AGU100 ADVANCING EARTH AND SPACE SCIENCE

Water Resources Research

RESEARCH ARTICLE
10.1029/2017WR022205

Near-Real-Time Assimilation of SAR-Derived Flood Maps for Improving Flood Forecasts

Renaud Hostache[1], Marco Chini[1], Laura Giustarini[1], Jeffrey Neal[2], Dmitri Kavetski[3], Melissa Wood[1,2,4], Giovanni Corato[1], Ramona-Maria Pelich[1], and Patrick Matgen[1]

[1]Department of Environmental Research and Innovation, Luxembourg Institute of Science and Technology, Esch sur Alzette, Luxembourg, [2]School of Geographical Sciences, University of Bristol, Bristol, UK, [3]School of Civil, Environmental and Mining Engineering, University of Adelaide, Adelaide, South Australia, Australia, [4]Faculty of Geosciences, Utrecht University, Utrecht, Netherlands

Key Points:
- Probabilistic flood maps are derived from SAR images
- Probabilistic flood maps are assimilated into a flood forecasting model cascade
- Water level forecast quality improves substantially in the assimilation time steps, and benefits persist for hours to days

Correspondence to:
R. Hostache,
renaud.hostache@list.lu

Citation:
Hostache, R., Chini, M., Giustarini, L., Neal, J., Kavetski, D., Wood, M., et al. (2018). 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

Received 9 NOV 2017
Accepted 27 JUN 2018
Accepted article online 2 JUL 2018
Published online 17 AUG 2018

Abstract Short- to medium-range flood forecasts are central to predicting and mitigating the impact of flooding across the world. However, producing reliable forecasts and reducing forecast uncertainties remains challenging, especially in poorly gauged river basins. The growing availability of synthetic aperture radar (SAR)-derived flood image databases (e.g., generated from SAR sensors such as Envisat advanced synthetic aperture radar) provides opportunities to improve flood forecast quality. This study contributes to the development of more accurate global and near real-time remote sensing-based flood forecasting services to support flood management. We take advantage of recent algorithms for efficient and automatic delineation of flood extent using SAR images and demonstrate that near real-time sequential assimilation of SAR-derived flood extents can substantially improve flood forecasts. A case study based on four flood events of the River Severn (United Kingdom) is presented. The forecasting system comprises the SUPERFLEX hydrological model and the Lisflood-FP hydraulic model. SAR images are assimilated using a particle filter. To quantify observation uncertainty as part of data assimilation, we use an image processing approach that assigns each pixel a *probability of being flooded* based on its backscatter values. Empirical results show that the sequential assimilation of SAR-derived flood extent maps leads to a substantial improvement in water level forecasts. Forecast errors are reduced by as much as 50% at the assimilation time step, and improvements persist over subsequent time steps for 24 to 48 hr. The proposed approach holds promise for improving flood forecasts at locations where observed data availability is limited but satellite coverage exists.

1. Introduction

Short- to medium-range flood forecasts are central to mitigating the impact of flooding, allowing stakeholders in flood management to better anticipate critical needs. Flood forecasts can help meet many of the foreseeable needs identified by the World Health Organization, including coordinating search and rescue, providing medical assistance, managing displacement of people and goods, and avoiding the interruption of critical supply chains that are known to cause extreme financial losses. Hydrological (rainfall-runoff) and hydraulic models are powerful tools in flood forecasting systems (Revilla-Romero et al., 2016). Yet, optimal decision making in emergency situations is hampered by forecast uncertainties (Hostache et al., 2011; Pappenberger et al., 2007). To mitigate uncertainties, the models comprising the forecasting chain are often calibrated using in situ streamflow and/or water level measurements. A well-established method of reducing forecast uncertainties and keeping predictions *on track* is to periodically adjust the forecast models, for example, by assimilating external observations (Neal et al., 2007). Ideally, this is achieved using data from distributed streamflow gauging stations, yet such data are often not available. Even where gauged data are available, it is seldom publicly accessible in near-real-time (Revilla-Romero et al., 2016); moreover, the number of gauging stations worldwide is in decline, and existing coverage is sparse (Biancamaria et al., 2011; Mason et al., 2012). A promising approach for improving flood forecasts that has gained recent attention is to assimilate hydrology-related data derived from globally and freely available Earth observation data sets. Flood extent maps derived from satellite images can then be either integrated with models on their own (Wood et al., 2016; e.g., via calibration or data assimilation) or intersected with a digital elevation model (DEM) to produce, after a number of postprocessing steps, shoreline water levels for integration with models (Mason et al., 2012).

Among available sensors mounted on spaceborne platforms, synthetic aperture radars (SARs) are arguably the most relevant for estimating flooded areas. SAR sensors are capable of imaging the Earth's surface day and night and under almost all weather conditions. This capability offers a major advantage over optical sensors, which are hindered by cloud cover and can only acquire synoptic images of the Earth's surface during the day. Due to the specular reflection of the incoming radar signal, detection of (smooth, open) water bodies in SAR images is relatively straightforward (Giustarini et al., 2013; Pulvirenti et al., 2016), although layover, vegetated, and shadow areas can mask water or induce false alarm in derived maps. Many studies on data assimilation into hydraulic models or forecasting systems integrate synthetic, in situ or remote sensing-derived observations of water levels. For example, see Table 1 in Revilla-Romero et al. (2016), and also Neal et al. (2007), Matgen et al. (2010), Hostache et al. (2010), Giustarini et al. (2011), Yoon et al. (2012), Andreadis and Schumann (2014), García-Pintado et al. (2015), Hostache et al. (2015), and Xu et al. (2017). Indeed, water level is a diagnostic variable of any hydraulic model and hence is more straightforward to assimilate than flood extent (Lai et al., 2014), which is a prognostic variable (diagnostic variables are defined as variables that are required to solve the model, that is, state variables, whereas prognostic variables are derived quantities). However, working with water levels from SAR images has several important limitations: (i) they offer information only at the flood extent shoreline; (ii) the estimation methods are neither straightforward nor automatic, and they typically require high-resolution topographic data (e.g., Lidar DEM) that are not globally available, as well as hydraulic expertise for interpretation (Hostache et al., 2009, 2010; Mason et al., 2012; Wood et al., 2016). Although the accuracy of globally available DEMs has improved appreciably (e.g., Yamazaki et al., 2017), the above-mentioned limitations arguably will continue hindering the use of SAR-derived water levels for automatic, near-real-time applications.

Methods that directly assimilate flood extent (rather than water levels) maps into flood forecasting chains are hence of interest. The main advantage is that flood extent mapping from SAR images can be implemented using fast automatic algorithms that provide information over the entire area of interest (Chini et al., 2017; Giustarini et al., 2013). Promising results have already been obtained by assimilating flood extents derived from satellite images into a hydraulic model (Lai et al., 2014) and into a forecasting system (Revilla-Romero et al., 2016). The study by Lai et al. (2014) was based on 4DVAR (variational) data assimilation, with no objective toward real-time forecasting; they showed that the assimilation of a flood extent map derived from MODIS data (250-m spatial resolution) allows for the optimization of a lumped friction parameter. Revilla-Romero et al. (2016) used the ensemble Kalman filter to assimilate low resolution ($0.1° \times 0.1°$) satellite-derived flood extents based on the Global Flood Detection System (http://www.gdacs.org/flooddetection/) into a global forecasting system composed of a hydrological model and a routing function, with an objective toward real-time forecasting; their study over 101 stations in Africa and South America shows that flood extent assimilation improves simulated streamflow at the majority of stream gauges, especially at the gauges with poorest skill scores on open-loop runs. One limitation of the study by Revilla-Romero et al. (2016) is the relatively coarse spatial resolution of the model results (i.e., 0.1°), which may not meet operational needs in a crisis management context.

This study develops an efficient framework for the assimilation of high-resolution flood extent information derived from SAR images, for the purposes of improving near real-time flood forecasts. An approach based on particle filtering (PF) (e.g., Giustarini et al., 2011; Hostache et al., 2015; Matgen et al., 2010) is developed to optimally combine observations and simulations, within a flood forecasting chain that couples a conceptual hydrological model with a 2-D hydraulic model.

The paper is organized as follows. Section 2 describes the algorithm for deriving probabilistic flood maps from SAR images and the flood forecasting model cascade. Next, section 3 presents a new approach for assimilating SAR-derived flood maps into the forecasting chain. Section 4 describes the study area and available data, as well as the case study experiments. Section 5 reports, evaluates, and discusses the results obtained using the proposed methods. Conclusions are given in section 6.

2. Flood Mapping and Flood Forecasting System

2.1. Probabilistic Flood Mapping From SAR Image

When attempting to assimilate flood extent maps into a model, it is necessary to estimate the uncertainty associated with the flood extent observations (Giustarini et al., 2015). These uncertainty estimates

are obtained using the method of Giustarini et al. (2016), which is a further development of previous studies (Giustarini et al., 2013; Hostache et al., 2012; Matgen et al., 2011).

The probability of a pixel corresponding to open water given its backscatter value in the SAR image is estimated using Bayes' theorem as follows:

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

In equation (1), $p(\sigma^0)$ is the marginal probability of recording the backscatter σ^0 for any pixel, $p(\sigma^0|w)$ is the conditional probability of recording the backscatter σ^0 if the pixel is water, $p(\sigma^0|nw)$ is the conditional probability of recording the backscatter σ^0 if the pixel is nonwater, $p(w)$ and $p(\mathrm{nw})$ are the prior probabilities of a pixel being water and nonwater, respectively. As no a priori information is available, we set $p(w) = p(\mathrm{nw}) = 0.5$ as suggested in Giustarini et al. (2016). The unknown quantities $p(\sigma^0|w)$ and $p(\sigma^0|nw)$ are estimated from the empirical distribution of backscatter values derived from the SAR image. The image histogram is assumed to represent a mixture of two distributions: (i) a distribution of backscatter values from water pixels that is approximated by a Gaussian probability density function (pdf); and (ii) a distribution of backscatter values from nonwater pixels, estimated as the difference between the total histogram and the estimated Gaussian distribution (Giustarini et al., 2016). The Hierarchical Split Based Approach (Chini et al., 2017) is used to estimate the parameters of the Gaussian pdf describing the distribution of water backscatter even when water pixels cover only a small fraction of the SAR image. This approach iteratively splits the image into tiles of decreasing size and identifies those that exhibit a bimodal histogram, allowing an automatic calibration of the assumed distribution of open water backscatter. These steps generate a probabilistic flood map where each pixel is assigned a probability $p \in (0, 1)$ of being water (i.e., flooded). The unknown quantities $p(\sigma^0|w)$ and $p(\sigma^0|nw)$ are estimated from the empirical distribution of backscatter values derived from the SAR image. The image histogram is assumed to represent a mixture of two distributions: (i) a distribution of backscatter values from water pixels that is approximated by a Gaussian pdf; and (ii) a distribution of backscatter values from nonwater pixels, estimated as the difference between the total histogram and the estimated Gaussian distribution (Giustarini et al., 2016). The Hierarchical Split Based Approach (Chini et al., 2017) is used to estimate the parameters of the Gaussian pdf describing the distribution of water backscatter even when water pixels cover only a small fraction of the SAR image. This approach iteratively splits the image into tiles of decreasing size and identifies those that exhibit a bimodal histogram, allowing an automatic calibration of the assumed distribution of open water backscatter. These steps generate a probabilistic flood map where each pixel is assigned a probability $p \in (0, 1)$ of being water (i.e., flooded).

2.2. Flood Forecasting Model Cascade

Flood forecasting systems often comprise coupled rainfall-runoff models and hydraulic models. This forecasting model setup is particularly relevant for achieving accurate predictions of flood extents and water depth over large floodplains, which is of primary importance for flood management (Grimaldi et al., 2016). In contrast, simplified channel flow routing models are generally not able to accurately predict floodplain inundation extents (Montanari et al., 2009; Schumann et al., 2013). In this study, the flood forecasting model cascade comprises the conceptual hydrological model SUPERFLEX loosely coupled with the grid-based hydraulic model Lisflood-FP. These models are generally able to capture floodplain and streamflow dynamics in a parsimonious and computationally efficient manner, making them very well suited for near-real-time flood forecasting across a wide range of locations.

SUPERFLEX (Fenicia et al., 2011) is a flexible modeling framework designed for hypothesis testing in hydrology. It is based on a combination of generic elements, namely reservoirs, lag functions, and connection functions. The reservoir elements are intended to represent hydrological processes such as rainfall interception, upper soil root zone storage, riparian zone storage, and fast/slow runoff generation. A variety of model structures can be built from these generic elements. Within each reservoir, the storage/flow relationships and shape of lag functions can be specified from a set of available options. In our application, SUPERFLEX uses as inputs time series of rainfall and air temperature and produces as outputs time series of simulated storage, subsurface flow, and surface runoff.

Lisflood-FP (Bates & de Roo, 2000) is a grid-based two-dimensional hydraulic model. In the floodplain, the model solves a simplified inertial version (neglecting convective acceleration) of the de Saint Venant equations over the two horizontal dimensions using a finite difference method (de Almeida et al., 2012).

A posteriori developments of the model include a subgrid channel routing (Neal et al., 2012) for simulating the flow within the channel. As usual in hydraulic modeling, Lisflood-FP requires downstream boundary conditions, the inflow for all river streams within the area of interest, the ground elevations in the floodplain, and the geometry of river streams. The main hydraulic parameter is the Manning friction coefficient, which can be specified as spatially distributed.

The streamflow time series simulated by the hydrological model SUPERFLEX are used as inputs into the hydraulic model Lisflood-FP. The model cascade outputs time series of water depth and water fluxes at each model grid cell.

As a working hypothesis, it is assumed that the main and eventually dominating source of uncertainty in the streamflow, water depth, and flood extent forecasts is the rainfall forecast. Following this assumption, an ensemble of perturbed rainfall forecasts is used to produce an ensemble of rainfall inputs, which is then propagated through the forecasting chain. This assumption is expected to be reasonable in an operational context (Laurent et al., 2010; Yatheendradas et al., 2008), though it is well understood that hydrological and hydraulic model parameters, as well as catchment properties such as river and floodplain geometries, can also contribute substantial uncertainty. Instances where these other sources of error are significant would require developing separate error models, for example, using aggregated approaches such as in McInerney et al. (2017). Alternatively, the perturbed rainfall could potentially be used to compensate for other sources of uncertainty; this compensation will reduce the interpretability of the perturbed rainfall (Renard et al., 2010; Renard, 2011) but could still lead to operationally adequate predictions.

3. Data Assimilation Framework

This section develops the new data assimilation framework for integrating SAR-derived probabilistic flood maps within the flood forecasting model cascade.

The major structural assumption is that the hydraulic model uncertainty originates solely from uncertainty in the boundary conditions (i.e., upstream streamflow), which in turn originates from the uncertainty in rainfall forecasts.

Using the outputs of the hydrological model as upstream boundary conditions, the hydraulic model generates an ensemble of water depth and flood extent maps. An output of the hydraulic model is then an ensemble of binary maps where each pixel is either flooded or nonflooded. The set of flooded/nonflooded pixels is then compared against the probabilistic flood map derived from a SAR image (section 2.1). To optimally combine observations and simulations, we adopt an approach based on PF. The prior and posterior probability distributions, which describe model state variables and flood extent across model grid cells before and after the assimilation event, respectively, are approximated by a set of *particles*. In this study, each particle represents the output from the forecast model with its own set of forcing data (rainfall amounts) and its own *weight*, which can be interpreted as the probability of that particular model output being *correct*. During an assimilation event, the PF is used to update the particle weights based on information from the satellite flood extent data. The proposed implementation of the PF is based on Sequential Importance Sampling (Plaza et al., 2012). The PF framework offers the key advantage of relaxing the assumption that observation errors are Gaussian (Moradkhani, 2007), making it better suited to data assimilation of probabilistic flood maps than the more widely used ensemble Kalman filters (EnKFs; e.g., García-Pintado et al., 2015; Neal et al., 2007; Revilla-Romero et al., 2016). Moreover, the Sequential Importance Sampling implementation of a PF has the advantage of avoiding an update of model states that may lead to hydraulic model instabilities.

In our PF implementation, the weight of a given particle (equation (4)) is obtained by aggregating (equation (3)) the weights associated with the individual pixels across the simulated area given the observed image being assimilated (equation (2)). The computation of these weights is described next.

The weight $w^i_{(k,j)}$ associated with the kth pixel of the jth particle (i.e., model ensemble member) is calculated for the ith observation as follows:

$$w^i_{(k,j)} = p^i_k(w|\sigma^0) \times \theta^i_{(k,j)} + \left(1 - p^i_k(w|\sigma^0)\right) \times \left(1 - \theta^i_{(k,j)}\right) \quad (2)$$

where $p^i_k(w|\sigma^0)$ is the probability of the kth pixel in the ith flood map being water (see Equation (1). The term $\theta^i_{(k,j)}$ is the corresponding model prediction, interpreted as follows: $\theta^i_{(k,j)} = 1$ if the pixel is predicted as flooded and $\theta^i_{(k,j)} = 0$ otherwise.

The spatial aggregation of individual local weights $w^i_{(k,j)}$ (i.e., weights for individual pixels within a given particle) to obtain the global particle weight ω^i_j makes the use of three pragmatic ideas: (i) prevent particles being given a weight of zero solely due to a mismatch at a few pixels (e.g., if a pixel estimated as being *certainly* flooded in the observed image but is simulated as dry by the model, and *vice versa*), (ii) reduce biases due to overprediction of flood extent (false positive) being penalized more strongly than underprediction; and (iii) reduce the risk of particle degeneration (Plaza et al., 2012), where weights for all but a few particles go to zero.

The first challenge, namely, the risk of discarding an entire particle due to mismatches at a few pixels, arises because pixels with probability values close to 0 or 1 in the SAR-derived flood map can lead to weights of zero if observation and model result disagree locally. The entire particle could then be discarded (given zero weight), even if all but one pixel provide a perfect match. To avoid this unreasonable behavior, probability values of individual pixels in the probabilistic flood map are constrained to lie in the range (0.001, 0.999); all lower and higher probability values are truncated to these min and max bounds, respectively.

The second challenge, namely, *asymmetric* penalties for overprediction and underprediction, arises because, most of the time, flooded pixels only cover a limited portion of the flood image. As a consequence, the satellite-derived probabilistic flood map tends to be dominated by values below 0.5, and overprediction of flood extent tends to be penalized much stronger than underprediction when computing the global weights (see equations (3) and (4) hereafter). To overcome this undesirable behavior, pixels with the lowest probabilities of being flooded are masked out prior to the assimilation, so that the numbers of assimilated pixels with probabilities below and above 0.5 are (approximately) balanced.

The third challenge, namely particle degeneration, can be illustrated by considering the (rather unrealistic) assumption that all pixel observations are independent and computing the global weight of a particle as the simple product of its local pixel weights. When the number of pixels in the assimilated image is large, the global weights for the majority of particles degenerate virtually to zero. The challenge of particle degeneracy is a well-known problem in the PF community, and no theoretically complete solution exists to our best knowledge (Li et al., 2014). In this work, we employ a pragmatic solution, which introduces an empirical rescaling factor α into the computation of the global weight ω^i_j (see equation (3)).

$$\omega^i_j = \prod_{k=1}^{N} \left(w^i_{(k,j)} \right)^{\frac{\alpha}{N}} \tag{3}$$

where N indicates the number of pixels within the assimilated image.

In this work, α is set to the maximum value such that the worst possible model (in complete disagreement with the SAR-derived probabilistic flood map) has a strictly positive (nonzero) global weight. A specific α value is set for each individual satellite observation. This approach prevents the particle degeneracy issue because all particles are guaranteed to retain a nonzero weight.

To ensure the particle posterior probabilities sum up to 1, the global weights are normalized

$$W^i_j = \frac{\omega^i_j}{\sum_{j=1}^{Np} \omega^i_j} \tag{4}$$

where W^i_j is the posterior probability of the jth particle given the ith observation. Once all particles and their weights are computed, they are used to obtain the expectation of any state variable of the modeling cascade. For example, consider the ensemble of simulated water depth $D^i_k = \left\{ d^i_{(k,j)}, j = 1, \ldots, Np \right\}$ on the kth grid cell and at time step i. The expectation of the water depth $Ex(D^i_k)$ at this grid cell and this time step is computed as follows:

$$Ex(D^i_k) = \sum_{j=1}^{Np} W^i_j \times d^i_{(k,j)} \tag{5}$$

The same formula (equation (5)) is used to compute expectation of streamflow. The expectation of water level and flood extent is derived from the expectation of water depth (equation (5)).

4. Study Area, Data, and Forecasting Model Setup

4.1. Study Area and Gauging Station Network

The study area is a 30.5 km by 52.4 km domain around the town of Tewkesbury (United Kingdom), located at the confluence of River Severn and River Avon. Figure 1 shows the model domain and river network. The green squares indicate model upstream boundaries (with associated gauging stations); the yellow stars indicate gauging stations in the interior of the hydraulic model domain, which are used to evaluate model results. Two of these evaluation stations are located upstream of the confluence, along River Avon (Bredon) and River Severn (Saxons Lode), and two others are located downstream of the confluence (Deerhurst and Haw Bridge). This gauging station network configuration is highly beneficial for this study, as it allows model predictions for the Rivers Severn and Avon to be evaluated both jointly and separately.

4.2. Hydrometeorological Data

All gauging stations (see Figure 1) provide time series of water level at 15-min intervals. In this study, streamflow time series at Evesham (River Avon) and Bewdley (River Severn) gauging stations are used as well.

Observed streamflow time series at Evesham (River Avon) and Bewdley (River Severn) for years 2004 to 2006 were used for the calibration of the SUPERFLEX models, and time series for years 2007 to 2010 were used for their validation.

The assimilation experiment considers four flood events. These flood events are of moderate to large magnitude and occurred in March and July 2007, January 2008, and January 2010 and have been described in detail by Wood et al. (2016).

Time series of rainfall and 2-m air temperature over the corresponding river basins are available at a 3-hourly time step and 0.75° spatial resolution from the ERA-Interim reanalysis data set (Dee et al., 2011). From this data set, based on the grid cells (or portions of the grid cells) located within each river basin, we computed spatially averaged rainfall and 2-m air temperature for the period 2003 to 2010. Next, the resulting time series were uniformly redistributed to an hourly time step. The potential evapotranspiration was estimated from the air temperature data using the Hamon formula (Hamon, 1963). Rainfall and potential evapotranspiration time series are used as inputs of the SUPERFLEX models.

4.3. Earth Observation Data Set

A total of 11 flood images of the 4 considered flood events (see section 4.2) are selected from the Envisat Wide Swath Mode imagery archive (European Space Agency). The satellite data are described in detail in Wood et al. (2016).

The flood event of July 2007 is of particular value for testing and evaluating the proposed methodology because an airborne campaign imaged the flood at a very high spatial resolution (50 cm) the day following the two Envisat image acquisitions on 23 July 2007 at 10:27 a.m. and 9:53 p.m. (GMTZ). The aerial photographs taken during this airborne campaign were photo-interpreted to delineate flood extents (see Giustarini et al., 2016 for more details). The resulting flood validation map allows to evaluate the flood extent forecasts using a data set independent from the SAR images.

Figure 2 shows all probabilistic flood maps derived from the SAR image database using Giustarini et al.'s (2016) method, together with the flood validation map manually delineated from the aerial photography campaign of 24 July 2007. As seen in Figure 2, these probabilistic flood maps correspond to markedly different flood extents. This makes the image database especially valuable for testing the proposed assimilation framework under diverse hydrological conditions. The probabilistic flood maps derived from the two images acquired in July 2007 were evaluated in earlier work (Giustarini et al., 2016), demonstrating a good reliability compared to the flood validation map acquired almost synchronously (see reliability diagrams in Figures 1 and 2 of Giustarini et al. (2016)).

4.4. Hydrologic and Hydraulic Models

Two SUPERFLEX hydrological models were calibrated to the two main river basins in order to predict streamflow hydrographs at Evesham (River Avon) and Bewdley (River Severn).

The calibration of the two SUPERFLEX models uses as input the ERA-Interim reanalysis data and the potential evapotranspiration (section 4.2). The same SUPERFLEX model structure (but with different parameter values) is assumed in both basins. This structure was set to comprise a total of four reservoirs intended to simulate, respectively, the interception of rainfall, the water storage in the upper root zone soil layer, and the fast

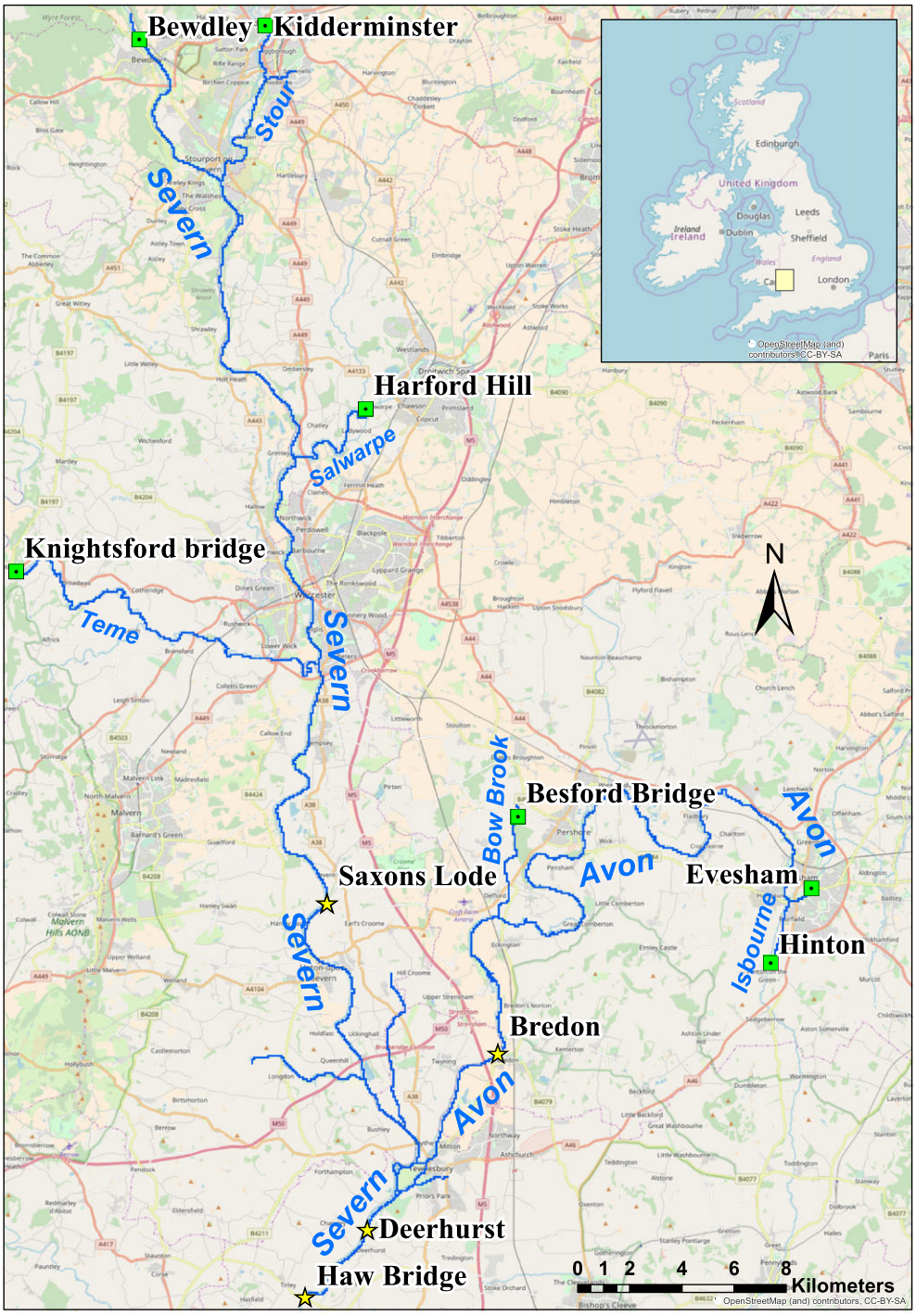

Figure 1. Study area and hydraulic model domain, showing the model boundary conditions (green squares) and the hydrometric gauging station used as a reference (yellow stars).

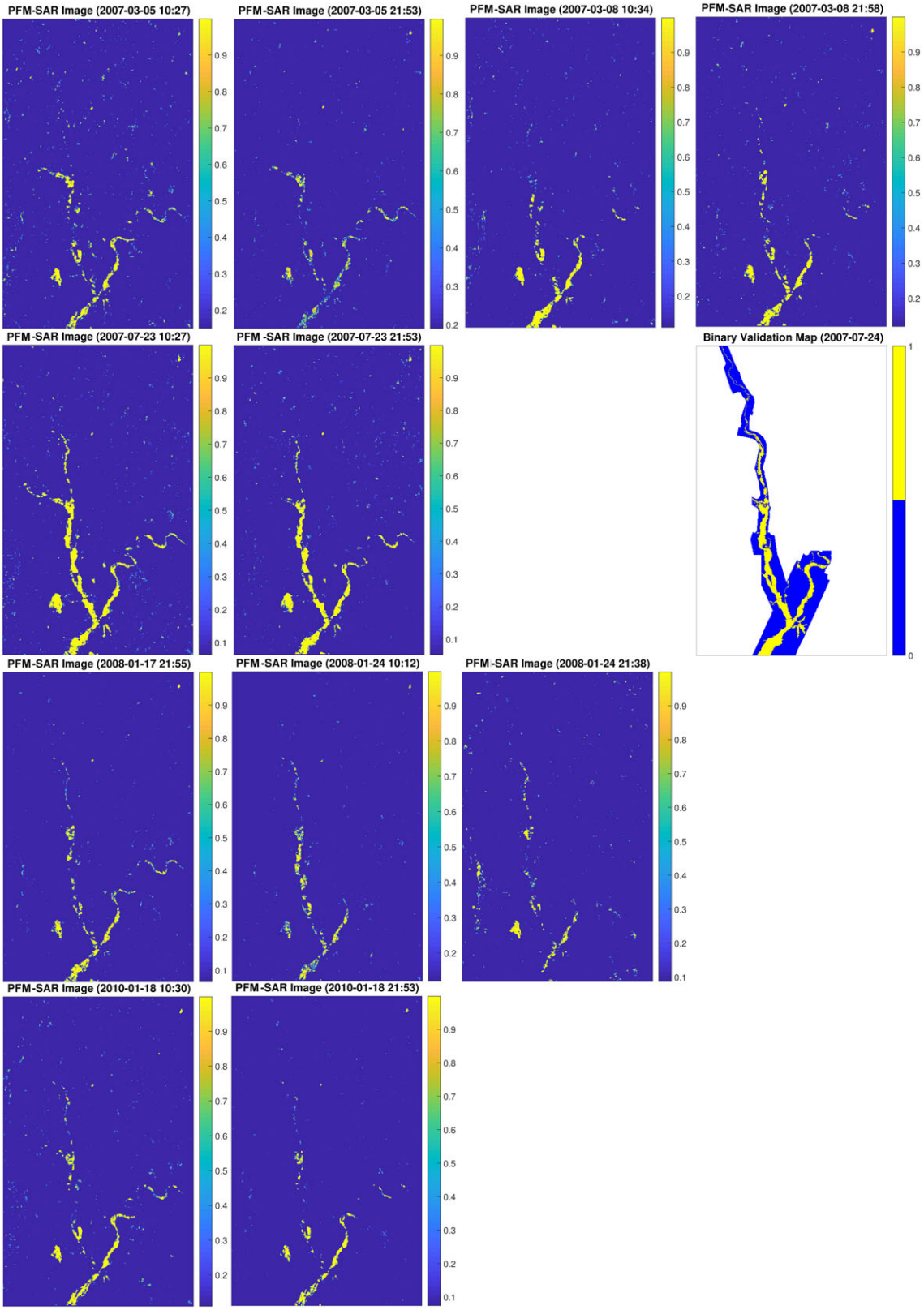

Figure 2. Probabilistic flood maps derived from the Envisat images (all events) and binary flood validation map derived from the aerial photographs (July 2007 flood event). SAR = synthetic aperture radar; PFM = probabilistic flood map.

and the slow runoff components. The SUPERFLEX models were calibrated to observed streamflow time series at Bewdley (Severn) and Evesham (Avon; section 4.2) using the weighted least squares technique, with inverse weights linearly dependent on the simulated flow magnitude (Kavetski & Fenicia, 2011; Evin et al., 2014). The Nash-Sutcliffe efficiency (Nash & Sutcliffe, 1970) achieved over the calibration period was 0.6 for the Avon model and 0.75 for the Severn model; over the validation period the Nash-Sutcliffe efficiencies at these locations were 0.63 and 0.74, respectively.

The hydraulic model used in this study is identical to the *observed* model described by Wood et al. (2016). This model was created using surveyed cross sections of the main rivers to determine channel width and depth, and fixing Mannings channel roughness coefficient to 0.038. Outside of the river channel, Mannings roughness coefficient was set to 0.06 (Wood et al., 2016). The downstream boundary condition is of uniform flow type with an energy slope set to 7×10^{-5}. The grid cell of this model is aligned with the SAR images, with a 75-m pixel spacing. The hydraulic model uses as input inflow for Rivers Severn and Avon the streamflow time series simulated by the corresponding SUPERFLEX models. Moreover, as Rivers Severn and Avon represent the two major tributaries within the hydraulic model domain, the streamflow hydrographs for the other tributaries shown in Figure 1 were constructed from gauged data without the use of a SUPERFLEX model.

4.5. Particle Ensemble Generation

As explained in section 2.2, rainfall inputs are assumed to be the main source of streamflow forecast uncertainty. This section describes the generation of particle ensembles via propagation of rainfall uncertainty. ERA-Interim uses a fixed version of a numerical weather prediction system to produce reanalyzed data. This system assimilates daily observations once they become available. The result of this assimilation is the starting point for the next forecast. Strictly speaking, the ERA-Interim predictions are not forecasts because they are produced at increasingly later times. In this study, we artificially emulate ensemble weather forecasts by perturbing the ERA-Interim data set. The control criteria proposed by de Lannoy et al. (2006) were used to obtain suitable prior particle probability distributions, and no additional quality assessment of the ERA-Interim rainfall product was carried out since de Leeuw et al. (2015) already showed a good correlation (0.91) with daily rainfall observations over England and Whales (covering our study area) despite underestimated daily rainfall amounts (~20% on average).

To emulate rainfall ensemble forecasts, the rainfall time series were perturbed using lognormal multiplicative noise. Each hourly rainfall value had an independently drawn random multiplier. These perturbed rainfall time series were then propagated through the SUPERFLEX models, generating a total of 128 streamflow hydrographs at Bewdley and 128 streamflow hydrographs at Evesham, for the period 2007–2010.

The parameters of the lognormal rainfall perturbation model were set based on the criteria proposed by de Lannoy et al. (2006) computed over the streamflow time series simulated by the SUPERFLEX models, namely, (i) the ratio of the averaged ensemble skill and the averaged ensemble spread (VM_1, see equations (6) and (7)), and (ii) the ratio of the average squared ensemble skill and the averaged root-mean-square error (RMSE) computed between simulation and observation (VM_2, see equations (6) and (8)).

$$\bar{\hat{q}}_i = \frac{1}{Np}\prod_{j=1}^{Np}\hat{q}_{i,j} \tag{6}$$

$$VM_1 = \frac{\langle \mathrm{ensk}_i\rangle_i}{\langle \mathrm{ensp}_i\rangle_i} = \frac{\frac{1}{Nt}\sum_{i=1}^{Nt}\left(\bar{\hat{q}} - q_i^o\right)^2}{\frac{1}{Nt}\sum_{i=1}^{Nt}\left(\frac{1}{Np}\sum_{j=1}^{Np}\left(\hat{q}_{i,j} - \bar{\hat{q}}_i\right)\right)} \tag{7}$$

$$VM_2 = \frac{\left\langle\sqrt{\mathrm{ensk}_i}\right\rangle_i}{\left\langle\sqrt{\mathrm{mse}_i}\right\rangle_i} = \frac{\frac{1}{Nt}\sum_{i=1}^{Nt}\sqrt{\left(\bar{\hat{q}} - q_i^o\right)^2}}{\frac{1}{Nt}\sum_{i=1}^{Nt}\sqrt{\frac{1}{Np}\sum_{j=1}^{Np}\left(\hat{q}_{i,j} - q_i^o\right)^2}} \tag{8}$$

In equations (6)–(8) the vertical bar indicates the ensemble mean (over the particles, see equation (6)), $\hat{q}_{i,j}$ is the streamflow simulated by particle j at time step i, Np is the number of particles, $\langle . \rangle_i$ indicates the average over the time steps with available streamflow observations, Nt the number of time steps with available streamflow observations, and q_i^o is the observed streamflow at time step i. The quantities ensk_i, ensp_i, and mse_i denote, respectively, the skill, spread, and mean squared error of the ensemble at time step i.

Table 1
Statistical Evaluation of SUPERFLEX Model Ensemble Based on de Lannoy et al. (2006) Recommendations

River station	VM_1	VM_2
Bewdley Severn	0.99	0.63
Evesham Avon	1.184	0.54
Target values (with 128 particles)	1	0.71

Criterion VM_1 assesses whether the ensemble spread and the model deviation to observation are of the same order of magnitude, and criterion VM_2 assesses whether the observation and the ensemble are statistically indistinguishable.

To define a suitable rainfall perturbation, many lognormal perturbations with various standard deviations were tested and the corresponding rainfall realizations were randomly generated. The set of perturbations that yielded VM_1 and VM_2 values closest to the target values (see Table 1) was selected. The rainfall perturbation can therefore be considered calibrated and site specific. The corresponding VM_1 and VM_2 values obtained for the two SUPERFLEX models are reported in Table 1 as well as the corresponding target values for 128 particles as given by de Lannoy et al. (2006). The values of VM_1 and VM_2 in Table 1 are reasonably close to their target values for both models, and hence, the generated ensemble is considered statistically satisfactory. For the River Severn, the lognormal perturbation mean and standard deviation were set to 1.0 and 0.2, respectively. For the River Avon, the mean and standard deviation were set to 1.0 and 0.35.

The ensemble of 128 perturbed rainfall amounts is used as forcings to the two calibrated SUPERFLEX models for the Rivers Severn and Avon. Figure 3 shows the ensemble of streamflow hydrographs generated by the two SUPERFLEX models for the considered flood events (see section 4.2). Each black line corresponds to a single particle simulation, and the red line indicates the corresponding observed streamflow hydrograph.

The ensembles of *forecast* hydrographs obtained for the two rivers generated using SUPERFLEX are then propagated through the Lisflood-FP hydraulic model to yield ensemble forecasts of streamflow, water depth, water level, and flood extent maps. The forecasts are therefore generated continuously, so that any time step can be considered as a forecast initialization date.

4.6. Assimilation of SAR-Derived Flood Maps

The method developed in section 3 for assimilating SAR-derived probabilistic flood maps is applied to the four flood events described in section 4.2. The forecast distributions are represented using 128 particles. Each time a SAR image is available, the corresponding probabilistic flood map is assimilated by computing the posterior probability of each particle (see equations (2)–(4)), updating the particle weights, and computing the resulting expectation (see equation (5)) of streamflow, water depth, water level, and flood extent forecasts.

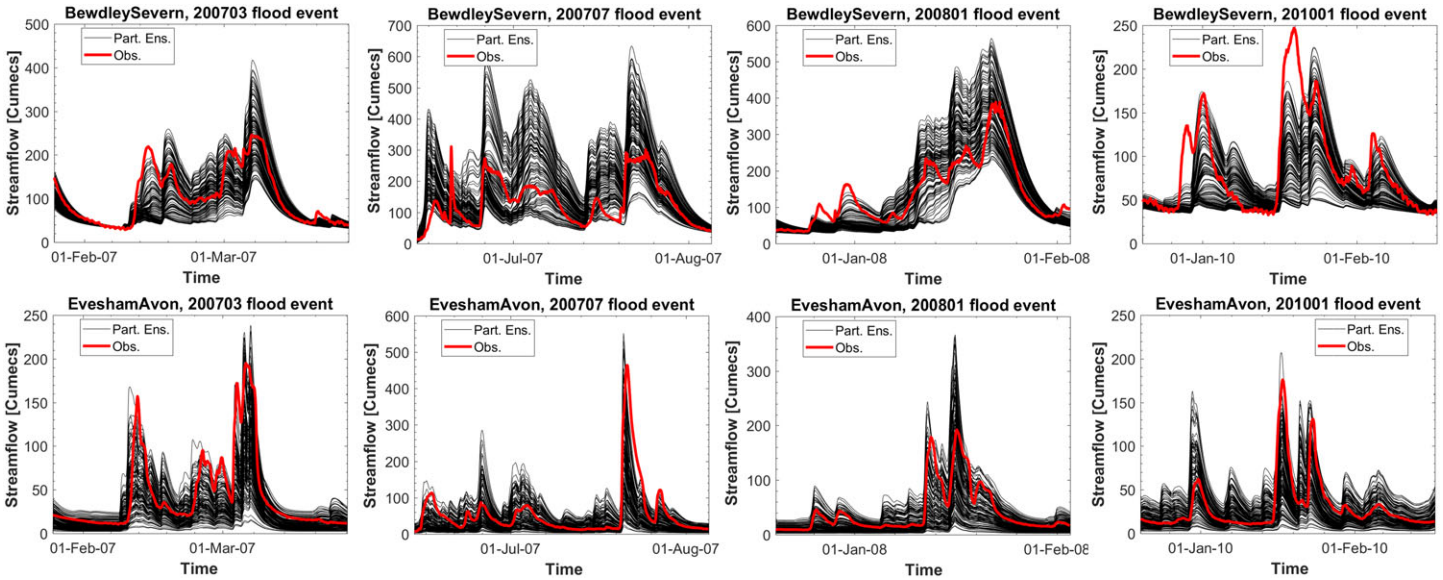

Figure 3. Ensembles of inflow time series (black lines) of Lisflood-FP based on SUPERFLEX simulation at Bewdley on the River Severn and Evesham on the River Avon. Red lines indicate the observed streamflow time series.

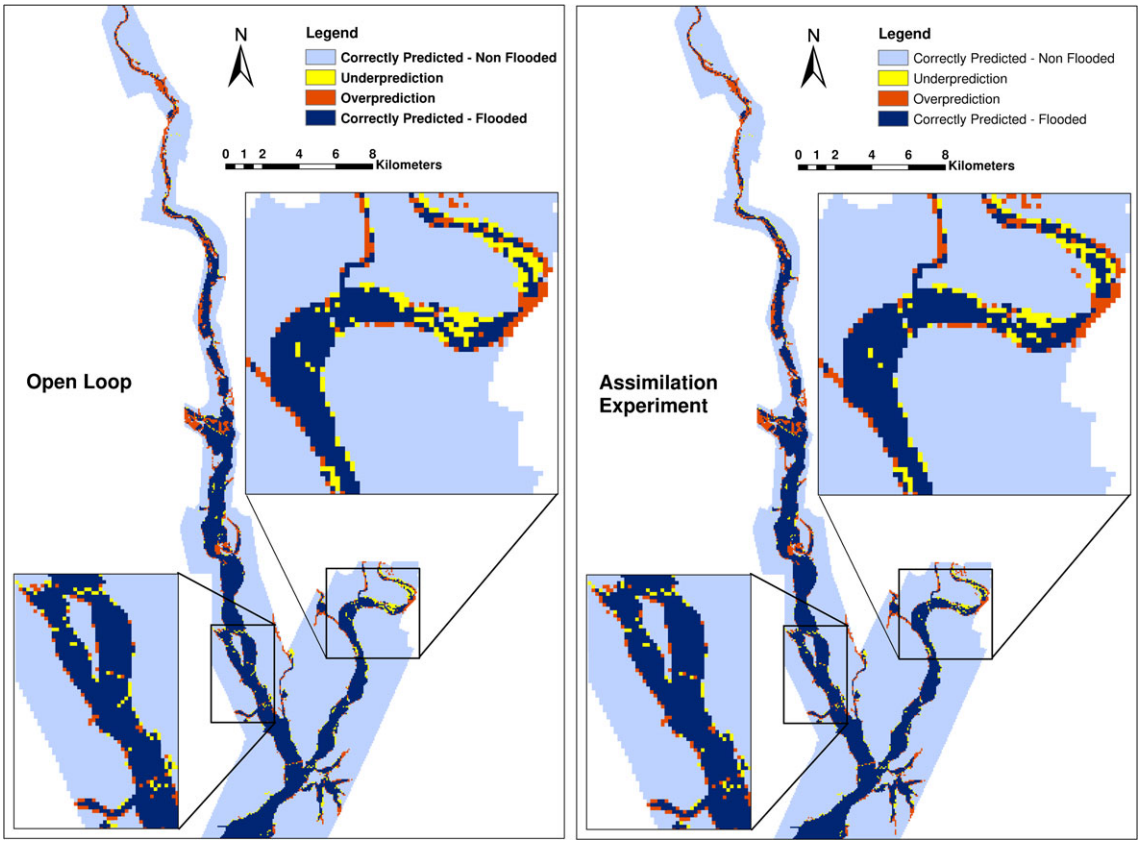

Figure 4. Comparison of forecast versus observed validation flood extent maps. Left-hand panel shows results of open-loop simulation; right-hand panel shows results obtained after the assimilation of the SAR image acquired on 23 July 2007 at 21:53.

4.7. Analyses Used to Test the Proposed Framework

The proposed framework for improving flood forecasting using SAR image assimilation is evaluated using a series of empirical tests. The tests are based on comparing the model expectation of flood extent, streamflow, and water levels against respective reference data sets, namely, the aerial photography-derived flood validation map (only available for the July 2007 flood event), the observed streamflow time series at Evesham and Bewdley, and the water level measurements at Saxons Lode, Bredon, Haw Bridge, and Deerhurst gauging stations. Two forecasting regimes are considered: (i) an open-loop simulation where no image is assimilated and all particles remain equally weighted at all time steps, and (ii) an assimilation experiment, where particle weights are updated each time an image is available as described in section 3.

First, we investigate the impact of the assimilation procedure on the forecast flood extent. The forecast flood extents obtained using the open-loop expectation and the assimilation expectation are plotted and compared to the observed validation flood extent on 24 July 2007 (see Figure 2, second row, right panel). The forecast flood extents are then further evaluated using the following performance coefficients (metrics): the confusion matrix (Jolliffe & Stephenson, 2011), overall accuracy (Stehman, 1997), critical success index (Donaldson et al., 1975), and Cohen's kappa (Cohen, 1960). These metrics are commonly used to assess forecast or image classification quality (Jolliffe & Stephenson, 2011; Stehman, 1997). The confusion matrix is composed of four values, which in this study are defined as follows: the number of pixel correctly predicted as *flooded*, the number of pixels associated with overprediction of flood extent (false positives), the number of pixels associated with underprediction (false negatives), and the number of pixels correctly predicted as *nonflooded*. The accuracy, critical success index, and kappa coefficients are computed from the confusion matrix. These metrics quantify goodness of fit; they attain their highest value of 1 when the predictions provide a perfect fit to the observations.

Second, the impact of SAR image assimilation on forecast quality is investigated for streamflow time series. This analysis is carried out by plotting observed, open loop, and assimilation experiment-derived

Table 2
Evaluation of the Forecast Flood Extent Maps Using as a Reference the Observed Flood Extent Map Derived From the Aerial Photographs on 24 July 2007

Experiment		Confusion Matrix FF	FN	Accuracy	CSI	Kappa Stat.
Open Loop	OF	7867	652	0.938	0.785	0.838
	ON	1506	24930			
Assimilation	OF	7975	544	0.938	0.787	0.839
	ON	1618	24818			

Note. CSI = critical success index; FF = forecast flooded; FN = forecast nonflooded; ON = observed nonflooded; OF = observed flooded.

streamflow time series at Bewdley (River Severn) and Evesham (River Avon). It is used to evaluate the hydrological forecasts.

Third, the impact of SAR image assimilation on the forecast quality and uncertainty is investigated for water level time series. This analysis is carried out by first plotting observed, open loop, and assimilation experiment-derived water level time series at Saxons Lode, Bredon, Deerhurst, and Haw Bridge. As these locations lie throughout the hydraulic model domain, this analysis allows to evaluate the complete modeling chain (SUPERFLEX + Lisflood-FP) forecasts. In addition, we compare the widths of the 90% uncertainty intervals of the ensemble forecasts of water levels generated using the open loop and assimilation experiments, at Saxons Lode, Bredon, Deerhurst, and Haw Bridge.

Fourth, a brief comparison of the forecast performance of the assimilation experiments in our study and the study by García-Pintado et al. (2015) is carried out. The study by García-Pintado et al. (2015) shares two similarities with our work: (i) information derived from a series of very high-resolution SAR images

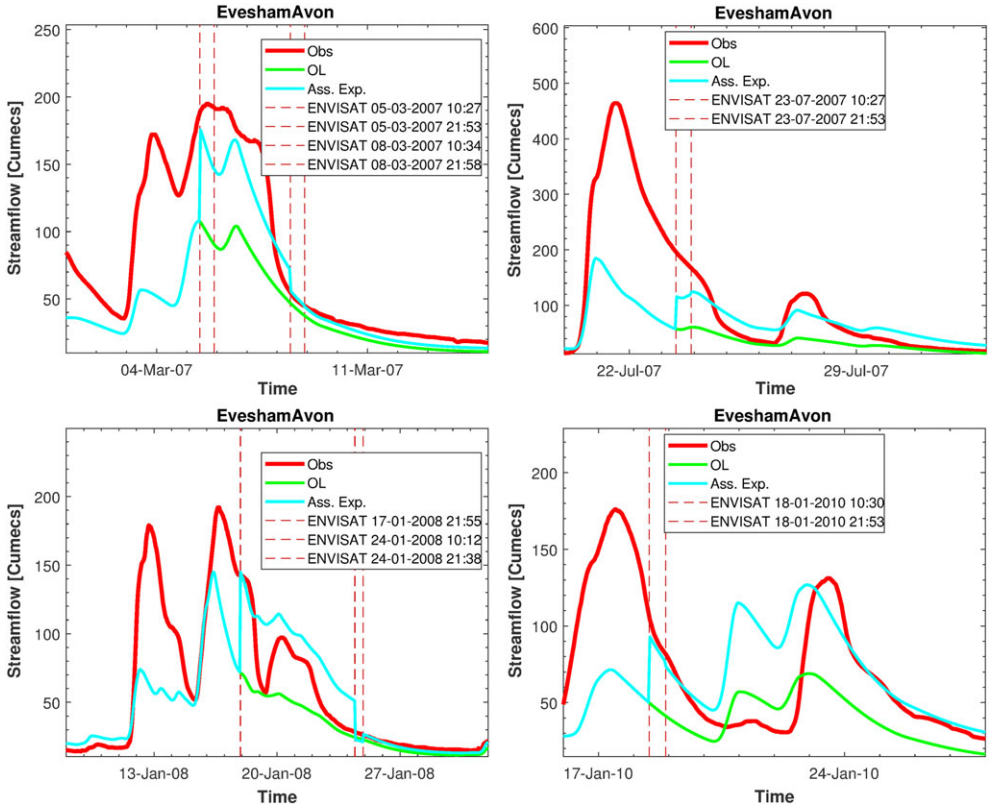

Figure 5. Streamflow time series at Evesham for the four flood events: open-loop expectation (green), assimilation experiment expectation (cyan), and observations (red).

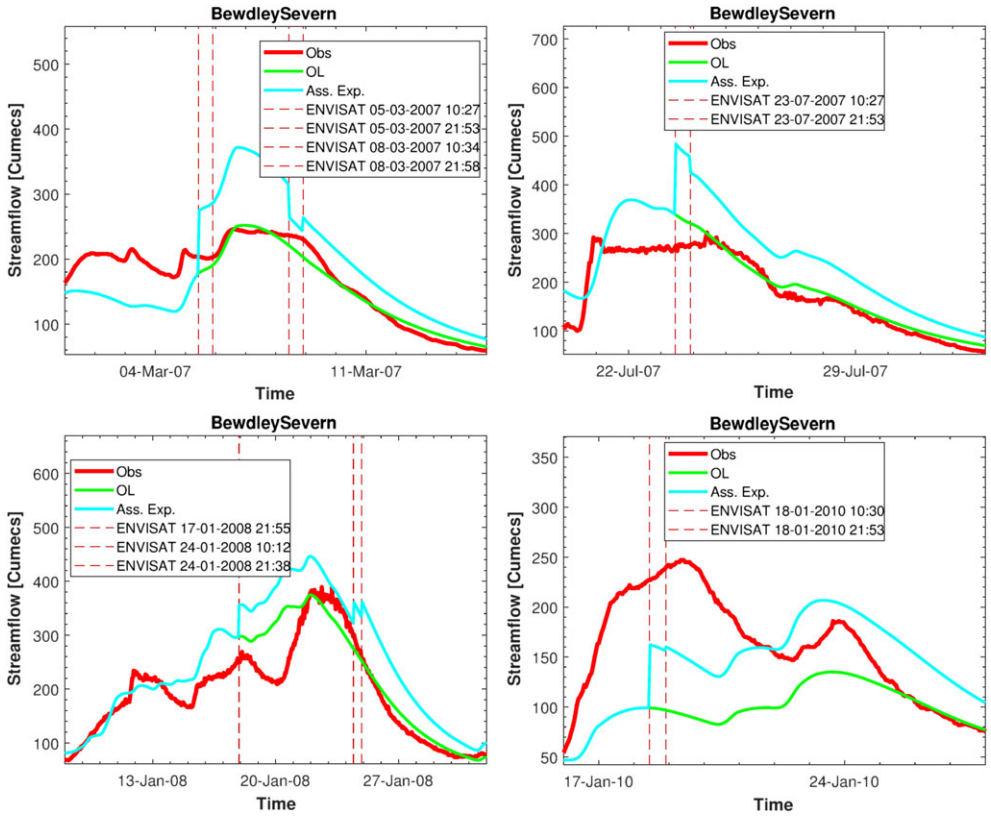

Figure 6. Streamflow time series at Bewdley for the four flood events: open-loop expectation (green), assimilation experiment expectation (cyan), and observations (red).

(namely, water levels derived from Cosmo-Skymed satellite images) is assimilated (using the local ensemble transform Kalman filter) into a flood forecasting chain; (ii) the same model domain, River Severn, is used—although for a different flood event from July 2011. For this comparison, we use in situ observations to compute the RMSE of streamflow forecasts at Bewdley and Evesham and the RMSE of water level forecasts at Bredon and Saxons Lode, and report the RMSE averaged over the four flood events considered in our case study (section 4.2). These forecast quality metrics are then compared to the metrics reported in García-Pintado et al. (2015) for the July 2011 flood event at the same locations and with a similar assimilation experiment setup.

Finally, we consider the persistence of the impact of data assimilation, in terms of the simulated water level time series at Saxons Lode, Bredon, Deerhurst, and Haw Bridge. This behavior is investigated by computing the averaged (over the 11 assimilation events) ratio of the forecast error of the open-loop simulation and the forecast error of the assimilation experiment, and reporting this ratio as a function of increasing prediction lead time from the assimilation event.

5. Results and Discussion

5.1. Assessment of Forecast Flood Extent

We begin by analyzing the flood event of July 2007. Figure 4 compares the observed flood validation map with the flood extent map derived, respectively, from the open loop and the assimilation experiment. Table 2 summarizes the comparison of forecast and observed (validation map) flood extent. In Table 2 the notations OF, ON, FF, and FN stand, respectively, for observed flooded, observed nonflooded, forecast flooded, and forecast nonflooded pixels.

Figure 4 and Table 2 show that assimilation experiment during the July 2007 flood event has a somewhat limited effect on the predicted flood extent. Indeed, the maps in Figure 4 are very similar, and the flood extent map derived from the open loop is already quite accurate and closely resembles the observed validation

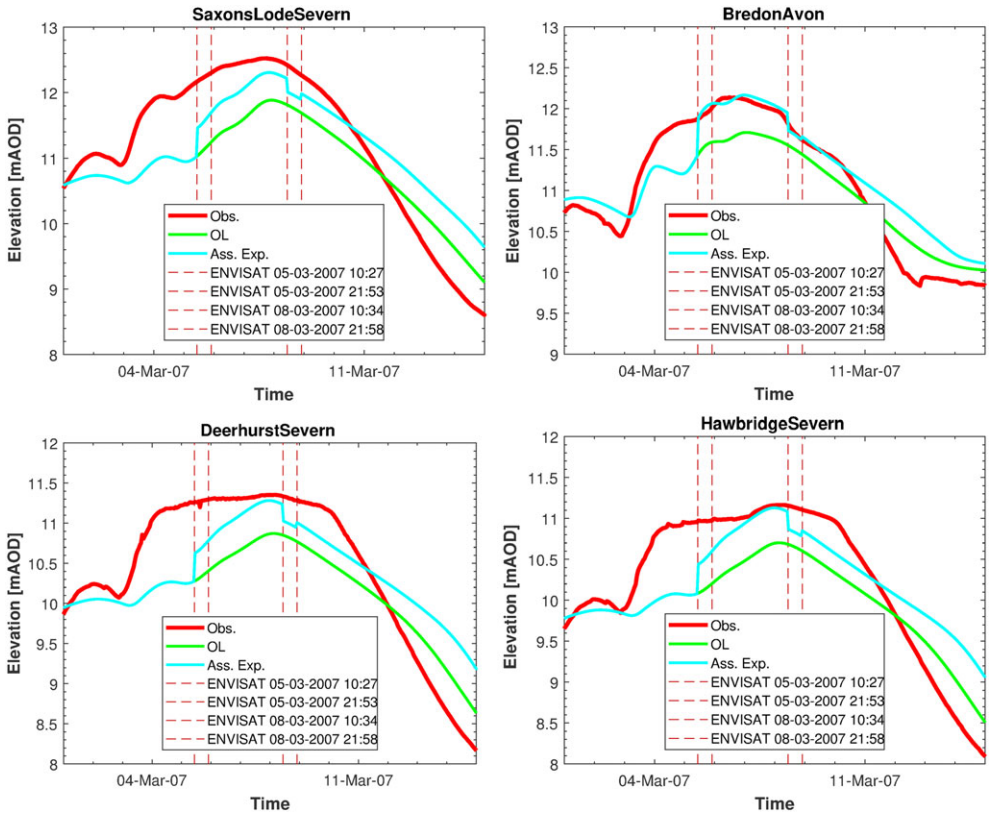

Figure 7. Water level time series at Saxons Lode, Bredon, Deerhurst, and Haw Bridge for March 2007 flood event: open-loop expectation (green), assimilation experiment expectation (cyan), and observations (red).

map. Figure 4 also shows that most of the overprediction (i.e., red pixels) occurs along the upstream part of River Severn, while most of the underprediction (i.e., yellow pixels) occurs along the upstream part of River Avon. In Table 2, a slight improvement is visible in all metrics. An in-depth examination of the confusion matrices in Table 2 indicates that only a limited number of pixels are actually updated during the assimilation. One promising result is that the assimilation substantially increases the number of correctly forecast flooded pixels (i.e., correct positives), from 7867 pixels in the open loop to 7975 pixels in the assimilation experiment. However, this positive effect is counterbalanced by an increase in overprediction (false positives), from 1506 pixels in the open loop to 1618 pixels in the assimilation experiment.

5.2. Assessment of Forecast Streamflow Time Series

The predicted streamflow time series at the hydraulic model upstream boundaries (Evesham and Bewdley) are examined next. Figures 5 and 6 show the streamflow hydrographs obtained for the considered flood events against the observed data. Each of these figures is dedicated to one gauging station, and each figure panel is dedicated to a single flood event. Within these figures, the vertical dashed lines indicate the acquisition time of the SAR images. It is worth noting that, for each flood event, the open loop and the sequentially updated simulations are identical before any satellite image has been assimilated, because prior weights of the particles are assumed uniform.

Results are clearly encouraging for Evesham (River Avon, Figure 5). Streamflow forecasts obtained using SAR image assimilation (cyan lines) move substantially closer to the observation than the open-loop simulations (green line) at all 11 assimilation events in all 4 flood events. In three of the four flood events, the assimilation procedure results in the expected streamflow forecast matching the observation almost exactly. On average over the 11 assimilation events, assimilation reduces the errors in forecast streamflow at Evesham by a factor of 5. The bottom left panel of Figure 5 demonstrates that assimilation corrects both overprediction and underprediction. The results obtained for the River Severn are more ambivalent, as shown in Figure 6.

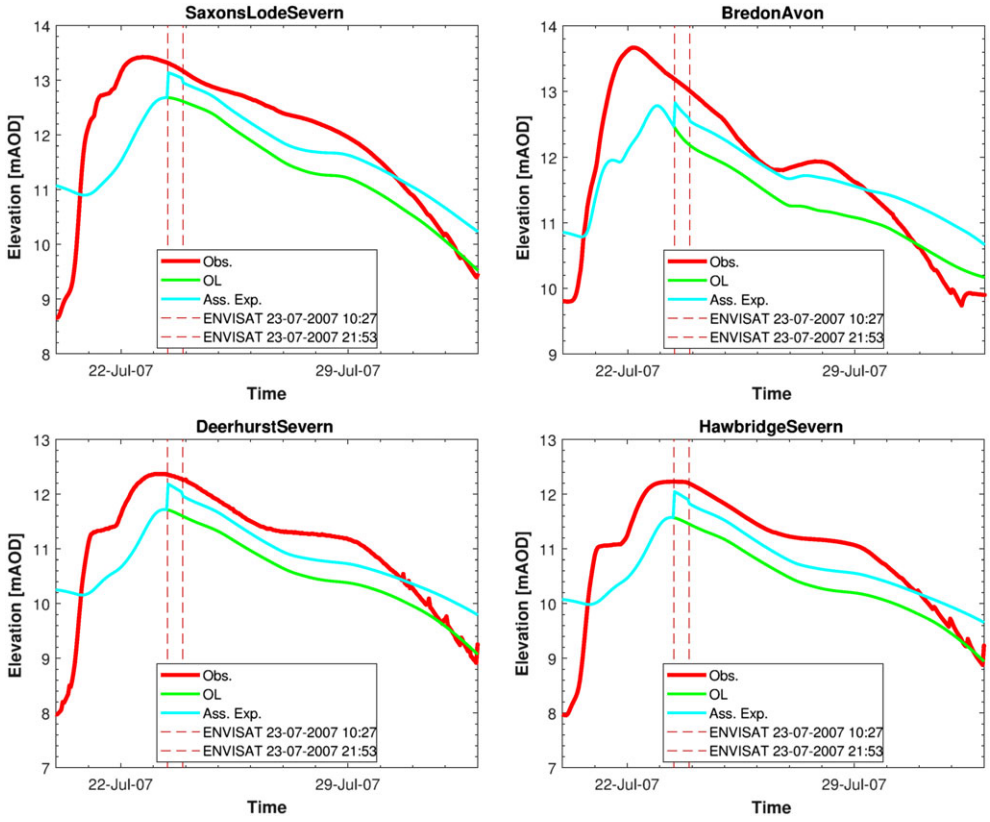

Figure 8. Water level time series at Saxons Lode, Bredon, Deerhurst, and Haw Bridge for July 2007 flood event: open-loop expectation (green), assimilation experiment expectation (cyan), and observations (red).

Here the assimilation leads to an overestimation of streamflow forecasts during the first three flood events at Bewdley (Figure 6). This result indicates that the hydraulic model would need more incoming water than is observed at River Severn upstream boundary condition to reproduce the observed flood extents. At this stage it is difficult to establish the reason for this streamflow overprediction. For example, it could be due to the simplified representation of riverbed geometry in the hydraulic model, errors in the rating curves used for converting water level into streamflow (Bates et al., 2006; Bermúdez et al., 2017; Neal et al., 2011), or an overestimation of the flood extent extracted from the SAR images. An additional reason for this overestimation could be the locally sourced flows due to rainfall within the hydraulic model domain, which are not taken into account as usual in hydraulic modeling. The study by Bermúdez et al. (2017) reported similar model overprediction in the same study area for the July 2007 flood event. It highlighted underestimation of river streamflow for River Teme at Knightsford Bridge gauge and for River Severn at Bewdley gauge especially for high streamflows. Moreover, it showed that the 2007 rainfall storm event occurred right over the hydraulic model domain and that it was necessary to take local rainfall-runoff processes to correctly simulate flood peak for this event over River Severn. Given this storm path, we can hypothesize that, in our study, the assimilation experiment overpredicts the upstream River Severn streamflow (at Bewdley) because it is trying to improve the forecasts downstream, while compensating for underestimated streamflow in other tributaries and for locally sourced flows.

5.3. Assessment of Forecast Water Level Time Series

Figures 7–10 show the assimilation results for the four flood events against observed water levels, at four gauging stations. Note that the results for the Deerhurst gauging station are not shown in Figure 9 because this station did not record the water levels during the January 2008 flood event. The symbology in Figures 7–10 is identical to that of Figures 5 and 6.

Figures 7–10 indicate that sequential assimilation of satellite observations improves the quality of water level forecasts. The forecasts obtained using assimilation (cyan lines) move substantially closer to the observation

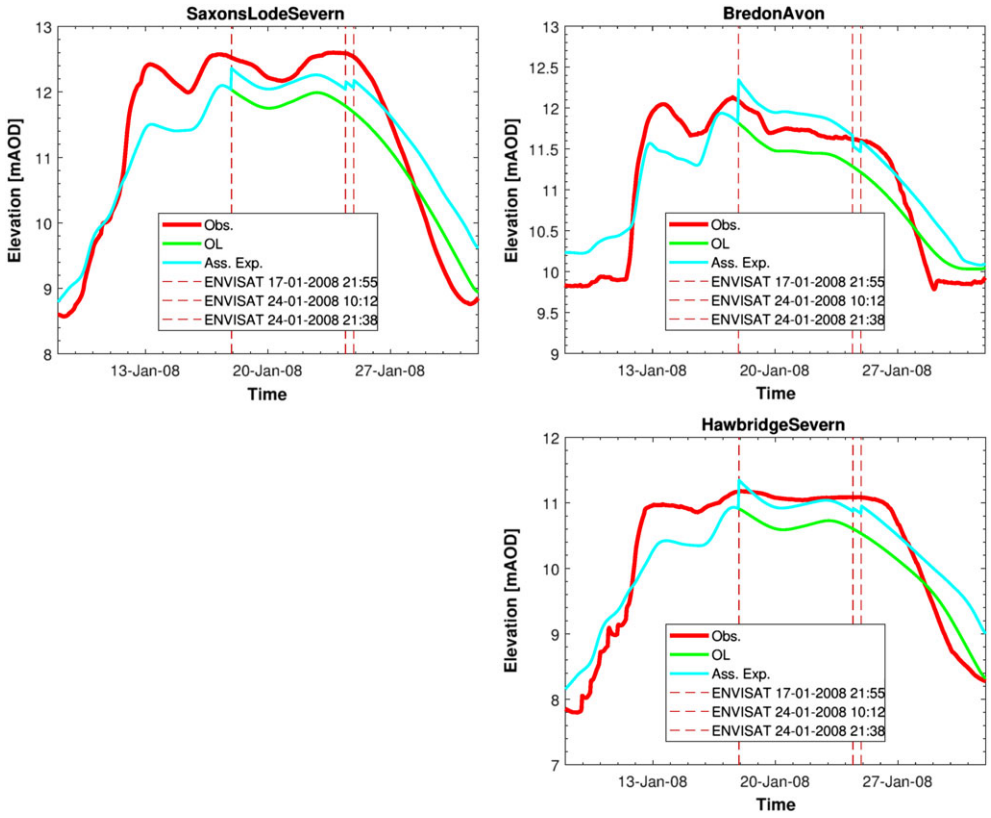

Figure 9. Water level time series at Saxons Lode, Bredon, and Haw Bridge for January 2008 flood event: open-loop expectation (green), assimilation experiment expectation (cyan), and observations (red).

than the open-loop simulation (green line), at all assimilation steps in all flood events, at all four gauging stations. On average, assimilation reduces the errors in predicted water levels by a factor of 2. This finding demonstrates that the assimilation is effective at correcting errors for both the River Severn and the River Avon. Results shown in Figures 7–10 are clearly encouraging: the assimilation of flood extent maps that characterize floodplain inundation effectively corrects simulated water levels within the main channel. This finding suggests that, although flood extent is not a diagnostic, but rather a prognostic variable of hydraulic models, it is a relevant source of information that can substantially improve predictions of water level over the entire model domain. Moreover, the systematic improvement of water level forecast lends further weight to the hypothesis that the overprediction of streamflow at Bewdley as a result of the assimilation (section 5.2) is due to a compensation for missing incoming water in the hydraulic model domain.

Figure 11 plots the widths of the 90% uncertainty intervals of the water level forecasts produced by the assimilation and open-loop approaches. Forecast uncertainty reduces systematically for every assimilation event, at all river stations. On average, assimilation tightens the 90% uncertainty interval by a factor of 3 compared to the open loop.

5.4. Comparison With the Similar Study by García-Pintado (2015)

The average (over the four flood events) performances of the streamflow and water level forecasts obtained in our assimilation experiments are as follows: the RMSE of streamflow forecasts at Bewdley and Evesham are 69.67 m^3/s and 56.80 m^3/s, respectively, and the RMSE of water level forecasts at Bredon and Saxons Lode are 0.47 m and 0.72 m, respectively. These RMSE values are similar to those reported by García-Pintado et al. (2015) in their Tables 3 and 4. Although the setup of the forecast models, the SAR images and the derived information, the assimilation filter, and the considered flood events in our study is different from those of García-Pintado et al. (2015), the model domain —River Severn— is the same. Hence, the similarity of RMSE values here and in the study by García-Pintado et al. (2015) suggests that the performance of the models

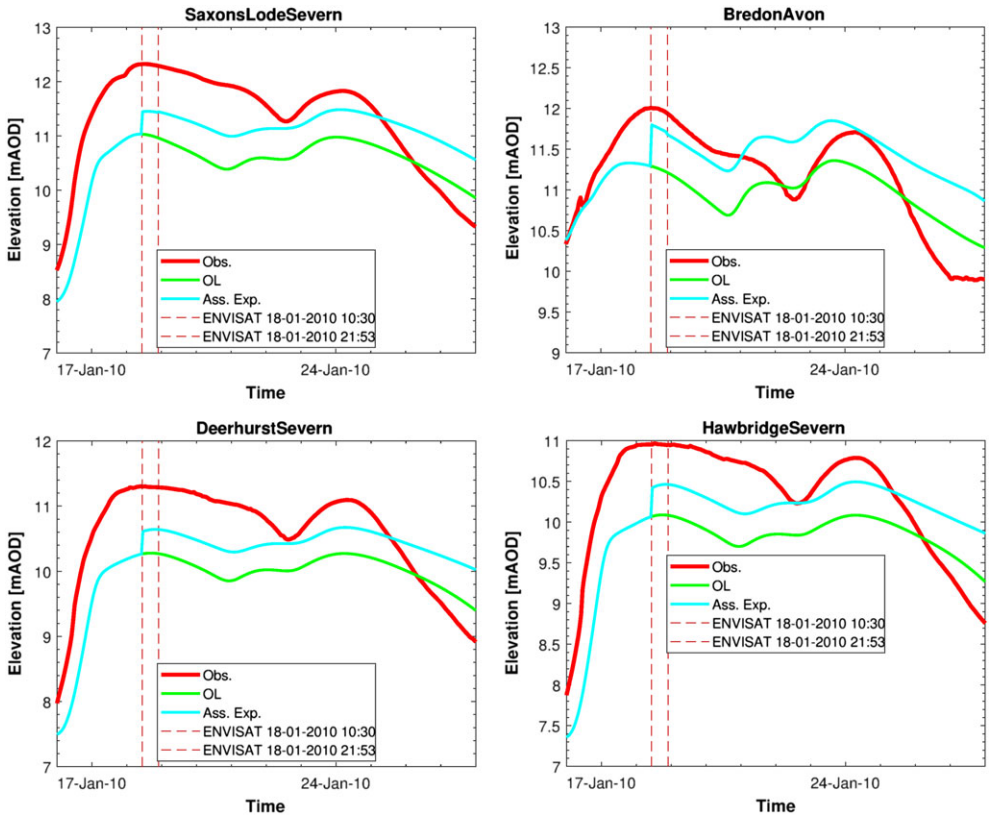

Figure 10. Water level time series at Saxons Lode, Bredon, Deerhurst, and Haw Bridge for January 2010 flood event: open- loop expectation (green), assimilation experiment expectation (cyan), and observations (red).

is not very sensitive to the assimilated variable—both the assimilation of water levels and flood extents derived from SAR images provide similar levels of performance, at least for the study area considered. One of the main advantages of assimilating flood extent rather than water levels is to avoid the nontrivial step of converting flood extent shorelines into water levels.

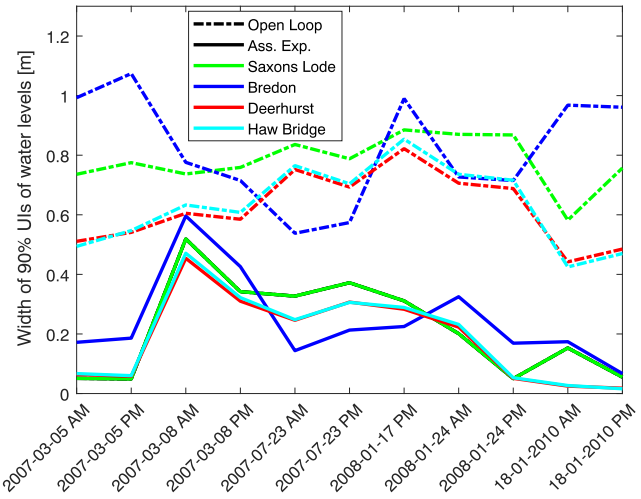

Figure 11. Widths of the 90% uncertainty intervals (UIs) of the water level forecasts produced by the assimilation and open-loop approaches. Results shown for Saxons Lode, Bredon, Deerhurst, and Haw Bridge.

5.5. Persistence of Assimilation Benefits

Figures 7–10 show that the assimilation improves model prediction not only at the assimilation time step but also at subsequent time steps. The improvement due to the assimilation lasts from a few hours to a few days, depending on the station and the flood event. However, the benefits of the update reduce over time and eventually the past assimilation event starts to have a *negative* effect. We attribute this behavior to the fact that the best-performing particles at the assimilation time step are not necessarily the best-performing particles long after this event, especially if the hydrological conditions change substantially (e.g., as the storm event hydrograph switches from its rising limb into its recession). Figure 10 illustrates this effect. In this figure, where the two satellite images were acquired at the flood peak, the improvement due to the assimilation persists for a few days, until the recession has become substantial.

The eventual reduction in forecast quality represents a limitation due to the currently rather long revisit time of satellites. Had other images been acquired sooner, a new assimilation step would have been carried out, and we hypothesize that the expectation would have most likely moved back toward the observation. Results in Figures 7–10 confirm this hypothesis: the recession is better reproduced by the expectation

Table 3
Average (Across Assimilation Events) Reduction of Error in Forecast Water Levels, as a Function of Prediction Lead Times

River Station	1 hr	2 hr	6 hr	12 hr	18 hr	24 hr	48 hr	72 hr	120 hr	168 hr
Saxons Lode	0.54	0.54	0.53	0.51	0.49	0.47	0.46	1.44	0.78	1.60
Bredon	0.29	0.29	0.28	0.24	0.23	0.24	1.06	1.36	2.60	7.29
Deerhurst	0.55	0.55	0.55	0.55	0.54	0.52	0.43	0.90	0.78	1.39
Haw Bridge	0.52	0.52	0.52	0.51	0.50	0.48	0.41	0.72	0.77	1.51

when images are assimilated during the recession. The overall picture provided by these figures is that the assimilation is most beneficial during the time steps following the data assimilation itself but continues to provide an increase in model performance as long as hydrological conditions do not change too rapidly.

To quantify the persistence of improvements due to assimilation steps, Table 3 reports the average reduction of forecast error in the water levels as a function of prediction lead times, for the Saxons Lode, Bredon, Deerhurst, and Haw Bridge gauging stations. A value of 0.54 in the first line, first column of Table 3 indicates that, on average, the assimilation of one SAR image reduces the 1-hr lead time forecast error by a factor of almost 2. Table 3 shows that, at the Saxons Lode, Deerhurst, and Haw Bridge gauging stations, all located on the River Severn, the forecast error is still reduced by a factor of almost 2 more than 2 days after the assimilation of a single image. In other words, the substantial positive effect of the assimilation of a single SAR image persists for more than 2 days. For Bredon on the River Avon, assimilation effects remain positive for 1 day but starts to become negative after 2 days. The shorter persistence of assimilation benefits on River Avon could be tentatively explained by the faster reaction of the Avon basin to rainfall inputs. This behavior can be seen in Figure 3, where the flood durations are clearly shorter in Evesham than in Bewdley. This finding lends further weight to the hypothesis that the persistence of the positive effect of the assimilation is closely linked to the reactivity of the basin, with positive effects persisting longer in slower basins. If this hypothesis is valid, satellite revisit time emerges as a major remaining limitation for using SAR image assimilation to improve flood extent forecasts. As satellite imagery becomes more accessible and frequent, for example, due to new initiatives such as the Sentinel missions of the Copernicus program (http://www.copernicus.eu/) that employ multiple satellites, we expect this limitation to gradually reduce. Using images from multiple satellites (including optical sensors during daytime and when cloud cover is limited) could also help in decreasing revisit times.

6. Conclusion

This study introduces a particle filter-based method for assimilating SAR-derived probabilistic flood maps into a flood forecasting chain composed of a conceptual hydrological model and a hydraulic model. The uncertainty in the flood maps derived from the SAR image is quantified using a Bayesian approach (Giustarini et al., 2016), which assigns all image pixels a probability of being flooded based on the observed backscatter value. The particles representing individual forecasts with different (perturbed) rainfall inputs are given weights computed based on the degree of match between forecast and observed flood extent at the assimilation time step.

Four flood events on the Rivers Severn and Avon (United Kingdom) were used as test cases. A set of 11 SAR flood images were processed for these events. The corresponding flood extent observations were assimilated sequentially. The performance of the assimilation approach was evaluated with reference to an independent flood validation map (derived from aerial photography), gauged streamflow time series at the upstream boundaries of the hydraulic model for the River Severn and the River Avon, and water level time series from four gauging stations at various locations across the Rivers Severn and Avon.

The following key conclusions can be drawn:

1. Satellite-derived flood extent observations represent a relevant source of information for improving flood forecasts, despite not being a state (diagnostic) variable of the model cascade. For example, compared to the open loop, the sequential assimilation of the 11 probabilistic flood maps shows a systematic and substantial improvement of the model cascade forecasts in terms of predicted water levels.
2. In most instances, flood extent assimilation improves forecasts at the assimilation time step. Using the observed flood validation map as a reference, we found that the predicted flood extents are slightly

improved by the data assimilation likely because the flood extent predicted by this open loop is already accurate. Furthermore, the predicted input streamflow is systematically well corrected by the assimilation for the River Avon while it is often overestimated for the River Severn. Forecast errors in the water levels are reduced, on average, by factors of 2 to 3, at the assimilation time steps.

3. The benefits of flood extent assimilation persist for several time steps after the assimilation event. Depending on the hydrological conditions, the improvements may persists from a few hours to a few days. However, improvements in performance wane some days after the assimilation event: the best-performing particles (simulations) at the assimilation time step do not maintain their high performance indefinitely. The degradation of forecast performance occurs more rapidly when hydrological conditions change substantially since the last assimilation time step.

Overall, the study provides consistent empirical evidence that the assimilation of flood extent data improves the performance of flood forecasting systems. The forecast performance in terms of water levels and streamflows is comparable to those obtained by García-Pintado et al. (2015) in a broadly similar study where SAR-derived water levels (rather than flood extent) were assimilated into a flood forecasting model cascade using a local ensemble transform Kalman filter. One main advantage of our method is to directly assimilate flood extent rather than water levels avoiding therefore the nontrivial step of converting flood extent shorelines into water levels and opening the door for a more straightforward application especially in near-real time. The main limitation of using satellite-derived flood extent to improve flood forecasting models appears to lie with the satellite revisit times. This limitation is expected to reduce in the future as the number of satellite images available in near-real-time increases.

Acknowledgments
The research reported herein was funded by the National Research fund of Luxembourg through the Mosquito (C15/SR/10380137) and Hydras+ (SR/8888433) projects. 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-Interim data set is freely available at http://apps.ecmwf.int/datasets/data/interim-full-daily/levtype=sfc/. The Envisat WSM images can be freely downloaded at https://earth.esa.int/.

References

Andreadis, K., & Schumann, G.-P. (2014). Estimating the impact of satellite observations on the predictability of large-scale hydraulic models. *Advances in Water Resources*, *73*, 44–54.

Bates, P., & de Roo, A. (2000). A simple raster-based model for flood inundation simulation. *Journal of Hydrology*, *236*, 54–77.

Bates, P. D., Wilson, M. D., Horrit, M. S., Mason, D. C., Holden, N., & Currie, A. (2006). Reach scale floodplain inundation dynamics observed using airborne synthetic aperture radar imagery: Data analysis and modelling. *Journal of Hydrology*, *328*, 306–318.

Bermúdez, M., Neal, J. C., Bates, P. D., Coxon, G., Freer, J. E., Cea, L., & Puertas, J. (2017). Quantifying local rainfall dynamics and uncertain boundary conditions into a nested regional-local flood modeling system. *Water Resources Research*, *53*, 2770–2785. https://doi.org/10.1002/2016WR019903

Biancamaria, S., Durand, M., Andreadis, K., Bates, P., Boone, A., Mognard, N., et al. (2011). Assimilation of virtual wide swath altimetry to improve Arctic River modeling.*Remote Sensing of Environment*, *115*, 373–381.

Chini, M., Hostache, R., Giustarini, L., & Matgen, P. (2017). A Hierarchical Split-Based Approach (HSBA) for parametric thresholding of SAR images: Flood inundation as a test case. *IEEE Transactions on Geoscience and Remote Sensing*, *55*(12), 6975–6988.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

de Almeida, G. A. M., Bates, P., Freer, J. E., & Souvignet, M. (2012). Improving the stability of a simple formulation of the shallow water equations for 2-D flood modeling. *Water Resources Research*, *48*, W05528. https://doi.org/10.1029/2011WR011570

de Lannoy, G. J. M., Houser, P. R., Pauwels, V. R. N., & Verhoest, N. E. C. (2006). Assessment of model uncertainty for soil moisture through ensemble verification. *Journal of Geophysical Research*, *111*, D10101. https://doi.org/10.1029/2005JD006367

de Leeuw, J., Methven, J., & Blackburn, M. (2015). Evaluation of ERA-Interim reanalysis precipitation products using England and Wales observations. *Quarterly Journal of the Royal Meteorological Society*, *141*(688), 798–806.

Dee, D., Uppala, S., Simmons, A., Berrisford, P., Poli, P., Kobayashi, S., et al. (2011). The ERA-Interim reanalysis: Configuration and performance of the data assimilation system.*Quarterly Journal of the Royal Meteorological Society*, *137*, 553–597.

Donaldson, R. J., Dyer, R. M., & Kraus, M. J. (1975). An objective evaluator of techniques for predicting severe weather events. In *Preprints, Ninth Conference on severe local storms, American Geological Society* (pp. 321–326). Norman,OK.

Evin, G., Thyer, M., Kavetski, D., McInerney, D., & Kuczera, G. (2014). Comparison of joint versus postprocessor approaches for hydrological uncertainty estimation accounting for error autocorrelation and heteroscedasticity. *Water Resources Research*, *50*, 2350–2375. https://doi.org/10.1002/2013WR014185

Fenicia, F., Kavetski, D., & Savenije, H. (2011). Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development. *Water Resources Research*, *47*, W11510. https://doi.org/10.1029/2010WR010174

García-Pintado, J., Mason, D. C., Dance, S. L., Cloke, H. L., Neal, J. C., Freer, J., & Bates, P. D. (2015). Satellite-supported flood forecasting in river networks: A real case study. *Journal of Hydrology*, *523*, 706–724.

Giustarini, L., Hostache, R., Kavetski, D., Chini, M., Corato, G., Schlaffer, S., & Matgen, P. (2016). Probabilistic flood mapping using synthetic aperture radar data. *IEEE Transactions on Geoscience and Remote Sensing*, *54*(12), 6958–6969.

Giustarini, L., Hostache, R., Matgen, P., Schumann, G. J. P., Bates, P. D., & Mason, D. C. (2013). A change detection approach to flood mapping in urban areas using TerraSAR-X. *IEEE Transactions on Geoscience and Remote Sensing*, *51*(4), 2417–2430.

Giustarini, L., Matgen, P., Hostache, R., Montanari, M., Plaza, D., Pauwels, V. R. N., et al. (2011). Assimilating SAR-derived water level data into a hydraulic model: A case study. *Hydrology and Earth System Sciences*, *15*, 2349–2365.

Giustarini, L., Vernieuwe, H., Verwaeren, J., Chini, M., Hostache, R., Matgen, P., et al. (2015). Accounting for image uncertainty in SAR-based flood mapping. *International Journal of Applied Earth Observation and Geoinformation*, *34*, 70–77.

Grimaldi, S., Li, Y., Pauwels, V. R. N., & Walker, J. P. (2016). Remote sensing-derived water extent and level to constrain hydraulic flood forecasting models: Opportunities and challenges. *Surveys in Geophysics*, *37*(5), 977–1034.

Hamon, W. (1963). Computation of direct runoff amounts from storm rainfall. *IAHS-AISH P*, *63*, 52–62.

Hostache, R., Lai, X., Monnier, J., & Puech, C. (2010). 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*(3-4), 257–268.

Hostache, R., Matgen, P., Giustarini, L., Teferle, F., Tailliez, C., Iffly, J.-F., & Corato, G. (2015). A drifting GPS buoy for retrieving effective riverbed bathymetry. *Journal of Hydrology*, *520*, 397–406.
Hostache, R., Matgen, P., Montanari, A., Montanari, M., Hoffmann, L., & Pfister, L. (2011). Propagation of uncertainties in coupled hydro-meteorological forecasting systems: A stochastic approach for the assessment of the total predictive uncertainty. *Atmospheric Research*, *100*(2–3), 263–274.
Hostache, R., Matgen, P., Schumann, G., Puech, C., Hoffmann, L., & Pfister, L. (2009). Water level estimation and reduction of hydraulic model calibration uncertainties using satellite SAR images of floods. *IEEE Transactions on Geoscience and Remote Sensing*, *47*(2), 431–441.
Hostache, R., Matgen, P., & Wagner, W. (2012). Change detection approaches for flood extent mapping: How to select the most adequate reference image from online archives? *International Journal of Applied Earth Observation and Geoinformation*, *19*(1), 205–213.
Jolliffe, I., & Stephenson, D. (2011). *Forecast verification: A practitioner's guide in atmospheric science* (2nd ed.). Chichester, West Sussex, England: John Wiley & Sons, Ltd. Publication.
Kavetski, D., & Fenicia, F. (2011). Elements of a flexible approach for conceptual hydrological modeling: 2. Application and experimental insights. *Water Resources Research*, *47*, W11511. https://doi.org/10.1029/2011WR010748
Lai, X., Liang, Q., Yesou, H., & Daillet, S. (2014). Variational assimilation of remotely sensed flood extents using a 2-D flood model. *Hydrology and Earth System Sciences*, *18*, 4325–4339.
Laurent, S., Hangen-Brodersen, C., Ehret, U., Meyer, I., Moritz, K., Vogelbacher, A., & Holle, F.-K. (2010). *Forecast uncertainties in the operational flood forecasting of the Bavarian Danube catchment* (pp. 367–387). Dordrecht, Netherlands: Springer.
Li, T., Sun, S., Sattar, T., & Corchado, J. (2014). Fight sample degeneracy and impoverishment in particle filters: A review of intelligent approaches. *Expert Systems with Applications*, *41*(8), 3944–3954.
Mason, D., Schumann, G.-P., Neal, J., Garcia-Pintado, J., & Bates, P. (2012). Automatic near real-time selection of flood water levels from high resolution synthetic aperture radar images for assimilation into hydraulic models: A case study. *Remote Sensing of Environment*, *124*, 705–716.
Matgen, P., Hostache, R., Schumann, G., Pfister, L., Hoffmann, L., & Savenije, H. (2011). Towards an automated SAR-based flood monitoring system: Lessons learned from two case studies. *Physics and Chemistry of the Earth, Part B*, *36*(7–8), 241–252.
Matgen, P., Montanari, M., Hostache, R., Pfister, L., Hoffmann, L., Plaza, D., et al. (2010). Towards the sequential assimilation of SAR-derived water stages into hydraulic models using the particle filter: Proof of concept. *Hydrology and Earth System Sciences*, *14*, 1773–1785.
McInerney, D., Thyer, M., Kavetski, D., Lerat, J., & Kuczera, G. (2017). Improving probabilistic prediction of daily streamflow by identifying Pareto optimal approaches for modeling heteroscedastic residual errors. *Water Resources Research*, *53*, 2199–2239. https://doi.org/10.1002/2016WR019168
Montanari, M., Hostache, R., Matgen, P., Schumann, G., Pfister, L., & Hoffmann, L. (2009). Calibration and sequential updating of a coupled hydrologic-hydraulic model using remote sensing-derived water stages. *Hydrology and Earth System Sciences*, *13*, 367–380.
Moradkhani, H. (2007). Hydrologic remote sensing and land surface data assimilation. *Sensors*, *8*, 2986–3004.
Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models. Part I: A discussion of principles. *Journal of Hydrology*, *10*(3), 282–290.
Neal, J. C., Atkinson, P. M., & Hutton, C. W. (2007). Flood inundation model updating using an ensemble Kalman filter and spatially distributed measurements. *Journal of Hydrology*, *336*, 401–415.
Neal, J., Schuman, G., Fewtrell, T., Budimir, M., Bates, P., & Mason, D. (2011). Evaluating a new LISFLOOD-FP formulation with data from the summer 2007 floods in Tewkesbury, UK. *Journal of Flood Risk Management*, *4*, 88–95.
Neal, J., Schumann, G., & Bates, P. (2012). A subgrid channel model for simulating river hydraulics and floodplain inundation over large and data sparse areas. *Water Resources Research*, *48*, W11506. https://doi.org/10.1029/2012WR012514
Pappenberger, F., Frodsham, K., Beven, K., Romanowicz, R., & Matgen, P. (2007). Fuzzy set approach to calibrating distributed flood inundation models using remote sensing observations. *Hydrology and Earth System Sciences*, *11*, 739–752.
Plaza, D. A., de Keyser, R., de Lannoy, G. J. M., Giustarini, L., Matgen, P., & Pauwels, V. R. N. (2012). The importance of parameter resampling for soil moisture data assimilation into hydrologic models using the particle filter. *Hydrology and Earth System Sciences*, *16*(2), 375–390.
Pulvirenti, L., Chini, M., & Pierdicca, N. (2016). Use of SAR data for detecting floodwater in urban and agricultural areas: The role of the interferometric coherence. *IEEE Transactions on Geoscience and Remote Sensing*, *54*(3), 1532–1544.
Renard, B. (2011). A Bayesian hierarchical approach to regional frequency analysis. *Water Resources Research*, *47*, W11513. https://doi.org/10.1029/2010WR010089
Renard, B., Kavetski, D., Kuczera, G., Thyer, M., & Franks, S. (2010). Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. *Water Resources Research*, *46*, W05521. https://doi.org/10.1029/2009WR008328
Revilla-Romero, B., Wanders, N., Burek, P., Salamon, P., & de Roo, A. (2016). Integrating remotely sensed surface water extent into continental scale hydrology. *Journal of Hydrology*, *543, Part B*, 659–670.
Schumann, G. J.-P., Neal, J. C., Voisin, N., Andreadis, K. M., Pappenberger, F., Phanthuwongpakdee, N., et al. (2013). A first large-scale flood inundation forecasting model. *Water Resources Research*, *49*, 6248–6257. https://doi.org/10.1002/wrcr.20521
Stehman, S. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, *62*(1), 77–89.
Wood, M., Hostache, R., Neal, J., Wagener, T., Giustarini, L., Chini, M., et al. (2016). 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.
Xu, X., Zhang, X., Fang, H., Lai, R., Zhang, Y., Huang, L., & Liu, X. (2017). A real-time probabilistic channel flood-forecasting model based on the Bayesian particle filter approach. *Environmental Modelling & Software*, *88*, 151–167.
Yamazaki, D., Ikeshima, D., Tawatari, R., Yamaguchi, T., O'Loughlin, F., Neal, J., et al. (2017). A high accuracy map of global terrain elevations. *Geophysical Research Letters*, *44*, 5844–5853. https://doi.org/10.1002/2017GL072874
Yatheendradas, S., Wagener, T., Gupta, H., Unkrich, C., Goodrich, D., Schaffner, M., & Stewart, A. (2008). Understanding uncertainty in distributed flash flood forecasting for semiarid regions. *Water Resources Research*, *44*, W05S19. https://doi.org/10.1029/2007WR005940
Yoon, Y., Durand, M., Merry, C. J., Clark, E. A., Andreadis, K. M., & Alsdorf, D. E. (2012). Estimating river bathymetry from data assimilation of synthetic SWOT measurements. *Journal of Hydrology*, *464–465*, 363–375.