

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Impacts of Spatiotemporal Gaps in Satellite Soil Moisture Data on Hydrological Data Assimilation

Khaled Mohammed ^{1,*}, Robert Leconte ¹ and Mélanie Trudel ¹

¹ Département de génie civil et de génie du bâtiment, Université de Sherbrooke, Sherbrooke, QC J1K 2R1, Canada

* Correspondence: khaled.mohammed@usherbrooke.ca

Abstract: Soil moisture modeling is necessary for many hydrometeorological and agricultural applications. One of the ways in which modeling of soil moisture (SM) can be improved is by assimilating SM observations to update the model states. Remotely sensed SM observations are prone to being riddled with data discontinuities, namely in the horizontal and vertical spatial, and temporal dimensions. A set of synthetic experiments were designed in this study to assess how much impact each of these individual components of spatiotemporal gaps can have on the modeling performance of SM as well as streamflow. Results show that not having root-zone SM estimates from satellite derived observations is most impactful in terms of modeling performance. Having temporal gaps and horizontal spatial gaps in the satellite SM data also impacts modeling performance, but to a lesser degree. Real-data experiments with the remotely sensed Soil Moisture Active Passive (SMAP) product generally brought improvements to the SM modeling performance in the upper soil layers, but not so much in the bottom soil layer. The updating of model SM states with observations also resulted in some improvements in the streamflow modeling performance during the synthetic experiments, but not during the real-data experiments.

Keywords: data assimilation; soil moisture; EnKF; SMAP; WRF-Hydro

1. Introduction

Soil moisture is a key variable in both hydrological and atmospheric modeling, as it influences the partitioning of water and energy fluxes between the land surface and the atmosphere. Accurate knowledge of soil moisture is important for many applications, such as numerical weather prediction, climate modeling, flood forecasting, drought monitoring and irrigation management [1,2]. While traditional in-situ soil moisture measurements may offer more accuracy, the global availability of these measurements is rather limited. Even in watersheds where soil moisture measurements are actively being taken, the network of point measurement locations is usually sparse, thereby unable to provide a proper representation of the spatial variability of soil moisture over larger areas [3].

In recent decades, multiple satellite platforms have started operating that help provide soil moisture estimates at a global scale, albeit at a relatively coarse spatial resolution and only for the uppermost layer of soil. Some of these remote sensing missions are even dedicated to soil moisture, namely the Soil Moisture and Ocean Salinity (SMOS) mission [4] and the Soil Moisture Active Passive (SMAP) mission [5]. Both SMOS and SMAP missions are using passive L-band (1.4 GHz) microwave radiometers to generate surface soil moisture estimates for approximately top 5 cm of soil, at a native spatial resolution of around 36 km and a temporal resolution of about 1-3 days [6]. There is also the upcoming NASA-ISRO Synthetic Aperture Radar (NISAR) mission [7] to be launched in 2023, which will use active L-band (1.26 GHz) backscatter measurements to provide global soil moisture estimates at a spatial resolution of 200 m every 6 days.

To get soil moisture estimates for both surface and subsurface soil layers (also known as root-zone soil moisture or profile soil moisture) with a higher spatiotemporal resolution, land surface models or distributed hydrological models can be used. However, these modeled estimates can have large uncertainties depending on the source of the forcing data, as well as the parameterization and structure of the model itself [8]. It has been found that some of these modeling errors can be reduced by integrating remotely sensed soil moisture information into the model through the process of data assimilation [9-12]. One of the things that can be done using data assimilation techniques is that internal model states can be updated with collocated observations in an optimal fashion at each observation time step, which may lead to better model prediction at the subsequent time steps.

One of the common practical issues that arise during the assimilation of remotely sensed soil moisture data is the question of how to address the spatiotemporal gaps within the data [13,14]. The spatial gaps may be split into two categories, gaps in the horizontal direction and in the vertical direction. In the horizontal direction, data discontinuity may be caused by the soil moisture retrieval algorithms not being able to accurately generate an estimate over some grid cells due to dense vegetation, hilly terrain, frozen soil, radio frequency interference etc. [15,16]. Also, especially in the case of data assimilation in models representing larger watersheds, it is probable that when a satellite passes over the watershed, its viewing angle does not cover the whole watershed, thus leaving a spatial gap in the soil moisture map for that overpass [17,18]. As for the gaps in the vertical direction, since passive microwave sensors are only effective at estimating soil moisture of the uppermost layer of soil, the root-zone layer of hydrological models cannot be updated directly using remotely sensed soil moisture estimations. Finally, the temporal gaps in remotely sensed soil moisture datasets are due to the geometry of the satellite orbits, which lead to longer revisit times over any specific location.

When sequential data assimilation methods are used, for example the Ensemble Kalman Filter (EnKF) [19], no additional steps are needed on account of the temporal gaps, because these methods pause the model simulation to make an update only when observations become available, and then resume the simulation until the next set of observations are available. Next, the issue of not having root-zone soil moisture observations during data assimilation can be dealt with in multiple ways. The simplest approach is to update the surface layer of the model with corresponding surface layer soil moisture estimates from remote sensing, and let the model propagate this added information downwards to the root-zone layer through the inherent model physics [20-22]. Results from these studies show that it is possible to improve soil moisture simulation of both layers by updating only the surface layer. Another approach is to apply an indirect update to the root-zone layer of the model, based on the update increment applied to the surface layer and the covariance between the soil moisture of different layers [23-25]. These studies show that the simulated soil moisture at varying depths can be improved by using this approach. When using the EnKF for example, this can be accomplished by including the root-zone layer in the state vector to be updated, in addition to including the surface layer.

The root-zone layer can even be directly updated along with directly updating the surface layer, if corresponding root-zone layer estimates based on remotely sensed surface layer measurements are generated prior to performing data assimilation (to be assimilated as 'observations'), using methods like the Soil Moisture Analytical Relationship (SMAR) or the exponential filter [26-27]. Some studies have even assimilated root-zone soil moisture data that have been previously generated using other land surface models, such as the publicly available H-SAF SM-DAS-2 product, produced by assimilating ASCAT soil moisture data into the ECMWF Land Data Assimilation System and distributed by the European Organization for the Exploitation of Meteorological Satellites (EUMETSAT) [28]. Even though the results from these studies too are positive, it is not quite clear which of the three abovementioned approaches are more effective given that usually only a single approach is employed in a single study.

Similar to the issue of whether and how to update the root-zone soil moisture, multiple approaches can be investigated when the remotely sensed soil moisture maps used in data assimilation are spatially incomplete in the horizontal direction. This may be less of a concern when lumped or semi-distributed models are used [29,30], because in this case, the irregularly shaped basin of the model is more likely to be larger in area than the spatial resolution of the observation data, thereby necessitating the averaging of multiple observation grid cells. So, it will still be possible to update the model state if observed data of a few grid cells are missing. But for finer resolution distributed hydrological models or land surface models whose grid spatial resolution is closer to that of the remotely sensed observations, this problem of having horizontal spatial discontinuity needs to be addressed before performing data assimilation. The simplest, and most commonly used approach is to update only the grid cells for which observations are available and let the remaining grid cells retain their model simulated values [31-33]. Alternatively, the soil moisture state of an unobserved grid cell may also be updated if other nearby grid cells have observations available and are correlated with the unobserved grid cell [25,34,35]. This approach is similar to the covariance-based approach described previously for vertical spatial gaps. It is also possible to estimate the soil moisture of unobserved grid cells prior to data assimilation using methods like geostatistical modeling, and then use these estimates as 'observations' for data assimilation [17].

All these methodologies that account for missing data in the horizontal spatial direction during data assimilation are shown to be able to improve modeling performance when compared to modeling performance without data assimilation. However, data assimilation studies are hard to find in the literature where the impacts of having spatial data discontinuities in both the horizontal and vertical directions of remotely sensed soil moisture, as well as temporal data gaps, are assessed within the same modeling framework. Looking at all these aspects of data discontinuity using the same model, datasets and study area will make it easier to compare which kind of data gap is more detrimental to modeling performance, and which kind of modeling approach is better suited to circumvent this problem of missing data. This study was therefore aimed at adding to the existing literature on this topic by carrying out multiple synthetic data assimilation experiments using the EnKF algorithm and the WRF-Hydro modeling system, to investigate how the ability of the model to simulate soil moisture may be affected by having spatiotemporal gaps in the observation data. To do this, spatiotemporal discontinuity information was extracted from SMAP datasets and then imposed on synthetically generated observation datasets to mimic the conditions found in actual remotely sensed datasets. The impact of these different soil moisture assimilation scenarios on the model's ability to accurately simulate streamflow was also investigated. Lastly, the data assimilation experiments were repeated with SMAP data as observations instead of synthetic observations.

2. Materials and Methods

2.1. Study Area

The 721-km long Susquehanna River is situated in the Northeastern United States, and its 71,432 km² drainage basin covers parts of the New York (NY), Pennsylvania (PA), and Maryland (MD) states [36]. With a mean annual flow of around 1,100 m³/s, it drains into the Atlantic Ocean through the Chesapeake Bay, and accounts for about 50 % of the freshwater inputs of the bay [37].

The Susquehanna River basin has a humid continental climate with a mean annual temperature of 9.7 °C and mean annual precipitation of 980 mm. The warmest months are June to August, and coldest months are January and February. As for precipitation, the highest amounts are seen in May to July, and lowest amounts in January and February. During winter months, snowfall occurrences are more prominent in the northern portion

of the watershed. And in the summer, higher temperatures are common along with locally intense convective storms. While in the late summer to fall seasons, the watershed becomes prone to floods brought about by tropical storms and hurricanes originating in both the Atlantic Ocean and the Gulf of Mexico [36,37]. Droughts have also affected the watershed in the past, although with a lesser frequency than floods [38].

The physiography of the watershed includes high plateaus, mountains, valleys, and ridges, and the soils of the watershed are predominantly silt loam and loam. The most common land cover category is forest (63 %), followed by cropland (19 %), pasture (7 %) and urban development (9 %) [36]. There are multiple large water infrastructures near the downstream end of the watershed, namely the Safe Harbor Dam, Holtwood Dam, and the Conowingo Dam. To avoid the complexities that would arise if these dams were incorporated into the hydrological model, only the drainage area upstream of Harrisburg, PA was considered for this study, whose area is about 60,600 km² (Figure 1).

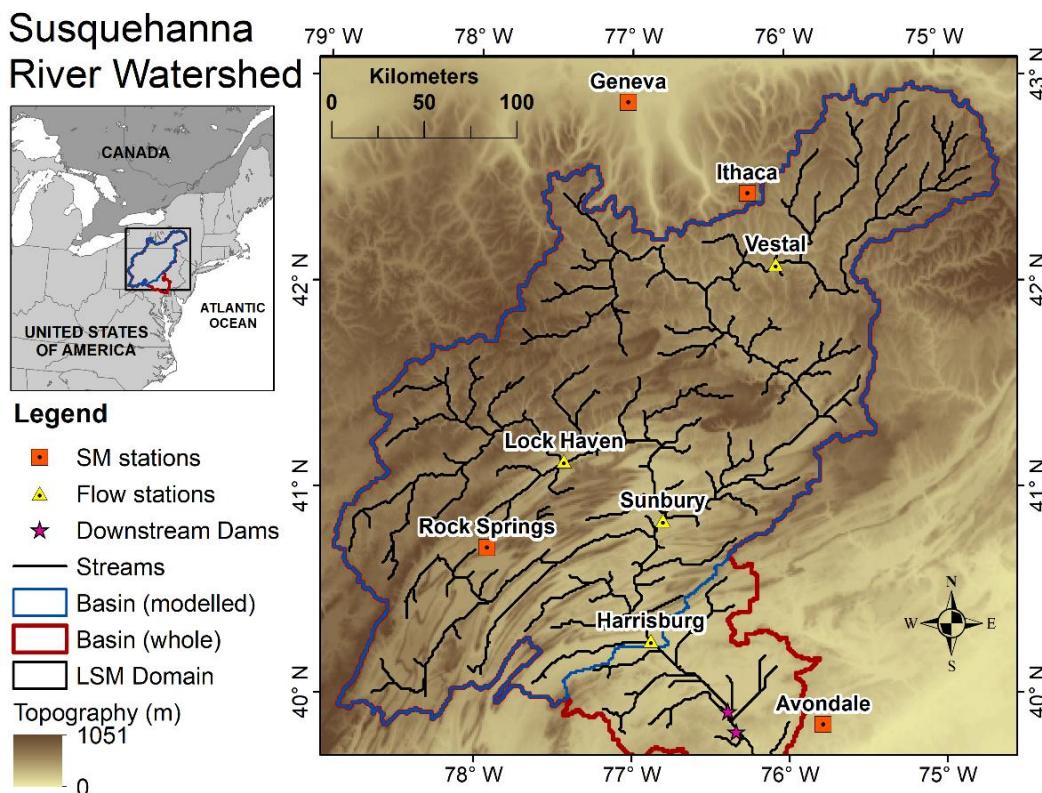


Figure 1. Topography of the Susquehanna River watershed and the locations of the in-situ streamflow and soil moisture measuring stations. The complete watershed is shown in red, and the modelled portion of the watershed is shown in blue which excludes the large downstream dams.

For the purposes of model parameter calibration and data assimilation performance assessment, in-situ measurements of soil moisture at different soil depths were collected from the International Soil Moisture Network (ISMN) database [3]. A total of four stations were selected, two of which (Geneva, NY and Rock Springs, PA) are part of the Soil Climate Analysis Network (SCAN) [39] and the other two (Ithaca, NY and Avondale, PA) are part of the U.S. Climate Reference Network (USCRN) [40]. All these measurement stations use Stevens HydraProbe sensors to measure soil moisture at 5, 10, 20, 50 and 100 cm soil depths. It should be noted that three out of these four in-situ stations are located outside of the Susquehanna River watershed's boundary. This did not pose any problems because they are still located within the land surface model (LSM) domain as shown in Figure 1. The LSM used for this study utilizes square grids to discretize a larger rectangular domain, and therefore all the grid cells within the LSM domain were included in the calibration process and provided soil moisture estimates during the assimilation

experiments. However, when quantifying basin-averaged soil moisture modelling performance later in the Results and Discussion section, grid cells outside of the modelled basin boundary were masked out.

In-situ measurements of streamflow were also collected for model parameter calibration, data assimilation performance assessment, and to help select the timeframes of model calibration/validation and assimilation experiments. The locations of these selected United States Geological Survey (USGS) measurement locations are shown in Figure 1, which are in Vestal, NY, Lock Haven, PA, Sunbury, PA, and Harrisburg, PA.

2.2. *Hydrological Modeling*

The WRF-Hydro modeling system [41], developed by the National Center for Atmospheric Research (NCAR), is a fully distributed system that consists of multiple modules, namely a column land surface module, surface overland and saturated subsurface lateral flow modules, channel routing and reservoir routing modules, and a conceptual baseflow module. The Noah-MP option [42] was selected for the column land surface module. Soil moisture was simulated in Noah-MP for four soil layers, having a total thickness of 200 cm. The thicknesses of the individual layers were defined to be 5, 35, 60 and 100 cm. The thickness of the top layer was chosen to be 5 cm to be compatible with remotely sensed soil moisture estimates. These thicknesses were uniform throughout the model domain.

It is possible to keep some of the other WRF-Hydro modules switched off, but all of them were activated for this study so that both soil moisture and streamflow are simulated. This way, the impact of updating soil moisture values on streamflow generation can be investigated, which is caused by the propagation of assimilated information through the lateral surface and subsurface terrain routing, and channel routing of water. The subsurface runoff in WRF-Hydro uses a quasi-3D flow equation as implemented in the Distributed Hydrology Soil Vegetation Model (DHSVM), the surface runoff calculation uses a fully unsteady diffusive wave formulation, and a one-dimensional, variable time-stepping, diffusive wave gridded routing method was used for channel routing. Readers are referred to [41] for complete technical descriptions of WRF-Hydro.

To set up the model domain for the study area, soil texture information was collected from the 16-category hybrid STATSGO/FAO soil texture map produced by NCAR, the land use information was that of the 20-category IGBP-modified MODIS land use dataset, and the 30 arc-second version of the HydroSHEDS data was used as elevation information. A 5-km grid size was chosen for the column land surface module and a 1-km grid size was used for the terrain routing and channel routing modules. The decision to use these horizontal resolutions was made based on a trade-off between satisfactory model performance and trying not to overwhelm available computing resources, as WRF-Hydro is a computationally intensive modeling system. As for the time steps of the different model components, the land surface module was run hourly, and both the terrain and channel routing modules had a time step of one minute.

WRF-Hydro was initially developed for easy coupling with the Weather Research and Forecasting (WRF) atmospheric modeling system [43], but it can also be used in an offline mode, i.e., not coupled with any atmospheric model. In this case, the meteorological forcings from any independent source need to be provided to WRF-Hydro, which are the incoming shortwave and longwave radiation, specific humidity, air temperature, surface pressure, near surface wind in two orthogonal directions, and liquid water precipitation rate. For this study, these meteorological data were sourced from the ERA5 reanalysis product [44]. ERA5 datasets are generated by the European Centre for Medium-Range Weather Forecasts (ECMWF) using their Integrated Forecast System (IFS), which combines model data with observations through a 4D-Var data assimilation scheme. The 10-member ensemble version of ERA5 was used for this study, which has a temporal resolution of 3 hours and a spatial resolution of 0.5 decimal degrees.

Prior to using the WRF-Hydro model for data assimilation experiments, the model was calibrated against in-situ soil moisture and streamflow data using the Pareto Archived Dynamically Dimensioned Search (PADDS) algorithm [45]. PADDS is the multi-objective version of DDS [46], which is a stochastic and heuristic global-search optimization algorithm. For single-objective problems, DDS starts searching for the optimum globally, and then narrows its search to local regions when the user-specified maximum number of model iterations is approaching. In the case of multi-objective calibration, PADDS tries to define the Pareto front between the objective functions, on which improving one objective function deteriorates the other(s). For this study, the hypervolume contribution selection metric was used in PADDS, and the neighborhood perturbation factor was set to the recommended default value of 0.2.

2.3. Remotely Sensed Soil Moisture

SMAP soil moisture data was used in this study for direct assimilation during the real-data experiments, as well as for using the spatiotemporal gap patterns of this dataset during the synthetic experiments. Specifically, the version 4 of SMAP Enhanced L2 Radiometer Half-Orbit 9 km EASE-Grid Soil Moisture was used. This dataset, provided by the National Aeronautics and Space Administration (NASA), has a 13-hour latency, and has the unit of volumetric soil moisture, which is the same as the WRF-Hydro model soil moisture outputs. Data from only the descending pass of the satellite orbits were used in this study, which has a retrieval time of 6 am (local time). The soil moisture retrieval algorithm assumes the surface soil, vegetation, and air to be in thermal equilibrium in the early morning, and so the morning retrievals are expected to be of slightly better quality.

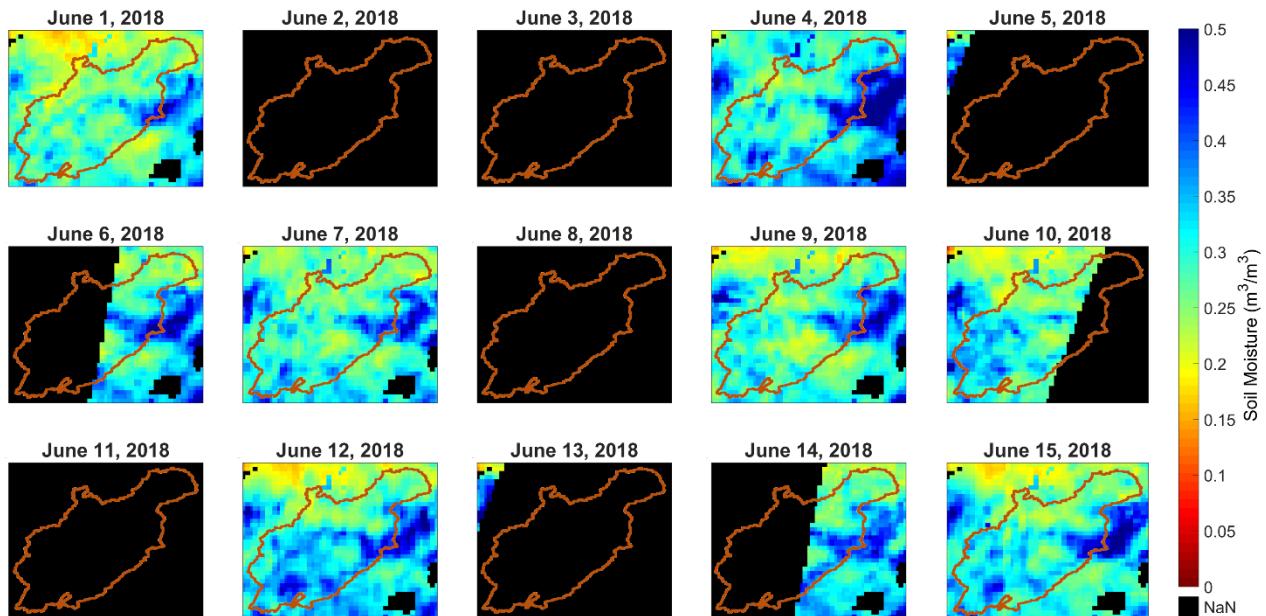


Figure 2. Spatiotemporal gaps in SMAP data over the study area from 1st June 2018 to 15th June 2018.

All the assimilation experiments were conducted for the summer-fall months to avoid the winter months where frozen soil and snow cover make it challenging to estimate soil moisture from satellites, and to avoid the spring months when soil moisture assimilation may have a lesser impact on snowmelt-driven streamflow. Therefore, filtering the SMAP data for snow covered or frozen soil conditions was not necessary. To illustrate the patterns of spatiotemporal gaps in the SMAP dataset, soil moisture maps over the study domain are presented in Figure 2 for the first 15 days of June 2018. When horizontal spatial coverage within the watershed boundary is considered, some days there is full coverage, some days have zero coverage, and some days have partial coverage. Within the experiment months of June-October, about 38 % of days have full

coverage, another 38 % have zero coverage, and the remaining 24 % of days have partial coverage.

2.4. Data Assimilation

2.4.1. Ensemble Kalman Filter (EnKF)

The EnKF was chosen as the preferred data assimilation method in this study because it was found to be the predominant method of choice in the existing soil moisture assimilation literature. This is so because the EnKF is well suited for high dimensional nonlinear problems, is computationally efficient and easy to implement albeit with a limitation of Gaussian error assumption [47]. Alternatives to EnKF include the Particle Filter (PF) which does not require any assumptions of Gaussian error and sometimes even slightly outperform EnKF. However, it is possible for PF to underperform EnKF and be generally comparable to EnKF at other times, all the while carrying a larger computational burden, leading to fewer users [48,49].

EnKF is a Monte Carlo based approach that allows model uncertainties to be estimated from a model ensemble spread that is assumed to be large enough to represent the true uncertainty of the simulation [50]. It works in two steps: a forecast step, and an analysis step. In the forecast step, ensembles are generated by either perturbing the forcing data, model states, model parameters, or any combination between them. Then the model is propagated to a future time step where observations are available. In the analysis step, uncertainty between the ensembles of model forecast and the observation is compared. If it is a state-updating scheme, then the model state at time t will be updated using the following equation, which gives more weight to the component between the model forecast and observation that has the least uncertainty:

$$x_t^a = x_t^b + K_t(y_t - H_t x_t^b), \quad (1)$$

where x_t^a is the updated state (a.k.a. analysis), x_t^b is the forecast state (a.k.a. background), K_t is the Kalman gain, y_t is the observation, and H_t is the observation operator. The analysis, background and observation operator took different forms in this study (either scalar, vector, or matrix) depending on the different scenarios and will be discussed in the subsequent section. The Kalman gain, which acts as a weighted average between the model forecast and the observation, is computed as follows:

$$K_t = P_t^b H_t^{-1} (H_t P_t^b H_t^{-1} + R_t)^{-1}, \quad (2)$$

where R_t is the observation matrix, and P_t^b is the model covariance matrix, which is calculated as:

$$P_t^b = \frac{1}{N-1} (x_t^b - \bar{x}_t^b) (x_t^b - \bar{x}_t^b)^{-1}, \quad (3)$$

where N is the number of ensemble members and \bar{x}_t^b is the ensemble mean of the background.

The model forecast uncertainties can be thought of to be primarily coming from the forcing data, model parameter and model structure. The variability within the 10-member ensemble of the ERA5 forcing data are assumed to contain a sufficient amount of forcing uncertainty. The remaining two categories of uncertainty, model parameter and structure, are represented jointly in this study by directly perturbing the soil moisture in the model background. All the experiments in this study were conducted with 24 ensemble members generated by combining the different sources of uncertainties together as follows.

First, nine out of the 10 ensemble members of ERA5 forcings (the tenth member was set aside to be used as synthetic truth, which is explained further in the next section) were duplicated into 24 members, of which only nine are unique members. Using these 24 forcing members to run the model 24 times provided a 24-member ensemble of soil moisture states, of which again only nine are unique members, and the rest are duplicates of those nine. Next, the soil moisture states of each of these 24 members were perturbed with unique random noise, leading to a model background of 24 unique members. As

perturbations, temporally and spatially uncorrelated additive Gaussian noise was used having a zero mean and standard deviation of $0.05 \text{ m}^3/\text{m}^3$. This value was decided upon after multiple trials to determine what magnitude of noise leads to maximum post-assimilation improvement in terms of root-mean-squared-error (RMSE).

For the synthetic experiments, the soil moisture observation uncertainty was estimated using a temporally and spatially uncorrelated zero mean Gaussian distribution having a standard deviation of $0.04 \text{ m}^3/\text{m}^3$, mirroring the baseline science requirements of the SMAP mission [5]. And for the real-data experiments, observation uncertainty was estimated using the triple collocation analysis method, which is discussed in Section 2.4.3. Also, all assimilation experiments had a lead time of 24 hours i.e., the model propagations were paused at 6 am (local time) every day to calculate the analysis only if observations were available, thereby coinciding with the time of SMAP data retrievals. Experimental model runs were commenced separately for 2018 (which had a relatively wet summer-fall season) and 2020 (a drier than average summer-fall season), with each simulation spanning from June 1st to October 31st of the corresponding year.

2.4.2. Synthetic Experiments

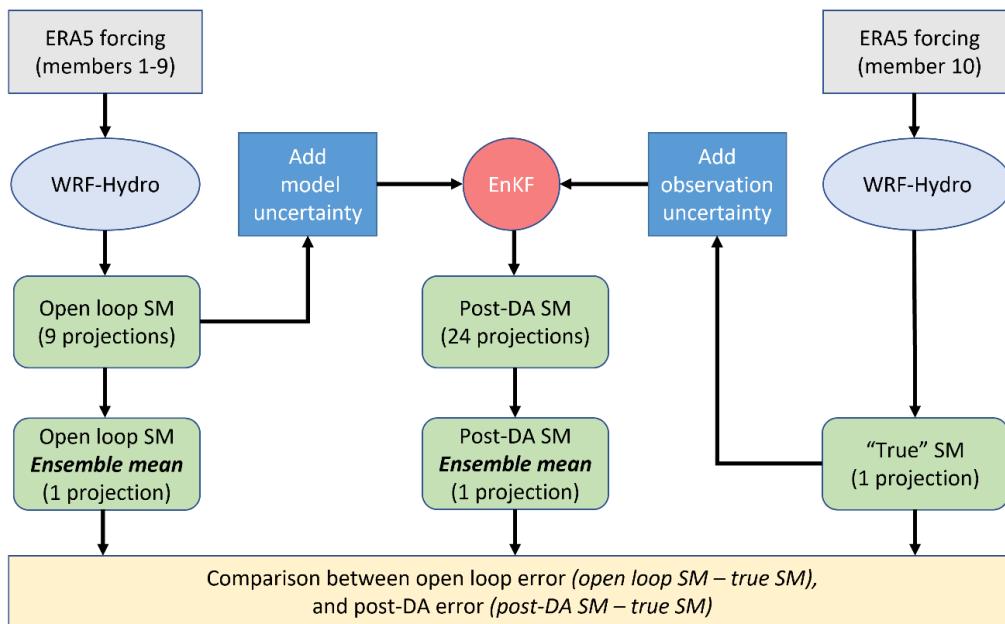


Figure 3. Flowchart describing the design of the synthetic experiments.

A set of synthetic experiments, a.k.a. Observation System Simulation Experiments (OSSE), were designed for this study, where assimilation was performed not with soil moisture observations from the real world, but rather with synthetically generated observations. The overall methodology of generating these synthetic observations and applying them in the EnKF is presented in Figure 3. First, the WRF-Hydro model was run with forcings from nine out of the 10 ERA5 ensemble members, generating nine projections of soil moisture to be used as the 'open loop', i.e., what happens if the model is run without any data assimilation. Next, the WRF-Hydro model was run with the remaining tenth member of the ERA5 forcings, generating another projection of soil moisture, which was considered to be the 'synthetic truth'. The goal of all the synthetic experiments was to apply EnKF in order to guide the open loop simulations closer to this synthetic truth.

In the natural world, these true values are never accurately known, because all observations are always prone to some type of error such as instrument error or operator error. Therefore, random Gaussian noise was added to this synthetic truth soil moisture (as described in the previous section) to prepare it for using as a synthetic observation for

the assimilation experiments. The EnKF algorithm then compared between the uncertainties of this synthetic observation and the model forecast uncertainties (as described in the previous section) to determine the final post-assimilation soil moisture. It should be pointed out that both the open loop and post-assimilation simulations have multiple ensemble members (nine and 24 respectively) while the synthetic truth has only a single member. Therefore, to consistently evaluate between the three types of data, only the ensemble mean of the open loop and post-assimilation soil moisture data was considered.

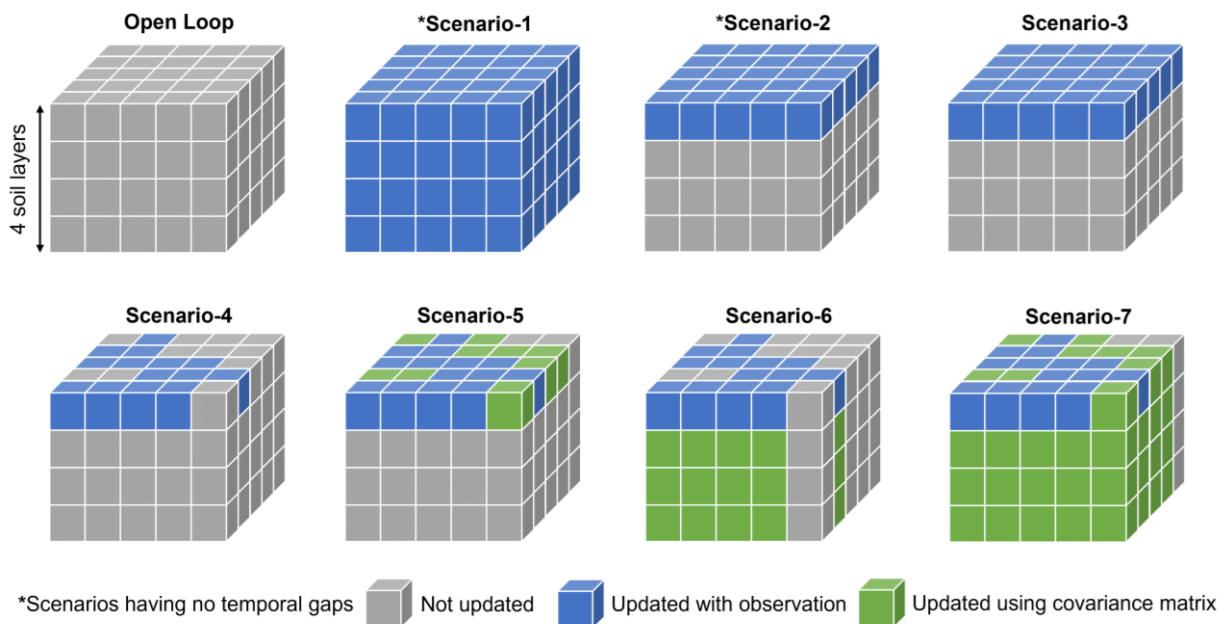


Figure 4. Different spatial configurations of soil moisture assimilation tested in this study.

A total of seven different scenarios were tested to investigate the impacts of spatiotemporal gaps in observed data on assimilation performance, by comparing these scenarios with the open loop (no assimilation) model runs. The spatial configurations of these scenarios are visualized in Figure 4. Scenario-1 is the most ideal situation, where observed data is available for all model grid cells in all soil layers. Also, data is available every day, i.e., there are no temporal gaps. Scenario-2 also has no temporal gaps and has observed data available for all model grid cells, but only for the topmost soil layer. This scenario is more realistic than Scenario-1 because satellite sensors are unable to detect soil moisture of root-zone layers. Scenario-3 is spatially similar to Scenario-2, the only difference being that temporal gaps are introduced here. This was done by not performing assimilation on any model grid cell during the days on which there are no SMAP data over the entire model domain. This scenario is intended to isolate the impacts of having temporal gaps in data, from the impacts of spatial gaps.

From Scenario-4 onwards, both spatial and temporal gaps available in the SMAP dataset were superimposed on the synthetic observations, by assuming that grid cells which do not have SMAP observations on a particular day, does not have synthetic observations either. In Scenario-4 only those grid cells were updated with synthetic observations, for which SMAP data is available in the corresponding day. In Scenario-5 on the other hand, some of the top layer grid cells without available observations were updated as well. This was accomplished using the following technique. In Scenarios 1-4, assimilation was point-based or zero-dimensional i.e., each grid cell of each soil layer was updated separately and independently. In other words, the background matrix x_t^b and observation matrix y_t contained the soil moisture value of only one grid cell, and therefore

the observation operator took the form of a scalar, $H_t = 1$. In Scenario-5 however, assimilation was performed two-dimensionally (in two of the horizontal directions).

For example, if a grid cell has no observation but three of its surrounding grid cells do, then the background matrix will have four components: soil moisture of the grid cell to be updated and those of the surrounding three grid cells. The observation matrix will have only three components, as the grid cell to be updated does not have any observations. And the observation operator will be the 3x4 matrix, $H_t = [0 \ 1 \ 0 \ 0; 0 \ 0 \ 1 \ 0; 0 \ 0 \ 0 \ 1]$. By setting up the matrices this way, the grid cell with missing observation will be updated based on the covariance between it and the surrounding three grid cells as calculated in Equation (3). It was determined through trial-and-error that the optimum number of surrounding grid cells to be utilized for this approach is a maximum of five grid cells having the same land use category and located within a radius of 25 km from the grid cell to be updated. Putting a constraint on the location of the utilized surrounding grid cells in this way, a.k.a. localization, helps prevent grid cells to be updated through spurious correlations between faraway grid cells [51]. It should be noted this led to some of the top layer grid cells not being updated as they did not have any available observations within a 25 km radius.

In Scenario-6, only for the grid cells that have observations available, a one-dimensional assimilation (in the vertical direction) approach was taken to update all four soil layers based on the correlation between the soil moisture of the top and bottom layers. In this case, the background matrix contained four components: soil moisture of the four soil layers, and the observation matrix had only one component: the observation at the top layer. The observation operator becomes the following vector in this case, $H_t = [1 \ 0 \ 0 \ 0]$. Finally, Scenario-7 combines both the approaches of Scenario-5 and Scenario-6. For top layer grid cells in which observations were available, one-dimensional assimilation was carried out on all four soil layers of those grid cells the same way as described for Scenario-6. Otherwise, for top layer grid cells that did not have corresponding observations but some of its surrounding grid cells within 25 km did have observations, a three-dimensional approach (two horizontal directions and one vertical direction) was undertaken. The background matrix in this case will have (assuming only two surrounding top layer grid cells have observations in this example) six components: soil moisture of all four layers of the grid cell to be updated, and surface soil moisture of the two surrounding grid cells that have observations. The observation matrix will have two components: observation over the two selected surrounding grid cells. And the observation operator will be the 2x6 matrix, $H_t = [0 \ 0 \ 0 \ 0 \ 1 \ 0; 0 \ 0 \ 0 \ 0 \ 0 \ 1]$.

To summarize the scenarios, Scenario-1 represents the configuration which may theoretically provide the maximum benefit from data assimilation (in terms of improvement of soil moisture simulation accuracy compared to the open loop), because synthetic observations are available over all layers in all grid cells to guide the model towards the synthetic truth. Scenarios 2-4 represent the gradual loss of this benefit due to not having observations for bottom layers, and the introduction of spatiotemporal gaps in the top layer observations. Finally, Scenarios 5-7 are meant to represent how much of these lost benefits in Scenarios 2-4 relative to the hypothetical optimum of Scenario-1, can be recovered through the application of the abovementioned technique of utilizing the covariance matrix of EnKF.

2.4.3. Real-data Experiments

In addition to all these synthetic experiments where the WRF-Hydro model was updated with synthetic observations of soil moisture with a goal of reaching closer to the synthetic truth, some real-data experiments, a.k.a. Observation System Experiments (OSE) were also performed where the WRF-Hydro model was updated with SMAP observations with a goal of reaching closer to the in-situ soil moisture observations. Some of the major differences between the synthetic and real-data experiments are as follows. In the case of real-data experiments, only Scenarios 4-7 were performed, because the observed data of all layers over all domain grid cells for every day which are required for Scenarios 1-3, is

not possible to be provided by SMAP. Also, during the synthetic experiments, assimilation performance was measured for all model grid cells because of the availability of synthetic truth data over all model grid cells. But in the case of real-data experiments, the assimilation performance was calculated only for four grid cells over which in-situ data was available.

Furthermore, the SMAP data was taken through a few pre-processing steps before being used for data assimilation purposes. First, the 9 km horizontal resolution of the dataset was resampled into 5 km using the nearest neighbor method, to match with the resolution of the model. Next, the SMAP data was rescaled to the model space (a.k.a. bias correction) by using a cumulative distribution function (CDF) matching method [52]. This is intended to correct any climatological differences between the SMAP data and the modelled soil moisture, because the EnKF can adjust random errors only and not systematic biases [53,54].

Lastly, the uncertainty information of the SMAP data, which is required for EnKF, was estimated using the triple collocation analysis method [55]. The calculation of error variance of the SMAP observations by triple collocation analysis requires collocated data from three independent datasets or triplets. The two other datasets used in this study in addition to SMAP were the ensemble mean of open loop model simulations and the SMOS L2 soil moisture product. The open loop data was chosen as the reference data for triple collocation analysis, meaning that the errors will be estimated in the model space. The benefit of choosing SMOS as the third dataset is that SMAP and SMOS data are both distributed in the same units of volumetric soil moisture, and so an additional step of unit conversion could be avoided. SMOS data also satisfy the independence requirement of triple collocation, as SMAP and SMOS retrievals are based on different algorithms applied on information from different satellites. And like the SMAP data, the SMOS data too was rescaled to the reference data using CDF matching to ensure that the errors of the triplets were unbiased relative to each other. The errors for the SMAP data were estimated for each grid cell separately, and the errors were assumed to be time-invariant considering the limited seasonal nature of this study.

3. Results and Discussion

3.1. Model Calibration and Validation

Calibration of the model parameters are not essential for conducting synthetic experiments because all the observations are synthetically generated using the model itself. But for setting up the real-data experiments where real observations will be assimilated into the model to get the model to behave more like the real world, it helps if the model parameters are tuned so that the model simulated soil moisture is as close to the in-situ soil moisture as possible. The aim of data assimilation would then be to improve the simulations further than what parameter calibration alone could achieve.

Calibrating the model for such a large watershed with in-situ soil moisture information from only four locations is a challenge, and so a multi-objective calibration approach was chosen to increase the robustness of the calibration process, where the model parameters were calibrated against both in-situ soil moisture and streamflow observations. An additional benefit of calibrating against streamflow is that the impact of soil moisture assimilation on the generation of streamflow could then be better analyzed. To reduce the risks of equifinality by calibrating a smaller set of model parameters, a sensitivity analysis was first performed on different model parameters of WRF-Hydro. This led to the choosing of the following four most influential parameters for calibration: soil porosity (MAXSMC), deep drainage coefficient (SLOPE), lateral saturated soil hydraulic conductivity (LKSATFAC), and slope of conductance to photosynthesis relationship (MP). These four parameters were automatically calibrated using 400 iterations of the PADDS algorithm. A two-year spin-up period was added to the different calibration and validation periods as mentioned above.

At the outlet of the modelled portion of the watershed, Harrisburg, it was found that in the decade to 2020, the mean summer-fall flows (averaging all the mean daily flows within June to October) ranged from 245 m³/s to 1,699 m³/s, with a 10-year average of 680 m³/s. Only two of the years within the decade stand out as wet outliers, 2011 (annual maximum daily flow of 16,357 m³/s) and 2018 (annual maximum daily flow of 8,603 m³/s). In the wettest year of 2011, Tropical Storm Lee resulted in a 100-year return period flood. As soil moisture data assimilation has the potential to improve peak-flow simulation depending on how well the antecedent soil wetness is represented, it was decided that one of the wet years would be used for the assimilation experiments.

Table 1. Calibration and validation performance of the model in terms of correlation coefficient (R), unbiased root-mean-squared-error (ubRMSE), bias, and Nash-Sutcliffe efficiency (NSE).

Calibration (2011/wet year)			Validation (2018/wet year)			Validation (2020/dry year)					
Soil Moisture		Flow	Soil Moisture		Flow	Soil Moisture		Flow			
R	ubRMSE (m ³ /m ³)	Bias (m ³ /m ³)	NSE	R	ubRMSE (m ³ /m ³)	Bias (m ³ /m ³)	NSE	R	ubRMSE (m ³ /m ³)	Bias (m ³ /m ³)	NSE
Layer 1	0.81	0.056	0.092		0.68	0.041	0.071		0.57	0.053	0.062
Layer 2	0.84	0.048	0.103	0.78	0.68	0.037	0.064	0.77	0.75	0.042	0.079
Layer 3	0.75	0.026	0.062		0.63	0.027	0.054		0.64	0.021	0.044

Since any year prior to 2015 could not be used for the data assimilation experiments (the SMAP data used in this study is available from 2015), the wettest year of 2011 was chosen for model calibration and the second wettest year of 2018 was chosen for both model validation and data assimilation experiments. The calibration/validation performances are presented in Table 1. The modeling performance of soil moisture compared against the in-situ observations are presented in terms of correlation coefficient (R), unbiased root-mean-squared-error (ubRMSE), and bias. Also, the values presented in Table 1 are the averaged values of four grids where in-situ data is available. The modeling performance of streamflow is presented in terms of Nash-Sutcliffe efficiency (NSE) calculated at Harrisburg, the basin outlet. Even though the model was calibrated for a wet year, it was found that the validation performance on a dry year (2020) was comparable to the validation performance on a wet year (2018). Thereafter it was decided to run the assimilation experiments on a dry year (2020) as well.

3.2. Synthetic Experiments

Spatially averaged (over the watershed) and temporally averaged (over June-October) improvements brought about by data assimilation are presented in Figure 5. Here, improvements in three evaluation metrics (ubRMSE, R, bias) are shown in terms of the difference between the values of those metrics during open loop model runs and the values of those metrics after assimilation. As intuitive, Scenario-1, where soil moisture of all grid cells in all soil layers were updated with respective observations every day, has the largest improvements out of all seven. In 2018, the layer-averaged improvements (defined as the average improvement of all four soil layers) for Scenario-1 are 0.0016 m³/m³ (ubRMSE), 0.025 (R) and 0.001 m³/m³ (bias). For the rest of the scenarios, the layer-averaged improvements are presented here in the following text as a percentage of the maximum Scenario-1 layer-averaged improvements, instead of in their original units as shown in Figure 5.

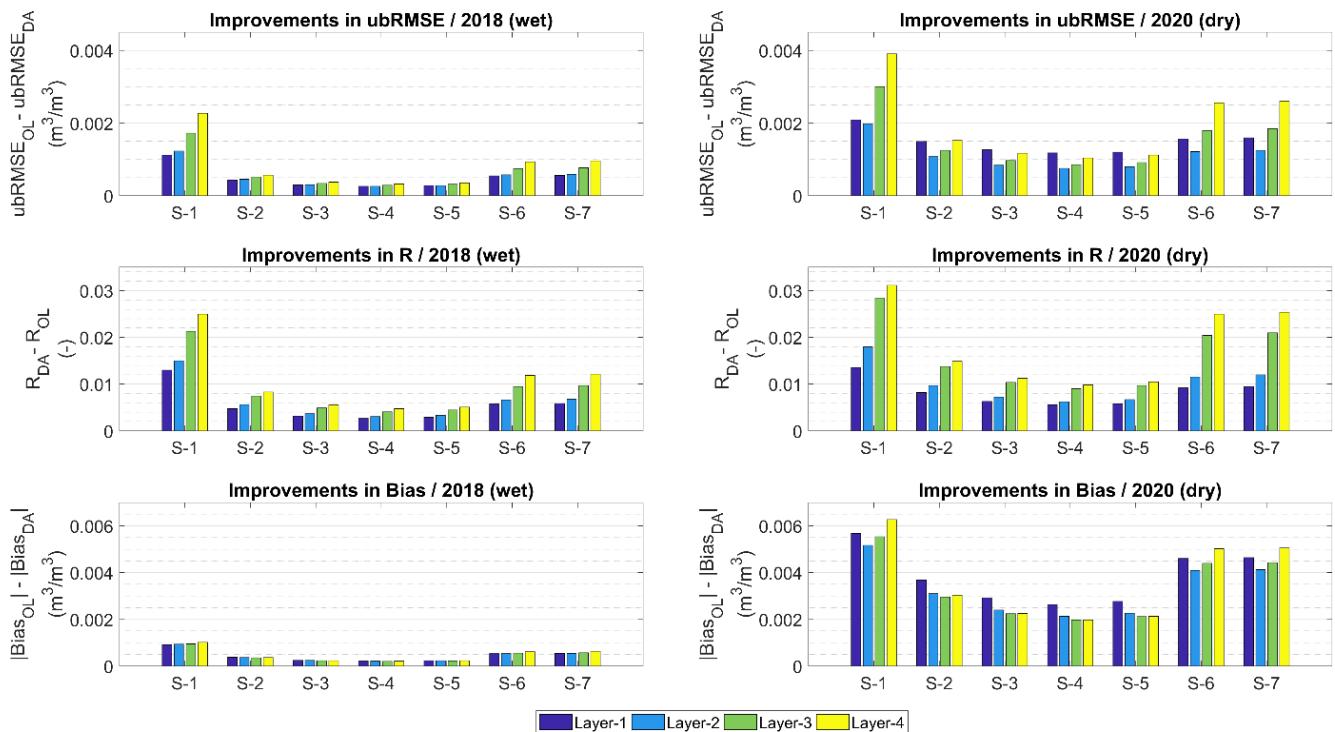


Figure 5. Basin-averaged soil moisture improvements in all scenarios for 2018 (left) and 2020 (right).

In Scenario-2, where soil moisture of all grid cells in only the surface soil layer was updated every day, the layer-averaged improvements are 31 % (ubRMSE), 35 % (R) and 38 % (bias) of the layer-averaged improvements that were achieved for Scenario-1 in 2018, and 48 % (ubRMSE), 51 % (R) and 56 % (bias) in the case of 2020. In Scenario-3, which is the same as Scenario-2 except that SMAP derived temporal gaps were introduced, the layer-averaged improvements drop further to 21 % (ubRMSE), 23 % (R), and 25 % (bias) of Scenario-1 in 2018, and 38 % (ubRMSE), 39 % (R), and 43 % (bias) in 2020.

In Scenario-4, where SMAP-derived horizontal spatial gaps were introduced in addition to SMAP-derived temporal gaps, the layer-averaged improvement is reduced even more to 18 % (ubRMSE), 20 % (R), and 22 % (bias) of Scenario-1 in 2018, and 35 % (ubRMSE), 34 % (R), and 38 % (bias) in 2020. To summarize, compared to the theoretical maximum improvements that can be achieved by Scenario-1, a substantial amount of that improvement is lost when the lower model soil layers are not updated. The second largest reduction in assimilation performance happens when the grid cells are not updated every day. And lastly, horizontal spatial gaps cause an even smaller amount of reduction.

So far Scenario-4 is the most realistic configuration, because remotely sensed soil moisture observations also suffer from missing data in vertical, horizontal, and temporal dimensions. And as previously discussed, a common workaround to this problem in the context of data assimilation is to update the unobserved model grid cells based on the covariance between soil moisture of the unobserved and nearby observed grid cells. Doing this in Scenario-5 only for the horizontal spatial dimension increases the layer-averaged improvements up to 20 % (ubRMSE), 21 % (R), and 24 % (bias) of Scenario-1 levels in 2018, and 37 % (ubRMSE), 36 % (R), and 41 % (bias) in 2020. In Scenario-6, where the covariance matrix of EnKF was used to update unobserved grid cells of lower soil layers instead of the surface soil layer, layer-averaged improvements are 44 % (ubRMSE), 45 % (R), and 59 % (bias) of Scenario-1 in 2018, and 65 % (ubRMSE), 73 % (R), and 80 % (bias) in 2020. In Scenario-7, where unobserved grid cells of all soil layers were updated, layer-averaged improvements reach 46 % (ubRMSE), 47 % (R), and 60 % (bias) of Scenario-1 in 2018, and 66 % (ubRMSE), 75 % (R), and 81 % (bias) in 2020. As it was already determined that not updating lower soil layers causes a significant reduction in

assimilation performance, it makes sense that updating unobserved grid cells in the lower soil layers is much more beneficial than updating unobserved grid cells in the surface soil layer.

Finally in Scenario-7, even after updating the unobserved grid cells in all soil layers using the covariance matrix of EnKF, the assimilation performance could not be brought up to the levels of Scenario-1. Part of the reason is that the temporal data gaps are still present in Scenario-7. Under real-data conditions, and if the modeling system is not needed for real-time purposes such as operational forecasting, higher level data products (with higher latency) may be used for data assimilation which usually have their temporal gaps (as well as spatial gaps) filled through external means. Another reason why Scenario-7 failed to reach the improvement levels of Scenario-1 could be because of inherent limitations of the covariance matrix technique. Future studies are recommended where soil moisture of all unobserved grid cells is estimated independently outside of the data assimilation framework, and then brought in to update the model states in a grid-by-grid fashion.

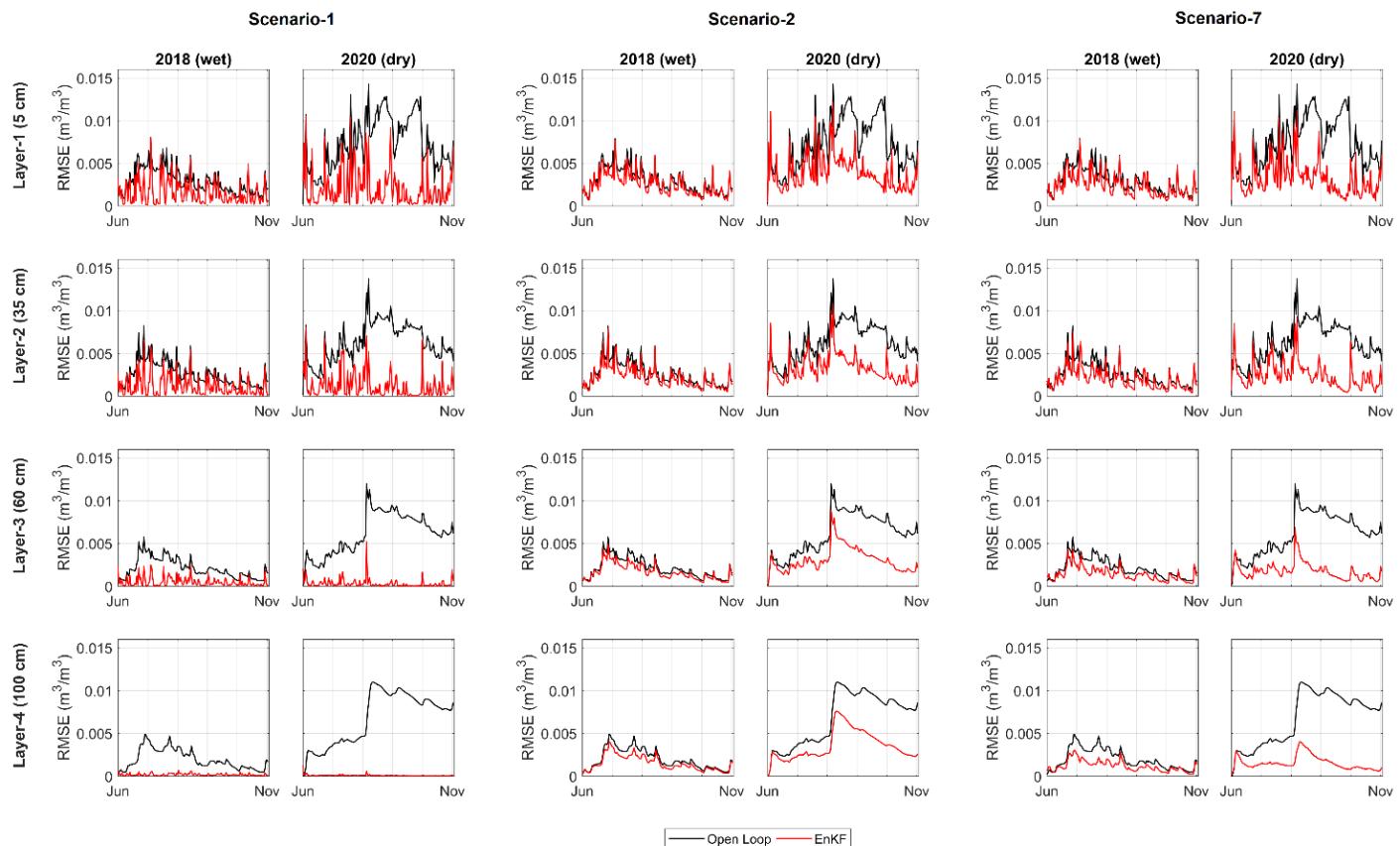


Figure 6. Temporal distribution of basin-averaged soil moisture simulation errors for all soil layers and for scenarios 1, 2 & 7.

Another key finding from Figure 5 is that the improvement magnitudes are much larger in 2020 (dry summer) compared to 2018 (wet summer). This difference may be better explained through Figure 6, where the temporal distribution of spatially averaged (over the watershed) RMSE is presented. Only three out of seven scenarios are presented for brevity. The term ‘improvement’ used so far is essentially the difference between the black and blue lines (i.e., open loop and post-assimilation model errors) in Figure 6. For each soil layer within each scenario, post-assimilation model errors have somewhat similar ranges between the two years. Rather it is the range of open loop model errors that are starkly different between the years. Therefore, the larger magnitude of open loop errors is what is mainly contributing to the larger improvements in 2020 compared to

2018. Incidentally, the forcings in the ERA5 ensemble member (out of the 10) that was randomly chosen to generate the synthetic truth is farther away from the ensemble mean in 2020 than in 2018, causing the larger open loop model errors, especially the bias component of errors (Figure 5). Regardless of the different magnitudes of open loop RMSE in the two years, the EnKF algorithm was able to bring down the open loop RMSE in both years to a similar level (Figure 6). This indicates that if there is available room for improvement, the EnKF can, at least under ideal situations such as this synthetic experiment, effectively improve modelling performances.

Looking at the improvements in the individual soil layers in Scenario-2 (Figures 5,6), it is interesting to observe that even though soil moisture in only the surface soil layer was updated, improvements occurred at all four soil layers. The changes being made to the surface soil moisture are therefore being propagated to the lower soil layers through model physics. However, the improvements to the bottom soil layers are greater when they are being actively updated either in a zero-dimensional (Scenario-1), one-dimensional (Scenario-6) or three-dimensional (Scenario-7) manner. In fact, actively updating the model states of the bottom layers appear beneficial for the top layer as well. For instance, the magnitude of the top layer improvement is higher in Scenario-1 than in Scenario-2, even though the only difference between these scenarios is whether bottom layers are actively updated or not.

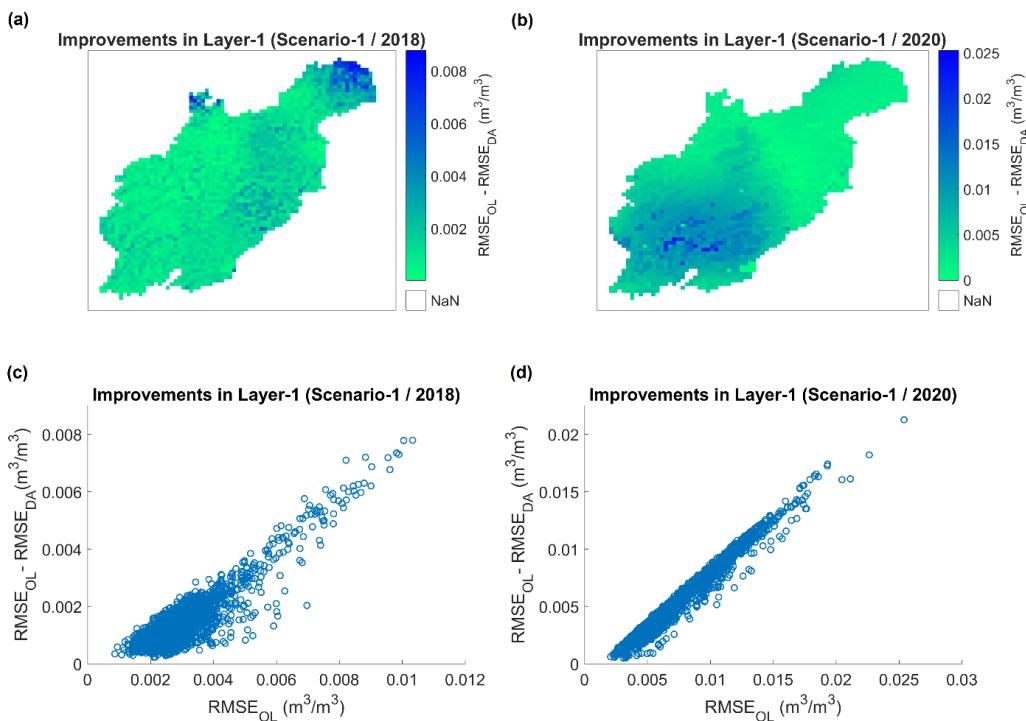


Figure 7: Spatial distribution of soil moisture improvements in Layer-1 of Scenario-1 for 2018 (a) and 2020 (b). Correlation between open loop errors and post-assimilation improvements for 2018 (c) and 2020 (d)

The spatial distribution of temporally averaged (over June-October) improvements of the surface soil moisture in Scenario-1 are shown in Figure 7 (a-b). The magnitudes of improvement are not spatially uniform. Model grid cells located at the northeast corner of the watershed have higher levels of improvement in 2018, and in the southwestern region in the case of 2020. To identify the factors creating these spatial patterns, correlations between the improvements of the watershed grid cells and different variables were plotted, including meteorological variables such as seasonal precipitation and temperature, as well as static physiographic variables such as soil type and land use. No significant correlation could be identified between the spatial patterns of the improvements with the spatial patterns of either the meteorological or physiographic

variables. Rather, the strongest predictor of the improvement patterns is the magnitude of the open loop model error, as shown in the Figure 7 (c-d). In other words, the grid cells in which open loop RMSE was higher, saw a larger improvement after assimilation. This is similar to the finding for watershed averaged open loop RMSE and improvements, as previously discussed in Figure 6.

3.3. Real-data Experiments

The experiment of Scenario-7 was repeated for real-data conditions, whereby instead of synthetic soil moisture observations, SMAP observations were used to update model soil moisture. It should be noted that all available SMAP data with a 'retrieval successful' flag was utilized for assimilation, regardless of whether the data also had a 'recommended quality' flag or not. Due to large portions of the study area being forested, if only data having 'recommended quality' flags were to be used, a vast majority of the dataset would be rendered unusable. Whereas in the synthetic experiments the goal was to guide the model soil moisture values towards the synthetic truth, in the real-data experiments the goal was to guide the model soil moisture values towards the in-situ observations. Another major difference between the synthetic and real-data experiments is that during the synthetic experiments, synthetic truth was available over all model grid cells and over all four soil layers for evaluating the post-assimilation model performance. But for the real-data experiments, in-situ soil moisture data was available only over four model grid cells.

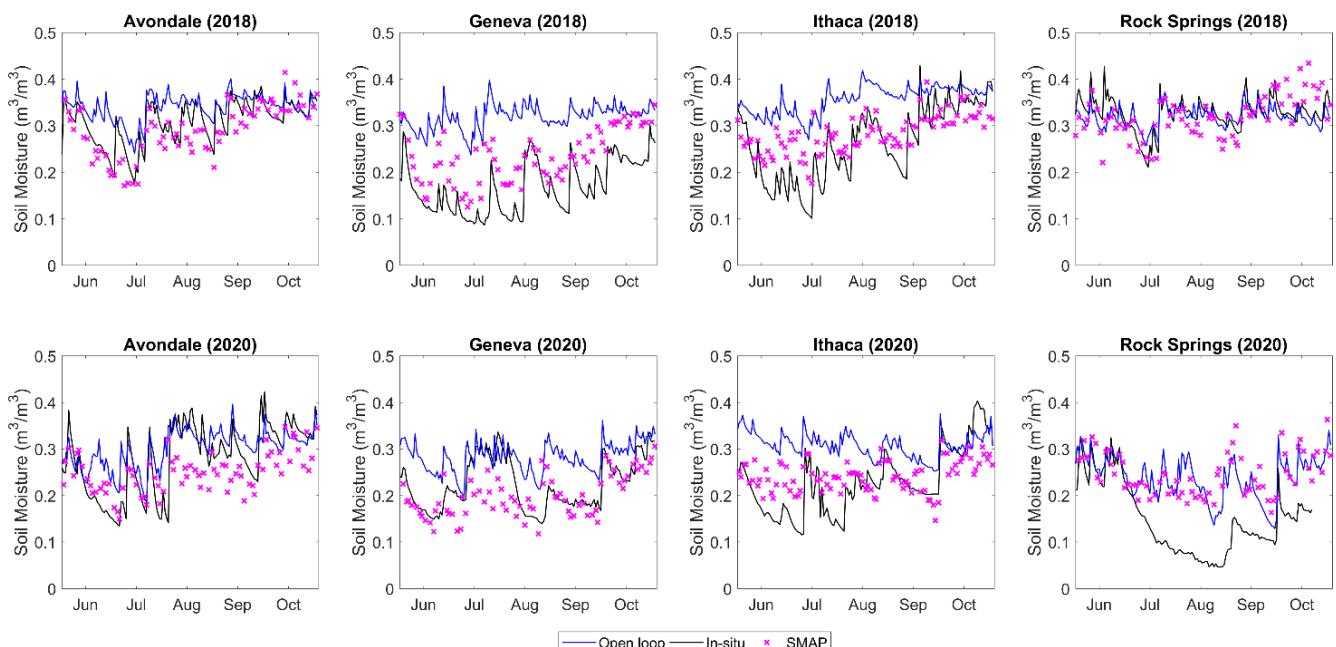


Figure 8. Soil moisture values of open loop model, in-situ observations and SMAP observations at four in-situ measurement stations in 2018 (top row) and 2020 (bottom row).

To compare the spatially averaged model soil moisture values with point-scale in-situ ones, the in-situ data was assumed to be representative of the soil moisture of the corresponding 25 km^2 model grid cell. Also as previously mentioned, in-situ observations were measured at 5, 10, 20, 50 and 100 cm soil depths. To compare depth-averaged model soil moisture values with point-scale in-situ ones, the in-situ data at 5 cm depth was assumed to be representative of the first model soil layer of 5 cm thickness (0-5 cm depth), the mean of the 10 and 20 cm in-situ data was used to represent the second model soil layer of 35 cm thickness (5-40 cm depth), and the mean of the 50 and 100 cm in-situ data was compared with the third model soil layer of 60 cm thickness (40-100 cm depth). Such

spatial scale mismatches (in both horizontal and vertical directions) are expected to inevitably introduce some errors to the estimates of the data assimilation performance.

The surface soil moisture values of the four available in-situ stations are plotted in Figure 8 along with the corresponding open loop model soil moisture and remotely sensed SMAP soil moisture. As there is a bias between the SMAP data and the model in most cases, the SMAP data was rescaled to the model climatology before performing data assimilation. However, because of this rescaling process, only the information about the relative variability within the SMAP data could be utilized through data assimilation, and not its absolute values [56]. Similarly, because of the existing bias seen between the in-situ and modelled data, the goal of the real-data assimilation experiments will not be to guide the model simulations of soil moisture towards in-situ data in terms of absolute values. Instead, the goal will be to modify the relative variability of the modelled soil moisture towards the relative variability of the in-situ soil moisture. The results from the real-data experiments (ubRMSE, R, and bias computed between the modelled and in-situ soil moisture) are presented in Table 2, along with the ubRMSE, R, and bias computed between the original SMAP data (prior to bias correction) and in-situ soil moisture observations over the surface layer. All the values in Table 2 are the averaged values over the four grid cells where in-situ data is available.

Table 2. Unbiased root-mean-squared-error (ubRMSE), correlation coefficient (R), and bias calculated between in-situ soil moisture observations and open loop model runs (OL), remotely sensed SMAP data (RS), and post-assimilation model runs (DA).

	Layer 1						Layer 2						Layer 3					
	2018 (wet)			2020 (dry)			2018 (wet)			2020 (dry)			2018 (wet)			2020 (dry)		
	OL	RS	DA	OL	RS	DA	OL	DA	OL	DA	OL	DA	OL	DA	OL	DA	OL	DA
ubRMSE (m ³ /m ³)	0.041	0.040	0.040	0.053	0.052	0.051	0.037	0.036	0.042	0.041	0.027	0.028	0.021	0.022				
R	0.68	0.73	0.72	0.57	0.61	0.66	0.68	0.71	0.75	0.78	0.63	0.64	0.64	0.60				
Bias (m ³ /m ³)	0.071	0.026	0.071	0.062	0.039	0.062	0.064	0.064	0.079	0.079	0.054	0.054	0.044	0.044				

In the first soil layer, assimilation of SMAP data caused small improvements in the simulated soil moisture for both 2018 and 2020. This is consistent with the fact that the SMAP data in both of those years is slightly better at representing in-situ data (in terms of ubRMSE and R) than the open loop model. Even though SMAP data is much better at representing in-situ data in terms of bias compared to open loop model, assimilation was unable to improve the model bias in any soil layer because of the bias-correction of SMAP data prior to assimilation. Minor improvements in the model simulation are also true for the second soil layer, but assimilation fails to improve soil moisture simulation in the third soil layer. In fact, assimilation causes a slight overall decrease in model performance in the third layer. It may be reminded that SMAP data was directly assimilated into the first layer, while the bottom two layers were updated based on the covariance between the top and bottom model layers. Therefore, improvements in the upper two layers, yet failure to do so in the bottom layer, may indicate a poor representation in the model of the cross correlations between the different soil layers that exist in nature, and which is embedded in the in-situ data.

The results for the four individual stations are also presented in Table 3. Here, only the correlation coefficient values are shown, since the bulk of the improvement (or degradation) caused by data assimilation is best expressed by this metric as seen in Table 2. There is no improvement in terms of bias since the SMAP observations were rescaled to the model space, and the improvements in terms ubRMSE are also minimal. Similar to Table 2, all the R values presented in Table 3 have either been computed between the modelled and in-situ soil moisture, or between the original SMAP data (prior to bias

correction) and in-situ soil moisture observations over the surface layer. Based on the R values of open loop, it could be stated that the calibrated model is generally well capable of simulating the temporal variability of the in-situ soil moisture data, with a few exceptions such as the first soil layer at Ithaca in 2020, or the third soil layer at Geneva in 2020.

Table 3. Correlation coefficient (R) values for individual in-situ measurement locations, calculated between in-situ soil moisture observations and open loop model runs (OL), remotely sensed SMAP data (RS), and post-assimilation model runs (DA).

	Layer 1						Layer 2						Layer 3					
	2018 (wet)			2020 (dry)			2018 (wet)			2020 (dry)			2018 (wet)			2020 (dry)		
	OL	RS	DA	OL	RS	DA	OL	DA	OL	DA	OL	DA	OL	DA	OL	DA	OL	DA
Avondale	0.77	0.82	0.81	0.76	0.70	0.75	0.73	0.76	0.76	0.71	0.33	0.24	0.84	0.86				
Geneva	0.50	0.78	0.54	0.67	0.61	0.70	0.47	0.52	0.71	0.74	0.47	0.57	0.16	0.27				
Ithaca	0.79	0.80	0.84	0.25	0.57	0.54	0.85	0.88	0.73	0.86	0.85	0.88	0.69	0.69				
Rock Springs	0.68	0.53	0.68	0.62	0.56	0.66	0.68	0.70	0.80	0.82	0.87	0.85	0.85	0.58				

After assimilating SMAP data into the model, the temporal variability of simulated soil moisture was generally improved, albeit slightly, again with a few exceptions. The relatively smaller improvements in real-data experiments as compared to the synthetic experiments, and the degradations in the few cases may be explained as follows. It was demonstrated through the synthetic experiments that if high-quality observations of the true soil moisture are assimilated into the model, then the model soil moisture will be nudged closer to the true soil moisture. In that case the observations, open loop model runs, and true soil moisture were all generated from the same source (WRF-Hydro model) and were well known when calculating the evaluation metrics. But in this real-data experiment, remotely sensed SMAP soil moisture is used as observations to nudge the modelled soil moisture towards in-situ soil moisture. Each of these datasets have conceptually different sources having different spatial scales: one estimated from remotely sensed microwave data, one generated based on physical equations in a model, and one measured physically in the field.

The errors between each of these datasets and the 'true' soil moisture actually occurring in nature is not well known. As a result, the improvement in terms of R between the post-assimilation model simulated soil moisture and in-situ soil moisture depended partially on how well SMAP data happened to agree with corresponding in-situ soil moisture data. Therefore, performing assimilation with an observation data of higher quality may have contributed to further improvements in model performance. Other detrimental factors, that can impact any real-data experiment, include errors in the in-situ dataset (which appears to be an issue for the 2020 data of Rock Springs, as noticeable in Figure 8), errors introduced during the spatial rescaling of data, using time-invariant observation errors in EnKF, the soil/vegetation properties of the model grid cell not being representative of the local conditions, and so on. Finally, these results are specific to the modeling framework and observation datasets that were used in this study, and therefore using a different model or a different remotely sensed soil moisture dataset may lead to different, and perhaps better simulations of these in-situ target datasets.

3.4. Streamflow Modeling Performance

The impacts of updating model soil moisture on the simulated streamflow during the synthetic experiment (Scenario-1) are presented in Figure 9. The first row shows the spatially averaged (over the watershed) 3-day average precipitation. In the second row are the changes in the watershed-averaged modelled soil moisture (defined as post-assimilation soil moisture minus the open loop soil moisture) due to the assimilation of

synthetic soil moisture observations. Similarly, the third row shows the changes in the streamflow at the watershed outlet point of Harrisburg (defined as post-assimilation streamflow minus the open loop streamflow) caused by the same assimilation of synthetic soil moisture observations. When soil moisture in the model is increased during data assimilation, extra water is essentially added to the water budget and vice versa. The similarity of temporal variations between Figure 9 (c) and Figure 9 (e) helps visualize this phenomenon for 2018 (wet summer). Consequently, when soil moisture is added to the system, streamflow at the watershed outlet is also increased, and vice versa.

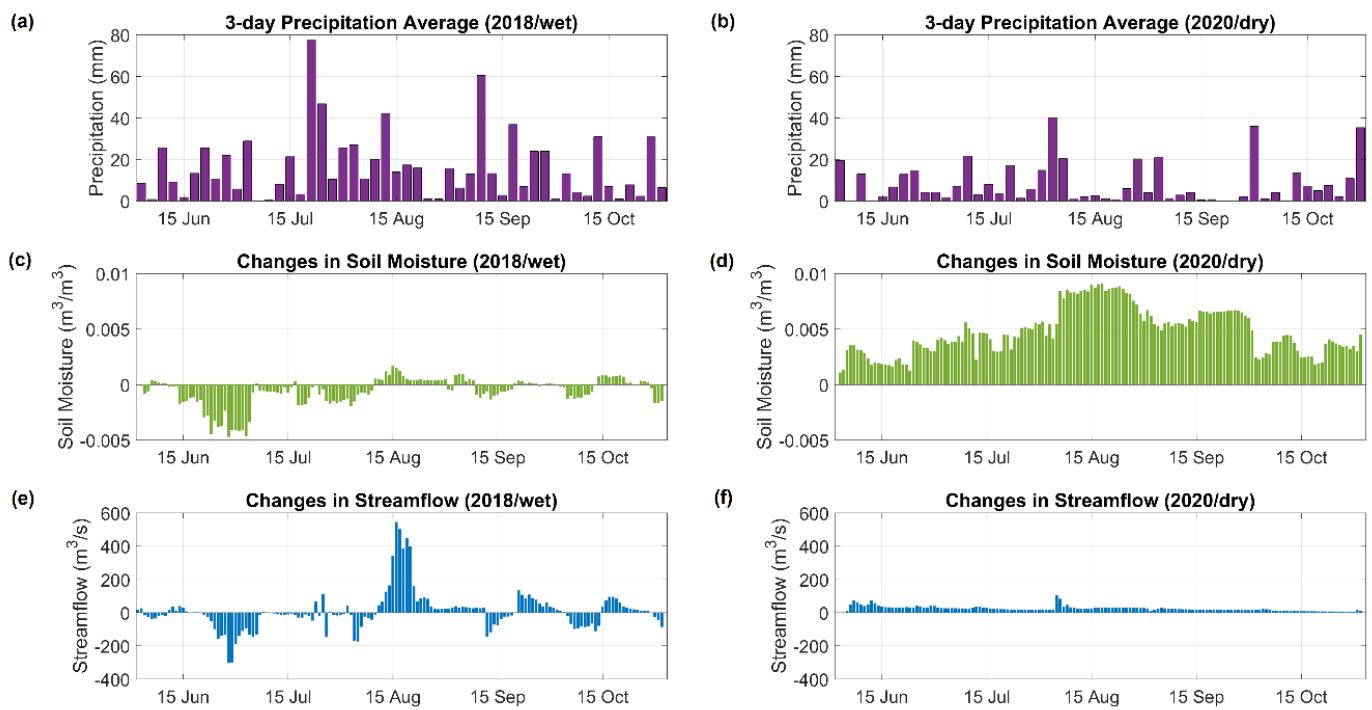


Figure 9. Watershed-averaged 3-day precipitation average, changes in watershed-averaged soil moisture due to data assimilation with Scenario-1, and changes in streamflow of watershed outlet due to data assimilation in 2018 (left column) and 2020 (right column).

Note that the magnitude of the changes in soil moisture does not correspond to a similar magnitude of change in streamflow. This is because in addition to adding or subtracting the total amount of water in the water budget, data assimilation of soil moisture causes redistribution of existing water within different fluxes. For example, when soil moisture is reduced during data assimilation, this increases the storage capacity of the soil and thus any concurrent precipitation event will generate less surface runoff than if data assimilation had not been performed. Therefore, some portion of the streamflow source will now be transformed from quick surface runoff to slower moving baseflow, which will be added to the stream at a later period. Similarly, if soil moisture is increased during a wet period when the soil is already near saturation, for example on around 15 August 2018, a larger portion of the precipitation event will now join the streamflow as direct surface runoff.

As previously discussed, for 2018 the open loop soil moisture and synthetic truth soil moisture were relatively close to each other in magnitude, and the open loop soil moisture both underestimated and overestimated the synthetic truth soil moisture at times, as depicted in Figure 9 (c). But as also discussed previously, for 2020 it just so happened that the open loop soil moisture highly underestimated the synthetic truth soil moisture all throughout the experiment period, as shown in Figure 9 (d). Therefore, data assimilation only increased the soil moisture of the open loop, resulting in only positive changes to the outlet streamflow. Even though the magnitude of soil moisture increases in 2020 are much

greater than the soil moisture increases in 2018, the magnitude of streamflow increases in 2020 are relatively lower than 2018. This can be explained by the fact that 2020 had a dry summer, with lower precipitation values and lower soil saturation percentages compared to 2018. This combination of unsaturated soil and lack of heavy rainstorms meant that much of the added soil moisture was gradually added to the stream via baseflow.

Now that it has been established that the soil moisture-streamflow interactions of this modeling framework are behaving realistically and as intuitively expected, the next question might be whether updating soil moisture with observations cause the simulated streamflow to shift towards the desired direction, i.e., closer to the true streamflow. In other words, if the goal of a modeling exercise is to improve the simulation of streamflow, will updating model soil moisture states alone with observations achieve this goal? The model simulated streamflow resulting from soil moisture data assimilation in the synthetic experiments are plotted in Figure 10, along with the open loop model run and synthetic truth. The NSE and log-NSE values presented were both calculated using the post-assimilation model streamflow and the synthetic truth version of streamflow.

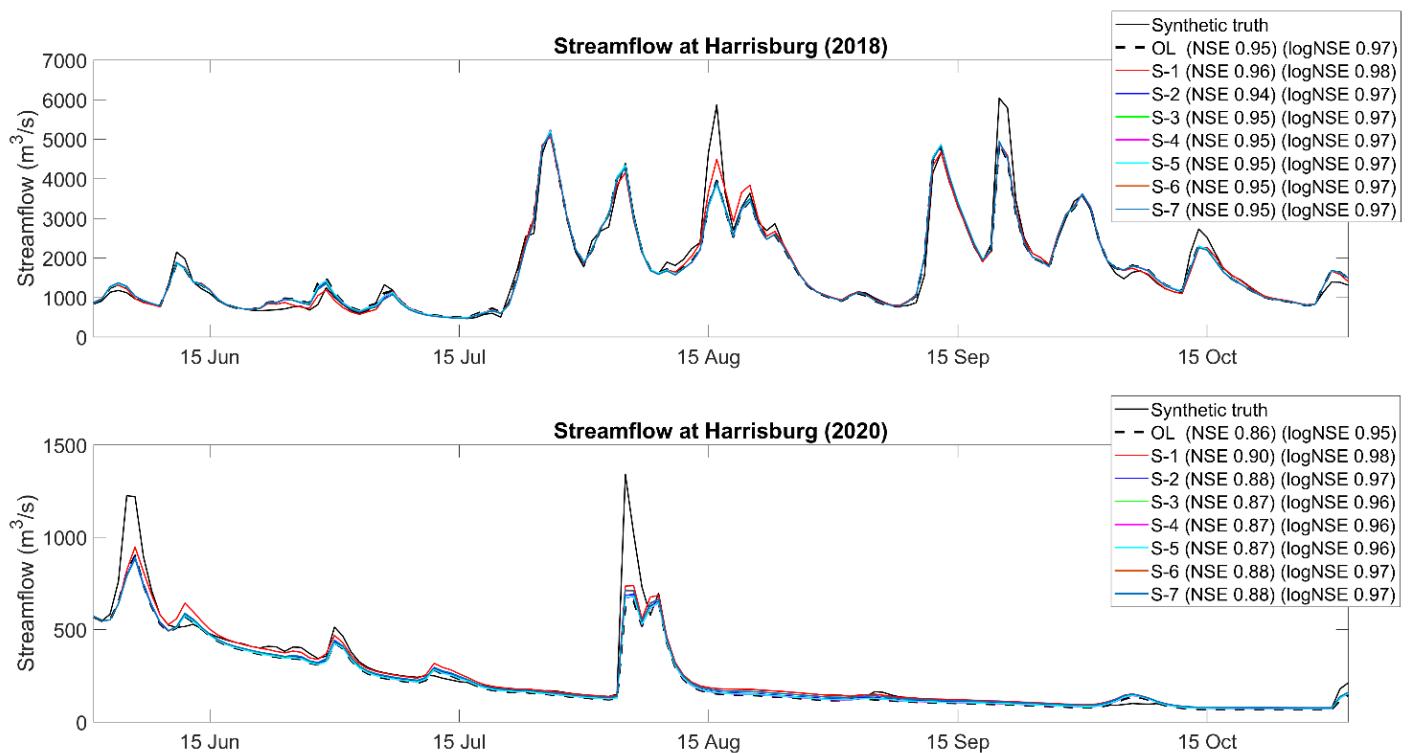


Figure 10. Simulated streamflow at the watershed outlet (Harrisburg) for all seven assimilation scenarios as well as the open loop model run, and synthetic truth model run in 2018 (top) and 2020 (bottom).

In the wet year of 2018 and compared to the open loop, assimilation of soil moisture causes no significant modifications to the overall streamflow (in terms NSE and log-NSE) in all of the synthetic scenarios, except minor improvements in Scenario-1. The NSE value increased from 0.95 (OL) to 0.96 (S-1), and the log-NSE value increased from 0.97 (OL) to 0.98 (S-1). However, in terms of absolute error, during the heavy rainfall period on around 15 August 2018, adjusting antecedent soil moisture caused an error reduction in Scenario-1 of $544 \text{ m}^3/\text{s}$ (also shown in Figure 9-e). To achieve such gains though, soil moisture of all grid cells of the model needs to be updated as in Scenario-1, because no such significant absolute error reduction occurred in any other scenario.

In the dry year of 2020, no such significant absolute error reduction occurred in any of the scenarios at all, as the modification to the streamflow tended to be somewhat uniform all throughout the season (also shown in Figure 9-f). As previously discussed,

this is expected because correcting soil moisture is more likely to improve peak flow simulation during wet periods when the soil state remains closer to saturation. But the overall streamflow simulation (in terms of NSE and log-NSE) did improve for all scenarios compared to the open loop, with the highest improvement seen in Scenario-1. The NSE value increased from 0.86 (OL) to 0.9 (S-1) and 0.88 (S-7). The log-NSE value increased from 0.95 (OL) to 0.98 (S-1) and 0.97 (S-7).

In the real-data experiments, the open loop model performances are much lower to begin with compared to the synthetic experiments. This is expected because as explained previously for soil moisture, the open loop and the truth were both generated by the model in the case of synthetic experiments. Whereas for real-data experiments, the open loop performance is the outcome of a complex parameter-tuning process trying to emulate the in-situ data. In the wet year of 2018 updating model soil moisture states with SMAP data in the Scenario-7 configuration resulted in NSE value staying unchanged at 0.77, and the log-NSE value increasing from 0.82 to 0.84. This is somewhat commensurate with the wet year assimilation gains seen in the synthetic experiments. But in the dry year of 2020, no gains at all are achieved contrary to the synthetic experiments. Rather, there is a decrease in model performance. The NSE value drops from 0.71 (OL) to 0.7 (S-7), while the log-NSE (which gives more weight to the baseflow simulation performance) dropped from 0.61 (OL) to 0.5 (S-7). One important factor that may partially explain the relatively poor dry year assimilation performance is that the model parameters being used for these experiments have been calibrated based on the data of a wet year and using this model for running dry year assimilation experiments was an afterthought. Perhaps using a model that is better calibrated against dry conditions may be able to improve the streamflow simulation for dry years as well.

Another possible cause behind the real-data experiments not leading to improved streamflow is that, while the hydrological processes connecting soil moisture and streamflow were exactly the same for both the hydrological model and synthetic truth in the synthetic experiments, such was not the case for real-data experiments. The exact mechanisms by which any changes in real-world soil moisture translates into the real-world streamflow is unknown and most likely different from that of the model. Finally, model structure also matters. A majority of hydrological models offer a rather crude conceptualization of baseflow/groundwater flow. WRF-Hydro is perhaps an exception, where saturated subsurface flow is based on the Dupuit-Forchheimer assumption. Yet, using groundwater flow models based on the transient groundwater flow equation, such as MODFLOW, would probably better simulate low streamflow dominated by baseflow conditions.

Other studies have indicated that assimilating soil moisture to improve streamflow is a hit-and-miss approach depending on what exact methods and datasets are used in the process, and that assimilating soil moisture alone may not be sufficient for this purpose [32,57-58]. Therefore, a trial-and-error strategy is required to determine which modeling framework is most beneficial to improve the streamflow simulation for a particular watershed, and whether to assimilate soil moisture, streamflow, or a combination of both. Also, this study took a deterministic approach by comparing between the ensemble mean of the 24-member post-assimilation outputs and that of the 9-member open loop outputs. For a probabilistic approach to data assimilation, performance metrics like the Continuous Ranked Probability Score (CRPS) may be recommended, which is able to provide information about the spread of the forecast ensemble as well.

4. Conclusions

A set of synthetic experiments were designed in this study to assess the impacts of spatiotemporal discontinuities in the remotely sensed soil moisture data on the performance of hydrological data assimilation. For this purpose, the WRF-Hydro model was set up over the Susquehanna River watershed. SMAP was selected as the example of a remotely sensed soil moisture product, and the EnKF was selected as the data

assimilation algorithm. The model was forced with the ERA5 10-member ensemble. Nine of those 10 members were used to generate open loop model runs and the remaining tenth member was considered as a synthetic truth. These nine different forcing projections were combined with soil moisture state perturbations to represent the overall modeling uncertainty in the EnKF assimilation framework to create 24 post-assimilation ensemble members in total.

The synthetic experiments consisted of the following seven different scenarios: 1) soil moisture states of all model grid cells in all soil layers were updated every day with synthetic soil moisture observations, 2) all grid cells of the surface soil layer were updated every day, 3) same as 2 but no updates were made on days in which SMAP data is not available, 4) same as 3 but no updates were made on model grid cells over which SMAP data is not available, 5) same as 4 but surface layer model grid cells with missing observations are updated based on the covariance between them and nearby grid cells that have observations, 6) same as 4 but grid cells in bottom soil layers with missing observations are updated based on the covariance between them and nearby surface grid cells that have observations, and 7) combining 5 and 6 i.e., updating all grid cells with missing observations based on nearby grid cells that have observations. All of these scenarios were then compared with the open loop scenario.

Results show that out of all the scenarios, the best improvement in simulated soil moisture is achieved when synthetic soil moisture observations are assimilated into the model in all grid cells of all soil layers. Introducing spatiotemporal discontinuities in the observation data reduces assimilation performance. The largest reduction happens because of the unavailability of root-zone observations, followed by temporal data gaps and horizontal spatial gaps. In practice, satellite soil moisture datasets have all three of these missing components. The reduction in data assimilation performance due to the presence of these data discontinuities can be somewhat offset by indirectly updating the states of unobserved model grid cells. The indirect update is made based on the covariance between the soil moisture of an unobserved grid cell and one or more nearby observed grid cells. Results also show that if high-quality observations are available, then the magnitude of improvements brought about by data assimilation will be mostly dictated by how much model error there is in the pre-assimilation open loop model runs.

Real-data experiments were also performed where SMAP data was assimilated into the model to try and help the variability of simulated soil moisture to match that of the in-situ soil moisture. Some parameters of the WRF-Hydro model were calibrated against in-situ soil moisture and streamflow to make the model mimic in-situ conditions as best as it can prior to data assimilation. In-situ soil moisture measurements were available for only four locations within the study area. Results indicate that data assimilation was able to generally improve the ubRMSE and correlation coefficient values between the model simulated and in-situ soil moisture for the top two model soil layers. One of the reasons behind the less than optimum performance (compared to something like a synthetic experiment) is that the SMAP observations are not a high-quality representation of the in-situ observations. Large portions of the study area are forested which negatively impacts microwave retrieval-based soil moisture products like SMAP.

The impact of updating model soil moisture on the simulation of streamflow was also analyzed. It was found that when soil moisture is added to the model through data assimilation during a wet period when the soil is already near saturation, this increases surface runoff after a heavy rainfall event and causes significant increases in the streamflow. Increasing soil moisture during a dry period does not have this effect and the newly added water is rather added to the streamflow via baseflow. On the other hand, reducing the model soil moisture increases soil storage capacity and therefore decreases post-rainfall surface runoff by encouraging more infiltration. Even though assimilating soil moisture into the model impacts the generation of streamflow, the timing and magnitude of changes imposed on the simulated streamflow does not necessarily improve the accuracy of streamflow. Some improvements in the NSE and log-NSE of streamflow simulations were achieved in the synthetic experiments where high-quality soil moisture

observations were assimilated. But for poorer-quality SMAP data assimilation scenarios, streamflow accuracy was rather reduced in terms of NSE and log-NSE for the dry experiment year, which may also have been impacted by the use of a model calibrated specifically for wet conditions.

Finally, if the goal is to improve the streamflow modeling performance of a particular study area through the assimilation of observed soil moisture, it may be recommended to explore through multiple modeling/assimilation strategies and multiple observation datasets to find the best fit. Direct assimilation of streamflow into the model in combination with soil moisture is another avenue for further improvement of streamflow modeling performance. It should be noted that all the results presented in this study are for a 24-hour lead time. Any improvements achieved through data assimilation at a shorter lead time is likely to peter out as the lead time of forecast increases. Since streamflow observations provide short-lived information about a flux but soil moisture observations have a longer memory of the soil water storage, future research may be recommended where the impacts of combined soil moisture-streamflow assimilation on the modeling performance at different lead times are investigated.

Author Contributions: K.M., R.L., and M.T. contributed to the conceptualization of the study and development of methodology. K.M. conducted the formal analysis, investigation, and visualization. K.M. prepared the original draft. R.L. and M.T. conducted the review and editing of the manuscript, as well as supervising the overall study. All authors have read and agreed to the published version of the manuscript.

Acknowledgements: This research was supported by the NSERC Industrial Research Chair on the Application of Hydrometeorological Data from Satellite Images to Improve Hydrological Forecasting, whose industrial partners are Hydro-Québec, Brookfield Renewable, and the City of Sherbrooke.

Data Availability Statement: The underlying code and data are available upon request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Peng, J.; Albergel, C.; Balenzano, A.; Brocca, L.; Cartus, O.; Cosh, M. H.; ...; Loew, A. A roadmap for high-resolution satellite soil moisture applications—confronting product characteristics with user requirements. *Remote Sens. Environ.* **2021**, *252*, 112162. <https://doi.org/10.1016/j.rse.2020.112162>
2. Gruber, A.; De Lannoy, G.; Crow, W. A Monte Carlo based adaptive Kalman filtering framework for soil moisture data assimilation. *Remote Sens. Environ.* **2019**, *228*, 105-114. <https://doi.org/10.1016/j.rse.2019.04.003>
3. Dorigo, W.; Himmelbauer, I.; Aberer, D.; Schremmer, L.; Petrakovic, I.; Zappa, L.; ...; Sabia, R. The International Soil Moisture Network: serving Earth system science for over a decade. *Hydrol. Earth Syst. Sci.* **2021**, *25*(11), 5749-5804. <https://doi.org/10.5194/hess-25-5749-2021>
4. Kerr, Y. H.; Waldteufel, P.; Wigneron, J. P.; Delwart, S.; Cabot, F.; Boutin, J.; ...; Mecklenburg, S. The SMOS mission: New tool for monitoring key elements of the global water cycle. *Proc. IEEE* **2010**, *98*(5), 666-687. <https://doi.org/10.1109/PROC.2010.2043032>
5. Entekhabi, D.; Njoku, E. G.; O'Neill, P. E.; Kellogg, K. H.; Crow, W. T.; Edelstein, W. N.; ...; Van Zyl, J. The soil moisture active passive (SMAP) mission. *Proc. IEEE* **2010**, *98*(5), 704-716. <https://doi.org/10.1109/PROC.2010.2043918>
6. Babaeian, E.; Sadeghi, M.; Jones, S. B.; Montzka, C.; Vereecken, H.; Tuller, M. Ground, proximal, and satellite remote sensing of soil moisture. *Rev. Geophys.* **2019**, *57*(2), 530-616. <https://doi.org/10.1029/2018RG000618>
7. Park, J.; Bindlish, R.; Bringer, A.; Horton, D.; Johnson, J. T. Soil moisture retrieval using a time-series ratio algorithm for the NISAR mission. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Brussels, Belgium, 11-16 July 2021. <https://doi.org/10.1109/IGARSS47720.2021.9554619>
8. Massari, C.; Brocca, L.; Tarpanelli, A.; Moramarco, T. Data assimilation of satellite soil moisture into rainfall-runoff modelling: A complex recipe? *Remote Sens.* **2015**, *7*(9), 11403-11433. <https://doi.org/10.3390/rs70911403>
9. Kolassa, J.; Reichle, R. H.; Draper, C. S. Merging active and passive microwave observations in soil moisture data assimilation. *Remote Sens. Environ.* **2017**, *191*, 117-130. <https://doi.org/10.1016/j.rse.2017.01.015>
10. Lievens, H.; Reichle, R. H.; Liu, Q.; De Lannoy, G. J.; Dunbar, R. S.; Kim, S. B.; ...; Wagner, W. Joint Sentinel-1 and SMAP data assimilation to improve soil moisture estimates. *Geophys. Res. Lett.* **2017**, *44*(12), 6145-6153. <https://doi.org/10.1002/2017GL073904>

11. Dumedah, G.; Walker, J. P.; Merlin, O. Root-zone soil moisture estimation from assimilation of downscaled Soil Moisture and Ocean Salinity data. *Adv. Water Resour.* **2015**, *84*, 14–22. <https://doi.org/10.1016/j.advwatres.2015.07.021>
12. Draper, C. S.; Reichle, R. H.; De Lannoy, G. J. M.; Liu, Q. Assimilation of passive and active microwave soil moisture retrievals. *Geophys. Res. Lett.* **2012**, *39*(4), L04401. <https://doi.org/10.1029/2011GL050655>
13. Blyverket, J.; Hamer, P. D.; Bertino, L.; Albergel, C.; Fairbairn, D.; Lahoz, W. A. An Evaluation of the EnKF vs. EnOI and the Assimilation of SMAP, SMOS and ESA CCI Soil Moisture Data over the Contiguous US. *Remote Sens.* **2019**, *11*(5), 478. <https://doi.org/10.3390/rs11050478>
14. Kumar, S. V.; Reichle, R. H.; Harrison, K. W.; Peters-Lidard, C. D.; Yatheendradas, S.; Santanello, J. A. A comparison of methods for a priori bias correction in soil moisture data assimilation. *Water Resour. Res.* **2012**, *48*(3), W03515. <https://doi.org/10.1029/2010WR010261>
15. Karthikeyan, L.; Pan, M.; Wanders, N.; Kumar, D. N.; Wood, E. F. Four decades of microwave satellite soil moisture observations: Part 1. A review of retrieval algorithms. *Adv. Water Resour.* **2017**, *109*, 106–120. <https://doi.org/10.1016/j.advwatres.2017.09.006>
16. Han, X.; Li, X.; Hendricks Franssen, H. J.; Vereecken, H.; Montzka, C. Spatial horizontal correlation characteristics in the land data assimilation of soil moisture. *Hydrol. Earth Syst. Sci.* **2012**, *16*(5), 1349–1363. <https://doi.org/10.5194/hess-16-1349-2012>
17. Yan, H.; Moradkhani, H. Combined assimilation of streamflow and satellite soil moisture with the particle filter and geostatistical modelling. *Adv. Water Resour.* **2016**, *94*, 364–378. <https://doi.org/10.1016/j.advwatres.2016.06.002>
18. Sahoo, A. K.; De Lannoy, G. J.; Reichle, R. H.; Houser, P. R. Assimilation and downscaling of satellite observed soil moisture over the Little River Experimental Watershed in Georgia, USA. *Adv. Water Resour.* **2013**, *52*, 19–33. <https://doi.org/10.1016/j.advwatres.2012.08.007>
19. Evensen, G. The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean Dyn.* **2016**, *53*(4), 343–367. <https://doi.org/10.1007/s10236-003-0036-9>
20. Rouf, T.; Girotto, M.; Houser, P.; Maggioni, V. Assimilating satellite-based soil moisture observations in a land surface model: The effect of spatial resolution. *J. Hydrol. X* **2021**, *13*, 100105. <https://doi.org/10.1016/j.hydroa.2021.100105>
21. Baguis, P.; Roulin, E. Soil moisture data assimilation in a hydrological model: a case study in Belgium using large-scale satellite data. *Remote Sens.* **2017**, *9*(8), 820. <https://doi.org/10.3390/rs9080820>
22. López López, P.; Wanders, N.; Schellekens, J.; Renzullo, L. J.; Sutanudjaja, E. H.; Bierkens, M. F. Improved large-scale hydrological modelling through the assimilation of streamflow and downscaled satellite soil moisture observations. *Hydrol. Earth Syst. Sci.* **2016**, *20*(7), 3059–3076. <https://doi.org/10.5194/hess-20-3059-2016>
23. De Santis, D.; Biondi, D.; Crow, W. T.; Camici, S.; Modanesi, S.; Brocca, L.; Massari, C. Assimilation of satellite soil moisture products for river flow prediction: An extensive experiment in over 700 catchments throughout Europe. *Water Resour. Res.* **2021**, *57*(6), e2021WR029643. <https://doi.org/10.1029/2021WR029643>
24. Massari, C.; Camici, S.; Ciabatta, L.; Brocca, L. Exploiting satellite-based surface soil moisture for flood forecasting in the Mediterranean area: State update versus rainfall correction. *Remote Sens.* **2018**, *10*(2), 292. <https://doi.org/10.3390/rs10020292>
25. Leroux, D. J.; Pellarin, T.; Vischel, T.; Cohard, J. M.; Gascon, T.; Gibon, F.; ...; Seguis, L. Assimilation of SMOS soil moisture into a distributed hydrological model and impacts on the water cycle variables over the Ouémé catchment in Benin. *Hydrol. Earth Syst. Sci.* **2016**, *20*(7), 2827–2840. <https://doi.org/10.5194/hess-20-2827-2016>
26. Patil, A.; Ramsankaran, R. A. A. J. Improved streamflow simulations by coupling soil moisture analytical relationship in EnKF based hydrological data assimilation framework. *Adv. Water Resour.* **2018**, *121*, 173–188. <https://doi.org/10.1016/j.advwatres.2018.08.010>
27. Cenci, L.; Pulvirenti, L.; Boni, G.; Chini, M.; Matgen, P.; Gabellani, S.; ...; Pierdicca, N. An evaluation of the potential of Sentinel 1 for improving flash flood predictions via soil moisture–data assimilation. *Adv. Geosci.* **2017**, *44*, 89–100. <https://doi.org/10.5194/adgeo-44-89-2017>
28. Ciupak, M.; Ozga-Zielinski, B.; Adamowski, J.; Deo, R. C.; Kochanek, K. Correcting satellite precipitation data and assimilating satellite-derived soil moisture data to generate ensemble hydrological forecasts within the HBV rainfall-runoff model. *Water* **2019**, *11*(10), 2138. <https://doi.org/10.3390/w11102138>
29. Patil, A.; Ramsankaran, R. A. A. J. Improving streamflow simulations and forecasting performance of SWAT model by assimilating remotely sensed soil moisture observations. *J. Hydrol.* **2017**, *555*, 683–696. <https://doi.org/10.1016/j.jhydrol.2017.10.058>
30. Meng, S.; Xie, X.; Liang, S. Assimilation of soil moisture and streamflow observations to improve flood forecasting with considering runoff routing lags. *J. Hydrol.* **2017**, *550*, 568–579. <https://doi.org/10.1016/j.jhydrol.2017.05.024>
31. Tian, S.; Renzullo, L. J.; Pipunic, R. C.; Lerat, J.; Sharples, W.; Donnelly, C. Satellite soil moisture data assimilation for improved operational continental water balance prediction. *Hydrol. Earth Syst. Sci.* **2021**, *25*(8), 4567–4584. <https://doi.org/10.5194/hess-25-4567-2021>
32. Mao, Y.; Crow, W. T.; Nijssen, B. A framework for diagnosing factors degrading the streamflow performance of a soil moisture data assimilation system. *J. Hydrometeorol.* **2018**, *20*(1), 79–97. <https://doi.org/10.1175/JHM-D-18-0115.1>
33. Fairbairn, D.; Barbu, A. L.; Napolé, A.; Albergel, C.; Mahfouf, J. F.; Calvet, J. C. The effect of satellite-derived surface soil moisture and leaf area index land data assimilation on streamflow simulations over France. *Hydrol. Earth Syst. Sci.* **2017**, *21*(4), 2015–2033. <https://doi.org/10.5194/hess-21-2015-2017>

34. De Lannoy, G. J.; Reichle, R. H. Assimilation of SMOS brightness temperatures or soil moisture retrievals into a land surface model. *Hydrol. Earth Syst. Sci.* **2016**, *20*(12), 4895-4911. <https://doi.org/10.5194/hess-20-4895-2016>

35. Ridler, M. E.; Madsen, H.; Stisen, S.; Bircher, S.; Fensholt, R. Assimilation of SMOS-derived soil moisture in a fully integrated hydrological and soil-vegetation-atmosphere transfer model in Western Denmark. *Water Resour. Res.* **2014**, *50*(11), 8962-8981. <https://doi.org/10.1002/2014WR015392>

36. Jackson, J.K.; Huryn, A.D.; Strayer, D.L.; Courtemanch, D.L.; Sweeney, B.W. Atlantic Coast Rivers of the Southeastern United States. In *Rivers of North America*, 1st ed.; Benke, A.C., Cushing, C.E., Eds.; Elsevier Academic Press: USA, 2005; pp. 20-71.

37. Ray, R. L.; Beighley, R. E.; Yoon, Y. Integrating runoff generation and flow routing in Susquehanna River Basin to characterize key hydrologic processes contributing to maximum annual flood events. *J. Hydrol. Eng.* **2016**, *21*(9), 04016026. [https://doi.org/10.1061/\(ASCE\)HE.1943-5584.0001389](https://doi.org/10.1061/(ASCE)HE.1943-5584.0001389)

38. DePhilip, M.; Moberg, T. Ecosystem flow recommendations for the Susquehanna River basin. The Nature Conservancy: Harrisburg, PA, USA, 2010.

39. Schaefer, G. L.; Cosh, M. H.; Jackson, T. J. The USDA natural resources conservation service soil climate analysis network (SCAN). *J. Atmos. Oceanic Technol.* **2007**, *24*(12), 2073-2077. <https://doi.org/10.1175/2007JTECHA930.1>

40. Bell, J. E.; Palecki, M. A.; Baker, C. B.; Collins, W. G.; Lawrimore, J. H.; Leeper, R. D.; ...; Diamond, H. J. US Climate Reference Network soil moisture and temperature observations. *J. Hydrometeorol.* **2013**, *14*(3), 977-988. <https://doi.org/10.1175/JHM-D-12-0146.1>

41. Gochis, D. J.; Barlage, M.; Cabell, R.; Casali, M.; Dugger, A.; FitzGerald, K.; ...; Zhang, Y. The WRF-Hydro modeling system technical description, Version (5.1.1). NCAR Technical Note, 2020.

42. Niu, G. Y.; Yang, Z. L.; Mitchell, K. E.; Chen, F.; Ek, M. B.; Barlage, M.; ...; Xia, Y. The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements. *J. Geophys. Res.: Atmos.* **2011**, *116*, D12109. <https://doi.org/10.1029/2010JD015139>

43. Skamarock, W. C.; Klemp, J. B.; Dudhia, J.; Gill, D. O.; Liu, Z.; Berner, J.; ...; Huang, X.Y. A description of the advanced research WRF version 4, NCAR Technical Note, 2019. <https://doi.org/10.5065/1dfh-6p97>

44. Hersbach, H.; Bell, B.; Berrisford, P.; Hirahara, S.; Horányi, A.; Muñoz-Sabater, J.; ...; Thépaut, J. N. The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.* **2020**, *146*(730), 1999-2049. <https://doi.org/10.1002/qj.3803>

45. Asadzadeh, M.; Tolson, B. Pareto archived dynamically dimensioned search with hypervolume-based selection for multi-objective optimization. *Eng. Optim.* **2013**, *45*(12), 1489-1509. <https://doi.org/10.1080/0305215X.2012.748046>

46. Tolson, B. A.; Shoemaker, C. A. Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. *Water Resour. Res.* **2007**, *43*, W01413. <https://doi.org/10.1029/2005WR004723>

47. Loizu, J.; Massari, C.; Alvarez-Mozos, J.; Tarpanelli, A.; Brocca, L.; Casali, J. On the assimilation set-up of ASCAT soil moisture data for improving streamflow catchment simulation. *Adv. Water Resour.* **2018**, *111*, 86-104. <https://doi.org/10.1016/j.advwatres.2017.10.034>

48. Sun, L.; Seidou, O.; Nistor, I.; Liu, K. Review of the Kalman-type hydrological data assimilation. *Hydrol. Sci. J.* **2015**, *61*(13), 2348-2366. <https://doi.org/10.1080/02626667.2015.1127376>

49. Montzka, C.; Pauwels, V. R. N.; Franssen, H. H.; Han, X.; Vereecken, H. Multivariate and multiscale data assimilation in terrestrial systems: A review. *Sensors* **2012**, *12*, 16291-16333. <https://doi.org/10.3390/s121216291>

50. Wanders, N.; Karssenberg, D.; De Roo, A.; De Jong, S. M.; Bierkens, M. F. P. The suitability of remotely sensed soil moisture for improving operational flood forecasting. *Hydrol. Earth Syst. Sci.* **2014**, *18*(6), 2343-2357. <https://doi.org/10.5194/hess-18-2343-2014>

51. Reichle, R. H.; Koster, R. D. Assessing the impact of horizontal error correlations in background fields on soil moisture estimation. *J. Hydrometeorol.* **2003**, *4*(6), 1229-1242. [https://doi.org/10.1175/1525-7541\(2003\)004<1229:ATIOHE>2.0.CO;2](https://doi.org/10.1175/1525-7541(2003)004<1229:ATIOHE>2.0.CO;2)

52. Reichle, R. H.; Koster, R. D. Bias reduction in short records of satellite soil moisture. *Geophys. Res. Lett.* **2004**, *31*, L19501. <https://doi.org/10.1029/2004GL020938>

53. Renzullo, L. J.; Van Dijk, A. I. J. M.; Perraud, J. M.; Collins, D.; Henderson, B.; Jin, H.; ...; McJannet, D. L. Continental satellite soil moisture data assimilation improves root-zone moisture analysis for water resources assessment. *J. Hydrol.* **2014**, *519*, 2747-2762. <https://doi.org/10.1016/j.jhydrol.2014.08.008>

54. Kumar, S. V.; Reichle, R. H.; Koster, R. D.; Crow, W. T.; Peters-Lidard, C. D. Role of subsurface physics in the assimilation of surface soil moisture observations. *J. Hydrometeorol.* **2009**, *10*(6), 1534-1547. <https://doi.org/10.1175/2009JHM1134.1>

55. Yilmaz, M. T.; Crow, W. T. The optimality of potential rescaling approaches in land data assimilation. *J. Hydrometeorol.* **2013**, *14*(2), 650-660. <https://doi.org/10.1175/JHM-D-12-052.1>

56. Lievens, H.; Tomer, S. K.; Al Bitar, A.; De Lannoy, G. J.; Drusch, M.; Dumedah, G.; ...; Pauwels, V. R. N. SMOS soil moisture assimilation for improved hydrologic simulation in the Murray Darling Basin, Australia. *Remote Sens. Environ.* **2015**, *168*, 146-162. <https://doi.org/10.1016/j.rse.2015.06.025>

57. Samuel, J.; Coulibaly, P.; Dumedah, G.; Moradkhani, H. Assessing model state and forecasts variation in hydrologic data assimilation. *J. Hydrol.* **2014**, *513*, 127-141. <https://doi.org/10.1016/j.jhydrol.2014.03.048>

58. Trudel, M.; Leconte, R.; Paniconi, C. Analysis of the hydrological response of a distributed physically-based model using post-assimilation (EnKF) diagnostics of streamflow and in situ soil moisture observations. *J. Hydrol.* **2014**, *514*, 192-201. <https://doi.org/10.1016/j.jhydrol.2014.03.072>