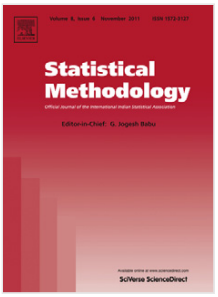
Accepted Manuscript



Process modeling for soil moisture using sensor network data

Souparno Ghosh , David M. Bell, James S. Clark, Alan E. Gelfand, Paul Flikkema

PII: S1572-3127(13)00063-4
DOI: http://dx.doi.org/10.1016/j.stamet.2013.08.002
Reference: STAMET 424

To appear in: *Statistical Methodology*

Received date: 6 June 2012
Revised date: 8 July 2013
Accepted date: 1 August 2013

Please cite this article as: S. Ghosh, D.M. Bell, J.S. Clark, A.E. Gelfand, P. Flikkema, Process modeling for soil moisture using sensor network data, *Statistical Methodology* (2013), http://dx.doi.org/10.1016/j.stamet.2013.08.002

Process Modeling for Soil Moisture using Sensor Network Data

Souparno Ghosh*, David M. Bell, James S. Clark, Alan E. Gelfand, and Paul Flikkema †

Abstract

The quantity of water contained in soil is referred to as the soil moisture. Soil moisture plays an important role in agriculture, percolation, and soil chemistry. Precipitation, temperature, atmospheric demand and topography are the primary processes that control soil moisture. Estimates of landscape variation in soil moisture are limited due to the complexity required to link high spatial variation in measurements with the aforesaid processes that vary in space and time. In this paper we develop an inferential framework that takes the form of data fusion using high temporal resolution environmental data from wireless networks along with sparse reflectometer data as inputs and yields inference on moisture variation as precipitation and temperature vary over time and drainage and canopy coverage vary in space. We specifically address soil moisture modeling in the context of wireless sensor networks.

Key words: data fusion; Euler discretization; hierarchical nonlinear model; partial differential equation; state space model

*Corresponding author: Tel: +1 806 742 2566

†S. Ghosh (*souparno.ghosh@ttu.edu*) is an assistant professor in the Department of Mathematics and Statistics, Texas Tech University, Lubbock, TX 79409, USA. A. E. Gelfand (*alan@stat.duke.edu*) is a professor in the Department of Statistical Science, Duke University, Durham, NC 27708, USA. D. M. Bell(*dbell9@uwyo.edu*) is a postdoctoral associate in the Department of Botany, University of Wyoming, Laramie, WY 82071, USA. J. S. Clark (*jimclark@duke.edu*) is a professor in the Nicholas School of Environment, Duke University, Durham, NC 27708, USA. P. Flikkema (*paul.flikkema@nau.edu*)is a professor in the department of Electrical Engineering, Northern Arizona University, Flagstaff, AZ 86011 , USA.

1 Introduction

The quantity of water contained in soil is referred to as the soil moisture. Soil moisture plays an important role in agriculture, percolation, and soil chemistry. Competition for soil moisture determines the structure and diversity of ecosystems, including variation across landscapes and over time (Korstian and Coile [18], Barton [2], Coomes and Grubb [6]). Precipitation inputs and atmospheric demand contribute unevenly to soil moisture additions and losses (Sturm et al. [25]). Topography controls spatial redistribution through drainage, and moisture is returned to the atmosphere through transpiration. Taken together these processes are responsible for diverse gradients in vegetation structure and composition across landscapes. Although these basic relationships have long been known, estimates of landscape variation in soil moisture are limited due to the complexity required to link high spatial variation in measurements with these processes that vary in space and time (Katul et al. [16]).

The contribution of this paper is to develop an inferential framework for soil moisture variation that takes, as its primary data source, environmental measurements from a wireless network as precipitation and temperature vary over time and drainage and canopy coverage vary in space. We fuse this with data from a portable data collection device. Jointly, we develop a dynamic nonlinear state space model driven by a latent specification for the foregoing processes. We employ a calibration model for the network data and a measurement error model for the data from the portable device. Our flexible stochastic framework merges often-used descriptive summaries of soil moisture with the more theoretical approach based upon partial differential equations. The result is the opportunity to carry out full inference with associated uncertainty for the soil moisture process. This mechanistically motivated statistical framewrok is similar in spirit to the methodolgies developed in Berliner [3], Wikle [27], Fuentes and Raftery [12]. A substantial challenge is that the soil moisture levels over time arise as a result of three primary processes: precipitation, drainage, and transpiration. We rarely observe the first at ground level in the forest and we never observe the second or third, rendering it difficult to allocate measured levels to these activities. A further challenge in working with such models is the high level of noise and the lack of calibration in the sensor network data. A useful assist is provided by the strong predictability of soil moisture behavior in the absence of a precipitation event.

The processes that control soil moisture change have long been known. Precipitation adds soil moisture up to a saturation level, at which point pore spaces are filled. Rapid drainage of moisture in excess of capillary forces occurs above *field capacity* (a one dimensional conceptual notion given as a soil moisture level). Subsurface inter-flow follows topography and directs moisture from high to low elevation (Western et al. [26], Qiu et al. [23], Penna et al. [22]). Below field capacity, moisture is withdrawn at a slower rate, depending on exposure to radiation, by evaporation and through the transpirational stream (Korstian and Coile [18], Kleb and Wilson [17]). Transpiration provides a conduit from soil to atmosphere at loss rates that depend on availability of moisture in the soil and atmospheric demand. As soil moisture approaches the *wilting point* (again, a conceptual soil moisture level), plants can no longer extract it. The rate of depletion varies with depth, with shallow layers reaching wilting point first and transpiration depending increasingly on deeper moisture stores (Sturm et al. [25]). Figure 1 visually describes the operation of transpiration and drainage with respect to wilting point and field capacity.

Elements for a minimal process model include several variables that must be measured (soil moisture, precipitation, temperature) and parameters for the relationships between them. These parameters arise from the regression specifications but also include the unknown field capacity and wilting point. Despite the general awareness of the important relationships, we are aware of no stochastic models that coherently connect them with the uncalibrated and highly variable soil moisture data.

To reiterate the inference challenge, these processes are only crudely described by simple models and depend on many sources of variation that cannot be measured. Measurements of soil moisture can vary at spatial scales as small as meters (Entin et al. [9], Zhou et al. [28]). Drainage rates depend on topographic variation, but water movement through soil depends on heterogeneity at scales that cannot be fully quantified. Transpiration varies spatially due to soil and vegetation heterogeneity. Due to the high cost of long-term monitoring, soil moisture measurements typically come from unevenly spaced probes under which spatial or temporal extrapolation is done empirically (Junior et al. [14], Zhou et al. [28]), i.e., without benefit of data on the processes that control it - temperature, precipitation, topography, and vegetation cover.

Furthermore, observation of soil moisture is, itself, a difficult task due to technology and high spatio-temporal variability. Conventional soil moisture

data collection is often done using a portable Time Domain Reflectometer (TDR) but at few locations and at low temporal resolution. Wireless sensing networks are employed in this study to provide data collection at high temporal resolution. However, they introduce new problems in terms of, for example, disruption due to animal movement, lack of calibration, and sensor failure (battery failure, transmission failure, suppression of transmission) (Clark et al. [5]). Here, we propose a fusion in order to work with both sources of collection, accommodating the misalignment in time scales for the collection.

In Section 2 we briefly describe the two main current approaches in the literature for modeling soil moisture and how our methodology differs from but, in fact, merges these approaches. Section 3 provides a description of the sensor networks deployed to monitor soil moisture. The posited model and the associated computational issues are discussed in Section 4. We apply our model to daily average soil moisture in Duke Forest observed in the year 2009. The performance of our model and associated inference is described in Section 5. The final section summarizes the findings of this study, notes some caveats, and provides direction to future research.

2 Background and modeling motivation

Most soil moisture studies provide descriptive statistics with respect to environmental controls, such as topography, precipitation, vegetation (Charpentier and Groffman [4], Famiglietti et al. [11], Crow and Wood [7]). An alternative path focuses more on mathematical formulations for dynamic processes. For example, Oldak et al. [21], and references therein, promote spectral scaling theory for statistical estimation of the variability in soil moisture at spatial scales. Katul et al. [15] develop a conservation equation model for soil moisture variability using a partial differential equation (PDE) approach. Rodriguez-Iturbe et al. [24] develop a model based on a stochastic differential equation for temporal dynamics of soil moisture. Albertson and Montaldo [1] also use a PDE based conservation equation model for soil moisture, but in addition obtain covariances between moisture fields and land surface flux fields and thereby develop a full fledged predictive framework for variation in soil moisture. Studies such as Junior et al. [14] and Zhou et al. [28] use kriging algorithms to make spatial and/or temporal prediction of soil moisture but ignore processes like temperature, precipitation and topography that

control it.

To merge the descriptive statistical approaches with the process-based approach, we propose a hierarchical nonlinear state-space model that captures the temporal dynamics of soil moisture in response to environmental controls. We assume that the true soil moisture content at a given point in time is an unknown state variable. The large number of processes contributing to the soil moisture dynamics and the uncertainty associated with them motivate us to assume a state space model for the evolution of the true soil moisture. This evolution is governed by a stochastic differential equation. The motivation for the stochastic differential equation is based on the rudimentary notion that the change in soil moisture reflects what goes in and what comes out. What comes in is precipitation. What goes out is captured in two terms, drainage above field capacity and and transpiration above the wilting point. In generic terms, we would have the stochastic differential equation,

$$dm(t) = (\text{Prec}(m(t);t) - \text{Drain}(m(t);t) - \text{Trans}(m(t);t))dt + \sigma dB(t). \quad (1)$$

Here $m(t)$ is the true soil moisture at time t, with Prec, Drain, and Trans being the precipitation, drainage and transpiration components, respectively, and $B(t)$ being Brownian motion with variance 1. Conceptually, we view soil moisture at a location as belonging to $[0, 1]$ since it is typically thought of as a proportion, i.e., volume of water per total volume. (In fact, we cannot do better since we never actually observe an *absolute* soil moisture.) Hence, we scale the foregoing terms to this interval.

We make the model explicit through specification of the component terms which attempt to be process-driven. We also specify the model at sampling site level. At site i, we take

$$\text{Prec}_i(t) = g(P_{i,t}; h_i), \quad (2)$$

where g is an increasing function in P on $[0, 1]$, $P_{i,t}$ is the precipitation at site i at time t, $h_i \in \mathcal{R}_+$ scales the precipitation according to its units and adjusts locally for canopy interception at site i. In fact, we set $\text{Prec}_i(t) = (1-e^{-h_i P_{i,t}})$.

We take

$$\text{Drain}_i(t) = \alpha_i (m_i(t) - fc_i)_+^{\theta_{fc}}. \quad (3)$$

This form implies hard thresholding for drainage, i.e., that it only occurs when m is above field capacity, fc. Further, the form is allometric with a local

scale ($\alpha_i \in [0,1]$) and field capacity but a common power θ_{fc} which controls the rate of drainage. As noted above, since we never measure drainage, it will be difficult to learn about a more complex specification for it.

Lastly, we take

$$\text{Trans}_i(t) = f(z_{i,t})(m_i(t) - \omega p_i)_+^{\theta_{\omega p}}. \tag{4}$$

This form also introduces hard thresholding, implying that transpiration only occurs when m is above the wilting point, ωp. Once again $\theta_{\omega p}$ controls the amount of transpiration above the wilting point. It provides an allometric form with a local wilting point and introduces f, a monotone function on $[0,1]$, which we take to be the inverse logit and enables scaling. We do the scaling locally, linking to local temperature.Here, $f(T_{i,t})$ is a local regression in temperature at site i at time t, $T_{i,t}$. Altogether, we set $f(T_{i,t}) = (\exp(\beta_{0,i} + \beta_{1,i}T_{i,t}))/(1 + \exp(\beta_{0,i} + \beta_{1,i}T_{i,t}))$. Again, since we never measure transpiration, it will be difficult to criticize this specification for it. The line diagram in Figure 1 clarifies the behavior of drainage and transpiration relative to ωp and fc.

The introduction of the f term in (4) is novel in the soil moisture literature. Most of the studies (Albertson and Montaldo [1], Rodriguez-Itrube et al. [24]) use linear terms to describe covariate effects. The two fold intuition behind the inclusion of the inverse logit in our model is: (i) the soil moisture readings are scaled to lie in the interval [0,1] so that the logit term scales the right hand side of the (1) to match the scale of the observed soil moisture, (ii) it allows us to preserve monotonicity in multiple dimensions with potential interactions.

In the sequel, we set $\theta_{\omega p} = 1$ and $\theta_{fc} = 3$. Fixing $\theta_{\omega p} = 1$ arises due to identifiability problems between this power and the regression for f; fixing $\theta_{fc} = 3$ is motivated by exploratory data analysis discussed in Section 4.2 below.

Inserting these forms into (1) implies the assumption that $m_i(t)$ exhibits complex drift but constant volatility. The dependence on elevation and canopy gap status can be introduced in an empirical fashion, to allow for the fact that drainage depends on topography, and transpiration rates are affected by the canopy. Here, we simply handle this by labeling and modeling the sites individually. We are ignoring depth in our model; this is usually justified because, typically, rocky substrate prohibits sampling at depth.

We apply a first order Euler discretization and obtain the working model

for the true soil moisture as

$$m_{i,t+\Delta_t} = m_{i,t} + \{(1 - e^{-h_i P_{i,t}}) - f(z_{i,t})(m_{i,t} - \omega p_i)_+^{\theta_{\omega p}} - \alpha_i (m_{i,t} - fc_i)_+^{\theta_{fc}} + \epsilon_{i,t}\}\Delta_t \quad (5)$$

where the $\epsilon_{i,t}$ are pure Gaussian errors with mean 0 and variance $\sigma^2_{\epsilon,i}$. With our data, Δ_t can be as fine as 2 hours. In fact, with data augmentation as in (Elerian et al. [8]; Eraker [10]) we can work at even finer resolution. However, our differential equation is primarily suggestive in order to motivate our discretized model. The differential equation does not reflect diurnal cycles which are revealed in the data at finer than daily resolution. Hence, we work at daily scale, acknowledging that capturing diurnal behavior is beyond the scope here. Note that in (1) and therefore in (5), we are modeling conditional on precipitation. Models for $P(t)$ are discussed in the literature; a common choice is a marked Poisson process for occurrence and a Gamma distribution for the amount of precipitation (Laio et al. [19], [20]). However, we are not interested in modeling precipitation here. Rather, we treat it as a covariate, fixing it at observed values. Besides precipitation, other prominent factors that add uncertainty in the soil moisture dynamics are air temperature, light availability (which affects the evapotranspiration) and soil characteristics (which affect interception and drainage). Were data available on the latter variables we could attempt to enrich our specification.

As noted above, we have two observational sources to provide the soil moisture measurements to inform about our assumed latent *true* process model in (1). One is the data provided by the sensor network, which requires calibration. The other is available through the TDR and is assumed to be accurate up measurement error.

3 Data description

This study used a subset of the data recorded as part of a wireless sensing and relay device network (WiSARDnet), first deployed in the Duke Forest in 2005. Two of the WiSARDnet deployments are located in a mature deciduous forest stand located in the Eno Division of the Duke Forest, Orange County, NC ($35^0$52' N, $80^0$00' W). The study area is characterized by rolling topography, with an elevation of approximately 130 m above sea level. Individual WiSARDs were located along a topographic gradient from a wet riparian area occurring along water courses to a dry rocky hilltop. Elevation

difference between the riparian area and hilltop was more than 40 m. In addition, experimental forest canopy treefall-gaps were created in January, 2009, by pulling down all trees within 15m of four of the WiSARD nodes. The forest is composed of eastern deciduous hardwood tree species with a variety of overstory tree species and an understory that is relatively sparse.

Measurements from the WiSARDnet deployment in the Duke Forest during the period March 18, 2009 through December 8, 2009 include volumetric soil moisture in the upper 10 cm of mineral soil (mm^3/mm^3), precipitation (mm), and ambient air temperature (^{0}C). These are the sensor data soil moisture measurements; they are taken every two hours at each site by each of *two* EC-20 soil moisture probes (Decagon Devices, Inc., Pullman, WA, USA). There are 16 sites altogether. Transmission issues arise - measurements sent but not received at the gateway and measurements which arrive corrupted so, as a result, across these sites there were approximate 4% missing or unobserved data. We have retained 14 of them, deleting the site with no TDR data. Four of the sites (WiSARD id: 133, 135, 141, 152) are *gap* sites; the remainder (WiSARD id: 146, 148, 151, 156, 157, 181, 301, 302, 303, 509) are understory. Spot measurements at 2-4 week intervals using the portable time-domain reflectometer (TDR; HydroSense, Campbell Scientific, Logan, UT, USA) also record volumetric soil moisture in the upper 10 cm. Precipitation was measured using a Vaisala WXT510 weather station (Vaisala, Vantaa, Finland) located atop a 30-m tower, recorded by a WiSARD, and transmitted to the base station. Air temperature was measured at a height of one meter at every WiSARD, using thermocouples, and the *average* daily air temperature was used in the model.

4 The hierarchical state space model

4.1 Full model specification

In this section we provide a detailed description of a state-space fusion model for soil moisture which is driven by (5). It captures assimilation between sensor network data and TDR data with the former providing the predominance of the data. The locations are widely separated from each other relative to the scale of soil variability. Preliminary exploration of association between sites suggested that there was no reason to introduce spatial dependence.

Let $w_{ij,t}$ be the soil moisture recorded by the jth probe (j =1, 2) at

location $i,\ i = 1, 2, ..., n$ at time $t,\ t = 1, 2, ..., T$. As noted below (1), since soil moisture is typically thought of as a proportion, we scale $w_{ij,t}$ at the outset so that the measurements lie between 0 and 1. (In fact, the probes record a proxy which is converted to a soil moisture value and then scaled.)

We assume that the wireless data at site i is related to true soil moisture at i via the model

$$w_{ij,t} = a_{0,ij} + a_{1,ij} m_{i,t} + \epsilon_{w,ij,t} \tag{6}$$

with $\epsilon_{w,ij,t} \sim N(0, \sigma^2_{w,i})$. That is, we have a calibration model that allows for additive and multiplicative bias.

Next, let $v_{il,t},\ l = 1, 2, ..., L$ denote the lth replicate of the TDR measurement of soil moisture taken at location i at time t. Once again we scale $v_{il,t}$ such that it lies between 0 and 1. We relate these TDR data to the true soil moisture at i via the measurement error model given by

$$v_{il,t} = m_{i,t} + \epsilon_{v,il,t} \tag{7}$$

where the error term $\epsilon_{v,il,t} \sim N(0, \sigma^2_{v,i})$. We note that, since the TDR data is measured at locations that can be several meters from where the probes are and $m_{i,t}$ is assumed to be the true value at the site of the probes, the measurement error in the TDR data is expected to be larger than if the latent m was at the location of the TDR measurement. (The reader may suggest that therefore a calibration model for the TDR data might be needed as well. However, two calibration models can not be identified so, we assume that the discrepancy is entirely pure error.) A related concern is that, according to the uncertainty in ϵ_w relative to that in ϵ_v, we can have the two data sources exert relatively more or less influence on the trajectory of the predicted m's. We can experiment with sensitivity to priors on the variances of these two errors but, in the end, faith in one source relative to the other will not be a statistical decision. Finally, (6) and (7) provide the observational or first stage for our state space model while, again, (5) provides the transitional or second stage of the model.

4.2 Exploratory data analysis

We describe brief exploratory data analysis with regard to choice of θ_{fc}. Again, let $w_{ij,t}$ be the soil moisture recorded by the jth probe of sensor at location i at time t so that $w_{i,t} = (w_{i1,t} + w_{i2,t})/2$ is the average soil moisture recorded by the sensor at location i at time t. Then, viewing (5) in terms of

the w's rather than m's, we see that the difference ($\Delta w_{i,t} = w_{i,t+\Delta t} - w_{i,t}$) has a log-linear relationship with $P_{i,t}$ and linearly related to $w_{i,t}^{\theta_{fc}}$. So, we first regress the $1/\log(1 - \Delta w_{i,t})$ obtained at each location on the $P_{i,t}$. Then we regress the residuals on $w_{i,t}^{\theta_{fc}}$ for various values of θ_{fc}. The value of θ_{fc} that yields the maximum likelihood (appropriate since all the competing models have same number of parameters) can be taken as an initial estimate of θ_{fc}. Figure 2 shows the scatter plot of the residuals (obtained from regressing $\Delta w_{i,t}$ on $P_{i,t}$) against the $w_{i,t}$ for four of the sites (WiSARD id 133, 152 and 181 and 302). Despite a large amount of noise, a non-linear relationship is suggested indicating, θ_{fc} should not be taken to be 1.

Minimum BIC is obtained for $\theta_{fc} = 3$. The fitted curves obtained by regressing the foregoing residuals on $w_{i,t}^3$, for the same four sites mentioned above, are overlaid on Figure 2.Thus, the EDA indicates that for $\theta_{fc} = 3$ will provide a satisfactory choice for the proposed state model (5). However, below we fit the model for $\theta_{fc} = 1, 2, 4, 5$ as well and see how each of these models performs on a test data set, choosing the one with best out-of-sample predictive performance.

4.3 Priors

To complete the hierarchical structure we specify the following priors for the parameters. All are quite vague.

$$\begin{aligned}
(\alpha_i) &\sim N(0,1)\mathcal{I}(0 < \alpha < 1),\ \forall i \\
h_i &\sim N(0,1)\mathcal{I}(h_i > 0),\ \forall i \\
(\beta_{0,i}, \beta_{1,i}) &\sim N(0, 100 \times I_2) \forall i \\
(a_{0,ij}, a_{1,ij}) &\sim N(0, 100 \times I_2) \\
\omega p_i &\sim Uniform(0.02, 0.35),\ \forall i \\
fc_i &\sim Uniform(0.20, 0.70),\ \forall i \\
\sigma^2_{\epsilon,i} &\sim Gamma(2, 0.05) \\
\sigma^2_{w,i} &\sim Gamma(2, 0.05) \\
\sigma^2_{v,i} &\sim Gamma(2, 0.05)
\end{aligned}$$

Since the TDR measurements are assumed to be well-calibrated and, since they are used extensively, we have available a fairly precise estimate of the

variation present in those measurements. Hence we center the variance of $\epsilon_{v,il,t}$ about this estimate and do not allow it vary widely about the center. We also center the variance $\epsilon_{w,ij,t}$ about 0.1 and do not allow its variance to be too large in order to preserve scaling of $w_{ij,t}$ between 0 and 1. Since ωp_i and fc_i are not well identified, we rely on informative priors to enable their estimation. The scaling of $w_{ij,t}$ and the weak identifiability of ecological parameters make the model sensitive to prior specification. As indicated from Section 3 we first fixed $\theta_{fc} = 3$ and carried out the analysis. Then, we fitted the model for $\theta_{fc} = 1, 2, 4, 5$ and compared their predictive performance on a hold-out data set. Let D_t denote the values of the WiSARD observation, $\{w_{ij,1}, ..., w_{ij,t}\}$ and the TDR observations $\{v_{il,1}, ..., v_{il,t}\}$ available up to time t and $\Theta = [\boldsymbol{\alpha}, \mathbf{h}, \boldsymbol{\beta}, \mathbf{a}_0, \mathbf{a}_1, \boldsymbol{\omega p}, \mathbf{fc}, \theta_{fc}]$ be set of all the model parameters, with $\mathbf{h} = [h_1, ..., h_n]$, $\boldsymbol{\beta} = [\beta_{0,1}, \beta_{1,1}..\beta_{0,n}, \beta_{1,n}]$, $\mathbf{a}_0 = [a_{0,11}, a_{0,12}..., a_{0,n1}, a_{0,n2}]$, $\mathbf{a}_1 = [a_{1,11}, a_{1,12}..., a_{1,n1}, a_{1,n2}]$. Then, given the information set D_t and Θ, the joint full conditional distribution of the state vector $\mathbf{m}_i = \{m_{i,1}, m_{i,2}, ..., m_{i,T}\}$ is given by

$$\begin{aligned} & p(m_{i,T}|D_T, \Theta) \times \prod_{t=1}^{T-1} p(m_{i,t}|m_{i,t+1}, ..., m_{i,T}, D_T, \Theta) \\ \propto \; & p(m_{i,T}|D_T, \Theta) \times \prod_{t=1}^{T-1} p(m_{i,t}|m_{i,t+1}, D_t, \Theta) \\ \propto \; & p(m_{i,T}|D_T, \Theta) \times \prod_{t=1}^{T-1} p(m_{i,t+1}|D_t, \Theta) p(m_{i,t}|D_t, \Theta) \end{aligned} \tag{8}$$

The forward filtering density in (8), $p(m_{i,t}|D_t, \Theta) \propto p(m_{i,t}|D_{t-1}, \Theta) \times \prod_j p(w_{ij,t}|m_{i,t}, D_{t-1}, \Theta) \times \prod_l p(v_{il,t}|m_{i,t}, D_{t-1}, \Theta)$, is Gaussian with mean $V_i \nu_i$ and variance V_i where,

$$\begin{aligned} V_i^{-1} &= \frac{1}{\sigma^2_{\epsilon,i}} + \frac{1}{\sigma^2_{w,i}} \sum_j a^2_{1,ij} + \frac{n_{v,i}}{\sigma^2_{v,i}} \\ \nu_i &= \frac{m^*_{i,t}}{\sigma^2_{\epsilon,i}} + \frac{1}{\sigma^2_{w,i}} \sum_j a_{1,ij}(w_{ij,t} - a_{0,ij}) + \frac{1}{\sigma^2_{v,i}} \sum_l v_{il,t} \end{aligned}$$

with

$$m^*_{i,t} = m_{i,t-1} - f(z_{i,t})(m_{i,t-1} - \omega p_i)^{\theta_{\omega p}}_+ - \alpha_i (m_{i,t-1} - fc_i)^{\theta_{fc}}_+ + (1 - e^{-h_i P_{i,t}}).$$

Then the state variables are sampled from the target density (8) using a Metropolis step. In order to obtain $m_{i,t}$s in the region where $w_{ij,t}$s are not observed, we treat the missing $w_{ij,t}$ as unknown and update them during the model fitting along with the model parameters Θ.

4.4 Cross validation

In order to assess the performance of the model, we perform a cross-validation on out-of-sample data points. That is, we do not know the true m's so we can only validate by holding out observed w's. The posterior predictive distribution required to perform this validation is obtained in the following fashion: Let $\mathbf{w}_t^*$ denote all the missing WiSARD data $w_{ij,t}^*$ up to time t, $\mathbf{w}_t = [w_{11,1}, w_{12,1}...w_{n1,t}, w_{n2,t}, v_{11,t}, ..., v_{nL,t}]$ denote all the WiSARD and TDR data observed up to time t and $\mathbf{w}$ denote the entire set of the WiSARD and TDR data in the training data set. For a new time point $t+1$, we have $w_{ij,t+1}|m_{i,t+1}, \Theta \sim N(a_{0,ij}+a_{1,ij}m_{i,t+1}, 1)$ and the process model is specified in (5). Let $m_{i,t+1}^{(1)}, ..., m_{i,t+1}^{(B_m)}, \mathbf{w}_t^{*(1)}, ...\mathbf{w}_t^{*(B_w)}, \Theta^{(1)}, ..., \Theta^{(B\Theta)}$ be the samples generated from the full posterior distribution $\pi(m_{i,t+1}, \mathbf{w}_t^*, \Theta|\mathbf{w})$. Then, the posterior predictive density of $w_{ij,t+1}$ is given by

$$\pi(w_{ij,t+1}|\mathbf{w}_t) \propto \int \pi(w_{ij,t+1}|\mathbf{m}_{t+1}, \Theta, \mathbf{w}_t^*, \mathbf{w}_t)\pi(\mathbf{m}_{t+1}, \mathbf{w}_t^*, \Theta|\mathbf{w}_t)d\mathbf{m}_{t+1}\, d\Theta\, d\mathbf{w}_t^* \tag{9}$$

Under model fitting using MCMC, the posterior predictive distribution in (9) is sampled by composition. Once we have samples from the posterior distribution of Θ and $\mathbf{w}_t^*$, we use the following algorithm to draw samples from the posterior predictive distribution (9).

1. Draw a sample $\theta^{(k)}, \mathbf{w}_t^{*(k)}$ from their posterior distribution.
2. Draw a sample of $m_{i,t+1}^{(k)}$ from its posterior distribution (8).
3. Finally draw $w_{ij,t+1}^{(k)}$ from $N(a_{0,ij}^{(k)} + a_{1,ij}^{(k)}m_{i,t+1}^{(k)}, 1)$.

5 Analysis of the Duke Forest data

To illustrate the performance of the posited hierarchical state-space model for soil moisture, we apply it to average daily soil moisture readings in the Duke

Forest as obtained from the WiSARD network and the TDR measurements during the growing season, from March 18, 2009 to December 8, 2009. We have run a single long MCMC chain. We generated 50000 MCMC samples and discarded the first 10000 as burn-in. Convergence was first assessed visually by looking at the trace plot and subsequently assessed quantitatively using Geweke's criterion. (Geweke, [13]). The average acceptance rate is around 40%. The computation time is roughly 2 hours. The results obtained from analyzing these data are summarized below.

As noted in Section 4.1, an important point in providing this analysis is the tension between the two data sources. That is, there is much more sensor data than TDR data. However, in principle, the TDR data is calibrated to the truth. How do we determine the appropriate balance? With a tight variance on the TDR data, the predictions will track the TDR; with a loose variance, the predictions will track the sensor data. We know that the sensor data can be unreliable due to the challenges of wireless sensor data collection implying substantial uncertainty. However, we also see that, at times, the TDR does not seem to respond to precipitation events (see Figure 5), as occurs when rainfall is intercepted by the canopy and does not penetrate to sensor depth. If the $m_{i,t}$ above are viewed as the true values at the sensor probes, then we can assume a larger error for the TDR measurements since they are not measured exactly at the probe locations.

Figure 3a shows the plot of the raw soil moisture measurements $w_{ij,t}$ (scaled by 100) for the aforementioned 4 nodes and for the available days. It also shows the plot of the TDR data (scaled by 100) observed during the period covered by the raw data. There is very little spatial variability of temperature and precipitation over the study region. Hence we propose to use the global values of these covariates, that is, we assume that temperature and precipitation vary with time only and not vary from node to node. So, we replace $P_{i,t}$ and $T_{i,t}$ is (2) and (4) by P_t and T_t, respectively. Figure 3b shows the plot of total daily precipitation (P_t) observed over the study region during the period covered by the raw soil moisture data. We see directly the response of soil moisture to precipitation. Figure 4 shows the plot of the daily average air temperature over the study period.

The plot of the raw soil moisture data found in the training set $w_{ij,t}$ and the estimated true measurements, $m_{i,t}$, for $\theta_{fc} = 3$ along with the TDR data for four of the sites- two gap sites (WiSARD id: 133, 152) and two understory sites (WiSARD id: 181, 302)- are shown in Figure 5. Note that the WiSARD sensors tend to provide much higher soil moisture measurements than the

TDR device, illuminating the need for calibration. Overlaid is the plot of daily precipitation amount (in cm). The posterior mean and the 95% credible interval for the parameters estimated from the training set for gap sites and understory sites are given in Table 2a and Table 2b, respectively.

We consider the posterior predictive distributions of the $w_{ij,t+1}$ corresponding to the hold-out data set using the method described in Section 4.4. In Figure 6 we plot the actual observed values of the WiSARD measurements, belonging to the hold-out dataset for $\theta_{fc} = 3$ along with their predicted values and the associated 95% predictive interval. We see that, generally, the prediction is good.

The value in the model fitting is that it allowed us: (i) to translate soil moisture dynamics from the sensor probes to the common scale represented by a portable standard, (ii) to gap-fill sequences through predictive distributions from the fitted state space model, and (iii) to infer the effect of temperature on transpiration rate in a dynamic model that contains process error and location-specific observation errors. Sensors can vary substantially, and they can drift between calibration events. By fusing data models for sensors at fixed locations with the portable standard we sidestep the calibration problems of individual sensors. Our latent soil moisture variable responds to the process model and wireless data, as calibrated by the portable standard.

Again, sensors fail frequently for a variety of reasons. Gap filling tends to be done on an ad hoc basis for most ecosystem data streams. Our model allows the latent soil moisture variable to be predicted by when sensors fail; our approach can be used fill out discontinuous data streams, with uncertainties, all stemming from data and process understanding.

The differences between gap and understory that emerge during the growing season (Figures 3a and 6) are attributed by the model to temperature effects on transpiration. The large positive temperature effects in the forest understory (e.g., β_1 estimates in Table 2b), where transpiration rates are high, are not identified in gap sites (Table 2a). More broadly, the integration of process-level understanding from the tradition of differential equations, with the inference possible under the hierarchical structure contributes to the objectives of environmental inference, i.e., estimation and prediction of states and quantification of the role of important input variables.

6 Summary and extensions

On the one hand, mathematical models of soil moisture have not been designed to accommodate uncertainty in process, data, and spatio-temporal variation. On the other hand, simple kriging of point measurements has been difficult to square with processes of transpiration, precipitation, and rainfall. The latter is critical for prediction in time and in space, for example, gap filling. Our novel data fusion model for soil moisture assimilates sensor network data with measurements from a portable TDR device. We formulate a highly nonlinear state space model motivated by a stochastic PDE like those used to relate soil moisture to dynamic processes. At the observation level, the sensor data need calibration while the TDR data introduce measurement error. We have fitted this hierarchical model using MCMC and validated it using hold-out data.

Again, we remark that we have specified a differential equation model primarily as motivation for the discretized-in-time analysis that we presented. Investigation at finer than daily scale would require modification of the modeling to capture diurnal behavior. This also discourages introduction of latent variables to achieve finer temporal resolution under the daily resolution for the data. We have alluded to the issue of prior sensitivity in terms of balance between the two data sources. As in any data assimilation approach, more confidence in one source relative to another will be reflected in more fidelity of predictions to that source. We have noted that soil moisture is difficult to measure. Hence, though we can envision richer model specifications, we would need better data than we currently have to justify pursuing them. Lastly, we note that the posited model is primarily predictive in nature, so inference about the model parameters is not key. Furthermore, we acknowledge that we need informative priors on key parameters. Consequently this model has not been fitted to simulated examples in order to study parameter inference.

Future work will find us looking into different data sources. One option is to adapt our approach to a manipulated setting, in particular a setting involving warming chambers. This will provide different temperature trajectories that are elevated compared to those in our current analysis and can further illuminate soil moisture behavior. Another problem involves seedling demography. Health of seedlings is very sensitive to available water; linking soil moisture at fine scale to seedling performance would be valuable.

Acknowledgments

The authors thank Gabriel Katul for useful discussions. This research was supported in part by grant DDDAS 054034L and 050414NL.

References

[1] J. D. Albertson, N. Montaldo, Temporal dynamics of soil moisture variability: 1. Theoretical basis, Water Resources Research 39 (2003) 1274, doi:10.1029/2002WR001616.

[2] A. M. Barton, Factors controlling plant distributions: drought, competition, and fire in montane pines in Arizona, Ecological Monographs 63 (1993) 367-397.

[3] L. M. Berliner, Physical-statistical modeling in geophysics, Journal of Geophysical Research 108 (2003) 15.

[4] M. Charpentier, P. Groffman, Soil moisture variability within remote sensing pixels, Journal of Geophysical Research 97 (1992) 18987–18995.

[5] J. S. Clark, P. Agarwal, D. M. Bell, P. Flikkema, A. E. Gelfand, X. Nguyen, E. Ward, J. Yang, Inferential ecosystem models, from network data to prediction, Ecological Applications 21 (2011) 1523–1536.

[6] D. A. Coomes, P. J. Grubb, Impacts of root competition in forests and woodlands: a theoretical framework and review of experiments, Ecological Monographs 70 (2000) 171–207.

[7] W. Crow, E. Wood Impact of soil moisture aggregation on surface energy flux prediction during SGP97, Geophysical Research Letters 29 (2002) 1008, doi:10.1029/2001GL013796.

[8] O. Elerian, S. Chib, N. Shepard, Likelihood inference for discretely observed nonlinear diffusions, Econometrica 4 (2001) 959–993.

[9] J. K. Entin, A. Robock, K. Y. Vinnikov, S. E. Hollinger, S. Liu, A. Namkhai, Temporal and spatial scales of observed soil moisture variations in the extratropics, Journal of Geophysical Research 105 (2003) 11865–11877.

[10] B. Eraker, MCMC analysis of diffusion models with applications to finance, Journal of Business and Economic Statistics 19 (2001) 177–191.

[11] J. J. Famiglietti, J. Rudnicki, M. Rodell , Variability in surface moisture content along a hillslope transect: Rattlesnake Hill, Texas, Journal of Hydrology 210 (1998) 259–281.

[12] M. Fuentes M, A. E. Raftery, Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models, Biometrics 66 (2005) 36–45.

[13] J. Geweke, Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In Bayesian Statistics 4 (ed JM Bernado, JO Berger, AP Dawid and AFM Smith) (1992) Clarendon Press, Oxford, UK.

[14] V. V. Junior, M. P. Carvalho, J. Dafonte,O. S. Freddi, E. V. Vazquez, O. E. Ingaramoc, Spatial variability of soil water content and mechanical resistance of Brazilian ferralsol, Soil and Tillage Research 85 (2006) 166–177.

[15] G. Katul, P. Todd, D. Pataki, Z. J. Kabala, R. Oren, Soil water depletion by oak trees and the inuence of root water uptake on the moisture content spatial statistics, Water Resources Research 33 (1997) 611–623.

[16] G. Katul, A. Porporato, R. Oren, Stochastic dynamics of plant-water interactions, Annual Review of Ecology, Evolution, and Systematics 38 (2007) 767–791.

[17] H. R. Kleb, S. D. Wilson, Vegetation effects on soil resource heterogeneity in prairie and forest, American Naturalist 150 (1997) 283–298.

[18] C.F. Korstian, T. S. Coile, Plant competition in forest stands, Duke University School of Forestry Bulletin 3 (1938) 125.

[19] F. Laio, A. Porporato, L. Ridolfi, I. Rodriguez-Iturbe, Plants in water-controlled ecosystems: active role in hydrologic processes and response to water stress II. Probabilistic soil moisture dynamics, Advances in Water Resources 24 (2001a) 707–723.

[20] F. Laio, A. Porporato, C. P. Fernandez-Illescas, I. Rodriguez-Iturbe, Plants in water-controlled ecosystems: active role in hydrologic processes and response to water stress IV. Discussion of real cases, Advances in Water Resources 24 (2001b) 745–762.

[21] A. Oldak, Y. Pachepsky, T. J. Jackson, W. J. Rawls, Statistical properties of soil moisture images revisited, Journal of Hydrology 255 (2002) 12–24.

[22] D. Penna, M. Borgaa, D. Norbiatoa, G. D. Fontana, Hillslope scale soil moisture variability in a steep alpine terrain, Journal of Hydrology 364 (2009) 311–327.

[23] Y. Qiu, B. Fua, J. Wang, L. Chen, Spatial variability of soil moisture content and its relation to environmental indices in a semi-arid gully catchment of the Loess Plateau, China, Journal of Arid Environments 49 (2001) 723–750.

[24] I. Rodriguez-Iturbe, A. Porporato, F. Laio, L .Ridolfi, Plants in water-controlled ecosystems: active role in hydrologic processes and response to water stress I. Scope and general outline, Advances in Water Resources 24 (2001) 695–705.

[25] N. Sturm, S. Reber, A. Kessler, J. D. Tenhunen, Soil moisture variation and plant water stress at the Hartheim Scots pine plantation, Theoretical and Applied Climatology 53 (1996) 123–133.

[26] A. W. Western, R. B. Grayson, G. Bloschl, G. R. Willgoose, T. A. McMahon, Observed spatial organization of soil moisture and its relation to terrain indices, Water Resources Research 35 (1999) 797–810.

[27] C. K. Wikle, Hierarchical Bayesian models for predicting the spread of ecological processes, Ecology 84 (2003) 1382–1394.

[28] X. Zhou, H. Lin, Q. Zhu, Temporal stability of soil moisture spatial variability at two scales and its implication for optimal field monitoring, Hydrology and Earth System Science Discussions 4 (2007) 1185–1214.

Tables and Figures

TABLE 1
MSPE obtained for various values of θ_{fc}

$\boldsymbol{\theta_{fc}}$	**MSPE**
1	0.00150
2	0.00081
3	0.00059
4	0.00068
5	0.00071

TABLE 2A
Posterior mean and 95% credible interval for model parameters for Gap sites

Parameters	**WiSARD id**	
	node 133	node 152
α_i	0.41 (0.03, 0.73)	0.56 (0.02,0.81)
ωp_i	0.27 (0.18, 0.34)	0.15 (0.11,0.20)
fc_i	0.53 (0.45, 0.57)	0.44 (0.40,0.53)
$\beta_{0,i}$	-12.32 (-29.20, -2.95)	-13.82 (-28.83, -4.94)
$\beta_{1,i}$	1.12 (-4.78, 9.74)	1.55 (-4.65, 10.00)
h_i	0.0021 (0.0001, 0.0047)	0.003 (0.0003, 0.0076)
$\sigma^2_{\epsilon,i}$	0.0009 (0.0007, 0.0013)	0.0009 (0.0007, 0.0014)
$\sigma^2_{w,i}$	0.0015 0.0012, 0.0019)	0.0079 (0.0065, 0.0086)
$\sigma^2_{g,i}$	0.0015 (0.0011, 0.0018)	0.0062 (0.0057, 0.0068)
$a_{0,i1}$	0.203 (0.17, 0.25)	0.157 (0.10, 0.18)
$a_{0,i2}$	-0.1761 (-0.21,-0.13)	0.045 (-0.01, 0.09)
$a_{1,i1}$	0.719 (0.63, 0. 78)	0.724 (0.64, 0.82)
$a_{1,i2}$	1.367 (1.28, 1.40)	0.851 (0.73, 0.91)

TABLE 2B
Posterior mean and 95% credible interval for model parameters for Understory sites

Parameters	**WiSARD id**	
	node 181	node 302
α_i	0.45 (0.24, 0.63)	0.75 (0.41, 0.89)
ωp_i	0.14 (0.10. 0.16)	0.13 (0.11, 0.16)
fc_i	0.29 (0.25, 0.32)	0.28 (0.26, 0.31)
$\beta_{0,i}$	-3.06 (-11.87, -1.59)	-1.51 (-10.30,-0.23)
$\beta_{1,i}$	3.96 (1.15, 16.32)	4.36 (1.35, 22.49)
h_i	0.0025 (0.0002, 0.0007)	0.0023 (0.0002,0.0073)
$\sigma^2_{\epsilon,i}$	0.0007 (0.0006, 0.0009)	0.0008 (0.0006, 0.0010)
$\sigma^2_{w,i}$	0.0078 (0.0051, 0.0091)	0.0063(0.0055, 0.0074)
$\sigma^2_{g,i}$	0.00027 (0.00021, 0.00034)	0.00078 (0.00072, 0.00085)
$a_{0,i1}$	0.027 (0.015, 0.042)	0.043 (0.029, 0.055)
$a_{0,i2}$	0.136 (0.11, 0.15)	0.103 (0.07, 0.16)
$a_{1,i1}$	0.986 (0.87, 1.06)	0.823 (0.76, 0.93)
$a_{1,i2}$	0.973 (0. 84, 1.08)	1.166 (1.03, 1.24)

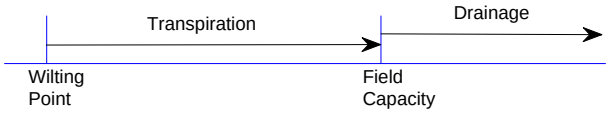


FIG. 1. A line diagram illustrating the operation of transpiration and drainage with respect to wilting point and field capacity.

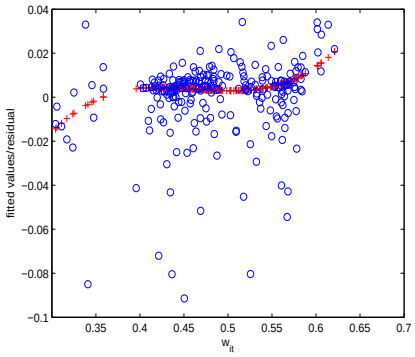


FIG. 2A.

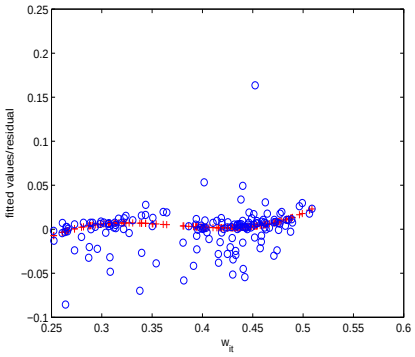


FIG. 2B.

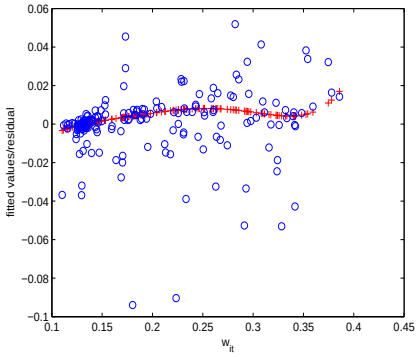


FIG. 2C.

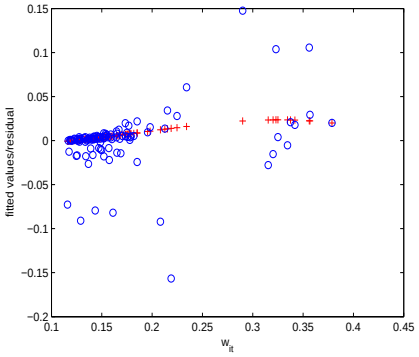


FIG. 2D.

FIG. 2. Plot of the first stage residuals (circles) and the curve fitted to these residuals (crosses), with $\theta_{fc} = 3$, against $w_{i,t}$ corresponding to WiSARD id's (A) 133 (B) 152 (C) 181 and (D) 302.

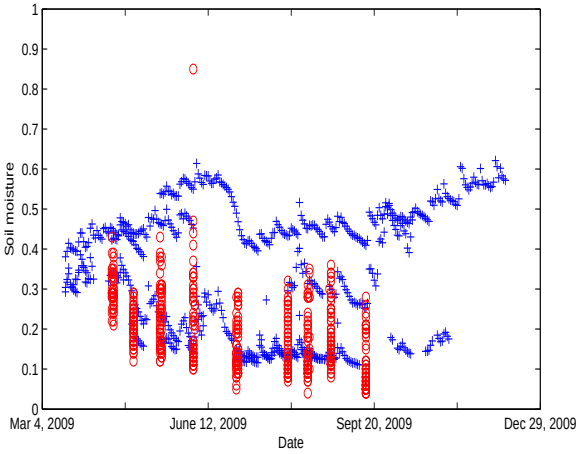


FIG. 3A. Plot of raw (scaled) WiSARD (crosses) and TDR (circles) data High values of WiSARD observations mostly correspond to gap sites while low values mostly correspond to understory sites

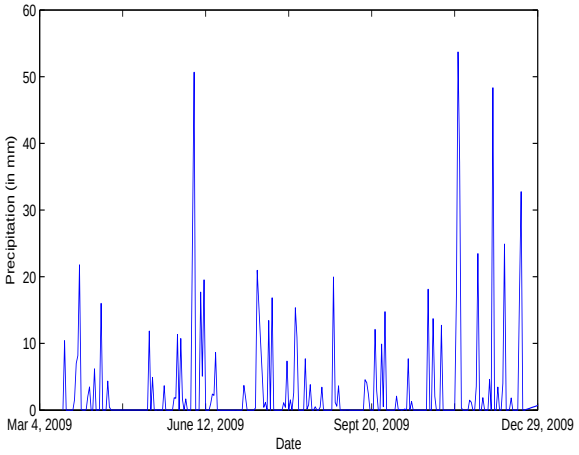


FIG. 3B. Plot of total daily precipitation amount during the study period

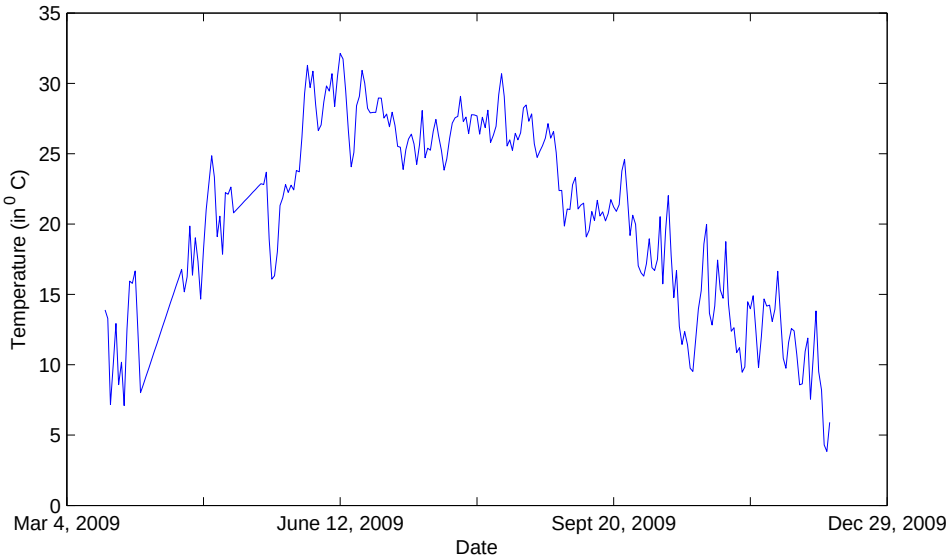


FIG. 4. Plot of daily-average air temperature

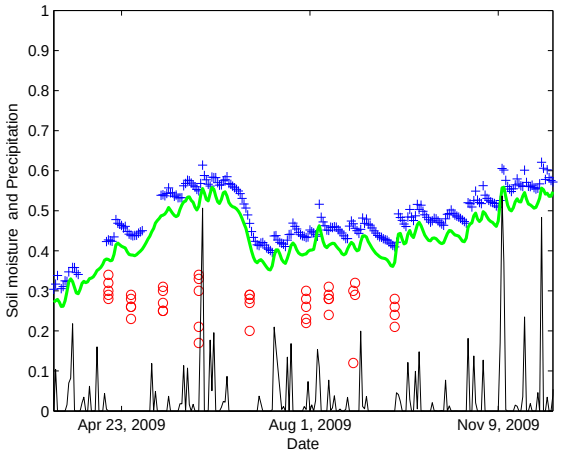


FIG. 5A.

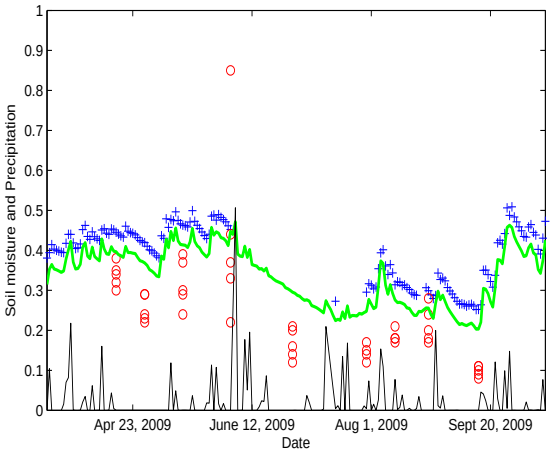


FIG. 5B.

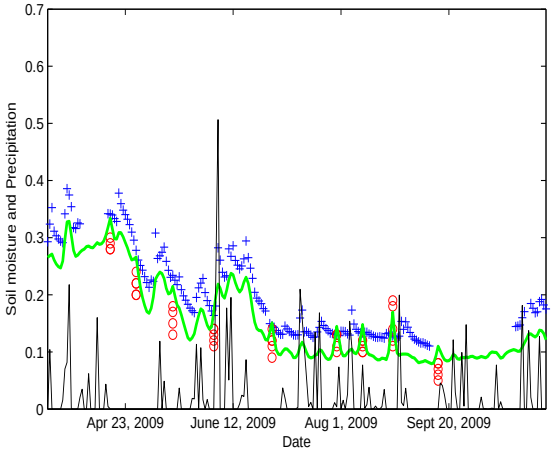


FIG. 5C.

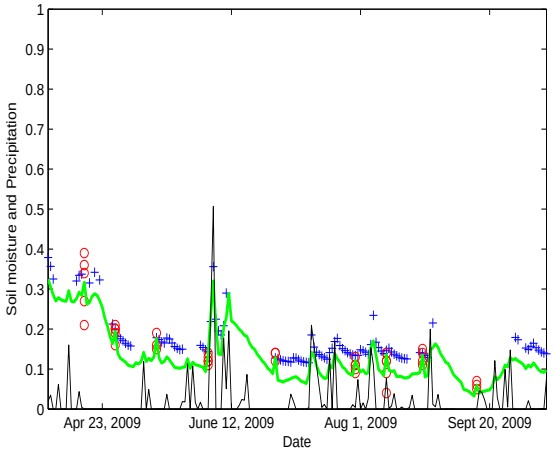


FIG. 5D.

FIG. 5. Plot of WiSARD (crosses) and TDR (circles) measurements along with estimated true measurements (solid thick) corresponding to WiSARD id's (A) 133 (B) 152 (C) 181 and (D) 302. The figures on the top panel correspond to gap sites while those below correspond to understory sites. Overlaid is the plot of precipitation amounts (solid thin) in cm

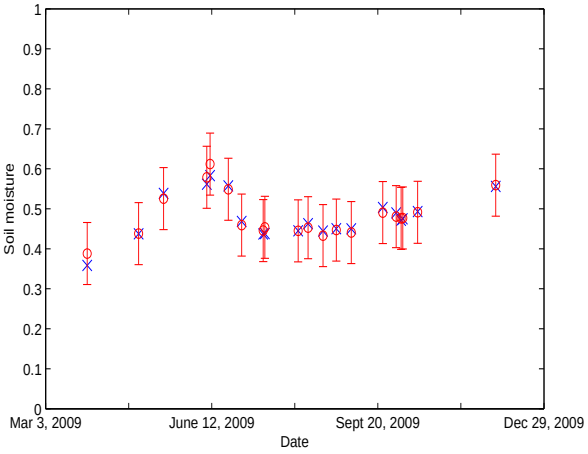


FIG. 6A.

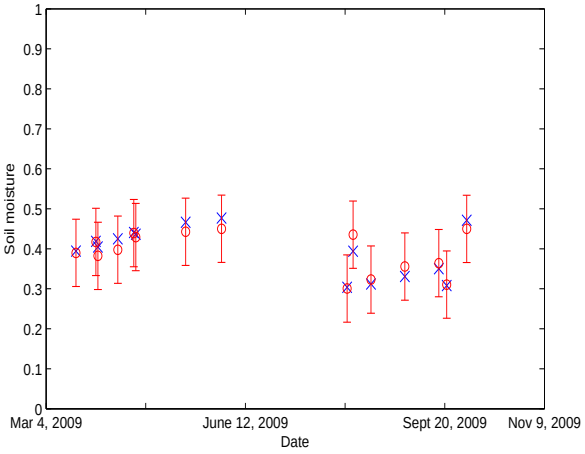


FIG. 6B.

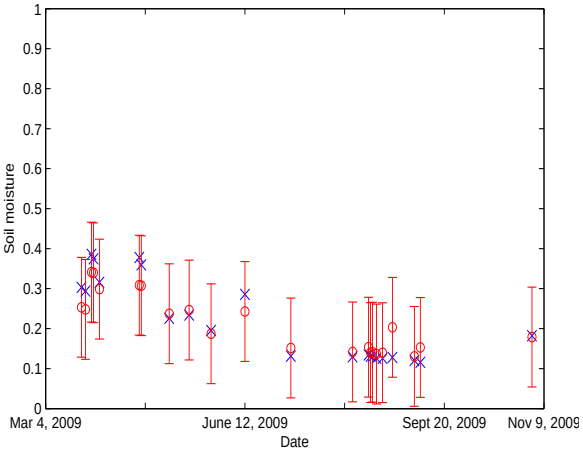


FIG. 6C.

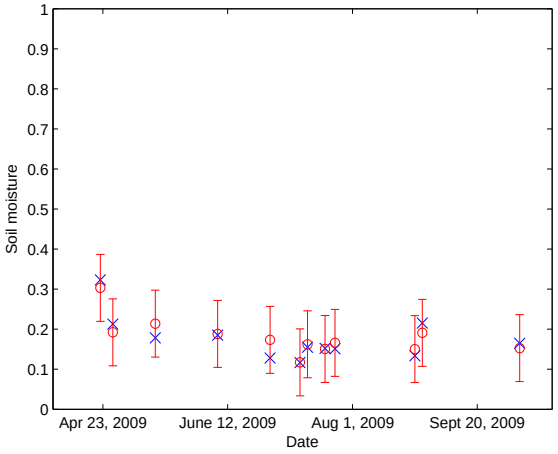


FIG. 6D.

FIG. 6. Plot of observed (crosses) and predicted (circles) $w_{ij,t}$ in the test dataset and 95% predictive interval for $\theta_{fc} = 3$ corresponding to WiSARD id's (A) 133 (B) 152 (C) 181 and (D) 302. The figures on the top panel correspond to gap sites while those below correspond to understory sites.