Authors' response to Anonymous Referee #1 on "Improving operational flood ensemble prediction by the assimilation of satellite soil moisture: comparison between lumped and semi-distributed schemes" by C. Alvarez-Garreton et al.

We are truly grateful for the anonymous reviewer's comments and interesting suggestions. We addressed each comment and followed most of the reviewer's suggestions, which we think improved significantly the quality of the manuscript and research outcomes. This document provides a comprehensive explanation of the approaches and decisions adopted to answer each of the reviewer's comments.

The document is structured as follows: i) the reviewer's comments in blue font, ii) the authors' reply in black font, iii) the changes made in the revised manuscript in black italic font.

OVERVIEW

The study investigates the assimilation of satellite soil moisture data into rainfall-runoff modelling with the purpose of improving streamflow prediction. Specifically, the comparison between a lumped and a semi-distributed version of the PDM model is carried out in terms of model performance with and without the assimilation of satellite soil moisture data.

GENERAL COMMENTS

The paper investigates a very important topic related to the assimilation of satellite soil moisture data for improving flood prediction. Being highly interested to this topic, I quickly and carefully read the paper that I found well written and well structured. I fully agree with the authors that there is a strong "...need for further studies focusing on SM–DA for the purposes of improving streamflow prediction from rainfall-runoff models". Indeed, besides the satellite observation that is considered, the assimilation of soil moisture into rainfall-runoff modelling involves several critical aspects (e.g. model and observation error, rainfall-runoff model structure, data assimilation technique, ensemble generation...) that significantly affect the final result and, hence, needs to be addressed carefully.

This manuscript addresses some important new aspects related to: 1) the generation of the ensemble, 2) the characterization of the temporal variability of the observation error, 3) the evaluation of the ensemble reliability through the rank histograms, and 4) the spatial discretization of the rainfall-runoff model.

We thank the reviewer for her/his appreciation of our work and for highlighting the importance of the topic and the novelty of some key techniques introduced in our research.

We now address the reviewer's comments:

1) The most important aspect is related to the analysis of the results. Overall, the assimilation of satellite soil moisture data improves the discharge simulation with respect to the open-loop ensemble prediction, but NOT against the model run in validation, without the assimilation. For instance, for the semi-distributed scheme, the NS-value is equal to 0.77, 0.28, and -1.89 for N7, N1, and N3 catchment, respectively, in the evaluation period. The corresponding NS-values after the assimilation are 0.73, 0.18, and -2.47, always worse. The improvements highlighted in the paper are very much related to the significantly lower performance of the

open-loop ensemble prediction (NS=0.53, -0.02, and -5.36). This point needs to be addressed, especially if the methodology is to be applied from operational purpose. Actually, in our analyses we didn't find this large deterioration of the model performance when the open-loop ensemble prediction is considered. What are the reasons for that? Is it due to a bias of the open-loop ensemble prediction with respect to the model prediction (the bias is not reported in the paper)? Could it be due to the procedure adopted for producing unbiased ensemble dealing with the upper and lower soil moisture limits? I am well aware on the difficulties of obtaining a robust ensemble with the Ensemble Kalman Filter applied to rainfall runoff modelling and to real cases (not synthetic). However, this represents a very important aspect that needs to be discussed in details. At least, it should be clarified in the paper that the assimilation deteriorates the model performance with respect to the model run in validation without the assimilation.

In the discussion paper, our results revealed that the assimilation of satellite soil moisture was able to improve the open-loop ensemble predictions; however, it did not improve the unperturbed model prediction. As the reviewer points out, these results were very much related to the lower performance of the open-loop ensemble prediction, compared to the unperturbed model run. And the reviewer is correct, this was due to a bias in the open-loop ensemble prediction with respect to the model prediction.

We have noted that unbiased input forcing and states ensemble can also cause biases in ensemble discharge values due to the nonlinear model structure, which results in degradation of both openloop and updated ensemble discharge predictions.

The perturbation biases in streamflow have been corrected in the revised manuscript, by applying the bias correction scheme identical to the one used for the state perturbation bias correction. This procedure ensures that the streamflow ensemble mean maintains the performance skill of the unperturbed (calibrated) model run. More importantly, artificial skill assessed to the SM-DA coming from a poor reference open-loop is discarded. This practical tool to avoid the degradation of the unperturbed model run, also avoids an overestimation of the SM-DA efficacy and, to our knowledge, it has not been applied in SM-DA studies. We described the scheme in the revised manuscript:

Revised Section 3.3 (Error model representation):
"Although the latter resulted in unbiased state ensemble, there are still some important but subtle effects that arise from the highly non-linear nature of hydrologic models that need to be guarded against in SM-DA. Representing model errors by adding unbiased perturbation to forcing, model parameters and/or model states can lead to a biased streamflow ensemble prediction (e.g., Ryu et al., 2009; Plaza et al., 2012), compared to the unperturbed model run. This biased streamflow ensemble prediction (open-loop hereafter) is degraded compared with the streamflow predicted by the unperturbed calibrated model. As a consequence, improvement of the open-loop after SM-DA will in part be due to the correction of bias introduced during the assimilation process itself.

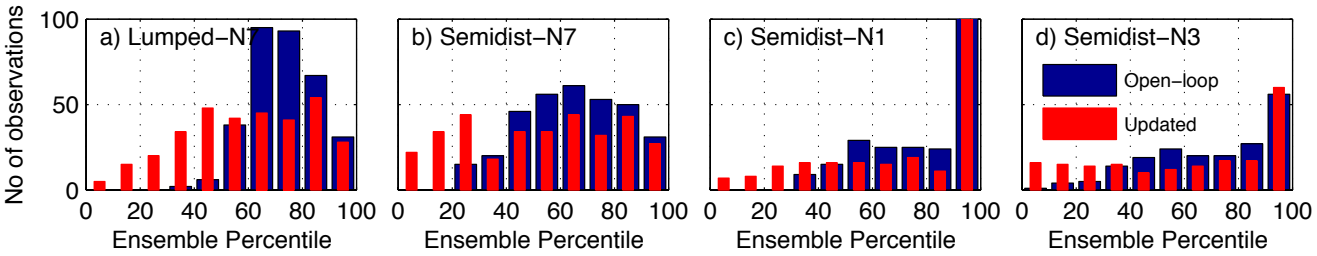
To avoid the overestimation of the SM-DA efficacy given the above, we applied the bias correction scheme proposed by Ryu et al., (2009) directly to the streamflow prediction. We used the unperturbed model run to estimate a mean bias in streamflow (following Eq. 12, but

using streamflow instead of soil moisture) and then corrected each ensemble member by subtracting this mean bias. This practical tool ensures that the streamflow ensemble mean maintains the performance skill of the unperturbed (calibrated) model run, thus avoiding the degradation of the unperturbed model run and, to our knowledge, it has not been applied in SM-DA studies."

The new results obtained after this streamflow bias correction scheme was added to the revised manuscript and the discussion was modified accordingly (see below). The new SM-DA results were also evaluated in terms of a new metric that quantifies the skill of an ensemble, the continuous rank probability score (CRPS). The CRPS description was added in the revised manuscript:

Revised Section 3.7 (Evaluation metrics):
"To evaluate the skill of the streamflow ensemble prediction before and after SM-DA, we calculated the continuous ranked probability score (CRPS; Robertson et al., 2013). CRPS is used as a measure of the ensemble errors. In the case of the model unperturbed run, CRPS reduces to the mean absolute error."

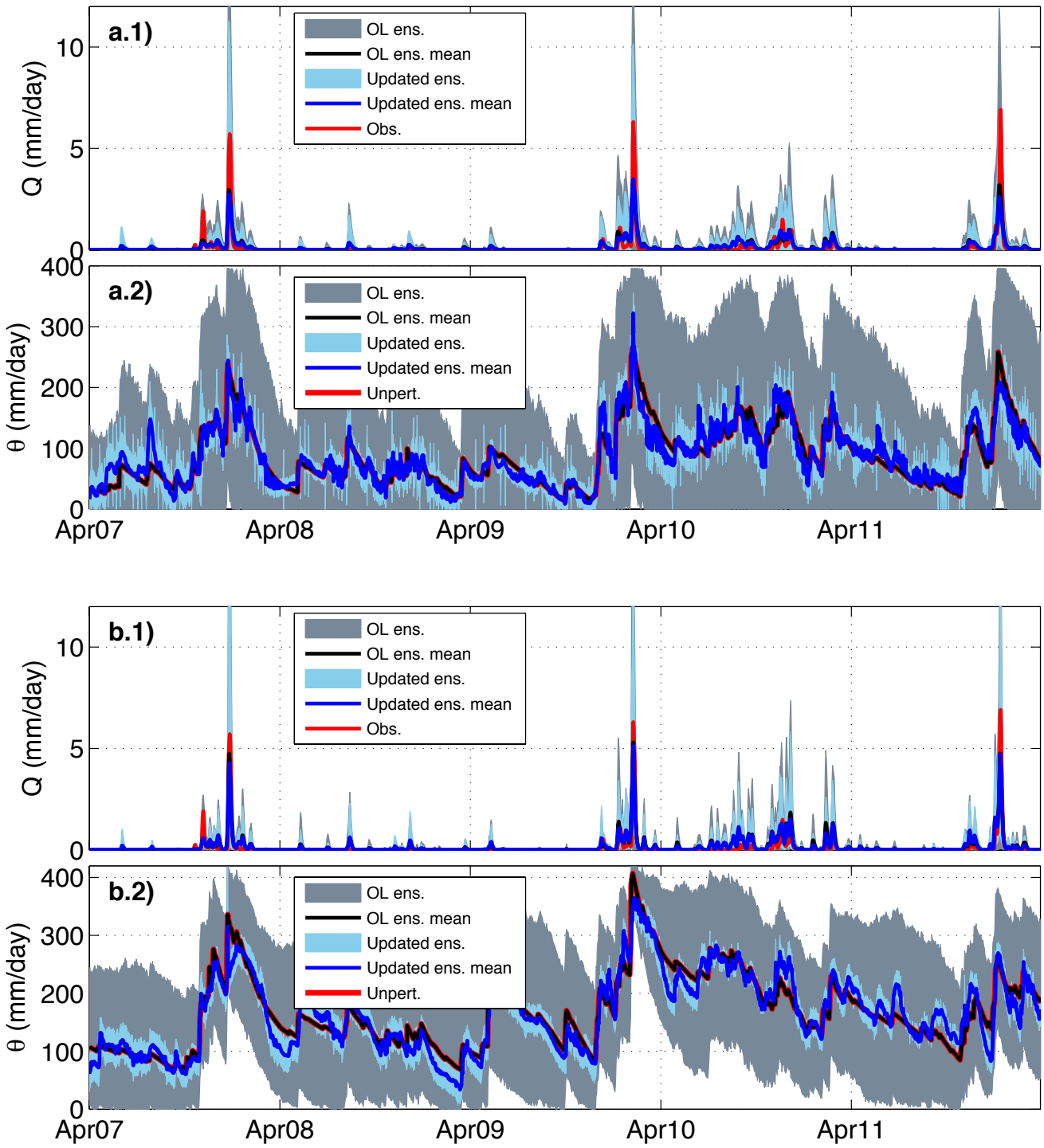
Revised Fig. 7: *Rank histograms of the open-loop and updated streamflow ensemble predictions. (a) presents the results from the lumped scheme at node N7. (b)-(d) present the results from the semi-distributed (semidist) scheme at nodes N7, N1 and N3.*



Revised Section 4.2 (Error model parameters and ensemble prediction):
"The rank histograms of the generated ensemble prediction (open-loop) are presented in Fig.7. The n-shaped and not centred histograms at the catchment outlet (N7), for both lumped and semi-distributed model schemes (Fig.7a and Fig.7b, respectively), suggest that the open-loop ensembles are slightly biased (with respect to the observed streamflow) and feature wider spread than an ideal ensemble. The width of the spread will be critical for the evaluation of SM-DA (Sect. 4.4) since any decrease of the spread would be considered as an improvement of the ensemble prediction."

"The ensemble predictions at the inner nodes N1 and N3 (Fig.7c and Fig.7d, respectively) feature high bias with respect to the observed streamflow (note that observations at N1 and N3 were not used to calibrate the error parameters). The large bias at these inner nodes owes to the large errors in the calibrated model in SC1 and SC3 (see Sect. 4.1)."

Revised Section 4.4 (Satellite soil moisture data assimilation):
Revised Fig. 10. *Streamflow (Q in mm/day) and soil moisture (θ in mm) ensemble prediction at the catchment outlet, before and after SM-DA for evaluation sub-period 2 (30 April 2007 - 02 March 2014), which had three major flooding events. (a.1) and (a.2) present the results for the lumped model. (b.1) and (b.2) present the results for the semi-distributed model.*



Revised Section 4.4 (Satellite soil moisture data assimilation):
"The rank histograms at N7, N1 and N3 are presented in Fig. 7. For all the evaluated nodes, the ensemble predictions are more reliable after SM-DA (flatter histograms compared to the open-loop). The consistent overestimation of the observed streamflow in the open-loop ensembles (diagonal histograms located towards the higher ensemble percentiles) is partially addressed by the SM-DA."

Revised Table 4: *SM-DA evaluation statistics calculated at the catchment outlet (N7) and at the inner catchments (N1 and N3).*

Statistic	Lumped scheme	Semi-distributed scheme		
	N7	N7	N1	N3
NRMSE	0.78	0.76	0.81	0.83
NS_{ol}	0.67	0.77	0.28	-1.75
NS_{up}	0.64	0.78	0.26	-1.39
POD_{ol}	0.96	0.92	0.56	0.69
POD_{up}	0.94	0.93	0.55	0.69
FAR_{ol}	0.11	0.11	0.07	0.12
FAR_{up}	0.10	0.10	0.06	0.11
PVE_{ol}	5.63	35.30	-96.87	56.42
PVE_{up}	-2.37	34.93	-109.66	40.71
$CRPS_{ol}$	0.32	0.26	0.74	0.20
$CRPS_{up}$	0.28	0.23	0.73	0.24

Revised Section 4.4 (Satellite soil moisture data assimilation):
"The performance of the ensemble mean was assessed by computing the NSol and NSup (Table 4). At the catchment outlet, the NS of the ensemble mean after SM-DA improved only for the semi-distributed scheme. At the ungauged catchments, SM-DA was effective at improving the performance of the ensemble mean only at N3, compared to the open-loop. However, the performance of the model in that catchment was still poor. This can be explained by the systematic errors of the model on those catchments before assimilation, which were not addressed by the SM-DA."

"The open-loop PVE was improved (lower PVE values) after SM-DA at N7 (for both the lumped and the semi-distributed scheme) and at N3. This was not the case however, for the inner node N1, in which the PEV was higher after SM-DA, compared to the open-loop. When compared to the unperturbed model run (Table 2), the assimilation of satellite soil moisture improved the performance of the model in terms of PVE, at all the nodes and for both the lumped and semi-distributed schemes."

"The skill of the ensembles after SM-DA (expressed by a reduction in CRPS) was improved at the catchment outlet by a 12% and 13% (for the lumped and semi-distributed scheme, respectively), and by a 17% at N1. The skill of the updated ensemble was also consistently higher than the unperturbed model run (Table 2)."

"To summarise the efficacy of the SM-DA, we take into account the characteristics of the ensemble predictions (open-loop and updated) in terms of the their mean, skill and reliability.

Based on this, we state that in overall, SM-DA was effective at improving streamflow ensemble predictions in the gauged and the ungauged catchments. By accounting for rainfall spatial distribution and routing process within the large study catchment, we improved the model performance at the outlet compared to a lumped homogeneous scheme, which in turn improved the performance of the SM-DA. The latter was achieved even though the relation between θ and the streamflow prediction was weaker in the semi-distributed scheme (Fig.6). The proposed SM-DA scheme therefore, has the merits of improving streamflow ensemble predictions by correcting the SM state of the model, even when rainfall appears to be the main driver of the runoff mechanism (see Sect. 4.1).

2) I believe that the PDM model is not the most suitable one for discharge prediction in the semiarid catchment considered in this study. Indeed, PDM was developed for simulation of discharge in humid climates, and its application in arid areas can be problematic. I was wondering if this could be the reason for the large deterioration open-loop ensemble prediction. Could the authors add some comments on this point?

It is a fact that the characteristics of the study catchment, such as its semi-arid climate and ephemeral flow regime, pose a major challenge for the simulation of its runoff mechanisms. In general, when the main drivers of runoff generation are rainfall intensity and antecedent wetness condition, a fair group of conceptual hydrologic models do a good job in representing both (including PDM). However, when rainfall intensity becomes the major factor in runoff generation, which is the case of the study catchment, these hydrologic models tend to have a sub-optimal performance. Indeed, simple event-based models could results in better streamflow prediction, provided with accurate information about losses. This is the context of this work, and is expressed in our first research question:

Discussion Paper (P10638, L27-29):
"1) While rainfall is presumably the main driver of flood generation in semi-arid catchments, can we effectively improve streamflow prediction by correcting the soil water state of the model?"

Regarding the limitations of PDM pointed out by the reviewer, we did test other more complex hydrologic models such as Sacramento Soil Moisture Accounting Model (SAC-SMA), in this study catchment plus other 7 catchments in the region, and found no improvement in the (unperturbed) streamflow prediction. Actually, PDM performed better (in terms of Nash-Sutcliffe efficiency) than SAC-SMA in most of the selected catchments.

3) The seasonal rescaling approach used in the paper allows the observations to be very close to the modelled data. It is evident looking at the correlation values reported in Table 2. Even though this is feasible, I believe that by doing this the impact of data assimilation will be very limited and I would like to know what the impact of this rescaling step is. For instance, what are the differences in the results if the rescaling is done for the whole period (instead of doing it separately for each season)?

4) Another important point is related to the observation error. This is the first paper that considers, in the context of rainfall-runoff modelling, the temporal variability of the observation error, usually assumed as constant in time. However, it should be shown the values of the

observation error used in the assimilation. How it varies in time? Could the authors show some plots of the observation error in time? What soil moisture product has the higher/lower error? Finally, what are the differences in the results if the observation error is considered as constant? I am well aware that a single paper can't analyse all the aspects of data assimilation, but some comments and suggestions should be provided (as it is done for the application of the SWI).

We answer comments 3 and 4 together since they are relevant to the same seasonal analysis in the discussion paper. The reviewer is correct in that this is the first paper considering seasonal rescaling and error estimation of the satellite soil moisture in the context of rainfall-runoff modelling. This brings many interesting questions like the ones expressed by the reviewer. However, a comprehensive evaluation of the effects of accounting for this seasonality in both, rescaling factors and error variance estimation, was not undertaken here since it fell beyond the scope of this work. Such evaluation would involve more comprehensive analyses of errors that are currently in progress, as a separate work. We added specific comments about these open questions in the conclusions of the revised manuscript:

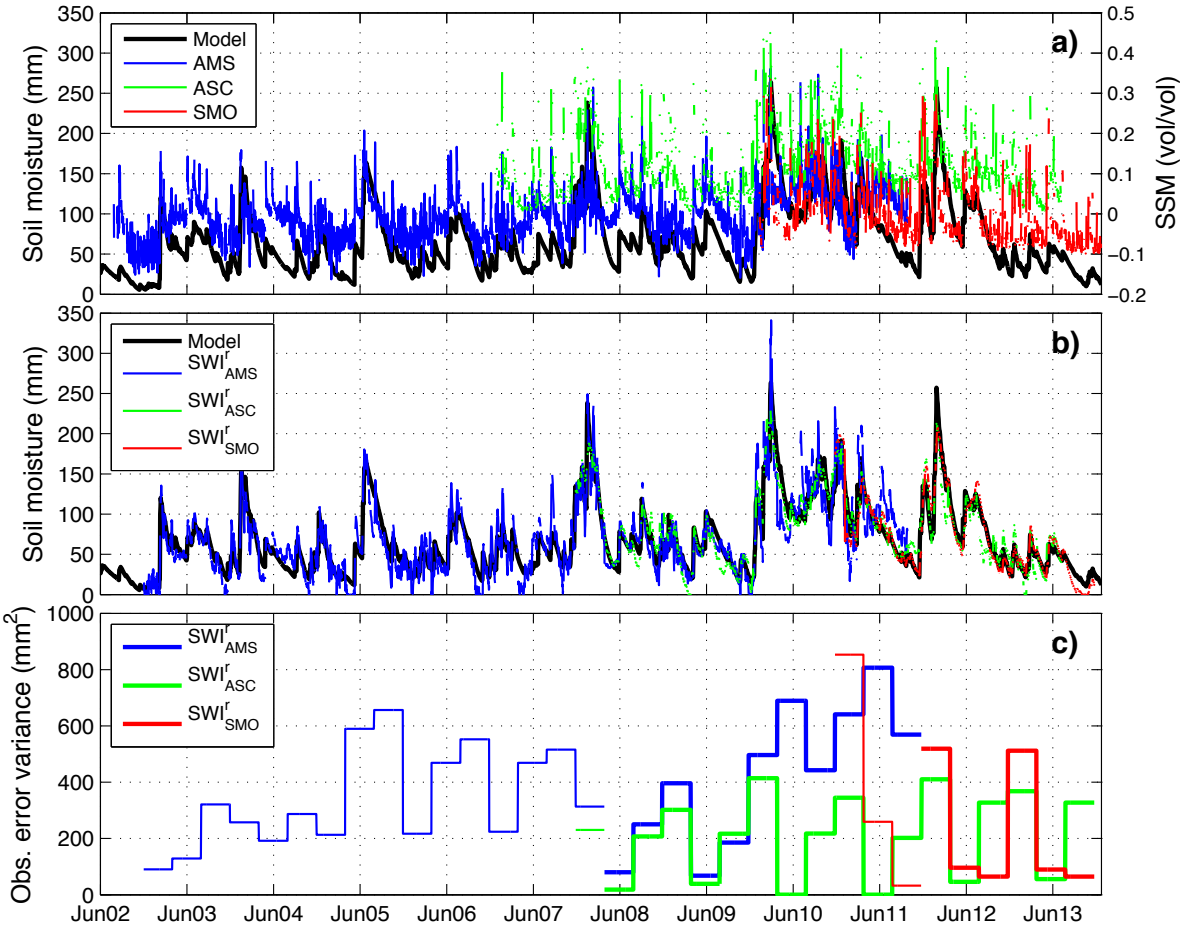
Revised Conclusions:
"In the rescaling and error estimation procedure, we applied seasonal TC and LV to avoid error-in-variable biases. Applying these to correct biases in the SWI, showed improved agreement between observed and modelled SM. This seasonal approach is novel in the context of SM-DA and tends to lead to closer agreement between model and observations. Further investigation is required to assess the impacts and importance of accounting for seasonality in rescaling and error estimation."

We followed the reviewer's suggestion of adding a plot of the observations error variance time series in the revised manuscript (Fig. 8), with a description of the results and a comparison with the typical constant value used in previous studies (standard deviation of 3% vol/vol):

Revised Section 4.3 (SWI, rescaling and error estimation):
"Figure 8c shows the seasonal observation error variance, and reveals a clear variation in the error with time. The variation of the seasonal error values is due to the alternative use of TC or LV, and to the increasing sample size of each seasonal pool (see Section 3.6), which should reduce the uncertainties coming from finite sample size. One limitation of this procedure is its assumption that the errors vary seasonally without inter-annual variability. Since there are inter-annual cycles (wet and dry years), one may also expect the errors to vary with year. Ideally, moving-window estimation with windows smaller than 3 months should be considered, but that would cause greater sampling uncertainties of TC or LV estimates. The inverse relationships between AMS and ASC errors at some times could be due, among other factors, to the passive retrieval by AMS compared with the active ASC.

A common error standard deviation value used in previous SM-DA studies is 3% vol/vol (e.g., Chen et al., 2011). This constant error, when transformed accordingly to the soil moisture storage capacity of the model and the soil porosity (see Section 3.5), gives an error variance of 667 (750) mm^2 for the lumped (semi-distributed) scheme. As a simple comparison, these values are within the range of the error variance estimated through seasonal LV/TC, however

a comprehensive analysis of the impacts of accounting for seasonality in SM-DA is not performed here since it falls beyond the scope of this work."



Revised Fig. 8. (a) shows the model soil water content on the left axis and the satellite soil moisture (SSM) observations on the right axis. (b) shows the soil moisture in the model space, after the three SSM datasets were transformed into a soil wetness index (SWI) and then rescaled by using TC or LV (SWI^r_{AMS}, SWI^r_{ASC} and SWI^r_{SMO}). (c) shows the rescaled SSM observations error variance using TC (thick line) and LV (thin line).

5) The description of the approach employed for the data assimilation in the semi-distributed scheme is not well described. Is the assimilation carried out separately for each sub-catchment? Is it considered the spatial cross-correlation of measurements? Could the authors specify better these aspects?

These aspects were explained more clearly in the revised manuscript:

Revised Section 3.2 (EnKF formulation):
"In the case of the semi-distributed scheme, during the updating steps described above, each sub-catchment was treated independently and no spatial cross-correlation in the satellite measurements was considered."

6) The results in the calibration period are not reported. What is the model performance at the outlet, and for the inner catchments? Is the model able to capture flood peaks satisfactorily

(from Figure 4a it seems that PDM always underestimate the highest peaks)? What are the differences in the performance between the calibration and the validation period? I also suggest including all the performance scores (e.g. POD and FAR) used for the assessment of the results before and after the assimilation as reported in Table 3. The addition of the errors on peak discharge and volume would be useful for the evaluation of the results in a context of flood prediction.

Following the reviewer's suggestion, in the revised manuscript we have added a new table (Table 2 in the revised manuscript) with a summary of the model evaluation statistics for the calibration and evaluation periods. We also added a measure of the peak volume error, calculated as the aggregated difference between simulated and observed streamflow (expressed as mm), for all days where the daily observed streamflow was above a moderate flood.

We added brief comments on the Table in the revised manuscript and we incorporated the description and equation to calculate PEV in Section 3.7 (Evaluation metrics) of the revised manuscript:

Revised Section 3.7 (Evaluation metrics):
"Finally, we calculated the aggregated peak volume error (PVE, in mm) of the ensemble mean, for days when the observed streamflow was above a minor flood classification (t days in Eq. 25. PVE, as an example for the open-loop, was calculated as"*

$$PVE_{ol} = \sum_{t*} \left(\overline{Q_{sim}^{ol}(t^*)} - Q_{obs}(t^*) \right)$$

Revised Table 2: "*Model evaluation at the catchment outlet (N7) and at the inner catchments (N1 and N3), for calibration and evaluation periods. RMSE and PVE statistics are in units of mm."*

Statistic	Lumped scheme	Semi-distributed scheme		
	N7	N7	N1	N3
$RMSE_{calib}$	0.19	0.18	-	0.30
$RMSE_{eval}$	0.21	0.18	0.53	0.46
NS_{calib}	0.52	0.59	-	0.39
NS_{eval}	0.67	0.77	0.28	-1.89
POD_{calib}	0.79	0.76	-	0.76
POD_{eval}	0.93	0.91	0.54	0.73
FAR_{calib}	0.09	0.10	-	0.15
FAR_{eval}	0.11	0.11	0.07	0.14
PVE_{calib}	-70.86	-39.99	-28.97	168.23
PVE_{eval}	1.30	34.75	-100.53	115.52
$CRPS_{calib}$	0.29	0.28	1.45	0.58
$CRPS_{eval}$	0.56	0.33	0.92	0.49

"Table 2 presents the evaluation statistics of the streamflow prediction in the calibration and evaluation periods, for the catchment outlet and the inner catchments (notice that N1 does not have data in the calibration period). The different statistics in this table consistently show that, at the catchment outlet, the semi-distributed has a consistently better performance than the lumped scheme in terms of RMSE, NS, PEV and CRPS. Both schemes improve their statistics in the evaluation period due to the higher flows."

A brief discussion regarding the results for the inner nodes N1 and N3 was also added in this Section (please see reply to Specific comment #4 below).

SPECIFIC COMMENTS

1) P10636, L5: "... we assimilate active and passive satellite soil moisture...". The name of the products should be given in the abstract.

Following the reviewer's suggestion, we added the name of the products in the revised abstract.

Revised Abstract:

"Within this context, we assimilate satellite soil moisture (SSM) retrievals from the Advanced Microwave Scanning Radiometer (AMSR-E), the Advanced Scatterometer (ASCAT) and the Soil Moisture and Ocean Salinity (SMOS) instrument, using an Ensemble Kalman filter to improve operational flood prediction within a large semi-arid catchment in Australia (>40,000km^2)."

2) P10644, L6: The satellite soil moisture observations are assimilated sequentially in this study. Is there an impact in the order of the products that are assimilated? I.e., first AMSR-E, then ASCAT and finally SMOS. If you change the order, do the results remain the same?

In order to answer the reviewer's question, we repeated the SM-DA experiments with different order of the products and found that results did not vary significantly. We added a comment about this in the revised Section 3.2 (EnKF formulation):

"The selection of the order of the products assimilated in steps 1 to 3 was arbitrary; however, we checked that different orders did not significantly affect the SM-DA results."

Below are the evaluation statistics for each case:

Order 1: AMSR-E/ASCAT/SMOS (as presented in the manuscript).

Statistic	Lumped scheme	Semi-distributed scheme		
	N7	N7	N1	N3
NRMSE	0.78	0.76	0.81	0.83
NS_{ol}	0.67	0.77	0.28	-1.75
NS_{up}	0.64	0.78	0.26	-1.39
POD_{ol}	0.96	0.92	0.56	0.69
POD_{up}	0.94	0.93	0.55	0.69
FAR_{ol}	0.11	0.11	0.07	0.12
FAR_{up}	0.10	0.10	0.06	0.11
PVE_{ol}	5.63	35.30	-96.87	56.42
PVE_{up}	-2.37	34.93	-109.66	40.71
$CRPS_{ol}$	0.32	0.26	0.74	0.20
$CRPS_{up}$	0.28	0.23	0.73	0.24

Order 2: AMSR-E/ASCAT/SMOS

Statistic	Lumped scheme	Semi-distributed scheme		
	N7	N7	N1	N3
NRMSE	0.77	0.76	0.81	0.84
NS_{ol}	0.67	0.77	0.28	-1.75
NS_{up}	0.64	0.78	0.26	-1.38
POD_{ol}	0.96	0.92	0.56	0.69
POD_{up}	0.94	0.93	0.55	0.69
FAR_{ol}	0.11	0.11	0.07	0.12
FAR_{up}	0.10	0.10	0.06	0.11
PVE_{ol}	5.63	35.30	-96.87	56.42
PVE_{up}	-2.21	34.71	-109.88	40.56
$CRPS_{ol}$	0.32	0.26	0.74	0.20
$CRPS_{up}$	0.28	0.23	0.73	0.24

Order 3: AMSR-E/ASCAT/SMOS

Statistic	Lumped scheme	Semi-distributed scheme		
	N7	N7	N1	N3
NRMSE	0.77	0.76	0.81	0.84
NS_{ol}	0.67	0.77	0.28	-1.75
NS_{up}	0.64	0.78	0.26	-1.39
POD_{ol}	0.96	0.92	0.56	0.69
POD_{up}	0.94	0.93	0.55	0.68
FAR_{ol}	0.11	0.11	0.07	0.12
FAR_{up}	0.10	0.10	0.06	0.11
PVE_{ol}	5.63	35.30	-96.87	56.42
PVE_{up}	-2.29	34.80	-109.73	40.64
$CRPS_{ol}$	0.32	0.26	0.74	0.20
$CRPS_{up}$	0.28	0.23	0.73	0.24

3) P10650, L9: The use of only one year for the calibration of T parameter is likely not sufficient.

We agree with the reviewer, and added a comment highlighting this issue in the revised manuscript:

Revised Section 3.5 (Profile soil moisture estimation):

"*This calibration period was selected to maximise the independent evaluation period (see Section 3.7), however more representative values are likely to be obtained if a longer period was used for calibration."*

4) P10655, L11: It is very good to show that negative NS-values are obtained, usually this is not done. However, I believe that some investigations on the reasons for these bad performance should be given. Is it due to the PDM model (see General Comments), or to the input data, or to the model parameterization?

Our decision to evaluate the SM-DA in "ungauged catchments" (i.e., N1 and N3 observations were not used for calibration) was based on the scope of our paper, defined by our research questions 2 and 3:

"2) What is the impact of accounting for channel routing and the spatial distribution of forcing data on SM-DA performance? 3) What are the prospects for improving streamflow within ungauged inner catchments using SSM?"

This ungauged scenario had many implications, like those poor NS values. We do agree with the reviewer that such poor NS values are worthy of further investigation. Accordingly, we did set up a calibration scheme in which the observations of N1 and N3 were used to get specific set of optimal parameters for those two sub-catchments. Our results revealed that the model was able to adequately simulate streamflow in those sub-catchments (NS above 0.69). Further details of this scheme setup are provided in our reply to the second reviewer Dr. Uwe Ehret. Based on these results, we argue that the problem for such bad performance is mainly attributed to errors in model parameters and less likely to errors in the input data and model structure. These parameter errors are mainly due to the issue that the integrated catchment streamflow response is poor at informing about catchment heterogeneity. We did not show the details of these results in the discussion paper to maintain the manuscript concise and to be consistent with the "ungauged" nature of the inner catchments. However, we added a comment regarding this:

Revised Section 4.1 (Model calibration):
"To explore the reasons of such bad performance, we separately calibrated the model parameters in those sub-catchments by using all the available N7, N1 and N3 observations. The results (not shown here) revealed that in this case, the model was able to adequately simulate streamflow in those sub-catchments (NS in evaluation period of 0.78, 0.69 and 0.84 at N1, N3 and N7 nodes, respectively). Based on this, we argue that the problem for the poor model performance in the ``ungauged'' inner catchments is mainly attributed to sub-optimal parameter estimation (due to the limited information about catchment heterogeneity provided by the integrated catchment streamflow response) and less likely to errors in the input data and model structure."

5) P10656, L20: The rank histogram of the soil moisture ensemble might be also analysed here.
As the reviewer suggests, we also considered adding analyses of rank histograms for soil moisture on the manuscript; however, based on the considerations listed below, we decided not to present them in the revised manuscript:

- We processed the satellite data to make it comparable to the model soil moisture (SM), however there is still significant higher noise in the rescaled SWI compared to the model soil moisture (Fig. 8).
- In our opinion, the higher noise in the observed SM time series, in our opinion, makes them unsuitable for checking the reliability of the SM ensemble predictions. It is unclear for us that a reliable SM ensemble should contain SM observations with such a degree of noise. In contrast, for the streamflow ensemble prediction, we have an observed time series with similar degree of "noise", which we do aim to envelop with a reliable ensemble.
- In the case of the SM open-loop, we are summarising all the sources of error by perturbing only three components of the model, therefore we expect that the ensemble

spread of the SM is indicative errors in SM prediction, as well as other errors. This means that the ensemble SM is not directly comparable to the observed SM.

- Moreover, there are 3 different observed SM datasets and therefore three different cases to present and analyse. We believe this would enlarge the manuscript with aspects that fall beyond of our main scope.

6) P10657, L21-26: A T-value equal to 40 days is obtained for SMOS. This is not expected, as the SMOS soil moisture product should be the one with the higher penetration depth and, hence, the lower T value. It is the opposite. Could the authors add some explanations for that? Could this value be attributed to the noise of satellite data (e.g. due to RFI)?

To answer this question, we focus on the differences between SMOS and AMSR-E, since they are both passive retrievals. Indeed, it is true that the penetration depth of the C-band of AMSR-E is shallower than that of the L-band of SMOS. However, the underlying assumptions of their retrievals pertaining to spatial heterogeneity are quite different. For instance, the LPRM algorithm that produced the AMSR-E data set assumes a homogeneous surface and globally constant roughness and vegetation scattering albedo, whereas the parameterization of SMOS L-MEB is based on the dominant land cover and its surrounds.

While RFI in C-, X-, and L-bands are small over Australia, the presence of noise can influence the operation of the SWI filter. In our synthetic analysis (unpublished), increased noise increases the optimal T and also its uncertainty. To our best knowledge, the existing studies examining the dependence of T on soil depth are usually based on a single satellite product against in situ measurements at variable depths. Hence it is difficult for these studies to appreciate the complexity increased by noise and different sensing and retrieval methods.

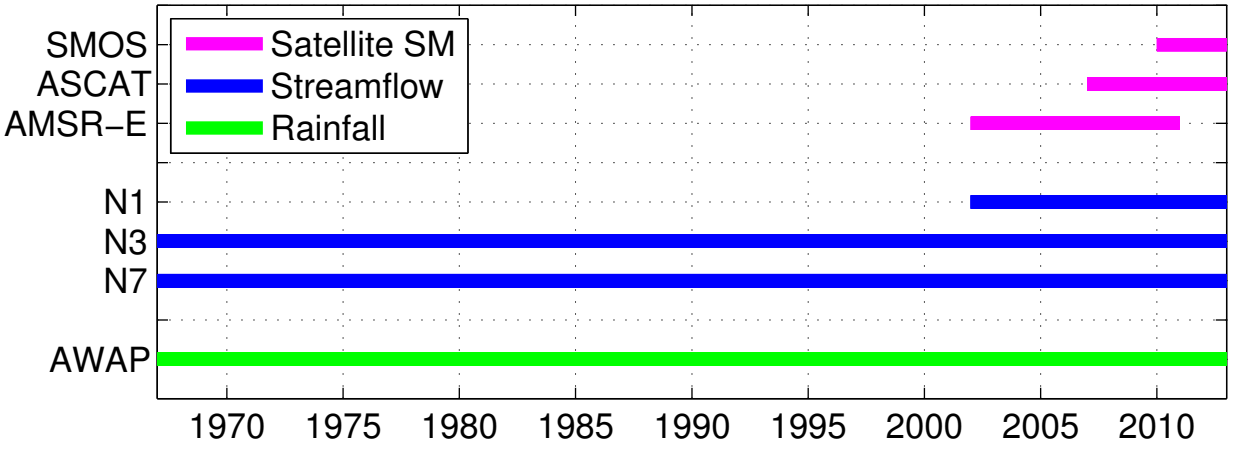
We added a brief comment about this in the revised Section 4.3 (SWI, rescaling and error estimation):
"Previous studies have shown that the optimal T value increases with layer depth (e.g., Brocca et al., 2010). Results presented here show an increased T value for SMO, which would be inconsistent with L-band having a deeper penetration than AMS C-band (to limit the comparison within passive retrievals). We speculate that these differences might be due various factors, including the different retrieval methods (which have quite different assumptions pertaining to spatial heterogeneity) and the influence that radio-frequency interference (RFI) noise. Moreover, to the best of our knowledge, the existing studies examining the dependence of T on the soil depth are usually based on a single satellite product against in situ measurements at variable depths. Hence it is difficult for these studies to elucidate the complexity increased by noise and different sensing and retrieval methods."

7) P106568, L27: "...seems to be less sensitive to these violated assumptions...". What is the proof of this sentence? The reference to another study performing data assimilation is, at least for me, not enough.
We agree with the reviewer in that two studies are not enough to support our statement. What we aimed to express in this paragraph is the urgent need to advance the study of SM-DA in rainfall-runoff models. We think that what has been found in previous studies using SM-DA in land-surface models and analysing soil moisture prediction may not be directly applicable in

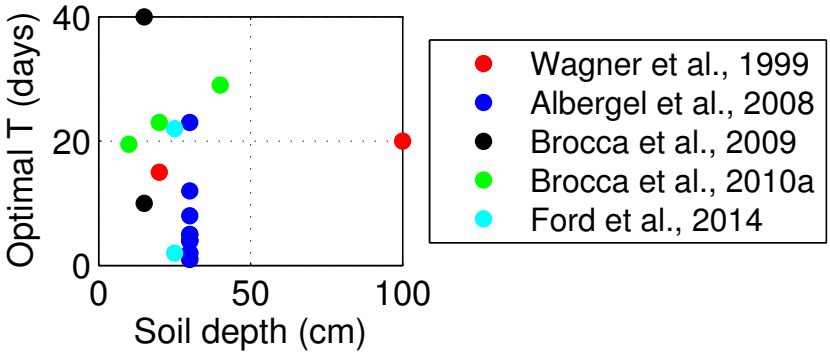
the case of rainfall-runoff models. We modified the paragraph in the revised manuscript to better explain this:

Revised Section 4.3 (SWI, rescaling and error estimation):
In this context, the performance of the SM-DA with respect to the improvement in streamflow has been under-investigated. Alvarez-Garreton et al. (2013, 2014) show that in terms of streamflow prediction, SM-DA seems to be less sensitive to these violated assumptions. This lower sensitivity and apparent contradiction with previous studies analysing soil moisture prediction performance, highlights the need for further studies focusing on SM-DA for the purposes of improving streamflow prediction from rainfall-runoff models.

8) Figure 2: SMOS should start in 2010. From the figure it seems starting in 2009. Please check.
The Figure was corrected in the revised manuscript.



9) Figure 9: The reference Brocca et al. (2010) should be the paper on Remote Sensing of Environment (Brocca, L., Melone, F., Moramarco, T., Wagner, W., Hasenauer, S. (2010). ASCAT Soil Wetness Index validation through in-situ and modeled soil moisture data in central Italy. Remote Sensing of Environment, 114 (11), 2745-2755.), not the one on HESS.
The Figure was corrected in the revised manuscript.



Authors' response to Uwe Ehret (Referee #2) on "Improving operational flood ensemble prediction by the assimilation of satellite soil moisture: comparison between lumped and semi-distributed schemes" by C. Alvarez-Garreton et al.

We are truly grateful for the constructive and interesting comments and suggestions provided by Dr Uwe Ehret. We strived to address each comment and to provide a comprehensive explanation of the approaches and decisions adopted to answer them.

The document is structured as follows: i) the reviewer's comments in blue font, ii) the authors' reply in black font, iii) the changes made in the revised manuscript in black italic font.

SCOPE

The article is within the scope of HESS

SUMMARY

This study investigates the benefits of assimilating satellite-derived soil moisture data into ensemble streamflow predictions in sparsely monitored catchments at the example of the semi-arid 40.000 km² Warrego catchment in Australia. To this end, a conceptual hydrological model (PDM) is calibrated by streamflow observations at a single outlet gauge in a lumped and semi-distributed (7 sub-catchments) configuration. For 3 components of the model (rainfall input for forcing, a conceptual store retention constant for parameters and the soil water storage for states), error models were formulated and their parameters found by calibration. The satellite observations were transformed to estimates of profile soil moisture, bias-corrected and then assimilated into the model predictions with an Ensemble Kalman filter approach. Model performance was then evaluated a) lumped vs. semi-distributed and b) unperturbed model vs. open loop ensemble vs. updated ensemble predictions (updated by data assimilation of satellite-derived soil-moisture estimates) by normalized RMSE (NRMSE), Nash-Sutcliffe efficiencies (NS), probability of detection (POD) and false alarm ratio (FAR). The authors' main conclusions are:

- The main limitation of the proposed ensemble model scheme was the too large ensemble spread of the open-loop model predictions
- Estimation of profile-average (here ~ 1 m depth) soil moisture from satellite data is difficult and involves large uncertainties
- Nevertheless, the data assimilation predictions outperformed the open-loop ensemble predictions, the data assimilation predictions with the semi-distributed model outperformed those based on the lumped model, and data assimilation predictions at uncalibrated gauges within the catchment outperformed the open-loop predictions.

The authors conclude with the recommendation to focus efforts on ensuring adequate hydrological models given the available data.

OVERALL RANKING

The work is ranked 'major revision'.

GENERAL EVALUATION

This is a thoroughly conducted study, where all of the (many) crucial assumptions that needed to be made on the way (e.g. the choice of error models or the profile soil moisture estimation) are discussed.

We thank the reviewer for its appreciation of our work and for highlighting the special attention we gave to comprehensibly discussing the assumptions needed in a real satellite soil moisture data assimilation experiment.

1) As the goal is to evaluate the additional value of assimilating remotely-sensed soil moisture data, it should be done with an 'as-good-as-possible' hydrological model, i.e. a model that has been set up making as good as possible use of the standard available data. I assume the author's last statement in the conclusions points exactly in this direction. So instead of evaluating the benefit of satellite-derived soil-moisture data assimilation (SM-DA) against a lumped model and a semidistributed model that was calibrated with the outlet gauge only, it should be evaluated against a model calibrated on all 3 gauges.

The reviewer is right about the main goal of this paper. Within this main goal however, the scope of this work was defined by 3 research questions: *1) While rainfall is presumably the main driver of flood generation in semi-arid catchments, can we effectively improve streamflow prediction by correcting the soil water state of the model? 2) What is the impact of accounting for channel routing and the spatial distribution of forcing data on SM-DA performance? 3) What are the prospects for improving streamflow within ungauged inner catchments using SSM?*

The evaluation of an “as-good-as-possible” model (benchmark model hereafter), which uses all available streamflow information to calibrate the model parameters, does not consistently fit within questions 2 and 3 (Q2 and Q3, respectively). Regarding Q2 cited above, the benchmark model accounts for the two aspects the question refers to (spatial distribution of forcing and channel routing), but more importantly, it adds additional observations to estimate model parameters, which means that is not directly comparable to the lumped case. We therefore consider that the benchmark model cannot be used to consistently address this question. Regarding Q3, the benchmark model does not allow us to test ungauged scenario since it uses the inner gauges.

The use of a benchmark model with an “as-good-as-possible” set of model parameters could provide important information about, for example, the effects that the parameter quality has in SM-DA efficacy (by comparing the benchmark model with the semi-distributed “ungauged” model). Although these results would be very interesting to explore, and they could potentially support our last statement in the conclusions regarding focusing effort on ensuring adequate models (as the reviewer correctly mentions), they address a different research question than the ones we defined. Therefore, we consider that adding a third case (first case is the lumped model, second case is the “ungauged” semi-distributed model) for evaluation of SM-DA falls beyond the scope of this work.

Notwithstanding the above, and to attend the reviewer’s (as well as our) interest in evaluating such a model, mostly with the purpose of investigating the poor performance of the semi-distributed scheme in the “ungauged” the inner catchments (in the Discussion paper), we calibrated an “as-good-as-possible” model. Results are presented in this response document (see Table 1 below), but they are not included in the revised manuscript.

The “as-good-as-possible” model was calibrated by using all the available streamflow information of N1, N3 and N7. Note that this in practice is not consistent with the calibration period defined in the discussion paper since the gauge N1 only has data in the evaluation period (see Fig. 2 of the discussion paper). Results are summarised below:

Table 1: Benchmark model results in evaluation period

	Unperturbed model		
	N7	N1	N3
RMSE (mm)	0.15	0.30	0.15
NS	0.84	0.78	0.69
POD	0.89	0.79	0.83
FAR	0.08	0.02	0.05
PVE (mm)	-22.90	-21.06	-15.69

*In the case of the open-loop and updated ensemble predictions, NS, POD and FAR statistics are calculated using the ensemble mean. RMSE statistic is calculated as the mean of each ensemble member’s RMSE (see Sect. 3.7 of the discussion paper).

A brief mention of these results is provided in the revised manuscript.

Revised Section 4.1 (Model calibration):
“To explore the reasons of such bad performance, we separately calibrated the model parameters in those sub-catchments by using all the available N7, N1 and N3 observations. The results (not shown here) revealed that in this case, the model was able to adequately simulate streamflow in those sub-catchments (NS in evaluation period of 0.78, 0.69 and 0.84 at N1, N3 and N7 nodes, respectively). Based on this, we argue that the problem for the poor model performance in the ``ungauged" inner catchments is mainly attributed to sub-optimal parameter estimation (due to the limited information about catchment heterogeneity provided by the integrated catchment streamflow response) and less likely to errors in the input data and model structure.”

Also, the performance of the hydrological model could potentially be considerably improved by either improvement or calibration of the evapotranspiration (ET) module, especially so as this study focuses on soil moisture states, which are strongly influenced by the ET scheme. So please include a description of the ET module, explain whether it has been used in calibration or not (and if not consider using it for calibration), explain its influence on the quality of the model predictions, and also consider including it in your model error model. Then conduct the evaluation of SM-DA again with the 'as-good-as-possible' hydrological model.

In the PDM formulation, the actual evapotranspiration (ET) is calculated for each time step based on potential evapotranspiration (PET, which is an input data), the soil moisture content (S), and a parameter b_e as follows:

$$\frac{ET}{PET} = 1 - \left(\frac{S_{max} - S(t)}{S_{max}}\right)^{b_e} \quad \text{Eq.1}$$

Where S_{max} is the total available soil moisture storage, calculated as:

$$S_{max} = \frac{b\ c_{min} + c_{max}}{b + 1} \qquad \text{Eq.2}$$

Where C_{min} and C_{max} are the minimum and maximum soil moisture store capacities, respectively. Parameter b is the exponent of the Pareto distribution controlling the spatial variability of store capacity.

In our experiments, parameters b_e, C_{min}, C_{max} and b are calibrated by maximising the Nash-Sutcliffe Efficiency of the streamflow prediction at the catchment outlet. Therefore, and answering the reviewer's comment, the ET module has been calibrated.

As the reviewer indicates, this ET scheme has a great influence in the quality of model predictions by accounting for the only loss term in the water balance. This is critical in our study catchment, where the runoff coefficient is very low.

Regarding the inclusion of the ET scheme in the model error representation suggested by the reviewer, by perturbing the model soil moisture (S in Eq. 1), we are implicitly accounting for errors in ET. We could potentially include error in parameters b_e, C_{min}, C_{max} and/or b, which would also account for errors in the ET estimation, however, in the current state of this work, we decided to represent model error parameters by perturbing parameter k_1 (time constant of surface storages S_{21} and S_{22} of the model, see Figure 1 of this document). This decision was made based on the direct effect that surface runoff has in the streamflow prediction. Adding further parameters into the error model representation would aggravate the highly underdetermined condition of this problem.

We do agree however, that different error model configurations should be explored and we have highlighted it as part of our limitations (and recommendations) in different sections of the original and revised manuscript:

Revised Conclusions:
"The open-loop ensembles at the catchment outlet provide key information about prediction uncertainty, which is required for assessing risks associated with water management decisions (Robertson et al., 2013). These ensembles showed a slight bias with respect to the observed streamflow and featured a wide spread. Further exploration of model error representation (sources of error and the structure of those errors) and error parameter estimation is required to improve the characteristics of the open-loop ensemble prediction."

2) So far, the semi-distributed, non-ensemble model outperformed all others (NS = 0.77). So as a flood forecaster, given the choice of all presented model configurations (lumped/semidistributed, open-loop/assimilation), I would choose the former (even though I recognize the additional benefit of a probabilistic prediction). It may well be that with the 'as-good-as-possible' hydrological model mentioned above, this may be even more the case. So for me the main message of the study is to focus on the set-up of the hydrological model and the associated error models rather than SM-DA, if better (and probabilistic) stream flow predictions are the goal. The authors have mentioned this in their study, but it should be

stated more clearly. Also, for the study this means that before applying SM-DA to the model, the focus should be on improving the error models of the hydrological model (which, as the authors correctly state, is a highly underdetermined problem), until the performance of the open-loop model ensemble mean is comparable to that of the non-ensemble model. So far, it is considerably worse (NS 0.61 and 0.53 for lumped and semi-distributed model).

The worse performance of the open-loop with respect to the unperturbed model was due to a bias in the open-loop streamflow introduced in the ensemble generation process. We corrected this bias in the revised manuscript by applying a perturbation bias correction identical to the one used for soil moisture ensemble. Please see please the details of this procedure and the revised results in our reply to reviewer 1 (general comment 1).

We agree with the reviewer in that focusing on a robust ensemble generation is a key message this work delivers. This involves further exploration of error model representation schemes and error model parameter calibration techniques. We have added clearer states regarding this in the revised manuscript (please see the Revised Conclusions paragraph cited in our previous response).

While acknowledging the limitations of our open-loop ensemble predictions, we recognise the added value and the need of having probabilistic predictions (as the reviewer also states). This probabilistic scenario is where we focus the key messages of this study:

Conclusions – Discussion paper:
"The evaluation of the SM-DA results led to several insights. 1) The SM-DA was successful at improving the open-loop ensemble prediction at the catchment outlet, for both the lumped and the semi-distributed case. 2) Accounting for spatial distribution in the model forcing data and for the routing processes within the large study catchment improved the skill of the SM-DA at the catchment outlet. 3) The SM-DA was effective at improving streamflow prediction at the ungauged locations, compared to the open-loop. However, the updated prediction in those catchments was still poor, because the systematic errors before assimilation are not addressed by a SM-DA scheme."

"This work provides new evidence of the efficacy of SM-DA to improve streamflow ensemble prediction in sparsely instrumented catchments. We demonstrate that SM-DA skill can be enhanced if the spatial distribution of forcing data and routing processes within the catchment are accounted for in large catchments. We show that SM-DA performance is directly related to the model quality before assimilation, therefore we recommend that efforts should be focused on ensuring adequate models, while evaluating the trade-offs between more complex models and data availability."

SPECIFIC COMMENTS
• 10641/1-5: what is the spatial resolution of the satellite data? Also, it is not clear yet what the satellites actually observe (penetration depth etc.). Please include a reference to section 3.5
Specifications about the satellite products were added to the revised Section 2 (Study area and data) and a reference to this information was added in the revised Section 3.5 (Profile soil moisture estimation).

Revised Section 2 (Study area and data):
"Three SSM products are used here. The first is the Advanced Microwave Scanning Radiometer - Earth Observing System (AMSR-E, AMS hereafter) version 5 VUA-NASA LPRM (Land Parameter Retrieval Model) Level 3 gridded product (Owe et al., 2008). AMS uses C- (6.9 GHz) and X-band (10.65 and 18.7 GHz) radiance observations to derive near-surface soil moisture (2 to 3 cm depth) using a land-surface radiative transfer model. The product used is in units of volumetric water content (m^3 m^{-3}) and has a regular grid of 0.25^o

The second product is the TU-WIEN (Vienna University of Technology) ASCAT (ASC hereafter) data produced using the change-detection algorithm (Water Retrieval Package, version 5.4) (Naeimi et al., 2009). ASC transmits and measures electromagnetic waves in C-band (5.3Gz) and has a nominal spatial resolution varying from 25 to 50 km. The change-detection algorithm assumes that land surface characteristics are relatively static over long time periods under a given incident angle. Based on this, the differences between instantaneous backscatter coefficients and the historical highest and lowest values, are related to changes in soil moisture (Wagner et al., 1999). The product is provided in relative terms as the degree of saturation.

The third SSM product is the Soil Moisture and Ocean Salinity (SMOS) satellite (SMO hereafter), version RE01 (Re-processed 1-day global soil moisture product) provided by Centre Aval de Traitement des Donnees. SMO uses L-band (1.4 GHz) detectors, which have a penetration depth of approximately 5 cm and a spatial resolution of approximately 43 km. Near-surface soil moisture is obtained in units of volumetric water content (m^3 m^{-3}), by using a forward physical model inversion, described by Kerr et al. (2012)."

• 10642/14: Are the k-parameters really time-dependent, i.e. time-variable?
Parameters k_1 and k_2 are the time constant parameters of the two surface storages S_{21} and S_{22} (see Figure 1). These parameters are fixed in time and estimated through calibration. We corrected the sentence the reviewer mentions, which led to this misunderstanding:

Revised Section 3.1 (Lumped and semi-distributed model schemes):
"The time constant parameters of the storages S_{21}, S_{22} and S_3 (k_1, k_2 and k_b, respectively) were scaled by the area of each sub-catchment."

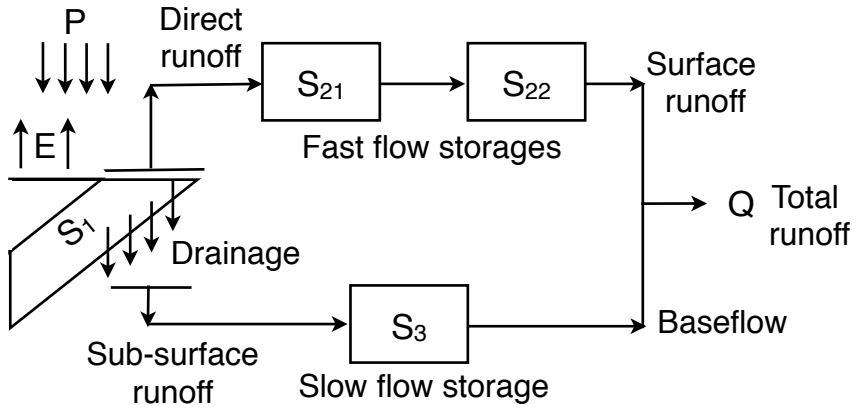


Figure1: PDM scheme

• 10645 pp.: Please justify in more detail the choice of your error functions, especially for k

As detailed in Section 3.4 of the Discussion paper, the three main sources of error in model predictions (forcing error, parameter error and structural error) were represented by perturbing rainfall data, parameter k1 and soil moisture state (water content in S1).

We consider that the selection of the error structures of rainfall and structural error was adequately justified in the discussion paper. We do agree with the reviewer however, that further justification can be provided regarding the parameter error model. Here we summarise the justification provided for the choice of forcing and structural error schemes. We also added further justification of parameter error choice in the revised manuscript:

Section 3.3 (Error model representation) – Discussion paper:

- Error in rainfall data: we based our choice on the findings of previous studies (e.g., McMillan et al., 2011; Tian et al., 2013) that show that multiplicative error model is suitable for rainfall observations. This error structure (log-normally distributed multiplicative error with mean one) indicates that higher rainfall will have higher observation error and it has been widely used in several SM-DA studies (e.g., Chen et al., 2011; Brocca et al., 2012; Alvarez-Garreton et al., 2014).

- Error in model structure: we represented the structural error by perturbing the soil moisture prediction with an additive random error (Gaussian distribution with mean zero). We followed the scheme used in most SM-DA experiments (e.g., Reichle et al., 2008; Crow and Van den Berg, 2010; Chen et al., 2011; Hain et al., 2012).

Revised Section 3.3 (Error model representation):

The parameter uncertainty was represented by perturbing the time constant parameter k_1 for store S_{21}, a highly sensitive parameter of the model that directly affects the streamflow generation by influencing the water stored in both surface storages S_{21} and S_{22} (note that in the PDM formulation used, the time constant k_2 is calculated as a function of k_1). Given the lack of a priori information about the structure of the parameter error, we followed previous SM-DA studies working with rainfall-runoff models (Brocca et al., 2010b, 2012) and adopted a normally distributed multiplicative error with unit mean and standard deviation of s_k."