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. An additional bias-aware assimilation for reducing the analysis bias is also investigated. The results of the assimilation process suggest that the inflation approach effectively reduces the analysis error from 6.70% to 2.00% in shallow layers, but increases from 6.38% to 12.49% in deep layers. The vertical localization approach leads to 6.59% of the analysis error in deep layers, and the bias-aware assimilation scheme further reduces to 6.05% . The spatial average of the water balance residual is 0.0487 mm of weakly constrained EnKF scheme, and 0.0737 mm of weakly constrained EnKF with inflation and localization scheme, which are much smaller than 0.1389 mm of the EnKF scheme.

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

38   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 is suggested 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.

90   In this study, a scheme for assimilating synthetic observations of the soil

91   moisture content into land models was developed based on EnKF method, which can

92   provide a foundation for further satellite data assimilation. For the synthetic

93   experiment, the Version 4.0 of the Community Land Model (CLM 4.0, (Lawrence *et*

94   *al.* 2011; Oleson *et al.* 2010)) was used to generate the "true values" and the Common

95   Land Model (CoLM, (Dai *et al.* 2003)) was selected as the forecast operator. The

96   differences in these two models are referred to the model error in an imperfect land

97   surface model. The inflation factors are estimated at every observation time step

98   during the assimilation process by minimizing the -2log-likelihood of the difference

99   between the forecast and the observation (Liang *et al.* 2012; Zheng 2009). For

100   assimilating observations near the surface only, such inflation approach can improve

101   the estimates of the forecast error statistics in shallow soil layers but may artificially

102   enlarge the forecast error statistics in deep soil layers. To avoid the possibility of

103   decreasing the quality of the estimates in deep soil layers, a vertical localization with

104   weighting of observations is adopted (Janjić *et al.* 2011). In this approach, a

105   localization function multiplies the weights on the components of the state vector

106   according to the distance from state layer to the observation. Moreover, the method

107   based on the maximum likelihood estimation was proposed to estimate the optimal

108   localization scale factor.

109   A major objective of soil moisture data assimilation is to address biases in

110   models and observations (Koster *et al.* 2009; Reichle and Koster 2004). In this study,

111   we only assume that models could be biased, while the soil moisture observations are

112   assumed to be unbiased. Moreover, the soil moisture observations are restricted in

113   shallow layer, so there is no observation available to directly correct the modeled soil

114   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,

165 monsoon processes and the prediction of river flows. More details can be found at

166 http://www.ceop.net.

167

168 2.2 Forcing data

169 In this study, synthetic experiments were conducted to explore the accuracy of the

170 assimilation schemes. The simulations were driven by forcing data (including

171 radiation, wind, pressure, humidity, precipitation and temperature) from the

172 0.125°x0.125° ERA-Interim dataset (Dee *et al.* 2011) that had been scaled down to

173 provide a temporal resolution of one hour.

174

175 2.3 Models

176 The Common Land Model (CoLM) developed by Dai et al. (2003) is a

177 third-generation land surface model. It combines the best features of three successful

178 models: the Land Surface Model (LSM, (Bonan 1996)), the Biosphere-Atmosphere

179 Transfer Scheme (BATS, (Dickinson *et al.* 1993)) and the 1994 version of the Chinese

180 Academy of Sciences/Institute of Atmospheric Physics model (IAP94, (Dai *et al.*

181 2003)), and is being further developed. The primary characteristics of the model

182 include 10 unevenly spaced soil layers (see Table 1), one vegetation layer, 5 snow

183 layers (depending on the snow depth), explicit treatment of the mass of liquid water,

184 ice and phase changes within the system of the snow and soil, runoff parameterization

185 following the TOPMODEL concept, a tiled treatment of the sub-grid fraction of the

186 energy and water budget balance (Dai *et al.* 2003) and a canopy

187 photosynthesis-conductance mode that describes the simultaneous transfer of CO_2 and

188 water vapor into and out of the vegetation. The model parameters include data on the

189 global terrain, elevation, land use, vegetation, land-water mask and hybrid

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

191   arc seconds.

192   Version 4.0 of the Community Land Model (CLM 4.0) (Lawrence *et al.* 2011;

193   Oleson *et al.* 2010) is the land surface parameterization used with the Community

194   Atmosphere Model (CAM 4.0) and the Community Climate System Model (CCSM

195   4.0). The CLM 4.0 includes bio-geophysics, the hydrologic cycle, biogeochemistry

196   and the dynamic vegetation. CLM 4.0 simulates the bio-geophysical processes in each

197   sub-grid unit independently and maintains its own prognostic variables. The

198   parameters used in the CLM4.0 differ from those used in the CoLM. For example, the

199   soil texture data are derived from the IGBP soil data, and the land use data are derived

200   from the UNH Transient Land Use and Land Cover Change Dataset

201   (http://luh.umd.edu/).

202   In addition to using different parameters, the two models have different structures.

203   For example, a model of groundwater-soil water interactions (Niu *et al.* 2007; Niu *et*

204   *al.* 2005) has been incorporated into the CLM 4.0, while zero water flux at the bottom

205   of a soil column is assumed in the CoLM. Besides, the CLM 4.0 has the same vertical

206   discretization scheme as the CoLM (see Table 1), which makes comparing the results

207   of the two models convenient.

208

209   **3. Methods**

210   3.1 Forecast and observation systems

211   Using notation similar to that used by Yilmaz et al. (2011), the forecast system

212   can be written as

213
$$\mathbf{y}_{n,t}^{f} = \mathbf{M}_{n,t-1}\left(\mathbf{y}_{n,t-1}^{a}\right),$$
(1)

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

215 study, the ensemble size is set to 100), $\mathbf{M}_{n,t-1}$ is a CoLM forced by the n-th perturbed

216 atmospheric forcing, and $\mathbf{y}$ is a state vector containing 126 variables. The superscript

217 "f" and "a" specify the forecast and analysis, respectively.

218   Let $\mathbf{x}$ be the state variables related to the water budget, that comprises of $\mathbf{SM}$

219 and $\mathbf{SIC}$ (the soil moisture content and the soil ice content in % at the 10 vertical

220 levels listed in Table 1), CWC and SWE (the canopy's water content and the snow

221 water equivalent in kg/m^2). In this study, only $\mathbf{x}$ is updated by data assimilation, while

222 the model propagates changes to the other variables over time.

223   For the traditional EnKF, the forecast error covariance matrix $\mathbf{P}_t$ is

224 obtained from the ensemble of their anomalies,

225
$$\mathbf{P}_t = \frac{1}{N-1}\sum_{n=1}^{N}\left(\mathbf{x}_{n,t}^f - \mathbf{x}_t^f\right)\left(\mathbf{x}_{n,t}^f - \mathbf{x}_t^f\right)^{\mathrm{T}} . \tag{2}$$

226 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

227 ensemble mean of $\mathbf{x}_{n,t}^f$. To avoid overestimation of the co-variability between shallow

228 observations and soil moistures deeper than a threshold layer s (see section 3.2 for the

229 estimation of s), the following vertical localization function with weighting of

230 observations $\mathbf{\rho}_s$ (Janjić $et\ al.$ 2011) will be applied on $\mathbf{P}_t$, i.e.,

231
$$\mathbf{\rho}_s(l) = \exp\left(-\mu_s\left|d_l - d_o\right|\right) \tag{3}$$

232 where l represents for the l-level soil layer, d_l and d_o represent the depths of

233 l-level soil layer and observation, respectively. $\left|d_l - d_o\right|$ is the Euclidian distance

234 between the two layers. μ_s is estimated by minimizing the following mean square

235 error between vertical localization function Eq (3) and a step function with threshold

236 layer s,

237 $$M(\mu) = \sum_{l \leq s} \left[\exp(-\mu|d_l - d_o|) - 1 \right]^2 + \sum_{l > s} \left[\exp(-\mu|d_l - d_o|) \right]^2 \qquad (4)$$

238 The estimated μ_s is listed in Table 2.

239 The observations of the soil moisture content are collected at a depth of 3 cm at

240 6:00 am every day (denoted by o_t). The observation system is defined as

241 $$o_t = \mathbf{h}\mathbf{x}_t + \varepsilon_t, \qquad (5)$$

242 where observational operator $\mathbf{h}$ is a 22-dimensional vector which linearly interpolated

243 the soil moisture at depths of 2.8 cm and 6.2 cm to depth of 3 cm, $\mathbf{x}_t$ represents the

244 true values of the state variables related to the water budget at the time step t and ε_t

245 is the observational error with mean zero and variance R_t. Since, the main objective

246 of this study is for methodology related to linear observational operators. Choosing

247 the linear interpolation as observational operator is only for convenience.

248

249 3.2 Assimilation with water budget constraint

250 Assimilating data on the soil moisture content usually results in an imbalance in

251 the water budget. To reduce this imbalance, a weak constraint on the water budget

252 (Yilmaz *et al.* 2011) is adopted in this study. The ensemble water budget residual at

253 time step t can be expressed as

254 $$r_{n,t} \equiv \beta_{n,t} - \mathbf{c}^{\mathrm{T}}\mathbf{x}_{n,t}^a, \qquad (6)$$

255 where

256 $$\beta_{n,t} = \mathbf{c}^{\mathrm{T}}\mathbf{x}_{n,t-1}^a + Pr_t - Ev_{n,t}^f - Rn_{n,t}^f, \qquad (7)$$

257 where $\mathbf{c}$ is a 22-dimensional vector that converts the units to millimeters (*mm*) and

258 adds up the states in $\mathbf{x}$, the diagnostic variables Pr_t, $Ev_{n,t}^f$ and $Rn_{n,t}^f$ (*mm*) are

259 scalars specifying the states of the precipitation, evapotranspiration and runoff,

260   respectively, in each pixel.

261   The cost function used to estimate the state variables with the weak water budget

262   constraint (Eq. (6)) is

263

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

264   where

265

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

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

267   covariance matrix defined by

268

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

269   where $\mathbf{P}_t$ is defined as Eq. (2); $\left[\boldsymbol{\rho}_s\right]$ is a diagonal matrix which localizes the soil

270   moisture error (i.e. it is $\boldsymbol{\rho}_s$ defined by Eq. (3) for the soil moisture contents and 1 for

271   other variables). $\left[\sqrt{\lambda_t}\right]$ is also a diagonal matrix which inflates the forecast soil

272   moisture error (i.e. it is a scalar λ_t for the soil moisture contents and 1 for other

273   variable). λ_t is estimated by minimizing the -2log-likelihood of the difference

274   between the forecast and the observation (Dee and Da Silva 1999; Liang *et al.* 2012;

275   Zheng 2009),

276

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

277   The estimated forecast error inflation factor is denoted as $\hat{\lambda}_t$. The perturbed analysis

278   states of the variables related to water budget can be derived by minimizing Eq. (8),

279   which has the analytic form

280 $$\mathbf{x}_{n,t}^{a} = \mathbf{x}_{n,t}^{f} + \mathbf{P}_{t}^{a}\mathbf{h}^{\mathrm{T}}R_{t}^{-1}\left(o_{t} + \varepsilon_{n,t} - \mathbf{h}\mathbf{x}_{n,t}^{f}\right) + \mathbf{P}_{t}^{a}\mathbf{c}\varphi_{t}^{-1}\left(\beta_{n,t} - \mathbf{c}^{\mathrm{T}}\mathbf{x}_{n,t}^{f}\right), \tag{12}$$

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

282 and

283 $$\mathbf{P}_{t}^{a} = \left(\mathbf{h}^{\mathrm{T}}R_{t}^{-1}\mathbf{h} + \mathbf{P}_{s,t}^{-1} + \mathbf{c}\varphi_{t}^{-1}\mathbf{c}^{\mathrm{T}}\right)^{-1}, \tag{13}$$

284 its analysis error covariance matrix.

285 For estimating the optimal threshold layer, define the -2log-likelihood of the total

286 difference between the forecasts and the observations,

287 $$L_{s} \equiv \sum_{t=1}^{T}(-2L_{s,t}(\hat{\lambda}_{t})). \tag{14}$$

288 The optimal threshold layer $\hat{s}$ is selected as the smallest number s such that L_{s} is

289 the minimum of $\left\{L_{2}, L_{3}, \cdots, L_{s+1}\right\}$. The final analysis state is the selected

290 corresponding to the optimal threshold layer $\hat{s}$. The complete assimilation procedure

291 with water budget constraint is shown in Figure 2.

292

293 3.3 Bias-aware assimilation

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

295 the analysis bias.

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

297 $$\mathbf{b}_{t} = \mathbf{b}_{t-1} - \gamma\tilde{\mathbf{P}}_{s,t}\mathbf{h}^{\mathrm{T}}\left(\mathbf{h}\tilde{\mathbf{P}}_{s,t}\mathbf{h}^{\mathrm{T}} + R_{t}\right)^{-1}\left(o_{t} - \mathbf{h}\left(\tilde{\mathbf{x}}_{t}^{f} - \mathbf{b}_{t-1}\right)\right). \tag{15}$$

298 where the scalar parameter γ that controls the magnitude of the forecast bias is

299 estimated following Dee and Todling (2000) (see Eqs (A5)-(A6) of Appendix A), $\tilde{\mathbf{x}}_{t}^{f}$

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

301 $\tilde{\mathbf{x}}_{n,t-1}^{a}$, $\tilde{\mathbf{P}}_{s,t}$ is the corresponding adjusted forecast error covariance (see Eq. (A2) of

302 Appendix A).

303   Then the perturbed assimilated states are

304
$$\tilde{\mathbf{x}}_{n,t}^{a} = \tilde{\mathbf{x}}_{n,t}^{f} - \mathbf{b}_{t-1} + \tilde{\mathbf{P}}_{t}^{a}\mathbf{h}^{\mathrm{T}}R_{t}^{-1}\left(o_{t} + \varepsilon_{n,t} - \mathbf{h}\left(\tilde{\mathbf{x}}_{n,t}^{f} - \mathbf{b}_{t-1}\right)\right)$$
$$+ \tilde{\mathbf{P}}_{t}^{a}\mathbf{c}\tilde{\varphi}_{t}^{-1}\left(\tilde{\beta}_{n,t} - \mathbf{c}^{\mathrm{T}}\left(\tilde{\mathbf{x}}_{n,t}^{f} - \mathbf{b}_{t-1}\right)\right)$$
.   (16)

305 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.

306

307 **4. Synthetic experiments**

308 4.1 Experimental design

309   To investigate the performance of the WCEnKF-based methods that incorporate

310 inflation, vertical local localization and bias-awre assimilation, synthetic experiments

311 were performed using the CoLM. Unlike the "perfect model" assumption used in

312 Yilmaz et al. (2011), the assumptions of this study are accounted for the error in the

313 model, especially the structural error. Because there were structural differences in the

314 models of the water cycle (see section 2.3) used in the two models, CLM 4.0 was used

315 to generate the "true values" (i.e., to perform a reference run) for the synthetic

316 experiments and CoLM was selected as the forecast operator (i.e., to perform an

317 open-loop run). Therefore, the CLM 4.0 and the CoLM were both integrated on a

318 $0.125°$ grid (see Figure 1 for the locations) with a time step of one hour. The

319 assimilation time was set to 6:00 am every day. The assimilation experiments were

320 conducted with 5 scenarios: the traditional ensemble Kalman filter (EnKF), a weakly

321 constrained ensemble Kalman filter (WCEnKF), a weakly constrained ensemble

322 Kalman filter with inflation (WCEnKF-Inf), a weakly constrained ensemble Kalman

323 filter with inflation and localization (WCEnKF-Inf-Loc) and a weakly constrained

324 ensemble Kalman filter with inflation, localization and bias-aware assimilation

325 (WCEnKF-Inf-Loc-BA).

326   Synthetic observations were obtained by interpolating $\mathbf{SM}_t$ to a depth of 3 cm

327   and adding noise with a normal distribution ($N(\mu=0,\sigma=0.5\%)$). The initial state

328    $\mathbf{x}_0$, was generated by running the CoLM from October 1, 2002 to June 1, 2003. Each

329   component of the initial state was perturbed using an independent standard Gaussian

330   random variable times 5% of magnitude of the component. The forcing data were

331   perturbed in the manner described in Yilmaz et al. (2011). The synthetic experiments

332   were conducted from June 1, 2003 to October 1, 2003. The state variables for each

333   pixel were updated independently.

334

335   4.2 Validation statistics

336   4.2.1 Model error and bias

337   The model errors are defined as the difference between the actual values and the

338   model's predictions based on true initial values, and the bias is the average of the error

339   in the model during the relevant period. Let x_t denote the true values of the soil

340   moisture content at time t for a location and vertical soil layer. x_t^M denotes the model

341   predicted soil moisture from the true state at the previous time step t-1. The model's

342   bias and error variance for one step can be written as

343
$$b_M = \frac{1}{a_{ts}} \sum_{t=1}^{a_{ts}} \left(x_t^M - x_t \right),$$
(17)

344
$$v_M = \frac{1}{a_{ts}} \sum_{t=1}^{a_{ts}} \left(x_t^M - x_t \right)^2,$$
(18)

345   where a_{ts} is the number of time steps over which the observations made at 6:00 am

346   each day are assimilated.

347   4.2.2 Validation of analysis soil moisture

348   The true soil moisture content values from 7:00 am to 5:00 am next day are used

349   to validate analysis states. For a location and vertical soil layer, let $x_{t,h}$ be the true

350   soil moisture content at hour h on day t, and $x_{t,h}^{f}$ represent the forecasted soil

351   moisture content at hour h from analysis state x_{t}^{a} at 6:00 am on day t. The analysis

352   bias is defined as

353
$$b_a = \frac{1}{23a_{ts}} \sum_{t=1}^{a_{ts}} \sum_{h=7}^{29} \left(x_{t,h}^{f} - x_{t,h} \right). \tag{19}$$

354   The analysis error variance is defined as

355
$$\begin{aligned} v_a &= \frac{1}{23a_{ts}} \sum_{t=1}^{a_{ts}} \sum_{h=7}^{29} \left(x_{t,h}^{f} - x_{t,h} \right)^{2} \\ &= \frac{1}{23a_{ts}} \sum_{t=1}^{a_{ts}} \sum_{h=7}^{29} \left(x_{t,h}^{f} - x_{t,h} - b_a \right)^{2} + b_a^{2} \end{aligned} \tag{20}$$

356   (See Appendix B for the proof)

357   4.2.3 Water balance

358   Following Yilmaz (2011), the water budget imbalance at location is evaluated

359   using the water balance residual,

360
$$R = \frac{1}{Na_{ts}} \sum_{t=1}^{a_{ts}} \sum_{n=1}^{N} r_{n,t}. \tag{21}$$

361   where N is the ensemble size, a_{ts} is the number of assimilation time steps, and $r_{n,t}$ is

362   the ensemble water budget residual at time step t as defined in Eq. (6).

363

364   **5. Results**

365   In the synthetic experiments, the magnitudes of the model's bias and error were

366   calculated using Eqs (17) and (18), respectively, and are shown in Figure 3. It shows

367   that the model's bias was almost negative from Figure 3a. The negative bias in the

368   surface layer was the result of a combination of a lower surface roughness and a larger

369   leaf area index in the CoLM; these values led to more soil evaporation and more

370   canopy interception and could result in a smaller amount of water infiltrating the soil

371   than the amount modeled using the CLM 4.0. In the CoLM, the porosity of each layer

372   was less than it was in the CLM 4.0, which retained less water and contributed to the

373   negative bias of the upper 9 layers. However, the magnitude of the bias increased to 2%

374   in the bottom layer. The significant difference between the two models at the bottom

375   layer could be ascribed to their different boundary conditions. Interactions between

376   the soil moisture content and the ground water at the bottom of the soil column were

377   modeled in the CLM 4.0 (Oleson *et al.* 2010) but not in the CoLM. The error in each

378   model (Figure 3b) fluctuated in a manner similar to that of the model's bias. Unbiased

379   observations are necessary for correcting bias in a model, which is not possible in

380   many realistic applications, especially in assimilating remote sensing retrievals. Since

381   satellite observations of the soil moisture content of deep layers are unavailable, only

382   removing the bias in shallow layers would introduce error in model dynamics.

383

384   5.1 Forecast error inflation and vertical localization

385   In the synthetic experiments, the study domain comprised 40 pixels. At each point

386   in the grid-scale threshold layer, the localization scale factor μ_s, was determined

387   independently. Therefore, totally 9 sets of experiments with different localization

388   scale factor (see Table 2) were conducted separately. Among these experiments, the

389   "optimal" case for each pixel was defined as the case in which the column averaged

390   analysis error (Eq. (20)) was minimized (shown in Figure 4). According to Figure 4a,

391   the corresponding threshold layer s of μ_s was generally between 5 and 6 in both

392   cases, which could be ascribed to the homogeneous soil texture and land cover. In the

393   WCEnKF-Inf-Loc, there were 19 pixels in which the threshold layers were "optimal,"

394 and the layers selected in the other pixels were suboptimal (most were roughly one

395 layer away from the "optimal" case). As shown in Figure 4b, the spatial average of the

396 root analysis error variance (Eq. (20)) of the WCEnKF-Inf-Loc (4.09%) was

397 comparable with the optimal value (3.84%) even though s was not selected on the

398 basis of minimizing the analysis error.

399 The spatial average of the root analysis error variance in each layer in the

400 schemes with (WCEnKF-Inf-Loc and WCEnKF-Inf) and without (WCEnKF)

401 inflation are displayed in Figure 5a. Above 36.6 cm, the analysis errors of the schemes

402 without inflation (6.70%) were substantially larger than those of the schemes with

403 inflation (2.00%) for the synthetic experiments. This suggested that inflation provided

404 a better estimate in the layers close to the observation. When no inflation was

405 performed, the accuracy of the soil moisture content was barely improved over that of

406 the open-loop (not shown here).

407 By comparing the schemes with (WCEnKF-Inf-Loc) and without (WCEnKF-Inf)

408 vertical localization, the impact of this approach on the assimilation accuracy in each

409 layer is shown in Figure 5a. Because the threshold layer of the localization function

410 $\boldsymbol{\rho}_s$ was layer 6 (36.6 cm) for 28 of the pixels (see Figure 4a), the spatial average of

411 root analysis error variance of the results of the WCEnKF-Inf-Loc is almost identical

412 to that of the results of the WCEnKF-Inf for depths above 36.6 cm. In contrast,

413 inflation increased the analysis error in the soil moisture content of the deep layers in

414 the WCEnKF-Inf from 6.38% to 12.49%. In this model, the sample error covariances

415 of the moisture contents of shallow and deep soil were inflated by a factor greater than

416 6 (the average inflation factor was 6.25). This could lead to larger assimilation errors

417 for deep soil moisture profiles in the WCEnKF-Inf. Therefore, inflation should be

418 used with vertical localization to reduce the spurious covariance resulting from the

419 covariance inflation-based approach.

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

421 was helpful in avoiding erroneous estimates of the soil moisture contents at lower

422 levels (in the WCEnKF-Inf). A comparison of the analysis error at a depth of 3 cm

423 (i.e., the depth of the assimilated observations was 3 cm) in the models with

424 (WCEnKF-Inf and WCEnKF-Inf-Loc) and without (WCEnKF) inflation showed that

425 the inflation technique significantly reduces the analysis error at the depth at which

426 observations are made.

427 To investigate the role of bias correction, the spatial averaged root analysis error

428 variance (Eq. (20)) of WCEnKF-Inf-Loc-BA and WCEnKF-Inf-Loc were compared.

429 According to Figure 5a, the spatial averaged root analysis error variances of the two

430 schemes were comparable with each other (2.12% for the WCEnKF-Inf-Loc-BA and

431 2.16% for the WCEnKF-Inf-Loc) in the layers that were shallower than 36.6 cm. This

432 could be due to that the observations are closer to the shallow layers and the vertical

433 localization approach is reasanable effective to reduced the bias. However, for the

434 layers that were deeper than 62.0 cm, the averaged root analysis error of the

435 WCEnKF-Inf-Loc-BA (6.05%) was less than that of the WCEnKF-Inf-Loc (6.59%).

436

437 5.2 The water budget constraint

438 In the synthetic experiment, the weak constraint on the water budget reduced the

439 water balance residual significantly in each pixel and the results are shown in Figure 6.

440 It shows that, the spatial average of the water balance residual of WCEnKF scheme

441 was 0.0487 mm, which was much smaller than that of the EnKF scheme (0.1389 mm).

442 Therefore, the assimilation scheme with water budget constraint can indeed reduce the

443 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. If the focus of a study or

experiment is reducing water balance, WCEnKF could be a better choice and computationally faster than WCEnKF-Inf and WCEnKF-Inf-Loc schemes. Accordingly, it is plainly obvious that the water balance residual of the scheme WCEnKF-Inf is larger than that of the scheme WCEnKF. However, the objective in this study is to reduce water balance without significantly increasing the analysis error. Since the analysis errors for WCEnKF in the layers shallower than 36.6 cm are significantly larger than those for the schemes with inflation, WCEnKF is not preferred.

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 2nd and 3rd 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

494   used, the soil moisture contents of the deep layers are not significantly updated.

495   Consequently, larger errors are avoided in the deep layers (see WCEnKF-Inf-Loc in

496   Figure 5a).

497   Comparing to traditional EnKF without inflation and localization, although

498   mainly the soil moisture contents of layers above the threshold layer (usually the 5th or

499   6th layer) were updated at each time step during the assimilation process when the

500   WCEnKF-Inf-Loc was used, Figure 5a shows that the soil moisture contents of the

501   layers below the threshold layer, especially the 6th and 7th layers, are also improved.

502   This may be because the model propagates changes in the shallow layers downward,

503   adjusting the soil moisture contents of the deep layers. Because the soil moisture

504   content of layers above the threshold layer was improved during the previous time

505   step, this process results in better predictions of the soil moisture contents of layers

506   below the threshold layer, and therefore, reduces the analysis error in layers below the

507   threshold layer.

508

509   6.2 Bias correction

510   Geophysical models are never perfect and usually produce estimates with biases

511   that vary in time and in space (Reichle 2008). Therefore, bias correction is important

512   for assimilating data into models. In this study, only soil moisture in shallow layers

513   can be observed (in order to mimic the satellite observation), so the bias for the soil

514   moisture in deeper layers can not be entirely removed only using the observations.

515   However, bias can be detected by monitoring statistics of observation-minus-forecast

516   residual in the assimilation systems. Therefore the bias-awre assimilation proposed by

517   Dee (2005) was further applied to reduce the bias of soil moisture in all layers.

518   For further evaluating the efficacy of the bias-awre assimilation scheme, the

519   analysis error variance was decomposed to a short-lived component (Figure 5b) and a

520   bias component (Figure 5c) for the synthetic experiment. It shows that for the

521   bias-blind data assimilation scheme (WCEnKF-Inf-Loc), both short-lived errors and

522   biases reduce in the layers close to observation, while maintain the similar levels as

523   those for EnKF for the deeper layers. The covariance inflation can play an important

524   role in bias reduction. Bias can only be seen during long assimilation period. At an

525   instant time, bias and error are mixed. For the traditional EnKF, the forecast error

526   covariance matrix obtained from the ensemble of their anomalies (Eq. (2)) mainly

527   represents short-lived error, so it has to be inflated to include error related to bias.

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

529   There are other bias estimation approaches in data assimilation. For example,

530   treading bias as model variables and estimate in assimilation (De Lannoy *et al.* 2007;

531   Dee and Da Silva 1998), adjusting the state variable of the forecast model not only

532   their covariance matrix in each forecast step (Zhang *et al.* 2014; Zhang *et al.* 2015),

533   addressing the biases in the model and observations by rescaling their cumulative

534   distribution functions (Koster *et al.* 2009; Reichle and Koster 2004). The scheme

535   proposed here can provide a base line to validate the efficacy of these approaches and

536   could be further improved after these bias corrections.

537

538   6.3 Broader implications

539   In our schemes, the canopy's water content was directly updated by the soil

540   moisture observations, following the approach of previous studies (Yilmaz *et al*. 2011;

541   Yilmaz *et al*. 2012). The canopy's water content (CWC) and snow water equivalent

542   (SWE) are related to the water budget. If the water budget constraint is absent, they

543   are normally not updated and the vegetation module transports the water into the

544   vegetation layer. However, the present study focused on the assimilation with the

545   water budget constraint, then updating CWC and SWE would help to reduce the water

546   budget residuals.

547   For the assimilation with the water budget constraint but without update of CWC

548   and SWE, the state variables related to the water budget are decomposed as

549      $\mathbf{x} = \left({_1}\mathbf{x}, {_2}\mathbf{x} \right)$ where ${_1}\mathbf{x}$ comprises of SM and SIC, ${_2}\mathbf{x}$ comprises of CWC and SWE.

550      $\mathbf{c} = \left({_1}\mathbf{c}, {_2}\mathbf{c} \right)$ converts the units of $\mathbf{x} = \left({_1}\mathbf{x}, {_2}\mathbf{x} \right)$ to millimeters (mm). The assimilation

551   for not update of ${_2}\mathbf{x}$ can be achieved by substituting $\mathbf{x}$ and $\beta_{n,t}$ in section 3.2 by

552      ${_1}\mathbf{x}$ and ${_1}\beta_{n,t}$ respectively, that is

553
$$ {_1}\beta_{n,t} = {_1}\mathbf{c}^\mathrm{T} {_1}\mathbf{x}_{n,t-1}^a + {_2}\mathbf{c}^\mathrm{T} {_2}\mathbf{x}_{n,t-1}^f + Pr_t - Ev_{n,t}^f - Rn_{n,t}^f, \tag{22} $$

554   where Pr_t, $Ev_{n,t}^f$ and $Rn_{n,t}^f$ are diagnostic variables specifying the states of the

555   precipitation, evapotranspiration and runoff, respectively. By this way, the canopy's

556   water content are not updated and the vegetation module transports the water into the

557   vegetation layer. In this study, the range of the estimated CWCs for all assimilations

558   with or without update of ${_2}\mathbf{x}$ is only about 0.005 mm. Considering the estimated

559   water budget residuals are between 0.05 mm and 0.14 mm and there is no SWE in the

560   summer peried, we conclude that update of CWC has a little impact on water balance

561   in this study.

562   The most computational cost in the assimilation system is on computing the

563   localization function at each model grid cell. For the synthetic experiments with

564   CoLM model and 40 grids, it takes about 24 hours running on the personal

565   workstation. For global data assimilation with 2° resolution it could take about 3

566   months. However, the super server and parallel computation can significantly shorten

567   the computational time. A regional scale using soil texture or climate regimes can also

568 be used to delineate different regions. By this way, the computational time of global

569 data assimilation can be further reduced.

570 In the near future, we plan to validate the major conclusions under different soil

571 conditions and land cover types. Vertical localization, which uses adjacent

572 observations, should also be tested in future work. More detailed analyses of the bias

573 correction for assimilating remote sensing retrievals should be performed. The

574 response of the analytic soil moisture content to weather predictions also needs to be

575 investigated. Completing these studies should improve the state of research into

576 land-atmosphere interactions.

577

578 **7. Conclusions**

579 In this study, observations of the soil moisture content at a depth of 3 cm were

580 assimilated using an ensemble Kalman filter with several improvements. Firstly, an

581 adaptive forecast error inflation based on maximum-likelihood estimation was

582 adopted to reduce the analysis error. This study supports the idea that the proper form

583 of the forecast error covariance matrix is crucial for reducing the analysis error near

584 the layers in which observations are made. Secondly, an adequate vertical localization

585 for the ensemble-based filter was proposed associated with the forecast error

586 covariance inflation, to avoid misestimates of the soil moisture contents of deep layers.

587 Lastly, a constraint on the water balance was used in this study to reduce the water

588 budget residual substantially without significantly changing the assimilation accuracy.

589 The experiment results of synthetic study show that the WCEnKF-Inf-Loc

590 assimilation scheme can reduce the analysis error from 6.70% to 2.00% in the shallow

591 layers, with both the short-lived analysis error and the analysis bias reduced. It also

592 leads to a rational water budget residual with spatial average 0.0737 mm, which is

593 much smaller than 0.1389 mm of the EnKF scheme. The bias-aware assimilation

594 scheme is potentially useful to further reduce the analysis error arising from model

595 bias.

596

597 **Data availability** The soil moisture observations are available at http://www.ceop.net.

598 The ERA-interim forcing data used in the synthetic experiments is obtained from

599 https://apps.ecmwf.int/datasets.

600

601 **Author Contributions** BD performed the simulations and assimilations. XZ and GW

602 designed the research. GW analyzed the results. TL collected and preprocessed the

603 data. GW and XZ prepared the manuscript with contributions from all co-authors.

604

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

606

607 **Acknowledgement** This study was funded by the the National Key R&D Program of

608 China (2019YFC1510002, 2015CB953703) and the National Natural Science

609 Foundation of China (41930970, 42077421, 41705086). We would like to thank the

610 Editor and two anonymous reviewers for their insightful comments in improving the

611 manuscript. We also thank Drs. Yongjiu Dai and Qingyun Duan for their help in land

612 surface model.

613

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

615   For correcting the bias of the analysis states $\mathbf{x}^a_{n,t}$ in Eq. (12), the bias-aware

616   assimilation (Dee 2005) is appied.

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

618
$$\mathbf{b}_t = \mathbf{b}_{t-1} - \gamma \tilde{\mathbf{P}}_{s,t} \mathbf{h}^{\mathrm{T}} \left(\mathbf{h} \tilde{\mathbf{P}}_{s,t} \mathbf{h}^{\mathrm{T}} + R_t \right)^{-1} \left(o_t - \mathbf{h} \left(\tilde{\mathbf{x}}^f_t - \mathbf{b}_{t-1} \right) \right). \tag{A1}$$

619   where $\tilde{\mathbf{x}}^f_t$ is the ensemble mean of the perturbed forecast states $\tilde{\mathbf{x}}^f_{n,t}$ predicted from

620   the perturbed analysis state at previous time step $\tilde{\mathbf{x}}^a_{n,t-1}$, the forecast error covariance

621   matrix is in the form

622
$$\tilde{\mathbf{P}}_{s,t} = \left[\sqrt{\tilde{\lambda}_t} \right] [\boldsymbol{\rho}_s] \tilde{\mathbf{P}}_t [\boldsymbol{\rho}_s] \left[\sqrt{\tilde{\lambda}_t} \right], \tag{A2}$$

623   where the localization threshold s is adopted from the bias-blind scheme documented

624   in section 3.2,

625
$$\tilde{\mathbf{P}}_t = \frac{1}{N-1} \sum_{n=1}^{N} \left(\tilde{\mathbf{x}}^f_{n,t} - \tilde{\mathbf{x}}^f_t \right) \left(\tilde{\mathbf{x}}^f_{n,t} - \tilde{\mathbf{x}}^f_t \right)^{\mathrm{T}}, \tag{A3}$$

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

627
$$-2\tilde{L}_{s,t}(\tilde{\lambda}_t) = \ln\left(\mathbf{h}\tilde{\mathbf{P}}_{s,t}\mathbf{h}^{\mathrm{T}} + R_t \right) + \left(o_t - \mathbf{h}\tilde{\mathbf{x}}^f_t \right)^{\mathrm{T}} \left(\mathbf{h}\tilde{\mathbf{P}}_{s,t}\mathbf{h}^{\mathrm{T}} + R_t \right)^{-1} \left(o_t - \mathbf{h}\tilde{\mathbf{x}}^f_t \right). \tag{A4}$$

628   The scalar parameter γ in Eq. (A1) that controls the magnitude of the forecast

629   bias estimates, is derived by

630
$$\gamma = \frac{\mu}{1-\mu} \left(R_t + \mathbf{h}\mathbf{P}_t\mathbf{h}^{\mathrm{T}} \right) \left(\mathbf{h}\mathbf{P}_t\mathbf{h}^{\mathrm{T}} \right)^{-1}, \tag{A5}$$

631   where μ is estimated by minimizing the following objective function (Dee and

632   Todling 2000)

633
$$f(\mu) = \sum_n n^2 \left\{ \left[1 - \mu / \left(1 - (1-\mu)e^{-2\pi i \Delta t / n} \right) \right] \left[\sum_t (o_t - \mathbf{h}\mathbf{x}_t^f)e^{-2\pi i \Delta t / n} \right]^2 \left(R_t + \mathbf{h}\mathbf{P}_t\mathbf{h}^{\mathrm{T}} \right)^{-1} \right| - 1 \right\}^2 \quad \text{(A6)}$$

634   Then the perturbed analysis states is calculated as

635
$$\begin{aligned}
\tilde{\mathbf{x}}_{n,t}^a &= \tilde{\mathbf{x}}_{n,t}^f - \mathbf{b}_{t-1} + \tilde{\mathbf{P}}_t^a \mathbf{h}^{\mathrm{T}} R_t^{-1} \left(o_t + \varepsilon_{n,t} - \mathbf{h}\left(\tilde{\mathbf{x}}_{n,t}^f - \mathbf{b}_{t-1} \right) \right) \\
&\quad + \tilde{\mathbf{P}}_t^a \mathbf{c} \tilde{\varphi}_t^{-1} \left(\tilde{\beta}_{n,t} - \mathbf{c}^{\mathrm{T}} \left(\tilde{\mathbf{x}}_{n,t}^f - \mathbf{b}_{t-1} \right) \right)
\end{aligned} \quad \text{(A7)}$$

636   where

637
$$\tilde{\beta}_{n,t} = \mathbf{c}^{\mathrm{T}}\tilde{\mathbf{x}}_{n,t-1}^a + Pr_t - Ev_{n,t}^f - Rn_{n,t}^f, \quad \text{(A8)}$$

638
$$\tilde{\varphi}_t = \frac{1}{N-1} \sum_{n=1}^{N} \left(\tilde{\beta}_{n,t} - \frac{1}{N}\sum_{j=1}^{N} \tilde{\beta}_{j,t} \right) \times \left(\tilde{\beta}_{n,t} - \frac{1}{N}\sum_{j=1}^{N} \tilde{\beta}_{j,t} \right)^{\mathrm{T}} \quad \text{(A9)}$$

639   and

640
$$\tilde{\mathbf{P}}_t^a = \left(\mathbf{h}^{\mathrm{T}} R_t^{-1} \mathbf{h} + \tilde{\mathbf{P}}_{s,t}^{-1} + \mathbf{c}\tilde{\varphi}_t^{-1}\mathbf{c}^{\mathrm{T}} \right)^{-1}, \quad \text{(A10)}$$

641

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

643   For a location and vertical soil layer, the analysis error variance in the synthetic

644   experiment is defined as

645
$$\begin{aligned}
v_a &= \frac{1}{23 a_{ts}} \sum_{t=1}^{a_{ts}} \sum_{h=7}^{29} \left(x_{t,h}^f - x_{t,h} \right)^2 \\
&= \frac{1}{23 a_{ts}} \sum_{t=1}^{a_{ts}} \sum_{h=7}^{29} \left(x_{t,h}^f - x_{t,h} - b_a + b_a \right)^2 \\
&= \frac{1}{23 a_{ts}} \sum_{t=1}^{a_{ts}} \sum_{h=7}^{29} \left(x_{t,h}^f - x_{t,h} - b_a \right)^2 + b_a^2 + \frac{2 b_a}{23 a_{ts}} \sum_{t=1}^{a_{ts}} \sum_{h=7}^{29} \left(x_{t,h}^f - x_{t,h} - b_a \right)
\end{aligned} \quad \text{(B1)}$$

646   From the definition of analysis bias (Eq. (19)), the last term on the right hand side of

647   is zero, so Eq. (20) is proved.

648

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

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

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

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

702   Dee, D.P., Uppala, S.M., Simmons, A.J., Berrisford, P., Poli, P., Kobayashi, S., Andrae,

703   U., Balmaseda, M.A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A.C.M.,

704   van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M.,

705   Geer, A.J., Haimberger, L., Healy, S.B., Hersbach, H., Hólm, E.V., Isaksen, L.,

706   Kållberg, P., Köhler, M., Matricardi, M., McNally, A.P., Monge-Sanz, B.M.,

707   Morcrette, J.J., Park, B.K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut,

708   J.N. and Vitart, F., 2011. The ERA-Interim reanalysis: configuration and

709   performance of the data assimilation system. *Quarterly Journal of the Royal*

710   *Meteorological Society*, 137(656): 553-597.

711   Delworth, T.L. and Manabe, S., 1988. The influence of potential evaporation on the

712   variabilities of simulated soil wetness and climate. *Journal of Climate*, 1(5):

713   523-547.

714   Dickinson, R.E., Henderson-Sellers, A. and Kennedy, P.J., 1993. Biosphere

715   Atmosphere Transfer Scheme (BATS) Version le as Coupled to the NCAR

716   Community Climate Model.

717   Dorigo, W.A., Wagner, W., Hohensinn, R., Hahn, S., Paulik, C., Xaver, A., Gruber, A.,

718   Drusch, M., Mecklenburg, S., van Oevelen, P., Robock, A. and Jackson, T.,

719   2011. The International Soil Moisture Network: a data hosting facility for

720   global in situ soil moisture measurements. *Hydrology and Earth System*

721   *Sciences*, 15(5): 1675-1698.

722   Dumedah, G. and Walker, J.P., 2014. Evaluation of Model Parameter Convergence

723   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.

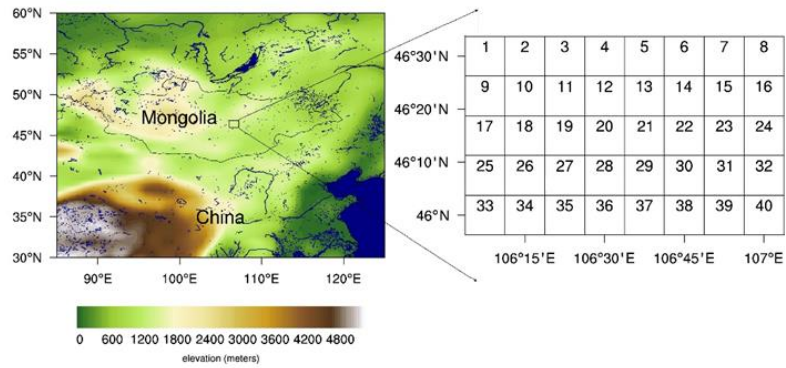
Zheng, X., 2009. An adaptive estimation of forecast error covariance parameters for Kalman filtering data assimilation. *Advances in Atmospheric Sciences*, 26(1):

874   154-160.

875

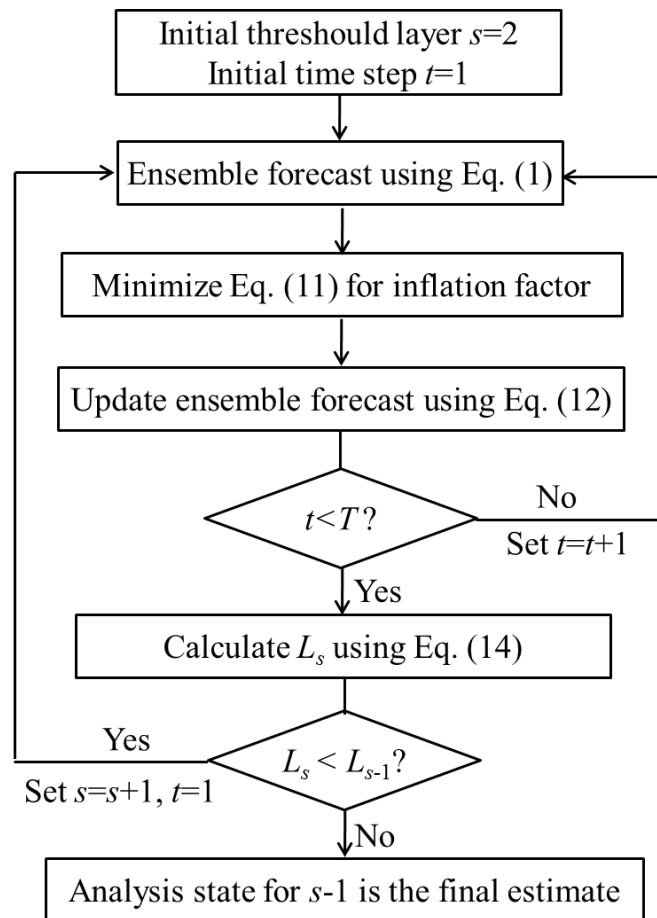
876

Figure captions

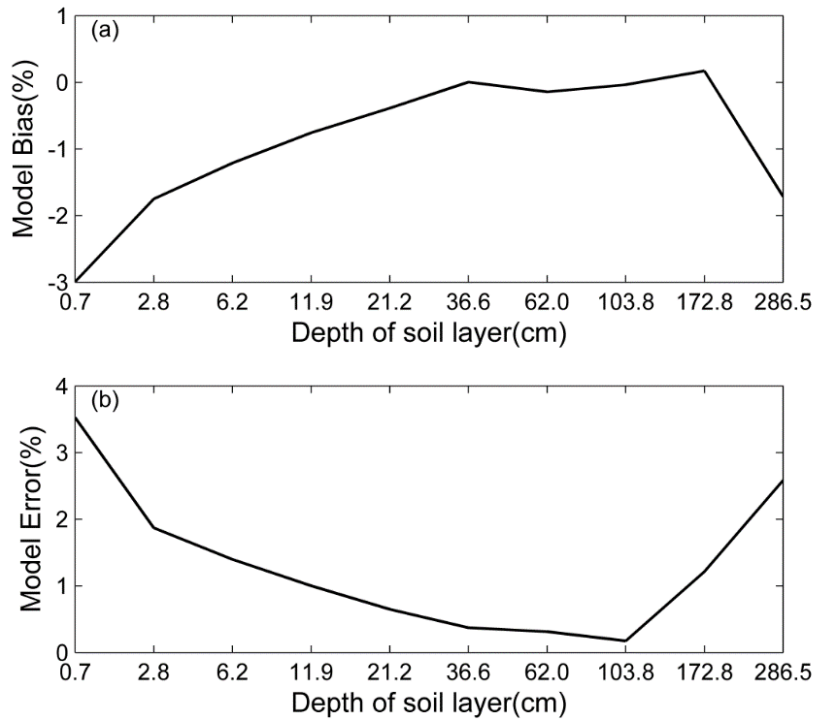
878

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

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

882

883

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

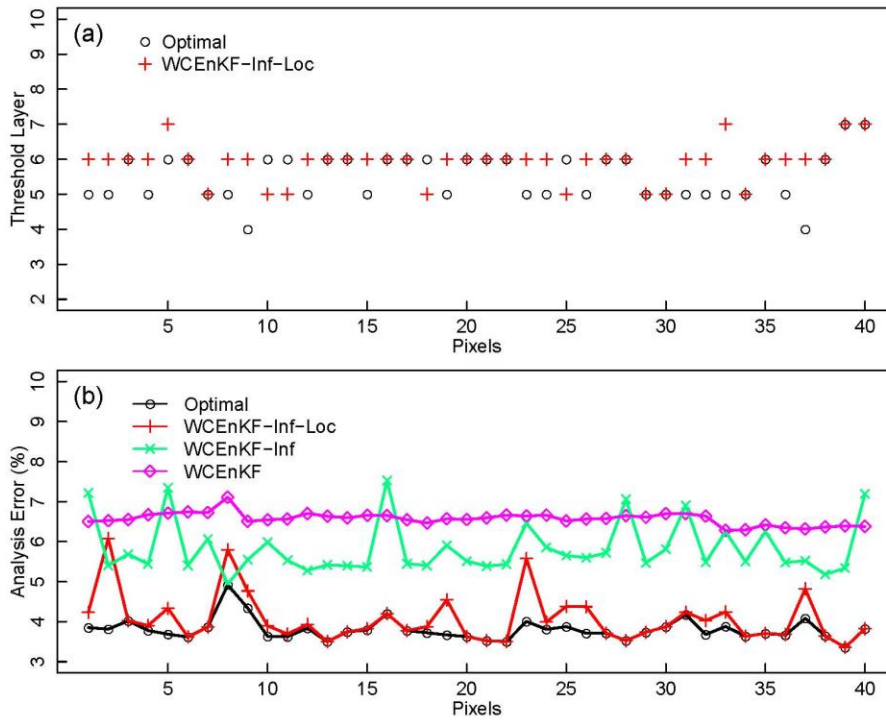
886 experiments. All of the equations are in accordance with that described in the text.

887

888

889 Figure 3. The areal average of the model's bias (a) and error (b) for one step in the soil

890 moisture content between the CoLM and the CLM 4.0. The horizontal axis represents

891 the layer depth.

892

893

894 Figure 4. The threshold layers and analysis error for each pixel in the synthetic

895 experiment. Graph (a) illustrates the optimal and WCEnKF-Inf-Loc threshold layers

896 of each pixel. Graph (b) shows the column RSME of each pixel in different schemes

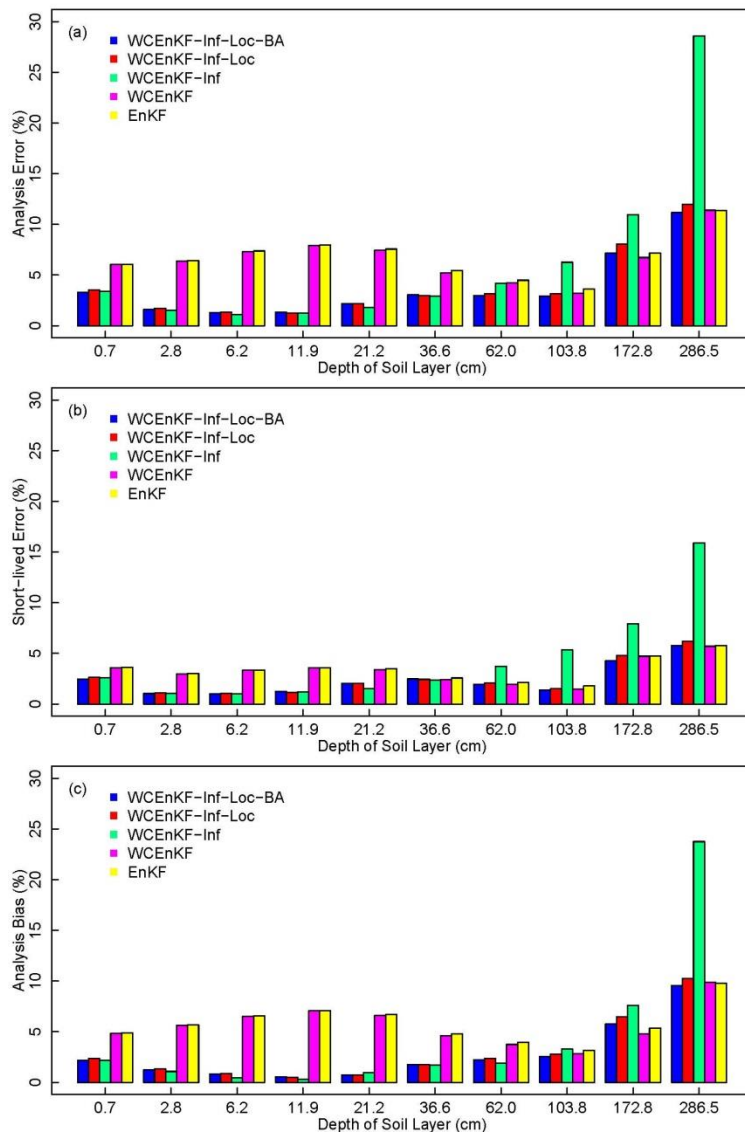
897 with water balance constraint (Optimal, WCEnKF-Inf-Loc, WCEnKF-Inf and

898 WCEnKF). The horizontal axes of (a) and (b) represent the 40 pixels in the study

899 domain.

900

901

902

903

904 Figure 5. The assimilation results in each layer for the five schemes: a weakly

905 constrained bias-aware ensemble Kalman filter with forecast error inflation and

906 vertical localization (WCEnKF-Inf-Loc-BA), a weakly constrained ensemble Kalman

907 filter with forecast error inflation and vertical localization (WCEnKF-Inf-Loc), a

908 weakly constrained ensemble Kalman filter with forecast error inflation

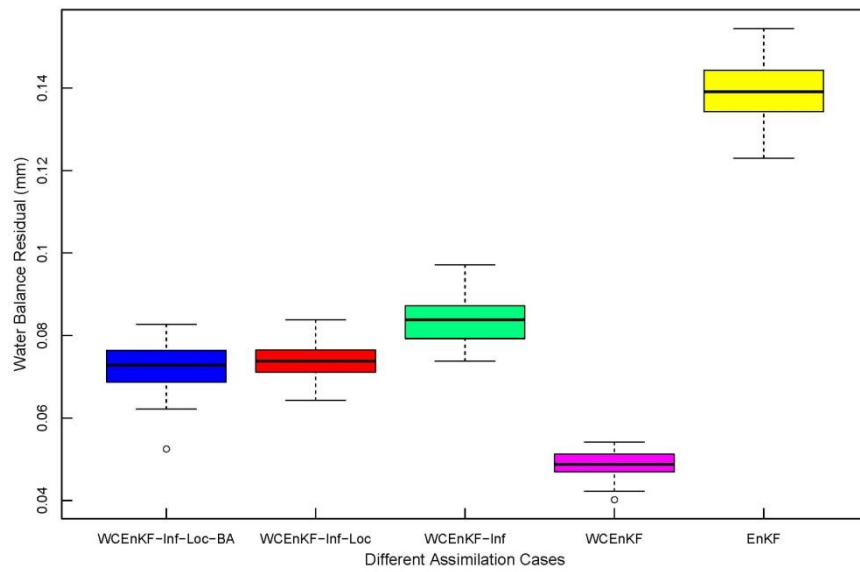
909 (WCEnKF-Inf), a weakly constrained ensemble Kalman filter (WCEnKF), and the

910 traditional assimilation (EnKF). Graphic (a) is for spatial averaged analysis error of

911 the soil moisture content, (b) is for the short-lived error and (c) is for the analysis bias.

912

913

914

915 Figure 6. The box plot of the water balance residual in all 40 pixels for the

916 WCEnKF-Inf-Loc-BA, WCEnKF-Inf-Loc,WCEnKF-Inf, WCEnKF and EnKF

917 assimilation schemes.

918

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

920

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

921

922

923

924   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

925