Dear Niko Verhoest,

We are very pleased to inform you that the revised manuscript has been uploaded.

The major changes in the manuscript include:
- The addition of a more detailed description of the model calibration in a way that the paper is stand alone (mainly section 2.5).
- More detailed information on the modification in the LISFLOOD model (as requested by both reviewer, section 2.2 and 2.5).
- Two extra scenarios as requested by reviewer #1 (Table 1-2, Section 2.6 and 3.4). A Figure with time series of update soil moisture simulation and the observations from different sensors.
- As requested by reviewer #2 we made a better distinction between EFAS and LISFLOOD.
- A flowchart is added describing the entire model setup to allow the reader to have a better understanding of the assimilation and forecasting procedures.
- More detailed information on the assimilation procedure (Section 2.4)
- Additional figure including a example time serie of soil moisture simulations (Appendix Figure 10)

Minor changes include:
- Adjustment of some figure captions for more clarification
- Corrections of sentences by reviewers

Please find attached the response to the reviewers which was submitted in February. Where possible line or section number were added.

Kind regards,
Niko Wanders

Comments to the Author:
Dear

based on the excellent reviews of both reviewers, I believe that your paper has large potential for being published in HESS. However, your manuscript requires a major revision that accounts for the comments raised by both reviewers.
I invite you to submit a revised version which will be sent out for review again.
Kind regards,
Niko Verhoest

Reviewer #1

Overview
The study investigates the assimilation of satellite soil moisture and in situ discharge observations into the European Flood Awareness System (EFAS). The Upper Danube River at Bratislava (135000 km2) is used as case study where a dense network of gauging stations (23) is available. The performance of EFAS with and without the assimilation of soil moisture and discharge is evaluated both for simulating and forecasting (lead times up to 10 days) discharge and by using different configurations (also for model calibration).

General Comments
The paper investigates a very important topic related to the use of satellite soil moisture data for improving operational flood forecasting. Being highly interested to this topic, I quickly and carefully read the paper that I found well written and structured. Moreover, for the first time the assimilation of BOTH DISCHARGE AND SOIL MOISTURE is tested for a LARGE RIVER BASIN (135000 km2) into a DISTRIBUTED and OPERATIONAL hydrological model. All these aspects (in capitals) should be probably better underlined in the paper. Finally, the amount of elaborations and analyses behind the results shown in the paper is quite impressive.

The authors are grateful for the kind words and constructive comments from the reviewer. We are pleased to note that the reviewer acknowledges the potential impact and importance of the paper. The questions raised by the reviewer are answered below. We separated comments related to soil moisture and to discharge. We will put more emphasis on the unique aspects of this study as suggested by the reviewer.

However, I found some important issues that should be addressed before the publication. The paper only shows the final results obtained after the assimilation (or not) of soil moisture and discharge observations into EFAS. However, also the intermediate results should be given to understand the reasons for which improvements (or not) are obtained. Specifically, I listed below the results that should be added (in my opinion) to help understanding the content of the paper (note that some of them could be reported in an Appendix section).

We will answer the questions of the reviewer in the section below. Information will be added to the paper and we will perform some additional analysis.

Soil Moisture
Three different satellite products (from ASCAT, SMOS and AMSR-E) are assimilated. However, they are obtained with different algorithms and sensors (active and passive microwave, C- and L-band) and, usually, they show temporal patterns quite dissimilar in terms of dynamics (not in absolute terms because all the products are correctly rescaled to the same range before the assimilation). Therefore, I expect that their mutual assimilation may generate some issues. How are the products integrated? Are they assimilated at the satellite overpass time or by computing daily averages? This information should be clarified.

The three microwave products are indeed obtained by different sensors and are computed by different retrieval algorithms. As a result, the temporal evolution and spatial patterns of soil moisture will sometimes be different between the satellite products. The satellite products are assimilated simultaneously as daily averages. Satellite products are integrated in the observation matrix of the EnKF. However, the correlation between the errors of different sensors is conserved and used in the assimilation and the error covariance matrix of the observations. The relations for the errors are obtained from a previous study by Wanders et. al. 2012. Variograms are calculated for the estimated errors of the observations over Spain as

well as the cross-variograms of the error (see attached figure).
For the scenarios that include discharge, the discharge observations are simultaneously assimilated on a daily basis. It was decided to assimilate daily average discharge values, because the exact time of the discharge observations is unknown and the model uses daily meteorological input. We acknowledge that this might result in a small temporal mismatch. However, due to the uncertainty in the exact observation times and the temporal resolution of the meteorological input we decided to assimilate daily averages of the observations. This information will be added to the text of the revised version.
The model prediction errors are directly calculated from the model ensemble and are not correlated to the observation errors.

Moreover, some figures showing the comparison between modelled and observed soil moisture data, also subdivided by sensors, should be included. Specifically, it could be very interesting to see the soil moisture dynamic for the surface layer (where soil moisture data are directly assimilated) and the root-zone before and after the assimilation. In fact, recent studies (Chen et al., 2011; Brocca et al., 2012) have obtained that the assimilation of surface soil moisture has a very limited impact on the root-zone. Consequently, the assimilation has little impact on discharge simulation that is mainly driven from the root-zone soil moisture.
As suggested by the reviewer a figure will be added in which timeseries of soil moisture dynamics of the surface layer are shown as well as the observations from the different soil moisture sensors. The impact of the update on the shallow layers of the model (0-2 and 2-5 cm) will be shown in the same figure. These two shallow layers correspond with the layers observed by the microwave sensors ASCAT and AMSR-E: 0-2 cm and SMOS: 0-5 cm. (Appendix Figure 10)

Reading the paper, it seems that the assimilation of soil moisture has a significant impact (in contrast with previous studies). This depends on the assumptions made for the observations and modelling errors. However, little information is given on these errors. For instance, which is the relation between modelling and observation errors? Why is the impact significant? Which is the correlation between the surface and root-zone soil moisture? Which is the depth of the soil layers used in the model? An answer to all these questions should be provided.
The errors of the different sensors are obtained from Wanders et. al. 2012. In this study we calculate the individual satellite observation errors as well as the correlation between the errors of different sensors and their spatial correlation. This provides us with the unique opportunity to fully calculate the multi-sensor error covariance matrix without making assumptions on the intercorrelation of different sensors.
The correlation between the different soil moisture layers of the model is relevant information and it is by definition positive: a wetter topsoil will result in wetter subsoil due to increased percolation, although this correlation is highly dependent on soil texture and other soil parameters. Soil moisture has quite some impact for areas where discharge is mostly determined by catchment-scale runoff instead of travel time through the rivers, i.e. higher up in the basin. As catchment-scale runoff in LISFLOOD is dependent on soil wetness (through surface runoff and interflow), changing surface soil moisture has an impact in these areas. The authors decided not to include maps of these correlations in the paper, since they are dependent on calibration/assimilation scenario and will change during assimilation.
A more detailed description of the model modifications (compared to the original LISFLOOD model) will be provided. In this study we added two additional soil layers. This first layer represents a soil depth of 0-2 cm and the second layer represents 2-5 cm. The depth of these layers is chosen in such a way that the soil moisture in these layers can directly be compared to the observations of SMOS (0-5cm), ASCAT (0-2cm) and AMSR-E (0-2cm). We will include this information in the revised paper (Section 2.2 and 2.3.1)

The assimilation of only soil moisture data (without discharge) is only considered for the configuration where no discharge data are used for model calibration (Q0sat). For really understand the impact of soil moisture assimilation, the configuration where the model is well calibrated (with 1 or 7 discharge stations) and ONLY soil moisture is assimilated should be considered. This is missing in the paper. At the same time, the benchmark simulations should be done by using 1 or 7 stations for the calibration, and without the assimilation of discharge. Also this configuration is missing in the paper.
To obtain a more in-depth analysis of the soil moisture performance we agree that the assimilation of soil moisture into a model calibrated on discharge and soil moisture is an important test. This information will be included as one of the scenarios in the Results section. The scenario where the model is calibrated on 7 discharge stations and soil moisture will be used for the assimilation of only soil moisture. In the terminology of the paper this will be a calibration based on Q7Sat and a forecasting scenario with Q0Sat. We will compare this scenario with the proposed benchmark scenario. We will add the results to Table 2. Results will be described and discussed in the main text (Section 2.6, 3.4 and Table 2).

Discharge
The simulation that considers the discharge observed at Bratislava for the model calibration (Q1) shows a consistent overestimation for the whole period (Figure 3). I do not expect this as after the calibration the modelled discharge should be closer (and unbiased) with respect to observations. Do the authors have some explanations for that? I believe that more information can be found in the paper submitted on WRR (Wanders et al., 2013) that is not available to reviewers. I suggest adding this paper in future submissions of the paper, as it appears to be relevant for understanding the content of the current paper.
The constant overestimation of discharge in the Q1 scenario is related to the time period selected for the hindcasting experiment. The model was calibrated on the period 2010-2011 and no biases were found for this period during the calibration. However, for 2011 we found some overestimation of the simulated discharge compared to the observation at the outlet. For other locations in the catchment this is not the case. So overall discharge is not overestimated, with and exception at the outlet, for the selected period. We will shortly explain this in the revised manuscript (Section 3.1).

Moreover, it is not clear if the assimilation of discharge is used for correcting the soil moisture states of the model. If yes (as I expect), which are the soil layers for which the assimilation has a significant effect? Which function/operator is used to update soil moisture states from discharge observations? Is it considered a time lag between discharge observations and soil moisture states? Can the authors address these issues?
The Ensemble Kalman filter used is capable of adjusting soil moisture in all layers. This is done by including soil moisture for all model layers in the state vector. In this set up used, assimilation of discharge mainly results in an adjustment of soil moisture in the upper two layers (up to a depth of 5 cm); below this depth, the adjustment is negligible. We will add this information to the manuscript (Appendix Figure 10).
A time-lag has not been used to update the soil moisture with discharge observations, since the correlation length between discharge and soil moisture varies throughout the catchment and would require additional assumptions on the relation between soil moisture and discharge. A flowchart will be added to describe the assimilation procedure in more detail. The flowchart is also added to this rebuttal as a figure (Figure 3).

Finally, the calibration, validation and assimilation periods are coincident. This is usually not good and clearly does not represent the real-time configuration when the model is run for future periods. Is it possible to consider the model calibration in a different time period?

We agree with the reviewer that ideally the calibration, validation and assimilation period should be separated. However, due to the failure of AMSR-E and the problems for SMOS with RFI over Europe, the period available for assimilating the soil moisture data of all three sensors is limited. With the launch of AMSR-2 the assimilation and validation of the hindcasting experiment could be extended to mid-2012 up to mid-2013. However, since AMSR-2 data is new and no studies have been done into the errors of this instrument it is not feasible to use these observations for this experiment.
Thus, we opted for the next best solution, i.e. to still have a spatial split sample approach for calibration and validation. Since different stations have been used for the calibration and validation, information is only used once. This will still ensure some measure of independence between the calibration and validation (Section 2.1).

Moreover, probably I missed something, but I didn’t found how the model performs for the discharge stations not used for model calibration (shown in Figure 1). Can the authors show these results?
The performance of the EnKF in the validation (Figure 5, 6, 7 and 8 and Table 2) is calculated using information at the validation locations only. These are different from the locations used in the calibration. This is mentioned in the caption of most Figures and Tables; however we will also include this in section 2.1.

In the Specific Comments I reported a number of corrections/explanations that are required.
On this basis, I feel that the paper deserve to be published on HESS as it addresses a very important and new topic but a major revision is required.

Specific Comments/ Technical Corrections (P: page, L: line or lines)

P13785, L20: "... correct incorrect ...". Please revise.
We will modify the text accordingly.

P13786, L15-16: Bolten et al. (2010) and Liu et al. (2011) did not consider the discharge simulation. Likely they are not appropriate here. On the other hand, some recent papers could be mentioned and discussed (Chen et al., 2011; Matgen et al., 2012). Unfortunately, only few studies on this topic are available so far.
We agree with the reviewer. It is difficult to find literature on this topic. We will add the references provided by the reviewer.

P13786, L21-22: "The potential to improve flood forecast...". This sentence is not well connected to the previous one. Please revise.
We will revise the text.

P13786, L22: "...studies mainly study...". Please revise.
We will revise the text.

P13786, L29: Actually, the assimilation of both discharge and soil moisture for a real case study was only considered by Aubert et al. (2003) but using in situ soil moisture observations. To my knowledge, the assimilation of discharge and satellite soil moisture data has not been studied so far. I suggest changing the sentence "... not been extensively explored".
We will modify the text according to the reviewer’s suggestion

P13787, L5: The research questions are three, not two.
We will modify the text.

P13787, L25: It should be km2 .
We will modify the text according to the reviewer’s suggestion.

P13788, L8: Change "because" with "become".
We will modify the text according to the reviewer’s suggestion.

P13789, L21: The revisit time of satellite soil moisture data should be 1 day.
We will modify the text according to the reviewer’s suggestion.

P13791, L23: Satellite soil moisture data are not always available, how are they assimilated (see General Comments)?
We use the data when available. When observations from a satellite are not available at a particular day or location the data from that satellite are only excluded on that particular day or location. We will modify the text and explain this in a better way.

P13793, L6: Add "soil moisture" between satellite and observations "satellite soil moisture observations".
We will modify the text according to the reviewer’s suggestion

P13795, L11: It should be F...(x, t), t is missing.
We will modify the text according to the reviewer’s suggestion

P13797, L5-7: It is highly expected that the assimilation reduces the spread of the simulations. With the assimilation, the model is constrained to follow observations and, hence, the spread reduces. This result is not an added-value of the assimilation, please revise.
We will modify the text according to the reviewer’s suggestion

P13797, L16: Change "of 0.08" with "to 0.08".
We will modify the text according to the reviewer’s suggestion

P13797, L23: Change "none" with "not".
We will modify the text according to the reviewer’s suggestion

P13797, L28: Remove "the" from "that this the method".
We will modify the text according to the reviewer’s suggestion

P13798, L20: Change "assimilation" with "assimilated".
We will modify the text according to the reviewer’s suggestion

P13810, L10: Remove "are reduced", it is repeated in the following line.
Figure 3: The last sentence of the captions should be moved in that of Figure 4. The opposite for the last sentence of the caption of Figure 4.
Figure 8: This figure should be explained better both in the caption and in the text of the manuscript.
We will improve the text to make it better understandable to the reader.

Additional References
F. Chen, W. T. Crow, P. J. Starks, and D. N. Moriasi, "Improving hydrologic predictions of catchment model via assimilation of surface soil moisture," Adv. Water Resour., vol. 34, pp. 526-535, 2011.
P. Matgen, F. Fenicia, S. Heitz, D. Plaza, R. de Keyser, V. R. N. Pauwels, W. Wagner,

and H. Savenije, "Can ASCAT-derived soil wetness indices reduce predictive uncer-
tainty in well-gauged areas? A comparison with in situ observed soil moisture in an assimilation application," Adv. Water Resour., 44, 49-65, 2012

Reviewer #2

Review "The suitability of remotely sensed soil moisture for improving operational flood forecasting" by Wanders et al

The manuscript describes a case study for assimilating discharge and satellite soil moisture observations into a Lisflood model for the Upper Danube river and evaluates the performance on the forecast quality (one year).

Main Comments

The manuscript reads like a feasibility study. The manuscript is in parts inaccurate and unclear and overstates and generalizes the conclusions (as maybe typical for a feasibility studies applied to one study area) too much. This is also due to the fact that the term EFAS is often used, while the use of the term Lisflood model for the Upper Danube is more appropriate. The manuscript raises more questions than it answers.

The authors do not fully agree with the reviewer. This is the first study that simultaneously assimilates observed discharge and remotely sensed soil moisture data for a large-scale river basin into a spatially distributed operational hydrological modelin retro-active forecasting mode. Due to limitation in available spatially distributed discharge data and computation times, it is not easy to extent this study to other areas. We answer the questions raised by the reviewer below and will modify the script in such a way that the questions raised by the reviewer will be answered by the manuscript wherever useful and possible.

Abstract

page 13784

Line 2-4: Replace EFAS with Lisflood model for the Upper Danube (throughout the abstract)

Line 4-5: Remove line about EFAS not sure why this needs to be in the abstract

Line 14: Replace show by suggest

We agree with the reviewer on the questions raised above, that it might be confusing when we use EFAS too often. Therefore we propose to use EFAS in the context of flood forecasting and not when talking about data assimilation. The entire purpose of the work described in this manuscript is eventually to improve operational forecasting in EFAS (or similar systems) by using data assimilation of satellite soil moisture and discharge.

In the model runs described in the manuscript we used exactly the same input (meteorological and catchment characteristics) and setup as defined in the EFAS system. This setup was provided by the JRC EFAS team. Also the input data for the ensemble forecast are identical to the data used by EFAS at JRC. Since the EFAS system is developed at JRC and because we use identical input data for our experiment we used the term EFAS in the paper. J. Thielen (also working at EC-JRC, in the operational unit for EFAS) checked the manuscript for appropriate use of the terms EFAS and LISFLOOD. In addition, co-author A. De Roo is both the developer of LISFLOOD and the initial EFAS system. Therefore, the authors believe that the broader system can be called EFAS in the paper and this thus does not need to be changed throughout the manuscript. Additionally, detailed information on EFAS can be found in other literature giving the reader more insight into the exact forecasting procedure, frequency and performance.

We agree with the reviewer that it would be more appropriate to talk about LISFLOOD when

dealing with the assimilation. Observations are assimilated into the LISFLOOD model and not into EFAS. We modified the text accordingly and hope that the reviewer agrees with our point of view on the terminology (Section 2.2 has been completely revised).

Line 14-25: I don think this remarks are valid for Q1sat so you can not generalize
We agree. We will modify the text to be more specific on the exact performance of the satellites under different calibration scenarios.

Page 13785
Line 1 Again I think show is to strong (there are too many thing unclear/not understood/ explained etc and validation is very limited)
Please see answer to previous question.

Introduction
page 13785

Line 11: Maybe also good to mention several other forecasting systems used in Europe (England, Scotland, France, Switserland, Austria, The Netherlands, etc.)
We will mention other forecasting systems in the introduction as suggested by the reviewer (Introduction)

Line 12/14: I find the statement "national forecasting systems are often not sufficient and transboundary forecasting systems are required" strange. What is a transboundary forecasting system? What do you mean by often (which rivers)? As an example, several countries (i.e. The Netherlands, Germany) run a flood/flow forecasting system in which transboundary rivers are modelled and used to generate forecasts for their own national domain.
We agree with the reviewer that positive examples exist where the cooperation between countries is good to excellent and in such cases the need for a system like EFAS is limited. However, the added value of EFAS has been proven in many transboundary flood events like the recent flooding of Eastern Germany and Poland or the recent flooding in the Danube, where many countries were affected. EFAS will have added value to national systems and can be used in emergency situations to provide useful additional information, especially for longer lead times (Introduction).

Line 13/15: I think the EFAS system was not developed to full fill the need described here, but was developed in support of crisis management at the European level.
We will add the international crisis management in the objectives of EFAS (Introduction).

Line 24: For Example. What do the authors mean here? This is a not an example related to the lines above on state updating. Please remove this sentence as it is not appropriate/relevant here
We will remove the sentence.

Line 26: However, it is difficult to obtain these measurements in real-time in a way they can be used in EFAS. Can the authors specify in more detail what the problem is (in a way they can be used in EFAS?)? Is the data not available in real-time? Or is there another issue? Please clarify
The problem is that there are practical barriers to obtaining discharge observations in real-time and getting them into the forecasting system in real-time. Often some delay exists between observation and data provision, and also authorities are not always willing to provide all observations in real-time. This issue should be addressed in the future. Note that satellite observations do not have this limitation. They are always near-real time available and are often free. We will clarify this in the text (Introduction).

page 13786
Line 11: The revisit time is 1-3 days how does this relate to the availability or usefulness of the discharge observations (in a way they can be used?)?
At the latitude of the Upper Danube the revisit time would be more like 1-2 days. With three sensors this results in almost two soil moisture observations per day from different sensors. Most days we even get three observations from different sensors. These observations are spatially distributed and the total number of observations is much higher compared to discharge. We will indicate that at the latitude of the Upper Danube observations are more frequent and we will include a table with the total number of observations from satellite sensors (Section 2.3.1).

Page 13788
Line25: One of the most difficult steps in a data assimilation setup is to determine the input and model uncertainty (PQR problem, see also Liu et al., HESS 2012 also for other relevant references). Here a setup is chosen which make use of 300 parameter sets, without any consideration what has been used by others even though this is the most critical step in the whole process. Because I don't know the study by Wanders et al. 2013 WRR I really cannot judge if this is correct, what the consequences are, etc I assume the bucket sizes vary between the different parameters set? How is this handled? Do you make use of maximum and minimum bounds in the data assimilation scheme? It is clear that this choice/assumption requires much more justification.
Is this also the way the operational system is being envisaged to run?
We fully agree that it is very important to correctly identify the relation between observation uncertainty (R) and model uncertainty (P) in the data assimilation. Without a correct assessment of R either updates are too strong or too weak, depending on the ratio between R and P. In this study we have done a detailed calculation of R, explicitly including the spatial correlation of the errors, the error cross correlation between the different sensors as well as their spatial intercorrelation. A figure with the error cross variogram is included in the rebuttal. Details on the calculation of R can be found in Wanders et al. (2012). Zero correlation has been assumed between the discharge observations and their correlation with the soil moisture observations. The standard error on the discharge observations has been put at 30% of the observed value, which is also confirmed by Di Baldassarre & Montanari (2009) (Page 13793 line 5). In our opinion, we have the best possible description of R. The calculation of P is done within the EnKF. Each of the members uses a different parameter set, which is determined by the calibration of the model for the Upper Danube. No upper or lower boundaries have been applied to the parameter calibration. State variables are limited to field capacity and wilting point for the soil moisture and zero discharge and groundwater (no upper limit).

Is it correct that per setup Q0, Q1, Q7 300 different parameter sets are being used at least this is the way I read it and if not I don't understand the setup? What is the difference between those sets in terms of states/spreads/correlation in space/time etc.
The reviewer is correct. Different parameter sets are used per forecasting scenario as described in Table 1. Scenarios could be seen as forecasting situations where either soil moisture observations, discharge observations or both observation types are available. Therefore we also calibrate the model with the same available data used for forecasting. So, for instance, when only remote sensing data can be used for the forecasting, only remote sensing data is used in the calibration, representing a catchment for which only remote sensing data is available (no discharge). This is interesting for data-poor catchments, where these particular scenarios could occur (Section 2.6, 3.4 and Table 2).

Line 5/6 "while the current EFAS uses fixed initial conditions for the hydrological forecast" I

assume this is not correct or does EFAS use fixed initial conditions for each hydrologic forecast?
We will modify the text to be more clear on this. EFAS uses fixed initial conditions for each ensemble member. Only meteorological forcing is different per ensemble member, based on the ECMWF-EPS. We wil modify the text to “while the current EFAS uses fixed initial conditions for each of the individual ensemble members of the hydrological forecast” .

Page 13790
Line 17-26: Here the assumption is made that the satellite soil moisture measurements (with 5cm or 2cm depth support) can be used for comparison with the Lisflood soil moisture bucket (theta_WP / theta_FC of the topsoil? How deep is the top soil?) through some scaling. What is the foundation/rational for doing so?
A more detailed description of the modifications to the model will be provided in the revised paper. In this study we added two additional soil layers to the original LISFLOOD model used in EFAS. The first layer represents a depth from 0-2 cm and the second layer from 2-5 cm. The depth of these layers is chosen in such a way that they can directly be compared to the observations of SMOS (0-5cm), ASCAT (0-2cm) and AMSR-E (0-2cm). The other two soil layers are: the third layer from 5cm to rooting depth (varies between 5 -180 cm depth below the surface) and the fourth layer representing rooting depth to soil depth (varies between 30 – 200cm below the surface) (Section 2.2)

Page 13791
Line5-10: Similar experimental setup using multiple interior discharge observation stations was used by Rakovec et al. 2012 and Lee et al 2012 (see http://www.hydrol-earth-syst-sci.net/16/2233/2012/hess-16-2233-2012.pdf)
We kindly thank the reviewer for the suggestions and we will add the suggested references in the Conclusions.

Page 13793
The experimental setup is not very clear given the objectives of the manuscript.
The effect of data assimilation is very much depending on the perturbation of the model (already mentioned above)...
Is the operational setup also based on running the model with 300 parameter sets? Or is the setup here ad-hoc or opportunity based? If the latter it is probably better to mention this upfront instead of presenting it as they way the operational setup will work;
Why was the choice made to run the forecasts over the calibration period (not knowing what the calibration entails) a more independent testing of the setup seems more appropriate.
The operational setup is not based on 300 parameters sets, since this is not required for the current set-up of EFAS. However, all the other input data and forecasting frequency is identical to the original set-up (as stated above). We will make it more clear in the text that this slightly deviates from the original set-up of EFAS. However, the reader can find al details on EFAS in the two papers mentioned in the manuscript.
The choice to use the calibration period also for the forecasting was based on the fact that the overlapping period of the satellite data was limited. Due to the fact that this period is limited we are forced to use the same period. The authors acknowledge that this is not ideal. However, other alternatives are even less attractive. The use of only two sensors would reduce the number of observations and would possibly also have an effect on the possible impact of the assimilation. We wanted to study the full potential, a goal which would not be achieved by only using two sensors. Please note that we did use a split sample approach in the spatial domain: different discharge locations were used for calibration and validation.
At the moment observations are available from all three sensors because a new AMSR-E II sensor has been launched. However, the length of the timeseries is not sufficient to make the data set useful for our study. Future studies could use these observations when the new

AMSR-E II observations are released to the general public. We contacted the VU Amsterdam and they are still busy with validation of the results, so it is unfortunately not an option for this study.

Line 5 The error on the discharge measurement is set according to expert knowledge. How did others treat this uncertainty their DA setup. Maybe better to refer what has been used and state why and if the authors deviated from this.
The observational error of discharge has been set according to a study of Di Baldassarre and Montanari, 2009. Other studies report values as low as 0.1. We believe that it is better to not underestimate errors in the discharge observations in the Upper Danube catchments, because detailed information on the observations routines and associated uncertainty lacks. Therefore, the authors believe that a discharge observation error of 0.3 is not unrealistic. The reference to the paper of Di Baldassarre and Montanari, 2009 (Section 2.4)

Page 13794
Line 1-10: By using only two (random) parameter sets the remaining initial condition uncertainty is basically removed in the forecast, in other words after the analysis we fall back to a more or less deterministic model in forecast mode driven by EPS? Why 2 random parameter sets? And not 1, 4 or 6 etc?
The authors did a test (not included in the original mansucript) to determine the required number of runs needed to fully describe the distribution in the discharge hindcast. A full set of 15300 forecasts was performed for scenario Q7 for the entire hindcast period. Using this complete set a boot strap procedure was performed to identify the minimum number of initial conditions per meteorological forecasting required to correctly describe the probability density function at any moment in time. Using a Kolmogorov-Smirnoff test it was shown that the distribution was identical to the full set (alpha 0.05). This test indicated that the initial conditions have the largest impact for the short lead times (up to 4 days), hereafter the initial conditions did not impact the discharge distribution anymore and meteorological conditions are more dominant. We will add this additional information to the revised version of the paper, describing the experiment in a number of lines.
The authors believe that it would of course be preferable to run a more extended set of forecasts. However, due to long computation times per forecast and the large number of scenarios this has not been done in this study. Although larger samples are often better, the KS-test indicated that we have an appropriate sample size to compare our scenarios and this would also make it still doable computation time wise in an operational setting (Section 2.5).

Page 13795
Q90 and Q80 are chosen for evaluation? How many Q90 and Q80 events (for one measurement location) are included in this one year period?
This slightly depends on the location in the catchment. In total 1035 events are detected for the entire Upper Danube. However, some of these events last for more than one day bringing it down to approximately 10 events on average per location for the Q90. A total of approximately 16 events is detected for the Q80. The total number of 1035 is included in Figure 8 of the submitted HESSD paper (Section 2.7).

Page 13797
I am interested understanding what happens to the model states also in spatial sense. Can you provide insight into where what happens to which model state in spatial sense?
What is the reason why the flood peaks are overestimated for the Q1sat case?
What happens there (also in spatial sense to the updated model states)?
Why are all Skill Scores for the Q1sat case lower than for the Q1 case this seems very unlikely when looking at Figure 3 and Figure 4 especially for Q80/Q90?

Does averaging the results over the (scaled) measurements give a false sense of accuracy as the events are the same? How statistically significant are the results per location/leadtime/threshold? Is the hindcast period not to short and does it contain enough model-observation pairs to justify any statement for higher thresholds?
The reason peak flows are overestimated in the Q1sat case is that the model will do everything to correctly simulate discharge at the outlet. Thereby internal catchment processes are neglected and intermediate discharge stations are not taken into account. Additionally the model is updated to correctly simulate the soil moisture. This will result in high overland flow compared to the groundwater flow. This is the fastest way for the model to correctly simulate soil moisture and also soil moisture has the highest impact on overland flow generation.
The reason that Q1sat has lower skill scores (better flood forecasting performance) is caused by the fact that interior discharge observations are often better simulated. Especially discharge stations situated in the headwater show a better flood forecasting skill. The overland flow component (which is mostly affected by the soil moisture) has a larger contribution in the locations compared to the main stream, where the performance of Q1sat seems lower compared to the other scenarios (Figure 3 and 4).
For Figure 5-8 the score for each station is calculated over 365 observations per lead time. Each boxplot contains the 16 stations. This should be sufficient to analyze the performance of the assimilation framework, especially when compared to most other studies where only one objective is used to validate different scenarios. Often only the outlet is used to validate the general performance for different scenarios. We try to also assess the uncertainties in the skill scores with our approach, instead of only providing the reader with average results.
As stated in the caption of Figure 8 the total number of timesteps exceeding the Q90 is 1035, with an average of 10 flood events per station. We would not study higher thresholds, but we believe Q90 still contains enough observations to be able to be used for analysis. This is also the reason why in Figure 7 both the Q90 and Q80 are given. They give similar results indicating that it is still appropriate to use Q90.

When zooming in on Figure 4 I see that for several Q1sat is shifted away from the discharge observation at T0 (at the analysis) what happens there? What is the reason for the large spread with Q1sat in the forecast? Even when the peak discharge arrives at T0 (last panel), state updating is not able to draw the model towards the observation. What is the reason for this? What happens here?
The reviewer is correct and something is wrong in the plot (the small shift). This shift is the result of a bug in the script used to create the figure. We will improve the figure.
The reason for the large spread in this particular case is that the uncertainty in the model simulation is very small, such that the impact of the discharge observation is limited. However, the plotted analysis is the model results after assimilation (ψ^a). So the update is already included in the figure. The intermediate model results (ψ^f) are not shown in this figure.

Conclusions page 13799
Line 24/25 remove EFAS and use Lisflood model it suggests that results are valid for the modelled area (Upper Danube) .
As explained in the first reply we will mention EFAS in the context of forecasting and LISFLOOD when talking of assimilation. Hence we will not replace EFAS in this line. However, we will replace EFAS on page 13800, line 14 and replace it by LISFLOOD hydrological model (also page13801, line 12).

Conclusions page 13800
Line 8 “We show that the assimilation of remotely sensed soil moisture improves the flood forecasting especially” Is this true I would state “Our results suggest that the assimilation of

remotely sensed soil moisture improves the flood forecasting only". "we show" is, I think, overstated and I would use "suggest" (throughout manuscript and abstract)
We agree with the reviewer that this statement is only true when a combination of discharge and soil moisture data is used. We will change the sentence to "Our results indicate that the assimilation of remotely sensed soil moisture improves the flood forecasting only". Figure 6, 7 and 8 in combination with Table 2 indicate that an improvement in the flood forecasting skill is found when both discharge and soil moisture are used. An exception exists for the CRPS of Q1sat compared to Q1. All other values in the matrix indicate an improvement when the combination of discharge and soil moisture is used.
We will check the manuscript and modify incorrect statements of a similar kind.

Line 14/15 This is not true Q0sat gives worse results (see results/figures etc)
We will add that this is true except for Q0sat.

Line 19-27 Complete unclear. "This will ensure that the parameterization of the model is optimal for the correct simulation of the hydrological variables used in the assimilation framework" Was this shown? Is it an assumption by the authors? It is certainly not a conclusion in my eyes.
The reviewer is right. This paragraph does not belong in the conclusions and we will remove it.

Page 13801
Line 10-18 why coming with these results in the conclusions please move to the results section
We agree with the reviewer that this paragraph is out of place. We will remove this section and combine it with the suggestions additional scenario from reviewer #1 (calibrate on discharge, assimilate only satellites data). A new subsection in the results section will be created to describe these new scenarios which will be beneficial for the paper as a whole as well.

Line 19-27 This only valid for this Lisflood model for the Upper Danube, this model calibration, and the input data used and this setup. If the authors would use a model that was not biased, used different (better?) input data, the results might be completely different
Line 20 replace EFAS by Lisflood model for the Upper Danube
Line 21-23 "The addition of remotely sensed soil moisture will reduce the number of false positive flood alerts and thereby increase the reliability of the flood awareness system." Again only when many discharge observations are being used and these are not available in real-time in a way they can be used in EFAS (see introduction)
We will rephrase to: "In conclusion, we show that the uncertainty in the flood forecasts is reduced when discharge observations and satellite data are assimilated *into hydrological model of* the EFAS system *for the Upper Danube*. The addition of remotely sensed soil moisture to existing discharge observations reduces the number of false positive flood alerts and thereby increases the reliability of the flood awareness system. Although the number of the data available via satellite retrievals still remain a challenge in an operational system, the potential benefits could lead to a significant reduction in the false flood alerts, possibly also *for other catchments*. This will reduce the number of unnecessary precautions taken by the responsible governments and increase the confidence and willingness to act upon these *flood alerts*."

Overall
The authors do state in the introduction that "it is difficult to obtain these measurements

(discharge) in real-time in a way they can be used in EFAS” (maybe the same holds for the satellite data). Given the fact that the discharge observations are not usable, the authors remain very positive about the use of satellite data even when the results are not positive (and do not explain why) when only satellite data are used. I would expect a more balanced conclusion.

We agree with the reviewer that the findings on the satellite only scenario Q0sat, are overstated and will modify the text accordingly. However, it is the authors believe that the potential benefits of these products have been shown for the other scenarios. This is one of the major findings of this study. Since this is the first study using real observed remotely sensed soil moisture data in a large-scale catchment, we believe this is an important finding which should not be removed (Conclusions have been modified).

Additional references:

Di Baldassarre, G. and Montanari, A., 2009, Uncertainty in river discharge observations; a quantitative analysis, HESS, 13, 913-921

Wanders, N., Karssenberg, D., Bierkens, M., Parinussa, R., de Jeu, R., van Dam, J. and de Jong, S.,2012; Observation uncertainty of satellite soil moisture products determined with physically-based modeling, Remote Sensing of Environment, Volume 127, Pages 341-356, doi:10.1016/j.rse.2012.09.004.