Advances in Water Resources 98 (2016) 122–131

Contents lists available at ScienceDirect

Advances in Water Resources

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

# Using machine learning to produce near surface soil moisture estimates from deeper *in situ* records at U.S. Climate Reference Network (USCRN) locations: Analysis and applications to AMSR-E satellite validation

Evan J. Coopersmith[a,*], Michael H. Cosh[a], Jesse E. Bell[b,c], Ryan Boyles[d]

[a] USDA-ARS-Hydrology and Remote Sensing Laboratory, Beltsville, MD, USA
[b] Cooperative Institute for Climate and Satellites - NC, Asheville, NC, USA
[c] NOAA-National Centers for Environmental Information, Asheville, NC, USA
[d] Southeast Climate Science Center, N.C.State University. Raleigh, NC, USA

ARTICLE INFO

*Article history:*
Received 5 April 2016
Revised 3 September 2016
Accepted 7 October 2016
Available online 18 October 2016

*Keywords:*
Soil moisture
Remote sensing
Modeling
USCRN
AMSR-E
ECONet

ABSTRACT

Surface soil moisture is a critical parameter for understanding the energy flux at the land atmosphere boundary. Weather modeling, climate prediction, and remote sensing validation are some of the applications for surface soil moisture information. The most common *in situ* measurement for these purposes are sensors that are installed at depths of approximately 5 cm. There are however, sensor technologies and network designs that do not provide an estimate at this depth. If soil moisture estimates at deeper depths could be extrapolated to the near surface, *in situ* networks providing estimates at other depths would see their values enhanced. Soil moisture sensors from the U.S. Climate Reference Network (USCRN) were used to generate models of 5 cm soil moisture, with 10 cm soil moisture measurements and antecedent precipitation as inputs, via machine learning techniques. Validation was conducted with the available, *in situ*, 5 cm resources. It was shown that a 5 cm estimate, which was extrapolated from a 10 cm sensor and antecedent local precipitation, produced a root-mean-squared-error (RMSE) of 0.0215 $m^3/m^3$. Next, these machine-learning-generated 5 cm estimates were also compared to AMSR-E estimates at these locations. These results were then compared with the performance of the actual *in situ* readings against the AMSR-E data. The machine learning estimates at 5 cm produced an RMSE of approximately 0.03 $m^3/m^3$ when an optimized gain and offset were applied. This is necessary considering the performance of AMSR-E in locations characterized by high vegetation water contents, which are present across North Carolina. Lastly, the application of this extrapolation technique is applied to the ECONet in North Carolina, which provides a 10 cm depth measurement as its shallowest soil moisture estimate. A raw RMSE of 0.028 $m^3/m^3$ was achieved, and with a linear gain and offset applied at each ECONet site, an RMSE of 0.013 $m^3/m^3$ was possible.



## 1. Introduction

Soil moisture estimates can be obtained by *in situ* networks and by satellite remote sensing. These estimates have been employed at a variety of spatial and temporal scales for applications ranging from droughts (Sheffield et al., 2004) to trafficability for agricultural decision support (Coopersmith et al., 2014a). In hydrologic models, subsurface flows at the watershed scale play crucial roles in modeling. These flows require soil moisture time series data (*e.g.* Grayson et al., 1997; Bell et al., 2010). Many of these hydrologic models are, in turn, utilized by General Circulation Models (GCMs) as mechanisms with which to address uncertainty–these require inputs of soil moisture (*e.g.* Koster and Milly, 1997; Belair et al., 2005; De Rosnay et al., 2013; Campoy et al., 2013; Joetzjer et al., 2013). These models have also been applied to Department of Defense Applications (Jones et al., 2010).

These important soil moisture inputs, when retrieved from satellites, typically approximate soil moisture at the 5 cm depth. These estimates can, in turn, be compared and validated against *in situ* networks where sensors are commonly available at that depth. Jackson et al. (2010) compared AMSR-E soil moisture products to a group of USDA Agricultural Research Service Experimental

* Corresponding author.
*E-mail address:* ecooper2@gmail.com (E.J. Coopersmith).

http://dx.doi.org/10.1016/j.advwatres.2016.10.007
0309-1708/

watersheds, demonstrating soil moisture could be estimated with an accuracy of approximately 0.05 $m^3/m^3$ RMSE. The Soil Moisture Ocean Salinity Mission (SMOS) was validated with these same watersheds and an accuracy of 0.04 $m^3/m^3$ was achieved (Jackson et al., 2012). AMSR-E data were compared against U.S. Climate Reference Network (USCRN in Coopersmith et al. (2015a). However, validation programs are currently limited to networks that produce 5 cm (or shallower) soil moisture estimates. For instance, the ECONet in North Carolina provides soil moisture estimates at the 10 cm depth as the shallowest measurement. This analysis proposes a methodology to produce 5 cm estimates from 10 cm *in situ* measurements and antecedent local precipitation, enabling the generation of 5 cm *in situ* products at locations where only 10 cm sensors are available. This would be the case for certain networks as well as a replacement method for sensor failures as shallow depths.

The methodology will be addressed in three parts. Firstly, a machine learning model is developed at local USCRN locations containing *in situ* measurements at both the 5 cm and 10 cm depths. Models calibrated at 10 cm, in concert with a simple implementation of the K-nearest-neighbor algorithm (Fix and Hodges, 1951), vertically extrapolate 10 cm estimates to produce 5 cm estimates. Validation is conducted using the *in situ*, 5 cm time series at the USCRN locations.

Secondly, the extrapolated 5 cm estimates will be compared to AMSR-E estimates at those locations. By comparing *in situ* observations to AMSR-E and model estimates to AMSR-E side-by-side, the amount of error introduced by the extrapolation model can be quantified. This will be critical to future work related to deeper networks, as knowing error estimates are necessary in any performance metric for a satellite calibration/validation program.

Lastly, the extrapolation method will be applied to a ‘deep’ network to compare with a satellite product. North Carolina's Environment and Climate Observing Network (NC ECONet) will be used as a test case. ECONet *in situ* soil moisture sensors are an ideal choice for such a test, as they measure soil moisture at 10 cm depths across a range of soil types and in a humid environment with generally dense vegetation. The extrapolated estimates will be compared to the AMSR-E satellite values for appropriate locations. These three steps will demonstrate the viability of extrapolating soil moisture estimates at the 10 cm depth via machine learning, and deploying these results in lieu of *in situ* measurements at 5 cm where such measurements are unavailable.

## 2. Methodology

### 2.1. The AMSR-E satellite

The Advanced Microwave Scanning Radiometer – Earth Observing System (AMSR-E) satellite, is a passive-microwave radiometer system measuring polarized brightness temperatures from which soil moisture estimates are inferred at approximately 60 km resolution. These C-band observations represent estimates at the top ~1 cm of depth and are delivered via the NASA algorithm, available from www.nsidc.org. AMSR-E delivered such estimates between 2002 and 2011. It is important to note that although USCRN's installation sites were selected to minimize the impacts of anthropogenic interference, spatial heterogeneity of land-cover, soil texture, and topography, may impact the representativeness of the USCRN point estimates to the AMSR-E spatial estimate to which it is compared in Coopersmith et al. (2015c) and as an application later in this manuscript, when ECONet site estimates are compared with AMSR-E values.

Though AMSR-E produces estimates at the top 1 cm depth, to retain robustness of installations, national *in situ* networks for soil moisture (*e.g.* USCRN, SCAN) place their shallowest sensors at the 5 cm depth. Thus, though a model could be developed for a 1cm-depth estimate, there would be no *in situ* record against which to validate it. The seminal work for AMSR-E (Jackson et al., 2010) addresses this in greater detail. As the focus of the current paper is 5 cm *in situ* estimation rather than AMSR-E, the chosen depth for analysis will be 5 cm.

### 2.2. The diagnostic soil moisture equation

At each of these 1075 profiles, the six parameters of the diagnostic soil moisture equation have been calibrated via genetic algorithm. The diagnostic soil moisture equation is presented below in Eq. 1.

$$\theta_{est} = \theta_{re} + (\phi_e - \theta_{re})\left(1 - e^{-c_4\beta}\right) \tag{1}$$

In Eq. 1, in calculating our soil moisture estimate, $\theta_{est}$, we utilized three parameters. The first, $\theta_{re}$, denotes the residual soil moisture – a value below which soil moisture levels will not fall even after a prolonged absence of rainfall. Though similar to “wilting point,” residual soil moisture represents a hard minimum value for the model, often calibrated as the lowest value a given sensor can record. The second, $\varphi_e$, signifies the porosity of the soil, the maximum soil moisture at saturation. The third, $c_4$, describes the soil's rate of drainage. Note that a high value implies soil that drains extremely slowly, and a low value implies a soil that drains rapidly. Finally, the $\beta$ value describes the convolution of an antecedent precipitation time series – it is presented in Eq. 2.

$$\beta = \sum_{i=2}^{i=n-1}\left[\frac{P_i}{\eta_i}\left(1 - e^{-\frac{\eta_i}{z}}\right)e^{-\sum_{j=1}^{j=i-1}\left(\frac{\eta_j}{z}\right)}\right] + \frac{P_1}{\eta_1}\left(1 - e^{-\frac{\eta_1}{z}}\right) \tag{2}$$

Eq. 2 presents the calculation of the $\beta$-series at depth $z$, using precipitation values at previous time stamps (the values of $P_i$). The values of $\eta_i$ represent the losses due to evapotranspiration and deep drainage – this $\eta$-series is assumed to be sinusoidal. The remaining three parameters define that sinusoid – amplitude, horizontal shift, and vertical shift (the period is fixed at one year). Additional descriptions of the model can be found in Pan et al. (2003) and Pan (2012). Further description of its calibration are found) in Coopersmith et al. (2014b).

### 2.3. Model approach: the K-Nearest-Neighbor algorithm

As this study assumes that there are no 5 cm data available, the diagnostic soil moisture equation cannot be used directly. However, the models that are calibrated at the 10 cm depth still provide some insights via the use of the K-Nearest-Neighbors algorithm (KNN). KNN is a simple machine learning algorithm that has been deployed in a number of hydrological contexts (*e.g.*, Kumar et al., 2006, Meliker et al., 2008, McRoberts et al., 2007; Nemes et al., 2008, and Coopersmith et al., 2011). KNN is an intuitively satisfying approach for classification, analysis, and forecasting. The algorithm employs current features to locate the most similar examples from historical data (where 5 cm values are actually known) and, in turn, utilizes those similar examples to estimate the 5 cm soil moisture values where those values are unavailable.

Similarity is determined via a simple Euclidian distance function in attribute space. Consider a hypothetical input vector as shown in Eq. 3:

$$X = (x_1, x_2, x_3, \ldots, x_n) \tag{3}$$

As each independent variable's distribution may occur on widely different numerical scales, normalization occurs by transforming each non-normalized variable, $z_i$ as shown in Eq. 4.

$$x_i = \frac{z_i - \mu_i}{\sigma_i} \tag{4}$$