

Journal: Hydrology and Earth System Sciences

Title: Assimilating Shallow Soil Moisture Observations into Land Models with a Water Budget Constraint

Authors: Bo Dan, Xiaogu Zheng, Guocan Wu, and Tao Li

MS NO.: hess-2019-696

MS Type: Research Article

The authors highly appreciate the anonymous reviewer for his/her very helpful and insightful comments that lead to the considerable improvement of the quality of this manuscript. We have checked our work carefully according to these comments and made the requested changes. The main improvement is the adoption of the bias-aware assimilation scheme and discussions on how this impacts the results.

Below we indicate the comments and use blue font for our responses. The corresponding revised texts are also used blue font in the revised version of our manuscript.

Reviewer #1

It is with great interest that I read this paper, which investigates a number of different approaches to improve the assimilation of near-surface soil moisture observations. Specifically, the authors use an Ensemble Kalman Filter (EnKF) data assimilation (DA) framework to investigate the effect of forecast error inflation, vertical localization, and a weak water balance constraint. They find that the forecast error inflation helps to reduce the analysis error in upper soil layers that are close to the assimilated observations, however, leads to large analysis errors in deeper soil layers. They conclude that the introduction of a vertical localization function can mitigate the increased analysis error in the deeper soil layers. Finally, the authors concluded that the introduction of a weak water constraint helped to reduce the water residual after the assimilation.

Overall, this is a well-written manuscript that presents a novel and scientifically valuable contribution to the field of land data assimilation. I have a few minor comments that I would invite the authors to address before the publication of this manuscript.

Response: [Thank you very much for your thorough reviewing and valuable comments.](#)

General comments:

1. One concern I have is the authors' choice not to implement any bias correction. The argument made is that the removing the model bias can lead to issues, but what is lacking is an argument for not implementing the (arguably more traditional) approach of bias correcting the observations to match the model's climatology. By not

addressing the bias between the observations and model, you are ultimately violating the assumptions of your DA framework and I would at the very least like to see a discussion on how this impacts the results.

Response: Thank you for your comments. Following it and the major comment of another reviewer, the bias-aware data assimilation proposed by Dee (2000) was applied to further correct the bias of the analysis states assimilated using WCEnKF-Inf-Loc. This scheme was named as WCEnKF-Inf-Loc-BA, and the corresponding results were added in Figures 5-6.

Figure 5 shows that, the spatial averaged root analysis error variances of WCEnKF-Inf-Loc and WCEnKF-Inf-Loc-BA were comparable (2.12% for the WCEnKF-Inf-Loc-BA and 2.16% for the WCEnKF-Inf-Loc) for the layers that are shallower than 36.6 cm. This could be due to that the observations are closer to the shallow layers and the vertical localization approach is reasonably effective to reduce the bias. However, for the layers that are deeper than 62.0 cm, the averaged root analysis error of the WCEnKF-Inf-Loc-BA (6.05%) was less than that of the WCEnKF-Inf-Loc (6.59%). This indicates that the bias correction is useful for this experiment, especially for the soil moistures in deeper layers. (Lines 420-428)

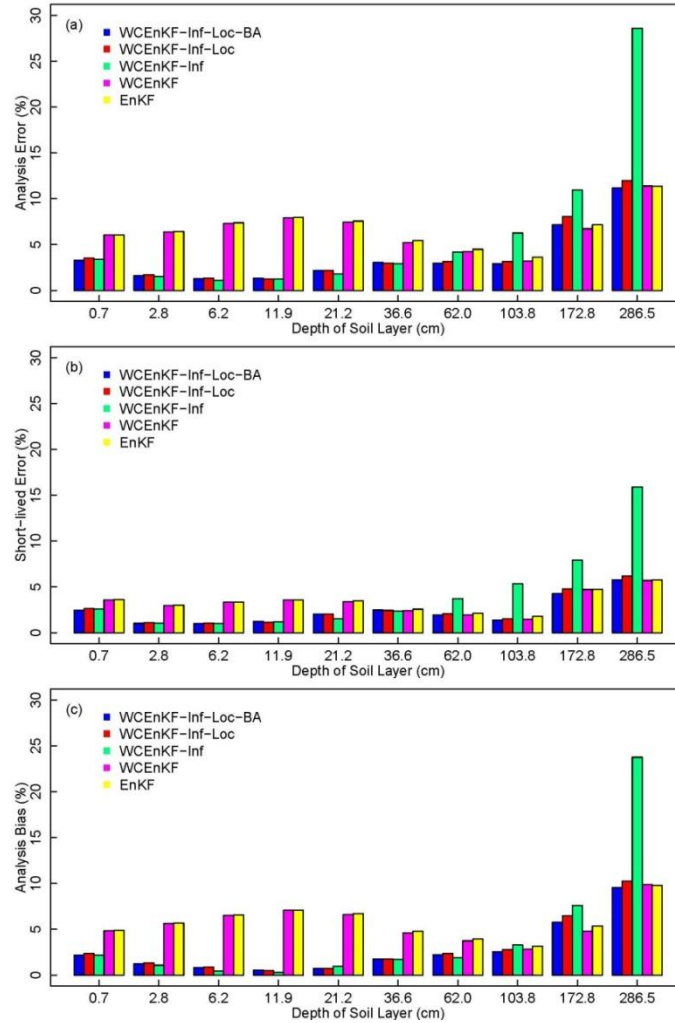


Figure 5. The assimilation results in each layer for the five schemes: a weakly constrained bias-aware ensemble Kalman filter with forecast error inflation and vertical localization (WCEnKF-Inf-Loc-BA), a weakly constrained ensemble Kalman filter with forecast error inflation and vertical localization (WCEnKF-Inf-Loc), a weakly constrained ensemble Kalman filter with forecast error inflation (WCEnKF-Inf), a weakly constrained ensemble Kalman filter (WCEnKF), and the traditional assimilation (EnKF). Graphic (a) is for spatial averaged analysis error of the soil moisture content, (b) is for the short-lived error and (c) is for the analysis bias.

A secondary effect of the lack of bias correction is that it probably aggravates the non-closure of the water balance. While assimilating bias-corrected observations can

still lead to a water imbalance, the effect would likely be reduced and requiring less ‘intervention’ by the water balance constraint introduced in the ‘WC’ experiments. In this context, it is also worth considering that the nature of observed and modeled soil moisture can be very different (see e.g. Koster et al. 2009), which further necessitates a bias correction step.

Koster, R.D., Guo, Z., Yang, R., Dirmeyer, P.A., Mitchell, K. and Puma, M.J., 2009. On the nature of soil moisture in land surface models. *Journal of Climate*, 22(16), pp.4322-4335.

Response: Thank you for your comments. In the revised version, the water budget residuals of different assimilation schemes were shown in Figure 6. The spatial average of the water balance residuals for WCEnKF-Inf-Loc-BA scheme was 0.0723 mm, which was slightly smaller than that for WCEnKF-Inf-Loc scheme (0.0737 mm). The small improvement on water balance residuals may be due to the small improvement on analysis bias by the additional bias-aware assimilation, but it suggests a tendency of the bias correction to further reduce the water balance budget. (Lines 442-445)

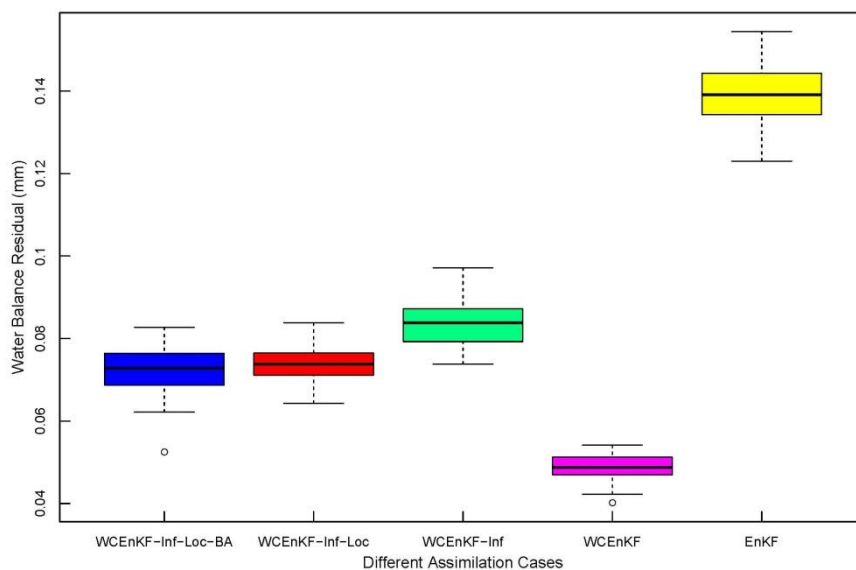


Figure 6. The box plot of the water balance residual in all 40 pixels for the WCEnKF-Inf-Loc-BA, WCEnKF-Inf-Loc, WCEnKF-Inf, WCEnKF and EnKF assimilation schemes.

2. I would like to see some discussion of the feasibility of the proposed approach for global data assimilation. As the authors discuss, in particular the vertical localization function can be strongly location dependent. I am wondering whether expanding to the global domain would require computing the localization function at each model grid cell and whether that would be computationally feasible? Or would you compute a localization function at a regional scale using soil texture or climate regimes to delineate different regions? Some discussion of the transferability of the presented approach to global scales would be valuable.

Response: Thank you for your suggestions. As you pointed out, the most computational cost in the assimilation system is on computing the localization function at each model grid cell. For the synthetic experiments with CoLM model and 40 grids, it takes about 24 hours running on the personal workstation. For global data assimilation with 2° resolution it could take about 3 months to finish if running on the personal workstation, but the super server and parallel computation can significantly shorten the computational time. We also agree with you that “a regional scale using soil texture or climate regimes can be used to delineate different regions”. By this way, the computational time of global data assimilation can be further reduced. (Lines 525-532)

Detailed Comments:

11. 14-15: This sentence is somewhat redundant, I would recommend rephrasing it.

Response: We have revised this sentence to “Assimilating observations of shallow soil moisture content into land models is an important step in estimating soil moisture content.” (Line 14-15)

l. 28: ‘effectively reduces. . .’

Response: Revised.

l. 42. Actually, the sub-seasonal to seasonal time scale is probably the one where the land states have the largest impact, so I would mention it here.

Response: We have added “at sub-seasonal to seasonal time scale”.

l. 63: ‘land model predictions’

Response: Revised.

l. 67: ‘can successfully increase’ or ‘successfully increases’

Response: The sentence has been deleted following the next comment.

ll. 65-74: This paragraph focuses a lot on the assimilation of brightness temperatures, while your study is actually investigating methods to improve soil moisture assimilation. So I would include a discussion of soil moisture assimilation here or even a short discussion of soil moisture vs. brightness temperature assimilation.

Response: Thanks for your comment. This paragraph has been revised as follows.
(Lines 60-69)

Many studies indicated that a better approach to improving the estimates of soil moisture contents on regional scales is to constrain land model predictions by

assimilating surface soil moisture data (Crow and Loon 2006; Crow and Wood 2003; Reichle and Koster 2005). It can provide better estimates of the true soil moisture content column states than the model forecasts (Crow *et al.* 2017; Lu *et al.* 2012; Lu *et al.* 2015), and can further improve land surface model initial conditions for coupled short-term weather prediction (Chen *et al.* 2014; Santanello *et al.* 2016; Yang *et al.* 2016). Especially, surface soil moisture data can be provided by in-situ observations and passive microwave measurements (brightness temperatures) observed by remote sensing.

ll. 115-119: Please see general comment on bias correction of observations.

Response: Thanks for your comment. In the revised version, the sentence “Therefore in this study, we still use traditional bias-blind data assimilation framework.” was replaced as “However, bias can be detected by monitoring observation-minus-forecast statistics in the assimilation system (Dee and Todling 2000). Then a bias-aware assimilation method can be designed to estimate and correct the systematic errors sequentially with the model state variables (Dee 2005). Such bias correction method is adopted in this study to detect the performance among different assimilation schemes.” (Lines 109-114)

ll.128-130: I think you are making an argument here in favor of removing the bias between model and observations.

Response: Yes, we have added the bias correction schemes in the revised version.

ll.178-182: Just to clarify, the synthetic experiments are conducted over 40 pixels, whereas the real data are conducted only over the two ground station sites, correct?

Response: Yes.

II. 230-231: I am surprised to see that you are letting the assimilated soil moisture observations update the canopy water content directly, rather than only updating the soil moisture and letting your vegetation module transport the water into the vegetation layer. Is this approach also taken for the real data assimilation experiments, when you are assimilating in situ soil moisture?

Response: We agree that not update the canopy water content is an option. The approach in this study is adopted from Yilmaz *et al.* (2011; 2012) where the canopy water content was updated. This approach was also taken when we are assimilating in situ soil moistures at the two Mongolia stations.

I. 395: Do you mean ‘At each point..’?

Response: Yes, the words have been revised.

I. 415: When you say ‘the simulation case’, do you mean the open-loop?

Response: Yes it is open-loop. However, the results for the open-loop are not shown in Figures 5 and 6 in the revised version.

II. 446-449: This paragraph needs some language editing.

Response: We have removed the observation study in the revised version. This is because the paper is too long after adding the bias-aware data assimilation experiment, so we would like to shorten it. Also the observations are not available in all layers, and then it is more difficult to validate the proposed assimilation methods. Since these paragraphs were related to observation study, they were deleted in the revised version.

l. 511: 'threshold'

Response: The typo has been revised.

Again, thank you very much for your thorough reviewing and valuable comments.

The references in this reply are listed as follows, while some of them have already in the previous version of the manuscript.

Chen, F., Crow, W.T. and Ryu, D., 2014. Dual Forcing and State Correction via Soil Moisture Assimilation for Improved Rainfall-Runoff Modeling. *Journal of Hydrometeorology*, 15(5): 1832-1848.

Crow, W.T., Chen, F., Reichle, R.H. and Liu, Q., 2017. L band microwave remote sensing and land data assimilation improve the representation of prestorm soil moisture conditions for hydrologic forecasting. *Geophysical Research Letters*, 44(11): 5495-5503.

Crow, W.T. and Loon, E.V., 2006. Impact of incorrect model error assumptions on the sequential assimilation of remotely sensed surface soil moisture. *Journal of Hydrometeorology*, 7: 421-432.

Crow, W.T. and Wood, E.F., 2003. The assimilation of remotely sensed soil brightness temperature imagery into a land surface model using Ensemble Kalman filtering: a case study based on ESTAR measurements during SGP97. *Advances in Water Resources*, 26: 137-149.

Dee, D.P., 2005. Bias and data assimilation. *Quarterly Journal of the Royal*

Meteorological Society, 131: 3323-3343.

Dee, D.P. and Todling, R., 2000. Data assimilation in the presence of forecast bias:

The GEOS moisture analysis. *Monthly Weather Review*, 128(9): 3268-3282.

Lu, H., Koike, T., Yang, K., Hu, Z.Y., Xu, X.D., Rasmy, M., Kuria, D. and

Tamagawa, K., 2012. Improving land surface soil moisture and energy flux simulations over the Tibetan plateau by the assimilation of the microwave remote sensing data and the GCM output into a land surface model.

International Journal of Applied Earth Observation and Geoinformation, 17: 43-54.

Lu, H., Yang, K., Koike, T., Zhao, L. and Qin, J., 2015. An Improvement of the

Radiative Transfer Model Component of a Land Data Assimilation System and Its Validation on Different Land Characteristics. *Remote Sensing*, 7(5): 6358-6379.

Reichle, R.H. and Koster, R.D., 2005. Global assimilation of satellite surface soil moisture retrievals into the NASA Catchment land surface model.

Geophysical Research Letters, 32.

Santanello, J.A., Kumar, S.V., Peters-Lidard, C.D. and Lawston, P.M., 2016. Impact

of Soil Moisture Assimilation on Land Surface Model Spinup and Coupled Land-Atmosphere Prediction. *Journal of Hydrometeorology*, 17(2): 517-540.

Yang, K., Zhu, L., Chen, Y., Zhao, L., Qin, J., Lu, H., Tang, W., Han, M., Ding, B.

and Fang, N., 2016. Land surface model calibration through microwave data assimilation for improving soil moisture simulations. *Journal of Hydrology*,

533: 266-276.

Yilmaz, M.T., Delsole, T. and Houser, P.R., 2011. Improving land data assimilation performance with a water budget constraint. *Journal of Hydrometeorology*, 12: 1040-1055.

Yilmaz, T.M., Delsole, T. and Houser, P.R., 2012. Reducing water imbalance in land data assimilation: Ensemble filtering without perturbed observations. *Journal of Hydrometeorology*, 13(1): 413-420.

Reviewer #2

The paper titled "Assimilating Shallow Soil Moisture Observations into Land Models with a Water Budget Constraint" presented several modifications to the EnKF data assimilation (DA) that potentially improve DA performance in soil moisture (SM) modeling using shallow-layer observations. A forecast error covariance matrix inflation approach to avoid filter divergence due to underestimated sampling and modeling errors is shown to improve modeling accuracy of SM in layers close to the observation, but leads to increased error in the deeper layers. A vertical localization method is applied to limit the updates to the deep layers to mitigate the errors introduced in the deeper layers. A weak constraint on water balance is able to reduce the water balance residual which is increased due to the forecast error covariance inflation at the price of small increase in the analysis error. Overall the results indicate potential usefulness of such modifications in improving soil moisture assimilation accuracy of surface soil moisture observations.

Response: [Thank you very much for your thorough reviewing and valuable comments.](#)

However, there is a major issue in the experiment design that raises my concern, i.e. the lack of observation bias-correction. I found the authors' reasoning behind adopting the "traditional bias-blind data assimilation framework" (line 112-117) unconvincing, as there is no evidence to support the "observations" are unbiased relative to the model background in both the synthetic and real-data experiments in this study. Also, it is well known that remotely sensed soil moisture (the intended application of the proposed modifications) and modeled soil moisture often exhibit different dynamic ranges which warrants the use of a "bias-aware" approach instead

(see e.g. Kumar et al. 2012, doi:10.1029/2010WR010261).

Response: Thank you for your comment. Following it and the major comment of the other reviewer, the bias-aware data assimilation proposed by Dee (2005) was applied to further correct the bias of the analysis states assimilated using WCEnKF-Inf-Loc. This scheme was named as WCEnKF-Inf-Loc-BA, and the corresponding results were added in Figures 5-6.

Figure 5 shows that, the spatial averaged root analysis error variances of WCEnKF-Inf-Loc and WCEnKF-Inf-Loc-BA were comparable (2.12% for the WCEnKF-Inf-Loc-BA and 2.16% for the WCEnKF-Inf-Loc) for the layers that are shallower than 36.6 cm. However, for the layers that are deeper than 62.0 cm, the averaged root analysis error of the WCEnKF-Inf-Loc-BA (6.05%) was less than that of the WCEnKF-Inf-Loc (6.59%). This indicated that the bias correction is useful for this experiment, especially for the soil moistures in deeper layers. (Lines 420-428)

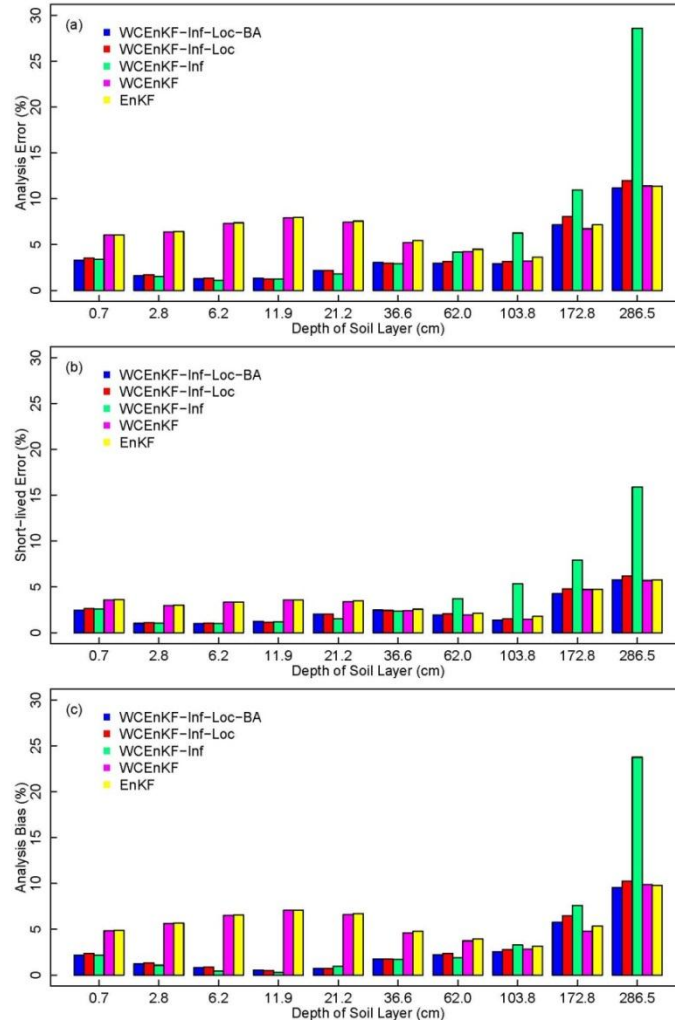


Figure 5. The assimilation results in each layer for the five schemes: a weakly constrained bias-aware ensemble Kalman filter with forecast error inflation and vertical localization (WCEnKF-Inf-Loc-BA), a weakly constrained ensemble Kalman filter with forecast error inflation and vertical localization (WCEnKF-Inf-Loc), a weakly constrained ensemble Kalman filter with forecast error inflation (WCEnKF-Inf), a weakly constrained ensemble Kalman filter (WCEnKF), and the traditional assimilation (EnKF). Graphic (a) is for spatial averaged analysis error of the soil moisture content, (b) is for the short-lived error and (c) is for the analysis bias.

In addition, there is an apparent misunderstanding of the Koster et al. (2009) and Reichle and Koster (2004) works where the authors stated that "A major objective of

soil moisture data assimilation is to address biases in models and observations" (line 110-111). In fact, both publications indicated the importance of removing the bias in the statistical moments in the observations relative to the model background prior to data assimilation. The major objective of data assimilation is not to remove the bias in model states but to reduce the random, mean-zero noise in the model states, with the model state climatology respected. Even if the observations are considered unbiased, it is recommended that the observations be "scaled" to match the statistical moments of the model states (with long enough time-series). It is well known that directly assimilation of raw observations likely causes model integration to drift, i.e. introduce further bias to the model states. Therefore, the model water balance residual after the soil moisture update in the experiments in this study may be partly attributed to assimilating observations without bias-correction (relative to model), and the true effect of the weak water balance constraint is not accurately revealed.

I would like suggest that the DA experiments repeated with a more robust "bias-aware" approach, to rule out the impact of observation bias in the analysis errors so that the effects of the proposed modifications are better isolated.

Response: Thank you for your comments. We agree that the model state climatology should be respected and directly assimilation of raw observations likely causes model integration to drift. We also agree that a more robust "bias-aware" approach is necessary, because it respects the model state climatology and uses the estimated bias to prevent model integration to drift. In the revised version, the bias-aware data assimilation proposed by Dee (2005) was investigated.

In the revised version, the water budget residuals of different assimilation schemes were shown in Figure 6. The spatial average of the water balance residuals for WCEnKF-Inf-Loc-BA scheme was 0.0723 mm, which was slightly smaller than

that for WCEnKF-Inf-Loc scheme (0.0737 mm). The small improvement on water balance residuals may be due to the small improvement on analysis bias by the additional bias-aware assimilation, but it suggests a tendency of the bias correction to further reduce the water balance budget. (Lines 442-445)

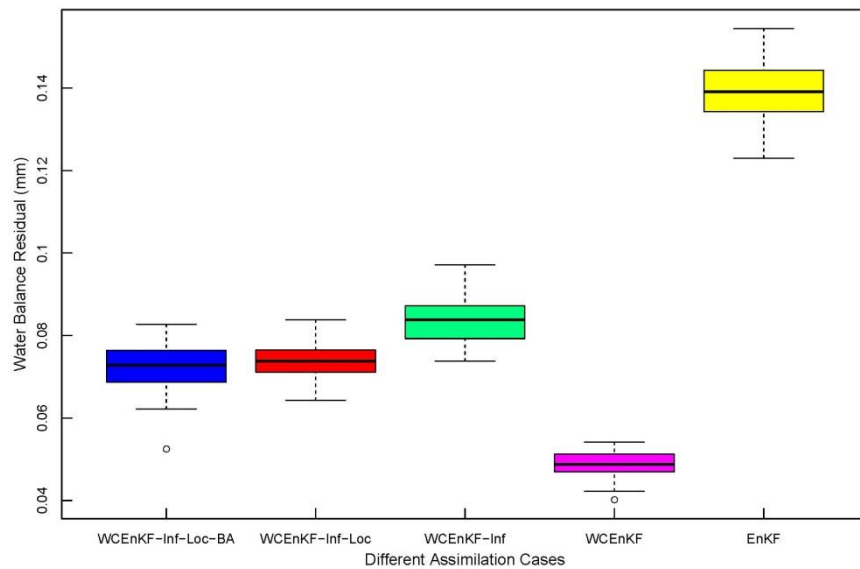


Figure 6. The box plot of the water balance residual in all 40 pixels for the WCEnKF-Inf-Loc-BA, WCEnKF-Inf-Loc, WCEnKF-Inf, WCEnKF and EnKF assimilation schemes.

Other minor comments:

Lines 65-74: irrelevant to the topic of the paper and should be removed.

Response: Thanks for the comment. Following it and the comment of the other reviewer, lines 65-74 were removed and the paragraph is revised as follows:

Many studies indicated that a better approach to improving the estimates of soil moisture contents on regional scales is to constrain land model predictions by assimilating surface soil moisture data (Crow and Loon 2006; Crow and Wood 2003; Reichle and Koster 2005). It can provide better estimates of the true soil moisture

content column states than the model forecasts (Crow *et al.* 2017; Lu *et al.* 2012; Lu *et al.* 2015), and can further improve land surface model initial conditions for coupled short-term weather prediction (Chen *et al.* 2014; Santanello *et al.* 2016; Yang *et al.* 2016). Especially, surface soil moisture data can be provided by in-situ observations and passive microwave measurements (brightness temperatures) observed by remote sensing. (Lines 60-69)

Line 229: why directly update canopy water content and snow water equivalent when these two variables are not regulated by near-surface soil moisture?

Response: We agree that not update the canopy water content and snow water equivalent is an option. The approach in this study is adopted from Yilmaz *et al.* (2011; 2012) where the canopy water content and snow water equivalent were updated.

Line 448: remove "the" following "cover"

Response: Revised.

Line 478: "the different experiments" -> "different cases"

Response: This paragraph was related to cases of the DGS and BTS stations, and was removed in the revised version.

Lines 479-480: for better understanding of the magnitude of improvement, use a percentage scale for water balance residuals

Response: Thank you for your suggestion. In the revised version, the real data experiments were deleted due to the length of the manuscript.

Line 494 "deflation <of> the water balance ..."; "plain" ->"plainly"

Response: Revised.

Line 511 "shreshold" ->"threshold"

Response: Revised.

Sec 7.2 Again, one should be careful to use data assimilation to achieve "bias correction" in model states. This is another example of misunderstanding the major objective of DA in this work. Seemingly reduced systematic bias in modeled soil moisture may be an artifact due to biased observation relative to model background.

Response: Thank you for your comment. We have added the experimentation with bias correction method. The results shows that the bias-ware assimilation schemes can further reduce the analysis error and water budget residuals.

Again, thank you very much for your thorough reviewing and valuable comments. The references in this reply are listed as follows, while some of them have already in the previous version of the manuscript.

Chen, F., Crow, W.T. and Ryu, D., 2014. Dual Forcing and State Correction via Soil Moisture Assimilation for Improved Rainfall-Runoff Modeling. *Journal of Hydrometeorology*, 15(5): 1832-1848.

Crow, W.T., Chen, F., Reichle, R.H. and Liu, Q., 2017. L band microwave remote sensing and land data assimilation improve the representation of prestorm soil moisture conditions for hydrologic forecasting. *Geophysical Research Letters*,

44(11): 5495-5503.

Crow, W.T. and Loon, E.V., 2006. Impact of incorrect model error assumptions on the sequential assimilation of remotely sensed surface soil moisture. *Journal of Hydrometeorology*, 7: 421-432.

Crow, W.T. and Wood, E.F., 2003. The assimilation of remotely sensed soil brightness temperature imagery into a land surface model using Ensemble Kalman filtering: a case study based on ESTAR measurements during SGP97. *Advances in Water Resources*, 26: 137-149.

Lu, H., Koike, T., Yang, K., Hu, Z.Y., Xu, X.D., Rasmy, M., Kuria, D. and Tamagawa, K., 2012. Improving land surface soil moisture and energy flux simulations over the Tibetan plateau by the assimilation of the microwave remote sensing data and the GCM output into a land surface model. *International Journal of Applied Earth Observation and Geoinformation*, 17: 43-54.

Lu, H., Yang, K., Koike, T., Zhao, L. and Qin, J., 2015. An Improvement of the Radiative Transfer Model Component of a Land Data Assimilation System and Its Validation on Different Land Characteristics. *Remote Sensing*, 7(5): 6358-6379.

Reichle, R.H. and Koster, R.D., 2005. Global assimilation of satellite surface soil moisture retrievals into the NASA Catchment land surface model. *Geophysical Research Letters*, 32.

Santanello, J.A., Kumar, S.V., Peters-Lidard, C.D. and Lawston, P.M., 2016. Impact of Soil Moisture Assimilation on Land Surface Model Spinup and Coupled

Land-Atmosphere Prediction. *Journal of Hydrometeorology*, 17(2): 517-540.

Yang, K., Zhu, L., Chen, Y., Zhao, L., Qin, J., Lu, H., Tang, W., Han, M., Ding, B. and Fang, N., 2016. Land surface model calibration through microwave data assimilation for improving soil moisture simulations. *Journal of Hydrology*, 533: 266-276.

Yilmaz, M.T., Delsole, T. and Houser, P.R., 2011. Improving land data assimilation performance with a water budget constraint. *Journal of Hydrometeorology*, 12: 1040-1055.

Yilmaz, T.M., Delsole, T. and Houser, P.R., 2012. Reducing water imbalance in land data assimilation: Ensemble filtering without perturbed observations. *Journal of Hydrometeorology*, 13(1): 413-420.

The main changes are listed as follows.

(1) Lines 60-69: Revised the discussion about soil moisture assimilation and removed the irrelevant texts to the topic.

(2) Lines 109-114: Revised the description of the bias correction scheme.

(3) Lines 288-300: Added the algorithm of the bias-aware assimilation scheme.

(4) Lines 420-428: Added the analysis error of the weakly constrained ensemble Kalman filter with inflation, localization and bias-aware assimilation (WCEnKF-Inf-Loc-BA).

(5) Lines 442-445: Added the water budget residuals of the weakly constrained ensemble Kalman filter with inflation, localization and bias-aware assimilation (WCEnKF-Inf-Loc-BA).

(6) Lines 525-532: Added the discussion about the computational time of global data assimilation.

(7) Some concepts were clarified and the English text was polished in the revised version.

1 **Assimilating Shallow Soil Moisture Observations into Land Models**
2 **with a Water Budget Constraint**

3

4 Bo Dan¹, Xiaogu Zheng², Guocan Wu^{3*}, and Tao Li⁴

5

6 ¹ National Marine Data and Information Service, Tianjin, China

7 ²Key Laboratory of Regional Climate-Environment Research for East Asia, Institute
8 of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China

9 ³College of Global Change and Earth System Science, Beijing Normal University,
10 Beijing, China

11 ⁴Institute of Statistics, Xi'an University of Finance and Economics, Xi'an, China

12

*Corresponding author: Guocan Wu
E-mail: gcwu@bnu.edu.cn

13 **Abstract**

14 **Assimilating** observations of shallow soil moisture content into land models is an
15 important step in **estimating** soil moisture content. In this study, several modifications
16 of an ensemble Kalman filter (EnKF) are proposed for improving this assimilation. It
17 was found that a forecast error inflation-based approach improves the soil moisture
18 content in shallow layers, but it can increase the analysis error in deep layers. To
19 mitigate the problem in deep layers while maintaining the improvement in shallow
20 layers, a vertical localization-based approach was introduced in this study. During the
21 data assimilation process, although updating the forecast state using observations can
22 reduce the analysis error, the water balance based on the physics in the model could
23 be destroyed. To alleviate the imbalance in the water budget, a weak water balance
24 constrain filter is adopted.

25 The proposed weakly constrained EnKF that includes forecast error inflation and
26 vertical localization was applied to a synthetic experiment. The results of the
27 assimilation process suggest that the inflation approach effectively **reduces** both the
28 short-lived analysis error and the analysis bias in shallow layers, while the vertical
29 localization approach avoids increase in analysis error in deep layers. **Finally, an**
30 **additional bias-aware assimilation for reducing the analysis bias is investigated.**

31
32 **Keywords** soil moisture, water balance, data assimilation, forecast error inflation,
33 vertical localization

35 **1. Introduction**

36 Soil moisture content is one of the most important variables that affect the water
37 cycle and energy balance through land-atmosphere interactions, especially
38 evaporation and precipitation (Han *et al.* 2014; Kumar *et al.* 2014; McColl *et al.* 2019;
39 Pinnington *et al.* 2018). Adequate knowledge of the horizontal and vertical
40 distributions of soil moisture [at sub-seasonal to seasonal time scale](#) could improve
41 weather and climate predictions (Delworth and Manabe 1988; Pielke 2001).
42 Alongside snow cover, soil moisture content is an important component of the
43 meteorological memory of the climate system over land (McColl *et al.* 2019; Robock
44 *et al.* 2000; Zhao and Yang 2018). It is also a primary water resource for the terrestrial
45 ecosystem and affects runoff (GUSEV and Novak 2007).

46 There are several ways to estimate the soil moisture content. Land surface
47 models can provide temporally and spatially continuous estimates of the soil moisture
48 content, but [limited](#) by the uncertainty in the models' parameters, errors in the forcing
49 data and imperfect physical parameterizations (Bonan 1996; Dai *et al.* 2003;
50 Dickinson *et al.* 1993; Oleson *et al.* 2010; Yang *et al.* 2009). Compared with the
51 results of models, in-situ observations of the soil moisture content provide more
52 accurate profiles (Bosilovich and Lawford 2002; Dorigo *et al.* 2011; Robock *et al.*
53 2000); however, networks of in-situ observations are usually too sparse to estimate the
54 soil moisture content on a regional scale (Gruber *et al.* 2018; Loizu *et al.* 2018).
55 Satellite remote sensing retrievals could provide soil moisture content data on regional
56 scales (Bartalis *et al.* 2007; Crow *et al.* 2017; Entekhabi *et al.* 2010; Kerr *et al.* 2010;
57 Lu *et al.* 2015; Njoku *et al.* 2003), but they are only available for the shallow layer of
58 the soil and the quality is poor in vegetated area (Pinnington *et al.* 2018; Yang *et al.*
59 2009).

60 Many studies indicated that a better approach to improving the estimates of soil
61 moisture contents on regional scales is to constrain land model predictions by
62 assimilating surface soil moisture data (Crow and Loon 2006; Crow and Wood 2003;
63 Reichle and Koster 2005). It can provide better estimates of the true soil moisture
64 content column states than the model forecasts (Crow *et al.* 2017; Lu *et al.* 2012; Lu
65 *et al.* 2015), and can further improve land surface model initial conditions for coupled
66 short-term weather prediction (Chen *et al.* 2014; Santanello *et al.* 2016; Yang *et al.*
67 2016). Especially, surface soil moisture data can be provided by in-situ observations
68 and passive microwave measurements (brightness temperatures) observed by remote
69 sensing.

70 A good estimate of the forecast error covariance matrix is crucial for the
71 compromise between uncertain observations and imperfect model predictions in data
72 assimilation (Anderson and Anderson 1999; Miyoshi 2011; Miyoshi *et al.* 2012; Wang
73 and Bishop 2003). For the Ensemble Kalman Filter (EnKF) assimilation method, the
74 forecast error covariance matrix is estimated using the sample covariance matrix of
75 the ensemble forecasts (Dumedah and Walker 2014; Evensen 1994; Han *et al.* 2014).
76 However, it is usually underestimated due to sampling and model errors, which can
77 eventually results in filter divergence (Anderson and Anderson 1999; Constantinescu
78 *et al.* 2007; Yang *et al.* 2015). To address this problem, it suggests that the forecast
79 covariance matrix be multiplied by a factor (Dee and Da Silva 1999; Dee *et al.* 1999;
80 Li *et al.* 2012; Zheng 2009). This approach is referred to as inflation, and it becomes
81 particularly important when the error in the model is large (Bauser *et al.* 2018; El
82 Gharamti *et al.* 2019; Liang *et al.* 2012; Raanes *et al.* 2019; Wu *et al.* 2013).
83 Therefore, it could work well in this situation because of the enormous errors in the
84 land model.

85 In this study, a scheme for assimilating **synthetic** observations of the soil
86 moisture content into land models was developed based on EnKF method, which can
87 provide a foundation for further satellite data assimilation. For the synthetic
88 experiment, the Version 4.0 of the Community Land Model (CLM 4.0, (Lawrence *et*
89 *al.* 2011; Oleson *et al.* 2010)) was used to generate the “true values” and the Common
90 Land Model (CoLM, (Dai *et al.* 2003)) was selected as the forecast operator. The
91 differences in these two models are referred to the model error in an imperfect land
92 surface model. The inflation factors are estimated at every observation time step
93 during the assimilation process by minimizing the -2log-likelihood of the difference
94 between the forecast and the observation (Liang *et al.* 2012; Zheng 2009). For
95 assimilating observations near the surface only, such inflation approach can improve
96 the estimates of the forecast error statistics in shallow soil layers but may artificially
97 enlarge the forecast error statistics in deep soil layers. To avoid the possibility of
98 decreasing the quality of the estimates in deep soil layers, a vertical localization with
99 weighting of observations is adopted (Janjić *et al.* 2011). In this approach, a
100 localization function multiplies the weights on the components of the state vector
101 according to the distance from state layer to the observation. Moreover, the method
102 based on the maximum likelihood estimation was proposed to estimate the optimal
103 localization scale factor.

104 A major objective of soil moisture data assimilation is to address biases in
105 models and observations (Koster *et al.* 2009; Reichle and Koster 2004). In this study,
106 we only assume that models could be biased, while the soil moisture observations are
107 assumed to be unbiased. Moreover, the soil moisture observations are restricted in
108 shallow layer, so there is no observation available to **directly** correct the modeled soil
109 moisture biases in deep layers. **However, bias can be detected by monitoring**

110 observation-minus-forecast statistics in the assimilation system (Dee and Todling
111 2000). Then a bias-aware assimilation method can be designed to estimate and correct
112 the systematic errors sequentially with the model state variables (Dee 2005). Such
113 bias correction method is adopted in this study to detect the performance among
114 different assimilation schemes. Furthermore, the analysis error is decomposed to a
115 short-lived error (random error) and a bias (system error). It demonstrates that the
116 proposed scheme can reduce the both for soil moisture in shallow layers. These
117 improvements steps can also result in a reasonable estimates of the soil moisture
118 content in the deep layers.

119 In addition to improve assimilation accuracy, this study also focuses on the
120 imbalance in the water budget that occurs during the process of assimilating the soil
121 moisture data. The terrestrial water budget is a key part of the global hydrologic cycle.
122 A better understanding of the budget can help us to improve our knowledge of
123 land-atmosphere water exchange and related physical mechanisms and therefore, can
124 improve our ability to develop models (Pan and Wood 2006). Generally speaking,
125 analyses do not conserve the water budget due to inconsistencies between predictions
126 made by models and observations (Li *et al.* 2012; Pan and Wood 2006; Wei *et al.*
127 2010; Yilmaz *et al.* 2011; Yilmaz *et al.* 2012). It is really a problem if the water
128 balance is violated in a systematic manner (for example, model is biased), which
129 suggests a trouble in data assimilation. Pan and Wood (2006) proposed a method
130 based on a strong constraint to reincorporate the water balance. However, this method
131 redistributes the error among the different terms in the water budget, which could
132 result in unrealistic estimates (Pan and Wood 2006; Yilmaz *et al.* 2011).

133 To overcome this shortcoming, Yilmaz *et al.* (2011) proposed using a weakly
134 constrained ensemble Kalman filter (WCEnKF) to reduce the imbalance in the water

135 budget. In a synthetic study, they concluded that the accuracy of a WCEnKF-based
136 analysis is close to that of an EnKF-based analysis but the water budget balance
137 residuals are much smaller than that of an unconstrained filter. Nevertheless, the
138 observations of the soil moisture content cover the entire column, and a perfect model
139 was used in their studies. This is not generally true, especially when only satellite
140 observations are assimilated. In this study, the experiments were further designed to
141 assimilate surface observations into an imperfect land model.

142 The structure of this paper is arranged as follows: The data and models used in
143 this study are described in section 2. The details of the WCEnKF-based methods that
144 incorporate inflation, vertical localization and bias-aware assimilation are provided in
145 section 3. The experimental designs and evaluations of synthetic experiments are set
146 in sections 4. The primary results are given in section 5. The discussion and
147 conclusion comprise sections 6 and 7.

148

149 **2. Models and data**

150 2.1 Study area

151 The study area is located in the Mongolian Plateau and comprises approximately
152 9352 square kilometers between 46° and 46.5° N and between 106.125° and 107° E.
153 The dominant biome is grassland, and no river flows through the area (see Figure 1).

154 The soil moisture content and related meteorological and hydrological parameters
155 are monitored by automatic stations maintained by the Coordinated Enhanced
156 Observing Period Asian Monsoon Project (CEOP AP) (Bosilovich and Lawford 2002;
157 Lawford *et al.* 2004). The CEOP AP was launched by the World Climate Research
158 Programme (WCRP) to develop an integrated global dataset that can be used to
159 address issues relating to water and energy budget simulations and predictions,

160 monsoon processes and the prediction of river flows. More details can be found at
161 <http://www.ceop.net>.

162

163 2.2 Forcing data

164 In this study, *synthetic* experiments were conducted to explore the accuracy of the
165 assimilation schemes. The simulations were driven by forcing data (including
166 radiation, wind, pressure, humidity, precipitation and temperature) from the
167 0.125°x0.125° ERA-Interim dataset (Dee *et al.* 2011) that had been scaled down to
168 provide a temporal resolution of one hour.

169

170 2.3 Models

171 The Common Land Model (CoLM) developed by Dai *et al.* (2003) is a
172 third-generation land surface model. It combines the best features of three successful
173 models: the Land Surface Model (LSM, (Bonan 1996)), the Biosphere-Atmosphere
174 Transfer Scheme (BATS, (Dickinson *et al.* 1993)) and the 1994 version of the Chinese
175 Academy of Sciences/Institute of Atmospheric Physics model (IAP94, (Dai *et al.*
176 2003)), and is being further developed. The primary characteristics of the model
177 include 10 unevenly spaced soil layers (see Table 1), one vegetation layer, 5 snow
178 layers (depending on the snow depth), explicit treatment of the mass of liquid water,
179 ice and phase changes within the system of the snow and soil, runoff parameterization
180 following the TOPMODEL concept, a tiled treatment of the sub-grid fraction of the
181 energy and water budget balance (Dai *et al.* 2003) and a canopy
182 photosynthesis-conductance mode that describes the simultaneous transfer of CO₂ and
183 water vapor into and out of the vegetation. The model parameters include data on the
184 global terrain, elevation, land use, vegetation, land-water mask and hybrid

185 FAO/STATSGO soil types from the USGS, which are available at a resolution of 30
186 arc seconds.

187 Version 4.0 of the Community Land Model (CLM 4.0) (Lawrence *et al.* 2011;
188 Oleson *et al.* 2010) is the land surface parameterization used with the Community
189 Atmosphere Model (CAM 4.0) and the Community Climate System Model (CCSM
190 4.0). The CLM 4.0 includes bio-geophysics, the hydrologic cycle, biogeochemistry
191 and the dynamic vegetation. CLM 4.0 simulates the bio-geophysical processes in each
192 sub-grid unit independently and maintains its own prognostic variables. The
193 parameters used in the CLM4.0 differ from those used in the CoLM. For example, the
194 soil texture data are derived from the IGBP soil data, and the land use data are derived
195 from the UNH Transient Land Use and Land Cover Change Dataset
196 (<http://luh.umd.edu/>).

197 In addition to using different parameters, the two models have different structures.
198 For example, a model of groundwater-soil water interactions (Niu *et al.* 2007; Niu *et*
199 *al.* 2005) has been incorporated into the CLM 4.0, while zero water flux at the bottom
200 of a soil column is assumed in the CoLM. Besides, the CLM 4.0 has the same vertical
201 discretization scheme as the CoLM (see Table 1), which makes comparing the results
202 of the two models convenient.

203

204 **3. Methods**

205 3.1 Forecast and observation systems

206 Using notation similar to that used by Yilmaz *et al.* (2011), the forecast system
207 can be written as

$$208 \quad \mathbf{y}_{n,t}^f = \mathbf{M}_{n,t-1}(\mathbf{y}_{n,t-1}^a), \quad (1)$$

209 where $t=1, \dots, T$ is the time index, $n=1, \dots, N$ represents an ensemble member (in this

210 study, the ensemble size is set to 100), $\mathbf{M}_{n,t-1}$ is a CoLM forced by the n -th perturbed
 211 atmospheric forcing, and \mathbf{y} is a state vector containing 126 variables. The superscript
 212 “ f ” and “ a ” specify the forecast and analysis, respectively.

213 Let \mathbf{x} be the state variables related to the water budget, that comprises of **SM**
 214 and **SIC** (the soil moisture content and the soil ice content in % at the 10 vertical
 215 levels listed in Table 1), CWC and SWE (the canopy’s water content and the snow
 216 water equivalent in kg/m²). In this study, only \mathbf{x} is updated by data assimilation, while
 217 the model propagates changes to the other variables over time.

218 For the traditional EnKF, the forecast error covariance matrix \mathbf{P}_t is
 219 obtained from the ensemble of their anomalies,

$$220 \quad \mathbf{P}_t = \frac{1}{N-1} \sum_{n=1}^N (\mathbf{x}_{n,t}^f - \mathbf{x}_t^f)(\mathbf{x}_{n,t}^f - \mathbf{x}_t^f)^T. \quad (2)$$

221 where $\mathbf{x}_{n,t}^f$ is the component of $\mathbf{y}_{n,t}^f$ related to the water budget, \mathbf{x}_t^f is the ensemble
 222 mean of $\mathbf{x}_{n,t}^f$. To avoid overestimation of the co-variability between shallow
 223 observations and soil moistures deeper than a threshold layer s (see section 3.2 for the
 224 estimation of s), the following vertical localization function with weighting of
 225 observations ρ_s (Janjić *et al.* 2011) will be applied on \mathbf{P}_t , i.e.,

$$226 \quad \rho_s(l) = \exp(-\mu_s |d_l - d_o|) \quad (3)$$

227 where l represents for the l -level soil layer, d_l and d_o represent the depths of
 228 l -level soil layer and observation, respectively. $|d_l - d_o|$ is the Euclidian distance
 229 between the two layers. μ_s is estimated by minimizing the following mean square
 230 error between vertical localization function Eq (3) and a step function with threshold
 231 layer s ,

232
$$M(\mu) = \sum_{l \leq s} [\exp(-\mu|d_l - d_o|) - 1]^2 + \sum_{l > s} [\exp(-\mu|d_l - d_o|)]^2 \quad (4)$$

233 The estimated μ_s is listed in Table 2.

234 The observations of the soil moisture content are collected at a depth of 3 cm at
235 6:00 am every day (denoted by o_t). The observation system is defined as

236
$$o_t = \mathbf{h}\mathbf{x}_t + \varepsilon_t, \quad (5)$$

237 where observational operator \mathbf{h} is a 22-dimensional vector which linearly interpolated
238 the soil moisture at depths of 2.8 cm and 6.2 cm to depth of 3 cm, \mathbf{x}_t represents the
239 true values of the state variables related to the water budget at the time step t and ε_t
240 is the observational error with mean zero and variance R_t . Since, the main objective
241 of this study is for methodology related to linear observational operators. Choosing
242 the linear interpolation as observational operator is only for convenience.

243

244 3.2 Assimilation with water budget constraint

245 Assimilating data on the soil moisture content usually results in an imbalance in
246 the water budget. To reduce this imbalance, a weak constraint on the water budget
247 (Yilmaz *et al.* 2011) is adopted in this study. The ensemble water budget residual at
248 time step t can be expressed as

249
$$r_{n,t} \equiv \beta_{n,t} - \mathbf{c}^T \mathbf{x}_{n,t}^a, \quad (6)$$

250 where

251
$$\beta_{n,t} = \mathbf{c}^T \mathbf{x}_{n,t-1}^a + Pr_t - Ev_{n,t}^f - Rn_{n,t}^f, \quad (7)$$

252 where \mathbf{c} is a 22-dimensional vector that converts the units to millimeters (mm) and
253 adds up the states in \mathbf{x} , the diagnostic variables Pr_t , $Ev_{n,t}^f$ and $Rn_{n,t}^f$ (mm) are
254 scalars specifying the states of the precipitation, evapotranspiration and runoff,

255 respectively, in each pixel.

256 The cost function used to estimate the state variables with the weak water budget
 257 constraint (Eq. (6)) is

$$258 \quad J_{n,t}(\mathbf{x}) = (o_t - \mathbf{h}\mathbf{x})^T R_t^{-1} (o_t - \mathbf{h}\mathbf{x}) + (\mathbf{x} - \mathbf{x}_{n,t}^f)^T \mathbf{P}_{s,t}^{-1} (\mathbf{x} - \mathbf{x}_{n,t}^f) \\ + (\beta_{n,t} - \mathbf{c}^T \mathbf{x})^T \varphi_t^{-1} (\beta_{n,t} - \mathbf{c}^T \mathbf{x}), \quad (8)$$

259 where

$$260 \quad \varphi_t = \frac{1}{N-1} \sum_{n=1}^N \left(\beta_{n,t} - \frac{1}{N} \sum_{j=1}^N \beta_{j,t} \right) \times \left(\beta_{n,t} - \frac{1}{N} \sum_{j=1}^N \beta_{j,t} \right)^T \quad (9)$$

261 is an estimate of the variance of $\beta_{n,t}$ and $\mathbf{P}_{s,t}$ represents a forecast error
 262 covariance matrix defined by

$$263 \quad \mathbf{P}_{s,t} = \left[\sqrt{\lambda_t} \right] [\boldsymbol{\rho}_s] \mathbf{P}_t [\boldsymbol{\rho}_s] \left[\sqrt{\lambda_t} \right]. \quad (10)$$

264 where \mathbf{P}_t is defined as Eq. (2); $[\boldsymbol{\rho}_s]$ is a diagonal matrix which localizes the soil
 265 moisture error (i.e. it is $\boldsymbol{\rho}_s$ defined by Eq. (3) for the soil moisture contents and 1 for
 266 other variables). $[\sqrt{\lambda_t}]$ is also a diagonal matrix which inflates the forecast soil
 267 moisture error (i.e. it is a scalar λ_t for the soil moisture contents and 1 for other
 268 variable). λ_t is estimated by minimizing the -2log-likelihood of the difference
 269 between the forecast and the observation (Dee and Da Silva 1999; Liang *et al.* 2012;
 270 Zheng 2009),

$$271 \quad -2L_{s,t}(\lambda_t) = \ln(\mathbf{h}\mathbf{P}_{s,t}\mathbf{h}^T + R_t) + (o_t - \mathbf{h}\mathbf{x}_t^f)^T (\mathbf{h}\mathbf{P}_{s,t}\mathbf{h}^T + R_t)^{-1} (o_t - \mathbf{h}\mathbf{x}_t^f). \quad (11)$$

272 The estimated forecast error inflation factor is denoted as $\hat{\lambda}_t$. The perturbed analysis
 273 states of the variables related to water budget can be derived by minimizing Eq. (8),
 274 which has the analytic form

275
$$\mathbf{x}_{n,t}^a = \mathbf{x}_{n,t}^f + \mathbf{P}_t^a \mathbf{h}^T R_t^{-1} (o_t + \varepsilon_{n,t} - \mathbf{h} \mathbf{x}_{n,t}^f) + \mathbf{P}_t^a \mathbf{c} \varphi_t^{-1} (\beta_{n,t} - \mathbf{c}^T \mathbf{x}_{n,t}^f), \quad (12)$$

276 where $\varepsilon_{n,t}$ is generated from a normal distribution with mean zero and variance R_t ,
 277 and

278
$$\mathbf{P}_t^a = \left(\mathbf{h}^T R_t^{-1} \mathbf{h} + \mathbf{P}_{s,t}^{-1} + \mathbf{c} \varphi_t^{-1} \mathbf{c}^T \right)^{-1}, \quad (13)$$

279 its analysis error covariance matrix.

280 For estimating the optimal threshold layer, define the -2log-likelihood of the total
 281 difference between the forecasts and the observations,

282
$$L_s \equiv \sum_{t=1}^T (-2L_{s,t}(\hat{\lambda}_t)). \quad (14)$$

283 The optimal threshold layer \hat{s} is selected as the smallest number s such that L_s is
 284 the minimum of $\{L_2, L_3, \dots, L_{s+1}\}$. The final analysis state is the selected
 285 corresponding to the optimal threshold layer \hat{s} . The complete assimilation procedure
 286 with water budget constraint is shown in Figure 2.

287

288 3.3 Bias-aware assimilation

289 The bias-aware data assimilation proposed by Dee (2005) is adopted to correct
 290 the analysis bias.

291 Let \mathbf{b}_t is the estimated bias at time step t and set $\mathbf{b}_1 = 0$. For $t > 1$,

292
$$\mathbf{b}_t = \mathbf{b}_{t-1} - \gamma \tilde{\mathbf{P}}_{s,t} \mathbf{h}^T \left(\mathbf{h} \tilde{\mathbf{P}}_{s,t} \mathbf{h}^T + R_t \right)^{-1} \left(o_t - \mathbf{h}(\tilde{\mathbf{x}}_t^f - \mathbf{b}_{t-1}) \right). \quad (15)$$

293 where the scalar parameter γ that controls the magnitude of the forecast bias is
 294 estimated following Dee and Todling (2000) (see Eqs (A5)-(A6) of Appendix A), $\tilde{\mathbf{x}}_t^f$

295 is the ensemble mean of the perturbed forecast states $\tilde{\mathbf{x}}_{n,t}^f$ from the analysis state

296 $\tilde{\mathbf{x}}_{n,t-1}^a$, $\tilde{\mathbf{P}}_{s,t}$ is the corresponding adjusted forecast error covariance (see Eq. (A2) of

297 Appendix A).

298 Then the perturbed assimilated states are

$$\begin{aligned} 299 \quad \tilde{\mathbf{x}}_{n,t}^a = & \tilde{\mathbf{x}}_{n,t}^f - \mathbf{b}_{t-1} + \tilde{\mathbf{P}}_t^a \mathbf{h}^T R_t^{-1} \left(o_t + \varepsilon_{n,t} - \mathbf{h} \left(\tilde{\mathbf{x}}_{n,t}^f - \mathbf{b}_{t-1} \right) \right) \\ & + \tilde{\mathbf{P}}_t^a \mathbf{c} \tilde{\varphi}_t^{-1} \left(\tilde{\beta}_{n,t} - \mathbf{c}^T \left(\tilde{\mathbf{x}}_{n,t}^f - \mathbf{b}_{t-1} \right) \right) \end{aligned} \quad (16)$$

300 where $\tilde{\beta}_{n,t}$, $\tilde{\varphi}_t^{-1}$ and $\tilde{\mathbf{P}}_t^a$ are defined by Eqs (A7)-(A9) in Appendix A respectively.

301

302 4. Synthetic experiments

303 4.1 Experimental design

304 To investigate the performance of the WCEnKF-based methods that incorporate
305 inflation, vertical local localization and bias-aware assimilation, synthetic experiments
306 were performed using the CoLM. Unlike the “perfect model” assumption used in
307 Yilmaz et al. (2011), the assumptions of this study are accounted for the error in the
308 model, especially the structural error. Because there were structural differences in the
309 models of the water cycle (see section 2.3) used in the two models, CLM 4.0 was used
310 to generate the “true values” (i.e., to perform a reference run) for the synthetic
311 experiments and CoLM was selected as the forecast operator (i.e., to perform an
312 open-loop run). Therefore, the CLM 4.0 and the CoLM were both integrated on a
313 0.125° grid (see Figure 1 for the locations) with a time step of one hour. The
314 assimilation time was set to 6:00 am every day. The assimilation experiments were
315 conducted with 5 scenarios: the traditional ensemble Kalman filter (EnKF), a weakly
316 constrained ensemble Kalman filter (WCEnKF), a weakly constrained ensemble
317 Kalman filter with inflation (WCEnKF-Inf), a weakly constrained ensemble Kalman
318 filter with inflation and localization (WCEnKF-Inf-Loc) and a weakly constrained
319 ensemble Kalman filter with inflation, localization and bias-aware assimilation
320 (WCEnKF-Inf-Loc-BA).

321 Synthetic observations were obtained by interpolating \mathbf{SM}_t to a depth of 3 cm
322 and adding noise with a normal distribution ($N(\mu=0, \sigma=0.5\%)$). The initial state
323 \mathbf{x}_0 , was generated by running the CoLM from October 1, 2002 to June 1, 2003. Each
324 component of the initial state was perturbed using an independent standard Gaussian
325 random variable times 5% of magnitude of the component. The forcing data were
326 perturbed in the manner described in Yilmaz et al. (2011). The synthetic experiments
327 were conducted from June 1, 2003 to October 1, 2003. The state variables for each
328 pixel were updated independently.

329

330 4.2 Validation statistics

331 4.2.1 Model error and bias

332 The model errors are defined as the difference between the actual values and the
333 model's predictions based on true initial values, and the bias is the average of the error
334 in the model during the relevant period. Let x_t denote the true values of the soil
335 moisture content at time t for a location and vertical soil layer. x_t^M denotes the model
336 predicted soil moisture from the true state at the previous time step $t-1$. The model's
337 bias and error variance for one step can be written as

$$338 \quad b_M = \frac{1}{a_{ts}} \sum_{t=1}^{a_{ts}} (x_t^M - x_t), \quad (17)$$

$$339 \quad v_M = \frac{1}{a_{ts}} \sum_{t=1}^{a_{ts}} (x_t^M - x_t)^2, \quad (18)$$

340 where a_{ts} is the number of time steps over which the observations made at 6:00 am
341 each day are assimilated.

342 4.2.2 Validation of analysis soil moisture

343 The true soil moisture content values from 7:00 am to 5:00 am next day are used
 344 to validate analysis states. For a location and vertical soil layer, let $x_{t,h}$ be the true
 345 soil moisture content at hour h on day t , and $x_{t,h}^f$ represent the forecasted soil
 346 moisture content at hour h from analysis state x_t^a at 6:00 am on day t . The analysis
 347 bias is defined as

$$348 \quad b_a = \frac{1}{23a_{ts}} \sum_{t=1}^{a_{ts}} \sum_{h=7}^{29} (x_{t,h}^f - x_{t,h}). \quad (19)$$

349 The analysis error variance is defined as

$$350 \quad v_a = \frac{1}{23a_{ts}} \sum_{t=1}^{a_{ts}} \sum_{h=7}^{29} (x_{t,h}^f - x_{t,h})^2 \quad (20)$$

$$= \frac{1}{23a_{ts}} \sum_{t=1}^{a_{ts}} \sum_{h=7}^{29} (x_{t,h}^f - x_{t,h} - b_a)^2 + b_a^2$$

351 (See Appendix B for the proof)

352 4.2.3 Water balance

353 Following Yilmaz (2011), the water budget imbalance at location is evaluated
 354 using the water balance residual,

$$355 \quad R = \frac{1}{Na_{ts}} \sum_{t=1}^{a_{ts}} \sum_{n=1}^N r_{n,t}. \quad (21)$$

356

357 5. Results

358 In the synthetic experiments, the magnitudes of the model's bias and error were
 359 calculated using Eqs (17) and (18), respectively, and are shown in Figure 3. It shows
 360 that the model's bias was almost negative from Figure 3a. The negative bias in the
 361 surface layer was the result of a combination of a lower surface roughness and a larger
 362 leaf area index in the CoLM; these values led to more soil evaporation and more
 363 canopy interception and could result in a smaller amount of water infiltrating the soil

364 than the amount modeled using the CLM 4.0. In the CoLM, the porosity of each layer
365 was less than it was in the CLM 4.0, which retained less water and contributed to the
366 negative bias of the upper 9 layers. However, the magnitude of the bias increased to 2%
367 in the bottom layer. The significant difference between the two models at the bottom
368 layer could be ascribed to their different boundary conditions. Interactions between
369 the soil moisture content and the ground water at the bottom of the soil column were
370 modeled in the CLM 4.0 (Oleson *et al.* 2010) but not in the CoLM. The error in each
371 model (Figure 3b) fluctuated in a manner similar to that of the model’s bias. Unbiased
372 observations are necessary for correcting bias in a model, which is not possible in
373 many realistic applications, especially in assimilating remote sensing retrievals. Since
374 satellite observations of the soil moisture content of deep layers are unavailable, only
375 removing the bias in shallow layers would introduce error in model dynamics.

376

377 5.1 Forecast error inflation and vertical localization

378 In the synthetic experiments, the study domain comprised 40 pixels. At each point
379 in the grid-scale threshold layer, the localization scale factor μ_s , was determined
380 independently. Therefore, totally 9 sets of experiments with different localization
381 scale factor (see Table 2) were conducted separately. Among these experiments, the
382 “optimal” case for each pixel was defined as the case in which the column averaged
383 analysis error (Eq. (20)) was minimized (shown in Figure 4). According to Figure 4a,
384 the corresponding threshold layer s of μ_s was generally between 5 and 6 in both
385 cases, which could be ascribed to the homogeneous soil texture and land cover. In the
386 WCEnKF-Inf-Loc, there were 19 pixels in which the threshold layers were “optimal,”
387 and the layers selected in the other pixels were suboptimal (most were roughly one
388 layer away from the “optimal” case). As shown in Figure 4b, the spatial average of the

389 root analysis error variance (Eq. (20)) of the WCEnKF-Inf-Loc (4.09%) was
390 comparable with the optimal value (3.84%) even though s was not selected on the
391 basis of minimizing the analysis error.

392 The spatial average of the root analysis error variance in each layer in the
393 schemes with (WCEnKF-Inf-Loc and WCEnKF-Inf) and without (WCEnKF)
394 inflation are displayed in Figure 5a. Above 62.0 cm, the analysis errors of the schemes
395 without inflation were substantially larger than those of the schemes with inflation for
396 the synthetic experiments. This suggested that inflation provided a better estimate in
397 the layers close to the observation. When no inflation was performed, the accuracy of
398 the soil moisture content was barely improved over that of the open-loop (not shown
399 here).

400 By comparing the schemes with (WCEnKF-Inf-Loc) and without (WCEnKF-Inf)
401 vertical localization, the impact of this approach on the assimilation accuracy in each
402 layer is shown in Figure 5a. Because the threshold layer of the localization function
403 ρ_s was layer 6 (36.6 cm) for 28 of the pixels (see Figure 4a), the spatial average of
404 root analysis error variance of the results of the WCEnKF-Inf-Loc is almost identical
405 to that of the results of the WCEnKF-Inf for depths above 36.6 cm. In contrast,
406 inflation increased the analysis error in the soil moisture content of the deep layers in
407 the WCEnKF-Inf. In this model, the sample error covariances of the moisture contents
408 of shallow and deep soil were inflated by a factor greater than 6 (the average inflation
409 factor was 6.25). This could lead to larger assimilation errors for deep soil moisture
410 profiles in the WCEnKF-Inf. Therefore, inflation should be used with vertical
411 localization to reduce the spurious covariance resulting from the covariance
412 inflation-based approach.

413 As it was in the synthetic experiments, vertical localization (WCEnKF-Inf-Loc)

414 was helpful in avoiding erroneous estimates of the soil moisture contents at lower
415 levels (in the WCEnKF-Inf). A comparison of the analysis error at a depth of 3 cm
416 (i.e., the depth of the assimilated observations was 3 cm) in the models with
417 (WCEnKF-Inf and WCEnKF-Inf-Loc) and without (WCEnKF) inflation showed that
418 the inflation technique significantly reduces the analysis error at the depth at which
419 observations are made.

420 To investigate the role of bias correction, the spatial averaged root analysis error
421 variance (Eq. (20)) of WCEnKF-Inf-Loc-BA and WCEnKF-Inf-Loc were compared.
422 According to Figure 5a, the spatial averaged root analysis error variances of the two
423 schemes were comparable with each other (2.12% for the WCEnKF-Inf-Loc-BA and
424 2.16% for the WCEnKF-Inf-Loc) in the layers that were shallower than 36.6 cm. This
425 could be due to that the observations are closer to the shallow layers and the vertical
426 localization approach is reasonable effective to reduced the bias. However, for the
427 layers that were deeper than 62.0 cm, the averaged root analysis error of the
428 WCEnKF-Inf-Loc-BA (6.05%) was less than that of the WCEnKF-Inf-Loc (6.59%).

429

430 5.2 The water budget constraint

431 In the synthetic experiment, the weak constraint on the water budget reduced the
432 water balance residual significantly in each pixel and the results are shown in Figure 6.
433 It shows that, the spatial average of the water balance residual of WCEnKF scheme
434 was 0.0487 mm, which was much smaller than that of the EnKF scheme (0.1389 mm).
435 Therefore, the assimilation scheme with water budget constraint can indeed reduce the
436 water balance residuals relative to the assimilation scheme without water budget
437 constraint which is consistent with the results of previous studies (Yilmaz *et al.* 2011;
438 Yilmaz *et al.* 2012). The interquartile range of the water balance residuals in the 40

439 pixels for the WCEnKF scheme was 0.0042 mm, which was less than half of that for
440 the EnKF scheme (0.0098 mm). The reduced spread of the water balance residuals
441 signals a more stable water balance budget with the water budget constraint.

442 The spatial average of the water balance residual for WCEnKF-Inf,
443 WCEnKF-Inf-Loc and WCEnKF-Inf-Loc-BA was 0.0834 mm, 0.0737 mm and
444 0.0723 mm, respectively. The corresponding interquartile range was 0.0079 mm,
445 0.0051 mm and 0.0072 mm, respectively. They are still much smaller than those for
446 the EnKF scheme, despite there are bit increase than those for WCEnKF. This
447 demonstrate the weak water budget constraint is still effective in reducing magnitude
448 and spread of the water inbalance, despite of more complicated assimilation
449 approaches were associated.

450

451 **6. Discussion**

452 6.1 Covariance inflation and vertical localization

453 In this study, the cost function used to estimate the state variables with the weak
454 water budget constraint (Eq. (8)) consists of three parts, which are related with
455 observations, model forecasts and water residual (Yilmaz *et al.* 2012). It is represented
456 as a summation of three scalars, no matter how many observations are assimilated.
457 Therefore, inflating of one scalar (e.g., model forecasts) seems to have the similar
458 impact as deflating another one (e.g., water residual), particularly the weights
459 associated in this problem can be shown as function of the ratio of these three scalars.
460 Specifically, inflation of forecast error covariance has somewhat similar impact with
461 deflation of the water balance residual covariance. Accordingly, it is plainly obvious
462 that the water balance residual of the scheme WCEnKF-Inf is larger than that of the
463 scheme WCEnKF. According to Figure 5a, the covariance inflation improved the

464 estimates of the soil moisture content in the shallow layers independently of whether
465 vertical localization was used. This is primarily because the observation operator, \mathbf{h} , is
466 the linear operator that was used to interpolate the soil moisture content at depths of
467 2.8 cm and 6.2 cm to a depth of 3 cm. Then, the likelihood function for the inflation
468 factor (Eq. (11)) depends only on the observations and predictions of the soil moisture
469 content in the 2nd and 3rd layers. The mean value of the inflation factor is 6.25 for
470 WCEnKF-Inf, indicating that the initial forecast spread is not large enough. This leads
471 to an improvement in the forecast error statistics in the shallow layers, and to further
472 improvements in the [assimilated](#) soil moisture contents of those layers.

473 However, the soil moisture contents of the deep layers are not directly related to
474 the inflation factor. Inflating the forecast errors in the deep layers leads to an
475 overestimation of the corresponding forecast error covariance, and could lead to larger
476 analysis errors in the deep layers (see WCEnKF-Inf in Figure 5a). Therefore in this
477 study, the vertical localization approach was developed to prevent soil moisture over
478 fitting for deep layers. Using all observations for [threshold](#) s is only for model
479 selection (from the 10 layers), not for fitting parameter. When vertical localization is
480 used, the soil moisture contents of the deep layers are not significantly updated.
481 Consequently, larger errors are avoided in the deep layers (see WCEnKF-Inf-Loc in
482 Figure 5a).

483 Comparing to traditional EnKF without inflation and localization, although
484 mainly the soil moisture contents of layers above the threshold layer (usually the 5th or
485 6th layer) were updated at each time step during the assimilation process when the
486 WCEnKF-Inf-Loc was used, Figure 5a shows that the soil moisture contents of the
487 layers below the threshold layer, especially the 6th and 7th layers, are also improved.
488 This may be because the model propagates changes in the shallow layers downward,

489 adjusting the soil moisture contents of the deep layers. Because the soil moisture
490 content of layers above the threshold layer was improved during the previous time
491 step, this process results in better predictions of the soil moisture contents of layers
492 below the threshold layer, and therefore, reduces the analysis error in layers below the
493 threshold layer.

494

495 6.2 Bias correction

496 Geophysical models are never perfect and usually produce estimates with biases
497 that vary in time and in space (Reichle 2008). Therefore, bias correction is important
498 for assimilating data into models. In this study, only soil moisture in shallow layers
499 can be observed (in order to mimic the satellite observation), so the bias for the soil
500 moisture in deeper layers can not be entirely removed only using the observations.
501 However, bias can be detected by monitoring statistics of observation-minus-forecast
502 residual in the assimilation systems. Therefore the bias-aware assimilation proposed by
503 Dee (2005) was further applied to reduce the bias of soil moisture in all layers.

504 For further evaluating the efficacy of the bias-aware assimilation scheme, the
505 analysis error variance was decomposed to a short-lived component (Figure 5b) and a
506 bias component (Figure 5c) for the synthetic experiment. It shows that for the
507 bias-blind data assimilation scheme (WCEnKF-Inf-Loc), both short-lived errors and
508 biases reduce in the layers close to observation, while maintain the similar levels as
509 those for EnKF for the deeper layers. The covariance inflation can play an important
510 role in bias reduction. Bias can only be seen during long assimilation period. At an
511 instant time, bias and error are mixed. For the traditional EnKF, the forecast error
512 covariance matrix obtained from the ensemble of their anomalies (Eq. (2)) mainly
513 represents short-lived error, so it has to be inflated to include error related to bias.

514 Moreover, the bias could be further reduced by the additional bias-aware assimilation.

515 There are other bias estimation approaches in data assimilation. For example,
516 treating bias as model variables and estimate in assimilation (De Lannoy *et al.* 2007;
517 Dee and Da Silva 1998), adjusting the state variable of the forecast model not only
518 their covariance matrix in each forecast step (Zhang *et al.* 2014; Zhang *et al.* 2015),
519 addressing the biases in the model and observations by rescaling their cumulative
520 distribution functions (Koster *et al.* 2009; Reichle and Koster 2004). The scheme
521 proposed here can provide a base line to validate the efficacy of these approaches and
522 could be further improved after these bias corrections.

523

524 6.3 Notes

525 The most computational cost in the assimilation system is on computing the
526 localization function at each model grid cell. For the synthetic experiments with
527 CoLM model and 40 grids, it takes about 24 hours running on the personal
528 workstation. For global data assimilation with 2° resolution it could take about 3
529 months. However, the super server and parallel computation can significantly shorten
530 the computational time. A regional scale using soil texture or climate regimes can also
531 be used to delineate different regions. By this way, the computational time of global
532 data assimilation can be further reduced.

533 In the near future, we plan to validate the major conclusions under different soil
534 conditions and land cover types. Vertical localization, which uses adjacent
535 observations, should also be tested in future work. More detailed analyses of the bias
536 correction for assimilating remote sensing retrievals should be performed. The
537 response of the analytic soil moisture content to weather predictions also needs to be

538 investigated. Completing these studies should improve the state of research into
539 land-atmosphere interactions.

540

541 **7. Conclusions**

542 In this study, observations of the soil moisture content at a depth of 3 cm were
543 assimilated using an ensemble Kalman filter with **several** improvements. Firstly, an
544 adaptive forecast error inflation based on maximum-likelihood estimation was
545 adopted to reduce the analysis error. This study supports the idea that the proper form
546 of the forecast error covariance matrix is crucial for reducing the analysis error near
547 the layers in which observations are made. Secondly, an adequate vertical localization
548 for the ensemble-based filter was proposed associated with the forecast error
549 covariance inflation, to avoid misestimates of the soil moisture contents of deep layers.
550 Lastly, a constraint on the water balance was used in this study to reduce the water
551 budget residual substantially without significantly changing the assimilation accuracy.
552 The experiment results of synthetic study show that the WCEnKF-Inf-Loc
553 assimilation scheme can reduce both the short-lived analysis error and the analysis
554 bias in the shallow layers, which also lead to a rational water budget residual. **The**
555 **bias-aware assimilation scheme is potentially useful to further reduce the analysis**
556 **error arising from model bias.**

557

558 **Data availability** The soil moisture **observations** are available at <http://www.ceop.net>.
559 The ERA-interim forcing data used in the synthetic experiments is obtained from
560 <https://apps.ecmwf.int/datasets>.

561

562 **Author Contributions** BD performed the simulations and assimilations. XZ designed
 563 the research. GW analyzed the results. TL collected and preprocessed the data. GW
 564 and XZ prepared the manuscript with contributions from all co-authors.

565

566 **Conflicts of Interest** The authors declare that they have no conflict of interest.

567

568 **Acknowledgement** This study was funded by the the National Key R&D Program of
 569 China (2019YFC1510002, 2015CB953703) and the National Natural Science
 570 Foundation of China (41705086). We would like to thank the Editor and two
 571 anonymous reviewers for their insightful comments in improving the manuscript. We
 572 also thank Drs. Yongjiu Dai and Qingyun Duan for their help in land surface model.

573

574 **Appendix A. A bias-aware assimilation scheme**

575 For correcting the bias of the analysis states $\mathbf{x}_{n,t}^a$ in Eq. (12), the bias-aware
 576 assimilation (Dee 2005) is applied.

577 Let \mathbf{b}_t is the forecast bias at time step t, and set $\mathbf{b}_1 = \mathbf{0}$. Then

$$578 \quad \mathbf{b}_t = \mathbf{b}_{t-1} - \gamma \tilde{\mathbf{P}}_{s,t} \mathbf{h}^T (\mathbf{h} \tilde{\mathbf{P}}_{s,t} \mathbf{h}^T + R_t)^{-1} (o_t - \mathbf{h}(\tilde{\mathbf{x}}_t^f - \mathbf{b}_{t-1})). \quad (\text{A1})$$

579 where $\tilde{\mathbf{x}}_t^f$ is the ensemble mean of the perturbed forecast states $\tilde{\mathbf{x}}_{n,t}^f$ predicted from
 580 the perturbed analysis state at previous time step $\tilde{\mathbf{x}}_{n,t-1}^a$, the forecast error covariance
 581 matrix is in the form

$$582 \quad \tilde{\mathbf{P}}_{s,t} = \begin{bmatrix} \sqrt{\tilde{\lambda}_t} \\ \tilde{\lambda}_t \end{bmatrix} [\boldsymbol{\rho}_s] \tilde{\mathbf{P}}_t [\boldsymbol{\rho}_s] \begin{bmatrix} \sqrt{\tilde{\lambda}_t} \\ \tilde{\lambda}_t \end{bmatrix}, \quad (\text{A2})$$

583 where the localization threshold s is adopted from the bias-blind scheme documented
 584 in section 3.2,

$$585 \quad \tilde{\mathbf{P}}_t = \frac{1}{N-1} \sum_{n=1}^N (\tilde{\mathbf{x}}_{n,t}^f - \tilde{\mathbf{x}}_t^f) (\tilde{\mathbf{x}}_{n,t}^f - \tilde{\mathbf{x}}_t^f)^T, \quad (\text{A3})$$

586 and the inflation factor $\tilde{\lambda}_t$ is estimated by minimizing

$$587 \quad -2\tilde{L}_{s,t}(\tilde{\lambda}_t) = \ln(\mathbf{h}\tilde{\mathbf{P}}_{s,t}\mathbf{h}^T + R_t) + (o_t - \mathbf{h}\tilde{\mathbf{x}}_t^f)^T (\mathbf{h}\tilde{\mathbf{P}}_{s,t}\mathbf{h}^T + R_t)^{-1} (o_t - \mathbf{h}\tilde{\mathbf{x}}_t^f). \quad (\text{A4})$$

588 The scalar parameter γ in Eq. (A1) that controls the magnitude of the forecast
 589 bias estimates, is derived by

$$590 \quad \gamma = \frac{\mu}{1-\mu} (R_t + \mathbf{h}\mathbf{P}_t\mathbf{h}^T) (\mathbf{h}\mathbf{P}_t\mathbf{h}^T)^{-1}, \quad (\text{A5})$$

591 where μ is estimated by minimizing the following objective function (Dee and
 592 Todling 2000)

$$593 \quad f(\mu) = \sum_n n^2 \left\{ \left[\frac{1-\mu}{1-(1-\mu)e^{-2\pi i\Delta t/n}} \right] \left[\sum_t (o_t - \mathbf{h}\tilde{\mathbf{x}}_t^f) e^{-2\pi i\Delta t/n} \right]^2 (R_t + \mathbf{h}\mathbf{P}_t\mathbf{h}^T)^{-1} - 1 \right\}^2 \quad (\text{A6})$$

594 Then the perturbed analysis states is calculated as

$$595 \quad \tilde{\mathbf{x}}_{n,t}^a = \tilde{\mathbf{x}}_{n,t}^f - \mathbf{b}_{t-1} + \tilde{\mathbf{P}}_t^a \mathbf{h}^T R_t^{-1} (o_t + \varepsilon_{n,t} - \mathbf{h}(\tilde{\mathbf{x}}_{n,t}^f - \mathbf{b}_{t-1})) \\ + \tilde{\mathbf{P}}_t^a \mathbf{c} \tilde{\varphi}_t^{-1} (\tilde{\beta}_{n,t} - \mathbf{c}^T (\tilde{\mathbf{x}}_{n,t}^f - \mathbf{b}_{t-1})) \quad (\text{A7})$$

596 where

$$597 \quad \tilde{\beta}_{n,t} = \mathbf{c}^T \tilde{\mathbf{x}}_{n,t-1}^a + P r_t - E v_{n,t}^f - R n_{n,t}^f, \quad (\text{A8})$$

$$598 \quad \tilde{\varphi}_t = \frac{1}{N-1} \sum_{n=1}^N \left(\tilde{\beta}_{n,t} - \frac{1}{N} \sum_{j=1}^N \tilde{\beta}_{j,t} \right) \times \left(\tilde{\beta}_{n,t} - \frac{1}{N} \sum_{j=1}^N \tilde{\beta}_{j,t} \right)^T \quad (\text{A9})$$

599 and

$$600 \quad \tilde{\mathbf{P}}_t^a = \left(\mathbf{h}^T \mathbf{R}_t^{-1} \mathbf{h} + \tilde{\mathbf{P}}_{s,t}^{-1} + \mathbf{c} \tilde{\varphi}_t^{-1} \mathbf{c}^T \right)^{-1}, \quad (\text{A10})$$

601

602 **Appendix B. Proof of Eq. (20)**

603 For a location and vertical soil layer, the analysis error variance in the synthetic
604 experiment is defined as

$$\begin{aligned}
 v_a &= \frac{1}{23a_{ts}} \sum_{t=1}^{a_{ts}} \sum_{h=7}^{29} (x_{t,h}^f - x_{t,h})^2 \\
 605 \quad &= \frac{1}{23a_{ts}} \sum_{t=1}^{a_{ts}} \sum_{h=7}^{29} (x_{t,h}^f - x_{t,h} - b_a + b_a)^2 \quad (\text{B1}) \\
 &= \frac{1}{23a_{ts}} \sum_{t=1}^{a_{ts}} \sum_{h=7}^{29} (x_{t,h}^f - x_{t,h} - b_a)^2 + b_a^2 + \frac{2b_a}{23a_{ts}} \sum_{t=1}^{a_{ts}} \sum_{h=7}^{29} (x_{t,h}^f - x_{t,h} - b_a)
 \end{aligned}$$

606 From the definition of analysis bias (Eq. (19)), the last term on the right hand side of
607 is zero, so Eq. (20) is proved.

608

609 **References**

- 610 Anderson, J.L. and Anderson, S.L., 1999. A Monte Carlo implementation of the
611 nonlinear filtering problem to produce ensemble assimilations and forecasts.
612 *Monthly Weather Review*, 127: 2741-2758.
- 613 Bartalis, Z., Wagner, W., Naeimi, V., Hasenauer, S., Scipal, K., Bonekamp, H., Figa, J.
614 and Anderson, C., 2007. Initial soil moisture retrievals from the METOP-A
615 Advanced Scatterometer (ASCAT). *Geophysical Research Letters*, 34(20).
- 616 Bauser, H.H., Berg, D., Klein, O. and Roth, K., 2018. Inflation method for ensemble
617 Kalman filter in soil hydrology. *Hydrology and Earth System Sciences*, 22(9):
618 4921-4934.
- 619 Bonan, G.B., 1996. Land surface model (LSM version 1.0) for ecological,
620 hydrological, and atmospheric studies: Technical description and users guide.
621 Technical note, National Center for Atmospheric Research, Boulder, CO
622 (United States). Climate and Global Dynamics Div.
- 623 Bosilovich, M.G. and Lawford, R., 2002. Coordinated enhanced observing period
624 (CEOP) international workshop. *Bulletin of the American Meteorological*
625 *Society*, 83(10): 1495-1499.
- 626 Chen, F., Crow, W.T. and Ryu, D., 2014. Dual Forcing and State Correction via Soil
627 Moisture Assimilation for Improved Rainfall-Runoff Modeling. *Journal of*
628 *Hydrometeorology*, 15(5): 1832-1848.
- 629 Constantinescu, E.M., Sandu, A., Chai, T. and Carmichael, G.R., 2007.
630 Ensemble-based chemical data assimilation I: general approach. *Quarterly*
631 *Journal of the Royal Meteorological Society*, 133: 1229-1243.
- 632 Crow, W.T., Chen, F., Reichle, R.H. and Liu, Q., 2017. L band microwave remote
633 sensing and land data assimilation improve the representation of prestorm soil

634 moisture conditions for hydrologic forecasting. *Geophysical Research Letters*,
635 44(11): 5495-5503.

636 Crow, W.T. and Loon, E.V., 2006. Impact of incorrect model error assumptions on the
637 sequential assimilation of remotely sensed surface soil moisture. *Journal of*
638 *Hydrometeorology*, 7: 421-432.

639 Crow, W.T. and Wood, E.F., 2003. The assimilation of remotely sensed soil brightness
640 temperature imagery into a land surface model using Ensemble Kalman
641 filtering: a case study based on ESTAR measurements during SGP97.
642 *Advances in Water Resources*, 26: 137-149.

643 Dai, Y., Zeng, X., Dickinson, R.E., Baker, I., Bonan, G.B., Bosilovich, M.G., Denning,
644 A.S., Dirmeyer, P.A., Houser, P.R., Niu, G., Oleson, K.W., Schlosser, C.A. and
645 Yang, Z.-L., 2003. The Common Land Model. *Bulletin of the American*
646 *Meteorological Society*, 84(8): 1013-1023.

647 De Lannoy, G.J.M., Reichle, R.H., Houser, P.R., Pauwels, V.R.N. and Verhoest,
648 N.E.C., 2007. Correcting for forecast bias in soil moisture assimilation with
649 the ensemble Kalman filter. *Water Resources Research*, 43(9): n/a-n/a.

650 Dee, D.P., 2005. Bias and data assimilation. *Quarterly Journal of the Royal*
651 *Meteorological Society*, 131: 3323-3343.

652 Dee, D.P. and Da Silva, A.M., 1998. Data assimilation in the presence of forecast bias.
653 *Quarterly Journal of the Royal Meteorological Society*, 124(545): 269-295.

654 Dee, D.P. and Da Silva, A.M., 1999. Maximum-likelihood estimation of forecast and
655 observation error covariance parameters. Part I: Methodology. *Monthly*
656 *Weather Review*, 127(8): 1822-1834.

657 Dee, D.P., Gaspari, G., Redder, C., Rukhovets, L. and Da Silva, A.M., 1999.
658 Maximum-likelihood estimation of forecast and observation error covariance

659 parameters. Part II: Applications. *Monthly weather review*, 127(8): 1835-1849.

660 Dee, D.P. and Todling, R., 2000. Data assimilation in the presence of forecast bias:
661 The GEOS moisture analysis. *Monthly Weather Review*, 128(9): 3268-3282.

662 Dee, D.P., Uppala, S.M., Simmons, A.J., Berrisford, P., Poli, P., Kobayashi, S., Andrae,
663 U., Balmaseda, M.A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A.C.M.,
664 van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M.,
665 Geer, A.J., Haimberger, L., Healy, S.B., Hersbach, H., Hólm, E.V., Isaksen, L.,
666 Kållberg, P., Köhler, M., Matricardi, M., McNally, A.P., Monge-Sanz, B.M.,
667 Morcrette, J.J., Park, B.K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut,
668 J.N. and Vitart, F., 2011. The ERA-Interim reanalysis: configuration and
669 performance of the data assimilation system. *Quarterly Journal of the Royal*
670 *Meteorological Society*, 137(656): 553-597.

671 Delworth, T.L. and Manabe, S., 1988. The influence of potential evaporation on the
672 variabilities of simulated soil wetness and climate. *Journal of Climate*, 1(5):
673 523-547.

674 Dickinson, R.E., Henderson-Sellers, A. and Kennedy, P.J., 1993. Biosphere
675 Atmosphere Transfer Scheme (BATS) Version 1e as Coupled to the NCAR
676 Community Climate Model.

677 Dorigo, W.A., Wagner, W., Hohensinn, R., Hahn, S., Paulik, C., Xaver, A., Gruber, A.,
678 Drusch, M., Mecklenburg, S., van Oevelen, P., Robock, A. and Jackson, T.,
679 2011. The International Soil Moisture Network: a data hosting facility for
680 global in situ soil moisture measurements. *Hydrology and Earth System*
681 *Sciences*, 15(5): 1675-1698.

682 Dumedah, G. and Walker, J.P., 2014. Evaluation of Model Parameter Convergence
683 when Using Data Assimilation for Soil Moisture Estimation. *Journal of*

684 *Hydrometeorology*, 15(1): 359-375.

685 El Gharamti, M., Raeder, K., Anderson, J. and Wang, X.G., 2019. Comparing
686 Adaptive Prior and Posterior Inflation for Ensemble Filters Using an
687 Atmospheric General Circulation Model. *Monthly Weather Review*, 147(7):
688 2535-2553.

689 Entekhabi, D., Njoku, E.G., O'Neill, P.E., Kellogg, K.H., Crow, W.T., Edelstein, W.N.,
690 Entin, J.K., Goodman, S.D., Jackson, T.J. and Johnson, J., 2010. The soil
691 moisture active passive (SMAP) mission. *Proceedings of the IEEE*, 98(5):
692 704-716.

693 Evensen, G., 1994. Sequential data assimilation with a nonlinear quasi-geostrophic
694 model using Monte Carlo methods to forecast error statistics. *Journal of*
695 *Geophysical Research*, 99: 10143-10162.

696 Gruber, A., Crow, W.T. and Dorigo, W.A., 2018. Assimilation of Spatially Sparse In
697 Situ Soil Moisture Networks into a Continuous Model Domain. *Water*
698 *Resources Research*, 54(2): 1353-1367.

699 GUSEV, Y. and Novak, V., 2007. Soil water–main water resources for terrestrial
700 ecosystems of the biosphere. *J. Hydrol. Hydromech*, 55(1): 3-15.

701 Han, E., Crow, W.T., Holmes, T. and Bolten, J., 2014. Benchmarking a Soil Moisture
702 Data Assimilation System for Agricultural Drought Monitoring. *Journal of*
703 *Hydrometeorology*, 15(3): 1117-1134.

704 Janjić, T., Nerger, L., Albertella, A., Schröter, J. and Skachko, S., 2011. On Domain
705 Localization in Ensemble-Based Kalman Filter Algorithms. *Monthly Weather*
706 *Review*, 139(7): 2046-2060.

707 Kerr, Y.H., Waldteufel, P., Wigneron, J.-P., Delwart, S., Cabot, F., Boutin, J.,
708 Escorihuela, M.-J., Font, J., Reul, N. and Gruhier, C., 2010. The SMOS

709 mission: New tool for monitoring key elements of the global water cycle.
710 *Proceedings of the IEEE*, 98(5): 666-687.

711 Koster, R.D., Guo, Z.C., Yang, R.Q., Dirmeyer, P.A., Mitchell, K. and Puma, M.J.,
712 2009. On the Nature of Soil Moisture in Land Surface Models. *Journal of*
713 *Climate*, 22(16): 4322-4335.

714 Kumar, S.V., Peters-Lidard, C.D., Mocko, D., Reichle, R., Liu, Y.Q., Arsenault, K.R.,
715 Xia, Y.L., Ek, M., Riggs, G., Livneh, B. and Cosh, M., 2014. Assimilation of
716 Remotely Sensed Soil Moisture and Snow Depth Retrievals for Drought
717 Estimation. *Journal of Hydrometeorology*, 15(6): 2446-2469.

718 Lawford, R., Stewart, R., Roads, J., Isemer, H., Manton, M., Marengo, J., Yasunari, T.,
719 Benedict, S., Koike, T. and Williams, S., 2004. Advancing global-and
720 continental-scale hydrometeorology: Contributions of GEWEX
721 hydrometeorology panel. *Bulletin of the American Meteorological Society*,
722 85(12): 1917-1930.

723 Lawrence, D.M., Oleson, K.W., Flanner, M.G., Thornton, P.E., Swenson, S.C.,
724 Lawrence, P.J., Zeng, X., Yang, Z.-L., Levis, S., Sakaguchi, K., Bonan, G.B.
725 and Slater, A.G., 2011. Parameterization improvements and functional and
726 structural advances in Version 4 of the Community Land Model. *Journal of*
727 *Advances in Modeling Earth Systems*, 3(3).

728 Li, B., Toll, D., Zhan, X. and Cosgrove, B., 2012. Improving estimated soil moisture
729 fields through assimilation of AMSR-E soil moisture retrievals with an
730 ensemble Kalman filter and a mass conservation constraint. *Hydrology and*
731 *Earth System Sciences*, 16(1): 105-119.

732 Liang, X., Zheng, X., Zhang, S., Wu, G., Dai, Y. and Li, Y., 2012. Maximum
733 likelihood estimation of inflation factors on error covariance matrices for

734 ensemble Kalman filter assimilation. *Quarterly Journal of the Royal*
735 *Meteorological Society*, 138: 263-273.

736 Loizu, J., Massari, C., Alvarez-Mozos, J., Tarpanelli, A., Brocca, L. and Casali, J.,
737 2018. On the assimilation set-up of ASCAT soil moisture data for improving
738 streamflow catchment simulation. *Advances in Water Resources*, 111: 86-104.

739 Lu, H., Koike, T., Yang, K., Hu, Z.Y., Xu, X.D., Rasmy, M., Kuria, D. and Tamagawa,
740 K., 2012. Improving land surface soil moisture and energy flux simulations
741 over the Tibetan plateau by the assimilation of the microwave remote sensing
742 data and the GCM output into a land surface model. *International Journal of*
743 *Applied Earth Observation and Geoinformation*, 17: 43-54.

744 Lu, H., Yang, K., Koike, T., Zhao, L. and Qin, J., 2015. An Improvement of the
745 Radiative Transfer Model Component of a Land Data Assimilation System and
746 Its Validation on Different Land Characteristics. *Remote Sensing*, 7(5):
747 6358-6379.

748 McColl, K.A., He, Q., Lu, H. and Entekhabi, D., 2019. Short-Term and Long-Term
749 Surface Soil Moisture Memory Time Scales Are Spatially Anticorrelated at
750 Global Scales. *Journal of Hydrometeorology*, 20(6): 1165-1182.

751 Miyoshi, T., 2011. The Gaussian approach to adaptive covariance inflation and its
752 implementation with the local ensemble transform Kalman filter. *Monthly*
753 *Weather Review*, 139: 1519-1534.

754 Miyoshi, T., Kalnay, E. and Li, H., 2012. Estimating and including observation-error
755 correlations in data assimilation. *Inverse Problems in Science & Engineering*,
756 32: 1-12.

757 Niu, G.-Y., Yang, Z.-L., Dickinson, R.E., Gulden, L.E. and Su, H., 2007.
758 Development of a simple groundwater model for use in climate models and

759 evaluation with Gravity Recovery and Climate Experiment data. *Journal of*
760 *Geophysical Research*, 112(D7).

761 Niu, G.Y., Yang, Z.L., Dickinson, R.E. and Gulden, L.E., 2005. A simple
762 TOPMODEL - based runoff parameterization (SIMTOP) for use in global
763 climate models. *Journal of Geophysical Research: Atmospheres (1984–2012)*,
764 110(D21).

765 Njoku, E.G., Jackson, T.J., Lakshmi, V., Chan, T.K. and Nghiem, S.V., 2003. Soil
766 moisture retrieval from AMSR-E. *Geoscience and Remote Sensing, IEEE*
767 *Transactions on*, 41(2): 215-229.

768 Oleson, K.W., Lawrence, D.M., Gordon, B., Flanner, M.G., Kluzek, E., Peter, J.,
769 Levis, S., Swenson, S.C., Thornton, E. and Feddema, J., 2010. Technical
770 description of version 4.0 of the Community Land Model (CLM).

771 Pan, M. and Wood, E.F., 2006. Data assimilation for estimating the terrestrial water
772 budget using a constrained ensemble Kalman filter. *Journal of*
773 *Hydrometeorology*, 7(3): 534-547.

774 Pielke, R.A., 2001. Influence of the spatial distribution of vegetation and soils on the
775 prediction of cumulus Convective rainfall. *Reviews of Geophysics*, 39(2):
776 151-177.

777 Pinnington, E., Quaife, T. and Black, E., 2018. Impact of remotely sensed soil
778 moisture and precipitation on soil moisture prediction in a data assimilation
779 system with the JULES land surface model. *Hydrology and Earth System*
780 *Sciences*, 22(4): 2575-2588.

781 Raanes, P.N., Bocquet, M. and Carrassi, A., 2019. Adaptive covariance inflation in the
782 ensemble Kalman filter by Gaussian scale mixtures. *Quarterly Journal of the*
783 *Royal Meteorological Society*, 145(718): 53-75.

784 Reichle, R.H., 2008. Data assimilation methods in the Earth sciences. *Advances in*
785 *Water Resources*, 31: 1411-1418.

786 Reichle, R.H. and Koster, R.D., 2004. Bias reduction in short records of satellite soil
787 moisture. *Geophysical Research Letters*, 31(L19501).

788 Reichle, R.H. and Koster, R.D., 2005. Global assimilation of satellite surface soil
789 moisture retrievals into the NASA Catchment land surface model. *Geophysical*
790 *Research Letters*, 32.

791 Robock, A., Vinnikov, K.Y., Srinivasan, G., Entin, J.K., Hollinger, S.E., Speranskaya,
792 N.A., Liu, S. and Namkhai, A., 2000. The global soil moisture data bank.
793 *Bulletin of the American Meteorological Society*, 81(6): 1281-1299.

794 Santanello, J.A., Kumar, S.V., Peters-Lidard, C.D. and Lawston, P.M., 2016. Impact of
795 Soil Moisture Assimilation on Land Surface Model Spinup and Coupled
796 Land-Atmosphere Prediction. *Journal of Hydrometeorology*, 17(2): 517-540.

797 Wang, X. and Bishop, C.H., 2003. A comparison of breeding and ensemble transform
798 kalman filter ensemble forecast schemes. *Journal of the Atmospheric Sciences*,
799 60: 1140-1158.

800 Wei, J., Dirmeyer, P.A., Guo, Z., Zhang, L. and Misra, V., 2010. How Much Do
801 Different Land Models Matter for Climate Simulation? Part I: Climatology
802 and Variability. *Journal of Climate*, 23(11): 3120-3134.

803 Wu, G., Zheng, X., Wang, L., Zhang, S., Liang, X. and Li, Y., 2013. A New Structure
804 for Error Covariance Matrices and Their Adaptive Estimation in EnKF
805 Assimilation. *Quarterly Journal of the Royal Meteorological Society*, 139:
806 795-804.

807 Yang, K., Koike, T., Kaihotsu, I. and Qin, J., 2009. Validation of a dual-pass
808 microwave land data assimilation system for estimating surface soil moisture

809 in semiarid regions. *Journal of Hydrometeorology*, 10: 780-793.

810 Yang, K., Zhu, L., Chen, Y., Zhao, L., Qin, J., Lu, H., Tang, W., Han, M., Ding, B. and
811 Fang, N., 2016. Land surface model calibration through microwave data
812 assimilation for improving soil moisture simulations. *Journal of Hydrology*,
813 533: 266-276.

814 Yang, S.-C., Kalnay, E. and Enomoto, T., 2015. Ensemble singular vectors and their
815 use as additive inflation in EnKF. *Tellus A*, 67.

816 Yilmaz, M.T., DelSole, T. and Houser, P.R., 2011. Improving Land Data Assimilation
817 Performance with a Water Budget Constraint. *Journal of Hydrometeorology*,
818 12(5): 1040-1055.

819 Yilmaz, M.T., DelSole, T. and Houser, P.R., 2012. Reducing Water Imbalance in Land
820 Data Assimilation: Ensemble Filtering without Perturbed Observations.
821 *Journal of Hydrometeorology*, 13(1): 413-420.

822 Zhang, S., Yi, X., Zheng, X., Chen, Z., Dan, B. and Zhang, X., 2014. Global carbon
823 assimilation system using a local ensemble Kalman filter with multiple
824 ecosystem models. *Journal of Geophysical Research-Biogeosciences*, 119(11):
825 2171-2187.

826 Zhang, S., Zheng, X., Chen, J., Chen, Z., Dan, B., Yi, X., Wang, L. and Wu, G., 2015.
827 A global carbon assimilation system using a modified ensemble Kalman filter.
828 *Geoscientific Model Development*, 8: 805-816.

829 Zhao, L. and Yang, Z.L., 2018. Multi-sensor land data assimilation: Toward a robust
830 global soil moisture and snow estimation. *Remote Sensing of Environment*,
831 216: 13-27.

832 Zheng, X., 2009. An adaptive estimation of forecast error covariance parameters for
833 Kalman filtering data assimilation. *Advances in Atmospheric Sciences*, 26(1):

834 154-160.

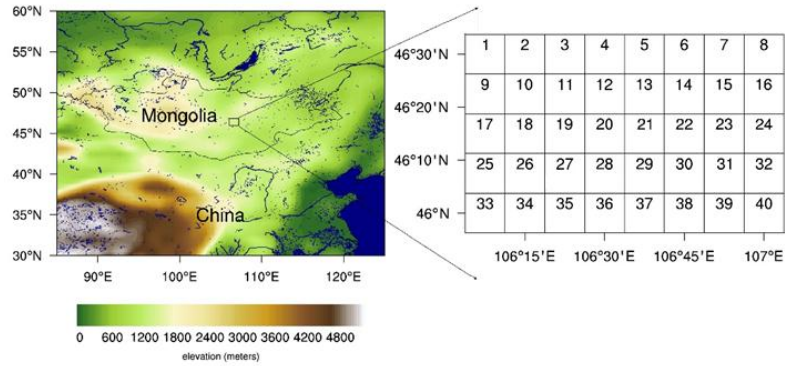
835

836

837 **Figure captions**

838

839



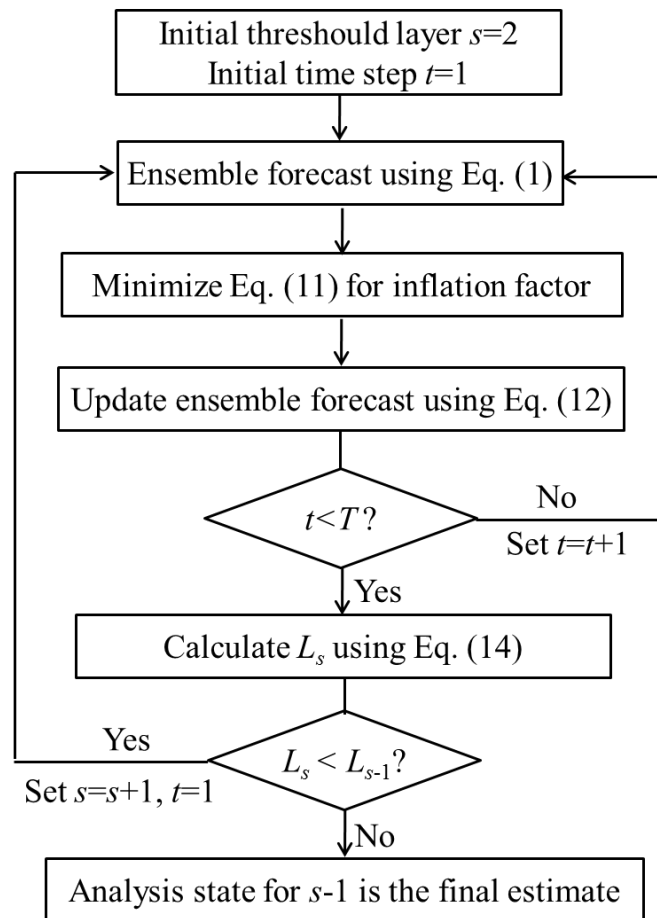
840 Figure 1. The topography and river distribution (left plot) and the geographical

841 location of the synthetic study area (right plot).

842

843

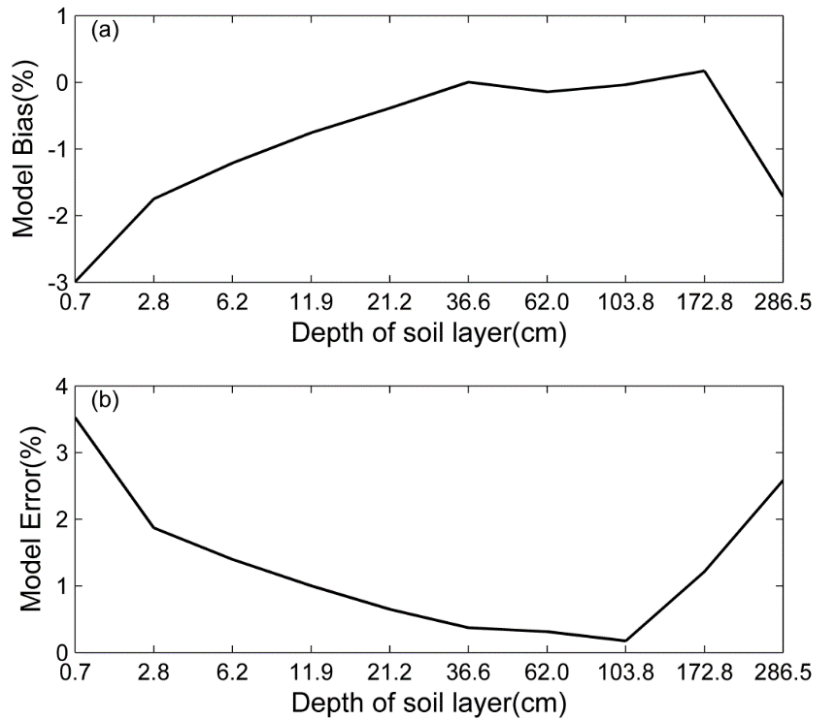
844



845 Figure 2. The assimilation procedure and localization scale factor estimation in the

846 experiments. All of the equations are in accordance with that described in the text.

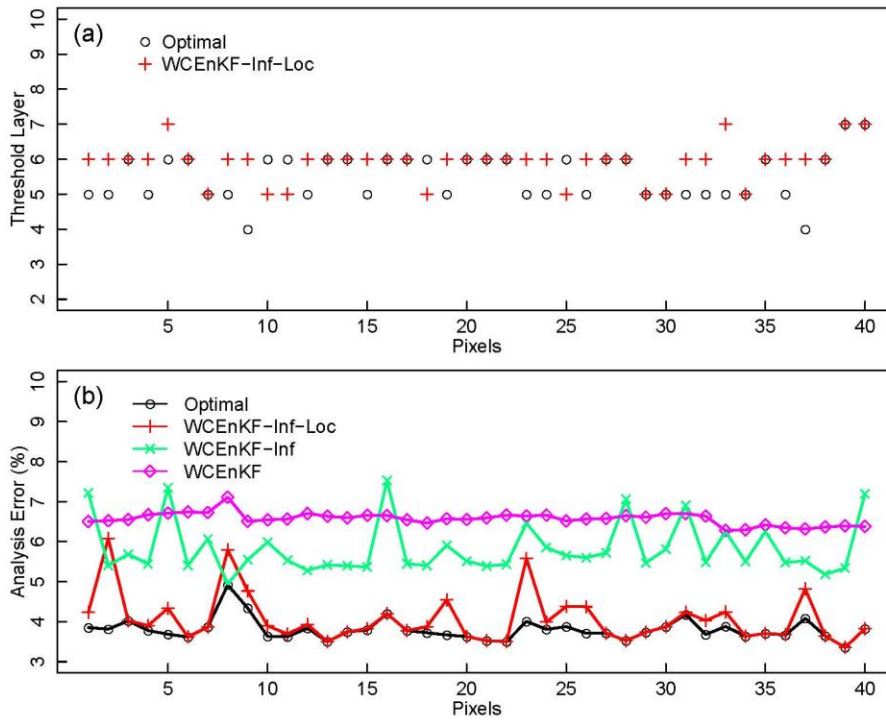
847



848

849 Figure 3. The areal average of the model's bias (a) and error (b) for one step in the soil
 850 moisture content between the CoLM and the CLM 4.0. The horizontal axis represents
 851 the layer depth.

852



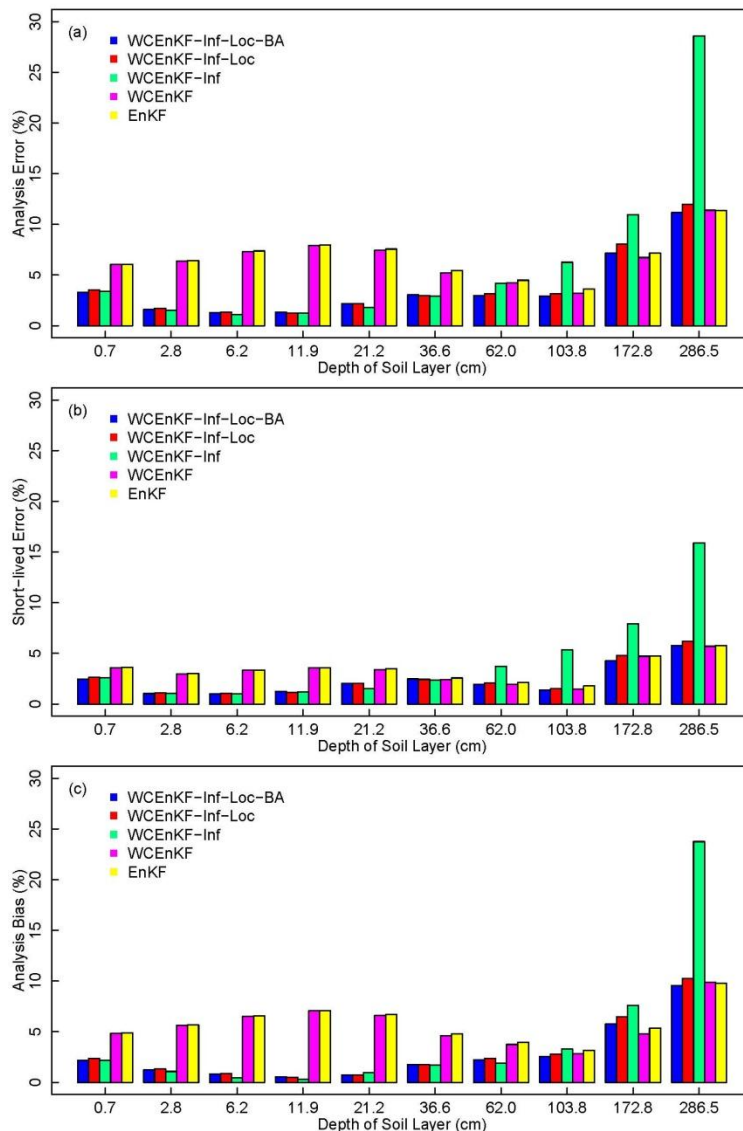
853

854 Figure 4. The threshold layers and analysis error for each pixel in the synthetic
 855 experiment. Graph (a) illustrates the optimal and WCEnKF-Inf-Loc threshold layers
 856 of each pixel. Graph (b) shows the column RSME of each pixel in different schemes
 857 with water balance constraint (Optimal, WCEnKF-Inf-Loc, WCEnKF-Inf and
 858 WCEnKF). The horizontal axes of (a) and (b) represent the 40 pixels in the study
 859 domain.

860

861

862



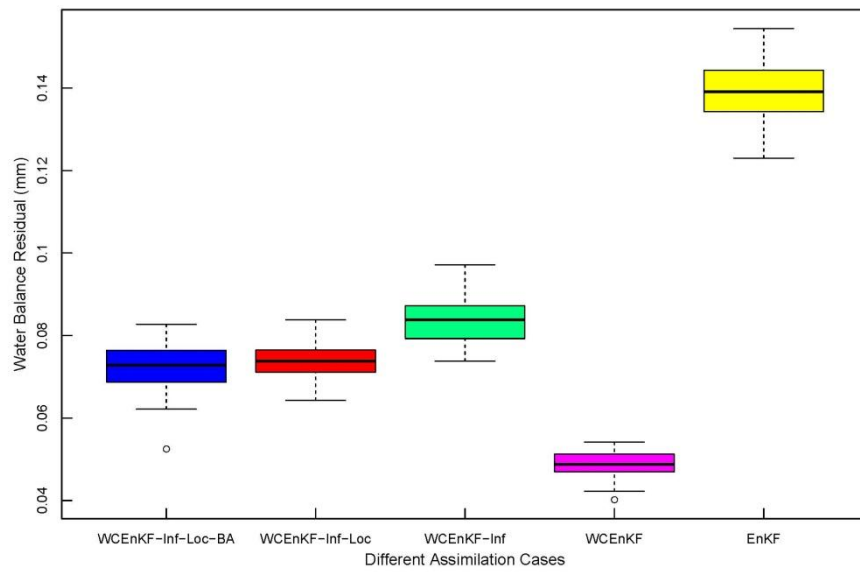
863

864 Figure 5. The assimilation results in each layer for the five schemes: a weakly
865 constrained bias-aware ensemble Kalman filter with forecast error inflation and
866 vertical localization (WCEnKF-Inf-Loc-BA), a weakly constrained ensemble Kalman
867 filter with forecast error inflation and vertical localization (WCEnKF-Inf-Loc), a
868 weakly constrained ensemble Kalman filter with forecast error inflation
869 (WCEnKF-Inf), a weakly constrained ensemble Kalman filter (WCEnKF), and the
870 traditional assimilation (EnKF). Graphic (a) is for spatial averaged analysis error of
871 the soil moisture content, (b) is for the short-lived error and (c) is for the analysis bias.

872

873

874



875 Figure 6. The box plot of the water balance residual in all 40 pixels for the
876 [WCEnKF-Inf-Loc-BA](#), WCEnKF-Inf-Loc, WCEnKF-Inf, WCEnKF and EnKF
877 assimilation schemes.

878

879 Table 1. The node depths (cm) of the 10 soil layers in the CoLM model.

880

Layer	1	2	3	4	5	6	7	8	9	10
Depth (cm)	0.7	2.8	6.2	11.9	21.2	36.6	62.0	103.8	172.8	286.5

881

882

883

884 Table 2. Estimated localization scale factor for different cases.

Layer	2	3	4	5	6	7	8	9	10
μ_s	0.2824	0.1256	0.0587	0.0300	0.0163	0.0093	0.0053	0.0025	0.0001

885