Journal of Hydrology 590 (2020) 125445

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

Sub-daily soil moisture estimate using dynamic Bayesian model averaging

Yong Chen[a,b], Huiling Yuan[a,*], Yize Yang[c], Ruochen Sun[d]

[a] School of Atmospheric Sciences and Key Laboratory of Mesoscale Severe Weather/Ministry of Education, Nanjing University, Nanjing, China
[b] Research Applications Laboratory, National Center for Atmospheric Research, Boulder, CO, USA
[c] Key Laboratory of Meteorological Disaster/Ministry of Education, Nanjing University of Information Science and Technology, Nanjing, China
[d] State Key Laboratory of Hydrology - Water Resources and Hydraulic Engineering, College of Hydrology and Water Resources, Hohai University, Nanjing, China

ARTICLE INFO

This manuscript was handled by Emmanouil Anagnostou, Editor-in-Chief, with the assistance of Viviana Maggioni, Associate Editor

Keywords:
Soil moisture
Dynamic Bayesian model averaging
Reanalysis
Land data assimilation system
CLDAS

ABSTRACT

Accurate estimation of soil moisture (SM) from satellite products and model simulations at sub-daily timescale remains a challenge. This study proposes a general dynamic Bayesian model averaging (BMA) framework for merging sub-daily model products. Compared to the traditional BMA method, this study introduces adaptive weights (dynamically variant with time) for BMA members. Based on the previous evaluation work, a subset of model products is selected from eight model products as BMA members. The dynamic BMA experiment is performed for the surface SM (0–10 cm) model products at sub-daily (6-h) timescale in 2017 over the Yangtze-Huaihe river basin. The results are compared with the automatic SM observations (ASMOs) with unprecedented high spatial and temporal resolution (up to 7 stations within a 10^4 km^2 pixel; hourly). Because weather pattern and model performance change over time, the determination of an optimal training period is critical to obtain adaptive BMA weights for rapid weather regime changes. The sensitivity of training length (days) is then examined, and the optimum data length used in the BMA training period proves to be about 80 days. With deterministic and probabilistic verification metrics, the dynamic BMA estimated SM is comprehensively evaluated against the ASMOs, eight global model products, and the CMA's (China Meteorological Administration) regional Land Data Assimilation System (CLDAS) product. To better compare the probability distribution of different products, the cumulative distribution function (CDF) consistency histogram and a more objective metric consistency deviation (CD) are proposed to diagnose the consistency of two SM CDFs (e.g., the BMA estimated and the observed CDF). In terms of both the deterministic (the Kling-Gupta efficiency, correlation, system bias, and bias adjusted root-mean square error) and probabilistic verification methods (CD, QQ-plots, and reliability), the dynamic BMA estimated SM outperforms any BMA members and even the benchmark product CLDAS. This study demonstrates that the dynamic BMA framework provides a new solution for merging SM model products. The merged SM and the BMA combined probability distribution can be further used for drought monitoring and prediction.

1. Introduction

As a critical state variable in Earth systems, soil moisture (SM) plays a significant role in hydrological prediction (Brocca et al., 2010b; James and Roulet, 2009; Koster et al., 2010), drought monitoring (Enenkel et al., 2016; Rahmani et al., 2016; Zhang et al., 2017), and weather forecast (de Rosnay et al., 2012, 2013; Dharssi et al., 2011; Zhong et al., 2020). Multiple SM data sources including in-situ monitoring networks, satellite remote sensing missions, and numerous modeling capabilities can provide local (point), regional or global SM estimates to meet the need of applications in hydrology and meteorology. However, the point measurement representativeness is limited to few square meters (Brocca et al., 2007; Penna et al., 2009; Susha Lekshmi et al., 2014). Although satellite products can monitor surface SM (usually 0–5 cm) at a larger spatial scale, they are suffered from limited spatial resolution along with retrieval errors (Brocca et al., 2011, 2010a; Wang and Qu, 2009). In recent years, satellite retrieved SM has been successfully applied to agriculture and drought monitoring (Abbaszadeh et al., 2019b; Entekhabi, 2010), while fine resolution model products can provide additional information for deeper soil layers. Owing to the atmospheric or/and land data assimilation system (LDAS), the SM products from numerical models can serve as alternatives to provide spatio-temporally continuous SM estimates (Robock et al., 2000; Sheffield, 2004).

* Corresponding author at: School of Atmospheric Sciences and Key Laboratory of Mesoscale Severe Weather/Ministry of Education, Nanjing University, Nanjing 210023, China.
E-mail address: yuanhl@nju.edu.cn (H. Yuan).

https://doi.org/10.1016/j.jhydrol.2020.125445
Received 17 May 2020; Received in revised form 4 August 2020; Accepted 20 August 2020
Available online 27 August 2020

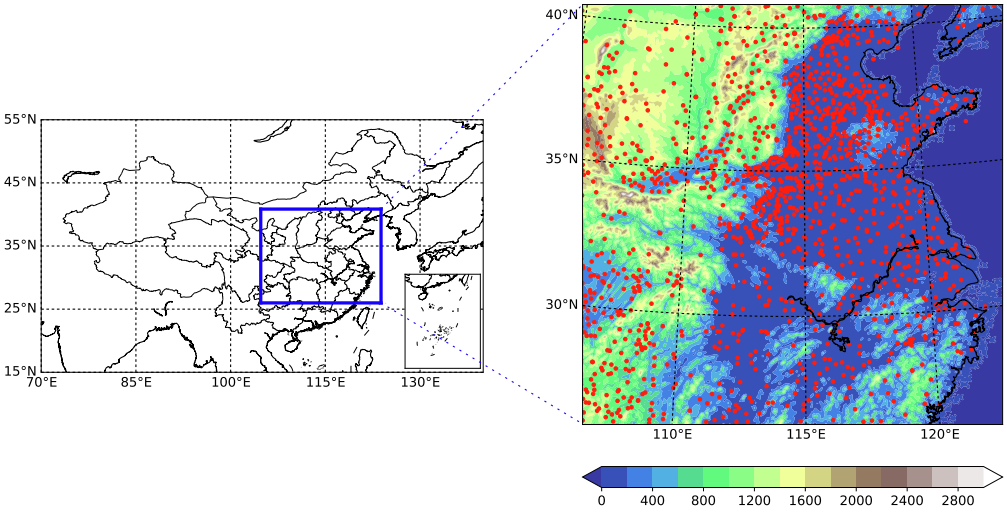

Fig. 1. Spatial distribution of in-situ soil moisture stations (red dots) over the Yangtze-Huaihe river basin. The shaded contour shows the terrain height (m). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The SM state obtained by land surface model (LSM) simulation is highly model-dependent (Koster et al., 2009), which means models could show large inconsistency with each other. Large differences can be found in the SM products generated by different LSM models (Chen and Yuan, 2020), including (re)analysis and LDAS products, even when the models are forced with the same meteorological dataset (Dirmeyer et al., 2006; Chen and Yuan, 2020). The inconsistency among the modeled SM is attributed to not only the model structure, the land data assimilation techniques, the quality of assimilated observations, but also the meteorological forcing, especially precipitation (Sakov and Sandery, 2017; Leach and Coulibaly, 2019). Furthermore, these SM model products could have significant systematic biases in their horizontal, temporal and/or vertical supports (Dirmeyer et al., 2004; Koster et al., 2009). Based on sub-daily (6-h) high-resolution in-situ observations in China, Chen and Yuan (2020) pointed out that different SM model products have variable accuracy and distinct performance in different climate regions. Each model has its unique strength and weakness when representing SM climatology (e.g., extremely dry and wet conditions). Therefore, one approach that leverages the advantages of available model products, possibly also integrating in-situ and satellite SM observations, is expected to be designed for hydro meteorological applications.

The model structural uncertainties can be addressed by Bayesian model averaging (BMA; Hoeting et al., 1999), Ensemble Model Output Statistics (E-MOS; Gneiting et al., 2005), Hydrological Uncertainty Processor (HUP; Krzysztofowicz and Kelly, 2000), and data assimilation approaches (Abbaszadeh et al., 2019a; Pathiraja et al., 2018). The E-MOS is based on multiple linear regression and the predictive mean is a weighted average of the ensemble members (usually bias-corrected). The HUP is based on based on a probabilistic quantitative precipitation forecast (PQPF) to quantify the hydrologic prediction uncertainty. The BMA is an ensemble post-processing approach that assigns different weights to individual probability density function (PDF) of each member over the training period. Therefore, it has the advantage in representing the distribution of a quantity of interest. The BMA weights reflect the members' relative contributions to predictive skill and thus can be used to assess the value of the candidate members. The BMA method was firstly proposed (Raftery et al., 2005) to calibrate surface temperature and sea level pressure of ensemble weather forecasts, and then has been successfully applied to improve forecast of precipitation (Sloughter et al., 2007), wind speed (Sloughter et al., 2013), and streamflow (Duan et al., 2007; Jiang et al., 2014; Roy et al., 2017; Sun et al., 2018; Yang et al., 2018). However, there is limited research on merging SM model products based on the BMA method to compensate the limitations of using an individual model and explore the value of available model products. Kim et al. (2015) adopted multiple sets of BMA weights for multi-model SM simulations under different wetness conditions (e.g., wet and dry conditions) to obtain an effective SM estimate. Applying BMA with varying weights in time or space (dynamic BMA) rather than static ones is possibly more suitable for identifying SM variability, because SM predictions in LSMs vary with antecedent wetness conditions (SM memory). Recent studies using dynamic BMA method are mainly focused on atmospheric variables (Ma et al., 2018; Raftery et al., 2005; Sloughter et al., 2013, 2007) and streamflow data (Duan et al., 2007; Vrugt et al., 2008) by training BMA weights for different categories of the verificaiton samples. It is of interest to apply dynamic BMA method (time varying) to merge sub-daily SM model data, considering the advantages of individual models under different weather regimes and climate scenarios.

However, challenges still exist in applying the BMA method into merging SM model products. On the one hand, calculating the weights for individual models requires SM observations within the region of interest. Due to the lack of in-situ observations, Kim et al. (2015) applied dynamic BMA method to three hydrologic model simulated SM at only two in-situ sites. The results were validated using daily in-situ SM observations at 0–5 cm for less than one month. Longer length of validation period and more in-situ sites are required to achieve more general and convincing conclusions, when applying the BMA method to SM model products in the domain of interest. On the other hand, while the variation of SM in models is largely dependent on antecedent SM condition and the meteorological forcing, it's unknown how long the training period is needed to obtain the optimum dynamic BMA weights. Using a longer training period is expected to better estimate the BMA weights (Raftery et al., 2005). However, weather pattern and model strengths change over time, which requires flow-dependent weights and a trade-off between the stable training samples and dynamic conditions. Therefore, the determination of an optimal training period is critical to obtain adaptive BMA weights for rapid weather regime changes. This study examines the sensitivity of training length (days) to obtain the best weights over the Yangtze-Huaihe river basin (Fig. 1) in China, which is suffered from frequent floods and droughts due to the hydrological impacts of precipitation extremes (Yang et al., 2016). The automatic SM observations (ASMO) with unprecedented high spatial and temporal resolution (up to 7 stations within a 10^4 km^2 pixel; hourly) in China (Chen and Yuan, 2020) are used for training and validation of the dynamic BMA.

The first purpose of this paper is therefore to design a general dynamic BMA framework for merging sub-daily SM model products. Four

Table 1
Overview of the eight SM model products (1–8) and CLDAS (9). Products (4), (5), (6), and (8) are selected as the BMA members.

Product (producer)	Spatial resolution	Temporal resolution	Soil depths (cm)	Reference
(1) GFS (NCEP)	0.5°, global	6-hourly, 2007-present	0–10, 10–40, 40–100, 100–200	Environmental Modeling Center (2016)
(2) NCEP2 (NCEP)	2.5°, global	6-hourly, 1979-present	0–10, 10–200	Kanamitsu et al. (2002)
(3) ERA Interim (ECMWF)	0.75°, global	6-hourly, 1979-present	0–7, 7–28, 28–100, 100–255	Dee et al. (2011)
(4) ERA5 (ECMWF)	~31 km, global	1-hourly,1979-present	0–7, 7–28, 28–100, 100–255	Copernicus Climate Change Service (2017)
(5) GLDAS Noah (NASA)	0.25°, global	3-hourly, 2000-present	0–10, 10–40, 40–100, 100–200	Rodell et al. (2004)
(6) GLDAS CLM (NASA)	1°, global	3-hourly, 1979-present	0–1.8, 1.8–4.5, 4.5–9.1, 9.1–16.6, 16.6–28.9, 28.9–49.3, 49.3–82.9, 82.9–138.3, 138.3–229.6, 229.6–343.3	Rodell et al. (2004)
(7) GLDAS VIC (NASA)	1°, global	3-hourly, 1979-present	0–10, 10–160, 160–190	Rodell et al. (2004)
(8) GLDAS Mosaic (NASA)	1°, global	3-hourly, 1979-present	0–2, 2–150, 150–350	Rodell et al. (2004)
(9) CLDAS (CMA)	0.0625°, regional	1-hourly, 2008-present	0–5, 0–10, 10–40, 40–100, 100–200	Shi et al. (2014)

(re)analysis products and four LDAS products have been evaluated by Chen and Yuan (2020) and their performance are reported to vary over space and time. Therefore, the selection of model products could impact the performance of dynamic BMA estimates, which means whether merging all model products or a subset of those products could achieve better performance remains unknown. The CMA's (China Meteorological Administration) Land Data Assimilation System (CLDAS; Table 1) obtains the best performance over most climate regions in China, especially over the Yangtze-Huaihe river basin. However, this product has the shortest data record which starts from 2008 compared to other products (Table 1), which limits the scientific research that requires long-term SM data. It is of interest to explore whether the dynamic BMA method can generate deterministic forecasts better than or comparable to CLDAS based on model products with long-term record. The second purpose is to apply dynamic BMA method to sub-daily probabilistic forecast for drought events, which is very useful for drought monitoring and forecast. Drought monitoring and forecast are currently relied on drought indicators, which are usually computed by statistical methods or dynamical model simulations (DeChant and Moradkhani, 2014; Kumar et al., 2014; Mao et al., 2015; Wood and Lettenmaier, 2008; Zhu et al., 2019). Statistical methods could make up the deficiency of using individual dynamic model simulation which can exhibit high uncertainty in drought prediction (Madadgar and Moradkhani, 2013). It has been demonstrated that probabilistic drought forecasting methods can have good seasonal drought forecasting skill (Chen et al., 2015; Madadgar and Moradkhani, 2014b; Mishra and Desai, 2005). The main advantage lies in its probabilistic features in analyzing future droughts.

In this study, dynamic Bayesian framework is proposed to merge a subset of SM products based on the evaluation by Chen and Yuan (2020) and additional test experiments. Compared to traditional BMA method, the adaptive weights (dynamically variant with time) for BMA members are introduced. This is to accommodate weather regime and soil condition changes. The dynamic BMA method proves to be the most robust in terms of deterministic and probabilistic verifications, and the results is even better than the benchmark product CLDAS. This paper is organized as follows. Section 2 describes the study region and gives a brief introduction to the data. Section 3 provides the details of designed dynamic BMA framework, and then describes the evaluation metrics. Results and conclusions can be found in Section 4 and 5 respectively.

2. The study region and data

2.1. Study region

This study focuses on the Yangtze-Huaihe river basin (Fig. 1), located in the east of China. It is one of the regions with the most agricultural production and highest population density in China. The occurrence of drought and flood disasters in this region is a serious hazard to agricultural production. The resultant economic losses can account for 39% of the total losses caused by natural disasters. Due to the rapid economic development of this region, droughts and floods have caused greater and greater losses to people's lives, property and economics. Therefore, it is of great significance to strengthen the research on drought and flood over the Yangtze-Huaihe river basin.

2.2. Hourly in situ observations

This research uses the ASMOs provided by CMA National Meteorological Information Center (NMIC). Due to the data management policy in China, the original ASMO data at 2776 stations are not available via public repository. However, anyone who would like to obtain these data could contact CMA/NMIC (http://data.cma.cn/en) for detailed information of data acquisition. There are 1812 observation stations within the study region (Fig. 1) and the density of observations is estimated 6.29 stations/10^4 km^2. The network of ASMOs is shown in Fig. 1. SM is observed hourly at the depths of 0–10 cm, 10–20 cm, 20–30 cm, 30–40 cm, 40–50 cm, 50–60 cm, 60–80 cm and 80–100 cm. Only the ASMO data at 0–10 cm soil layer are available in this study. The ASMO data are subject to strict quality control (Chen and Yuan, 2020).

2.3. Sub-daily model products

Considering consistent spatial resolution, high temporal resolution, and long-term records, eight sub-daily model products including four (re)analysis products (product 1–4; Table 1) and four LDAS products (product 5–8; Table 1) are compared with dynamic BMA results. Four of them (product 4, 5, 6, and 8) are selected as dynamic BMA members based on the evaluation work by Chen and Yuan (2020). Additional test experiments including using all global model products (1–8) and using a subset of BMA products in 2016 has been conducted to select the best dynamic BMA members. The selected members show the best KGE performance in 2016, and then they are further used as BMA members in 2017. The eight model products are released by three different agencies in the world, and adopts different LSM models to generate SM products. Large inconsistencies are found among these model products, especially in winter (Chen and Yuan, 2020). The CLDAS (product 9; Table 1) which generally obtains the best performance over the Yangtze-Huaihe river basin is treated as a benchmark. An overview of these products can be found in Table 1 and detailed introductions are as follows.

2.3.1. Analysis and reanalysis products

The Global Forecast System (GFS; Environmental Modeling Center, 2016) is a global weather forecast model produced by the National Centers for Environmental Prediction (NCEP). It's a coupled model using Noah LSM (Chen and Dudhia, 2001) with four vertical soil layers (Table 1) to represent land-surface processes. The 0.5° gridded GFS analysis with a temporal resolution of 6 h is obtained from NOAA's National Centers for Environmental Information (https://www.ncdc.noaa.gov/data-access/model-data/model-datasets/global-forcast-system-gfs).

The NCEP/DOE Reanalysis 2 (NCEP2; https://rda.ucar.edu/

datasets/ds091.0) is the first-generation (Kanamitsu et al., 2002) global reanalysis developed by NCEP, which fixes the known errors (Kanamitsu et al., 2002) of NCEP/DOE Reanalysis 1 (NCEP1). Both NCEP1 and NCEP2 adopt the Oregon State University (OSU) LSM (Pan and Mahrt, 1987) as its land surface component with two layer thicknesses of 10 cm and 190 cm (Table 1). The spatial resolution is 2.5° and the temporal resolution is 6 h. The NCEP2 data are available from 1979 to present.

ERA-Interim (Dee et al., 2011) is a third-generation global atmospheric reanalysis produced by the European Centre for Medium-Range Weather Forecasts (ECMWF). It stopped being produced on 31 August 2019 and has been replaced by ERA5. ERA5 is the fourth-generation reanalysis (Copernicus Climate Change Service, 2017), which is the latest high-resolution reanalysis also released by ECMWF. Compared to ERA Interim, ERA5 assimilates more historical observations and provides finer temporal (1 h vs. 6 h) and spatial (~31 km vs. ~ 75 km) estimates for atmospheric, land, and oceanic states. The ECMWF datasets can be downloaded from https://www.ecmwf.int/en/forecasts/datasets/browse-reanalysis-datasets.

2.3.2. LDAS products

All the LDAS products used as dynamic BMA members are from the Global Land Data Assimilation System (GLDAS; Rodell et al., 2004) jointly developed by the National Aeronautics and Space Administration (NASA) Goddard Space Flight Center (GSFC) and NOAA/NCEP. The NASA Goddard Earth Sciences Data and Information Services Center (https://disc.gsfc.nasa.gov/datasets?keywords = GLDAS) provides access to LDAS datasets. The differences among the four GLDAS products are mainly the model parameterizations schemes and the forcing data (Rodell et al., 2004). The spatial resolution for GLDAS CLM, GLDAS VIC and GLDAS Mosaic from GLDAS version 1.0 (GLDAS-1) is 1°, while the GLDAS Noah from GLDAS version 2.0 (GLDAS-2) has a higher spatial resolution of 0.25°. All the GLDAS products are at a temporal resolution of 3 h.

CLDAS (Shi et al., 2014) is a spatially and temporally high resolution regional LDAS product, which obtains the best performance in most of the regions over China (Chen and Yuan, 2020), especially over the Yangtze-Huaihe river basin. It is therefore used as a benchmark in this study. The CLDAS product is developed by CMA and has been operationally run by CMA/NMIC since July 2013, and is available from 2008 to present. It assimilates regional hourly surface temperature observations and rain gauge data (Shen et al., 2014), covering the Asian region (0–65°N, 60–160°E). CLDAS has the highest spatial resolution of 0.0625° and temporal resolution of 1 h among the nine model product listed in Table 1. CLDAS data are shared by the China Meteorological Data Service Center (http://data.cma.cn/en).

3. Methods

3.1. Dynamic Bayesian model averaging

BMA is a statistical post-processing approach that can combine inferences and forecasts based on multiple competing models (Raftery et al., 2005) to get a more skillful and reliable probabilistic ensemble. In this study, the dynamic BMA algorithm (with time-varying weights) is used to improve the estimation of surface SM (at the depth of 0–10 cm) by adjusting the PDF to obtain a good fitness to the ASMOs. Compared to the traditional BMA approach, the dynamic BMA method designed in this study assigns dynamic weights for the model products. The dynamic here means that the weight varies with time in order to make the dynamic BMA algorithm adapt to the change of weather pattern and antecedent soil condition. The dynamic BMA approach is briefly described in the following.

The BMA method gives a combined PDF of the individual model products. The BMA predictive PDF is given by:

$$p(y|f_1,\cdots,f_N) = \sum_{i=1}^{N} p(f_i|O)p_i(y|f_i, O) \quad (1)$$

where y indicates the merged SM; N is the number of BMA members, i.e., the number of model products to be merged in this study; $O = [y_1^{obs}, y_2^{obs}, \ldots, y_m^{obs}]$ denotes the observed y in the training period with the length of m; f_i (i = 1, …, 8) indicates the SM value of ith model product; $p(f_i|O)$ denotes the posterior probability of f_i, also known as the likelihood of ith model product; $p(y|f_i,O)$ is the posterior distribution of y given SM f_i estimated by model products, and observed SM, O. The likelihood of ith model product $p(f_i|O)$ actually reflects the products' relative contribution to BMA predictive skill over the training period, and can be denoted as a weight parameter, w_i. The posterior model probabilities $p(f_i|O)$ are nonnegative and add up to one, so that $\sum_{i=1}^{N} w_i = 1$. For dynamic BMA approach used for merging SM model products in this study, the w_i varies with time (a function of time, T) considering the change of weather patterns and soil memory. Thus, for the dynamic BMA approach, Eq. (1) could be expressed as:

$$p[y(T)|f_1(T),\cdots,f_N(T)] = \sum_{1}^{N} w_i(T)p_i[y(T)|f_i(T), O(T)] \quad (2)$$

The conditional PDF $p(y|f_i,O)$ is approximated by normal distribution with mean $\bar{f_i}$ and standard deviation σ_i. Using normal distribution could be inappropriate for SM, because it is mainly driven by precipitation in LSMs which has a skewed distribution (Sloughter et al., 2007). Gamma distribution seems more reasonable to represent the PDF of SM. However, after testing both normal and gamma distribution, normal distribution is found to improve more the skill of dynamic BMA predicted SM. This corresponds with Kim et al. (2015) and Vrugt and Robinson (2007) who also found the assumption of normality exhibits better predictive skill. In this case, the situation can be denoted as:

$$y|f_i \sim N(a_i + b_i f_i, \sigma_i^2) \quad (3)$$

where a_i and b_i is estimated by simple linear regression of SM observation O on model product f_i for the training data. This can be viewed as a simple bias-correction process (Raftery et al., 2005). The standard deviation σ_i and BMA weights w_i are estimated using the expectation–maximization (EM) algorithm (Dempster et al., 1977; Raftery et al., 2005). The posterior mean (E) and variance (Var) of the BMA prediction (y) are then expressed as:

$$E(y|f_1, ..f_N) = E\left[\sum_{i=1}^{N} w_i p_i(y|f_i, O)\right] = \sum_{i=1}^{N} w_i(a_i + b_i f_i) \quad (4)$$

$$Var(y|f_1, ..f_N) = \sum_{i=1}^{N} w_i\left[(a_i + b_i f_i) - \sum_{i=1}^{N} w_i(a_i + b_i f_i)\right]^2 + \sum_{i=1}^{N} w_i\sigma_i^2 \quad (5)$$

3.2. Experimental design

The general framework of dynamic BMA algorithm for merging SM model products is shown in Fig. 2. The "dynamic" primarily refers to the BMA weights varying with time T, which assumes that the model performance and the soil condition could vary temporally.

The first step is to generate the training and validation datasets. The training period is a moving window with the length of 80 days preceding every validation period (Fig. 2b). As mentioned before, the length of training period could affect the BMA predictive skill. It's unknown how many days should be used for the training datasets to achieve the most robust multi-model merged SM estimate. For time T at every 6 h in 2017, model products during the period 80 days preceding time T are treated as training datasets (Fig. 2b). There is no automatic way to determine the length of training period. Section 4.1 will further discuss how to determine the appropriate days used in the training period. Only 6-h model data are used in this study, because the coarsest temporal resolution of model products is 6 h. For both the training and

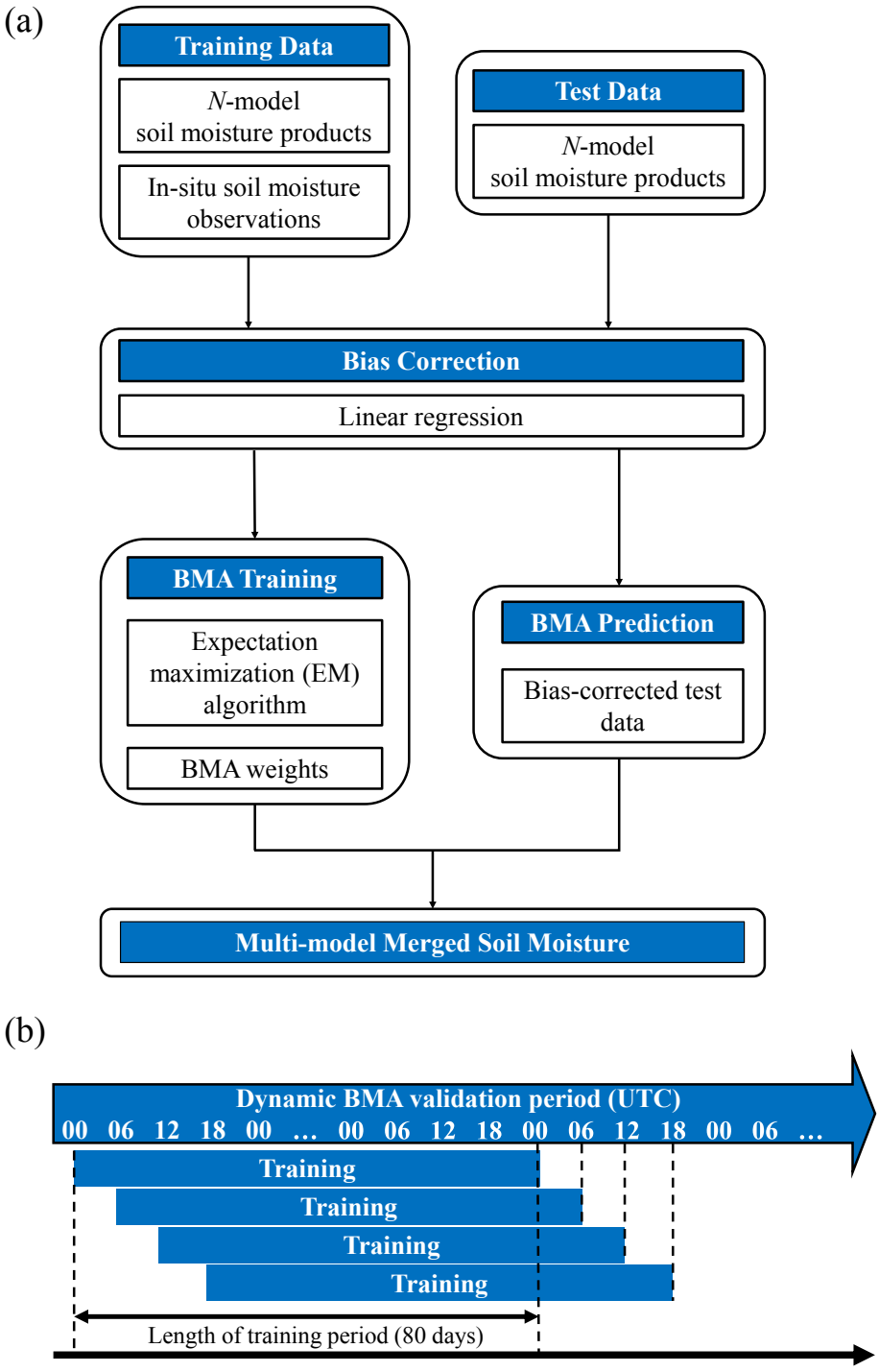

Fig. 2. General framework of the dynamic Bayesian model averaging algorithm for merging *N* -model soil moisture. The training period is a moving window with the length of 80 days preceding each validation time (at every 6 h in this study).

validation process, model products are horizontally interpolated to in-situ stations using bilinear interpolation and vertically interpolated to 0–10 cm to match the ASMO depth.

Second, the model products in training period are bias-corrected using linear regression. In this procedure, a_i and b_i described in Section 3.1 are estimated using data in the training period and then applied to validation datasets at time T to remove the system biases (Fig. 2b). That is because all the model products are characterized by different bias features and model assumptions that can cause a very different range of variability of the SM estimates.

Then the BMA weights for the four model products (4, 5, 6, and 8; Table 1) are fitted using the training datasets. As mentioned in Section 3.1, BMA weights w_i (i = 1, 2, 3, 4) are estimated using the expectation–maximization (EM) algorithm (Dempster et al., 1977; Raftery et al., 2005). The estimated BMA weights at each time step are further applied to the multi-model SM products over the whole study domain. Next, the BMA merged SM in 2017 is evaluated using both deterministic and probabilistic verification methods in Section 3.3.

The example experiment over the Yangtze-Huaihe river basin in is presented in 2017 during which the CLDAS product is actually available. However, the dynamic BMA estimates is designed to serve as an alternative when CLDAS data are not available (i.e., years out of 2008-

present). The assumption is that only multiple model products and observations except for CLDAS are available in 2017 (during which the CLDAS data is actually available). The selection of year 2017 is to compare dynamic BMA results with the benchmark product CLDAS.

3.3. Evaluation metrics

To comprehensively evaluate the BMA estimate against the nine model products, especially CLDAS, both deterministic and probabilistic verification methods are used. The deterministic verification metrics include Kling-Gupta efficiency (KGE; Gupta et al., 2009; Kling et al., 2012), Pearson's correlation coefficient (CC), systematic bias (BIAS), and adjusted root mean square error (aRMSE). The probabilistic verification methods include cumulative distribution function (CDF) consistency histogram (CCH), quantile-quantile (QQ; Wilk and Gnanadesikan, 1968) plot with the corresponding quantified metric π_{reliab}, and Brier skill scores (BSS).

KGE is an objective performance metric that combines correlation, bias and variability. The definition of KGE metric is as follows:

$$KGE = 1 - \sqrt{(CC-1)^2 + (\beta-1)^2 + (\gamma-1)^2}, \tag{6}$$

where CC is the Pearson's correlation coefficient, β indicates the ratio of estimated and observed means (i.e., frequency bias), and γ is represented by the ratio of the estimated and observed coefficients of variation. KGE is calculated only if CC passes the 0.05 significance test. The β and γ given by:

$$\beta = \frac{\mu_e}{\mu_o}, \tag{7}$$

$$\gamma = \frac{\sigma_e/\mu_e}{\sigma_o/\mu_o}, \tag{8}$$

where μ and σ indicate the mean and the standard deviation, respectively. The subscripts e and o are the estimate and observation, respectively. Therefore, KGE is perfect at unity, and smaller KGE represents worse performance.

The formulas of the other three metrics, including correlation (CC), system bias (BIAS), and bias adjusted root mean square error (aRMSE), are given by:

$$\mathrm{CC} = \frac{\sum_{i=1}^{n}(O_i-\bar{O})(E_i-\bar{E})}{\sqrt{\sum_{i=1}^{n}(O_i-\bar{O})^2}\sqrt{\sum_{i=1}^{n}(E_i-\bar{E})^2}} \tag{9}$$

$$\mathrm{BIAS} = \frac{1}{n}\sum_{i=1}^{N}(E_i - O_i) \tag{10}$$

$$\mathrm{aRMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[(E_i-\bar{E}) - (O_i-\bar{O})]^2} \tag{11}$$

where n is the number of observations; E_i and O_i are the ith pairs of SM estimate and observation. Their corresponding mean values are expressed as $\bar{E}$ and $\bar{O}$, respectively. BIAS represents the system bias, while aRMSE is used to represent the random errors.

For probabilistic verification, the CDF consistency histogram is designed to diagnose the consistency of two CDFs. CCH shows the distribution of PIT values over equally sized bins. The PIT value is defined as:

$$\mathrm{PIT} = \mathrm{F}(O_i) \tag{12}$$

where O_i is the ith observation, F is the CDF of SM estimates. The interval [0,1] is divided into 10 equally sized bins. The CCH is approximately flat for a perfect SM estimate with equal bin frequency of $\frac{1}{m}$, where m is the number of bins. When the CCH is not flat, its shape can reflect problems within the SM estimate. As for a U-shape CCH, it indicates the estimated PDF has inadequate spread or underdispersion, which means the SM is usually overestimated. Oppositely, a humpback-shape CCH indicates the estimated PDF has overdispersion, which means the SM is usually underestimated. Note that a flat CCH is not a sufficient condition for a reliable SM estimate, because a combination of positively and negatively biased SM estimate can also yield a flat CCH.

The metric consistency deviation (CD) is then proposed to measure the degree of deviation from a flat CCH, and can be viewed as a more objective metric than CCH. The formula can be expressed as:

$$\mathrm{CD} = \frac{2m-2}{m}\sum_{i=1}^{m}\left|bin_i - \frac{1}{m}\right| \tag{13}$$

where bin_i the bin frequency of the ith bin. The CD is normalized and ranges from 0 (optimum) and 1 (worst).

QQ plot is a graphical method to compare the probability distribution of estimated SM with observations. In the QQ plot, the set of CDF values of observed SM within the estimated distribution is compared to the cumulative uniform distribution, U[0, 1]. If the two distributions are similar, the curve matches the diagonal line. Considering the difference between the diagonal line and QQ-plot curve, the quantitative reliability metric $\pi_{reliab}$is then derived as:

$$\pi_{reliab} = \frac{2}{n}\sum_{i=1}^{n}|F_U - F_E(O_i)| \tag{14}$$

Where F_U is the uniform CDF and $F_E(O_i)$ indicates the estimated CDF value for ith observation O_i. The range of π_{reliab} is [0,1] with the perfect value of 0 and the worst value of 1.

The BSS measures the accuracy of SM probabilistic estimate relative to a SM climatological estimate given a threshold. The BSS is defined as:

$$\mathrm{BSS} = 1 - \frac{BS_E}{BS_{Climate}} \tag{15}$$

where BS_E and $BS_{Climate}$ are the Brier score (BS) of the SM probabilistic estimate and the climatological estimate, respectively. The BS is calculated by:

$$\mathrm{BS} = \frac{1}{n}\sum_{i=1}^{n}(p_i - o_i)^2 \tag{16}$$

where p_i is the estimated probability of ith event, and o_i is the observed probability. The BSS score compares the BS to the reference BS (such as climatological estimate). If BSS equals 0, the SM estimate has the comparable performance (no skill to the reference) to the climatology. A positive (negative) BSS indicates the SM estimate is more (less) skillful than the climatological estimate.

4. Results

4.1. Length of training period

To address the appropriate length of training days used in training period, KGE, CC, BIAS, and aRMSE are computed for the dynamic BMA estimates using a set of training lengths, 10, 20, …, 120 days. The training period is a sliding window before time T (Fig. 2). The assessment of dynamic BMA estimates is performed using SM data from 2016 and the appropriate training length will then be applied to SM data from 2017.

Making the length of training days longer than 80 days will add little improvement to the dynamic BMA estimated SM (Fig. 3). Fig. 3 presents KGE, CC, BIAS, and aRMSE against different training days used in training period. These performance statistics show little variation if the days used in training period is more than 80 days. When setting training period as 80 days, KGE, BIAS, and aRMSE are the best among the performance of SM estimates using the twelve training lengths ranging from 10 to 120 days (Fig. 3). CC is a little lower than the performance of SM estimates using 100-day training length (Fig. 3). As a result, the appropriate length of days used in training period is ~80 days. The use of 80-day training period is then applied to the SM

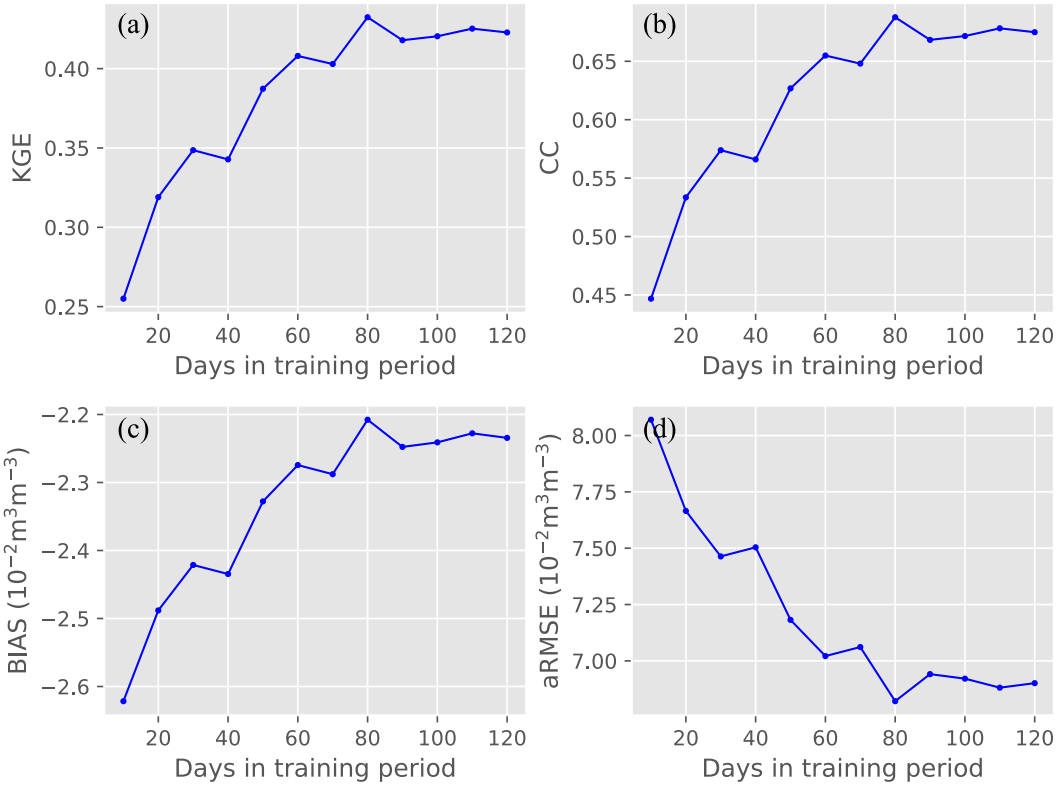

Fig. 3. Comparison of training period lengths: (a) KGE, (b) CC, (c) BIAS, and (d) aRMSE of BMA estimated SM.

datasets of 2017, and the resultant dynamic BMA estimates are further evaluated using both deterministic and probabilistic verification methods.

4.2. Deterministic verification

Dynamic BMA is applied to SM model products (product 4, 5, 6, and 8; Table 1) at every 6 h (denoted as *T*) in 2017 over the Yangtze-Huaihe river basin. The training period is a sliding window of 80-days preceding time *T* based on the result in Section 4.1. The dynamic BMA weights are estimated based on ASMOs and model data in the training period, and then applied to time *T* for the whole study domain. The BMA multi-model merged SM estimates are produced for ASMO locations.

Fig. 4 presents the resultant BMA weights at each time *T* in 2017 for the four BMA members (product 4, 5, 6, and 8; Table 1). The BMA weights can vary a lot over time, which indicates the relative contribution of BMA members to the skill of SM estimates is not static. In winter (December to February), GLDAS CLM, GLDAS Noah, and GLDAS Mosaic show relatively larger BMA weights than ERA5. In summer (June to August) a smaller weight which is mostly less than 0.1 is assigned to GLDAS Mosaic. ERA5 has a very small weight mostly less than 0.2 from November to March, but could be assigned larger weights than 0.25 for the other months. GLDAS Mosaic has relatively larger weights from March to early June and in November and December. In May and June, this region is frequently suffered from severe drought events while the models usually show positive biases. In November and December, this region receives much less precipitation than other months, which could also lead to the dry condition of the soil. Therefore, higher weights for GLDAS Mosaic could reduce the systematic biases of BMA prediction, especially during dry period. Including a model product with negative biases as a BMA member could make up the deficiency of other models in representing dry conditions. This demonstrates that using a dynamic BMA with a varying weight over time for SM model products is necessary.

The dynamic BMA estimated SM is then compared with the benchmark product CLDAS. Fig. 5 shows the mean sub-daily (6-h) SM of CLDAS, dynamic BMA estimate and ASMO observations. Both CLDAS and BMA well captures the variation of observed SM. CLDAS overestimates the mean SM, especially in summer (June to August), while BMA shows smaller bias for mean SM than CLDAS and underestimates the mean SM a little. Obviously, the mean SM of BMA is comparable to that of CLDAS in winter and better than CLDAS in summer with much smaller biases.

To comprehensively compare the BMA estimated SM with CLDAS, the four deterministic metrics including KGE, CC, BIAS, and aRMSE are calculated at each ASMO stations (Fig. 6). This is to evaluate their performance in capturing the SM variability over time. Fig. 6 presents the spatial distribution of the four evaluation metrics for BMA estimated SM and CLDAS using the sub-daily (6-h) ASMOs as a reference. BMA exhibits comparable KGE, CC, and aRMSE to CLDAS, while the BMA better represents the magnitude of observed SM in terms of BIAS. This corresponds with the results of mean SM that CLDAS underestimates the SM with a smaller negative bias than CLDAS.

Fig. 7 shows the time series of KGE, CC, BIAS, and aRMSE calculated at each time *T* for the spatial distribution of dynamic BMA estimated SM and CLDAS. This shows their capability in capturing the spatial variability. The BMA have better KGE, CC, BIAS, and aRMSE than CLDAS throughout the year. The time series of BIAS shows that CLDAS generally has positive biases larger than 0.02 m^3m^{-3} throughout the year, while BMA has a relatively smaller and more stable biases ~0.01 m^3m^{-3}. The stable biases for BMA can be easily corrected. Therefore BMA has stronger capability in representing the spatial variability of SM than CLDAS. An overall evaluation using all samples of 6-h ASMOs is presented in Table 2. The KGE, CC, BIAS, and aRMSE are calculated for all the model products and the dynamic BMA estimated SM. The multi-model merged SM with dynamic BMA algorithm has the best KGE, CC, BIAS, and aRMSE among the nine model products.

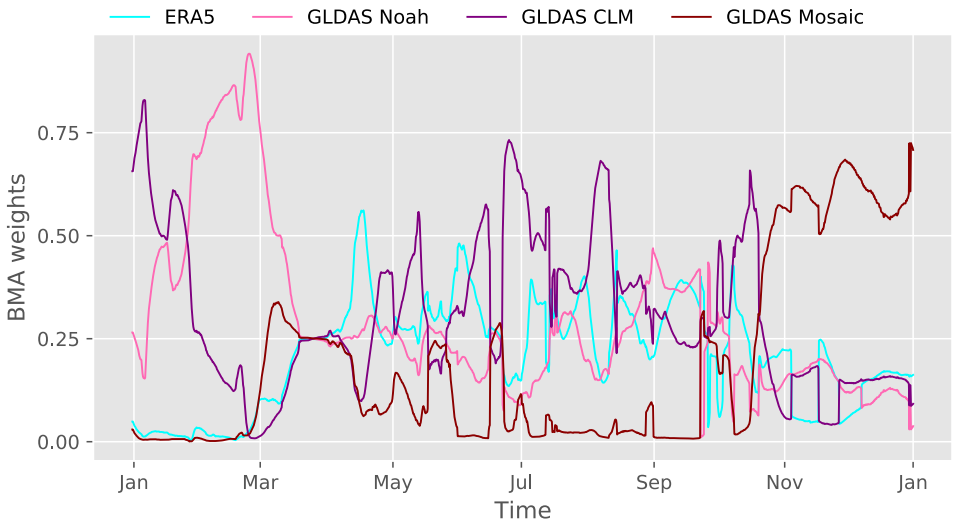

Fig. 4. BMA weights over the 80-day training period preceding each validation time in 2017.

4.3. Probabilistic verification

Fig. 8 presents the PDF of sub-daily (6-h) SM for observations, CLDAS, dynamic BMA estimated SM, and BMA members. The PDFs of BMA members are quite different with each other and none of the members well match the PDF of observations. The PDF of BMA members highlights the problem of their capability to represent extreme events. For example, all the BMA members underestimate the probability for SM less than 0.08 $m^3\ m^{-3}$, and overestimate the probability for SM between 0.2 and 0.4 $m^3\ m^{-3}$. Compared to CLDAS, BMA shows larger probability density for low SM values and more close to the probability density of observations. Therefore, multi-model merged SM using dynamic BMA might have better skill in capturing extreme drought events.

To quantitatively assess the performance of model products and BMA estimates in representing the probability distribution of observed SM, CDF consistency histogram are provided in Fig. 9. The CD values are calculated by Eq. (13). On each subplot, the reference line (dashed) of a perfectly flatted histogram. Fig. 9 clearly illustrates that most BMA members are highly unreliable in terms of the CDF consistency histogram. After applying dynamic BMA, the multi-model merged SM estimates effectively improves the reliability with the lowest CD of 0.100 (Fig. 9) and the lowest aRMSE of 0.083 $m^3\ m^{-3}$ (Table 2).

The QQ plot curve for dynamic BMA estimated SM is the closest to the diagonal line, indicating its SM estimate is reliable (Fig. 10). It obtains the best π_{reliab} value of 0.076 among the nine model products (Table 2), while CLDAS shows has a much worse π_{reliab} value of 0.202. The QQ plot curves for the four dynamic BMA members are quite diverse, and mostly deviates from the diagonal line a lot. GLDAS Mosiac has the worst reliability for representing the observed SM distribution (Fig. 10; Table 2) among the four products, but it could be assigned a relatively large weight for the dry periods (e.g., November and December) in the year (Fig. 4). This means that the weights of BMA members are not necessarily showing their skill in representing the SM distribution, but they could represent their relative contribution to the skill of BMA estimated SM.

4.4. Case study for a drought event in May 2017

Drought is a widespread meteorological disaster in China, the economic losses caused by which over the Yangtze-Huaihe river basin often ranks first among various natural disasters. Since China is located in a monsoon climate zone, precipitation mainly occurs in summer. Spring is therefore a transitional season in which the transition from winter

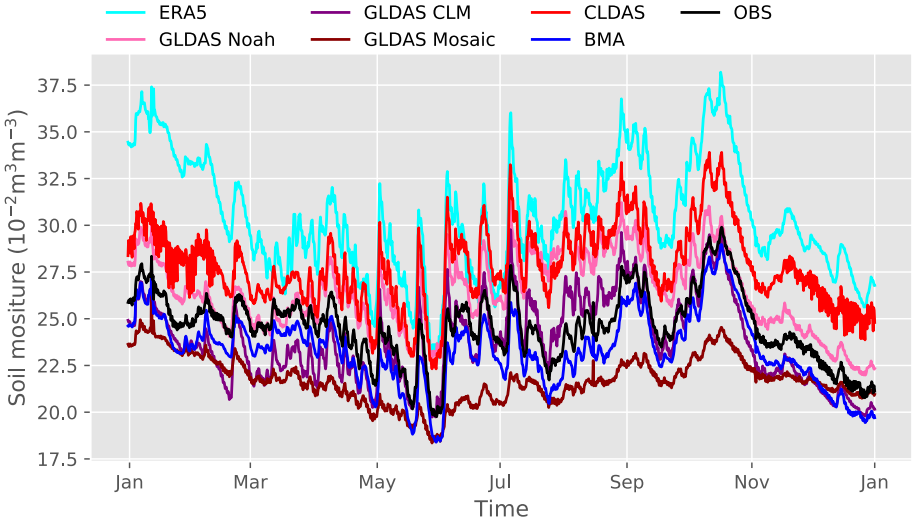

Fig. 5. Mean sub-daily (6-h) SM of CLDAS (red), dynamic BMA estimate (blue), the four BMA members, and observations (black) in 2017. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

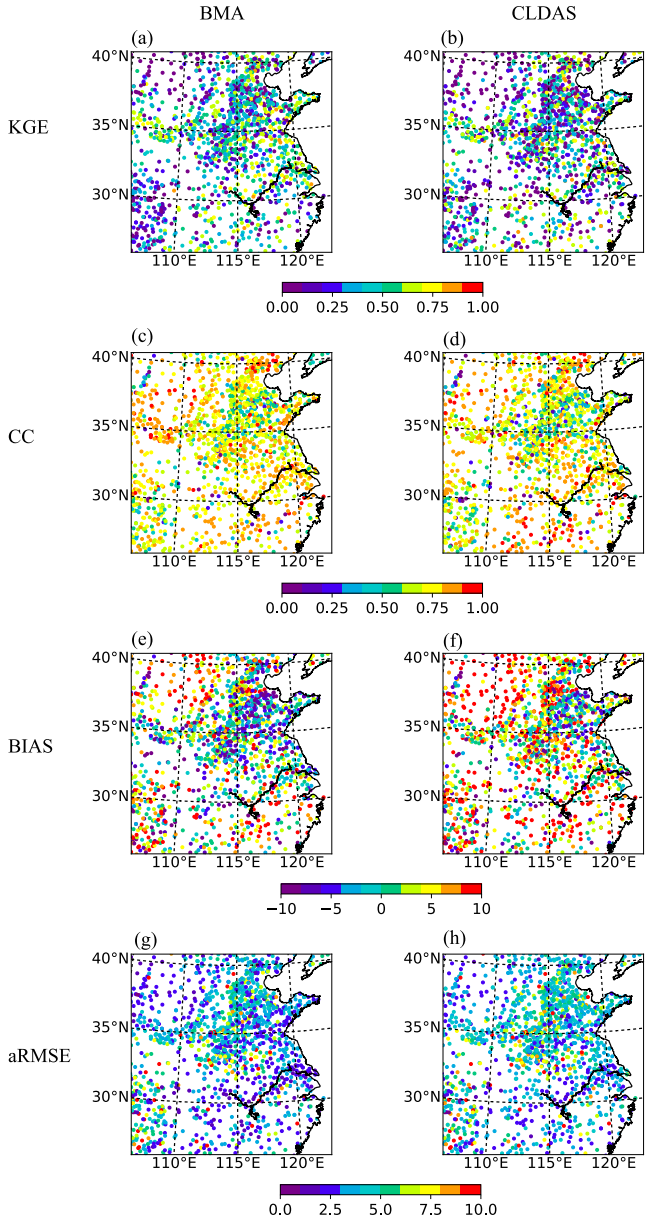

Fig. 6. (a-b) KGE, (c-d) CC, (e-f) BIAS (10^{-2} m^3 m^{-3}), and (g-h) aRMSE (10^{-2} m^3 m^{-3}) of BMA estimate (left) and CLDAS (right) using the sub-daily (6-h) ASMOs as a reference.

monsoons to summer monsoons happens. During the transitional season, northwest winds prevail in most parts of the northern regions, causing low precipitation amount and greater variability in space and time. Therefore, the spring drought is mainly caused by the abnormally low precipitation amount. In addition, the abnormally high temperature also aggravates the degree of drought to a certain extent.

In May 2015, the rainfall amount over the Yangtze-Huaihe river basin is abnormally low, which is 20–50% less than usual. The drought event in this region developed at the beginning of the month, and the degree of drought continued to increase till the end of the month. This drought event is characterized by a large extension over space and a long duration in time. As reported by the government that till 19 May 2017, 23.9 million people had been suffered from this severe drought, of which 130,000 people need help because of difficulties in water supply. The economic losses were estimated 400 million RMB.

The U.S. Drought Monitor (Svoboda et al., 2002) classified the magnitude of drought into five levels based on percentile chance: D0 (abnormally drought; 30%), D1 (moderate drought; 20%), D2 (severe drought; 10%), D3 (extreme drought; 5%), and D4 (exceptional drought; 2%). In this study, moderate drought is characterized with the surface SM at the 20% percentile (0.133 m^3 m^{-3}; Mao et al., 2015; Shukla et al., 2011; Wang et al., 2011a, 2011b; Svoboda et al., 2002). The probability of drought at ASMO locations is estimated using the CDF of dynamic BMA merged SM.

To investigate the capability of dynamic BMA in representing this drought event, SM from AMSOs, CLDAS, and dynamic BMA estimate on 30 May 2017 is shown in Fig. 11. The spatial distribution of ASMOs shows that the north half of the region is characterized by a severe drought condition with SM less than 0.133 m^3 m^{-3} (20% quantile; Fig. 11a). Both the BMA and CLDAS well capture this pattern even more extreme conditions (e.g. 2–5% quantile), though with a little overestimation at some local region (e.g., the northeast of the study domain). The multi-model mean (from product 4, 5, 6 and 8, i.e., BMA members; Table 1) generally captures the spatial distribution of moderate drought (20% quantile), but loses more extreme values (e.g. 2–5% quantile). For this dry event, the dynamic BMA exhibits better KGE, CC, BIAS, and aRMSE than CLDAS (Table 3). Overall, for the spatial pattern on 30 May, the performance of BMA is superior to other models, and dynamic BMA captures the spatial pattern and the magnitude of SM pretty well.

Based on the multi-model merged SM using dynamic BMA, the spatial distribution of moderate drought probability is shown in Fig. 11d. The BMA estimated probability is compared with the multi-model equal probability (Fig. 11c), which is defined as:

$$P = \frac{N_d}{N} \tag{17}$$

where N_d is the number of models that is reporting a dry event and N is the total number of models. Most of the stations in the north of the study region have a drought probability larger than 50%, and the probability could be up to 80% for some local regions (Fig. 11f). This corresponds with the ASMOs shown in Fig. 11a. The probabilistic estimate using multi-model equal probability (Fig. 11e) reports much lower possibility for moderate drought, and for most in-situ sites the probabilities are less than 50%. Table 3 summarizes the deterministic and probabilistic metrics for the model products and dynamic BMA estimated SM in May 2017. The BMA shows the best performance in terms of probabilistic verification metrics and deterministic verification metrics. By assigning larger weights to GLDAS Mosaic in May (Fig. 4) than some wetter months (e.g., July), BMA achieves the best BIAS of -0.236×10^{-2} m^3 m^{-3} (Table 3), suggesting the necessity of using dynamic BMA for drought monitoring with SM.

To quantitatively verify probabilistic estimates generated from equal-weighted multi-model and the dynamic BMA, the BSS for detecting the drought events is compared with the benchmark skill of CLDAS (Fig. 12). Obviously, the BMA estimated SM obtains the highest probabilistic skill at 2%, 5%, 10%, 20%, and 30% quantiles. CLDAS shows a slightly lower BSS than BMA, while the multi-model equal probability exhibits the lowest BSS and even unskillful skill (negative BSS) at 2% quantile. Overall, the dynamic BMA achieved the best skill in capturing drought events from light to severe drought thresholds.

5. Conclusions and discussion

A general dynamic BMA framework is proposed for multiple SM model products fusion. The experiment is performed over the Yangtze-Huaihe river basin with a dense ASMO network (1812 stations). Four model products (product 4, 5, 6, and 8), including three GLDAS products and a reanalysis product ERA5, are merged based on the dynamically variant BMA weights in 2017. Deterministic and probabilistic verification methods are used to evaluate the merged SM against the eight model products (Table 1) and the benchmark CLDAS. To better compare the probability distribution of different products, the CDF consistency histogram and a more objective metric consistency

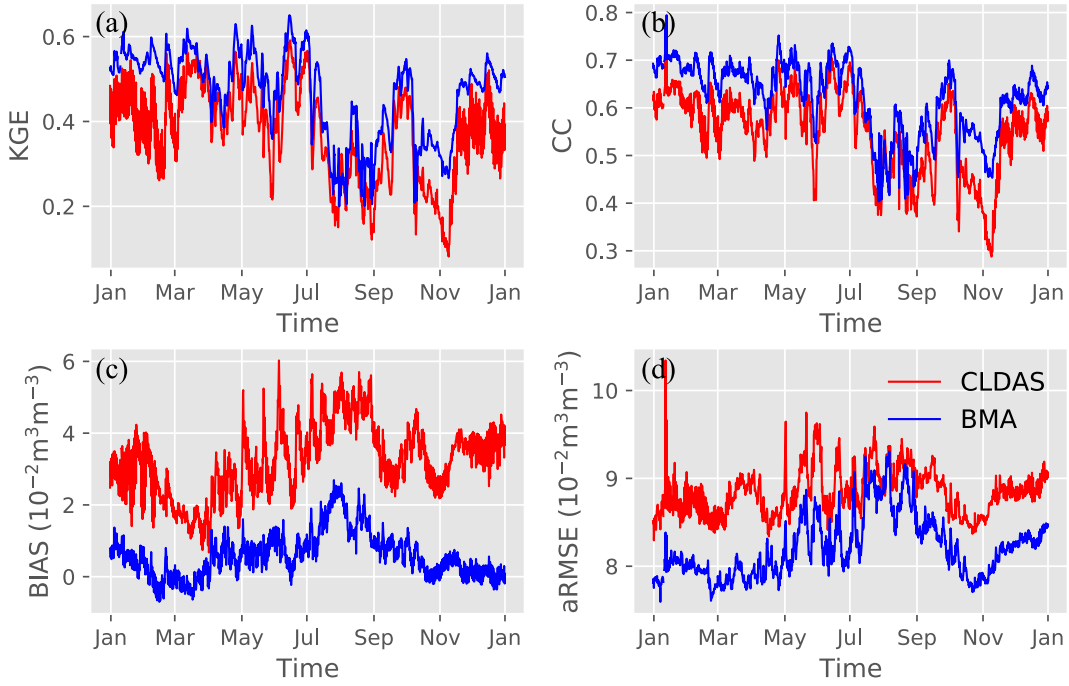

Fig. 7. Time series of (a) KGE, (b) spatial CC, (c) BIAS, and (d) aRMSE of BMA estimate (blue) and CLDAS (red) in 2017. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

deviation (CD) are proposed to diagnose the consistency of two SM CDFs (e.g., the BMA estimated and the observed CDF).

The optimum data length used for dynamic BMA in the training period is about 80 days. It is related to the soil memory (Targulian and Bronnikova, 2019) and the change of weather pattern. Using training data length longer than 80 days will not yield significant improvements for the BMA estimated SM. The framework of dynamic BMA uses weights changing over time to make them quickly adjusted according to the weather patterns and soil conditions. The four model products are assigned with temporally varying weights. In summer, when soil condition is wet, GLDAS Mosaic receives much smaller weights (< 0.05) than others. While in dry periods, e.g., November and December, GLDAS Mosaic could be assigned the largest weight over 0.5. This indicates that although GLDAS Mosaic is assigned with relatively smaller weights for some periods, it could contribute a lot to the skill of multi-model merged SM based on dynamic BMA in dry periods.

BMA is proved to be effective in both deterministic and probabilistic SM estimates, compared with the benchmark product CLDAS. First, the BMA estimated SM has improved performance than any of the BMA members, with the best KGE, CC, and aRMSE among the eight models (Table 2). Therefore, with the dynamic BMA algorithm, a better product than CLDAS could be generated, which can serve as an alternative high-quality SM product, especially for the period without the CLDAS product. The dynamic BMA method can be applied to other years or regions even with sparse observations. Additional experiment (Table 2; BMA with 10% OBS) shows that the dynamic BMA is also doable with only 10% amount (~180 in-situ stations) of observations used in this study.

More importantly, BMA has much improvement in representing the probability distribution of observed SM than CLDAS, with much better probabilistic skill (BSS, CD and π_{reliab}). This highlights the superiority of BMA in the probabilistic estimate. BMA can leverage the strength and weakness of each model product by combining the PDF of each SM model product. The dynamic BMA framework designed for merging sub-daily model products can be further used for drought monitoring and prediction, based on the combined PDF based on model products. The showcase of drought detection in May 2017 indicates the usefulness of the BMA estimated probabilistic SM. This study can be further extended to generate merged SM products at the model grids by applying the BMA framework to SM model products, which requires the appropriate downscaling methods for generating high-resolution merged SM products (Wei et al., 2019). This could encourage the SM

Table 2
Deterministic (KGE, CC, BIAS, and aRMSE) and probabilistic (CD and π_{reliab}) verification statistics for nine SM products (1–9), the BMA merged SM with the full and 10% in-situ observations (10% OBS) in 2017. The best (bold underscore), second best (*), and worst (bold) values in each column are highlighted. All the correlations (CC) are significant at 0.05 level.

Product	KGE	CC	BIAS (10^{-2} m^3 m^{-3})	aRMSE (10^{-2} m^3 m^{-3})	CD	π_{reliab}
(1) GFS	0.337	0.433	1.569	10.128	0.157	0.099
(2) NCEP2	**0.099**	**0.386**	4.351	9.915	**0.434**	0.306
(3) ERA Interim	0.390	0.497	6.000	10.586	0.228	0.312
(4) ERA5	0.265	0.438	0.932	9.736	0.330	0.177
(5) GLDAS Noah	0.391	0.516	1.837	9.646	0.184	0.118
(6) GLDAS CLM	0.348	0.520	−0.725	9.183	0.308	0.176
(7) GLDAS VIC	0.262	0.527	**6.067**	9.180	0.347	**0.375**
(8) GLDAS Mosaic	0.348	0.463	−2.881	**10.760**	0.263	0.199
(9) CLDAS	0.410	0.573	3.255	8.907	0.195	0.202
BMA	**0.494**	**0.640**	−0.613*	**8.275**	**0.100**	**0.076**
BMA (10% OBS)	0.449*	0.584*	**0.033**	8.831*	0.141*	0.087*

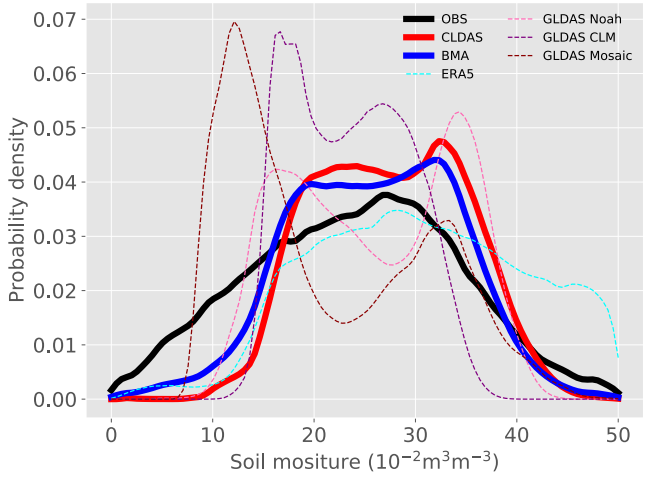

Fig. 8. Probability density distribution of sub-daily (6-h) SM for observations (black), CLDAS (red), dynamic BMA estimate (blue), and the four BMA members. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

related research based on hydrologic and meteorological simulations, which are both in urgent need for high-quality gridded SM initialization field. Ma et al. (2018) used the ordinary Kriging approach to map the BMA weight to grid points focusing on the precipitation in the Tibetan

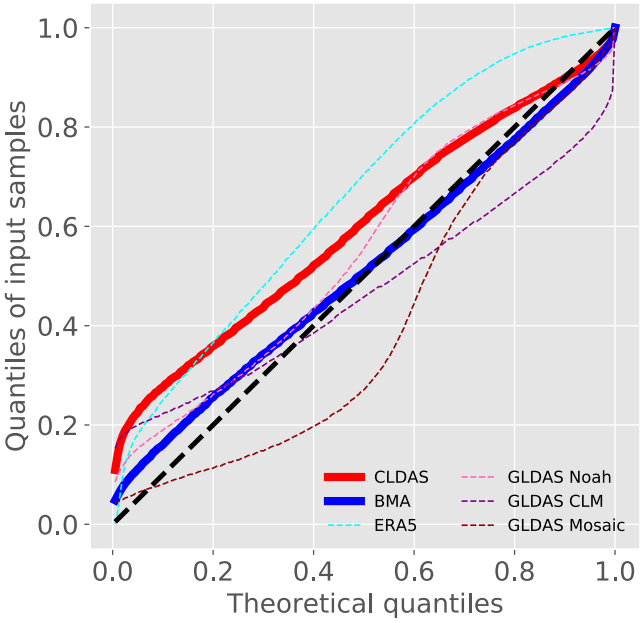

Fig. 10. QQ plots for the four BMA members, CLDAS, and BMA estimate.

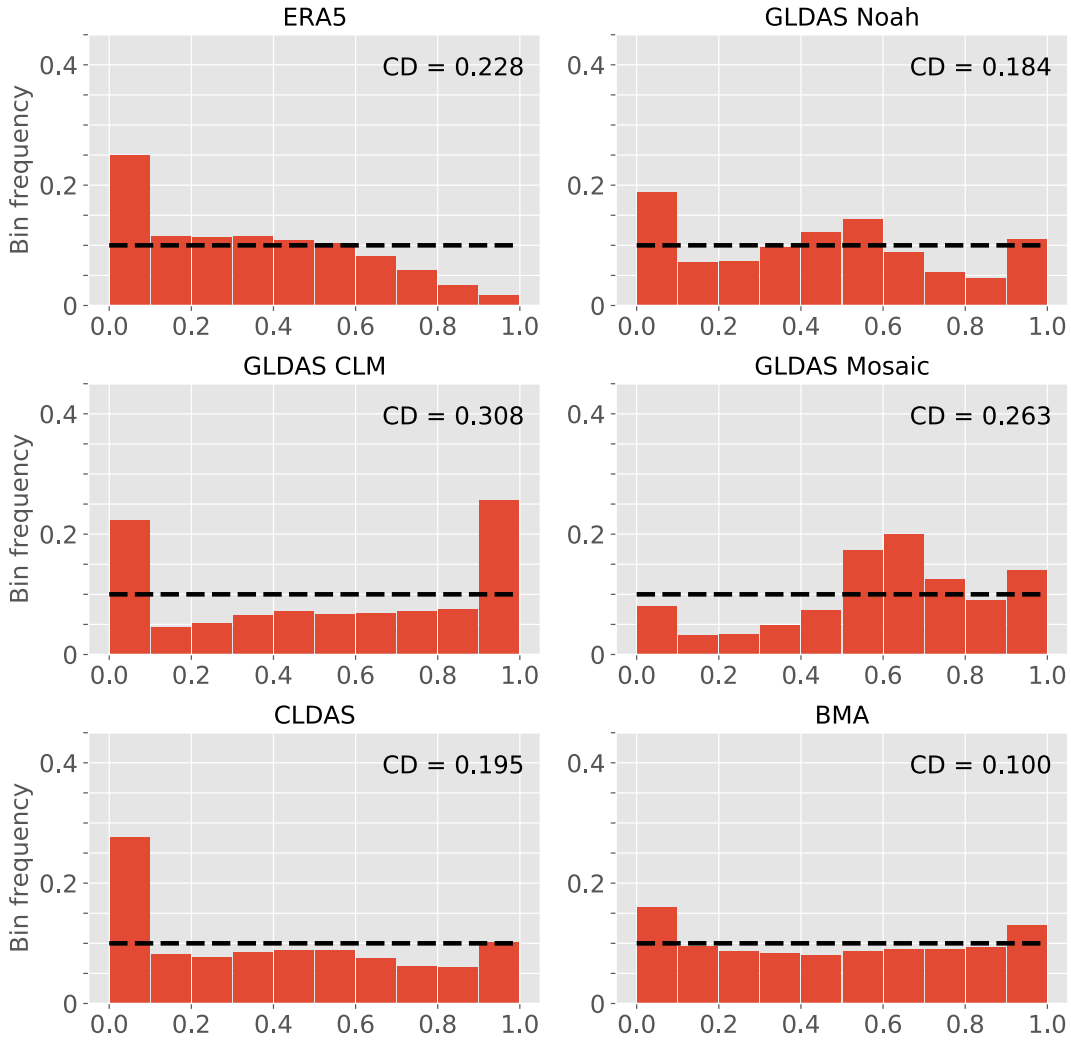

Fig. 9. Cumulative distribution function (CDF) consistency histogram of the four BMA members, CLDAS, and BMA estimate.

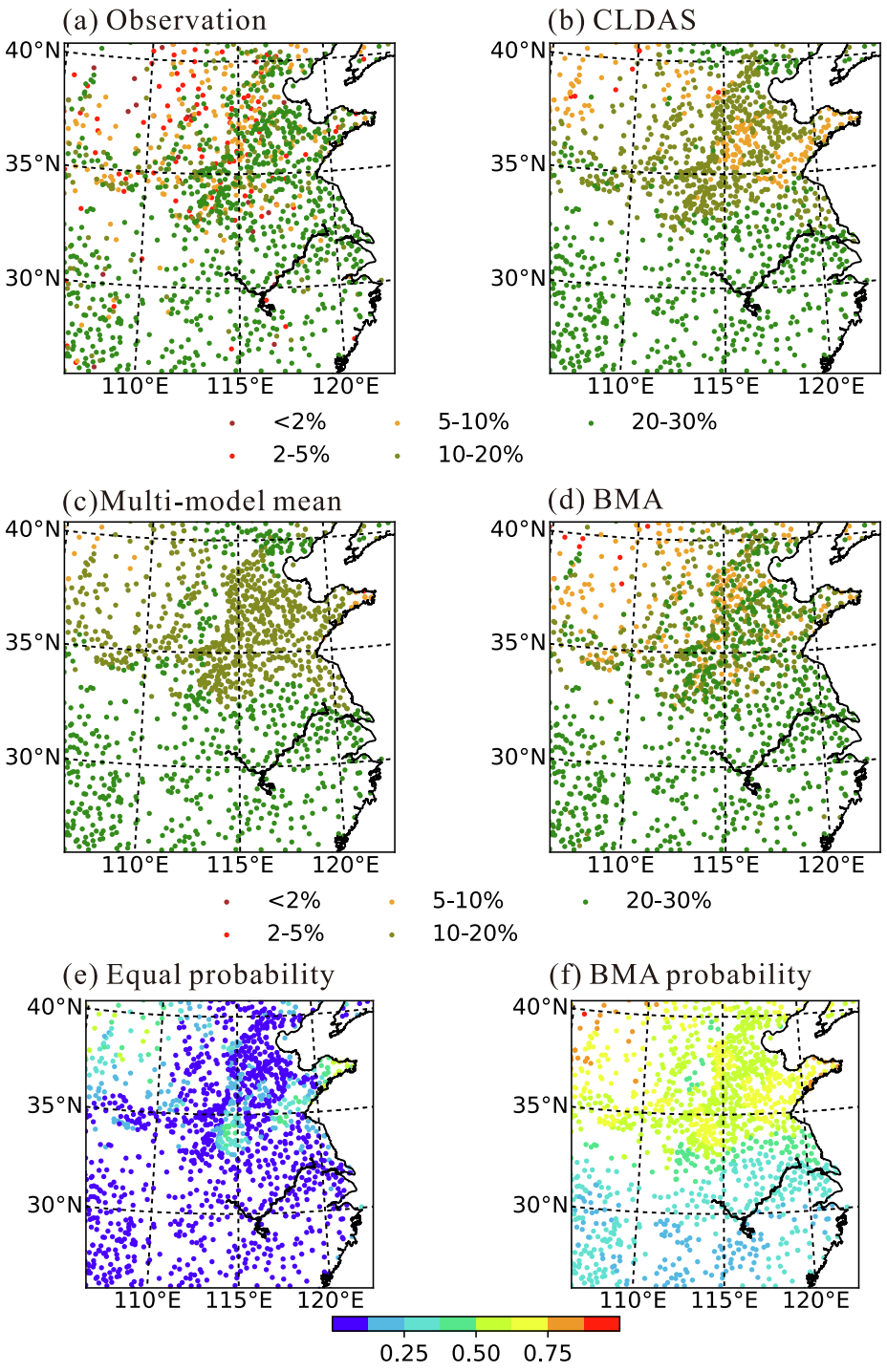

Fig. 11. Soil moisture (10^{-2} m^3 m^{-3}) from (a) observations, (b) CLDAS, (c) multi-model (the four BMA members) mean, and (d) dynamic BMA estimate. Moderate drought (defined at 20% quantile) probabilities generated from (e) equal weighted multi-model and (f) dynamic BMA on 30 May 2017.

Plateau (TP) region. Wang et al. (2014) firstly extended a Multiscale Kalman Smoother-based approach to fuse precipitation data at different spatial resolutions. Since SM is quite different from precipitation characterized with high inhomogeneity over space and time, further validation of applying these methods into SM data fusion is still needed. Furthermore, the TP region is characterized by permafrost and seasonally frozen soil (Wang et al., 2019), which poses a great challenge for applying BMA approach in this region. The ASMO sensors have large uncertainty when the temperature is around 0 °C, during which ASMO data are removed by quality control procedures in this study.

The BMA member independency and the performance of BMA members could impact the performance and usefulness of the dynamic BMA approach. Additional experiment shows that using all the model products would not significantly improve the performance, because of the high correlation between GLDAS products and the use of models with bad performance. When only one best GLDAS product (GLDAS Noah) along with ERA5, NCEP2, and GFS are used as BMA members based on the evaluation by Chen and Yuan (2020) to remove the highly correlated models, the results improve the KGE by 6.7% than using all models. However, in this study, using the four products with the best performance (Chen and Yuan, 2020), i.e., EAR5, GLDAS Noah, GLDAS CLM, and GLDAS Mosaic, achieves the best dynamic BMA estimate with

Table 3
The same as Table 2 but for May 2017.

Product	KGE	CC	BIAS (10^{-2} m^3 m^{-3})	aRMSE (10^{-2} m^3 m^{-3})	CD	π_{reliab}
(1) GFS	0.452	0.509	1.931	9.220	0.132	0.185
(2) NCEP2	**0.057**	**0.414**	5.462	9.111	**0.399**	**0.560**
(3) ERA Interim	0.520	0.564	4.546	**10.231**	0.205	0.280
(4) ERA5	0.349	0.532	1.042	8.475	0.193	0.391
(5) GLDAS Noah	0.514	0.560	2.074	8.911	0.138	0.244
(6) GLDAS CLM	0.439	0.566	−0.638*	8.264	0.175	0.304
(7) GLDAS VIC	0.300	0.562	**6.469**	8.363	0.398	0.369
(8) GLDAS Mosaic	0.477	0.549	−3.161	9.776	0.208	0.305
(9) CLDAS	0.548*	0.598*	2.659	8.215*	0.125*	0.184*
BMA	**0.581**	**0.612**	**−0.236**	**5.897**	**0.112**	**0.096**

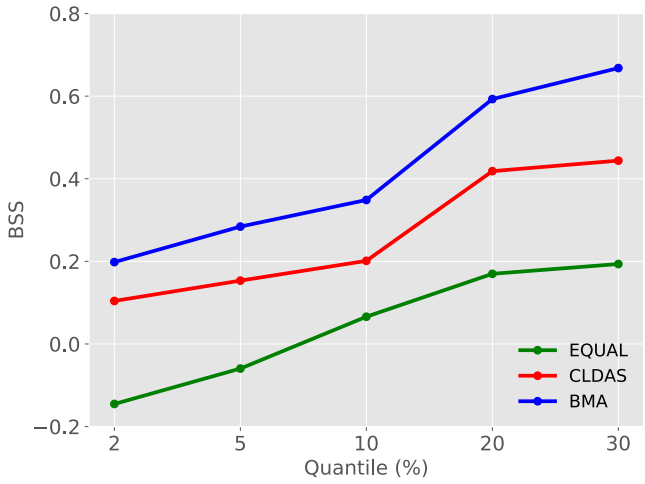

Fig. 12. Brier skill scores for CLDAS (red), and the probabilistic SM generated from equal-weighted multi-model (green) and BMA (blue) in May 2017. Thresholds are set using the observed CDF quantiles of 2%, 5%, 10%, 20%, and 30%. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

improved KGE of 18.8% compared to using all models. It's due to the fact that the best model products are used and one GLDAS product is replaced with ERA5, which partly addresses the issue of high correlation between models. As three GLDAS products are used in this study, the uncertainty of predictive distribution could be overestimated. The challenge for selecting independent models (mutually exclusive) still remains.

The dynamic BMA has the same limitations as traditional standard BMA, since the overall structure of both algorithms is similar. The conditional PDF of each model is assumed to be a particular probability distribution (e.g., Gaussian, gamma, etc.). Bias correction methods are strongly recommended before applying dynamic BMA. Integrating Copula with BMA (Cop-BMA) can be a reliable approach to overcome these limitations (Madadgar and Moradkhani, 2014a). The Cop-BMA relaxes any assumption on the shape of conditional PDFs). Furthermore, copulas are effective tools in correcting model biases. In the future, dynamic Cop-BMA could be designed to overcome the limitations of dynamic BMA proposed in this study.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Key R&D Program of China [grant number 2018YFC1507405], and the National Natural Science Foundation of China [grant number 41675109]. In particular, we would like to acknowledge Fei Chen and Michael Barlage from the National Center for Atmospheric Research (NCAR) for their constructive suggestions. Y. Chen thanks the China Scholarship Council (CSC) for the financial support. We thank the CMA/NMIC for providing the ASMO data and CLDAS data. We also thank the Jiangsu Collaborative Innovation Center for Climate Change, China. Computing resources were provided by the High Performance Computing Center at Nanjing University and the Climate Simulation Laboratory at NCAR's Computational and Information Systems Laboratory.

References

Abbaszadeh, P., Moradkhani, H., Daescu, D.N., 2019a. The quest for model uncertainty quantification: A hybrid ensemble and variational data assimilation framework. Water Resour. Res. 55 (3), 2407–2431. https://doi.org/10.1029/2018WR023629.
Abbaszadeh, P., Moradkhani, H., Zhan, X., 2019b. Downscaling SMAP radiometer soil moisture over the CONUS using an ensemble learning method. Water Resour. Res. 55, 324–344. https://doi.org/10.1029/2018WR023354.
Brocca, L., et al., 2011. Soil moisture estimation through ASCAT and AMSR-E sensors: An intercomparison and validation study across Europe. Remote Sens. Environ. 115 (12), 3390–3408. https://doi.org/10.1016/j.rse.2011.08.003.
Brocca, L., Melone, F., Moramarco, T., Wagner, W., Hasenauer, S., 2010a. ASCAT soil wetness index validation through in situ and modeled soil moisture data in central Italy. Remote Sens. Environ. 114 (11), 2745–2755. https://doi.org/10.1016/j.rse.2010.06.009.
Brocca, L., et al., 2010b. Improving runoff prediction through the assimilation of the ASCAT soil moisture product. Hydrol. Earth Syst. Sci. 14 (10), 1881–1893. https://doi.org/10.5194/hess-14-1881-2010.
Brocca, L., Morbidelli, R., Melone, F., Moramarco, T., 2007. Soil moisture spatial variability in experimental areas of central Italy. J. Hydrol. 333 (2–4), 356–373. https://doi.org/10.1016/j.jhydrol.2006.09.004.
Chen, F., Dudhia, J., 2001. Coupling an advanced land surface–hydrology model with the Penn State–NCAR MM5 modeling system. Part II: Preliminary model validation. Mon. Weather Rev. 129 (4), 587–604. https://doi.org/10.1175/1520-0493(2001)129<0587:Caalsh>2.0.Co;2.
Chen, Y., Yuan, H., 2020. Evaluation of nine sub-daily soil moisture model products over China using high-resolution in situ observations. J. Hydrol. 588, 125054. https://doi.org/10.1016/j.jhydrol.2020.125054.
Chen, Y.D., Zhang, Q., Xiao, M., Singh, V.P., Zhang, S., 2015. Probabilistic forecasting of seasonal droughts in the Pearl River basin, China. Stoch. Env. Res. Risk Assess. 30 (7), 2031–2040. https://doi.org/10.1007/s00477-015-1174-6.
Copernicus Climate Change Service Retriedved on 26 April 2020 from https://cds.climate.copernicus.eu/cdsapp#!/home.
de Rosnay, P., et al., 2013. A simplified Extended Kalman Filter for the global operational soil moisture analysis at ECMWF. Q. J. R. Meteorolog. Soc. 139 (674), 1199–1213. https://doi.org/10.1002/qj.2023.
de Rosnay, P., Balsamo, G., Albergel, C., Muñoz-Sabater, J., Isaksen, L., 2012. Initialisation of land surface variables for numerical weather prediction. Surv. Geophys. 35 (3), 607–621. https://doi.org/10.1007/s10712-012-9207-x.
DeChant, C.M., Moradkhani, H., 2014. Toward a reliable prediction of seasonal forecast uncertainty: Addressing model and initial condition uncertainty with ensemble data assimilation and Sequential Bayesian Combination. J. Hydrol. 519, 2967–2977. https://doi.org/10.1016/j.jhydrol.2014.05.045.
Dee, D.P., et al., 2011. The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. Q. J. R. Meteorolog. Soc. 137 (656), 553–597. https://doi.org/10.1002/qj.828.
Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc.: Ser. B (Methodol.) 39 (1), 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.
Dharssi, I., Bovis, K.J., Macpherson, B., Jones, C.P., 2011. Operational assimilation of ASCAT surface soil wetness at the Met Office. Hydrol. Earth Syst. Sci. 15 (8), 2729–2746. https://doi.org/10.5194/hess-15-2729-2011.
Dirmeyer, P.A., et al., 2006. GSWP-2: Multimodel analysis and implications for our perception of the land surface. Bull. Am. Meteorol. Soc. 87 (10), 1381–1398. https://doi.

org/10.1175/bams-87-10-1381.
Dirmeyer, P.A., Guo, Z., Gao, X., 2004. Comparison, validation, and transferability of eight multiyear global soil wetness products. J. Hydrometeorol. 5 (6), 1011–1033. https://doi.org/10.1175/jhm-388.1.
Duan, Q., Ajami, N.K., Gao, X., Sorooshian, S., 2007. Multi-model ensemble hydrologic prediction using Bayesian model averaging. Adv. Water Resour. 30 (5), 1371–1386. https://doi.org/10.1016/j.advwatres.2006.11.014.
Enenkel, M., et al., 2016. A combined satellite-derived drought indicator to support humanitarian aid organizations. Remote Sens. 8 (4), 340. https://doi.org/10.3390/rs8040340.
Entekhabi, B.D., et al., 2010. The Soil Moisture Active Passive (SMAP) mission. Proc. IEEE 98 (5), 704–716. https://doi.org/10.1109/JPROC.2010.2043918.
Environmental Modeling Center, 2016. The Global Forecast System (GFS) - Global Spectral Model. Retrieved on 9 November 2019 from https://www.emc.ncep.noaa.gov/GFS/doc.php.
Gneiting, T., Raftery, A.E., Westveld, A.H., Goldman, T., 2005. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Mon. Weather Rev. 133 (5), 1098–1118. https://doi.org/10.1175/mwr2904.1.
Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. J. Hydrol. 377 (1–2), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003.
Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: A tutorial. Statistical Science 14 (4), 382–417. https://doi.org/10.1214/ss/1009212519.
James, A.L., Roulet, N.T., 2009. Antecedent moisture conditions and catchment morphology as controls on spatial patterns of runoff generation in small forest catchments. J. Hydrol. 377 (3–4), 351–366. https://doi.org/10.1016/j.jhydrol.2009.08.039.
Jiang, S., et al., 2014. Improvement of multi-satellite real-time precipitation products for ensemble streamflow simulation in a middle latitude basin in South China. Water Resour. Manage. 28 (8), 2259–2278. https://doi.org/10.1007/s11269-014-0612-4.
Kanamitsu, M., et al., 2002. NCEP–DOE AMIP-II Reanalysis (R-2). Bull. Am. Meteorol. Soc. 83 (11), 1631–1644. https://doi.org/10.1175/bams-83-11-1631.
Kim, J., Mohanty, B.P., Shin, Y., 2015. Effective soil moisture estimate and its uncertainty using multimodel simulation based on Bayesian Model Averaging. J. Geophys. Res.: Atmos. 120 (16), 8023–8042. https://doi.org/10.1002/2014jd022905.
Kling, H., Fuchs, M., Paulin, M., 2012. Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. J. Hydrol. 424–425, 264–277. https://doi.org/10.1016/j.jhydrol.2012.01.011.
Koster, R.D., et al., 2009. On the nature of soil moisture in land surface models. J. Clim. 22 (16), 4322–4335. https://doi.org/10.1175/2009jcli2832.1.
Koster, R.D., Mahanama, S.P.P., Livneh, B., Lettenmaier, D.P., Reichle, R.H., 2010. Skill in streamflow forecasts derived from large-scale estimates of soil moisture and snow. Nat. Geosci. 3 (9), 613–616. https://doi.org/10.1038/ngeo944.
Krzysztofowicz, R., Kelly, K.S., 2000. Hydrologic uncertainty processor for probabilistic river stage forecasting. Water Resour. Res. 36 (11), 3265–3277. https://doi.org/10.1029/2000wr900108.
Kumar, S.V., et al., 2014. Assimilation of remotely sensed soil moisture and snow depth retrievals for drought estimation. J. Hydrometeorol. 15 (6), 2446–2469. https://doi.org/10.1175/jhm-d-13-0132.1.
Leach, M.J., Coulibaly, P., 2019. An extension of data assimilation into the short-term hydrologic forecast for improved prediction reliability. Adv. Water Resour. 134, 103443. https://doi.org/10.1016/j.advwatres.2019.103443.
Ma, Y., et al., 2018. Performance of optimally merged multisatellite precipitation products using the dynamic Bayesian model averaging scheme over the Tibetan plateau. J. Geophys. Res.: Atmos. 123 (2), 814–834. https://doi.org/10.1002/2017jd026648.
Madadgar, S., Moradkhani, H., 2013. A Bayesian framework for probabilistic seasonal drought forecasting. J. Hydrometeorol. 14 (6), 1685–1705. https://doi.org/10.1175/jhm-d-13-010.1.
Madadgar, S., Moradkhani, H., 2014a. Improved Bayesian multimodeling: Integration of copulas and Bayesian model averaging. Water Resour. Res. 50, 9586–9603. https://doi.org/10.1002/2014WR015965.
Madadgar, S., Moradkhani, H., 2014b. Spatio-temporal drought forecasting within Bayesian networks. J. Hydrol. 512, 134–146. https://doi.org/10.1016/j.jhydrol.2014.02.039.
Mao, Y., Nijssen, B., Lettenmaier, D.P., 2015. Is climate change implicated in the 2013–2014 California drought? A hydrologic perspective. Geophys. Res. Lett. 42 (8), 2805–2813. https://doi.org/10.1002/2015gl063456.
Mishra, A.K., Desai, V.R., 2005. Drought forecasting using stochastic models. Stoch. Env. Res. Risk Assess. 19 (5), 326–339. https://doi.org/10.1007/s00477-005-0238-4.
Pan, H.L., Mahrt, L., 1987. Interaction between soil hydrology and boundary-layer development. Bound.-Layer Meteorol. 38 (1–2), 185–202. https://doi.org/10.1007/bf00121563.
Pathiraja, S., Moradkhani, H., Marshall, L., Sharma, A., Geenens, G., 2018. Data-driven model uncertainty estimation in hydrologic data assimilation. Water Resour. Res. 54 (2), 1252–1280. https://doi.org/10.1002/2018WR022627.
Penna, D., Borga, M., Norbiato, D., Dalla Fontana, G., 2009. Hillslope scale soil moisture variability in a steep alpine terrain. J. Hydrol. 364 (3–4), 311–327. https://doi.org/10.1016/j.jhydrol.2008.11.009.
Raftery, A.E., Gneiting, T., Balabdaoui, F., Polakowski, M., 2005. Using Bayesian model averaging to calibrate forecast ensembles. Mon. Weather Rev. 133 (5), 1155–1174. https://doi.org/10.1175/mwr2906.1.
Rahmani, A., Golian, S., Brocca, L., 2016. Multiyear monitoring of soil moisture over Iran through satellite and reanalysis soil moisture products. Int. J. Appl. Earth Obs. Geoinf. 48, 85–95. https://doi.org/10.1016/j.jag.2015.06.009.
Robock, A., et al., 2000. The global soil moisture data bank. Bull. Am. Meteorol. Soc. 81 (6), 1281–1299. https://doi.org/10.1175/1520-0477(2000)081<1281:Tgsmdb>2.3.Co;2.
Rodell, M., et al., 2004. The global land data assimilation system. Bull. Am. Meteorol. Soc. 85 (3), 381–394. https://doi.org/10.1175/bams-85-3-381.
Roy, T., Serrat-Capdevila, A., Gupta, H., Valdes, J., 2017. A platform for probabilistic Multimodel and Multiproduct Streamflow Forecasting. Water Resour. Res. 53 (1), 376–399. https://doi.org/10.1002/2016wr019752.
Sakov, P., Sandery, P., 2017. An adaptive quality control procedure for data assimilation. Tellus A: Dyn. Meteorol. Oceanogr. 69 (1), 1318031. https://doi.org/10.1080/16000870.2017.1318031.
Susha Lekshmi, S.U., Singh, D.N., Shojaei Baghini, M., 2014. A critical review of soil moisture measurement. Measurement: J. Int. Measure. Conf. 54, 92–105. https://doi.org/10.1016/j.measurement.2014.04.007.
Sheffield, J., 2004. A simulated soil moisture based drought analysis for the United States. J. Geophys. Res. 109, D24. https://doi.org/10.1029/2004jd005182.
Svoboda, M., et al., 2002. The drought monitor. Bull. Am. Meteorol. Soc. 83 (8), 1181–1190. https://doi.org/10.1175/1520-0477-83.8.1181.
Shen, Y., Zhao, P., Pan, Y., Yu, J., 2014. A high spatiotemporal gauge-satellite merged precipitation analysis over China. J. Geophys. Res.: Atmos. 119 (6), 3063–3075. https://doi.org/10.1002/2013jd020686.
Shi, C., Jiang, L., Tao, Z., Gong, W., Han, S., 2014. Status and plans of CMA Land Data Assimilation System (CLDAS) project. The 19th International TOVS Study Conference. Jeju Island, South Korea.
Shukla, S., Steinemann, A.C., Lettenmaier, D.P., 2011. Drought monitoring for Washington state: Indicators and applications. J. Hydrometeorol. 12 (1), 66–83. https://doi.org/10.1175/2010jhm1307.1.
Sloughter, J.M., Gneiting, T., Raftery, A.E., 2013. Probabilistic wind vector forecasting using ensembles and Bayesian model averaging. Mon. Weather Rev. 141 (6), 2107–2119. https://doi.org/10.1175/mwr-d-12-00002.1.
Sloughter, J.M., Raftery, A.E., Gneiting, T., Fraley, C., 2007. Probabilistic quantitative precipitation forecasting using Bayesian model averaging. Mon. Weather Rev. 135 (9), 3209–3220. https://doi.org/10.1175/mwr3441.1.
Sun, R., Yuan, H., Yang, Y., 2018. Using multiple satellite-gauge merged precipitation products ensemble for hydrologic uncertainty analysis over the Huaihe River basin. J. Hydrol. 566, 406–420. https://doi.org/10.1016/j.jhydrol.2018.09.024.
Targulian, V.O., Bronnikova, M.A., 2019. Soil memory: Theoretical basics of the concept, its current state, and prospects for development. Eurasian Soil Sci. 52 (3), 229–243. https://doi.org/10.1134/s1064229319030116.
Vrugt, J.A., Robinson, B.A., 2007. Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging. Water Resour. Res. 43, W01411. https://doi.org/10.1029/2005wr004838.
Vrugt, J.A., ter Braak, C.J.F., Clark, M.P., Hyman, J.M., Robinson, B.A., 2008. Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation. Water Resour. Res. 44, W01411. https://doi.org/10.1029/2007wr006720.
Wang, A., Lettenmaier, D.P., Sheffield, J., 2011a. Soil moisture drought in China, 1950–2006. J. Clim. 24 (13), 3257–3271. https://doi.org/10.1175/2011jcli3733.1.
Wang, L., Qu, J.J., 2009. Satellite remote sensing applications for surface soil moisture monitoring: A review. Front Earth Sci. China 3 (2), 237–247. https://doi.org/10.1007/s11707-009-0023-7.
Wang, T., et al., 2019. Spatial distribution and changes of permafrost on the Qinghai-Tibet Plateau revealed by statistical models during the period of 1980 to 2010. Sci. Total Environ. 650 (Pt 1), 661–670. https://doi.org/10.1016/j.scitotenv.2018.08.398.
Wang, S., Liang, X., Nan, Z., 2011b. How much improvement can precipitation data fusion achieve with a Multiscale Kalman Smoother-based framework? Water Resour. Res. 47, W00H12. https://doi.org/10.1029/2010wr009953.
Wei, Z., Meng, Y., Zhang, W., Peng, J., Meng, L., 2019. Downscaling SMAP soil moisture estimation with gradient boosting decision tree regression over the Tibetan Plateau. Remote Sens. Environ. 225, 30–44. https://doi.org/10.1016/j.rse.2019.02.022.
Wilk, M.B., Gnanadesikan, R., 1968. Probability plotting methods for the analysis for the analysis of data. Biometrika 55 (1), 1–17. https://doi.org/10.1093/biomet/55.1.1.
Wood, A.W., Lettenmaier, D.P., 2008. An ensemble approach for attribution of hydrologic prediction uncertainty. Geophys. Res. Lett. 35, L14401. https://doi.org/10.1029/2008gl034648.
Yang, M., Chen, X., Cheng, C.S., 2016. Hydrological impacts of precipitation extremes in the Huaihe River Basin. China. SpringerPlus 5 (1), 1731. https://doi.org/10.1186/s40064-016-3429-1.
Yang, Y., Yuan, H., Yu, W., 2018. Uncertainties of 3D soil hydraulic parameters in streamflow simulations using a distributed hydrological model system. J. Hydrol. 567, 12–24. https://doi.org/10.1016/j.jhydrol.2018.09.042.
Zhang, X., Tang, Q., Liu, X., Leng, G., Li, Z., 2017. Soil moisture drought monitoring and forecasting using satellite and climate model data over southwestern China. J. Hydrometeorol. 18 (1), 5–23. https://doi.org/10.1175/jhm-d-16-0045.1.
Zhong, J.Q., Lu, B., Wang, W., Huang, C.C., Yang, Y., 2020. Impact of soil moisture on winter 2-m temperature forecasts in northern China. J. Hydrometeorol. 21 (4), 597–614. https://doi.org/10.1175/jhm-d-19-0060.1.
Zhu, Q., Luo, Y., Xu, Y.-P., Tian, Y., Yang, T., 2019. Satellite soil moisture for agricultural drought monitoring: Assessment of SMAP-derived Soil Water Deficit Index in Xiang River Basin China. Remote Sens. 11 (3), 362. https://doi.org/10.3390/rs11030362.