Application of Data Assimilation with the Root Zone Water Quality Model for Soil Moisture Profile Estimation in the Upper Cedar Creek, Indiana

Eunjin Han¹, Venkatesh Merwade²* and Gary C. Heathman³

¹Graduate Research Assistant, School of Civil Engineering, Purdue University, West Lafayette, IN. e-mail: ehan@purdue.edu

²Assistant Professor, School of Civil Engineering, Purdue University, West Lafayette, IN; Corresponding author; e-mail: vmerwade@purdue.edu

³Soil Scientist, USDA-ARS, National Soil Erosion Research Laboratory, West Lafayette, IN; email: Gary.Heathman@ars.usda.gov
Abstract

Data assimilation techniques have been proven as an effective tool to improve model forecasts by combining information about observed variables in many areas. This paper examines the potential of assimilating surface soil moisture observations into a field-scale hydrologic model, the Root Zone Water Quality Model, to improve soil moisture estimation. The Ensemble Kalman Filter (EnKF), a popular data assimilation technique for non-linear systems, was applied and compared to a simple direct insertion method. In situ soil moisture data at four different depths (5 cm, 20 cm, 40 cm and 60 cm) from two agricultural fields (AS1 and AS2) in northeastern Indiana were used for assimilation and validation purposes. Through daily update, the EnKF improved soil moisture estimation compared to the direct insertion method and model results without assimilation, having more distinct improvement at the 5 cm and 20 cm depths than for deeper layers (40 cm and 60 cm). Local vertical soil property heterogeneity in AS1 deteriorated soil moisture estimates with the EnKF. Removal of systematic bias in the forecast model was found to be critical for more successful soil moisture data assimilation studies. The study also demonstrates that more frequent update generally contribute to enhance the open loop simulation, however, large forecasting error can prevent more frequent update from providing better results. In addition, results indicate that various ensemble sizes make little difference in the assimilation results. An ensemble of 100 members produced results that were comparable to results obtained from larger ensembles.

Key words: Soil Moisture; Root Zone Water Quality Model (RZWQM); Data Assimilation; Direct Insertion; Ensemble Kalman Filter; Cedar Creek Watershed
1. Introduction

Estimation of soil moisture has received considerable attention in the areas of hydrology, agriculture, meteorology and environmental studies because of its role in the partitioning of water and energy at the land surface, specifically, precipitation into runoff and infiltration, and energy into latent and sensible heat fluxes. Better understanding and prediction of soil moisture in the root zone is beneficial in irrigation planning and crop management, flooding and drought prediction, water quality management, climate change and weather prediction. However, despite its significance, the expense associated with field measurements and high spatio-temporal variability of soil properties, have placed limitations in obtaining in situ measurements of profile soil moisture. Recent developments in remote sensing technology have provided possibilities to overcome these limitations through soil moisture retrieval from remote sensing images (Schmugge and Jackson, 1994; Jackson et al., 1995; Jackson and Vine, 1996; Jackson et al., 2002; Njoku et al., 2003; Verstraeten et al., 2006; Baup et al., 2007). However, in most cases, the actual sensing depth is limited to a few centimeters allowing soil moisture estimates for the top 0-5 cm and 1 cm surface layer using the L-band (~1-2 GHz) and C-band (~4-8 GHz), respectively (Jackson and Schmugge, 1989; Njoku et al., 2003).

Shallow soil moisture estimates from remotely sensed data has led researchers to integrate measured surface data and hydrologic models to obtain more accurate estimates of soil moisture content in the root zone through data assimilation techniques (Houser et al., 1998; Li and Islam, 1999; Hoeben and Troch, 2000; Walker et al., 2001b; Heathman et al., 2003a). Various data assimilation techniques have been used in soil moisture studies including the Kalman Filter (Hoeben and Troch, 2000; Walker et al., 2001b; Crosson et al., 2002), direct insertion method (Walker et al., 2001b; Heathman et al., 2003a), extended Kalman Filter (Reichle et al., 2002b; Seuffert et al., 2004; Draper et al., 2009) and Ensemble Kalman Filter (Reichle et al., 2002a, 2002b; Zhang et al., 2006). Among these techniques, the Ensemble Kalman Filter (EnKF) has been widely adopted because of its strength in handling non-linear systems and computational efficiency (Crow and Wood, 2003; Huang et al., 2008). As one of surface measurements from a remote platform, brightness temperature has been assimilated into hydrologic models for better estimation of soil moisture (Margulis et al., 2002; Crow and Wood, 2003; Huang et al., 2008). Recently, satellite-based surface soil moisture observation (0-5 cm)
from the Scanning Channel Microwave Radiometer (SMMR) and the Advanced Microwave
Scanning Radiometer – Earth Observing System (AMSR-E) have been integrated with land
surface models through data assimilation (Reichle and Koster, 2005; Ni-Meister et al., 2006;
Reichle et al., 2007; Draper et al., 2009).

Most previous investigations that explored assimilation of remotely sensed surface soil moisture
involved regional or global scales for more accurate climate forecasts by improving temporal and
spatial soil moisture estimations (Walker and Houser, 2001; Reichle et al., 2002b; Ni-Meister,
2008; Draper et al., 2009). However, Troch et al. (2003) stressed the necessity of extending the
applicability of data assimilation, which has so far been studied in land surface models using
synthetic datasets, to many other important hydrological issues at catchment scale for a variety of
water resources management problems. Therefore, there is a strong need for more application-
 focused research to determine the operational potentials of data assimilation within local scale
hydrologic modeling using actual observations.

There have been a few studies which have applied soil moisture data assimilation techniques for
improved hydrologic predictions at field or catchment scales. Walker et al. (2002) demonstrated
feasibility of near-surface soil moisture assimilation to retrieve profile soil moisture in a 6 ha
catchment using a three dimensional distributed soil moisture model. In-situ soil moisture
observations in an agricultural field (21 ha) were used for soil moisture assimilation with the
EnKF, specifically for correcting bias (De Lannoy et al., 2007a; De Lannoy et al., 2007b) and
for applying an adaptive EnKF (De Lannoy et al., 2009), with the Community Land Model.
Recently, a catchment scale hydrologic model, Soil and Water Assessment Tool, was used to
investigate the performance of the EnKF for hydrologic predictions (Xie and Zhang, 2010; Chen
et al., 2011).

Heathman et al. (2003a) used a physically based and field-scale agricultural model, the Root
Zone Water Quality Model (RZWQM) to explore the possibility of assimilating observed surface
soil moisture for better estimates of root zone soil water content. Based on substantial field
measured data, they showed that the direct insertion data assimilation technique produced
dynamics of model simulation results in the top 30 cm layers better than the model simulation
without assimilation. Since they focused on using field measured surface soil moisture to
estimate root zone soil water content rather than synthetic analyses, their study used the simple
direct insertion method. As an extension to Heathman et al. (2003), this paper examines the
benefit of assimilating surface soil moisture through the EnKF to improve model simulated soil
moisture in the root zone using the RZWQM. In addition, contrary to many previous studies
which assimilated remotely sensed surface soil moisture into land surface models at the regional
or global scale, this study applies field measured surface soil moisture to a point scale model
(RZWQM) for the purpose of data assimilation.

Field scale application of the EnKF in this study provides the fundamental groundwork
necessary to advance our knowledge of surface soil moisture data assimilation in hydrologic
modeling. First, field scale, deterministic models such as RZWQM can serve as the basis for the
watershed scale hydrologic or water quality models (Abrahamson et al., 2006) and has been
directed to extend the modeling methodology from one dimension to two or three dimensions
(Walker et al., 2001a; Ma et al., 2007a). Second, in situ soil moisture observations have fewer
uncertainties than the remotely sensed soil moisture data, which is better for testing different data
assimilation schemes and validating the assimilation results more effectively. This is because
observation errors for in situ soil moisture data can be estimated in a more reasonable way and
thus, the in situ observations can be considered to represent the true system state with certain
errors. Third, a field scale assimilation study can be based on more accurate soil characteristics
data which are essential for water flow simulation in the unsaturated zone possibly reducing
systematic model bias before assimilation. Walker et al., (2001a) discussed that incorrect
information about soil porosity or residual soil moisture content restrained the soil moisture
model from correct retrieval of profile soil moisture, even with an assimilation scheme. Lastly,
it is possible with the field scale studies to investigate the impacts of local horizontal or vertical
soil heterogeneity on data assimilation results, which is generally ignored in large scale studies.

RZWQM is one of the most representative agricultural system models with components for
water movement, plant growth, and chemical transport with management effects (Ma et al.,
2007a). Since its release in 1992 (Ahuja et al., 2000), RZWQM has been widely applied and
tested in different areas for various purposes (Ma et al., 2001): to determine soil hydraulic
properties (Cameira et al., 2000), to simulate tile drainage and leached nitrate (Abrahamson et al.,
Among major components of RZWQM, soil hydrological components have performed satisfactorily (Cameira et al., 2000; 2005; Kozak et al., 2007) and soil water content were reasonably simulated (Farahani et al., 1999; Ma et al., 2003; 2007b; 2007c). In addition, RZWQM has been successfully used to induce saturated hydraulic conductivities for spatially distributed soils by coupling with microwave remotely sensed soil moisture (Mattikalli et al., 1998).

Thus, this study has the following three objectives: (i) apply EnKF to RZWQM using point scale soil moisture data; (ii) compare the results from EnKF application with simple direct insertion method; and (iii) explore the effect of ensemble size and update interval on model output. The above objectives were accomplished by using point scale profile soil moisture data collected for Upper Cedar Creek Watershed (UCCW) in northeast Indiana. Details of the study area including data acquisition are provided in the next section.

2. Study area and data

The Matson Ditch sub-catchment in the upper Cedar Creek Watershed located in northeastern Indiana (Fig. 1) was selected as the test bed for this study. Since 2004, the National Soil Erosion Research Laboratory (NSERL) of the USDA – Agricultural Research Service (ARS) has established an extensive soil moisture and weather monitoring network in the upper Cedar Creek Watershed as shown in Fig. 1. The weather stations in this network collect the following meteorological data for model simulations: ten minute rainfall, air temperature, solar radiation, wind speed, and relative humidity.

Among the ten soil monitoring sites within the upper Cedar Creek network, AS1 and AS2 field-size watersheds were selected for RZWQM simulations in this study because of the availability of measured soil properties data (e.g., soil texture, bulk density) and crop management information. The drainage area of the AS1 and AS2 field sites is 2.23 ha and 2.71 ha, respectively. Both sites are fairly flat (slope < 5%), and are in agriculture production (alternative cropping of corn and soybean) with AS1 being in no-till, and AS2 under a rotational tillage system. The major soil types found at these sites are Glynwood (GnB2) silt loam (Fine, illitic,
mesic Aquic Hapludalfs) for AS1 and Blount (BaB2) silt loam (Fine, illitic, mesic, Aeric
Epiaqualfs) for AS2.

The soil sensors in both fields are Stevens SDI-12 Hydra Probes that measure soil moisture and
temperature every ten minutes by propagating an electromagnetic signal into the soil, which is a
Frequency Domain Reflectometry (vs Time Domain Reflectometry). Out of four factory
calibration equations for soil moisture measurements, the most current calibration equation
(Equation 4) was used in this study. The calibration equation is based on the linear relationship
between volumetric water content and the real component of the complex dielectric permittivity,
which has been proven as reasonably accurate for many soils (Topp et al., 1980; Topp and Davis,
1985; Heathman et al., 2003b; Seyfried and Murdock, 2004). In addition, Seyfried et al. (2005)
concluded that Equation 4 outperformed the other factory-supplied calibrations. According to
texture analysis in our laboratory and Seyfried et al (2005), the calibration coefficients of the
Equation 4 for loam soil type were determined. Since the Loam setting is applicable for Loam,
Clay Loam, and Silty Clay Loam textures according to the Hydra Probe manual (Stevens Water
Monitoring Systems, Inc., Portland, Oregon, USA, 2007), the selection of the calibration
coefficients for loam is suitable for the soil types in this study (Table 1).

Soil moisture measurements were obtained at four depths by individual sensors installed at 5 cm,
20 cm, 40 cm and 60 cm. Accuracy of the soil moisture measurement is reported as ±0.03 water
fraction by volume in typical soil by the manufacturer. Seyfried et al. (2005) demonstrated the
dielectric loss corrected calibration could reduce soil moisture measurement errors. Considering
that measured soil moisture in this study was not corrected with the dielectric loss and near
saturated soil condition has the largest errors (Seyfried et al., 2005), actual observation error of
the soil moisture may be slightly higher than 0.03m³m⁻³.

In this study, the measured surface soil moisture data (5 cm) are used for data assimilation, and
the data at 5, 20, 40 and 60 cm are used for validation purposes. In order to separate the
observations that are assimilated into the model from the observations that are used to verify
subsequent data assimilation results, especially for the 5 cm observations, the observations (Y_k
in Equation (4)) are compared to the model predicted soil moisture before assimilation step (X_k⁻¹
in Equation (6)) for results analysis. That is, the two sets of observations are considered completely mutually exclusive. The measured soil moisture and temperature data are also used to set up the initial conditions in RZWQM.

Measured soil moisture data during the simulation period (April – October 2007) at the two study sites are displayed in Fig. 2. Observed meteorological data was collected at AS1, and used as model input for both sites since the fields are less than 500 m apart. The top soil layers (0-5 cm), especially for AS1, show very dynamic variations in soil water content with rainfall; whereas deeper layers show more stable soil water content conditions.

3. Methodology

In this experiment, we first ran the RZWQM with the input data described in the previous section without any assimilation. This run is called the open loop simulation. Then the model was run for the same period with two different data assimilation techniques: the direct insertion method (DIR) and the Ensemble Kalman Filter (EnKF). The ability of the EnKF in improving the simulated soil moisture profile is evaluated by comparing the results of the open loop and the DIR. Time series graphs and statistical indices such as correlation coefficient, root mean square error and mean bias error are used for the evaluation.

In this section, we begin with a brief description of the RZWQM focusing on hydrologic processes, which is a model operator in the data assimilation system. A brief review of the EnKF is given in section 3.2. Section 3.3 describes how the data assimilation algorithm is implemented in the RZWQM to integrate observed surface soil moisture with the model prediction.

3.1 Root Zone Water Quality Model (RZWQM)

RZWQM is a physically based, one-dimensional deterministic model that uses fundamental flow equations to simulate infiltration and redistribution in the subsurface region. During a rainfall event, water infiltration is simulated by a modified form of the Green-Ampt equation (Equation (1)) given below after (Ahuja et al., 2000).
\[ V = K_s \frac{\tau_c + H_0 + Z_{wf}}{Z_{wf}} \]  \hspace{1cm} (1)

where \( V \) is the infiltration rate at any given time (cm h\(^{-1}\)), \( K_s \) is the effective average saturated hydraulic conductivity of the wetting zone (cm h\(^{-1}\)), \( \tau_c \) is the capillary drive or suction head at the wetting front (cm), \( H_0 \) is the depth of surface ponding (cm), and \( Z_{wf} \) is the depth of the wetting front (cm).

Redistribution between storm events is simulated by a mixed form of the Richards’ equation as given by Equation (2) after (Ahuja et al., 2000).

\[
\frac{\partial \theta}{\partial t} = \frac{\partial}{\partial z} \left[ K(h, z) \frac{\partial h}{\partial z} - K(h, z) \right] - S(z, t) \]  \hspace{1cm} (2)

where \( \theta \) is the volumetric soil water content (cm\(^3\) cm\(^{-3}\)), \( t \) is time (h), \( z \) is the soil depth (cm), \( h \) is the soil-water pressure head (cm), \( K \) is the unsaturated hydraulic conductivity (cm h\(^{-1}\)), and \( S(z, t) \) is the sink term (h\(^{-1}\)).

RZWQM adopts a general mass-conservative numerical solution of Celia et al. (1990) to solve the Richards’ equation. The sink term takes into account plant water uptake and tile drainage, if present. All other biological and chemical processes for plant growth and movement of nutrients and pesticides in RZWQM are simulated following the physical water flow process. More detailed information for biological and chemical processes in RZWQM can be found in Ahuja et al. (2000).

Measured soil physical and hydraulic properties at the two sites (AS1 and AS2) were obtained from field samples analyzed at the NSERL lab or determined in situ and used as model input (Table 1). RZWQM provides various options for the estimation of soil hydraulic properties depending on the availability of the data (Ahuja et al., 2000). In this study, minimum necessary input (soil texture, -33 kPa water content and bulk density for each soil layer) are provided and other properties are estimated by the model (Table 1). For the constitutive relationship between \( h-\theta-K \) which is a key for the numerical solution of the Richards’ equation, RZWQM uses functional forms of those relationships based on the modified Brooks-Corey equation. The
parameters for the Brooks-Corey equation were compiled by Rawls et al. (1982) for major USDA soil textures. Validation of this estimation technique can be found in Starks et al. (2003) who showed that the limited input data option, based on textural class name, predicted soil water content as well as the input from more detailed laboratory measurements. Thus, the parameter values for Brooks-Corey equation are used from Rawls et al. (1982) in this study.

3.2 Ensemble Kalman Filter (EnKF) Assimilation

The main concept of EnKF, which includes forecasting the error statistics using Monte Carlo methods, was first introduced by Evensen (1994), and has been applied to various fields, especially in oceanography and meteorology (Evensen and Leeuwen, 1996; Burgers et al., 1998; Keppenne, 2000; Houtekamer and Mitchell, 2001). There are slight differences between various ensemble-based filter approaches. For instance, Houtekamer and Mitchell (1998) used two ensembles of model states for forecasting, and updated each ensemble with error covariance from the other ensemble. Whitaker and Hamill (2002) suggested a variant of the EnKF by avoiding perturbation of observations. This study follows the algorithm proposed by Evensen (2003,2004) and notation presented by Reichle et al. (2002b). In this approach, a non-linear system model can be expressed with the following generic form:

\[ X_{k+1} = f_k(X_k) + w_k \]  \hspace{1cm} (3)

where \( X_k \) is the system state of interest (in our case soil moisture water content), \( f_k(\cdot) \) indicates non-linear model operator at time step \( k \), and \( w_k \) is the system error that accounts for all uncertainties in the model physics or forcing data. The \( w_k \) term is treated as a normally distributed random variable with zero mean and covariance \( Q_k \).

Observed soil moisture values (\( Y_k \)) are related to the true state (\( X_k \)) through a measurement operator (\( H_k \)) as shown in Equation (4):

\[ Y_k = H_k X_k + v_k \]  \hspace{1cm} (4)
Because measured soil moisture data are available in this study, $H_k$ is an identity matrix. The $v_k$ term represents errors in the measurement instrument and procedures, and conforms to a Gaussian distribution with zero mean and covariance $R$. The $v_k$ term represents errors in the measurement instrument and procedures, and conforms to a Gaussian distribution with zero mean and covariance $R$.

The Kalman filter mainly consists of forecasting and updating steps (Gelb, 1974). Before starting the forecasting or updating of EnKF, an initial ensemble of size $N$ is created by adding pseudorandom noise $e_i$ with zero mean and covariance $P$ to get the first “guess” state $X_0$ as shown in Equation (5) below.

$$X_0^{i+} = X_0 + e_i, \quad e_i \sim N(0, P) \quad i=1, \ldots, N.$$  \hspace{1cm} (5)

As a next step, forecasting is performed to estimate the state at a future time by integrating the initially created ensemble through the non-linear model operator, which in our study is redistribution or the infiltration process.

$$X_k^{i-} = f_{k-1}(X_{k-1}^{i+}) + w_k^i, \quad i=1, \ldots, N.$$  \hspace{1cm} (6)

In Equation (5) and (6), the superscripts ‘-’ and ‘+’ refer to state variables that are forecasted and updated, respectively. In the standard Kalman filter algorithm, uncertainties in the system are expressed with state error covariance. Contrary to the standard Kalman filter or extended Kalman filter, the EnKF does not require the propagation of state error covariance $P_k^-$ explicitly.

Propagation of state error covariance is computationally expensive for a large system, and its exclusion is one of the advantages offered by the EnKF. Instead of forecasting the state error covariance $P_k^-$ from the covariance of the previous time step $P_{k-1}^+$, the EnKF estimates error covariance from the forecasted ensemble of the state and their mean as shown in Equation (7) below.

$$P_k^- = \frac{1}{N-1} D_k D_k^T$$  \hspace{1cm} (7)

where $D_k = [X_k^{1-} - X_k^-, \ldots, X_k^{N-} - X_k^-]$ and $X_k^- = \frac{1}{N} \sum_{i=1}^N X_k^{i-}$.
Whenever measured data are available, forecasted system variables are updated through the weighted sum of forecasted variables and observed data. The Kalman gain, $K_k$ works as a weight for the updating step as shown below.

$$K_k = P_k^{-}H_k^{T}\left[H_kP_k^{-}H_k^{T} + R_k\right]^{-1} \quad (8)$$

$$X_k^{i+} = X_k^{i-} + K_k[Y_k - H_kX_k^{i-} + v_k^i] \quad i = 1,...,N \quad (9)$$

As Equation (9) shows, in the update step, that the ensemble of observation size $N$ are created by adding random perturbations ($v_k^i$) with the mean equal to zero and observation variance $R_k$. After each member of ensemble variables are updated, a best estimate of the system variable at time $k$ is found by averaging the updated ensemble members.

### 3.3 Application of EnKF to RZWQM

The basic framework of RZWQM, and how EnKF is incorporated into RZWQM processes is presented in Fig. 3. Model initial conditions are set up using the observation data from the previous day of the first simulation date. Initial ensemble members are generated by adding random errors to the initial condition (Equation (5)). The initial covariance $P_i$ is determined by averaging measurement variances at four measurement depths in the soil profile

$$([V_{5cm} + V_{20cm} + V_{40cm} + V_{60cm}]/4)$$

from the actual ten minute data on the same day as the initial condition. 100 ensemble members are used for all base simulations, but the effect of different ensemble size is also investigated by changing the number of ensemble members from 50 to 500.

Most of the subroutines in the RZWQM are based on a daily time step except the physical processes, which are based on a sub-hourly time step (Fig. 3). In addition, considering that remotely sensed data from satellites are available on a daily time step or longer, the minimum update interval for this study was taken as one day, even though application of more frequent update interval is possible from the in situ ten minute observations. In order to find the optimum frequency of surface soil moisture assimilation, different update intervals ranging from one day to two weeks were examined in this study with 100 ensemble size.
To evaluate the performance of EnKF compared to other data assimilation techniques, a simple data assimilation scheme, direct insertion method, was also applied in this study. The direct insertion method involves simply the substitution of available observed data for forecasted system state variable. In this study, it was assumed that the top three numerical layers included in the top 5 cm have the same observed soil moisture as the one measured at the 5 cm depth by Hydra Probe.

Among the measured soil profile moisture data from four different depths (5 cm, 20 cm, 40 cm and 60 cm), data for the top 5 cm depth was used for assimilation. This arrangement mimics the process of assimilating remotely sensed soil moisture information which is also available only for the top few centimeters of the soil profile. In assimilating data at the daily time step by using measurements collected at ten minute time intervals, the data collected at 11:50 PM is used to update and reinitialize the model simulation at the end of the each day.

In the EnKF application, it is important to have a good knowledge of model and observation errors to define the error covariance ($w_k$ in Equation (3) and $v_k$ in Equation (4)). In studies that use synthetic data, these errors are pre-defined, but in experiments involving actual measurements, these errors are difficult to estimate. In this study, observation errors are mainly from instrument error and error from using point measured soil moisture and ignoring field-variability of soil water characteristics. Model errors result from the uncertainties in model physics, soil characteristics, and forcing variables (precipitation and other meteorological data). Because these errors were difficult to estimate in this study, appropriate standard deviation of observation and model errors which minimize errors were determined by trial and error. Standard deviations of model errors of 9.58E-3 and 7.20E-3 m$^3$m$^{-3}$ were used for AS1 and AS2, respectively, and an equal standard deviation of observation errors of 7.07E-3 m$^3$m$^{-3}$ was used for both sites.

Forecasted soil moisture at each of the four measurement depths was compared with the measured data for evaluation before the updating step. This evaluation was conducted by using time series graphs and standard statistical measures. The correlation coefficient (R) was used to indicate the strength of linear association between the measured and predicted soil moisture.
values. In addition, mean bias error (MBE; Equation (10)) and root mean square error (RMSE; Equation (11)) were used to represent the prediction bias and error, respectively.

\[ \text{RMSE} = \sqrt{\frac{\sum (P-O)^2}{n}} \tag{10} \]

\[ \text{MBE} = \frac{\sum (P-O)}{n} \tag{11} \]

Where \( P \) and \( O \) are predicted and observed soil moisture, respectively, and \( n \) is the number of measurements.

4. Results and Discussion

4.1 Assimilation results

All modeling results presented in this study are for non calibrated conditions using a combination of measured, default and model estimated soil parameters or properties. We chose to perform non calibrated simulations to better determine the fundamental affects of data assimilation on the physically-based model performance of soil moisture dynamics in the 0-60 cm root zone. Non calibrated RZWQM simulations were conducted at field sites AS1 and AS2 under the following three conditions: (i) simulation without data assimilation (open loop); (ii) simulation with direction insertion data assimilation method (DIR); and (iii) simulation with EnKF data assimilation (EnKF). As mentioned in section 3.3, the soil moisture at 5 cm was used for assimilation, with model predictions compared to measured data at 5, 20, 40 and 60 cm depths. Simulation of RZWQM was conducted from April 1 (Julian day 91) to October 31 (Julian day 304) 2007 to include the effect of plant water uptake and to avoid periods of frozen soil water. Results at AS1 (beginning Julian day 125) and AS2 (beginning Julian day 115) are presented in Fig. 4 and Table 2. For the EnKF simulation, 100 ensemble members were created and the results shown in Fig. 4 and Table 2 are the average of ten10 simulations.

Application of the EnKF improved the open loop results more than the DIR, especially by increasing correlation coefficients to more than 0.90 in the top soil layer (5cm) for both AS1 and
AS2, and reducing prediction errors by 38 and 59% for AS1 and AS2, respectively, compared to open loop (Table 2). The AS2 simulation had relatively poor results from the open loop, and data assimilation using both direct insertion and EnKF produced better results (increased correlation coefficients and reduced errors) for the all layers. However, results from the AS1 simulation were more complicated and will be discussed in greater detail below. The open loop simulation did not capture the dynamic variation of soil moisture conditions in the top soil layer especially during the dry-down period. In most cases, the open loop simulation overestimated soil moisture in the top layer while surface soil moisture data assimilation improved simulated soil moisture dynamics and soil moisture estimates.

The consistent discrepancies between model forecasting and observations shown in Fig. 4 can be found in other previous studies (Walker et al., 2001a; Reichle and Koster, 2004; Reichle et al., 2004). Walker et al. (2001a) concluded that removal of the systematic bias in the model forecasts or observation is essential for correct retrieval of profile soil moisture using the Kalman filter assimilation scheme. In regard to remotely sensed soil moisture data, cumulative distribution function (CDF) matching has been used to reduce the systematic differences between satellite-based and model-based soil moisture estimates effectively (Reichle and Koster, 2004; Drusch et al., 2005). The need for the correction or rescaling of the discrepancies (bias), before assimilating soil moisture observations into model are confirmed in this study as well. Removal of the systematic forecast bias is expected to improve assimilation results especially during the dry-down period.

The observed soil moisture data at the AS1 site (Fig. 2) shows consistently higher values of soil moisture being measured at the 40 cm depth compared to those at 60 cm, which is not typical of most profile soil moisture distributions. This is because the AS1 sensor site has a dense till soil layer derived from unsorted glacial material at approximately 40 cm which makes the lateral flow more dominant than the vertical flow. As a result, the measured soil moisture at the 40 cm depth at AS1 is always higher than the soil moisture at 60 cm depth. Based on soil sampling data, this appears to be a local phenomenon near our sensors and not representative of the entire field. Thus, the model is unable to capture the unique phenomenon at AS1 (predicting higher water content at the 40 cm depth than at 60 cm) and leads to underestimation in soil moisture at 40 cm.
depth (high negative MBE in the Table 2). The EnKF reduces the open loop overestimated soil moisture at 5 and 20 cm and is closer to the measured values, but this change affects the soil moisture profile at 40 and 60 cm. As a result, the initially underestimated soil moisture at 40 cm and 60 cm was further reduced to give a lower correlation coefficient and higher error (Table 2). This local phenomenon, due to the highly heterogeneous soil characteristics, is ignored in large-scale data assimilation studies. However, this finding for the small field scale study points to issues that should be considered for future soil moisture data assimilation studies.

As expected from the previous studies (Walker et al., 2001b; Zhang et al., 2006), the EnKF is shown to be superior to the DIR except for the results of deeper layers at AS1. Walker et al. (2001b) illustrated the difference between the DIR and the EnKF graphically and provides a concise description of both techniques. They describe two limitations of the DIR. First, with the DIR, surface information from observations can be transferred into the deeper layers only through physical infiltration and exfiltration processes while the Kalman Filter modifies the entire soil profile using the covariances of both the surface measurements and model profile prediction. Second, DIR changes the soil moisture profile according to the difference between the measured and simulated soil moisture. The DIR results in this study revealed that the updated (reinitialized) simulated soil moisture, using the surface observations at the end of day, returns to soil moisture values very near those of the open loop after several sub-hourly simulation loops. However, the results of the EnKF are closer to measured values than the DIR (Fig. 4) resulting in better statistical results. The above-mentioned characteristics of the DIR approach had less of an effect than the EnKF at the 40 and 60 cm depths at AS1 (Table 2).

This study demonstrates the potential for reliable application of the EnKF for surface soil moisture assimilation. However, in spite of satisfactory model performance in this study with application of the EnKF, some precautions should be considered. One of the main reasons for the uncertainty with the EnKF appear to be the presence of bias in the model forecast and violation of the basic assumption of the EnKF (zero mean noise) because of nonlinear processes imbedded in the model and the bounded nature of soil moisture (porosity and residual water content of soil). To overcome these limitations of the EnKF with the real observations, it is desirable to adopt an appropriate bias-correction algorithm for the EnKF (De Lannoy et al., 2007a; De
Lannoy et al., 2007b; Ryu et al., 2009). Second, randomly generated ensemble numbers may interfere with the numerical algorithm to solve the Richards’ equation in the RZWQM and lead to an infinite routine. Therefore, cautious assignment of variances for the EnKF and proper restriction of random numbers are necessary.

There appears to be inconsistencies between the field-measured soil moisture and laboratory measured soil hydraulic properties as observed in Fig. 4. For both sites, field-measured soil moisture at the 5 cm depth during the dry period (less than 0.1) were lower than the laboratory-measured or model estimated water content at -1500 kPa, considered the wilting point in Table 2. The field measured soil moisture content is indicative of high evaporative demand during dry-down period. The RZWQM, however, does not allow the simulated soil moisture to fall below the wilting point when using laboratory measured values as input. Even though the model-estimated inputs for wilting point were used, the simulated soil moisture during the dry period did not go below 0.146 for AS1, and 0.153 for AS2 (Table 1). This implies that calibration and optimization of these values should be conducted for future model runs taking into consideration that laboratory measured hydraulic properties are determined based on point-scale soil cores that do not always characterize actual field conditions.

4.2 Effect of update interval

The effect of update interval was investigated for DIR and EnKF by varying the update interval from 1 to 14 days. AS1 simulation in Fig. 5 showed that for the 5 cm and 20 cm depths, less frequent update interval longer than four days, did not produce significant benefit from the assimilation scheme even though assimilated simulation results are slightly better than ones of the open loop (0.816 and 0.785, R values for 5 cm and 20 cm, respectively). Interestingly, deep layers (40 and 60 cm depths) have worse results with the frequent update (daily and every two day updates). This seems to be because of the large forecasting errors as Walker et al. (2001a) showed that the accuracy of the forecasting model is more important than the temporal resolution of observed data.

On the effect of update interval with AS2 data, even less frequent update interval improved model predictions compared to open loop simulations (Fig. 6). For instance, an update interval of
two weeks for the 5cm depth raised the correlation coefficient to 0.754 with DIR and 0.784 with
EnKF compared to 0.690 with the open loop. In addition, the effect of update interval is more
obvious with AS2 simulations. As the update interval increases, errors increased and correlation
coefficient decreased apparently except that the results of 60 cm depth does not vary much with
different update interval. This is due, in part, to relatively small changes in soil water content at
this depth.

Two previous studies (Walker et al., 2001b; Zhang et al., 2006) investigated the effect of update
interval on soil moisture profile retrieval in a desktop study using a one dimensional soil
moisture equation and synthetic data. In the work of Walker et al. (2001b), full retrieval of
profile soil moisture profile (1 m) was achieved approximately after 12 days with the DIR and 12
hours with the EnKF for the observation depth of 4 cm when the system was updated once every
day. Their conclusion, in regards to observation interval, was that the frequent observations are
more important than observation depth using the EnKF. Zhang et al. (2006) also showed similar
results for the update interval impact with a required time for full retrieval being 16 hours with
the EnKF and 12 days for the DIR when updated once every hour. With a more realistic “daily”
update interval, they found that the full profile retrieval took 15 days with the EnKF while the
DIR failed to retrieve the full profile. The studies mentioned above were based on synthetic
experimental data, whereas our current study explores the effect of update interval using actual
soil moisture measurements and different governing equations in the RZWQM to find the
optimal and effective update interval.

Contrary to the previous synthetic studies (Walker et al. 2001; Zhang et al. 2006), data
assimilation in this study was due to the number and types of physically-based processes utilized
in the RZWQM. In general, we found that more frequent updates contribute to improved model
predictions and that the EnKF was superior to the DIR method of data assimilation.

4.3 Effect of ensemble size

As mentioned in the section 3.2, the EnKF is based on the concept of Monte Carlo methods and
calculates error statistics from an ensemble of system states at the time of update. Therefore, it
requires a sufficient size of ensemble numbers to obtain satisfactory estimates. However for
practical and realistic applications of the EnKF, it is useful to know an appropriate size of ensemble.

In this study, different numbers of ensemble from 50 to 500 were applied to determine how the assimilation results were affected by ensemble size. With a relatively smaller ensemble size of 50, compared to the default value of 100, model performance improved with data assimilation compared to the open loop results; the correlation coefficient increased by 12% and 13% and RMSE was reduced by 38% and 13% for 5 cm and 20 cm depth, respectively for AS1 (Fig. 7). For the AS2 simulation which initially had relatively poor prediction from the open loop simulation, an even smaller ensemble sizes (e.g., 50) produced better results. The improvement with the EnKF at AS2 with 50 ensemble numbers is more apparent than with the DIR and open loop; RMSE were reduced by 59% from the open loop results and 24% from the DIR results for the 5 cm depth (not shown).

The sensitivity of EnKF to ensemble size can be seen in Fig. 7. For AS1 data, the EnKF showed little difference in the efficiency due to ensemble size for upper layers. In the case of deeper layers, smaller ensemble size (50) produced poorer results compared to larger ensemble sizes. Simulations using AS2 data did not show significant differences in predictions between different ensemble sizes (not shown).

Earlier studies for surface soil moisture assimilation with the EnKF (Reichle et al., 2002a; Reichle et al., 2002b) show that the EnKF estimates converge to the true states with different numbers of ensemble based on synthetic experimental results using a land surface model. Reichle et al. (2002b) presented that an ensemble size greater than ten outperforms the extended Kalman Filter. The work of Reichle et al. (2002a) showed that larger ensemble size improved results even though only 30 numbers of ensemble reduced actual errors by 55% compared to the open-loop simulation. In addition, they concluded that for robust estimation of error covariance, at least 500 ensemble numbers were required. Recently, Zhang, Li et al. (2006) suggested that 40 ensemble members were enough to represent the error covariance for their one dimensional model.

5. Conclusion
In this study, field measured surface soil moisture was assimilated into a one dimensional physically based model, RZWQM, using two different data assimilation techniques, the direct insertion method and the Ensemble Kalman Filter. A key difference of this study from previous studies is the use of real observational data instead of synthetic data for the EnKF assimilation into a field scale water quality model. In addition, this study provides a more practical and operational application of surface soil moisture assimilation using the EnKF by investigating the effect of update intervals and ensemble size. This is a significant aspect in terms of the eventual use of satellite soil moisture products and frequency of observations.

Overall, the results of this study indicate that daily assimilation of surface soil moisture resulted in more accurate soil moisture estimation in the upper more dynamic layers (5 and 20 cm). However, improvement of soil moisture prediction in deeper layers (40 and 60 cm) is less certain. Since the observed data are clustered on the top soil layers, there are limitations for the corrected information from the observed surface soil moisture to propagate to the deep layers. Nevertheless, unsuccessful predictions from the open loop for AS1 and AS2 were greatly improved for all layers with both the DIR and the EnKF, except for deeper layers in AS1.

Unique characteristics of soil profile properties and soil moisture distribution at AS1 caused inconsistent assimilation results for the deep layers. That is, the data assimilation techniques, especially the EnKF gave somewhat inadequate results at the 40 cm and 60 cm depths. Therefore, for more successful application of the EnKF, it is recommended to adjust or calibrate the model parameters before data assimilation so that the model prediction can simulate the unique field characteristics properly.

Even though soil hydraulic properties measured in the laboratory and estimated by the RZWQM model were used for the simulation, the open loop results of AS2 highly overestimated the measured soil moisture, thus failing to capture dynamic variations in the soil profile. Considering uncertainties in the laboratory analysis and the high spatial variability of field soil characteristics, it is recommended that the model parameters be optimized through calibration. In this study, however, the model was not calibrated because the purpose of this experiment was to investigate the fundamental effect of different data assimilation techniques on model output compared to the
open loop simulation. However, it is also found a need for a proper rescaling step to remove systematic errors between model forecasts and observed data which exist even after calibration.

The investigation into the effect of update interval in the range of 1 to 14 days found that shorter update intervals improve open loop simulation results better than the long update intervals. This linear relationship between the update interval and performance of the data assimilation was more apparent in the upper layers with the AS2 simulation. Results of AS1 do not show significant benefit from the frequent update because of its unique soil water distribution in the deep layers.

Theoretically, the EnKF requires sufficiently large ensemble numbers to obtain satisfactory results. In this study, however, the test of the effect of the ensemble size showed that larger ensemble sizes did not produce any significant improvement in the results, and an ensemble of 100 members produced results that were comparable to results from larger ensembles. Especially, the top layer at 5 cm depth did not show any sensitivity to the different ensemble numbers. For the other deep layers, smaller ensemble size of 50 deteriorated the statistical results slightly but still produced better results than the open loop and the DIR for AS2.

Although data assimilation was shown to contribute to better estimation of profile soil moisture, there are certain additional points to consider in future studies. First, data assimilation breaks up the water balance of the system. Contrary to the above-mentioned synthetic experiments, the RZWQM includes very sophisticated physical processes; varying boundary conditions due to surface interaction with atmosphere and fluctuating water table depth, and plant water uptake and tile drainage as sink terms. Artificial interruption from the surface soil moisture assimilation affects all these processes and further work is necessary to determine to what extent data assimilation impacts them. Therefore, future investigations involving data assimilation will include the assessment of other hydrologic variables such as ET, runoff, as well as water quality parameters and crop yield.
References


List of Figures.

Figure 1. Study Area- AS1 and AS2 sites.

Figure 2. Observed rainfall and soil water content at 5, 20, 40 and 60 cm depth (Observation period: May 5 to October 31 for AS1 and April 25 to October 31 for AS2 in 2007).

Figure 3. Framework of RZWQM and EnKF Scheme.

Figure 4. Comparison of observed and assimilated soil moisture at 5 (a, b) and 20 cm.

Figure 5. Effects of update intervals with DIR (a and b) and EnKF (c and d) for AS1.

Figure 6. Effects of update intervals with DIR (a and b) and EnKF (c and d) for AS2.

Figure 7. Effects of ensemble size on the correlation coefficient and RMSE for AS1.

List of Tables.

Table 1. Soil physical and hydraulic properties.

Table 2. Statistical data analysis (simulation period: April October, 2007).
Figure 1. Study area in Upper Cedar Creek Watershed and AS1 and AS2 field sites.
Figure 2. Observed rainfall and soil water content at 5, 20, 40 and 60 cm depth at (a) field AS1 and (b) field AS2. (Observation period: May 5 to October 31 for AS1 and April 25 to October 31 for AS2 in 2007).
Figure 3. Framework of RZWQM and EnKF Scheme.
Figure 4. Comparison of observed and assimilated soil moisture at 5 (a and c) and 20 cm (b and d).
Figure 5. Effects of update intervals with DIR (a and b) and EnKF (c and d) for AS1.
Figure 6. Effects of update intervals with DIR (a and b) and EnKF (c and d) for AS2.
Figure 7. Effects of ensemble size on the (a) correlation coefficient and (b) RMSE for AS1.
Table 1. Soil physical and hydraulic properties

<table>
<thead>
<tr>
<th>Site</th>
<th>Depth (cm)</th>
<th>Soil Texture</th>
<th>Measured in laboratory</th>
<th>Model estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Sand (%)</td>
<td>Silt (%)</td>
<td>Clay (%)</td>
</tr>
<tr>
<td>AS1</td>
<td>0-5</td>
<td>L</td>
<td>51.3</td>
<td>31.6</td>
</tr>
<tr>
<td></td>
<td>5-15</td>
<td>L</td>
<td>42.3</td>
<td>37.0</td>
</tr>
<tr>
<td></td>
<td>15-30</td>
<td>L</td>
<td>37.6</td>
<td>36.3</td>
</tr>
<tr>
<td></td>
<td>30-45</td>
<td>SCL</td>
<td>45.7</td>
<td>23.4</td>
</tr>
<tr>
<td></td>
<td>45-60</td>
<td>CL</td>
<td>36.4</td>
<td>27.1</td>
</tr>
<tr>
<td>AS2</td>
<td>0-5</td>
<td>L</td>
<td>30.8</td>
<td>45.4</td>
</tr>
<tr>
<td></td>
<td>5-15</td>
<td>L</td>
<td>27.7</td>
<td>46.0</td>
</tr>
<tr>
<td></td>
<td>15-30</td>
<td>CL</td>
<td>25.4</td>
<td>33.4</td>
</tr>
<tr>
<td></td>
<td>30-45</td>
<td>CL</td>
<td>26.3</td>
<td>38.0</td>
</tr>
<tr>
<td></td>
<td>45-60</td>
<td>CL</td>
<td>25.1</td>
<td>39.4</td>
</tr>
</tbody>
</table>

a) Acronym for soil texture: SCL=sandy clay loam, CL= clay loam, L= loam, C= clay
b) FC: field capacity, water content at -33 kPa
c) WP: wilting point, water content at -1500 kPa
d) Ks: Saturated hydraulic conductivity
<table>
<thead>
<tr>
<th>Site</th>
<th>Depth (cm)</th>
<th>Open Loop</th>
<th>Direct Insertion</th>
<th>EnKF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>R</td>
<td>RMSE</td>
<td>MBE</td>
</tr>
<tr>
<td>AS 1</td>
<td>5</td>
<td>0.816</td>
<td>0.055</td>
<td>0.038</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.785</td>
<td>0.035</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>0.836</td>
<td>0.086</td>
<td>-0.084</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>0.834</td>
<td>0.027</td>
<td>-0.018</td>
</tr>
<tr>
<td>AS 2</td>
<td>5</td>
<td>0.690</td>
<td>0.091</td>
<td>0.078</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.320</td>
<td>0.050</td>
<td>0.024</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>0.109</td>
<td>0.073</td>
<td>0.056</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>0.412</td>
<td>0.033</td>
<td>0.022</td>
</tr>
</tbody>
</table>