1
2
3
4
5
6
7
8
9
10

**Evaluating Uncertainties in Multi-layer Soil Moisture
Estimation with Support Vector Machines and Ensemble
Kalman Filtering**

Di Liu^{a,b}, Ashok K. Mishra^a, Zhongbo Yu^b

a. Glenn Department of Civil Engineering, Clemson University, 202 Lowry Hall,
Clemson, SC 29634, USA

b. State Key Laboratory of Hydrology-Water Resources and Hydraulic
Engineering, HoHai University, Nanjing 210098, PR China

1 **Abstract**

2 This paper examines the combination of support vector machines (SVM) and the
3 dual ensemble Kalman filter (EnKF) technique to estimate root zone soil moisture at
4 different soil layers up to 100 cm depth. Multiple experiments are conducted in a data
5 rich environment to construct and validate the SVM model and to explore the
6 effectiveness and robustness of the EnKF technique. It was observed that the
7 performance of SVM relies more on the initial length of training set than other factors
8 (e.g., cost function, regularization parameter, and kernel parameters). The dual EnKF
9 technique proved to be efficient to improve SVM with observed data either at each
10 time step or at a flexible time steps. The EnKF technique can reach its maximum
11 efficiency when the updating ensemble size approaches a certain threshold. It was
12 observed that the SVM model performance for the multi-layer soil moisture
13 estimation can be influenced by the rainfall magnitude (e.g., dry and wet spells).

14
15 **Keywords:** Support Vector Machines, Data Assimilation, Dual Ensemble Kalman
16 Filter, Soil Moisture Estimation, Uncertainty

1. Introduction

Soil moisture is a key hydrologic state variable that controls the land surface and biophysical processes, as well as atmospheric exchanges of sensible and latent heat (Castelli et al., 1999; Boni et al., 2001; Sun et al., 2011; Ridler et al., 2014). On a catchment scale, soil moisture controls the partitioning of rainfall into runoff and infiltration—thus impacting both groundwater recharge and river discharge (Moradkhani, 2008; Ni-Meister, 2008). For water resource managers, improved soil moisture estimates can improve agricultural water management, drought and flood forecasting (Milly et al., 2008). However, it is challenging to accurately estimate soil moisture that varies in depth, space and time.

Currently, soil moisture can be obtained by in situ networks, hydrological modeling and/or remote sensing techniques and each method has some limitations. The in situ networks are expensive and impractical for large areas, whereas hydrological modeling are uncertain due to poorly described model physics, imperfect parameterization, meteorological forcing data, and initial conditions. Developments in remotely sensed soil moisture retrievals (e.g., Advanced Microwave Scanning Radiometer (AMSR-E) (Njoku et al., 2003), Soil Moisture Ocean Salinity Satellite (SMOS) (Kerr et al., 2010), Soil Moisture Active Passive (SMAP) (Entekhabi et al., 2010)) can provide an unprecedented spatial and temporal resolution of soil moisture data across a range of scales, but is limited in terms of sensing at different depth. Some efforts have been done to retrieve root zone soil moisture profile from the surface values (Li et al., 2012; Tran et al., 2013; Ridler et al., 2014), but these studies often encountered obstacles, including the hydraulic parameterization, initial moisture condition, satellite measurement accuracy and the spatial and temporal scales (Vereecken et al., 2008; Montzka et al., 2011).

Data assimilation (DA) technique is considered to be a promising technique to optimally estimate the soil moisture by merging observed information into models (e.g., Kumar et al., 2008; Lu et al., 2010; Das et al., 2011; Li et al., 2012; Tran et al., 2013; Han et al., 2014; Kornelsen and Coulibaly, 2014; Yin et al., 2015). The integration of surface soil moisture measurements into a hydrological model through DA has proven a promising approach to predict root-zone soil moisture (Walker et al., 2001; Das and Mohanty, 2006; Lu et al., 2010; Dumedah and Coulibaly, 2012; Yu et al., 2012; Han et al., 2014; Mishra et al., 2015). However, the evolutionary DA technique often exhibits a high computational cost due to the multi-objective

evolutionary search strategy (Dumedah and Coulibaly, 2012). For instance, (Clark et al., 2008) found that the EnKF-based approach is not only computationally demanding but also has strong limitations for nonlinear systems.

Data-driven methods, such as support vector machines (SVMs), which can mine data for nonlinear interdependencies and have potential for estimating root zone soil moisture with prior atmospheric observations (e.g. solar radiation, relative humidity, air temperature) and hydrologic observations (e.g. soil temperature, soil moisture), therefore, can serve as a simple alternative technique to hydrological models. For soil moisture estimation, the SVMs are found to be efficient in both the training and estimating processes (e.g., Gill et al., 2006; Gill and McKee, 2007). Other uses of SVM can be found in (Kaheil et al., 2008; Liu et al., 2010).

As stated in previous literatures (e.g., Evensen, 2003; Samuel et al., 2014), ensemble Kalman filter (EnKF) technique uses statistical distributions to represent uncertainties of both observations and model errors and to generate ensembles of model forcing and other variables. In the forecast step, an ensemble of model states is propagated forward in time using the model. As a result, the accuracy of the sampled covariance depends on the ensemble size (Kumar et al., 2008). However, the approach to determine the ensemble size is usually limited by high computational costs. Thus, a large group of experiments are designed in this paper.

This paper aims to develop a multilayer soil moisture estimation model using a combination of SVM and a sequential DA method known as EnKF technique. The data used in this study is from the Blackville experiment site located in South Carolina of United States. The remainder of this paper is organized as follows. Study area and data is described in Section 2. The SVM and EnKF technique is introduced in Section 3. The experimental design, including the prior data analysis, SVM construction and validation experiments, the EnKF updating experiments as well as the evaluation criteria are illustrated in Section 4. The results from each experiment are analyzed in Section 5, while the discussions and conclusions are summarized in Section 6.

2. Study Area and Data

Blackville experimental site (latitude 33°21'18", longitude 81°19'40", elevation 317ft) is located in South Carolina, United States. In addition to the real time sensor based soil moisture monitoring station, this site also includes a state of the art NOAA

U.S. Climate Reference Network station (SC_Blackville_3W) with automated measurements of air temperature, humidity, solar radiation, rainfall, soil temperature and moisture at different soil depth from 5-100cm. Besides, it provides the opportunity to study a variety of crops (i.e., Cotton, Corn, Wheat, Soybean, Peanut, and Sorghum) and to develop drought indices for individual crops based on their water demand.

The daily time series data includes the maximum, minimum, mean and average air temperature (named as $t_{\max}$, $t_{\min}$, t_{mean} , t_{avg}), daily precipitation (P), total solar energy (SR), maximum, minimum and average infrared surface temperature (named as $sur_temp_{\max}$, $sur_temp_{\min}$ and sur_temp_{avg}), maximum, minimum and average relative humidity (named as $rh_{\max}$, $rh_{\min}$ and rh_{avg}), soil moisture at 5cm, 10cm, 20cm, 50cm and 100cm depth (named as SM1, SM2, SM3, SM4 and SM5) and soil temperature at 5cm, 10cm, 20cm, 50cm and 100cm depth (named as ST1, ST2, ST3, ST4 and ST5). The daily data between 1st September 2009 and 31st May 2015 (totally 2099 days) was obtained from the U.S. Climate Reference Network (USCRN) (<https://www.ncdc.noaa.gov/crn/qcdatasets.html>). The USCRN use high-quality instruments to measure temperature, precipitation, wind speed, soil conditions, and more.

3. Methodology

3.1 Support Vector Machines (SVMs)

The SVM was applied to estimate the soil moisture at different soil layers from surface to root zone. As described in previous literatures (Liu et al., 2010), the algorithm of SVM maps the input space in a high-dimensional feature space by utilizing kernels (Vapnik, 1995). SVM regression estimation generally involves the following training: (1) selection of a suitable kernel and kernel parameter, (2) specifying the penalty parameter, and (3) specifying the insensitive parameter. Since SVMs are inherently deterministic models, the evaluation of confidence bounds is a difficult but necessary task. In this paper, the SVM model was implemented using the concept provided by De Brabanter et al., (2010).

3.2 Ensemble Kalman Filter (EnKF)

The EnKF is a popular DA technique used in hydrology (Gill and McKee, 2007; Liu et al., 2010; Leisenring and Moradkhani, 2012; Yin et al., 2014; Mishra et al., 2015; Yin et al., 2015). It is a Monte Carlo approximation of a sequential Bayesian

filtering process, which alternates between an ensemble forecast step and a state variable update step (Reichle et al., 2002). The dual EnKF technique is adopted in this paper consists of two steps: generating an ensemble of model outputs and updating it when new observations become available. The differential equations for the generic nonlinear dynamic system are formulated as follows (Reichle et al., 2002; Leisenring and Moradkhani, 2012):

$$X_t = F(X_{t-1}, U_t, \theta_t) + W_t \quad (1)$$

$$Y_t = H(X_t) + V_t \quad (2)$$

Where $F(\cdot)$ is the model operator mapping the previous state X_{t-1} at time $t-1$ to state X_t at time t ; $H(\cdot)$ is the observation operator that converts state to observation; Here X_t is a vector of the uncertain state variable at time t , while Y_t is a vector of the measurement at time t . U_t is a vector of uncertain forcing inputs while θ_t is vector of the model parameters at time t . W_t represents the model errors while V_t is the measurement error. In most cases, W_t and V_t are assumed as independent and white noises with mean zero and covariance respectively for the state vector and measurement vector.

In general, the dual EnKF requires two separate state-space representation for the state and parameters through two interactive filters by updating model parameters and model states. In the study, model parameters are first updated and then states. The simulation results from each step are considered separately. The detailed procedure for the dual EnKF can be referred to (Samuel et al., 2014).

4. Experimental Design

The dual EnKF technique is applied to determine additional improvements that can be obtained to update the SVM for soil moisture estimation. Multiple experiments are designed to construct and validate the SVM model; then additional experiments are conducted to examine the effectiveness and robustness of EnKF technique to improve the performance of SVM.

4.1 Preliminary Data Analysis

It is well known that the soil moisture can be estimated using low-level atmospheric and meteorological inputs (Mahfouf, 1991; Gill and McKee, 2007; Yu et al., 2012). Many previous studies (e.g., Koster et al., 2004; Cook et al., 2006; Koster

et al., 2006; Liu et al., 2014b) indicated the importance of soil moisture during the process of land-atmosphere interactions. The changes of soil moisture would lead to a chain of reactions, including changes of albedo, net radiation, latent heat flux, sensible heat flux, boundary layer heat and moist static energy density.

The raw data are preprocessed to obtain the input variables for the SVM model. The monthly time series plot of mean daily soil moisture and soil temperature (within 1 meter depth), mean air temperature and total precipitation ([Figure S1, available in supplementary document](#)) indicates that the soil moisture has a positive but lagged time response to the precipitation and a negative response to the air and soil temperature. This is reasonable according to the land-atmosphere interaction mechanisms: the soil moisture has a direct impact on the surface air temperature through evapotranspiration while the relationship with precipitation is much more complicated (Koster et al., 2004; Cook et al., 2006; Koster et al., 2006; Liu et al., 2014b). The correlations between current daily soil moisture with subsequent 1-day lead monthly accumulated precipitation and average air temperature ([Table S1](#)) further emphasize the relationship between soil moisture and precipitation/air temperature. As shown, the correlations between soil moisture and precipitation are weak at each layer in each season accompanied with a negative soil moisture-air temperature signal in spring (MAM), summer (JJA) and winter (DJF) seasons, implying weak soil moisture-precipitation feedback at each season.

The principal component analysis (PCA) was applied to determine the input variables for SVM. This method has been greatly used in the hydrology (e.g., Westra et al., 2007; Ssegane et al., 2012; Dronova et al., 2015). It provides a representation of a multivariate data set using the information that is contained within the covariance matrix, so that the extracted components are mutually uncorrelated. The correlation coefficients combined with the multi-scale principal component analysis were used to select relevant input variables ([Table S2](#)). The t_max, SR, sur_temp_max, rh_max, rh_min, rh_avg, ST1, ST2, ST3, ST4 and ST5 variables stand out as the main factors for soil moisture estimation. Therefore, these variables are chosen as the input variables for the SVMs to estimate the soil moisture at multiple layers, (i.e., SM1, SM2, SM3, SM4 and SM5).

4.2 SVM Construction and Validation Experiments

A set of experiments(Exp1 - Exp6 shown in [Table 1](#)) with different inputs, model parameters, length of training set and cost function are designed to evaluate the performance of SVM in the multi-layer soil moisture estimation.

As shown, experiments 1-4 are conducted with the same cost and kernel function but different kernel parameters and regularization parameter (Exp2), cost function parameters (Exp3) and input variables (Exp4); Exp5 and Exp6 compare the model with different initial training lengths and cost functions. Exp5 uses the “cross-validation” function while Exp6 applies the “leave-one-out” function. To avoid the degradation of model performance, the models in Exp5 and Exp6 are updated every 10 time steps (days) by incorporating the estimated soil moisture in the training set. Meanwhile, the initial 10 steps training data is deleted to keep the same training length.

4.3 EnKF Updating Experiments

EnKF technique uses statistical distributions to represent uncertainties of both observations and model errors and to generate ensembles of model forcing and other variables (Evensen, 2003; Samuel et al., 2014). Four sensitivity experiments (Exp7 - Exp10 shown in [Table 1](#)) are conducted to illustrate the ability of EnKF to improve the SVM soil moisture estimation. Both the model state and parameters are updated with EnKF technique in each experiment. Initial 500 days (September 1, 2009 to January 13, 2011) data were used as the training sets in Exp7 - Exp9. These experiments are designed based on the following considerations: Exp7 aims to examine the effectiveness of EnKF technique to improve SVM estimations at each time step with observed soil moisture at each layer; Exp8 is designed to test the improvement of EnKF technique based on the time step when there is observed data; while Exp9 aims to reduce the computational resources by updating the SVM with EnKF technique after certain time steps with 50% ratio of observed soil moisture data within these time steps. To further investigate the effectiveness of EnKF technique to the initial length of SVM training set, Exp10 is conducted with different initial training lengths, including 100 days, 200 days, 300 days, and 400 days started from September 1, 2009, while the EnKF updating time scale is set to 50 days with 50% ratio of observed soil moisture data randomly selected within this 50-day. In this paper, the 50% ratio of observed soil moisture data at a fixed time scale means that if the fixed

time scale is 50-day, then there are 25 days of observed soil moisture data available within this 50-day. The procedure of each experiment is described in [Table 2](#).

The flowchart for the SVM model and EnKF technique is presented in [Figure 1](#). It contains three processes: training, estimating and updating. The whole data set are divided into two parts: training and the testing sets. There are two methods to update the SVM: one directly uses the soil moisture estimates in the later time periods to update the SVM (i.e., Exp5- Exp6), while the other applies the EnKF technique combined with the observed data to update the SVM (i.e., Exp7- Exp10).

The EnKF update process consists of two processes. Firstly, the model state and input forcing data are updated, which would give one estimation (Y_d), seen as the results from 1st update for the following analysis. Then the observed data is used to update the posterior Y_d with EnKF technique to get the final estimation (Y_{du}), which would be analyzed as the results from 2nd update. The estimation from the SVM model (Y_p) is marked as the results from the open loop. It should be noticed that the estimations from open loop are different from the SVM estimations without EnKF technique. The result from open loop has the model updating with EnKF technique in previous time steps.

In summary, the SVM model and EnKF update can be simply described in three steps. Firstly, the training process, which applies the atmospheric and soil temperature variables as inputs to train the SVM multi-layer soil moisture (as outputs) model at the same time steps. Here, the soil moisture data are not used as the input variables, which is different from previous studies (e.g., Gill and McKee, 2007). Secondly, the estimation process, where the SVM was applied to estimate soil moisture at different layers. Thirdly, the updating process which applies the dual EnKF method to update the soil moisture estimates from the second step with respect to the observed data.

4.4 Evaluation Methods

The goodness of fit (GOF) statistics: mean absolute error (MAE), root mean square error ($RMSE$), and correlation coefficient (R) are used to assess the performance of the SVM-EnKF model combinations. The MAE and $RMSE$ measure the accuracy of the mean estimations to the observations. These criteria are calculated according to formulas (3)-(5) separately.

$$MAE = \frac{\sum_{i=1}^n |y_i - y_i^*|}{n} \quad (3)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - y_i^*)^2}{n}} \quad (4)$$

$$R = \frac{\sum_{i=1}^n y_i y_i^* - \sum_{i=1}^n y_i \sum_{i=1}^n y_i^* / n}{\sqrt{\left[\sum_{i=1}^n y_i^2 - \left(\sum_{i=1}^n y_i \right)^2 / n \right]} \sqrt{\left[\sum_{i=1}^n (y_i^*)^2 - \left(\sum_{i=1}^n y_i^* \right)^2 / n \right]}} \quad (5)$$

Where n is the number of samples, and y_i and y_i^* are the observed and estimated values, respectively. The smaller the MAE and $RMSE$ are and the larger R is, the better the model performs. Here, if R is greater than 0.5, the model is considered satisfactory (Al-Hamdan and Cruise, 2010; Liu et al., 2010; Yu et al., 2012). The $RMSE$ less than 0.06 is considered reasonable in soil moisture estimation (Shibata et al., 2003).

5. Analysis of Results

5.1 SVM Model Construction and Validation

Experiments 1-6 are designed with different input variables, model parameters, ensemble size, and cost function to validate the SVM in the multi-layer soil moisture estimation. The GOF statistics for training and estimation processes from Exp1-Exp4 are given in Table 2. Based on the MAE , $RMSE$ and R criteria (Table 2), each model performs well to approximate the true observed soil moisture at each soil layer at the training step (initial 730 days, Date: September 1, 2009 to August 31, 2011) but has obvious bias at the estimating (validation) step (Date: September 1, 2011 to May 31, 2015). In the training step, each experiment presents high correlation coefficient (the magnitude of R higher than 0.8 for SM1, SM2, and SM3, while higher than 0.7 for SM4 and SM5) and low MAE (lower than 0.02 at each layer) and $RMSE$ (lower than 0.02 at each layer except SM5 where $RMSE$ lower than 0.04) values (Table 2). In the estimating step, the magnitude of R is lower than 0.5 at each layer except for SM1 while the MAE and $RMSE$ values are higher than that in the training step (greater than 0.02, but lower than 0.05 at each layer for both MAE and $RMSE$) (Table 2), indicates poor estimation ability.

The time series of soil moisture estimations from each experiment (Exp1-Exp4) versus the observations at each soil layer and the precipitation (mm/day) are plotted in Figure 2. As shown, the SVM works well in the training process (1-730 days) where

the estimated soil moisture almost resemble the similar trend and magnitude with the observed data for most of the soil layers except obvious bias exists at the 50cm (SM4) and 100cm (SM5) layers, where the SVM comparatively overestimates the soil moisture during the 400-500 days. Unlike the good performance in the training process, the performance in the estimating process (from 731 day to 2090 day) has obvious bias with the observations. The estimated soil moistures tend to underestimate the observed soil moisture during the time of higher precipitation while overestimate when there is less precipitation at most time steps for each layer. In reality, the precipitation has good connection with soil moisture. The soil moisture-precipitation feedback is complicated and involves many procedures in the land-atmosphere interactions (e.g., Cook et al., 2006; Koster et al., 2006; Liu et al., 2014a; Liu et al., 2014b). The influence of precipitation on soil moisture is direct and easy to understand while the impact of soil moisture on precipitation is complex. The evapotranspiration is generally seen as the linkage between soil moisture and precipitation (e.g., Guo et al., 2006). The sensitivity of soil wetness is related to the uncertainties in precipitation and radiation forcings (Wei et al., 2008). The precipitation sensitivity is dominated in cold climate; while in warm climate, the radiation is dominant or both are high. It is possible that soil moisture would be influenced under extreme precipitation conditions.

The time series of deviation between the estimated soil moisture from each experiment with the observations at each soil layer (Figure S2) further demonstrate the performance of SVM with different input variables and parameter settings. The results from different experiment agree well with each other. The deviation for each experiment is within absolute magnitude of 0.06 at the training step while it increases to be greater than 0.1 at several time steps during the estimating process (Figure S2).

As analyzed, the performance of SVM differs with different input variables, cost function, kernel parameter and regularization parameter. Therefore, the model from Exp1 with KBF kernel function, cross validation error function with *MSE* value set as default (10), regularization parameter (gam) and squared kernel parameter (sig2) auto selected are used in the following experiments.

Exp5 validates the performance of SVM with different lengths of training set (from 50-day to 2050-day with 50-daytime interval). The GOF statistics (*MAE*, *RMSE* and *R*) for different lengths of training set is shown in Figure 3. The magnitude of *R* gradually increases while the magnitude of *MAE* and *RMSE* slowly decrease (not as

obvious as R but stable) with the increased training length. This indicates the improved performance of SVM with the increased training length. As shown, the R value tends to decrease at the same training length with the deeper soil moisture layers, implying the ability of SVM to track the soil moisture would reduce when the soil depth increases. For instance, for the surface layer (i.e., SM1), the R value can reach 0.5 when the training length is longer than 1200 days; for the SM2, the R value can reach 0.5 with the training length longer than 1300 days. For SM3, SM4 and SM5, the training length should be longer than 1700 days, 1850 days and 2000 days, respectively, for the R value to be greater than 0.5. Compared to the R values, the difference for the MAE and $RMSE$ trends with training length increases are not as obvious as the R values. But both of them show better results with longer training length, especially for the root zone. These results demonstrate that the capability of SVM relies on the training length. As stated, to get an efficient SVM model, there should be long enough time series of initial training set, however, with the increase in soil depth, longer training time series are needed. In this way, the training length should be carefully considered when using SVM model.

The time series of estimated soil moisture with different lengths of training set (500, 1000, 1500 and 2000) versus the observations at each layer and the precipitation (shown in [Figure 4](#)) further analyzes the performance of SVM with different training length. Consistent with Exp1- Exp4, the simulations from each model with different training length have an overestimation trend in the drought period with little precipitation and underestimation trend during the higher precipitation events at most time steps at each soil layer. By comparison, the simulations with longer lengths of initial training set seem to reduce the bias.

Exp6 is designed to investigate the performance of SVM with leave-one-out cost function. The difference between Exp6 and Exp5 were not obvious assessed by the GOF statistics ([Tables S3-S4](#)) and the time series of deviation of soil moisture estimations at each soil layer ([Figure S3](#)). The absolute magnitude of bias between Exp6 and Exp5 is within 0.005 for MAE , $RMSE$ and R , respectively, with the same initial training length. But the computing time for the leave-one-out experiment is almost 10 times higher than the cross-validation cost function; therefore, the cross-validation cost function is more suitable in comparison to the leave-one-out.

5.2 EnKF Updating Results Analysis

We designed four experiments to evaluate dual EnKF DA technique to determine additional improvements that can be achieved beyond the SVM model.

(a) Effectiveness of EnKF Technique

Exp7 is conducted to update SVM estimations using EnKF technique at each time step; while Exp8 is updated at the time steps with available observed data. In reality, Exp8 is more reliable according to the limited observed environment. Moreover, Exp8 has the merit with less computational cost, which is one main advantage in complex spatial and temporal soil moisture estimation problems.

The GOF statistics for Exp7 and Exp8 from combined SVM and EnKF model are provided in Table 3. In each experiment, the EnKF technique reduces the *MAE* and *RMSE* magnitude while increases *R* magnitude for soil moisture at each layer at each updating step (including the open loop, 1st update, 2nd update) compared to the SVM control run, indicating the effectiveness of EnKF technique in the soil moisture estimation. For example in Exp7, the performance measure in the open loop, 1st update and 2nd update has lower *MAE* and *RMSE* values and higher *R* values than the results from SVM control run without DA. By comparison, the 2nd update is the best among the three updating steps, while the 1st update is better than the open loop in each experiment. This verifies that the dual EnKF technique can achieve better improvement than the EnKF updated only at the model state or the posterior estimation updating step.

Consistent with the GOF statistics, the time evolution of estimated soil moisture at each layer from Exp7 agrees well with the trend of observation based on the SVM control run (Figure 5). In addition, the 2nd update estimations are better than the 1st update and open loop, which almost resemble the true observed soil moisture data at each layer. The deviation between the estimated soil moisture from Exp7 at each updating state with the observed soil moisture at each soil layer (Figure S4) further indicate the improvements of dual EnKF technique in soil moisture estimation. As shown, the SVM control run has obvious bias (greater than 0.06) to the observation at different part of the model validation periods (1-1599 days from January 14, 2011 to May 31, 2015) (e.g., 400~800 days for SM1, 800~1600 days for SM2, 100~1600 days for SM4 and 600~1200 days for SM5). The bias seems to grow at deeper soil layer, for instance, the soil moisture bias at 100cm (SM5) is over 0.06 at most of the validation period, especially during 300~1600 days. By comparison, the updated estimation

from EnKF technique at each update and at each time step reduces the bias with observed data at each soil layer. This is especially for the 2nd update, which is much better than the 1st update and the open loop with the minor bias (within 0.06 at every estimating time step). [Figure S4](#) clearly demonstrates the effectiveness of EnKF techniques from the decrease of absolute deviation magnitude at almost every time step in the validation periods.

The scattered plot ([Figure S5](#)) demonstrates that the SVM control run deviates from the 1:1 line slope while EnKF updated estimations converge to fit the 1:1 line at each soil layer. Among the three updating steps of EnKF technique, the 2nd update is better than the 1st update while the 1st update is better than the open loop, confirming the advantage of dual EnKF in the improvement of soil moisture estimation at both surface and root zone. The results from Exp8 ([Figure S6- Figure S8](#)) also demonstrate the improvement of dual EnKF technique in the soil moisture estimation. Unlike Exp7, the updates from model state (open loop) to posterior (2nd update) differ little with the EnKF technique.

(b) Uncertainties with Updating Ensemble Size

Considering the higher computational time required for DA technique, Exp9 is designed to save computational time by updating the SVM only at a fixed time interval. Here, for each time interval, the observed data are randomly selected with a probability of 0.5; therefore, the updating EnKF ensemble size is half of the time interval. For instance, if the time interval is 10, there are only 5 observed data are selected for updating the EnKF at each time period.

The iteration results of GOF statistics at each updating time interval are given in [Figure 6](#). As shown, the results from the 2nd update are better than the open loop and 1st update for each soil layer and at each time interval based on the decrease in *MAE* (*RMSE*) and increase in *R* magnitude. The update from the open loop and 1st update differs little as the criteria at each time interval almost resemble the same magnitude. These results imply the efficiency of dual EnKF technique at flexible updating time period.

Based on [Figure 6](#), the *MAE* and *RMSE* tend to decrease while the *R* tends to increase at each soil layer from 5cm to 100cm with the increase in time interval. In addition, the gradient for each GOF criteria increases in the first five steps when the ensemble size changes from 10 to 20, to 30, to 40 and to 50, and then grows gradually at a slower rate. These results indicate that the efficiency of EnKF technique tends to

be stable with the increase in ensemble size. Besides, the EnKF achieve better improvement when the updating size increases from time interval 10 to 20 to 30 (with decreasing *MAE* and *RMSE* while increasing *R* magnitude) and then grows slowly since time interval from 50 to higher. As stated, the updating ensemble size at each updating is half the time interval; thus, the EnKF can reach its best efficiency at the size around 15 in this study. This finding is almost consistent with previous literature (Yin et al., 2015) which indicates that the EnKF technique can reach its maximum efficiency when the ensemble size is 12. Similarly, (Kumar et al., 2008) observed that ensemble sizes in excess of 10 provide little additional benefit with the EnKF's sensitivity to the size of the ensemble for the catchment soil moisture assimilation experiment. As stated in (Yin et al., 2015), this optimal ensemble size may have theoretical implications on the implementation of EnKF in other sequential data assimilation problems. According to (Reichle et al., 2002), the error variance for larger ensemble sizes does not decrease much further, implying that the EnKF estimates may not converge to the optimal smoothing solution when the ensemble size becomes too large. This study helps to illustrate the maximum effectiveness of EnKF technique with different ensemble size.

Consistent with the GOF criteria, the time evolution of estimated soil moisture with EnKF updating at the 10, 30, 100 and 200 time interval at each soil layer (Figure 7) further demonstrates the effectiveness of EnKF technique with the increased ensemble size. As shown, the time evolution of estimated soil moisture with different ensemble size matches well with each other at most of the time steps at each layer, though the results from 100 or 200 time interval resembled better with the observed data.

Figure 8 compares the soil moisture estimations from Exp7-Exp9 with the SVM outputs at each layer. Clearly, the dual EnKF technique improves the SVM model at each time step with different EnKF updating methods. The Exp7 and Exp8 outweigh the Exp9 as it matches well with the true observations at each layer. The deviation between estimated soil moisture from each experiment (Exp7-Exp9) with observation decreases at most time periods at each soil layer compared to the SVM control run (FigureS9). Among the three experiments, Exp7 approximates the observations greatly with the smallest absolute deviation and each deviation is within 0.06 during the whole time period at each layer; Exp8 has better performance than Exp9, which can capture the observation at most time steps with the absolute deviation within

0.06. Unlike Exp7 and Exp8, Exp9 has some obvious divergence with the observation, even though, Exp9 improves the SVM modeling by reducing the high model bias (e.g., SM1, SM4 and SM5 from 800 to 1600 time steps). The goodness of fit results of Exp9 at the time interval of 100 are given in Table 3. Compared with SVM control run, Exp9 absolutely improves the soil moisture estimation at each soil layer from 5cm (SM1) to 100cm (SM5) based on the decreased *MAE* (*RMSE*) and increase in *R* magnitude.

(c) Uncertainty Analysis with Different Initial SVM Training Size

Previous literatures (e.g., Ridler et al., 2014; Samuel et al., 2014) noted that the DA technique can improve model performance even with limited initial length of training set. To evaluate the ability of EnKF technique with limited training length, Exp10 was designed with four different initial training lengths (i.e., 100, 200, 300 and 400 days). According to the results from Exp9, the efficiency of EnKF differs little when the updating ensemble size exceeds 15. Here, to save the computational resources, the EnKF technique updating time interval is set to 50 days.

The criteria of *MAE*, *RMSE* and *R* from different training size with and without EnKF updating is shown in Figure 9. Clearly, the EnKF technique improves the SVM model at each update step from the open loop, 1st update to the 2nd update with different initial training lengths. Moreover, with the increase of training length, the effectiveness of EnKF technique tends to be better. This is obvious when the length increases from 100 days to 200 days. Consistent with Exp7 to Exp9, the 2nd update is better than the open loop and 1st update resulted from the smaller *MAE* and *RMSE* values and higher *R* values in each modeling with different training length.

The time series plot of estimated soil moisture from Exp10 with different lengths of initial training set versus the observations at each layer plotted in Figure 10, which indicates the effectiveness of EnKF technique with different initial training length. The performance improves greatly from 100 days to 200 days, but then slows down with no obvious improvement with increase in training size. It can be observed that the length of initial training set will have an influence on quality of soil moisture estimation (Figure S10). The deviations among the longer lengths (i.e., 300, 400 and 500 days) differ little while the deviations with shorter training length (i.e., 100 and 200 days) are higher at some time steps. The deviation for 100 days and 200 days training length have higher divergence (absolute bias more than 0.1 at some time steps) while the deviations with other initial SVM training lengths (300 days, 400 days and

500 days) are smaller and stable within an interval of 0.06 at most of the time steps for each soil layer. This implies the better performance of EnKF technique when the initial SVM training length exceeds certain amount, such as 300 days in this experiment.

As discussed above, the EnKF technique is effective to improve the SVM model in multi-layer soil moisture estimations with limited training length. But the performance is better when the length exceeds certain threshold because a longer time period of initial training data can help capture the inter-annual and intra-seasonal relationships between soil moisture and climate variables.

6. Conclusions

We found that the combination of SVM and dual EnKF is a robust statistical tool suitable for the multi-layer soil moisture estimation. Based on multiple experiments, the major findings are follows: (1) The SVM is good at simulating the soil moisture at different layer from surface to root zone during the training process but the performance during validation lacks confidence especially for the deeper soil layers. (2) The performance of SVM relies most on the length of initial training set compared with other factors (e.g., cost function, regularization parameter, and kernel parameters). The ability of SVM can be improved with longer period of training sets and specifically for the deeper soil layers. (3) The estimated soil moisture was influenced by the precipitation pattern. The soil moisture estimations at each layer tend to overestimate when there is little precipitation while underestimate when there is more precipitation. (4) The dual EnKF technique for the model state and parameter state is better than the DA update at the model state or parameter state only. (5) The EnKF technique can reach its maximum efficiency when the updating training length approaches around 15, which is consistent with the previous findings (Yin et al., 2015), and (6) The EnKF technique can improve the SVM even with limited initial training time series.

Acknowledgement

This study is supported by the United States Department of Agriculture (USDA) award 2015-68007-23210. We very much appreciate editor and three reviewer's valuable comment that helped us to improve the manuscript.

References

- Al-Hamdan, O.Z., Cruise, J.F., 2010. Soil moisture profile development from surface observations by principle of maximum entropy. *J. Hydrol. Eng.* 15, 327-337.
- Boni, G., Entekhabi, D., Castelli, F., 2001. Land data assimilation with satellite measurements for the estimation of surface energy balance components and surface control on evaporation. *Water Resour. Res.* 37, 1713-1722.
- Castelli, F., Entekhabi, D., Caporali, E., 1999. Estimation of surface heat flux and an index of soil moisture using adjoint-state surface energy balance. *Water Resour. Res.* 35, 3115-3126.
- Clark, M.P., Rupp, D.E., Woods, R.A., et al., 2008. Hydrological data assimilation with the ensemble Kalman filter: Use of streamflow observations to update states in a distributed hydrological model. *Adv. Water Resour.* 31, 1309-1324.
- Cook, B.I., Bonan, G., Levis, S., 2006. Soil moisture feedbacks to precipitation in southern Africa. *J. Climate* 19, 4198-4206.
- Das, N., Entekhabi, D., Njoku, E., 2011. Algorithm for merging SMAP radiometer and radar data for high resolution soil moisture retrieval. *IEEE T. Geosci. Remote Sens.* (IEEE Transactions in Geoscience and Remote Sensing) 49, 1504-1512.
- Das, N.N., Mohanty, B.P., 2006. Root zone soil moisture assessment using remote sensing and vadose zone modeling. *Vadose Zone J.* 5, 296-309.
- De Brabanter, K., De Brabanter, J., Suykens, J.A.K., et al., 2010. Approximate Confidence and Prediction Intervals for Least Squares Support Vector Regression. *IEEE T. Neural Networ.* (IEEE Transactions on Neural Networks) 22, 110-120.
- Dronova, I., Gong, P., Wang, L., et al., 2015. Mapping dynamic cover types in a large seasonally flooded wetland using extended principal component analysis and object-based classification. *Remote Sens. Environ.* 158, 193-206.
- Dumedah, G., Coulibaly, P., 2012. Evolutionary assimilation of streamflow in distributed hydrologic modeling using in-situ soil moisture data. *Adv. Water Resour.* 53, 231-241.
- Entekhabi, D., et al., 2010. The Soil Moisture Active Passive (SMAP) mission. *Proc. IEEE* 98, 704-716.
- Evensen, G., 2003. The ensemble Kalman filter: theoretical formulation and practical implementation. *Ocean. Dynam.* 53, 343-367.
- Gill, M.K., Asefa, T., Kemblowski, M.W., et al., 2006. Soil moisture prediction using support vector machines. *J. Am. Water Resour. Assoc.* 42, 1033-1046.
- Gill, M.K., McKee, M., 2007. Soil moisture data assimilation using support vector machines and ensemble Kalman filter. *J. Am. Water Resour. Assoc.* 43, 1004-1015.
- Guo, Z.C., Dirmeyer, P.A., Koster, R.D., et al., 2006. GLACE: The global land-atmosphere coupling experiment. Part II: Analysis. *J. Hydrometeor.* 7, 611-625.
- Han, X., Franssen, H.-J.H., Montzka, C., et al., 2014. Soil moisture and soil properties estimation in the Community Land Model with synthetic brightness temperature observations. *Water Resour. Res.*

555 50, 6081-6105.

556 Kaheil, Y.H., Gill, M.K., McKee, M., et al., 2008. Downscaling and assimilation of surface soil
557 moisture using ground truth measurements. *IEEE Transactions on Geoscience and Remote Sensing*
558 (*IEEE Trans Geosci Remote Sens*) 46, 1375-1384.

559 Kerr, Y.H., Waldteufel, P., Wigneron, J.-P., et al., 2010. The SMOS mission: New tool for monitoring
560 key elements of the global water cycle. *Proc. IEEE* 98, 666-687.

561 Kornelsen, K.C., Coulibaly, P., 2014. Root-zone soil moisture estimation using data-driven
562 methods. *Water Resour. Res.* 50, 2946-2962.

563 Koster, R.D., Guo, Z.C., Dirmeyer, P.A., et al., 2006. GLACE: The global land-atmosphere coupling
564 experiment. Part I: Overview. *J. Hydrometeor.* 7, 590-610.

565 Koster, R.D., Guo, Z.C., Dirmeyer, P.A., et al., 2004. Regions of strong coupling between soil
566 moisture and precipitation. *Science* 305, 1138-1140.

567 Kumar, S.V., Riechle, R.H., Peters-Lidard, C.D., et al., 2008. A land surface data assimilation
568 framework using the land information system: Description and applications. *Adv. Water Resour.* 31,
569 1419-1432.

570 Leisenring, M., Moradkhani, H., 2012. Analyzing the uncertainty of suspended sediment load
571 prediction using sequential data assimilation. *J. Hydro.* 468-469, 268-282.

572 Li, B., Toll, D., Zhan, X., et al., 2012. Improving estimated soil moisture fields through assimilation
573 of AMSR-E soil moisture retrievals with an ensemble Kalman filter and a mass conservation constraint.
574 *Hydrol. Earth Syst. Sci.* 16, 105-119.

575 Liu, D., Wang, G.L., Mei, R., et al., 2014a. Impact of initial soil moisture anomalies on climate
576 mean and extremes over Asia. *J. Geophys. Res.* 119, 1-14.

577 Liu, D., Wang, G.L., Mei, R., et al., 2014b. Diagnosing the strength of land-atmosphere coupling at
578 sub-seasonal to seasonal time scales in Asia. *J. Hydrometeor.* 15, 320-339.

579 Liu, D., Yu, Z., Lv, H.S., 2010. Data assimilation using support vector machines and ensemble
580 Kalman filter for multi-layer soil moisture prediction. *Water Science and Engineering* 3, 361-377.

581 Lu, H., Li, X., Yu, Z., et al., 2010. Using a H1filter assimilation procedure to estimate root zone soil
582 water content. *Hydrol. Processes* 24, 3648-3660.

583 Mahfouf, J.F., 1991. Analysis of Soil Moisture From Near-Surface Parameters: A Feasibility Study.
584 *Journal of Applied Meteorology* 30, 1534-1547.

585 Milly, P.C.D., Betancourt, J., Falkenmark, M., et al., 2008. Stationarity is dead: Whither water
586 management? *Science* 319, 573-574.

587 Mishra, A.K., Ines, A.V.M., Das, N.N., et al., 2015. Anatomy of a local-scale drought: Application of
588 assimilated remote sensing products, crop model, and statistical methods to an agricultural drought
589 study. *J. Hydro.* 526 15-29.

590 Montzka, C., Moradkhani, H., Weihermüller, L., et al., 2011. Hydraulic parameter estimation by
591 remotely-sensed top soil moisture observations with the particle filter. *J. Hydro.* 399, 410-421.

592 Moradkhani, H., 2008. Hydrologic remote sensing and land surface data assimilation. *Sensors* 8,

2986-3004.

Moradkhani, H., Hsu, K.-L., Gupta, H., et al., 2005. Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the particle filter. *Water Resour. Res.* 41, W05012.

Ni-Meister, W., 2008. Recent advances on soil moisture data assimilation. *Phys. Geogr.* 29, 19-37.

Njoku, E.G., Jackson, T.J., Lakshmi, V., et al., 2003. Soil moisture retrieval from AMSR-E. *IEEE Trans. Geosci. Remote Sens.* 41, 215-229.

Reichle, R.H., McLaughlin, D.B., Entekhabi, D., 2002. Hydrologic data assimilation with the ensemble Kalman filter. *Mon. Weather Rev.* 130, 103-114.

Ridler, M.-E., Madsen, H., Stisen, S., et al., 2014. Assimilation of SMOS-derived soil moisture in a fully integrated hydrological and soil-vegetation-atmosphere transfer model in Western Denmark. *Water Resour. Res.* 50, 8962-8981.

Samuel, J., Coulibaly, P., Dumedah, G., et al., 2014. Assessing model state and forecasts variation in hydrologic data assimilation. *J. Hydro.* 513, 127-141.

Shibata, A., Imaoka, K., Koike, T., 2003. AMSR/AMSR-E level 2 and 3 algorithm developments and data validation plans of NASDA. *IEEE Trans. Geosci. Remote Sens.* 41, 195-203.

Ssegane, H., Tollner, E.W., Mohamoud, Y.M., et al., 2012. Advances in variable selection methods I: Causal selection methods versus stepwise regression and principal component analysis on data of known and unknown functional relationships. *J. Hydro.* 438-439, 16-25.

Sun, J., Salvucci, G., Entekhabi, D., et al., 2011. Parameter Estimation of Coupled Water and Energy Balance Models Based on Stationary Constraints of Surface States. *Water Resour. Res.* 47.

Tran, A.P., Vanclooster, M., Lambot, S., 2013. Improving soil moisture profile prediction from ground-penetrating radar data: A maximum likelihood ensemble filter approach. *Hydrol. Earth Syst. Sci.* 17, 2543-2556.

Vapnik, V.N., 1995. *The nature of statistical learning theory*. Springer, New York.

Vereecken, H., Huisman, J.A., Boga, H., et al., 2008. On the value of soil moisture measurements in vadose zone hydrology: A review. *Water Resour. Res.* 44, W00D06.

Walker, J.P., Willgoose, G.R., Kalma, J.D., 2001. One-dimensional soil moisture profile retrieval by assimilation of near-surface measurements: A simplified soil moisture model and field application. *J. Hydrometeorol.* 2, 356-373.

Wei, J.F., Dirmeyer, P.A., Guo, Z.C., 2008. Sensitivities of soil wetness simulation to uncertainties in precipitation and radiation. *Geophys. Res. Lett.* 35.

Westra, S., Brown, C., Lall, U., et al., 2007. Modeling multivariable hydrological series: Principal component analysis or independent component analysis? *Water Resour. Res.* 43, W06429.

Yin, J., Zhan, X., Zheng, Y., et al., 2015. Optimal ensemble size of ensemble Kalman filter in sequential soil moisture data assimilation. *Geophys. Res. Lett.* 42.

Yin, J., Zhan, X., Zheng, Y., et al., 2014. Impact of quality control of satellite soil moisture data on their assimilation into land surface model. *Geophys. Res. Lett.* 41, 7159-7166.

631 Yu, Z., Liu, D., Lü, H., et al., 2012. A multi-layer soil moisture data assimilation using support
632 vector machines and ensemble particle filter. J. Hydro. 475 53-64.
633
634

Table Captions:

Table 1: Experiments construction.

Table 2: *MAE*, *RMSE* and *R* criteria results for Exp1 to Exp4.

Table 3: *MAE*, *RMSE* and *R* criteria results for Exp7 to Exp9.

Figure captions:

Fig.1: Flowchart of SVM modeling and EnKF technique updating schemes.

Fig.2: Time series of soil moisture estimations from Exp1 to Exp4 versus the observations at each soil layer from 5cm (SM1), 10cm (SM2), 20cm (SM3), 50cm (SM4) and 100cm (SM5). Exp1, Exp2, Exp3, Exp4 are shown in magenta line, blue line, red line and green line, respectively. Observed soil moisture is shown in black line. The blue line in bottom pattern is the precipitation (mm/day) during the same period. X-axis is the time step in days with the training length from 1 (September 1, 2009) to 730 (August, 31, 2011) (left side) and the estimation from 731 (September 1, 2011) to 2099 (May 31, 2015) (right side) divided by the red dashed line.

Fig.3: Criteria of *MAE*, *RMSE* and *R* for estimated soil moisture at each soil layer (5cm (SM1 black star line), 10cm (SM2 blue line), 20cm (SM3 red circle line), 50cm (SM4 green diamond line), 100cm (SM5 magenta squared line)) from Exp5 with the initial training lengths from 50-day to 2050-day with an increase interval of 50-day.

Fig.4: Time series of estimated soil moisture from Exp5 with different initial training lengths versus the observed soil moisture (black line) at multi-layers: 5cm (SM1), 10cm (SM2), 20cm (SM3), 50cm (SM4) and 100cm (SM5). The initial training lengths are 500-day (magenta line), 1000-day (cyan line), 1500-day (red line) and 2000-day (green line). The blue line in the bottom pattern is the precipitation (mm/day) during the same period. X-axis is time step for estimation started from 1 (January 14, 2011) to 1599 (May 31, 2015).

Fig.5: Time series of soil moisture estimations at each updating step (open loop (cyan line), 1st update (red line), 2nd update (green line)) from Exp7 with EnKF technique versus the estimations from SVM without EnKF technique (magenta line) and the observed soil moisture (black line) at multi-layers: 5cm (SM1), 10cm (SM2), 20cm (SM3), 50cm (SM4) and 100cm (SM5). The blue line in the bottom pattern is the precipitation (mm/day) during the same period. X-axis is time step for estimation started from 1 (January 14, 2011) to 1599 (May 31, 2015).

Fig.6: Criteria of *MAE*, *RMSE* and *R* for soil moisture estimations at each soil layer (5cm (SM1), 10cm (SM2), 20cm (SM3), 50cm (SM4), 100cm (SM5)) at each updating step (open loop (blue + line), 1st update (red circle line) and 2nd update (green diamond line)) from Exp9 with EnKF technique updated at different time interval from 10-day to 200-day with an increase interval of 10-day.

Fig.7: Time series of soil moisture estimations from Exp9 with EnKF technique updated at different time interval (10-day (magenta line), 30-day (cyan line), 100-day (red line) and 200-day (green line)) versus the observed soil moisture (black line) at multi-layers: 5cm (SM1), 10cm (SM2), 20cm (SM3), 50cm (SM4) and 100cm (SM5). The blue line in the bottom pattern is the precipitation (mm/day) during the same period. X-axis is time step for estimation started from 1 (January 14, 2011) to 1599

(May 31, 2015).

Fig.8: Time series of soil moisture estimations in the 2nd updating process from Exp7 (cyan line), Exp8 (red line) and Exp9 (green line) with EnKF technique versus the SVM estimations without EnKF technique (magenta line) and observed soil moisture (black line) at multi-layers: 5cm (SM1), 10cm (SM2), 20cm (SM3), 50cm (SM4) and 100cm (SM5). The blue line in the bottom pattern is the precipitation (mm/day) during the same period. X-axis is time step for estimation started from 1 (January 14, 2011) to 1599 (May 31, 2015).

Fig.9: Criteria of *MAE*, *RMSE* and *R* for estimated soil moisture at each updating step with EnKF technique (open loop (blue +line), 1st update (red circle line), 2nd update (green diamond line)) from Exp10 with the initial training lengths from 100-day to 500-day with an increase interval of 100-day versus the estimations of SVM without EnKF technique (black star line) at each soil layer (5cm (SM1), 10cm (SM2), 20cm (SM3), 50cm (SM4), 100cm (SM5)).

Fig.10: Time series of soil moisture estimations in the 2nd updating process from Exp10 with EnKF technique at different initial training lengths (100-day (magenta line), 200-day (cyan line), 300-day (red line), 400-day (green line) and 500-day (blue line)) versus the observed soil moisture (black line) at multi-layers: 5cm (SM1), 10cm (SM2), 20cm (SM3), 50cm (SM4) and 100cm (SM5). The black dot line in the bottom pattern is the precipitation (mm/day) during the same period. X-axis is time step for estimation started from 1 (January 14, 2011) to 1599 (May 31, 2015).

Tables:

Table 1: Experiments construction.

Name	Input	Output	Kernel Function	Kernel Parameter	Length of Training Set	Tuning	Regularization Parameter	Cost Function	Model Update
Exp1	B	AA	K	S1	730	T1	G1	C1	M1
Exp2	B	AA	K	S2	730	T1	G2	C1	M1
Exp3	B	AA	K	S1	730	T1	G1	C2	M1
Exp4	H	AA	K	S1	730	T1	G1	C1	M1
Exp5	B	AA	K	S1	50:50:2000	T1	G1	C1	M2
Exp6	B	AA	K	S1	500:500:2000	T2	G1	C1	M2
Exp7	B	AA	K	S1	500	T1	G1	C1	M31
Exp8	B	AA	K	S1	500	T1	G1	C1	M32
Exp9	B	AA	K	S1	500	T1	G1	C1	M33
Exp10	B	AA	K	S1	100:100:400	T1	G1	C1	M34

Abbreviation:

t2m_max: maximum daily air temperature (Celsius);

SR: total daily solar energy (WJ/m²);

sur_temp_max: maximum infrareded surface temperature measurement (Celsius);

rh_max, rh_min, rh_avg: maximum, minimum and average relative humidity (%), respectively;

ST1, ST2, ST3, ST4, ST5: soil temperature at 5cm, 10cm, 20cm, 50cm and 100cm depth (Celsius), respectively;

SM1, SM2, SM3, SM4, SM5: soil moisture at 5cm, 10cm, 20cm, 50cm and 100cm depth (mm³/mm³), respectively.

Baseline:

B={t_max, SR, sur_temp_max, rh_max, rh_min, rh_avg, ST1, ST2, ST3, ST4, ST5};

H={t_max, SR, sur_temp_avg, rh_avg, ST1, ST2, ST3, ST4, ST5};

AA={SM1, SM2, SM3, SM4, SM5};

K={RBF-kernel}; S1={auto selected}; S2={1};

T1={crossvalidation}; T2={leaveoneout}; G1={auto selected}; G2={10};

C1={mse, error=10}; C2={mse, error=5};

M1={none}; M2={10};

M31={EnKF, M=1}; M32={EnKF, M=1, random selected with 50% ratio}; M33={EnKF, M=10:10:200, random selected with 50% ratio in each time interval}; M34={EnKF, M=50, random selected with 50% ratio};

E={MAE, RMSE, R }.

Notes:

Length of Training Set:

Exp1 to Exp4: 730 days (from September 1, 2009 to August 31, 2011);

Exp5: 50:50:2000 days (the training length from 50-day to 2000-day with an increase interval of 50-day started from September 1, 2009);

Exp6: 500:500:2000 days (the training length from 500-day to 2000-day with an increase interval of 500-day started from September 1, 2009);

Exp7 to Exp9: 500 days (from September 1, 2009 to January 13, 2011);

Exp10: 100:100:400 days (training length from 100-day to 400-day with an increase interval of 100-day started from September 1, 2009).

Model Update: (days)

Table 2: *MAE*, *RMSE* and *R* criteria results for Exp1 to Exp4.

		MAE					RMSE					R				
		SM1	SM2	SM3	SM4	SM5	SM1	SM2	SM3	SM4	SM5	SM1	SM2	SM3	SM4	SM5
Training	Exp1	0.021	0.017	0.011	0.019	0.026	0.026	0.022	0.014	0.026	0.036	0.873	0.846	0.818	0.780	0.769
	Exp2	0.020	0.017	0.011	0.019	0.028	0.026	0.022	0.015	0.027	0.038	0.880	0.848	0.814	0.766	0.741
	Exp3	0.023	0.018	0.011	0.019	0.028	0.028	0.023	0.015	0.027	0.038	0.849	0.835	0.812	0.764	0.733
	Exp4	0.022	0.019	0.012	0.020	0.029	0.028	0.025	0.015	0.028	0.039	0.854	0.800	0.787	0.731	0.713
Estimation	Exp1	0.034	0.040	0.026	0.037	0.037	0.044	0.048	0.033	0.045	0.046	0.540	0.443	0.413	0.389	0.367
	Exp2	0.035	0.040	0.026	0.036	0.037	0.045	0.049	0.033	0.044	0.046	0.535	0.440	0.416	0.392	0.375
	Exp3	0.033	0.040	0.026	0.037	0.037	0.043	0.049	0.033	0.045	0.045	0.561	0.454	0.419	0.347	0.383
	Exp4	0.046	0.042	0.027	0.038	0.042	0.059	0.050	0.034	0.047	0.050	0.531	0.499	0.362	0.243	0.291

Table 3: *MAE*, *RMSE* and *R* criteria results for Exp7 to Exp9.

		MAE					RMSE					R				
N=500	EnKF	SM1	SM2	SM3	SM4	SM5	SM1	SM2	SM3	SM4	SM5	SM1	SM2	SM3	SM4	SM5
SVM	N/A	0.034	0.039	0.025	0.038	0.044	0.044	0.047	0.032	0.047	0.057	0.527	0.469	0.330	0.308	0.221
Exp7	open loop	0.024	0.024	0.017	0.023	0.024	0.030	0.030	0.022	0.030	0.031	0.789	0.780	0.779	0.747	0.744
	1 st update	0.019	0.019	0.014	0.019	0.020	0.024	0.025	0.018	0.025	0.025	0.886	0.885	0.888	0.858	0.839
	2 nd update	0.010	0.010	0.007	0.010	0.010	0.013	0.013	0.009	0.013	0.014	0.977	0.977	0.979	0.971	0.964
Exp8	open loop	0.022	0.019	0.012	0.013	0.016	0.030	0.026	0.018	0.021	0.024	0.804	0.820	0.850	0.877	0.860
	1 st update	0.017	0.014	0.009	0.010	0.012	0.024	0.021	0.015	0.017	0.020	0.871	0.885	0.904	0.921	0.905
	2 nd update	0.014	0.012	0.008	0.009	0.010	0.022	0.019	0.013	0.016	0.018	0.892	0.905	0.920	0.934	0.918
Exp9	open loop	0.032	0.034	0.024	0.036	0.043	0.040	0.040	0.031	0.045	0.053	0.570	0.592	0.417	0.329	0.107
	1 st update	0.032	0.033	0.024	0.036	0.042	0.039	0.040	0.030	0.044	0.051	0.578	0.602	0.457	0.356	0.163
	2 nd update	0.017	0.021	0.020	0.028	0.027	0.023	0.026	0.025	0.035	0.033	0.899	0.873	0.735	0.700	0.698

Figures:

Fig.1: Flowchart of SVM modeling and EnKF technique updating schemes.

Fig.2: Time series of soil moisture estimations from Exp1 to Exp4 versus the observations at each soil layer from 5cm (SM1), 10cm (SM2), 20cm (SM3), 50cm (SM4) and 100cm (SM5). Exp1, Exp2, Exp3, Exp4 are shown in magenta line, blue line, red line and green line, respectively. Observed soil moisture is shown in black line. The blue line in bottom pattern is the precipitation (mm/day) during the same

period. X-axis is the time step in days with the training length from 1 (September 1, 2009) to 730 (August, 31, 2011) (left side) and the estimation from 731 (September 1, 2011) to 2099 (May 31, 2015) (right side) divided by the red dashed line.

Fig.3: Criteria of *MAE*, *RMSE* and *R* for estimated soil moisture at each soil layer (5cm (SM1 black star line), 10cm (SM2 blue line), 20cm (SM3 red circle line), 50cm (SM4 green diamond line), 100cm (SM5 magenta squared line)) from Exp5 with the initial training lengths from 50-day to 2050-day with an increase interval of 50-day.

Fig.4: Time series of estimated soil moisture from Exp5 with different initial training lengths versus the observed soil moisture (black line) at multi-layers: 5cm (SM1), 10cm (SM2), 20cm (SM3), 50cm (SM4) and 100cm (SM5). The initial training lengths are 500-day (magenta line), 1000-day (cyan line), 1500-day (red line) and 2000-day (green line). The blue line in the bottom pattern is the precipitation (mm/day) during the same period. X-axis is time step for estimation started from 1 (January 14, 2011) to 1599 (May 31, 2015).

Fig.5: Time series of soil moisture estimations at each updating step (open loop (cyan line), 1st update (red line), 2nd update (green line)) from Exp7 with EnKF technique versus the estimations from SVM without EnKF technique (magenta line) and the observed soil moisture (black line) at multi-layers: 5cm (SM1), 10cm (SM2), 20cm (SM3), 50cm (SM4) and 100cm (SM5). The blue line in the bottom pattern is the precipitation (mm/day) during the same period. X-axis is time step for estimation started from 1 (January 14, 2011) to 1599 (May 31, 2015).

Fig.6: Criteria of *MAE*, *RMSE* and *R* for soil moisture estimations at each soil layer (5cm (SM1), 10cm (SM2), 20cm (SM3), 50cm (SM4), 100cm (SM5)) at each updating step (open loop (blue + line), 1st update (red circle line) and 2nd update (green diamond line)) from Exp9 with EnKF technique updated at different time interval from 10-day to 200-day with an increase interval of 10-day.

Fig.7: Time series of soil moisture estimations from Exp9 with EnKF technique updated at different time interval (10-day (magenta line), 30-day (cyan line), 100-day (red line) and 200-day (green line)) versus the observed soil moisture (black line) at multi-layers: 5cm (SM1), 10cm (SM2), 20cm (SM3), 50cm (SM4) and 100cm (SM5). The blue line in the bottom pattern is the precipitation (mm/day) during the same period. X-axis is time step for estimation started from 1 (January 14, 2011) to 1599 (May 31, 2015).

Fig.8: Time series of soil moisture estimations in the 2nd updating process from Exp7 (cyan line), Exp8 (red line) and Exp9 (green line) with EnKF technique versus the SVM estimations without EnKF technique (magenta line) and observed soil moisture (black line) at multi-layers: 5cm (SM1), 10cm (SM2), 20cm (SM3), 50cm (SM4) and 100cm (SM5). The blue line in the bottom pattern is the precipitation (mm/day) during the same period. X-axis is time step for estimation started from 1 (January 14, 2011) to 1599 (May 31, 2015).

Fig.9: Criteria of *MAE*, *RMSE* and *R* for estimated soil moisture at each updating step with EnKF technique (open loop (blue +line), 1st update (red circle line), 2nd update (green diamond line)) from Exp10 with the initial training lengths from 100-day to 500-day with an increase interval of 100-day versus the estimations of SVM without EnKF technique (black star line) at each soil layer (5cm (SM1), 10cm (SM2), 20cm (SM3), 50cm (SM4), 100cm (SM5)).

Fig.10: Time series of soil moisture estimations in the 2nd updating process from Exp10 with EnKF technique at different initial training lengths (100-day (magenta line), 200-day (cyan line), 300-day (red line), 400-day (green line) and 500-day (blue line)) versus the observed soil moisture (black line) at multi-layers: 5cm (SM1), 10cm (SM2), 20cm (SM3), 50cm (SM4) and 100cm (SM5). The black dot line in the bottom pattern is the precipitation (mm/day) during the same period. X-axis is time step for estimation started from 1 (January 14, 2011) to 1599 (May 31, 2015).