Soil & Tillage Research 181 (2018) 63–81

Contents lists available at ScienceDirect

Soil & Tillage Research

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

Ensemble committee-based data intelligent approach for generating soil moisture forecasts with multivariate hydro-meteorological predictors

Ramendra Prasad*, Ravinesh C. Deo*, Yan Li, Tek Maraseni

School of Agricultural, Computational, and Environmental Sciences Institute of Agriculture and Environment, University of Southern Queensland Springfield, Australia

ARTICLE INFO

Keywords:
Committee of models
Extreme learning machine
Random forest
Volterra
M5 tree
Soil moisture forecasting
Murray-Darling Basin

ABSTRACT

Soil moisture (*SM*) is a key component of the global energy cycle that regulates all domains of the natural environmental and the agricultural system. In this research, the challenge is to develop a low-cost data-intelligent *SM* forecasting model using climate dynamics (i.e., the climate indices, atmospheric and hydro-meteorological parameters) as the model inputs. A newly designed, multi-model ensemble committee machine learning approach based on the artificial neural network (ANN-CoM) is developed to forecast monthly upper layer (~0.2 m from the surface) and the lower layer (~0.2–1.5 m deep) *SM* at four agricultural sites in Australia's Murray-Darling Basin. ANN-CoM model is validated with respect to non-tuned second-order Volterra, M5 model tree, random forest, and an extreme learning machine (ELM) models. To construct the ANN-CoM model, the input variables comprised of the hydro-meteorological data from the Australian Water Availability Project, large-scale climate indices and atmospheric parameters derived from the Interim ERA European Centre for Medium-Range Weather Forecasting ECMWF reanalysis fields leads to a total of 60 potential predictors used for *SM* forecasting. To reduce the model input data dimensionality for accurate forecasts, the Neighborhood Component Analysis (NCA) based feature selection algorithm for regression purposes (*fsrnca*) is applied to determine the relative feature weights related to the targeted variable. The optimal predictor variables are then screened with an ELM model as the fitness function of the *fsrnca* algorithm to identify the set of most pertinent model variables. Extensive performance evaluation using statistical score metrics with visual and diagnostic plots show that the ensemble committee based, ANN-CoM model is able to effectively capture the nonlinear dynamics involved in the modeling of monthly upper and lower layer *SM* levels. Therefore, the ANN-CoM multi-model ensemble-based approach can be considered to be a superior *SM* forecasting tool, portraying as an amicable, integrated (or ensemble) machine learning stratagem that can be explored for soil moisture modeling and applications in agriculture and other hydro-meteorological phenomena.

1. Introduction

Being a vital component of the response loop within a climatic system and gas exchange mechanism, the soil moisture (*SM*) plays an important role in hydrological and agricultural processes (Tian et al., 2017). *SM* controls the partitioning of energy into sensible and latent heat fluxes, and precipitation into evapotranspiration and runoff (Brocca et al., 2017; Munro et al., 1998; Petropoulos, 2014). *SM* is not only important for agricultural production but is also imperative for biomass production, biophysical and ecological processes, runoff potential, soil erosion/slope failure, flood control mechanisms, reservoir management and water quality assessments. The *SM* levels are greatly influenced by vegetation cover, soil characteristics, climate dynamics and land use. In addition, the rising global temperature trend shows significant reductions in projected *SM* level within the Australian Murray-Darling Basin region (Cai et al., 2009; Timbal et al., 2015) which is expected to severely affect the hydrological cycle, agriculture, and human lives. Thus, forecasted *SM* is critical for an assessment and development of sustainable agricultural and hydrological management practices (Tian et al., 2017).

Advancements in measurement and estimation techniques have led to a variety of ways to quantify *SM*, some of which include in-situ measurements, remote sensing, and formulation of physical models. However, the spatial in-situ measurements and now-casts of *SM* are expensive in terms of the installation, calibration, and maintenance issues of apparatus, time-consuming and labor-intensive. This has resulted in limited spatial and temporal monitoring of *SM* using ground-based point measurement techniques (Grayson and Western, 1998; Walker et al., 2003). To increase spatial coverage, remote sensing of *SM* via satellites has been developed. Yet, satellites are only able to

* Corresponding authors.
E-mail addresses: ramendrap23@gmail.com (R. Prasad), ravinesh.deo@usq.edu.au (R.C. Deo).

https://doi.org/10.1016/j.still.2018.03.021
Received 1 February 2018; Received in revised form 30 March 2018; Accepted 31 March 2018
0167-1987/

estimate the *SM* in the top few centimeters of soil (1–5 cm) in areas away from large water bodies (e.g., ocean or lake) with low vegetation (Dharssi and Steinle, 2011; Du et al., 2000; Walker et al., 2003). The vegetation water content, dew, radiative fog and soil roughness add uncertainties in the satellite-derived observations (Dharssi and Steinle, 2011). As a result, significant vertical gradients in the *SM* can be overlooked.

Accordingly, the *WaterDyn* physical model has been developed to simulate *SM* and several other hydrological parameters across the Australian continent at a grid resolution of $0.05^{\circ} \times 0.05^{\circ}$ (Raupach et al., 2009, 2012). Developed under the Australian Water Availability Project (AWAP), the *WaterDyn* physical model incorporates meteorological forcing, *i.e.*, solar radiation, precipitation, minimum and maximum temperatures coupled with continental parameter maps, e.g., albedo, soil characteristics, seasonality of vegetation greenness to compute *SM* for the upper layer (up to a depth of 0.2 m from the surface) and the lower layer (0.2–1.5 m depth). The meteorological fields for this model are generated by the Australian Bureau of Meteorology (BOM) from its network of rain gauge and weather stations while solar irradiance data is obtained using geostationary satellites (Raupach et al., 2009, 2012). However, the main constraint faced by this physical model is the high spatial and temporal variability of meteorological data, which may not be appropriate for small-scale applications, such as 'on-farm' decision making. Another drawback is that the *WaterDyn* physical model is accustomed to determine instantaneous ('now-casts') *SM* level at the point in time when meteorological inputs are channeled, requiring a constant supply of input variables. More precisely, this physical model is hindcasting since the system operates using already recorded meteorological data. For instance, monthly *SM* levels are attained after the observation and the accumulation of all the essential meteorological parameters are completed at the end of the month. Despite the advancements in measurement techniques, delayed progress in *SM* forecasting is evident. Particularly, the forecasted value of *SM* at the local scale (e.g., at the farm level) is imperative for key decision making but the current limitations in *SM* forecasting tools present a significant challenge in this respect.

To ameliorate *SM* predictability issues, the forecasting ability of advanced data-driven models offer feasible alternatives at the local scale modeling of *SM*. The predictive models are able to 'learn' from historical data making it advantageous for practical applications (Zhang et al., 1998). Data intelligent models have been successfully applied in agricultural and soil science applications to forecast field capacity and permanent wilting point (Ghorbani et al., 2017), soil water retention and saturated hydraulic conductivity (Merdun et al., 2006; Schaap and Leij, 1998) and soil temperature (Samadianfard et al., 2018). Yet, *SM* forecasting applications are still in their nascent stages (Liu et al., 2014; Matei et al., 2017; Myers et al., 2009; Yang et al., 2017). Researchers argue that forecasting of hydro-climatic variables must explore hybrid (rather than standalone) models building on the strengths of individual data-driven models (Jain and Kumar, 2007; Maier et al., 2010; Tiwari and Adamowski, 2013). Consequently, a new two-stage multi-model ensemble committee of models constructed on the basis of artificial neural networks (ANN) is explored in this study. The notion is to extract the pertinent information simulated by standalone expert models and further optimize it via an ANN for a collective forecast. This overcomes the weaknesses of conventional simple averaging forecast combinations whereby the overall model performance is compromised by the worst performing model(s). This novel multi-model ensemble committee of models approach has to overcome the inherent drawbacks of individual standalone models, building on the aptness, and subsequently surpassing the individual performances (Barzegar et al., 2017; Hatampour, 2013). The key advantage is that the committee based model combination reaps the benefit of all expert models yielding better generalization and performance, i.e., obtains a comparable or lower error than simple averaging and individual best single expert models (Barzegar and Moghaddam, 2016; Barzegar et al., 2015; Chen and Lin, 2006). Although, varying degree of a standalone ANN has been successfully applied in *SM* forecasting (Huang et al., 2010; Yang et al., 2017), model combination techniques have been overlooked in environmental applications (Baker and Ellison, 2008). Related committee modeling approaches were successfully applied in preparing groundwater vulnerability maps (Barzegar et al., 2017), groundwater contamination risk assessment (Barzegar et al., 2015) and groundwater salinity forecasting (Barzegar and Moghaddam, 2016). However, to the best of the authors' knowledge, *SM* forecasting is yet to be performed using the novel two-stage ensemble committee of models.

To collate the relevant features, four-standalone expert data-intelligent models viz., 2nd order Volterra, M5 tree, random forest (RF) and an extreme learning machine (ELM) model have been used. The 2nd order Volterra performed well in forecasting of streamflow (Maheswaran and Khosa, 2012, 2015; Rathinasamy et al., 2013), however, *SM* forecasting has not been piloted. Likewise, the M5 model tree has not been applied so far in *SM* forecasting, although a similar regression tree algorithm (Cubist) was noted (Myers et al., 2009). The *SM* forecasting from the bootstrapped-aggregated tree approach, RF, has yielded good performance with a reasonable prediction accuracy in one study in Romania (Matei et al., 2017). Literature shows that ELM is also uncommon in *SM* forecasting as only one study by Liu et al. (2014) in Victoria, Australia applied ELM and support vector machine for a short period (14 months) ignoring the long-term dynamics. Hence, overall, the application of data-driven models in the area of *SM* forecasting has not been fully exploited.

The objective of this study is to develop a low cost (saving labor, time, energy, and money) *SM* forecasting model using climate dynamics, i.e., the climate mode indices, atmospheric and hydro-meteorological drivers as the model inputs. The other factors such as vegetation cover, soil characteristics, i.e., soil texture, soil structure, initial *SM*, hydraulic conductivity, and *SM* pressure and land-use are assumed to be site-specific and constant in this study. Thus, historical hydro-meteorological variables from AWAP, climate indices, and the Interim ERA European Centre for Medium-Range Weather Forecasting reanalysis derived atmospheric data are collated leading to sixty inputs. Consequently, salient inputs are screened using a two-stage feature selection technique via Neighborhood Component Analysis based feature weights and modeled minimum relative error criteria. A novel well-trained two-stage hybrid multi-model ensemble committee based on ANN (ANN-CoM) data intelligent model is developed in forecasting upper and lower layer *SM* within the Murray-Darling Basin region, Australia. The performance of the new ANN-CoM model is evaluated with various statistical measures together and the diagnostic plots and this is benchmarked against the four primary standalone models.

2. Materials and methodology

2.1. Machine learning algorithms used in developing ensemble committee model

2.1.1. 2nd order volterra model

The Volterra model is built upon the Taylor series expansion for nonlinear autonomous causal systems with memory. A second-order representation has been adopted, as substantiated by previous studies (Labat et al., 1999; Maheswaran and Khosa, 2012, 2015; Rathinasamy et al., 2013). With $z(t)$ as the model output and t as the *tth* instances, the 2nd-order Volterra expansion could be expressed as:

$$z(t) = \int_{\tau_1=0}^{\tau_1=t} k_1(\tau_1) X(\tau - \tau_1) d\tau_1 + \int_{\tau_2=0}^{\tau_2=t} \int_{\tau_1=0}^{\tau_1=t} k_2(\tau_1, \tau_2) X(\tau - \tau_1) X(\tau - \tau_2) d\tau_1 d\tau_2 \tag{1}$$

where $k_1(\tau_1)$ and $k_2(\tau_1, \tau_2)$ are the Volterra kernels. In a condensed