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

Bo Dan[1], Xiaogu Zheng[2], Guocan Wu[3*], and Tao Li[4]

[1] National Marine Data and Information Service, Tianjin, China

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

[3]College of Global Change and Earth System Science, Beijing Normal University, Beijing, China

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

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

Abstract

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

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

Keywords soil moisture, water balance, data assimilation, forecast error inflation, vertical localization

1. Introduction

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

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

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.

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

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

A major objective of soil moisture data assimilation is to address biases in models and observations (Koster *et al.* 2009; Reichle and Koster 2004). In this study, we only assume that models could be biased, while the soil moisture observations are assumed to be unbiased. Moreover, the soil moisture observations are restricted in shallow layer, so there is no observation available to directly correct the modeled soil moisture biases in deep layers. 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. Furthermore, the analysis error is decomposed to a short-lived error (random error) and a bias (system error). It demonstrates that the proposed scheme can reduce the both for soil moisture in shallow layers. These improvements steps can also result in a resonable estimates of the soil moisture content in the deep layers.

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

To overcome this shortcoming, Yilmaz et al. (2011) proposed using a weakly constrained ensemble Kalman filter (WCEnKF) to reduce the imbalance in the water budget. In a synthetic study, they concluded that the accuracy of a WCEnKF-based analysis is close to that of an EnKF-based analysis but the water budget balance residuals are much smaller than that of an unconstrained filter. Nevertheless, the observations of the soil moisture content cover the entire column, and a perfect model was used in their studies. This is not generally true, especially when only satellite observations are assimilated. In this study, the experiments were further designed to assimilate surface observations into an imperfect land model.

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

2. Models and data

2.1 Study area

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

The soil moisture content and related meteorological and hydrological parameters are monitored by automatic stations maintained by the Coordinated Enhanced Observing Period Asian Monsoon Project (CEOP AP) (Bosilovich and Lawford 2002; Lawford *et al.* 2004). The CEOP AP was launched by the World Climate Research Programme (WCRP) to develop an integrated global dataset that can be used to address issues relating to water and energy budget simulations and predictions, monsoon processes and the prediction of river flows. More details can be found at http://www.ceop.net.

2.2 Forcing data

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

2.3 Models

The Common Land Model (CoLM) developed by Dai et al. (2003) is a third-generation land surface model. It combines the best features of three successful models: the Land Surface Model (LSM, (Bonan 1996)), the Biosphere-Atmosphere Transfer Scheme (BATS, (Dickinson *et al.* 1993)) and the 1994 version of the Chinese Academy of Sciences/Institute of Atmospheric Physics model (IAP94, (Dai *et al.* 2003)), and is being further developed. The primary characteristics of the model include 10 unevenly spaced soil layers (see Table 1), one vegetation layer, 5 snow layers (depending on the snow depth), explicit treatment of the mass of liquid water, ice and phase changes within the system of the snow and soil, runoff parameterization following the TOPMODEL concept, a tiled treatment of the sub-grid fraction of the energy and water budget balance (Dai *et al.* 2003) and a canopy photosynthesis-conductance mode that describes the simultaneous transfer of CO_2 and water vapor into and out of the vegetation. The model parameters include data on the global terrain, elevation, land use, vegetation, land-water mask and hybrid

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

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

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

3. Methods

3.1 Forecast and observation systems

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

$$\mathbf{y}_{n,t}^{f} = \mathrm{M}_{n,t-1}\left(\mathbf{y}_{n,t-1}^{a}\right), \tag{1}$$

where t=1, ..., T is the time index, n=1, ..., N represents an ensemble member (in this study, the ensemble size is set to 100), $\mathrm{M}_{n,t-1}$ is a CoLM forced by the *n*-th perturbed atmospheric forcing, and **y** is a state vector containing 126 variables. The superscript "*f*" and "*a*" specify the forecast and analysis, respectively.

Let **x** be the state variables related to the water budget, that comprises of $\mathbf{SM}$ and $\mathbf{SIC}$ (the soil moisture content and the soil ice content in % at the 10 vertical levels listed in Table 1), CWC and SWE (the canopy's water content and the snow water equivalent in kg/m^2). In this study, only **x** is updated by data assimilation, while the model propagates changes to the other variables over time.

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

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

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 mean of $\mathbf{x}_{n,t}^{f}$. To avoid overestimation of the co-variability between shallow observations and soil moistures deeper than a threshold layer *s* (see section 3.2 for the estimation of *s*), the following vertical localization function with weighting of observations $\boldsymbol{\rho}_s$ (Janjić *et al.* 2011) will be applied on $\mathbf{P}_t$, i.e.,

$$\boldsymbol{\rho}_s(l) = \exp\left(-\mu_s\left|d_l - d_o\right|\right) \tag{3}$$

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

$$M(\mu) = \sum_{l \le s} \left[\exp\left(-\mu \left| d_l - d_o \right|\right) - 1 \right]^2 + \sum_{l > s} \left[\exp\left(-\mu \left| d_l - d_o \right|\right) \right]^2 \tag{4}$$

The estimated μ_s is listed in Table 2.

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

$$o_t = \mathbf{h}\mathbf{x}_t + \varepsilon_t , \tag{5}$$

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

3.2 Assimilation with water budget constraint

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

$$r_{n,t} \equiv \beta_{n,t} - \mathbf{c}^{\mathrm{T}} \mathbf{x}_{n,t}^{a} , \tag{6}$$

where

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

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

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

$$\begin{aligned} J_{n,t}(\mathbf{x}) = &\left(o_t - \mathbf{hx}\right)^{\mathrm{T}} R_t^{-1}\left(o_t - \mathbf{hx}\right) + \left(\mathbf{x} - \mathbf{x}_{n,t}^f\right)^{\mathrm{T}} \mathbf{P}_{s,t}^{-1}\left(\mathbf{x} - \mathbf{x}_{n,t}^f\right) \\ &+ \left(\beta_{n,t} - \mathbf{c}^{\mathrm{T}}\mathbf{x}\right)^{\mathrm{T}} \varphi_t^{-1}\left(\beta_{n,t} - \mathbf{c}^{\mathrm{T}}\mathbf{x}\right) \end{aligned}, \quad (8)$$

where

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

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

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

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

$$-2L_{s,t}(\lambda_t) = \ln\left(\mathbf{hP}_{s,t}\mathbf{h}^{\mathrm{T}} + R_t\right) + \left(o_t - \mathbf{hx}_t^f\right)^{\mathrm{T}}\left(\mathbf{hP}_{s,t}\mathbf{h}^{\mathrm{T}} + R_t\right)^{-1}\left(o_t - \mathbf{hx}_t^f\right). \quad (11)$$

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

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

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

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

its analysis error covariance matrix.

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

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

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

3.3 Bias-aware assimilation

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

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

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

where the scalar parameter γ that controls the magnitude of the forecast bias is estimated following Dee and Todling (2000) (see Eqs (A5)-(A6) of Appendix A), $\tilde{\mathbf{x}}_t^{f}$ is the ensemble mean of the perturbed forecast states $\tilde{\mathbf{x}}_{n,t}^{f}$ from the analysis state $\tilde{\mathbf{x}}_{n,t-1}^{a}$, $\tilde{\mathbf{P}}_{s,t}$ is the corresponding adjusted forecast error covariance (see Eq. (A2) of

Appendix A).

Then the perturbed assimilated states are

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

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.

4. Synthetic experiments

4.1 Experimental design

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

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

4.2 Validation statistics

4.2.1 Model error and bias

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

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

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

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

4.2.2 Validation of analysis soil moisture

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

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

The analysis error variance is defined as

$$\begin{aligned} v_a &= \frac{1}{23a_{ts}} \sum_{t=1}^{a_{ts}} \sum_{h=7}^{29} \left(x_{t,h}^f - x_{t,h} \right)^2 \\ &= \frac{1}{23a_{ts}} \sum_{t=1}^{a_{ts}} \sum_{h=7}^{29} \left(x_{t,h}^f - x_{t,h} - b_a \right)^2 + b_a^2 \end{aligned}. \tag{20}$$

(See Appendix B for the proof)

4.2.3 Water balance

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

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

5. Results

In the synthetic experiments, the magnitudes of the model's bias and error were calculated using Eqs (17) and (18), respectively, and are shown in Figure 3. It shows that the model's bias was almost negative from Figure 3a. The negative bias in the surface layer was the result of a combination of a lower surface roughness and a larger leaf area index in the CoLM; these values led to more soil evaporation and more canopy interception and could result in a smaller amount of water infiltrating the soil than the amount modeled using the CLM 4.0. In the CoLM, the porosity of each layer was less than it was in the CLM 4.0, which retained less water and contributed to the negative bias of the upper 9 layers. However, the magnitude of the bias increased to 2% in the bottom layer. The significant difference between the two models at the bottom layer could be ascribed to their different boundary conditions. Interactions between the soil moisture content and the ground water at the bottom of the soil column were modeled in the CLM 4.0 (Oleson *et al.* 2010) but not in the CoLM. The error in each model (Figure 3b) fluctuated in a manner similar to that of the model's bias. Unbiased observations are necessary for correcting bias in a model, which is not possible in many realistic applications, especially in assimilating remote sensing retrievals. Since satellite observations of the soil moisture content of deep layers are unavailable, only removing the bias in shallow layers would introduce error in model dynamics.

5.1 Forecast error inflation and vertical localization

In the synthetic experiments, the study domain comprised 40 pixels. At each point in the grid-scale threshold layer, the localization scale factor μ_s, was determined independently. Therefore, totally 9 sets of experiments with different localization scale factor (see Table 2) were conducted separately. Among these experiments, the "optimal" case for each pixel was defined as the case in which the column averaged analysis error (Eq. (20)) was minimized (shown in Figure 4). According to Figure 4a, the corresponding threshold layer s of μ_s was generally between 5 and 6 in both cases, which could be ascribed to the homogeneous soil texture and land cover. In the WCEnKF-Inf-Loc, there were 19 pixels in which the threshold layers were "optimal," and the layers selected in the other pixels were suboptimal (most were roughly one layer away from the "optimal" case). As shown in Figure 4b, the spatial average of the root analysis error variance (Eq. (20)) of the WCEnKF-Inf-Loc (4.09%) was comparable with the optimal value (3.84%) even though s was not selected on the basis of minimizing the analysis error.

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

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

As it was in the synthetic experiments, vertical localization (WCEnKF-Inf-Loc) was helpful in avoiding erroneous estimates of the soil moisture contents at lower levels (in the WCEnKF-Inf). A comparison of the analysis error at a depth of 3 cm (i.e., the depth of the assimilated observations was 3 cm) in the models with (WCEnKF-Inf and WCEnKF-Inf-Loc) and without (WCEnKF) inflation showed that the inflation technique significantly reduces the analysis error at the depth at which observations are made.

To investigate the role of bias correction, the spatial averaged root analysis error variance (Eq. (20)) of WCEnKF-Inf-Loc-BA and WCEnKF-Inf-Loc were compared. According to Figure 5a, the spatial averaged root analysis error variances of the two schemes were comparable with each other (2.12% for the WCEnKF-Inf-Loc-BA and 2.16% for the WCEnKF-Inf-Loc) in the layers that were 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 reasanable effective to reduced the bias. However, for the layers that were 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%).

5.2 The water budget constraint

In the synthetic experiment, the weak constraint on the water budget reduced the water balance residual significantly in each pixel and the results are shown in Figure 6. It shows that, the spatial average of the water balance residual of WCEnKF scheme was 0.0487 mm, which was much smaller than that of the EnKF scheme (0.1389 mm). Therefore, the assimilation scheme with water budget constraint can indeed reduce the water balance residuals relative to the assimilation scheme without water budget constraint which is consistent with the results of previous studies (Yilmaz *et al.* 2011; Yilmaz *et al.* 2012). The interquartile range of the water balance residuals in the 40 pixels for the WCEnKF scheme was 0.0042 mm, which was less than half of that for the EnKF scheme (0.0098 mm). The reduced spread of the water balance residuals signals a more stable water balance budget with the water budget constraint.

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

6. Discussion

6.1 Covariance inflation and vertical localization

In this study, the cost function used to estimate the state variables with the weak water budget constraint (Eq. (8)) consists of three parts, which are related with observations, model forecasts and water residual (Yilmaz *et al.* 2012). It is represented as a summation of three scalars, no matter how many observations are assimilated. Therefore, inflating of one scalar (e.g., model forecasts) seems to have the similar impact as deflating another one (e.g., water residual), particularly the weights associated in this problem can be shown as function of the ratio of these three scalars. Specifically, inflation of forecast error covariance has somewhat similar impact with deflation of the water balance residual covariance. Accordingly, it is plainly obvious that the water balance residual of the scheme WCEnKF-Inf is larger than that of the scheme WCEnKF. According to Figure 5a, the covariance inflation improved the estimates of the soil moisture content in the shallow layers independently of whether vertical localization was used. This is primarily because the observation operator, **h**, is the linear operator that was used to interpolate the soil moisture content at depths of 2.8 cm and 6.2 cm to a depth of 3 cm. Then, the likelihood function for the inflation factor (Eq. (11)) depends only on the observations and predictions of the soil moisture content in the 2^{nd} and 3^{rd} layers. The mean value of the inflation factor is 6.25 for WCEnKF-Inf, indicating that the initial forecast spread is not large enough. This leads to an improvement in the forecast error statistics in the shallow layers, and to further improvements in the assimilated soil moisture contents of those layers.

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

Comparing to traditional EnKF without inflation and localization, although mainly the soil moisture contents of layers above the threshold layer (usually the 5^{th} or 6^{th} layer) were updated at each time step during the assimilation process when the WCEnKF-Inf-Loc was used, Figure 5a shows that the soil moisture contents of the layers below the threshold layer, especially the 6^{th} and 7^{th} layers, are also improved. This may be because the model propagates changes in the shallow layers downward, adjusting the soil moisture contents of the deep layers. Because the soil moisture content of layers above the threshold layer was improved during the previous time step, this process results in better predictions of the soil moisture contents of layers below the threshold layer, and therefore, reduces the analysis error in layers below the threshold layer.

6.2 Bias correction

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

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

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

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

6.3 Notes

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^{o} resolution it could take about 3 months. However, the super server and parallel computation can significantly shorten the computational time. A regional scale using soil texture or climate regimes can also be used to delineate different regions. By this way, the computational time of global data assimilation can be further reduced.

In the near future, we plan to validate the major conclusions under different soil conditions and land cover types. Vertical localization, which uses adjacent observations, should also be tested in future work. More detailed analyses of the bias correction for assimilating remote sensing retrievals should be performed. The response of the analytic soil moisture content to weather predictions also needs to be investigated. Completing these studies should improve the state of research into land-atmosphere interactions.

7. Conclusions

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

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

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

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

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

Appendix A. A bias-aware assimilation scheme

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

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

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

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

$$\tilde{\mathbf{P}}_{s,t} = \left[\sqrt{\tilde{\boldsymbol{\lambda}}_t}\right]\left[\boldsymbol{\rho}_s\right]\tilde{\mathbf{P}}_t\left[\boldsymbol{\rho}_s\right]\left[\sqrt{\tilde{\boldsymbol{\lambda}}_t}\right], \tag{A2}$$

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

$$\tilde{\mathbf{P}}_t = \frac{1}{N-1}\sum_{n=1}^{N}\left(\tilde{\mathbf{x}}_{n,t}^f - \tilde{\mathbf{x}}_t^f\right)\left(\tilde{\mathbf{x}}_{n,t}^f - \tilde{\mathbf{x}}_t^f\right)^{\mathrm{T}}, \tag{A3}$$

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

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

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

$$\gamma = \frac{\mu}{1-\mu}\left(R_t + \mathbf{h}\mathbf{P}_t\mathbf{h}^{\mathrm{T}}\right)\left(\mathbf{h}\mathbf{P}_t\mathbf{h}^{\mathrm{T}}\right)^{-1}, \tag{A5}$$

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

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

Then the perturbed analysis states is calculated as

$$\begin{aligned}\tilde{\mathbf{x}}_{n,t}^a &= \tilde{\mathbf{x}}_{n,t}^f - \mathbf{b}_{t-1} + \tilde{\mathbf{P}}_t^a\mathbf{h}^{\mathrm{T}}R_t^{-1}\left(o_t + \varepsilon_{n,t} - \mathbf{h}\left(\tilde{\mathbf{x}}_{n,t}^f - \mathbf{b}_{t-1}\right)\right) \\ &\quad + \tilde{\mathbf{P}}_t^a\mathbf{c}\tilde{\varphi}_t^{-1}\left(\tilde{\beta}_{n,t} - \mathbf{c}^{\mathrm{T}}\left(\tilde{\mathbf{x}}_{n,t}^f - \mathbf{b}_{t-1}\right)\right)\end{aligned}. \tag{A7}$$

where

$$\tilde{\beta}_{n,t} = \mathbf{c}^{\mathrm{T}}\tilde{\mathbf{x}}_{n,t-1}^a + Pr_t - Ev_{n,t}^f - Rn_{n,t}^f, \tag{A8}$$

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

and

$$\tilde{\mathbf{P}}_t^a = \left(\mathbf{h}^{\mathrm{T}} R_t^{-1}\mathbf{h} + \tilde{\mathbf{P}}_{s,t}^{-1} + \mathbf{c}\tilde{\varphi}_t^{-1}\mathbf{c}^{\mathrm{T}}\right)^{-1}, \tag{A10}$$

Appendix B. Proof of Eq. (20)

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

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

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

References

Anderson, J.L. and Anderson, S.L., 1999. A Monte Carlo implementation of the nonlinear fltering problem to produce ensemble assimilations and forecasts. *Monthly Weather Review*, 127: 2741-2758.

Bartalis, Z., Wagner, W., Naeimi, V., Hasenauer, S., Scipal, K., Bonekamp, H., Figa, J. and Anderson, C., 2007. Initial soil moisture retrievals from the METOP-A Advanced Scatterometer (ASCAT). *Geophysical Research Letters*, 34(20).

Bauser, H.H., Berg, D., Klein, O. and Roth, K., 2018. Inflation method for ensemble Kalman filter in soil hydrology. *Hydrology and Earth System Sciences*, 22(9): 4921-4934.

Bonan, G.B., 1996. Land surface model (LSM version 1.0) for ecological, hydrological, and atmospheric studies: Technical description and users guide. Technical note, National Center for Atmospheric Research, Boulder, CO (United States). Climate and Global Dynamics Div.

Bosilovich, M.G. and Lawford, R., 2002. Coordinated enhanced observing period (CEOP) international workshop. *Bulletin of the American Meteorological Society*, 83(10): 1495-1499.

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.

Constantinescu, E.M., Sandu, A., Chai, T. and Carmichael, G.R., 2007. Ensemble-based chemical data assimilation I: general approach. *Quarterly Journal of the Royal Meteorological Society*, 133: 1229-1243.

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.

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

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

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

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

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

Dee, D.P., Gaspari, G., Redder, C., Rukhovets, L. and Da Silva, A.M., 1999. Maximum-likelihood estimation of forecast and observation error covariance parameters. Part II: Applications. *Monthly weather review*, 127(8): 1835-1849.

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.

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

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

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

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

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

Hydrometeorology, 15(1): 359-375.

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

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

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

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

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

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

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

Kerr, Y.H., Waldteufel, P., Wigneron, J.-P., Delwart, S., Cabot, F., Boutin, J., Escorihuela, M.-J., Font, J., Reul, N. and Gruhier, C., 2010. The SMOS mission: New tool for monitoring key elements ofthe global water cycle. *Proceedings of the IEEE*, 98(5): 666-687.

Koster, R.D., Guo, Z.C., Yang, R.Q., 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): 4322-4335.

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

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

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

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

Liang, X., Zheng, X., Zhang, S., Wu, G., Dai, Y. and Li, Y., 2012. Maximum likelihood estimation of inflation factors on error covariance matrices for ensemble Kalman filter assimilation. *Quarterly Journal of the Royal Meteorological Society*, 138: 263-273.

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

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.

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

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

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

Niu, G.-Y., Yang, Z.-L., Dickinson, R.E., Gulden, L.E. and Su, H., 2007. Development of a simple groundwater model for use in climate models and evaluation with Gravity Recovery and Climate Experiment data. *Journal of Geophysical Research*, 112(D7).

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

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

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

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

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

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

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

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

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

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

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

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.

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

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

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

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

in semiarid regions. *Journal of Hydrometeorology*, 10: 780-793.
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.
Yang, S.-C., Kalnay, E. and Enomoto, T., 2015. Ensemble singular vectors and their use as additive inflation in EnKF. *Tellus A*, 67.
Yilmaz, M.T., DelSole, T. and Houser, P.R., 2011. Improving Land Data Assimilation Performance with a Water Budget Constraint. *Journal of Hydrometeorology*, 12(5): 1040-1055.
Yilmaz, M.T., 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.
Zhang, S., Yi, X., Zheng, X., Chen, Z., Dan, B. and Zhang, X., 2014. Global carbon assimilation system using a local ensemble Kalman filter with multiple ecosystem models. *Journal of Geophysical Research-Biogeosciences*, 119(11): 2171-2187.
Zhang, S., Zheng, X., Chen, J., Chen, Z., Dan, B., Yi, X., Wang, L. and Wu, G., 2015. A global carbon assimilation system using a modified ensemble Kalman filter. *Geoscientific Model Development*, 8: 805-816.
Zhao, L. and Yang, Z.L., 2018. Multi-sensor land data assimilation: Toward a robust global soil moisture and snow estimation. *Remote Sensing of Environment*, 216: 13-27.
Zheng, X., 2009. An adaptive estimation of forecast error covariance parameters for Kalman filtering data assimilation. *Advances in Atmospheric Sciences*, 26(1):

154-160.

Figure captions

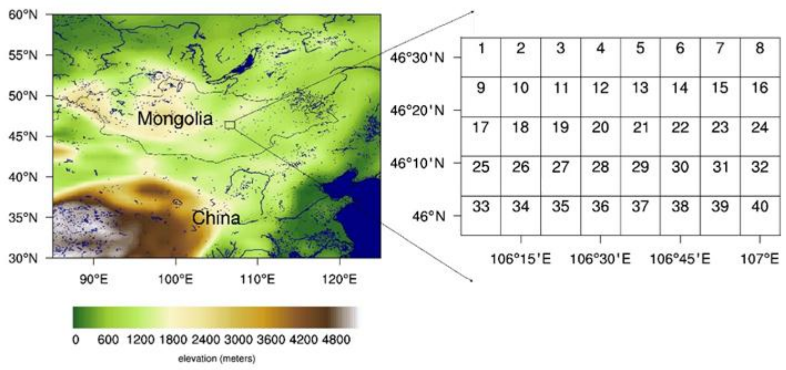

Figure 1. The topography and river distribution (left plot) and the geographical location of the synthetic study area (right plot).

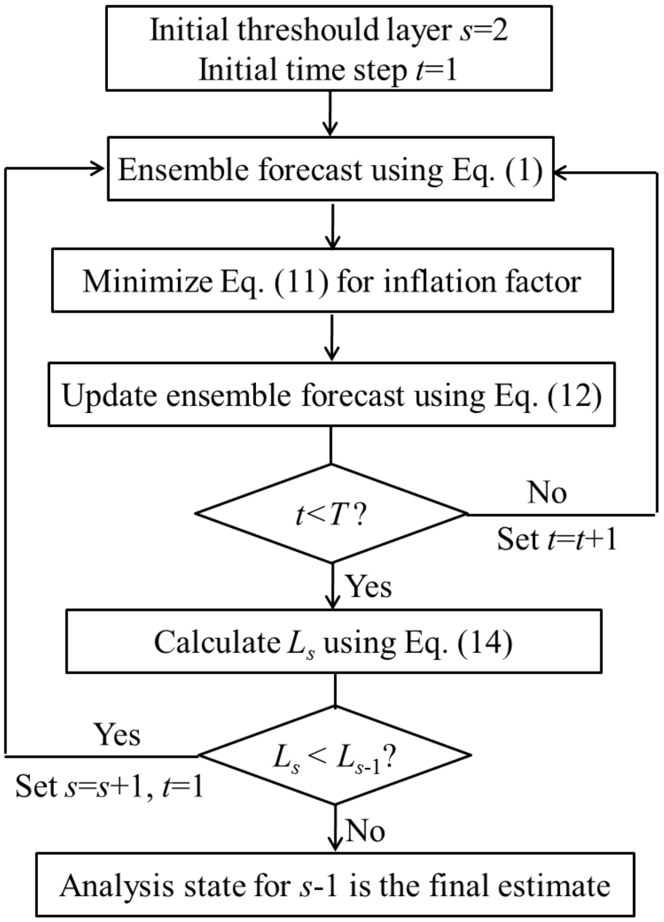

Figure 2. The assimilation procedure and localization scale factor estimation in the experiments. All of the equations are in accordance with that described in the text.

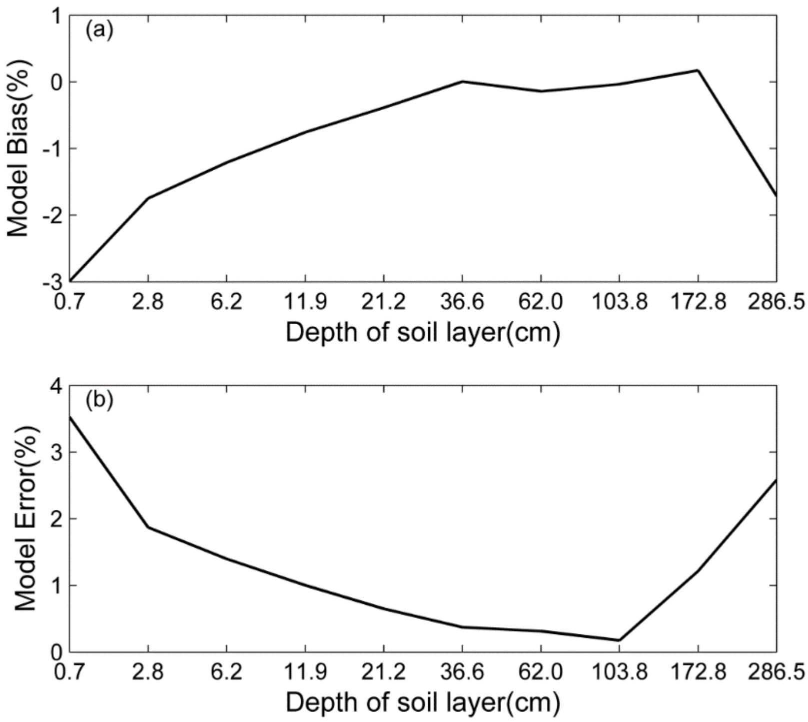

Figure 3. The areal average of the model's bias (a) and error (b) for one step in the soil moisture content between the CoLM and the CLM 4.0. The horizontal axis represents the layer depth.

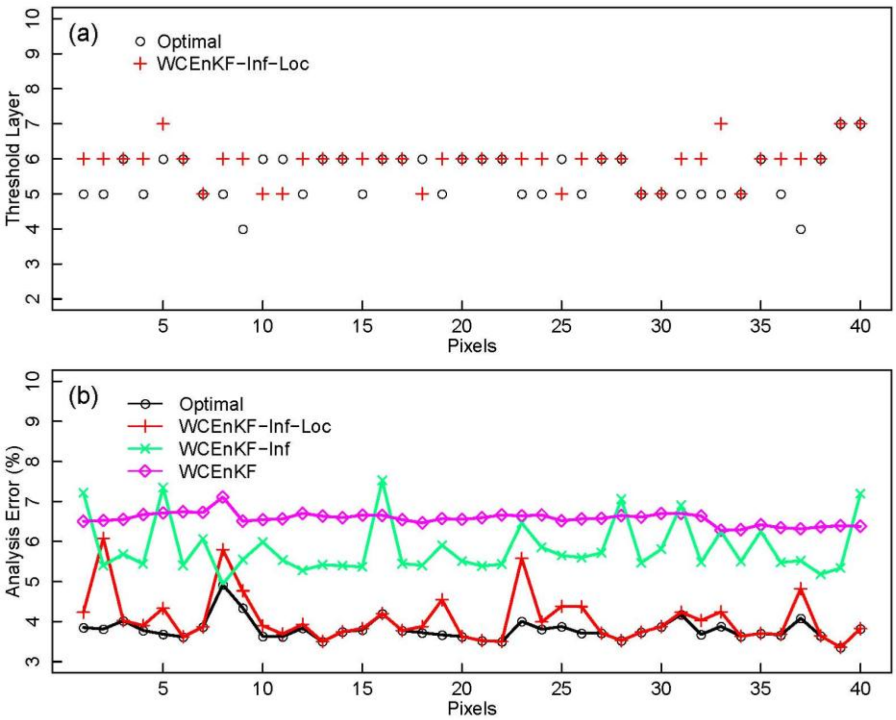

Figure 4. The threshold layers and analysis error for each pixel in the synthetic experiment. Graph (a) illustrates the optimal and WCEnKF-Inf-Loc threshold layers of each pixel. Graph (b) shows the column RSME of each pixel in different schemes with water balance constraint (Optimal, WCEnKF-Inf-Loc,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

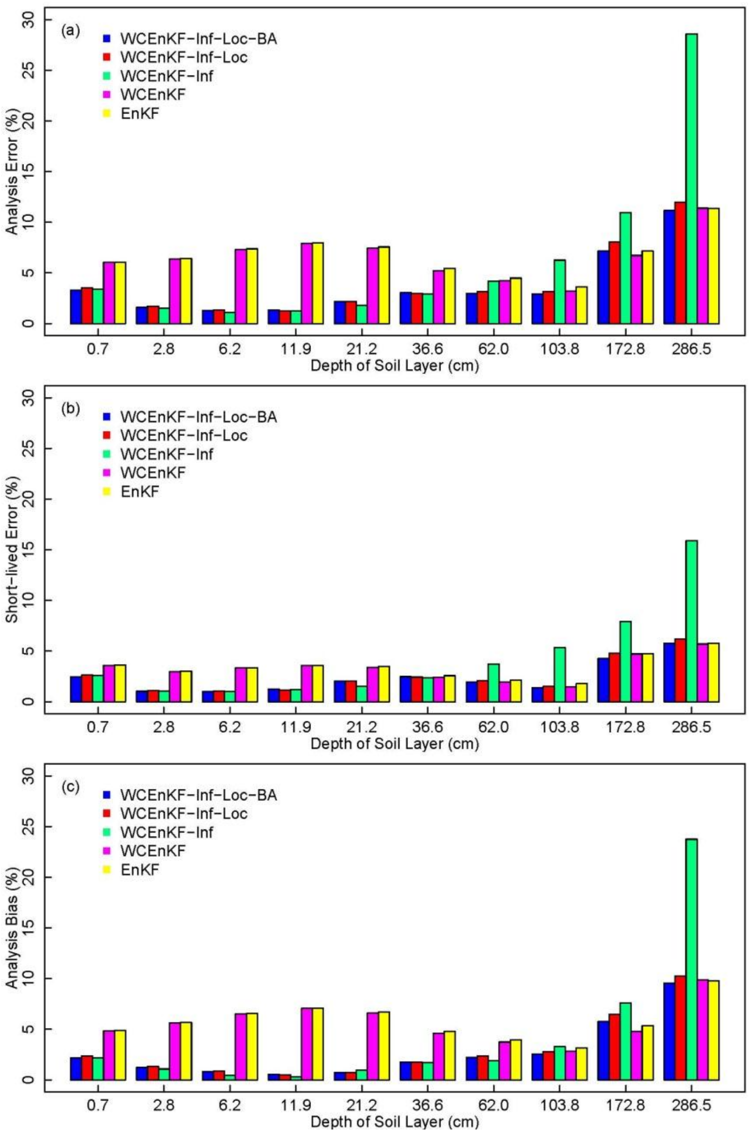

Figure 5. The assimilation results in each layer for the five schemes: a weakly constrained bias-aware ensemble Kalman filter with forecast error inflation and vertical localization (WCEnKF-Inf-Loc-BA), a weakly constrained ensemble Kalman filter with forecast error inflation and vertical localization (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.

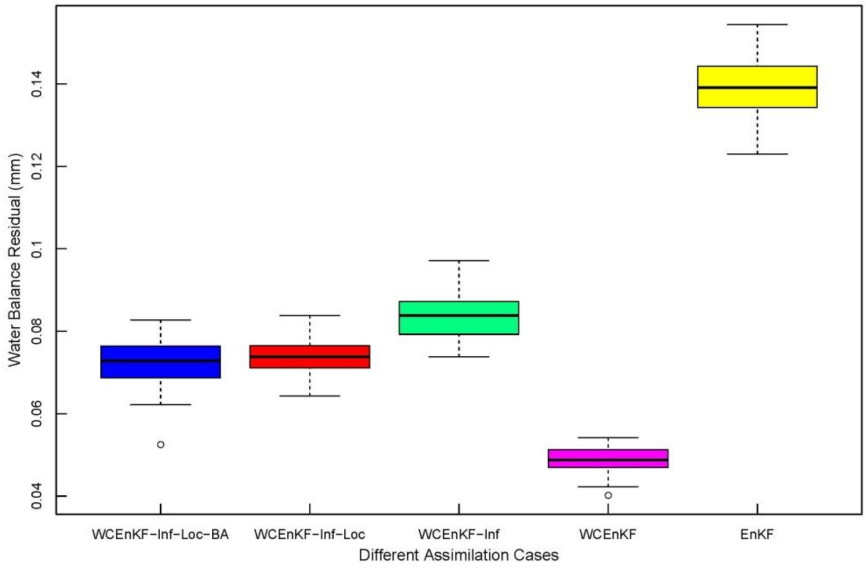

Figure 6. The box plot of the water balance residual in all 40 pixels for the WCEnKF-Inf-Loc-BA, WCEnKF-Inf-Loc,WCEnKF-Inf, WCEnKF and EnKF assimilation schemes.

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

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

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