Journal of Hydrology 538 (2016) 243–255

Contents lists available at ScienceDirect

Journal of Hydrology

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

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

ARTICLE INFO

Article history:
Received 13 January 2016
Received in revised form 6 April 2016
Accepted 8 April 2016
Available online 18 April 2016
This manuscript was handled by Geoff Syme, Editor-in-Chief

Keywords:
Support vector machines
Data assimilation
Dual ensemble Kalman filter
Soil moisture estimation
Uncertainty

SUMMARY

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

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

* Corresponding author.

http://dx.doi.org/10.1016/j.jhydrol.2016.04.021
0022-1694/

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 317 ft) 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 to 100 cm. 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 assur_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 5 cm, 10 cm, 20 cm, 50 cm and 100 cm depth (named as SM1, SM2, SM3, SM4 and SM5) and soil temperature at 5 cm, 10 cm, 20 cm, 50 cm and 100 cm 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, 2015; Mishra 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):

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

$$Y_t = H(X_t) + V_t \tag{2}$$

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, 2006; Cook 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.