected.Moreover,~~ Hostache et al. (2018) also highlighted that degeneracy may be a major issue of PFs: after the assimilation, the number of particles with high weights reduces to a few or only one particle so that the ensemble

loses statistical significance. To overcome this issue, Hostache et al. (2018) used a site-dependent tempering coefficient which inflates the posterior probability. In our study, we propose to adopt an enhanced tempering coefficient as a function of the desired effective ensemble size (EES) after the assimilation.

~~The detection of flooded areas in SAR images could be rather straightforward. However, for particular cases the SAR backscatter values of flooded and non-flooded areas are difficult to distinguish, leading to a wrong classification in the flood mapping results. Such errors could be due to particular atmospheric conditions (e.g. wind, snow, precipitation), to water-look like areas (e.g. asphalt, sand, shadow), to obstructing objects (e.g. dense canopy, buildings) or errors inherent to coherent imaging systems (e.g. speckle) as mentioned in (Giustarini et al., 2015). Detecting and removing these errors represent one of the main scientific challenges of using SAR data for the systematic, fully automated, and large-scale flood monitoring.~~

Moreover, in Hostache et al. (2018) speckle errors in the SAR observations, are taken into account through the Bayesian approach introduced by Giustarini et al. (2016). However, no conclusions are drawn concerning the effect of misclassified pixels. In fact, for some particular cases such as densely vegetated areas, the detection of floodwater from SAR imagery is known to be prone to errors. Detecting and removing such errors represents one of the main scientific challenges of using SAR data for a systematic, fully automated, large-scale flood monitoring (and prediction).

The main objective of the present study is to assess the strengths and the limitations of the DA framework previously proposed by Hostache et al. (2018). To do that we evaluate the DA framework in a fully controlled environment via synthetic twin experiments as this shall allow us drawing unambiguous and comprehensive conclusions. In addition, we conduct a sensitivity analysis of the DA framework with respect to the critical tempering coefficient that was recently introduced for tackling degeneracy more efficiently. We also aim to evaluate the effect of misclassified SAR pixels on DA. Therefore, errors are artificially added within the SAR image with the aim of getting a better understanding on how robust the proposed method is with respect to ~~the proportion of misclassified SAR pixels~~ this type of errors. Results are evaluated not only locally but also over the entire flood domain and for subsequent time steps to the assimilation. To carry out the experimental study we apply the DA framework to a forecasting system consisting of a loosely-coupled hydrologic model (SUPERFLEX) and hydraulic model (LISFLOOD-FP). The meteorological data that are used to run the experiments are derived from the ERA-5 archive with a spatial resolution of 25 km and a temporal resolution of 1 hour. The SAR data are synthetically generated with a ~~resolution~~ pixel spacing of 75 m.

2 Methods

The proposed methodology is based on numerical experiments conducted with synthetically generated data as illustrated in the flow chart given in Figure 1. In this framework, the following data inputs and models are employed:

1. *True* rainfall time series are used to generate the *true* hydrologic-hydraulic model simulation.
2. Synthetic SAR observations are generated from the *true* model run (i.e. from the simulated flood extent map).

3. The *true* rainfall time series are randomly perturbed ~~to obtain an ensemble of upstream boundary discharges via the hydrologic model~~ and used as inputs of the hydrologic model. The simulated discharge ~~data~~ are then used as boundary conditions to realize an ensemble of hydraulic model runs.

4. The synthetic SAR observations are assimilated into the coupled hydrologic-hydraulic model via different variants of the ~~SIS-based approach~~ Particle Filter (PF) .

The three conducted experiments are summarized as follows:

(a) An application of the standard PF where degeneracy occurs;

(b) An application of the adapted PF where a tempering coefficient is used to reduce degeneracy. We also investigated the sensitivity of the DA results to different values for the tempering coefficient, corresponding to ~~effective ensemble sizes~~ EES of 5, 10, 20 and 50%;

(c) An application of both proposed methods with artificially introduced known errors into the SAR image classification in order to evaluate the impact of these errors on the DA performance metrics.

2.1 Coupled hydrologic-hydraulic model: synthetic truth and ensemble

The coupled modelling system consists of a hydrologic model coupled with a hydraulic model. The hydrologic model is used to compute the run-off at the upstream boundaries of the hydraulic model. The hydrologic model used in this study is SUPERFLEX which is a framework for conceptual hydrologic modelling introduced by Fenicia et al. (2011). The model structure is a combination of generic components: reservoirs, connection elements and lag functions. In this study, a lumped conceptual model and its structure as a combination of three reservoirs are used: an unsaturated soil reservoir with storage S_{UR}, a fast reacting reservoir with storage S_{FR}, and a slow reacting reservoir with storage S_{SR}. A lag function has been added at the outlet of the slow and fast reacting reservoirs. The hydraulic model is based on LISFLOOD-FP (Bates and Roo, 2000; Neal et al., 2012) and simulates flood extent, water levels and streamflows along the river and on the floodplain. A sub-grid 1D kinematic solver is used for the channel flow. When the storage capacity of the river is exceeded, the water spills into the floodplain and a 2D diffusion wave scheme neglecting the convective acceleration (de Almeida and Bates, 2013; Bates et al., 2010) is used for the floodplain flow simulation. ~~Channel width, channel depth, slope of terrain, friction of the flood domain and channel bathymetry are defined in each cell of the model domain as described in Wood et al. (2016). A uniform flow condition is imposed downstream. No lateral inflow in the hydraulic model is assumed.~~

The *true* meteorological data (i.e., temperature and rainfall) are used as input of the hydrologic-hydraulic model to simulate streamflow and water level time series and to provide binary flood maps, where each pixel is classified as flooded (with value 1) or not flooded (with value 0), at each assimilation time step. These computational results represent the synthetic *truth* that will be used to evaluate the performance of the proposed assimilation framework. The *true* binary flood maps are also used to generate the synthetic SAR observations as described in the next section.

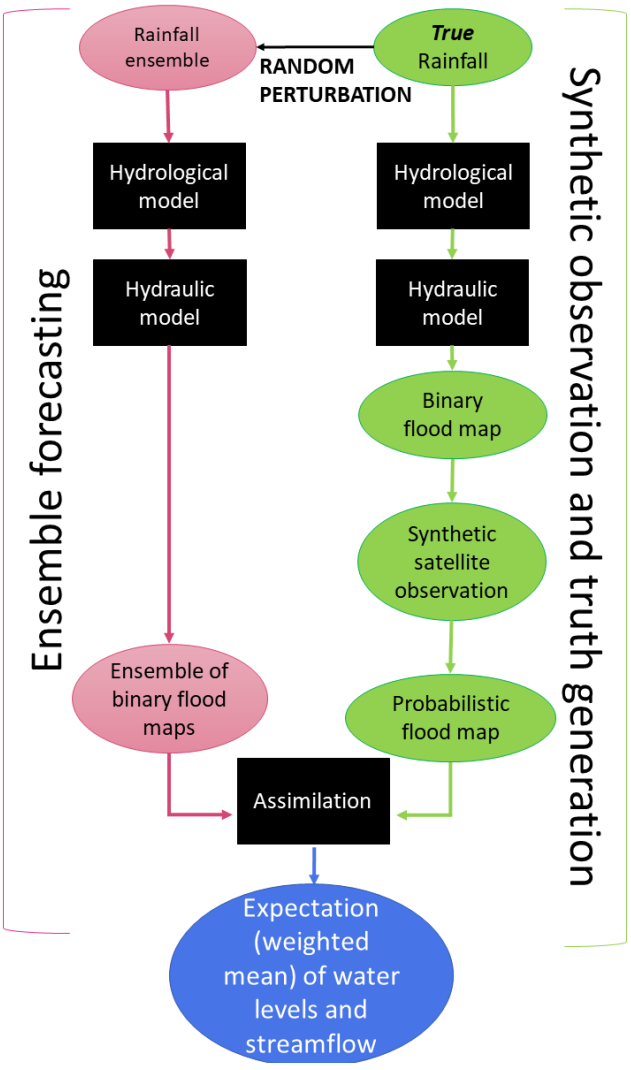

Figure 1. Flow chart of the synthetic experiment. The *true* rainfall is perturbed. The same flood forecasting model structure composed of a hydrologic model and a hydraulic model is used to obtain the probabilistic flood map and the ensemble of binary flood maps. The probabilistic flood map is assimilated into the ensemble of binary flood maps via the Particle Filter to obtain the weights with which the expectation of water levels, streamflow and flood extents are computed. ~~Flow chart of the synthetic experiment. The *true* rainfall is perturbed. The same flood forecasting model structure composed of a hydrologic model and a hydraulic model is used to obtain the probabilistic flood map, with the use of a reference SAR image, and an ensemble of binary flood maps. The probabilistic flood map is assimilated into the ensemble of binary flood maps to obtain the weights which are then used to compute the expected mean of water levels and discharge.~~

2.2 Synthetic observations

In the proposed synthetic experiment, we generate synthetic SAR images at each assimilation time step ~~corresponding to the real acquisition plan of the Sentinel-1 satellite constellation~~. The SAR images are generated with the same spatial resolution of the LISFLOOD-FP maps. Similarly to the Van Wesemael (2019) study, we make use of a real SAR image, acquired during a flood event in the past, and of the LISFLOOD-FP model to generate *true* binary flood maps. The histogram of the SAR image backscatter values can be approximated with two Gaussian curves relative to the flooded and not flooded pixel classes. Generally, the class of flooded pixels is often represented just by a fraction of the SAR image scenes. Therefore, to identify and characterize areas where the flooded and not flooded classes are more balanced, the hierarchical split based approach [HSBA, Chini et al. (2017)] is applied to the selected SAR image. The parameters of the Gaussian PDFs are determined by fitting the histogram values of the HSBA selected areas. Then random backscatter values, derived from the calibrated Gaussian PDFs, are associated to the pixels of the *true* binary flood map indicating the presence of water and no-water areas. Once the synthetic SAR images are generated, the Giustarini et al. (2016) procedure is applied and synthetic PFMs are derived. ~~This approach has been adopted to generate synthetic observations and to determine for each pixel of a SAR image its~~ The probability to be flooded given the recorded backscatter values $p(F|\sigma^0)$ for each pixel of a SAR image ~~This probability~~ is obtained via the Bayes' theorem:

$$p(F|\sigma^0) = \frac{p(\sigma^0|F)p(F)}{p(\sigma^0)} = \frac{p(\sigma^0|F)p(F)}{p(\sigma^0|F)p(F) + p(\sigma^0|\overline{F})p(\overline{F})} \quad (1)$$

In the equation 1, $p(\sigma^0|F)$ and $p(\sigma^0|\overline{F})$ represent, respectively, the probability of the backscatter values of the flooded and non-flooded pixels, $p(F)$ is the prior probability of a pixel being flooded and $p(\overline{F})$ is the prior probability of a pixel being non-flooded before any backscatter information is taken into account.The conditional probabilities are derived from the histogram of the image backscatter values estimated from the synthetically generated SAR image. The prior can be estimated from the flood extent model output or through visual interpretation of an aerial photography in real cases. However, in general such information is not always available and the prior probabilities are unknown. Consequently, Giustarini et al. (2016) set the prior probability of equation 1 to 0.5 so that both flooded and non flooded pixels are equally likely. ~~proposed to use 0.5 as default value~~. ~~In this synthetic experiment, the prior probability is derived from the *true* binary flood map as the ratio between the number of flooded pixels and the total number of pixels at each time step.~~ While this study is based on a synthetic experiment, *true* binary flood extent maps are available. Therefore, the assimilation is realized using both the estimated prior probability (as the ratio between the flooded area and the total area) and the prior probability equal to 0.5. Given the similarity of the results for both cases, in the following sections we only discuss the experiment using the estimated prior probability.

~~SAR observations are considered unbiased in the first part of the study. The method by Giustarini et al. (2016) aims at characterizing the speckle-induced uncertainty. However, it does not consider any other particular SAR-error causes, e.g generated by atmospheric conditions or specific ground features. Therefore, areas where such errors could occur should be masked out, otherwise the estimate of SAR-based flood extent could be compromised. In the first part of this study, SAR observations are~~

~~considered unbiased.~~ The method proposed by Giustarini et al. (2016) aims at characterizing the speckle-induced uncertainty. However, it does not consider any other phenomena leading to a wrong classification in SAR-based flood maps. Particular atmospheric conditions (e.g. wind, snow, precipitation), water-look-a-like areas (e.g. asphalt, sand, shadow) or obstructing objects (e.g. dense canopy, buildings), as mentioned in Giustarini et al. (2015), can lead to a wrong classification in the flood maps. Therefore, the areas where such errors could occur should be masked out from the SAR-based flood maps in order to provide a reliable flood detection.

In the first part of this study, SAR observations are considered without errors. In the second part, these kinds of errors are integrated in the synthetic SAR observations to evaluate their effect on the DA. Specifically, the pixels along the flood edge of each particle are selected. From this set, a given number of those pixels effectively flooded in the *true* binary flood map are artificially corrupted so that they belong to dry pixels ~~according to the magnitude of error introduced in the SAR observations.~~ The number of corrupted pixels depends on the magnitude of the error that we want to introduce in the SAR observations.

2.3 Ensemble generation

In a PF the prior and posterior PDFs are approximated by a set of particles. Here, we hypothesize that the rainfall is the only source of uncertainty affecting the model-based flood extent simulations. Due to this reason, ~~different~~ an ensemble of rainfall time series is used as input of the coupled hydrologic-hydraulic model. Each rainfall time series is obtained by perturbing, with a multiplicative random noise from a log-normal error distribution, the *true* rainfall time series following the approach proposed in Hostache et al. (2018). 128 rainfall time series are obtained and forwarded in time via the hydrologic model. It is important to note that the same hydrologic-hydraulic model in terms of structure, initial conditions and parameters is used for all model runs.

The reliability of the rainfall ensemble is verified with the statistical metrics proposed by De Lannoy et al. (2006). According to the verification measurement VM_1 in equation 2:

$$VM_1 = \frac{< ensk >}{< ensp >} \approx 1 \tag{2}$$

The ensemble spread in the equation 3 (where $x_{k,n}$ 3 represents the value of the variable x at time k for each pixel n)

$$ensp_k = \frac{1}{N}\sum_{n=1}^{N}(x_{k,n} - \overline{x}_k)^2 \tag{3}$$

has to be close to the ensemble skill (equation 4)

$$ensk_k = (\overline{x}_k - y_k)^2 \tag{4}$$

which is the difference between the mean $\overline{x}_k$ over the N particles of the ensemble and the observation y_k at time k. VM_2 (equation 5) verifies that the *truth* is statistically indistinguishable from the random samples of the ensemble.

$$VM_2 = \frac{< ensk >}{< mse >} \approx \frac{(N+1)}{2N} \tag{5}$$

with *mse* estimated as:

$$mse_k = \frac{1}{N}\sum_{n=1}^{N}(x_{k,n} - y_k)^2 \tag{6}$$

VM_1 and VM_2 are used to assess the *quality* of the discharge ensemble at the output of the hydrologic model.

2.4 Data assimilation framework

The DA framework consists of two main steps: *prediction*, i.e model simulations, and *analysis*, i.e update of particle probabilities when an observation is available. The prior probability of the model state *x* at a given time *k* is represented by a set of N independent random particles x_n sampled from the prior probability distribution *p(x)* as:

$$p(x) = \frac{1}{N}\sum_{n=1}^{N}\delta(x - x_n) \tag{7}$$

where δ is the Dirac delta function. In this study, the prior probability distribution is assumed ~~supposed~~ to be uniform. The observations of flooded/not flooded pixels y are related to the true state x^t according to the following equation:

$$y = H[x^t] + \epsilon \tag{8}$$

where H is the observation operator that maps the state vector into the observation space and ϵ represents the observation errors. According to the Bayes' theorem, the observations are assimilated by multiplying the prior PDF $p(x)$ and the likelihood $p(y \mid x)$, which is the probability density of the observation given the model state, and dividing by the total probability $p(y)$, resulting in:

$$p(x \mid y) = \frac{p(y \mid x)}{p(y)}p(x) \tag{9}$$

that is the posterior probability $p(x \mid y)$, i.e. the probability density function of the model state given the observations. By inserting the equation 7 into the equation 9 we obtain the following formula:

$$p(x \mid y) \approx \sum_{n=1}^{N} W_n \delta(x - x_n) \tag{10}$$

where W_n represents the relative importance in the probability density function (i.e. global weight) given by:

$$W_n = \frac{p(y \mid x_n)}{N \cdot p(y)} = \frac{p(y \mid x_n)}{N \cdot \int p(y \mid x)p(x)\delta x} \approx \frac{p(y \mid x_n)}{\sum_j p(y \mid x_j)} \tag{11}$$

In this study, the likelihood (global weight, W_n) is represented by the product of the pixel-based likelihood (local weight, w_i), assuming the L pixel observation errors to be independent from each other.

At time k of the observation, local weights $w_{i,n}$ are defined for each particle *n* and for each pixel i according to Hostache et al. (2018):

$$w_{i,n} = p_i(F \mid \sigma_0)M_{i,n} + [1 - p_i(F \mid \sigma_0)](1 - M_{i,n}) \tag{12}$$

$w_{i,n}$ is equal to the probability of a pixel being flooded as derived from the synthetically generated SAR image. $M_{i,n}$ is equal to "1" if the model predicts the pixel as flooded, whereas $M_{i,n}$ is equal to "0" if the model predicts the pixels as non-flooded. We convert the model-based water depth maps into binary flood extent maps by considering that a pixel is flooded if its water level is above 10 cm. $p_i(F \mid \sigma_0)$ equals the probability of a pixel being flooded according to the observations, viceversa $1-p_i(F \mid \sigma_0)$ equals the probability of not being flooded. By ~~doing so~~ applying the equation 12 we assign higher probabilities to those pixels where model predictions and observations agree. Next, W_n is estimated for each particle by the normalization of the product of the local weights ~~It is computed with the normalization of the local weights~~ ensuring that the sum of the global weights is equal to 1 (equation 13, *standard method*).

$$W_n = \frac{\prod_{i=1}^{L} w_{i,n}}{\sum_{n=1}^{N} \prod_{i=1}^{L} w_{i,n}} \tag{13}$$

The expectation of the OL is equivalent to the mean of the ensemble because the relative importance of each particle is the same. The global weights are used to compute the expectation of the streamflows (Q) and water levels (h) at time (k) and per pixel (i) (see equations 14, 15).

$$\overline{h}_i = \sum_{n=1}^{N} W_n \cdot h_{i,n} \tag{14}$$

$$\overline{Q}_i = \sum_{n=1}^{N} W_n \cdot Q_{i,n} \tag{15}$$

The particles keep these global weights until the next assimilation time. ~~With the application of the SIS,~~ Particles are then set to the same equal weight before a new *analysis* step is performed.

Unless the number of particles increases exponentially with the dimension of the system-state, the particle-filter is likely to degenerate because high probability is assigned to a single particle while all other members will result in small weights (van Leeuwen et al., 2019). PFs are often subject to degeneracy issues when, due to computational reasons, the number of particles is not sufficiently high (Zhu et al., 2016). After the application of the standard PF, the variance of the weights tend to increase and only a few particles of the ensemble have a non-negligible weight. To mitigate this problem, in Hostache et al. (2018), the global weight defined in the equation 13 has been adapted using a *tempering* coefficient (α, as described by the following equation 16).

$$W_n(\alpha) = \frac{\prod_{i=1}^{L} {w_{i,n}}^{\alpha}}{\sum_{n=1}^{N} \prod_{i=1}^{L} {w_{i,n}}^{\alpha}} \tag{16}$$

Since α and weights ~~have~~ values are lower than one, adding the power of α in the weights formula allows for shifting all weight values closer to one. This therefore decreases the variance of the weights and inflates the posterior probability. After the assimilation, the number of particles with significant weight depends on the α value. The smaller α, the higher the variance of the posterior PDF. Consequently, as argued in Hostache et al. (2018), when the α coefficient is small enough, this adaptation of

the PF helps reduce the degeneracy of the ensemble. ~~Using the tempering coefficient in this way leads to biased results because of the down-weighting of the observations by increasing their errors (van Leeuwen et al., 2019).~~ While in the previous study by Hostache et al. (2018), the α value was defined so that the worst model solution would have had a non-zero global weight, in this study we propose to define α based on the desired effective ensemble size (EES). ~~The definition of~~ The coefficient α in Hostache et al. (2018) ~~was~~ is site-dependent as it relies on the number of flooded pixels, whereas in this study α is a function of the EES which is a measure of degeneracy based on the global weights (Arulampalam et al., 2002):

$$EES(\alpha) = \frac{1}{\sum_{n=1}^{N}(W_n(\alpha))^2} \cdot \frac{1}{N} \cdot 100 \tag{17}$$

The EES is lower than N and its value indicates the level of degeneracy. α is equal to one when the standard method is used. Decreasing the α coefficient leads to an increase of the EES. ~~In this study, we evaluate the effect on the DA due to variations of α. In summary, different PFs are compared with the OL and the synthetic *truth*.: the SIS (with only a few particles from the ensemble potentially carrying non- negligible weight) and the adapted method with 5-10-20-50% EES (with the number of particles with non negligible weight increasing with the EES). This methodology leads to slightly biased estimates because the observation are down-weighted. We discuss this further in the concluding section~~.

2.5 Performance metrics

To carry out the evaluation of the PFM statistics ~~, we propose to adopt the following performance metrics.~~ we have used the reliability plots. The results of the different assimilation scenarios are ~~globally~~ evaluated on a spatio-temporal scale with the following performance metrics:

- Contingency maps and the confusion matrix;
- Critical success index (CSI);
- Root mean square error (RMSE);
- Discharge and water level time series.

2.5.1 Reliability plots

Reliability diagrams are employed to statistically evaluate the synthetically generated PFMs. In such diagrams, the probability range [0;1] is subdivided into intervals of average probability P_i and width ΔP_i. We identify the pixels Ω_i having a probability value of $P_i \pm \Delta P_i$ in the PFM. The fraction of Ω_i pixels effectively flooded in the binary *truth* map are identified with F_i. The reliability diagram plots P_i on the x-axis and F_i on the y-axis. A reliability diagram indicating an alignment of data points close to the 1:1 line means that the PFM is statistically reliable.

2.5.2 Contingency maps and confusion matrix

First, we use contingency maps to graphically compare the expected flood map with the synthetic *truth* map at each assimilation time step. Pixel classification errors can be of two types: overprediction (type error I) when the pixels in the *truth* map are not flooded but are predicted as flooded, and underprediction (type error II) in the opposite case. Then, the confusion matrix numerically summarizes the results of the contingency map. It is a 2 rows by 2 columns matrix that reports the number of false positives (type I error), false negatives (type II error), true positives and true negatives.

2.5.3 CSI

The CSI evaluates the goodness of fit between the *truth* map and the predicted flood extent map (Bates and Roo, 2000):

$$CSI = \frac{A}{A+B+C} \tag{18}$$

It represents the ratio between the number of pixels correctly predicted as flooded (A) over the sum of all the flooded pixels including the false positives (B, overdetection) and false negatives (C, underdetection). CSI ranges between 0 and 1 (best score). We also used it to evaluate the results at the assimilation time step and the effect of the assimilation at subsequent time steps. It has been also used to evaluate performances when errors are added in the SAR observations.

2.5.4 RMSE

The RMSE is considered an excellent error metric for numerical predictions. The RMSE measures the square root of the average square error of the predicted water levels (h_k^p) against the truth (h_k^t) per pixel k over the total number of pixels L of the flood domain.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{L}(h_i^p - h_i^t)^2}{L}} \tag{19}$$

In this study, the RMSE is a measure of the *global* accuracy of the flood forecasting model predictions of water levels allowing to compare prediction errors of the different assimilation scenarios over the flood domain. The RMSE is evaluated at the assimilation time and also at subsequent time steps. It has been also used to evaluate performances when errors are added in the SAR observations.

3 Study Area and Data

Our synthetic experiment is grounded on a real test site and an actual storm event: the river Severn in the middle-west of UK (figure 2) and the July 2007 flood event, rispectively. This area has experienced several floods along the river valleys (Environment Agency, 2009) generally due to intense precipitation. While seven upstream catchments contribute to the flow along the river Severn, in our study only one upstream catchment is considered: the Severn at the Bewdley gauging station. Our first objective is to evaluate whether the model correctly predicts the output in the simplest case, i.e. when a unique run-off input to the

hydraulic model determines the flood extent and no additional contributing tributaries interfere. The ERA5 dataset (Hersbach et al., 2019) referring to the period of July 2007 has been used in this experiment. ERA5 is a global atmospheric re-analysis dataset provided by the European Centre for Medium-Range Weather Forecasts (ECMWF). Rainfall and 2 m air temperature at a spatial resolution of approximately 25 km and a temporal resolution of 1 hour are used as input to the hydrologic model. The *true* rainfall time series is used to generate the *true* run-off before being perturbed in order to obtain 128 different particles as inputs to the hydrologic model. The boundary condition of the hydraulic model is imposed in correspondence to the red dots in Figure 2. Channel width, channel depth, slope of terrain, friction of the flood domain and channel bathymetry are defined in each cell of the model domain as described in Wood et al. (2016). A uniform flow condition is imposed downstream. ~~No later~~ No lateral inflow in the hydraulic model is assumed. Finally, at each time step a stack of 128 wet/dry maps is obtained. Discharges and water levels recorded at different gauging stations (corresponding to the existing ones, black, blue and yellow

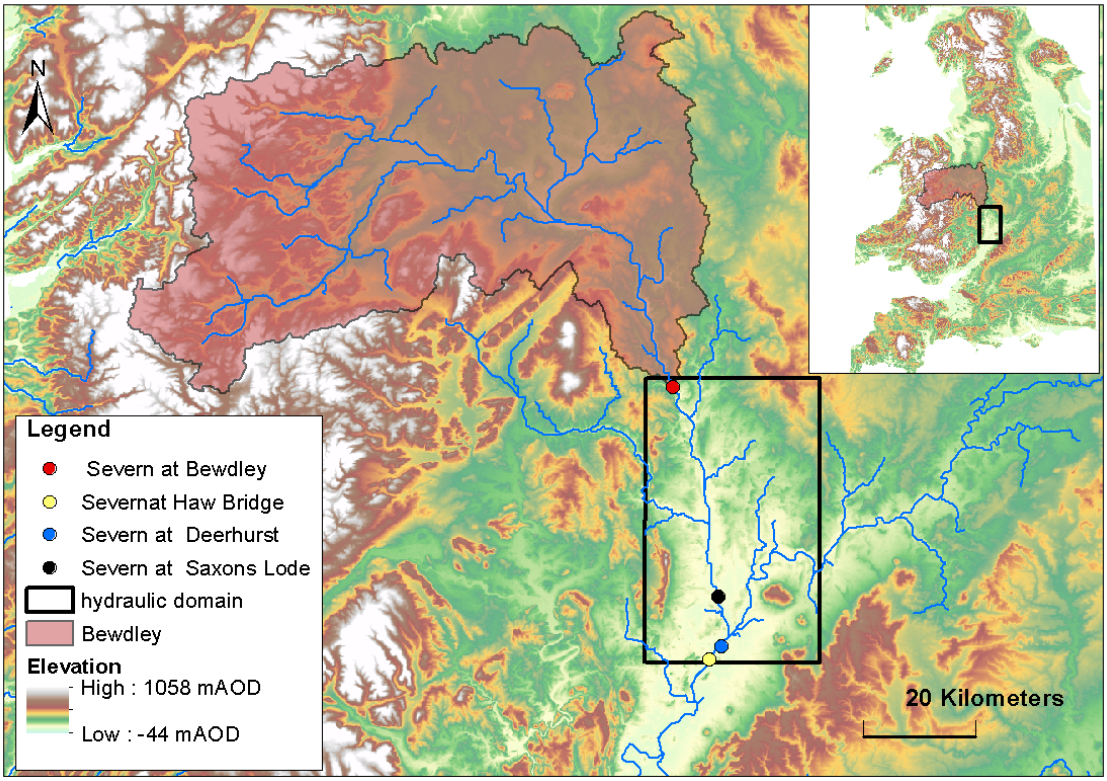

Figure 2. Study area: River Severn (UK). Only the boundary conditions in Bewdley is taken into account. Within the sub-catchment upstream of Bewdley a lumped hydrologic model is used to determine the input of the hydraulic model along the river Severn downstream. The dots represent the existing gauging stations where the performances of the DA framework are evaluated. The black square is the hydraulic domain where LISFLOOD-FP runs.

dots in figure 2) along the river are used to evaluate the performance of the DA.

4 Results

4.1 Synthetic SAR and ensemble generation and evaluation

The virtual satellite acquisition dates are aligned with the actual Sentinel-1 acquisition frequency. The revisit time over Europe, considering both ascending and descending orbits, is around 3-4 days meaning that on average 2 satellite images are available per week. In order to adopt a realistic Sentinel-1-like observation scenario we chose to assimilate four synthetic observations over a period of 10 days .

~~The prior is the probability to be flooded or not before any backscatter information is taken into account and it can be estimated from the flood extent model output or through visual interpretation of an aerial photography in real cases. However, such information are not always available. Therefore, Giustarini et al. (2016) set the prior probability of equation 9 to 0.5 so that both flooded and non flooded pixels are equally likely. In this study, being based on a synthetic experiment, *true* binary flood maps are available. Therefore, the assimilation is realized using both the estimated prior probability as the ratio between the flooded area and the total area, and the prior probability equal to 0.5. Given the similarity of the results for both cases, in the following sections we will discuss only the case with the estimated prior probability.~~ In the Figures 3 and 4 , the area corresponds to the hydraulic model domain. The hydrologic model, covering the upstream catchment, is used to compute the input boundary conditions of the hydraulic model. Results are computed and compared within the hydraulic model domain. The synthetic SAR observations are shown in Figure 3. The corresponding PFMs are shown in Figure 4 ~~at the top~~ and reliability plots are provided ~~at the bottom~~ in Figure 5. In the reliability plots, the points aligned along the 1:1 line indicate a statistically reliable PFM.

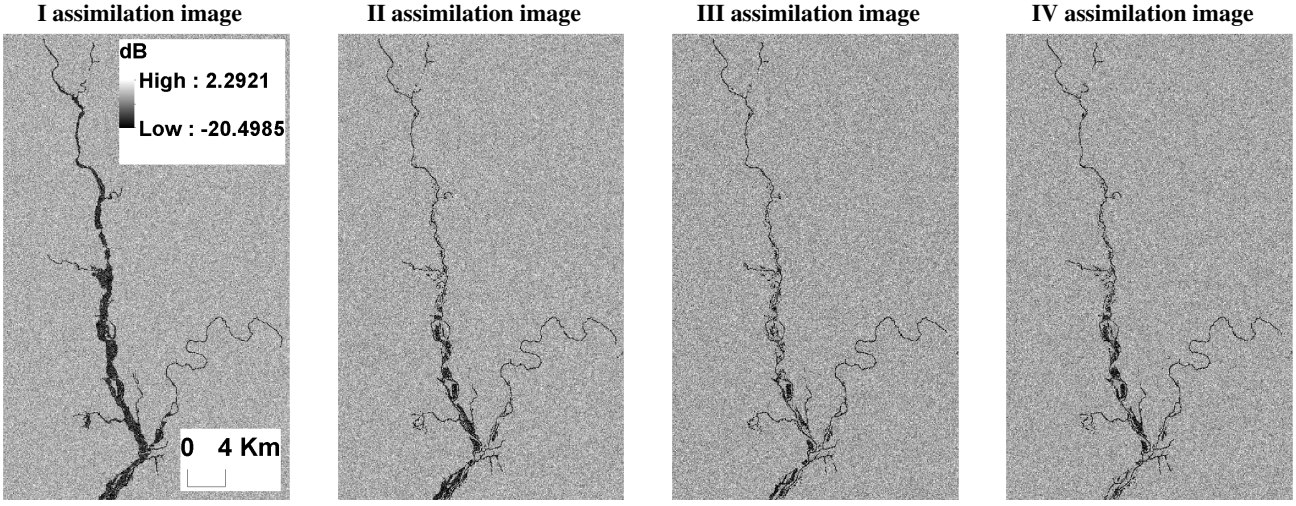

Figure 3. A detail of the synthetic SAR images corresponding to the 4 assimilation time steps. Darker pixels correspond to lower backscatter.

The verification measurements VM_1 and VM_2 (equations 2 and 5) of the ensemble discharge in Bewdley (figure 6) are equal to 1.047 and 0.527, respectively. These values are close to the ideal ones of 1 and 0.5.

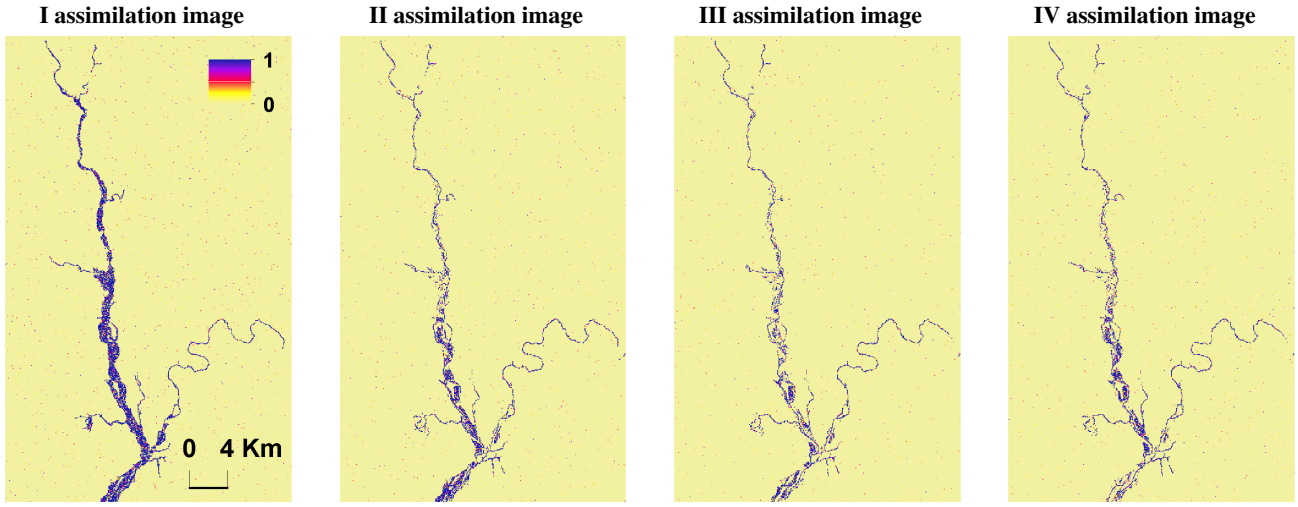

Figure 4. A detail of the synthetic probabilistic flood maps derived from synthetic SAR images. Probabilities to be flooded knowing the backscatter go from low value (yellow) to high values (blue).

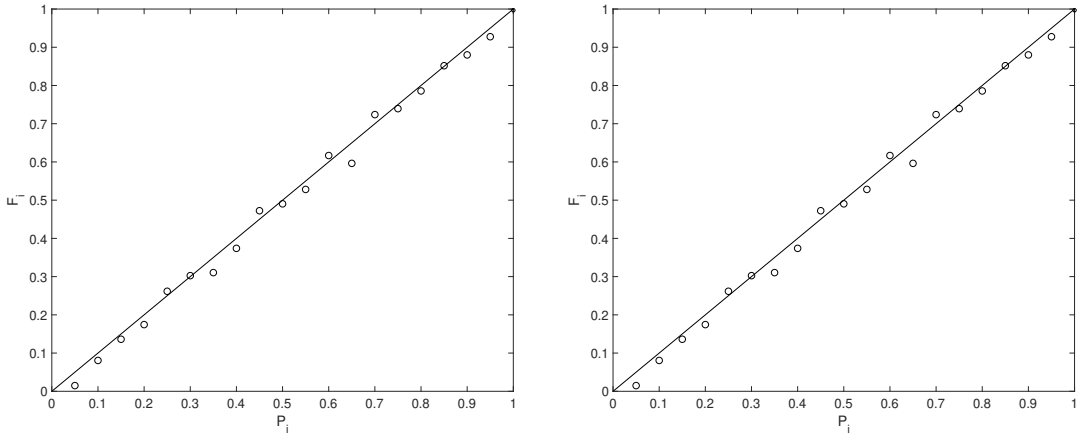

Figure 5. Example of the reliability plots for the verification of the synthetic probabilistic flood maps of the first two synthetic SAR images. On the x-axis probability range (P_i), on the y-axis the fraction of pixels within the probability range of the probabilistic flood map observed as flooded in the *true* binary flood map (F_i). The probabilistic flood maps are statistically reliable because the points align along the 1:1 line.

4.2 Evaluation of the flood extent map estimated at the assimilation time

~~CSI RMSE at the assimilation time~~

The CSI is computed over the entire hydraulic model domain at each assimilation time step.

The general trend of the assimilation effect is positive, as errors tend to decrease at all the assimilation steps with different assimilation methods. Even though the CSI is already high with the OL, the assimilation further improves the results and this

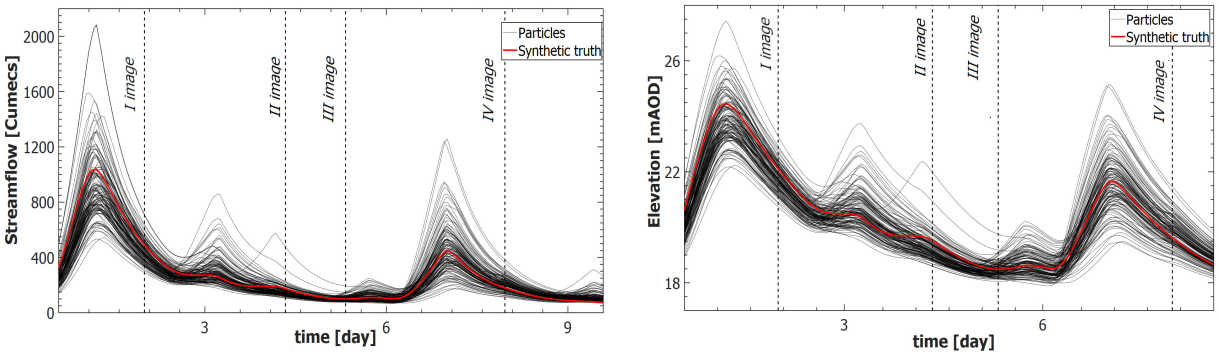

Figure 6. Streamflow time series (left) and water elevation time series (right) at the gauge station in Bewdley. Black lines represent the 128 particles while the red line corresponds to the synthetic truth.

Table 1. Critical success index values at each assimilation time step. The Open Loop (Figure 6) where no assimilation is computed is compared with the standard method and the adapted method with an increasing effective ensemble size (EES).

Assimilation times	**Open Loop**	**Assimilation**				
		standard	*5% EES*	*10% EES*	*20% EES*	*50% EES*
I image	0.9573	0.9887	0.9914	0.9866	0.9805	0.9779
II image	0.9202	0.9873	0.9800	0.9758	0.9658	0.9645
III image	0.9437	0.9921	0.9753	0.9690	0.9622	0.9636
IV image	0.7976	0.9881	0.9754	0.9638	0.9577	0.9610

becomes particularly clear at the last assimilation time step. From Table 1 it can be noticed that the CSI, approximately equal to 0.80 with the OL in the worst case (assimilation of the IV image), exceeds 0.96 for the different assimilation types and reaches the maximum value of 0.99 with the standard method. In Figure 7, we provide the contingency maps of the OL and of the 5% EES approach (results of the standard method are similar to those of the 5% EES approach and therefore not shown) . For each pair of images, we show on the left the results of the OL and on the right the results obtained after the assimilation. In this study, it can be observed that the OL has a tendency to over-detection; the number of red pixels is higher than the ~~orange~~ black ones and after the assimilation the number of over-detected pixels decreases confirming the results obtained with the CSI.

The confusion matrix given in Table 2 provides more details on the 4^{th} assimilation time step. On the one hand, the number of pixels wrongly predicted as flooded in the OL is 1196 and more than 90% of these are correctly classified as non flooded after the assimilation for both standard and 5% EES methods. On the other hand, a few pixels correctly predicted as flooded in the OL are classified not flooded after the assimilation. However, it can be argued that the number of 201 wrongly classified pixels after the assimilation is rather low compared to the 1253 pixels of the OL.

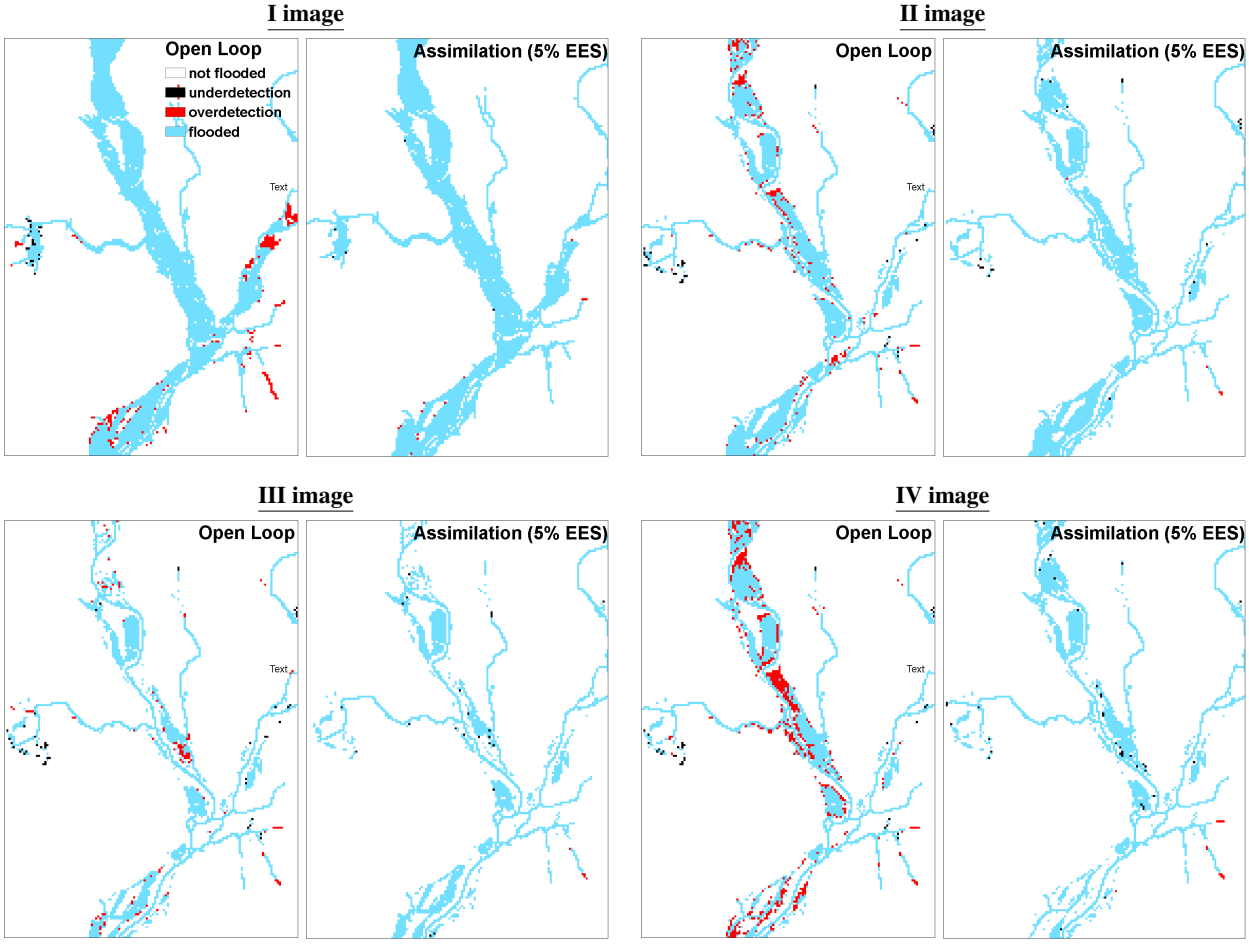

Figure 7. Contingency maps before (open loop) and after assimilation at 5 % EES at each time step. Two types of errors can be distinguished: overdetection (red pixels) when the model predicts the pixel as flooded but the pixel is observed as not-flooded and underdetection (black pixels) when the contrary occurs. In case model and observation agree pixels are correctly classified as not-flooded (white pixels) and flooded (blue pixels).

Table 2. Confusion matrix of the OL and of the 5% EES assimilation at the 4^{th} assimilation time step: OF=observed flooded pixels in the truth map, ON=observed non-flooded pixels in the truth map, PF= predicted flooded pixels, PN=predicted non-flooded pixels.

	Open Loop		Standard		Assimilation (5% EES)	
	PF	**PN**	**PF**	**PN**	**PF**	**PN**
OF	4826	57	4748	135	4815	68
ON	1196	264833	66	265963	41	265988

4.3 Evaluation of the flood map estimated in time

~~CSI RMSE in time~~

The flood is simulated using an hourly time step. Consequently, it is possible to evaluate the evolution of the performance metrics CSI (Figure 8). This figure shows that the OL's performance is consistently poor and the standard assimilation performs

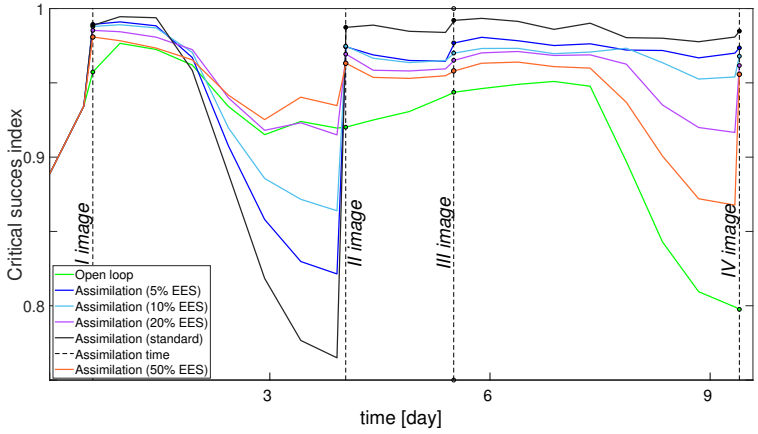

Figure 8. Time series of CSI of flood extent values for the different assimilation methods: open loop (green), standard assimilation (black), assimilations with 5% EES (blue), 10% EES (cyan), 20% EES (purple), 50% EES (orange).

best compared to the other assimilation runs at all the assimilation time steps. The assimilation runs with different EES values lie within these two extremes. It can be noted that the more particles are neglected, which is equivalent to say the lower is the EES, the higher is the performance at the assimilation time step. ~~The reason could be linked to the fact that results are biased since not the whole tempering scheme has been followed.~~

Moreover, markedly different CSI time series for the different assimilation experiments are shown in Figure 8 ~~shows~~ . 27 hours after the first assimilation, the performances of the standard and 5% EES methods, which perform better than the other methods, start decreasing. The lowest values are reached 54 hours after the assimilation. One explanation is that the weights assigned to the particles at the 1^{st} assimilation time are no longer valid when hydraulic conditions change and need to be recomputed. However, things change after the 2^{nd} assimilation, when the performances of the standard and the 5% EES assimilation methods remain stable until the end of the simulation time. The decrease of performances attributed to the standard and 5% assimilation methods after the 1^{st} time step is due to a drastic change in the flood extent. The total number of flooded pixels reduces from 8539 to 5494 because the flood started receding. The spread of the posterior PDF with the standard and 5% EES methods is small, meaning that only a few particles retain significant *importance* weight. Consequently, when the flood extent changes and particles evolve in time, it may happen that the uncertainty bounds of the posterior PDF do not encompass the *true* model state after several time steps. On the contrary, when more particles are considered (higher EES), more particles are used to draw the posterior PDF. This gives more chances to the ensemble to encompass the synthetic *truth* and increases the

overall robustness of the method. This becomes particularly relevant when the hydraulic boundary conditions change and no new observation is available.

4.4 Evaluation of the water levels in time over a global scale

The RMSE, reported in Table 3. decrease by factors larger than 2 and 3 with the standard assimilation and the 5% EES assimilation, respectively. After the 1^{st} assimilation, carried out close to the flood peak in Saxons Lode, the accuracy of the water level is improved by approximately 20 cm over the entire flood domain. The assimilation of the 2^{nd} and 4^{th} images has a negative effect in case the adapted method 50% EES of the assimilation particle filter is applied: the RMSE increases compared to the OL. As already shown in the Table 3 the standard assimilation and 5% EES predictions of water levels provide

Table 3. Root mean square error (RMSE [m]) at each assimilation time step. The Open Loop (figure 6) where no assimilation is computed is compared with the standard method and the adapted method with an increasing effective ensemble size (EES).

Assimilation times	**Open Loop**	**Assimilation**				
		standard	*5% EES*	*10% EES*	*20% EES*	*50% EES*
I image	0.2608	0.0742	0.0608	0.0785	0.1501	0.1762
II image	0.1246	0.0526	0.1046	0.1278	0.1553	0.1704
III image	0.1604	0.0645	0.1103	0.1665	0.2154	0.2270
IV image	0.1702	0.0541	0.0619	0.1084	0.1899	0.2205

more accurate results (figure 9). When moving away from the first assimilation, the RMSE of the best performing assimilation methods increases. For instance, after 54 hours the RMSE of the standard method is increased by 65% compared to the RMSE of the OL. In case different EES are considered, the RMSE values fluctuates significantly in between two assimilations and it becomes difficult to draw any general conclusions. As the number of *important* particles increases, water levels vary significantly, especially in the area close to the flood edge even though the flood extent does not change too much from a particle to another.

4.5 Evaluation of discharge and water level time series

~~Discharge and water level time series~~

The different assimilation runs are also compared considering the discharges and water levels at different gauge stations along the river Severn. In the right panels of Figures 10 and 11 the different assimilation experiments are compared against the synthetic truth (red line). In the left panels of Figures 10 and 11 the standard method and the 5% EES assimilation with the *important* particles and the synthetic truth are shown. The plotted *important* particles represent the 5% of the ensemble with the largest weight. All the 128 particles are equally weighted until the first observation is assimilated. After the first assimilation the number of important particles decreases. At the second assimilation time step, weights are recomputed and the new *important*

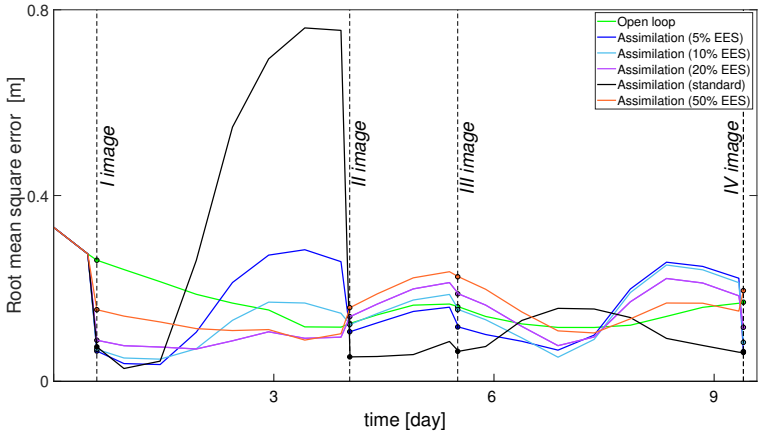

Figure 9. Time series of root mean square error (RMSE [m]) values for the different assimilation experiments: open loop (green), standard assimilation (black), assimilations with 5% EES (blue), 10% EES (cyan), 20% EES (purple), 50% EES (orange).

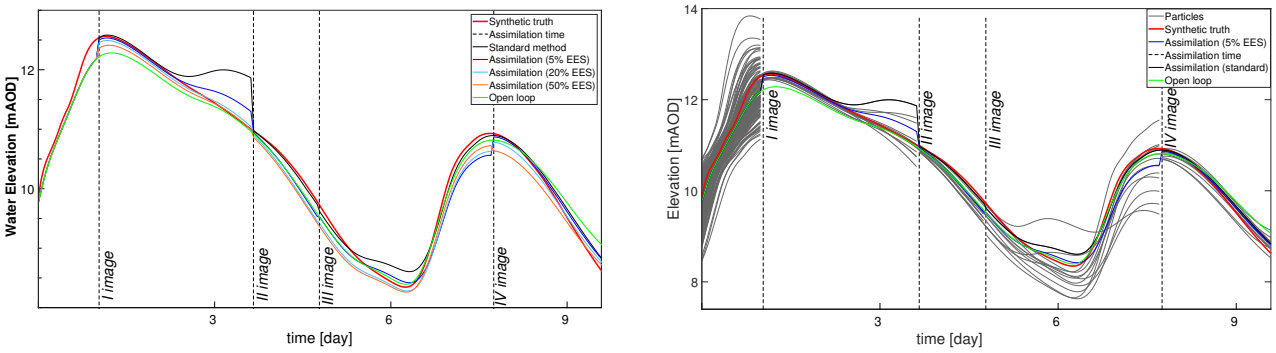

Figure 10. Water level time series at Saxons Lode. Left: assimilation runs with an EES of 5%(blue), 20% (cyan) and 50% (orange), OL (green), standard assimilation (black). Right: particles carrying significant weight after the assimilation at 5% EES (grey). Dashed lines correspond to the assimilation times.

particles are selected again and so on. The assimilation of the PFMs improves the predictions of water levels and streamflow at specific points of the river Severn, as in Bewdley and in Saxons Lode (Figure 10, 11), for the majority of the assimilation time steps in both underprediction and overprediction cases. The standard method and similarly the 5 % EES assimilation method are the most accurate in forecasting the values of water levels and streamflows. The improvements due to the assimilation persist for a long time: up to 27 hours after the first assimilation predictions are still close to the synthetic *truth*. The local results of water levels suggest that the inaccuracy of the global RMSE values in time is likely due to the evaluation over the entire flood domain.

4.6 Impact assessment of errors in SAR observations

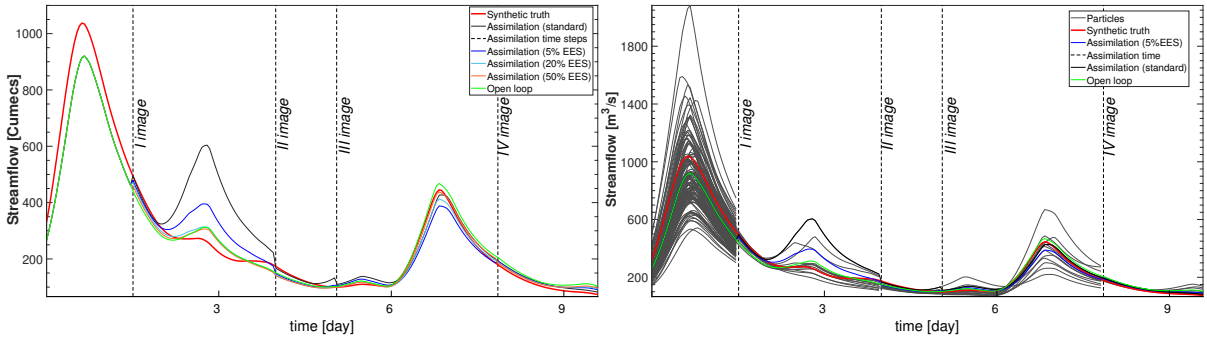

Figure 11. Streamflow time series at Bewdley. Left: assimilation runs with an EES of 5% (blue), 20% (cyan) and 50% (orange), OL (green), standard assimilation (black). Right: particles carrying significant weight after the assimilation at 5% EES (grey). Dashed lines correspond to the assimilation times.

~~**Errors in the observations**~~

In the previous section, speckle uncertainty in SAR observations is considered. However, in reality, SAR observations are also susceptible to errors due to the misclassification of wet/dry pixels caused by features on the ground as already mentioned ~~in the introduction~~. Therefore, ~~a bias~~ errors are added to the synthetic SAR observations as described in the methodology to investigate the impact on the DA assimilation framework. Figure 12 shows the RMSE and the CSI obtained at different assimilation time steps. The best performing assimilation methods (i.e. standard and 5% EES) with no ~~bias~~ error in the observations are compared with the ones where ~~bias~~ error is introduced. With the misclassification of 20% of the pixels, the assimilation still has beneficial effects: the CSI increases at each assimilation time step with respect to the OL. The RMSE values also tend to be satisfactory after each assimilation. With an increase in the error of 40% the performances of the DA framework start decreasing. The assimilation of the first image still has a positive effect on the predictions. In fact, CSI and RMSE are improved with respect to the OL even through the improvements are not as significant as in the previous cases. The explanation is arguably to be found in the high number of flooded pixels. It is large enough to counterbalance the misclassified pixels in the SAR image. Performances decrease with the assimilation of the remaining SAR observations when the number of flooded pixels is reduced by half.

5 Conclusions

Satellite images provide valuable information about flood extent that can complement or substitute in situ measurements. The fact that several space agencies provide free access to high resolution satellite Earth Observation data paves the way for improving Earth Observation-based flood forecasting and reanalyses worldwide. This study represents a follow-up of the previous real case study from Hostache et al. (2018) with the objective to further proceed in the evaluation of the proposed DA framework once the assumptions are effectively satisfied. This study has been set up in a controlled environment using a

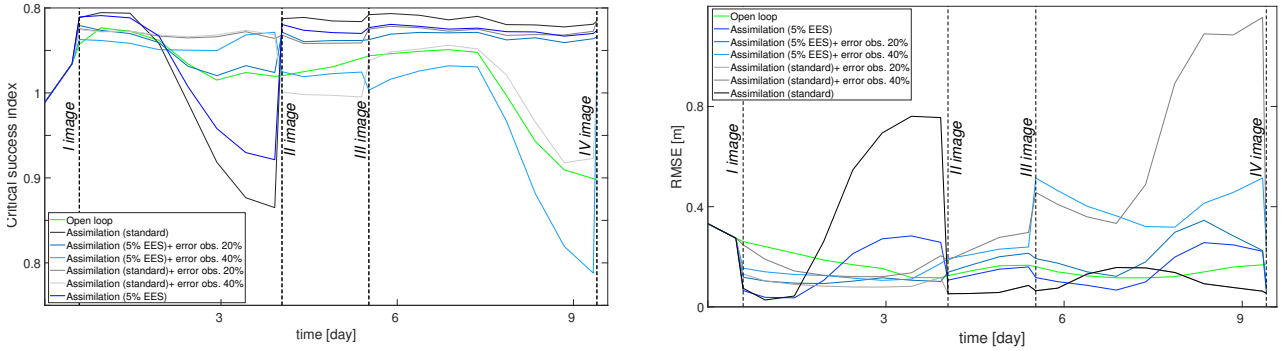

Figure 12. CSI values (on the left) and RMSE (on the rigth) after the standard assimilation of SAR observations free of errors (black), with 20% of errors (grey), with 40% of errors (ligth grey) and after the 5% EES assimilation free of errors (blue), with 20% of errors (ligth blue), with 40% of errors (cyan).

synthetically generated data-set in order to make sure that the rainfall and SAR observations are the only source of uncertainty. A common issue in Particle filters is degeneracy: the ensemble could collapse after the assimilation because higher probabilities are assigned to a limited number of particles. ~~Therefore,~~ The ~~enhanced~~ tempering coefficient ~~, based on the desired effective ensemble size after the assimilation, has been tested~~ can be used to reduce degeneracy because it inflates the posterior probability ~~to reduce degeneracy~~ and reduces the peak of the likelihood ~~and moves iteratively and smoothly from the prior to the posterior probability density function~~. In this study, we have evaluated the effect of variations of the α tempering coefficient on the DA performance. Different PFs are compared with the OL and the synthetic truth: the SIS (with only a few particles from the ensemble potentially carrying non- negligible weights) and the adapted method with 5-10-20-50% EES (with the number of particles with non-negligible weights increasing with the EES). This methodology leads to slightly biased estimates because the observation is down-weighted. In addition, we investigated the impact of ~~bias~~ error in the observations (i.e. errors in the SAR derived probabilistic flood maps due to dry water-look alike pixels or emerging objects) on the assimilation. Indeed, the main issue of using SAR observations in flood forecasting models is the difficulty of detecting flooded area for specific cases (e.g. urban or vegetated areas). At first, following the study from Hostache et al. (2018) only speckle uncertainty of the SAR image is taken into account in the pfms. In a second step, a ~~bias~~ error to reproduce misclassified pixels is introduced in the synthetic SAR observations.

The following key conclusions can be drawn from our experiments:

1. The best performing method is the standard method (i.e. SIS). Importance weights are assigned to a limited number of particles that better agree with the observations. At the time of the assimilation, results tend to be very accurate: the forecasts move close to the synthetic truth. The main weakness of the standard filter is to significantly suffer from degeneracy.

2. The 5% effective ensemble size assimilation (meaning that only the 5% of the ensemble will have a not-negligible weight after the assimilation) is slightly less accurate at the time of the assimilation but it has the advantage of reducing

the degeneracy problem. ~~Even though larger effective ensemble size prevents degeneracy results are less accurate and performances of the predictions are degraded.~~

3. ~~The persistence in time of the improvements depends on the rapidity with which hydrologic conditions change. A frequent image acquisition could help keep model predictions on track especially when the dynamic of the system is varying fast and the persistence in time of the assimilation is not long enough.~~

4. Our study further shows that it is important to characterize and mask out errors in the SAR observations. A large number of misclassified pixels substantially degrades the DA performance. In our case study, results suggest that an improvement of model simulations (i.e. water level and streamflow) in terms of CSI and RMSE performance metrics is achieved as long as errors in the observations are rather limited, i.e. when no more than 20% of the pixels are affected. However, if the misclassification goes beyond 40% of affected pixels, the assimilation has no effect and may even lead to a degradation of the model predictions. ~~Our study further shows that it is important to characterize and mask out errors in the SAR observations. An increasing number of wrongly classified pixels leads to a substantial reduction of the performances of the DA framework. In our case study, the improvement of model simulation (water levels and streamflow) and of performances (CSI and RMSE) with the assimilation is only possible when the errors in the observation is not larger than 20% of pixels in the SAR image.~~

6 Discussions

The results of our study confirm the effectiveness of the proposed DA framework when the hypothesis of the rainfall as the main source of uncertainty is verified. ~~from which it could be inferred that the wrong results, in the previous real case study by~~ Hostache et al. (2018) ~~, at some assimilation time steps may be eventually explained by additional sources of uncertainties not taken into account~~. Consequently, for those cases where rainfall represents the main source of uncertainty, more obviously but not only in poorly and un-gauged catchments and when using medium-range forecasting models, our study results indicate that the application of the approach described in the manuscript may lead to improved results of the model simulations. For those cases where the uncertainty of other sources becomes more relevant and may be even dominant, it is clear that such sources need to be taken into account explicitly. However, the required adaptations of the proposed DA framework still need to be developed. In this context it is also worth mentioning that the limitations identified in the previously published real case study by Hostache et al. (2018) were explained by additional sources of uncertainties not taken into account.

Using probabilistic flood maps or backscatter values increases the number of observations to be assimilated when compared to a method that only derives the flood edge from satellite observations as reported in Cooper et al. (2018a). Moreover, the nearly-direct use of the SAR information enables a faster end-to-end processing from the acquisition of the image to the assimilation of the SAR data into the model which is beneficial for an operational usage.

In our experiments, the improvements of model forecasts of water level and streamflow are significant at the assimilation time step ~~but they start decreasing after some hours~~ and the improvements persist over subsequent time steps (for example up to

27 hours after the first assimilation the model results outperform ~~worse compared to the open loop~~ the open loop simulation). ~~deviating from the synthetic truth~~.The persistence of these improvements depends on the flashiness of the flood event (i.e., the rapidity with which hydrologic conditions change). More frequent image acquisitions could help keep model predictions on track, especially when the system is highly dynamic. The update of a state variable of the forecasting model could as well increase the persistence of the improvements. In our study none of the model state variables is updated as only the particle weights are computed, based on the SAR observations and on the simulated flood extent maps and used to calculate the expectation of water levels and streamflow. In previous studies [Andreadis et al. (2007), Matgen et al. (2010), Cooper et al. (2018b)], inflow updating was identified as a condition leading to more persistent improvements. For instance, one of the conclusions from the study by Matgen et al. (2010) was that updating the fluxes at the upstream boundary conditions, rather than the water levels, is more effective because of the high uncertainty of the inflow due to the poorly known rainfall distribution over the catchment. Therefore, as a future perspective, we aim to update hydrologic model states because it might have a positive impact on the long-term runoff simulations and consequently on the persistence of DA benefits.

Some modifications of the DA framework are still required to fully overcome the issue of degeneracy. Although the use of a smaller tempering coefficient leads to a larger effective ensemble size (e.g. 50 %) and helps avoid degeneracy, the results are less accurate compared to the standard method or the adapted method with 5% EES. As described in Neal (1996) and in van Leeuwen et al. (2019), the tempering procedure consists of several steps, but in this study the tempering coefficient is applied only to flatten the likelihood, therefore down weighting the observations. This most likely explains why the data assimilation performs better when the effective ensemble size (the number of particles not negligible after the assimilation) is smaller. ~~This is maybe due to the way the tempering coefficient has been used , without applying the full tempering scheme, leading to biased results because it tends to down-weight the observations by increasing their errors. As described in and in , the tempering procedure consists of several steps, but in this study the tempering coefficient is applied only to flatten the likelihood, which maybe explains why data assimilation performs better when the effective ensemble size (the number of particles not negligible after the assimilation) is small~~. As already mentioned, the present study has the aim of assessing and validating the method proposed by Hostache et al. (2018) in a synthetic environment. Our DA framework can be applied to a variety of flood inundation forecasting chains. In fact, the forecast updating is carried out via a sequential importance sampling only (i.e. importance weights). Only the particle weights are updated based on the observations and used to compute the expectation (i.e. weighted mean) of the augmented state vector including hydraulic state variables of water depth, plus flood extent and boundary conditions. In this study the hydrologic and hydraulic models are loosely coupled with a one-way transfer of information as in many other studies [e.g., Peckham et al. (2013),Hoch et al. (2017),Rajib et al. (2020)]. The weights define the relative importance of the particles and thus of the inherent streamflow and stage along the entire river. We acknowledge that the observed flood extent is more closely linked to the past boundary conditions rather than the boundary conditions corresponding to the assimilation time steps. In spite of this limitation we argue that in this synthetic experiment, the particles that performed best in the past are also those that reach the highest performance level at the time of the assimilation. This is illustrated in the Figures 10 and 11 where the use of updated weights is shown to enable the correction of the state variables of the hydraulic model both upstream and downstream. However, we recognize that further improvements could be developed to address issues such as spurious relations

that may occur between SAR observations and model variables due to a rather small ensemble size. Enlarging the ensemble size could be necessary if this occurs.

We also argue that the method used in the manuscript has the potential to support EO-based modelling at large scale. This potential is particularly high in large, natural floodplains where flood inundation remains present over long time periods. In spite of the increased frequency of satellite observations, the persistence of a flood over many days increases the chance of its detection and mapping by satellite sensors. Another condition that needs to be satisfied is that there should be an unambiguous relationship between the flood extent observed by the spaceborne sensors and river discharge. This also means that areas where backscatter variations are not impacted by the appearance of floodwater (e.g. densely vegetated floodplains) should be rather small. Indeed, these constraints must be satisfied to enable a successful application of the proposed framework and to take advantage of the analysis carried out in this manuscript. As a conclusion, based on the above elements, we argue that our approach is valid regardless of the type of model coupling that is performed and is thus applicable to many different forecasting systems. However, more research is needed to fully understand the role of floodplain and water basin characteristics and SAR data properties on the DA performance. In a future study it is envisaged that to avoid degeneracy and keep a larger effective ensemble size, the full tempering scheme will be applied. Possible ways to adapt and advance the proposed DA framework are currently under development (e.g. updating a state variable of the model, using an enhanced version of the adapted filter).

Acknowledgements. The research reported herein was funded by the National Research fund of Luxembourg through the HyDRO-CSI projects. Funding from the Austrian Science Funds as part of the Vienna Doctoral Programme on Water Resources System (DK W1219-N22) is acknowledged. Peter Jan van Leeuwen thanks the European Research Council (ERC) for funding of the CUNDA ERC 694509 project under the European Unions Horizon 2020 research and innovation programme. Nancy Nichols was funded in part by the UK Natural Environmental Research Council (NERC) National Centre for Earth Observation (NCEO).

The Lisflood-FP model can be freely downloaded at http://www.bristol.ac.uk/geography/research/hydrology/models/lisflood. The river cross-section data, the digital elevation model, and the gauging station water level, streamflow, and rating curve data are freely available upon request from the Environment Agency (enquiries@environmentagency.gov.uk). The ERA-5 data set is freely available at https://confluence.ecmwf.int/display/CKB/ERA5.

References

Andreadis, K. M., Clark, E. A., Lettenmaier, D. P., and Alsdorf, D. E.: Prospects for river discharge and depth estimation through assimilation of swath-altimetry into a raster-based hydrodynamics model, Geophysical Research Letters, 34, https://doi.org/https://doi.org/10.1029/2007GL029721, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2007GL029721, 2007.

Arulampalam, M. S., Maskell, S., Gordon, N., and Clapp, T.: A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking, IEEE Transactions on Signal Processing, 50, 174–188, https://doi.org/10.1109/78.978374, 2002.

Bates, P. and Roo, A. D.: A simple raster-based model for flood inundation simulation, Journal of Hydrology, 236, 54 – 77, https://doi.org/https://doi.org/10.1016/S0022-1694(00)00278-X, http://www.sciencedirect.com/science/article/pii/S002216940000278X, 2000.

Bates, P. D., Horritt, M. S., and Fewtrell, T. J.: A simple inertial formulation of the shallow water equations for efficient two-dimensional flood inundation modelling, Journal of Hydrology, 387, 33–45, https://doi.org/https://doi.org/10.1016/j.jhydrol.2010.03.027, https://www.sciencedirect.com/science/article/pii/S0022169410001538, 2010.

Blöschl, G., Hall, J., Viglione, A., and Perdigão, R.: Changing climate both increases and decreases European river floods, Nature, 573, 108–111, https://doi.org/10.1038/s41586-019-1495-6, 2019.

Chini, M., Hostache, R., Giustarini, L., and Matgen, P.: A Hierarchical Split-Based Approach for Parametric Thresholding of SAR Images: Flood Inundation as a Test Case, IEEE Transactions on Geoscience and Remote Sensing, 55, 6975–6988, https://doi.org/10.1109/TGRS.2017.2737664, 2017.

Cooper, E., Dance, S., Garcia-Pintado, J., Nichols, N., and Smith, P.: Observation impact, domain length and parameter estimation in data assimilation for flood forecasting, Environmental Modelling Software, 104, 199–214, https://doi.org/https://doi.org/10.1016/j.envsoft.2018.03.013, https://www.sciencedirect.com/science/article/pii/S1364815217303602, 2018a.

Cooper, E. S., Dance, S. L., García-Pintado, J., Nichols, N. K., and Smith, P.: Observation operators for assimilation of satellite observations in fluvial inundation forecasting, Hydrology and Earth System Sciences Discussions, 2018, 1–32, https://doi.org/10.5194/hess-2018-589, https://www.hydrol-earth-syst-sci-discuss.net/hess-2018-589/, 2018b.

de Almeida, G. A. M. and Bates, P.: Applicability of the local inertial approximation of the shallow water equations to flood modeling, Water Resources Research, 49, 4833–4844, https://doi.org/https://doi.org/10.1002/wrcr.20366, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/wrcr.20366, 2013.

De Lannoy, G. J. M., Houser, P. R., Pauwels, V. R. N., and Verhoest, N. E. C.: Assessment of model uncertainty for soil moisture through ensemble verification, Journal of Geophysical Research: Atmospheres, 111, https://doi.org/10.1029/2005JD006367, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2005JD006367, 2006.

Environment Agency, E. A.: River Severn Catchment Flood Management Plan., https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/289103/River_Severn_Catchment_Management_Plan.pdf, 2009.

Fenicia, F., Kavetski, D., and Savenije, H. H. G.: Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development, Water Resources Research, 47, https://doi.org/10.1029/2010WR010174, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2010WR010174, 2011.

García-Pintado, J., Mason, D. C., Dance, S. L., Cloke, H. L., Neal, J. C., Freer, J., and Bates, P. D.: Satellite-supported flood forecasting in river networks: A real case study, Journal of Hydrology, 523, 706 – 724, https://doi.org/https://doi.org/10.1016/j.jhydrol.2015.01.084, http://www.sciencedirect.com/science/article/pii/S0022169415001031, 2015.

Giustarini, L., Matgen, P., Hostache, R., Montanari, M., Plaza, D., Pauwels, V. R. N., De Lannoy, G. J. M., De Keyser, R., Pfister, L., Hoffmann, L., and Savenije, H. H. G.: Assimilating SAR-derived water level data into a hydraulic model: a case study, Hydrology and Earth System Sciences, 15, 2349–2365, https://doi.org/10.5194/hess-15-2349-2011, https://www.hydrol-earth-syst-sci.net/15/2349/2011/, 2011.

Giustarini, L., Vernieuwe, H., Verwaeren, J., Chini, M., Hostache, R., Matgen, P., Verhoest, N., and Baets, B. D.: Accounting for image uncertainty in SAR-based flood mapping, International Journal of Applied Earth Observation and Geoinformation, 34, 70 – 77, https://doi.org/https://doi.org/10.1016/j.jag.2014.06.017, http://www.sciencedirect.com/science/article/pii/S0303243414001512, 2015.

Giustarini, L., Hostache, R., Kavetski, D., Chini, M., Corato, G., Schlaffer, S., and Matgen, P.: Probabilistic Flood Mapping Using Synthetic Aperture Radar Data, IEEE Transactions on Geoscience and Remote Sensing, 54, 6958–6969, https://doi.org/10.1109/TGRS.2016.2592951, 2016.

Grimaldi, S., Li, Y., Pauwels, V. R. N., and Walker, J. P.: Remote Sensing-Derived Water Extent and Level to Constrain Hydraulic Flood Forecasting Models: Opportunities and Challenges, Surveys in Geophysics, 37, 977–1034, https://doi.org/10.1007/s10712-016-9378-y, https://doi.org/10.1007/s10712-016-9378-y, 2016.

Hersbach, H., Bell, W., Berrisford, P., Horányi, A., J., M.-S., Nicolas, J., Radu, R., Schepers, D., Simmons, A., Soci, C., and Dee, D.: Global reanalysis: goodbye ERA-Interim, hello ERA5, pp. 17–24, https://doi.org/10.21957/vf291hehd7, https://www.ecmwf.int/node/19027, 2019.

Hoch, J. M., Neal, J. C., Baart, F., van Beek, R., Winsemius, H. C., Bates, P. D., and Bierkens, M. F. P.: GLOFRIM v1.0 – A globally applicable computational framework for integrated hydrological–hydrodynamic modelling, Geoscientific Model Development, 10, 3913–3929, https://doi.org/10.5194/gmd-10-3913-2017, https://gmd.copernicus.org/articles/10/3913/2017/, 2017.

Hostache, R., Lai, X., Monnier, J., and Puech, C.: Assimilation of spatially distributed water levels into a shallow-water flood model. Part II: Use of a remote sensing image of Mosel River, Journal of Hydrology, 390, 257 – 268, https://doi.org/https://doi.org/10.1016/j.jhydrol.2010.07.003, http://www.sciencedirect.com/science/article/pii/S0022169410004166, 2010.

Hostache, R., Chini, M., Giustarini, L., Neal, J., Kavetski, D., Wood, M., Corato, G., Pelich, R.-M., and Matgen, P.: Near-Real-Time Assimilation of SAR-Derived Flood Maps for Improving Flood Forecasts, Water Resources Research, 54, 5516–5535, https://doi.org/10.1029/2017WR022205, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2017WR022205, 2018.

Koussis, A. D., Lagouvardos, K., Mazi, K., Kotroni, V., Sitzmann, D., Lang, J., Zaiss, H., Buzzi, A., and Malguzzi, P.: Flood Forecasts for Urban Basin with Integrated Hydro-Meteorological Model., Journal of Hydrologic Engineering, 8, 1, http://proxy.bnl.lu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=iih&AN=8687464&site=ehost-live&scope=site, 2003.

Lai, X., Liang, Q., Yesou, H., and Daillet, S.: Variational assimilation of remotely sensed flood extents using a 2-D flood model, Hydrology and Earth System Sciences, 18, 4325–4339, https://doi.org/10.5194/hess-18-4325-2014, https://www.hydrol-earth-syst-sci.net/18/4325/2014/, 2014.

Matgen, P., Montanari, M., Hostache, R., Pfister, L., Hoffmann, L., Plaza, D., Pauwels, V. R. N., De Lannoy, G. J. M., De Keyser, R., and Savenije, H. H. G.: Towards the sequential assimilation of SAR-derived water stages into hydraulic models using the Parti-

cle Filter: proof of concept, Hydrology and Earth System Sciences, 14, 1773–1785, https://doi.org/10.5194/hess-14-1773-2010, https://www.hydrol-earth-syst-sci.net/14/1773/2010/, 2010.

Moradkhani, H., Hsu, K.-L., Gupta, H., and Sorooshian, S.: Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the particle filter, Water Resources Research, 41, https://doi.org/10.1029/2004WR003604, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2004WR003604, 2005.

Neal, J., Schumann, G., and Bates, P.: A subgrid channel model for simulating river hydraulics and floodplain inundation over large and data sparse areas, Water Resources Research, 48, https://doi.org/10.1029/2012WR012514, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2012WR012514, 2012.

Neal, R. M.: Sampling from multimodal distributions using tempered transitions, Statistics and Computing, https://doi.org/https://doi.org/10.1007/BF00143556, 1996.

Pappenberger, F., Beven, K. J., Hunter, N. M., Bates, P. D., Gouweleeuw, B. T., Thielen, J., and de Roo, A. P. J.: Cascading model uncertainty from medium range weather forecasts (10 days) through a rainfall-runoff model to flood inundation predictions within the European Flood Forecasting System (EFFS), Hydrology and Earth System Sciences, 9, 381–393, https://doi.org/10.5194/hess-9-381-2005, https://www.hydrol-earth-syst-sci.net/9/381/2005/, 2005.

Peckham, S. D., Hutton, E. W., and Norris, B.: A component-based approach to integrated modeling in the geosciences: The design of CSDMS, Computers Geosciences, 53, 3–12, https://doi.org/https://doi.org/10.1016/j.cageo.2012.04.002, https://www.sciencedirect.com/science/article/pii/S0098300412001252, modeling for Environmental Change, 2013.

Rajib, A., Liu, Z., Merwade, V., Tavakoly, A. A., and Follum, M. L.: Towards a large-scale locally relevant flood inundation modeling framework using SWAT and LISFLOOD-FP, Journal of Hydrology, 581, 124 406, https://doi.org/https://doi.org/10.1016/j.jhydrol.2019.124406, https://www.sciencedirect.com/science/article/pii/S0022169419311412, 2020.

Revilla-Romero, B., Wanders, N., Burek, P., Salamon, P., and de Roo, A.: Integrating remotely sensed surface water extent into continental scale hydrology, Journal of Hydrology, 543, 659 – 670, https://doi.org/https://doi.org/10.1016/j.jhydrol.2016.10.041, http://www.sciencedirect.com/science/article/pii/S0022169416306862, 2016.

UNISDR: United Nations Office for Disaster Risk Reduction. Making Development Sustainable: The Future of Disaster Risk Management. Global Assessment Report on Disaster Risk Reduction, www.unisdr.org/we/inform/publications/42809, 2015.

van Leeuwen, P. J.: Nonlinear data assimilation in geosciences: an extremely efficient particle filter, Quarterly Journal of the Royal Meteorological Society, 136, 1991–1999, https://doi.org/10.1002/qj.699, https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.699, 2010.

van Leeuwen, P. J., Künsch, H. R., Nerger, L., Potthast, R., and Reich, S.: Particle filters for high-dimensional geoscience applications: A review, Quarterly Journal of the Royal Meteorological Society, 145, 2335–2365, https://doi.org/10.1002/qj.3551, https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3551, 2019.

Van Wesemael, A.: Assessing the value of remote sensing and in situ data for flood inundation forecasts, Ph.D. thesis, Ghent University, 2019.

Wood, M., Hostache, R., Neal, J., Wagener, T., Giustarini, L., Chini, M., Corato, G., Matgen, P., and Bates, P.: Calibration of channel depth and friction parameters in the LISFLOOD-FP hydraulic model using medium-resolution SAR data and identifiability techniques, Hydrology and Earth System Sciences, 20, 4983–4997, https://doi.org/10.5194/hess-20-4983-2016, https://www.hydrol-earth-syst-sci.net/20/4983/2016/, 2016.

Zhu, M., van Leeuwen, P. J., and Amezcua, J.: Implicit equal-weights particle filter, Quarterly Journal of the Royal Meteorological Society, 142, 1904–1919, https://doi.org/10.1002/qj.2784, https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.2784, 2016.