Using similarity of soil texture and hydro-climate to enhance soil moisture estimation

E. J. Coopersmith,[1] B. S. Minsker,[1] and M. Sivapalan[1,2]

[1] *Department of Civil & Environmental Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA*

[2] *Department of Geography and Geographic Information Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA*

Abstract

Estimating soil moisture typically involves calibrating models to sparse networks of *in situ* sensors, which introduces considerable error in locations where sensors are not available. We address this issue by calibrating parameters of a parsimonious soil moisture model, which requires only antecedent precipitation information, at gauged locations and then extrapolating these values to ungauged locations via a hydro-climatic classification system. Fifteen sites within the soil climate analysis network (SCAN) containing multi-year time series data for precipitation and soil moisture are used to calibrate the model. By calibrating at one of these fifteen sites and validating at another, we observe that the best results are obtained where calibration and validation occur within the same hydro-climatic class. Additionally, soil texture data are tested for their importance in improving predictions between calibration and validation sites. Results have the largest errors when calibration/validation pairs differ hydro-climatically and edaphically, improve when one of these two characteristics are aligned, and are strongest when the calibration and validation sites are hydro-climatically and edaphically similar. These findings indicate considerable promise for improving soil moisture estimation in ungauged locations by considering these similarities.

Keywords: soil moisture, hydro-climatic, edaphic, similarity, prediction, operational models

1. Introduction

Soil moisture estimates are needed routinely for many practical applications, such as irrigation scheduling and operation of farm machinery. They are typically produced either through remote sensing or sparse networks of *in situ* sensors. Although recent remote sensing studies have confirmed that such measurements approximate *in situ* sensor networks (Jackson et al, 2012), satellite-based sensors provide measurements at a spatial resolution of several kilometers – too large for daily agricultural decision making. On the other hand, *in situ* sensor networks produce values that are difficult to generalize to locations with no proximal sensors. Under these circumstances, dynamic soil moisture evolution models are typically used for soil moisture estimation at the desired location, using information from the nearest available sensors. This method of soil moisture estimation immediately raises the issue regarding the type of model that is most appropriate for such an application. One could think of several different types of models that may be suitable.

The first group of soil moisture models considers only the variability of precipitation, as it has been shown that precipitation variability is the primary mechanism for wetting/drying (Entekhabi and Rodriguez-Iturbe, 1994). Many subsequent models employed an "antecedent precipitation index" (API), defining a pre-established temporal window for antecedent rainfall. This index is then used to estimate current levels of soil moisture (Saxton and Lenz, 1967) and has been implemented with recession modeling for soil water in agriculture (Choudhury and Blanchard, 1983) and also in weather prediction (Wetzel and Chang, 1988). Other precipitation-focused approaches utilize stochastic models to estimate the distributions of soil moisture values using an initialization of daily rainfall (Farago, 1985). Both the stochastic and API approaches require some initial condition for soil moisture at the forecast location – requiring either professional judgment or a sensor. While these issues can be addressed using a soil water balance model, this type of model must be recalibrated frequently, which most soil moisture models are not, as its errors are cumulative (Jones, 2004).

The second group of models adopts a process-based approach, estimating soil moisture from surface radiation and precipitation (Capehart and Carlson, 1994). These process-based models are typically forced by evapotranspiration demand and precipitation at their upper boundary and, if applicable, by groundwater at their lower boundary. More sophisticated models of this type, such as HYDRUS (Simunek et al, 1998), attempt to improve predictions via detailed knowledge of hydraulic soil parameters, information regarding root structures, soil temperature readings, ionic chemistry, CO_2 concentrations, solute transport data, and detailed atmospheric/meteorological information, which are not widely available, especially for routine applications envisaged here.

The third group of models are agriculturally-focused, building model projections outward from existing instrumentation and additional measurements. Gamache et al (2009) developed a soil drying model for which cone penetrometers and soil moisture sensors are required. At most remote sites, these data sources are not currently accessible. Another similar approach employs specific soil type information (theoretically, publicly available data), but ultimately requires proximal sensors to provide the needed soil moisture estimates (Chico-Santamaria, et al, 2009).

Pan et al (2003) and Pan 2012 addressed many of the shortcomings of the existing modeling approaches reviewed above by developing what they called a "diagnostic soil moisture equation" (i.e., model) in the form of a partial differential equation representing the lumped water balance of a vertical soil column, and representing the soil moisture at any moment in time as a function of the sum of a temporally decaying sequence of observed past rainfall events. The model has the advantage that initial soil moisture conditions are not required (only antecedent precipitation data), nor must the model be recalibrated periodically. However, this approach does require a soil moisture sensor at the relevant location for initial calibration of the model's parameters. This method has the disadvantage that the presence of soil heterogeneity could necessitate a large number of sensors to account for the spatial variation of soil moisture (Pan and Peters-Lidard, 2008). Furthermore, decision support often requires estimation at locations lacking sensors.

The aim of this paper is to present and test an approach that can help overcome the issues of calibration at ungauged locations associated with the Pan et al. soil moisture estimation model. The proposed solution involves calibrating the Pan (2012) diagnostic soil moisture equation (model) at gauged sites and then extrapolating the calibrated model to ungauged sites by invoking similarity. Similarity here is defined on the basis of hydro-climatic characteristics, using a classification system developed by Coopersmith et al (2012), as well as edaphic (soil) properties. The proposed new scheme maintains the advantage of Pan et al.'s parsimonious soil moisture model in that it does not require specification of initial soil moisture condition, and also there is no need to recalibrate periodically. The model's simplicity also permits implementation of the model in a manner that can easily be refit with new parameters, where necessary. Section 2 provides more details on the approach.

To calibrate and validate the model, we use data from the U.S. Department of Agriculture's (USDA) Soil Climate Analysis Network (SCAN). This national array of soil moisture sensors (with co-located precipitation sensors) delivers hourly data at a variety of publically-accessible sites throughout the United States. Fifteen sensor locations with numerous years of high-quality, minimally-interrupted data were selected for further analysis. These sites display considerable hydrologic diversity, which aids in demonstrating that the nationwide application of the proposed soil moisture model using precipitation data represents a feasible goal. With respect to agricultural decision-support, for energy-limited sites, the value of hourly soil moisture estimates is found in the determination of whether or not a field is trafficable – whether heavy equipment will damage fields or become mired. With respect to water-limited sites, the value of soil moisture estimates is found in devising optimal irrigation strategies that utilize limited water resources most efficiently. Of the fifteen SCAN sites examined, the three sites in New Mexico, the site in Colorado, the site in Nebraska, the site in Wyoming, and the two in Iowa are all water-limited (8 in total). The remaining sites (7 in total), located in Pennsylvania (2), Arkansas, Georgia, South Carolina, North Carolina, and Virginia, are all energy-limited. Results of the analysis are given in Section 3, followed by discussion in Section 4 to suggest further improvements and conclusions are presented in Section 5.

2. Methodology

The proposed modeling approach involves four steps, summarized in Figure 1 and described in more detail in the sections below. First, the diagnostic soil moisture model of Pan (2012) is calibrated at locations with ample data. Given that the focus of this study is on soil moisture estimation for agriculture, we only consider prediction during the growing season, which is appropriate given that the model does not address snow melt processes. Second, the predictions at these locations are improved using machine learning techniques for error correction. Third, the classification system proposed by Coopersmith et al. (2012) is used to generalize the parameters calibrated at each location, enabling its application at other sites characterized by the same hydro-climatic class. Fourth, sites are examined for edaphic (soil property) similarity in addition to hydro-climates. The results of these four steps are then examined to identify which approach to regionalization performs best.

Step 1: Calibration Using a Two-Layer Genetic Algorithm

Unlike the original diagnostic soil moisture calibrations, the ultimate objective of this work is to enable agricultural decision support in near real time. To this end, the daily model from Pan (2012) is first modified to yield an hourly model within the same framework. Genetic algorithms are then deployed to calibrate the model, enabling more efficient exploration of the parameter search space than the traditional Monte Carlo search, which was the approach taken by Pan (2012).

Genetic algorithms (GAs), a subset of evolutionary algorithms, were originally developed by Barricelli (1963) and have become increasingly common in environmental and water resources applications, including the calibration of hydrologic model parameters (e.g., Cheng et al, 2006; Singh et al, 2008; Zhang et al, 2009).

In this work, a simple genetic algorithm uses the operations of selection, crossover, and mutation (for reference, see Goldberg 1989) to search for parameters that minimize prediction errors from the diagnostic soil moisture equation (Pan, 2012):

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

Here θ_{est} represents the best estimate of soil moisture during a given hour. θ_{re} denotes residual soil moisture, the minimum quantity of moisture that is present regardless of the length of time without precipitation. ϕ_e, the soil's porosity, signifies the maximum possible soil moisture value, at which point the soil becomes saturated. Finally, c_4 is a parameter related to conductivity and drainage properties, essentially defining the rate at which soil can dry. If c_4 assumes a value of zero, the soil is permanently at its residual soil moisture value, θ_{re} - a soil that dries infinitely rapidly. Conversely, as c_4 becomes large, the soil will permanently assume the value of its porosity, ϕ_e – a soil that dries infinitely slowly. The β term in Equation 1 is calculated in Equation 2 below:

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

Here, P_i denotes the quantity of rainfall during hour i (day in the original presentation in Pan et al.). The soil depth at which an estimation occurs is given by z. This convolution summation has a temporal window of n hours for considering past precipitation. For instance, today's soil moisture is strongly influenced by yesterday's rainfall, influenced to a lesser degree by last week's rainfall, and not influenced at all by rainfall from ten years previous. Given the general limitation of our datasets and the fact that shallow-depth soil moisture is most relevant to decision-support, all of our analyses occur with measurements of two inch (~5cm) depth.

To choose the appropriate value for n, the value of β is calculated at each hour throughout the dataset – setting n to a very large value (2000 hours, denoted by M) initially. Next this "beta series" (where $n = M$) is correlated with a separate beta series, calculated where $n \ll M$. If the correlation coefficient between these two time series approaches unity, then the smaller value of n is selected. Otherwise, n is increased incrementally until the correlation between the $n \ll M$ beta series and the $n = M$ beta series approaches unity.

Finally, the estimated soil water loss at hour i, e.g. due to evapotranspiration or deep drainage, is expressed by the term, η_i. As this algorithm does not presume any more detailed knowledge of potential evaporation/drainage behaviors, this "eta series," representing losses due to evapotranspiration and deep drainage, is modeled as a sinusoid (Pan, 2012) with period 8,760 (the number of hours in a year). The eta (η) series is required to calculate the beta (β) series (Eq. 2), which is required to use the diagnostic soil moisture equation (Eq. 1). Thus, before any other parameters are chosen, a generalized sinusoidal form of η is estimated as given in Equation 3:

$$\eta = \alpha \sin(i - \delta) + \gamma \quad (3)$$

Here, α represents the sinusoid's amplitude, γ denotes the vertical shift, and δ signifies the necessary phase shift. These three parameters are fitted via the genetic algorithm such that the correlation between the beta series (using the eta series implied by α, γ, and δ) and the observed soil moisture series (θ_{obs}) is maximized. Once values for the eta series are established, the remaining three parameters of Equation 1 (θ_{re}, ϕ_e, and c_4) are then fitted by a second application of the genetic algorithm, this time minimizing the sum of squared errors between the estimated soil moisture series (θ_{est}) and the observed values (θ_{obs}).

Step 2: Error Correction Using The k-Nearest Neighbors Machine Learning Algorithm

After the parameters of the diagnostic soil moisture equation (Eq. 1) have been calibrated, the hourly precipitation time series is used to generate a soil moisture time series during the growing season months of interest. Discrepancies between the observed soil moisture values (θ_{obs}) and the estimated values(θ_{est}) are computed as shown in Equation 4:

$$\theta_{obs} = \theta_{est} + \varepsilon \quad (4)$$

where ε represents the error associated with any hour's soil moisture estimate.

To correct biases in these errors, the k-Nearest Neighbor algorithm (Fix and Hodges, 1951) is employed to predict ε using the characteristics from the training data. More specifically, the data are searched for the most similar matches in terms of time of day, day of year, θ_{est}, $\beta(n)$, and $\beta(M) - \beta(n)$. For example, if the model returns a prediction of $\theta_{est} =$ 0.35 at 2:00pm during July when rainfall has been heavy recently but drier over a longer period, KNN will search the training set for other estimates near 0.35 made on mid-summer afternoons where a similar recent rainfall pattern has been observed. Next, the algorithm averages the value of the error, ε, associated with those types of conditions, producing an estimated error, ε_{est}. Each validation estimate is then adjusted to be $\theta_{est} + \varepsilon_{est}$.This technique allows consistent model biases, such as underestimating wetter days and overestimating drier days, to be corrected.

This error correction model also accounts for diurnal soil moisture variations that were not considered in developing the diagnostic soil equation, which was designed to deliver daily soil moisture estimates. Consider a soil moisture estimate at 4pm, after soil has had a full day of sunlight (theoretically) to dry. As the diagnostic soil moisture equation only considers drainage and evapotranspiration losses on a daily basis, θ_{est} will be larger than θ_{obs}. Yet, because this type of mistake presumably occurred frequently throughout the training data, the algorithm will locate other 4pm estimates, each of which will be biased in the same direction, and our final soil moisture estimates will take this bias into account, improving the results as shown subsequently.

To assess the performance of the soil moisture models with and without machine learning, an R^2 value as defined in Eq. 5) is used, as this value represents the proportion of variance in soil moisture explained by the developed model.

$$R^2 = 1 - \frac{SSR}{SST} \qquad (5)$$

where SSR denotes the sum of squared residuals and the SST term signifies the total sum of squares, i.e. the sample's variance.

Step 3: Estimation by Hydro-climatic Similarity

This step tests the hypothesis that the classification system by Coopersmith et al. (2012) can be used to generalize the calibrated parameters for the diagnostic soil moisture equation using hydro-climatic similarity. If two locations are assigned the same hydro-climatic classification, then the calibrated parameters from one SCAN sensor within that class will be assumed to perform well at another.

This hypothesis was tested at fifteen SCAN sensors for which soil moisture and precipitation data are available hourly for a period of several years. These sensors are located in diverse geographic locations and hydro-climatic classes in Iowa, North Carolina, Pennsylvania, New Mexico, Arkansas, Georgia, Virginia, South Carolina, Nebraska, Colorado, and Wyoming. The data at each of these locations were divided into training/validation sets and parameters were calibrated using training data only. Next, these parameters were employed on the validation sets at the locations for which they were calibrated. The subsequent R^2 values (proportion of variance in soil moisture explained by the machine-learning-enhanced diagnostic soil moisture equation, see Steel and Torrie, 1960, for reference) defined a baseline level of performance for that site.

The process of cross-validation is detailed below:

1. Consider two sites, x and y, chosen from the fifteen available calibrated locations.
2. Estimate the soil moisture values in the validation dataset of site y, using the parameters calibrated from the training dataset at site x.
3. Record the difference between the R^2 baseline value at site y (obtained using parameters calibrated at site y) and the performance obtained at site y using parameters calibrated at site x.
4. Repeat steps 1-3 for all 210 possible (x, y) pairs where $x \neq y$

   Note: (x, y) and (y, x) are not equivalent. One signifies the performance of parameters calibrated at site x making predictions at site y, the other signifies the performance of parameters calibrated at site y making predictions at site x.

At this point, three types of (x, y) pairs emerge. If the hypothesis is correct, then the first type, when x and y fall within the same hydro-climatic class, should display limited losses in predictive power. The second type, when x and y fall within a "similar" hydro-climatic class (two classes differing by a single division of the classification tree developed in Coopersmith et al., 2012) should display greater losses of predictive power. Finally, the third type, when x and y fall in two unrelated classes, should display the largest loss of predictive power.

Step 4: Estimation by Hydro-climatic and Edaphic Similarity

The final step extends the hypothesis proposed in Step 3 by evaluating the impacts of soil texture and type on soil moisture predictive power. The fifteen sites from the SCAN network are examined based upon the soil textural information available from the Pedon soil reports that SCAN provides, as well as data from NRCS's soil survey database[1].

This information allows sites already deemed hydro-climatically similar to be further sub-divided into sites that are and are not edaphically similar. Analogous to the previous section, we consider pairs of sites, x and y, where parameters are calibrated at site x and validated at site y. In this case, four groups can be defined – the first, where x and y and hydroclimatically similar, the second, where x and y are hydroclimatically similar, but differ edaphically, the third, where x and y are edaphically similar, but differ hydroclimatically, and finally, where x and y are hydro-climatically and edaphically dissimilar.

3. Results

This section begins by presenting the results of the machine learning approach used in error correction during the initial calibration step (Section 3.1). Next, Section 3.2 presents results for the hydro-climatic similarity analysis, illustrating the performance of calibration/validation pairs within the same class and without. Finally, Section 3.3 shows how the predictive power improves when both hydro-climatic and edaphic similarity are considered.

3.1 Testing the Value of Machine Learning Error Correction for Soil Moisture Prediction Using the Diagnostic Soil Moisture Equation

Figure 2 shows the performance of the calibrated parameters for the 15 SCAN sites using only the diagnostic soil moisture equation (Step 1 of the methodology) along with the subsequent improvement in performance following machine learning error correction (Step 2). In each case, the six parameters required for the implementation of the diagnostic soil moisture equation are calibrated using training data from before 2010. Sensors with hourly precipitation and soil

[1] http://websoilsurvey.sc.egov.usda.gov/App/WebSoilSurvey.aspx

moisture time series data between 2004 and 2009 (inclusive) provide four to six years of training data (some sites are missing one or two years of data). Only days of the year where snow cover is unlikely are used to train the algorithm (from the 100th to 300th day of the year in all locations, for consistency). Validation data consist of days 100-300 for 2010 and 2011.

The results illustrate that in all fifteen test cases, performance within the validation sample is improved by machine learning modeling of residuals from the training set, in some cases, as much as 26.9% of the unexplained variance (site 2091) in soil moisture is corrected from by this technique. The average results (far right column, Fig. 2) illustrate that the diagnostic soil moisture equation explains just 69.2% of the variance in soil moisture ($\rho = 0.83$) before machine learning corrections occur, but explains 77.5% of the variance in soil moisture ($\rho = 0.88$) thereafter.

To explore these findings in more detail, three of the 15 SCAN sites, chosen to represent different hydro-climatic locations – New Mexico (#2015, hydroclimate IAQ/southwestern desert, Loamy Sand), Iowa (#2068, hydroclimate ISCJ/northern midwest plains, Silty Clay Loam), and Georgia (#2013, hydroclimate LWC/southeastern forest, Sandy Loam) are examined to illustrate how improvements from adding machine learning error models to the diagnostic soil moisture equation differ across sites. These three sites represent three distinct hydro-climatic classes, with significant differences in soil texture, seasonality of precipitation, aridity, timing of maximum precipitation, and timing of maximum runoff. Using error correction models for prediction at these sites increased R2-values by an average of 8.2%, which is similar to the 8.3% improvement in R2 averaged across all fifteen sites. Thus, these three locations are representative in terms of both hydro-climatic and edaphic diversity and their responsiveness to machine learning.

The base soil moisture model results from applying Step 1 at the three sites are displayed in Figures 3-5. These predictions are shown with the results produced by deploying the machine learning algorithm (KNN) in Step 2 to remove bias and correct errors. In each image, the blue line represents the observed soil moisture readings, the red line represents the estimates generated by the diagnostic soil moisture equation, and the green line represents those predictions after the machine learning algorithm has removed biases and corrected errors. Soil moisture values (y-axis) are presented as volumetric percentage (0-100).

In Figure 3, the diagnostic soil moisture equation is able to trace the general trend of the soil moisture time series (ρ = 0.860). However, during the middle of the time series, in which the observed soil moisture values fall below 5%, the benefits of machine learning error correction are most noteworthy. There are other hours scattered throughout the dataset where the green line (ML prediction) follows the blue line (observed values) much more closely than the red line (diagnostic soil moisture equation). The green line (ρ = 0.917) not only improves upon the correlation value of Pearson's Rho (the square root of the R^2 value in Eq. 5), but also displays marked improvement for those cases in which the diagnostic soil moisture equation produces significant errors.

During the validation period, specifically 2010, wetter conditions were observed than were present during calibration. At this SCAN site, before 2010, the average soil moisture value observed was 28.55%, with only 25% of values exceeding 35% volumetric soil moisture. However, in 2010, the average soil moisture value measured was 33.16% with 45% of values exceeding 35%. The machine learning driven error correction improves the diagnostic soil moisture equation (ρ = 0.846) significantly (ρ = 0.915), but fails to raise its forecasts to reach some of the wetter conditions experienced in validation. Underestimations of this nature, although detrimental in terms of numerical errors, are not necessarily a problem for decision support of agricultural or construction activities, for example. If a model warns that a site is very wet and in reality, it is even wetter than predicted, the user has still been given adequate warning not to attempt activity at that site. It is important to note that small errors are more significant in terms of decision support (specifically when and where to irrigate) during dry conditions. Generally, the model's errors are smaller, in absolute terms, during drier conditions. This analysis's approach to error correction, as it relies on previous errors to predict future errors, will not address long-term trends within the soil moisture record.

In Figure 5, a soil moisture series from Georgia is modeled by the diagnostic soil moisture equation. Even before adding any error correction, the equation performs well (ρ = 0.936) and the machine learning approach yields a smaller improvement (ρ = 0.941). It is worth noting that machine learning does not damage an already excellent performance, offering slight improvements when possible and essentially no correction when training data suggest the model has already performed adequately.

Table 1 presents all fifteen sites for which the diagnostic soil moisture equation has been calibrated, including information regarding their hydroclimatic class from Coopersmith et al (2012), their soil textural characteristics, and their performance before and after the KNN bias correction process.

3.2 Bias Correction – More Detailed Results

In addition to generalizing the parameters calibrated in the diagnostic soil moisture equation, the error correction approach allows for systematic biases to be removed by searching training data for similar conditions and then predicting the types of mistakes most likely to occur. Figure 6, by zooming in upon a 30-day period from Figure 2, illustrates how machine learning reduces errors by introducing a diurnal cycle into a model that previously lacked one. The remaining bias is likely explained by a slightly wetter training dataset as compared with the validation data. It is possible that the diurnal cycle at some locations reflects a soil moisture probe's dependency on electromagnetic properties driven by temperature change (apparent permittivity) rather than hydrologic processes (Rosenbaum et al, 2011). However, the model's ability to respond to these nuances would not compromise its performance were these nuances subsequently removed.

Any corrective algorithm will, over thousands of validation points, push the estimate away from the observed value in some cases. However, the results from Table 1 demonstrate that its overall performance represents an improvement at all sites, and thereby justifies its use. Regarding the issue of 'measurement artifacts,' whether the diurnal cycle is genuine or an idiosyncratic sensor output, the model is tasked with calibrating itself and correcting biases as defined by the empirically-reported data. Figure 6 illustrates its ability to do so. Were the sensors to no longer report such a diurnal pattern (i.e. it is merely a measurement artifact, and subsequently corrected), the machine learning step would no longer observe those biases, and consequently, no longer introduce such a pattern. The accuracy of the SCAN network is a relevant inquiry, but unfortunately, not within the scope of this paper.

By addressing such systematic biases, machine learning enables model performance to improve with each successive growing season as the training dataset expands. For instance, although the fields in Iowa endured flooding during the validation period and subsequently made errors, such errors would eventually populate the training data. The next time such flooding occurs, the model is likely to recognize the occurrence of those same conditions and adjust the diagnostic soil moisture equation's predictions accordingly. In this vein, model performance is likely to improve over time, especially with the models already showing reasonable accuracy using only a few years of training data.

Figure 7, 8, and 9 present these results in more detail for each of the three SCAN sites presented in Figures 3, 4, and 5. In each figure, the upper-left image presents the average bias correction (change in % soil moisture) for each hour of the day (0-23). At all three sites, bias corrections display a clear diurnal pattern – that is to say the removal of a diurnal cycle is a substantial role of machine learning under a variety of hydroclimatic and edaphic conditions. The upper-right image of each figure presents the bias correction as a function of the unadjusted soil moisture estimate – essentially, whether there exists a systemic over- or underestimation when values are high or low.

The first two sites (Figures 7 and 8) do not present a clear pattern, but Figure 9 displays a trend suggesting that the highest estimates of soil moisture tend to be overestimates and the lowest estimates of soil moisture tend to be underestimates – but these biases are removed via machine learning. The lower-left image presents bias correct as a function of the day of the year (from 100-300, the days of the year when the model is applied). At all three sites, the seasonal cycle does appear in terms of the patterns of bias correction, but the pattern is noisier than the diurnal cycle. The magnitude of the adjustments are largest in the monsoon-affected desert of New Mexico, a bit smaller in the Midwestern plains characterized by less extreme seasonal behavior, and smallest in the Southeast where seasonal variations are low.

Finally, the lower-right image relates bias correction to the beta series from the diagnostic soil moisture equation (Pan, 2012), a convolution of a decaying precipitation time series working backwards temporally from the current time. Stated differently, these charts relate bias correction to the amount of antecedent precipitation (with more recent precipitation weighted more heavily). In Figure 7 (Plains, Silty Clay Loam), the model tends to underestimate moisture when large quantities of antecedent rainfall are present, where in Figure 9 (Woods, Sandy Loam), once antecedent precipitation becomes non-trivial, displays the opposite pattern. This is consistent with the finer Midwestern soils' proclivity for ponding/flooding due to larger proportions of clay. In these cases, larger amounts of rain will soak soils from above, and capillary rise might further soak sensors from below, leading to underestimation from the diagnostic soil moisture equation and subsequent machine learning correction. By contrast, with sandier soils, drainage occurs easily, leading to higher rates of loss than the eta series (Pan, 2012) would predict (there is more available water to lose), leading to overestimation with large amounts of antecedent rainfall.

3.3 Cross-Validation Results for Hydro-climatic Similarity: Qualitative Findings and Significance Testing

To test the hypothesis that models calibrated in one location can be used in a hydro-climatically similar location, cross-validation was used as described in Step 3 of Section 2. The fifteen SCAN sites yield $15^2 = 225$ possible (x, y) pairs. Fifteen of these 225 pairs occur when $x = y$, establishing the baseline level of performance for a given site (validation performed using the parameters calibrated at that same location). Of the 210 remaining (x, y) pairs, 120 of them consist of paired catchments in which x and y are located in unrelated classes, 60 consist of paired catchments in which x and y are located in a "similar" class (different by a single split within the classification tree), and 30 consist of paired catchments in which x and y fall within the same hydroclimatic class (but x and y do not represent the same catchment). Figure 10 presents box plots illustrating the change in R^2 values for these three sets of pairs in a manner analogous to the differences shown in Figure 2. Table 2 presents the quantitative results, again averaging the deterioration of performance in terms of change in R^2.

These findings show that calibrating the model at one location and applying those parameters elsewhere within the same class (green) is preferable to applying those parameters in a similar, but not identical class (yellow) and vastly superior to applying those parameters in an unrelated class (red). The differences between any two clusters (same-class, similar-class, unrelated class) are all significant at the $\alpha = 0.01$ level ($p < .001$ in all cases) as calculated by a two-sample, heteroscedastic t-test (Welch, 1947).

3.4 Impact of Soils: Cross-validation Results for Edaphic and Hydro-climatic Similarity

To isolate the impacts of soil types (edaphic similarity) on soil moisture prediction, groups of sensor locations among the 15 SCAN sites that are hydro-climatically similar were analyzed, shown in Figure 11. The soil textural data for each of these fifteen sensors are plotted on a soil texture pyramid diagram in Figure 12. These data were obtained from either Pedon Soil Reports available through the SCAN network (which provide precise percentages of clay, silt, and sand), or, where this information was unavailable, from soil information in the national soil Web database[2].

Of the thirteen sensors from the four hydro-climatic classes with multiple SCAN sensors (light green, blue, dark green, and brown in Figures 11 and 12), 30 (x, y) pairs exist where the model can be calibrated at site x and its parameters applied at site y. Note that (x, y) is not equivalent to (y, x) as the sites for calibration and validation are reversed. Of these 30 pairs, 20 pairs are edaphically similar as well. However, 10 of them include a pair of points where the soil types or terrain types are notably misaligned (for example, light green dots in Figure 12 where two of the three sensors are in silty clay loam and the third is in sandy loam– a notably different soil). A similar analysis to the one presented in Figure 10 and Table 2 has been reproduced, comparing the loss in predictive power (R^2) for the 20 pairs with similar hydro-climates and soils against the loss for the 10 pairs in which either the soil texture (Figure 12) or type do not align. The average loss of 1.0% for the 20 very similar pairs is a much smaller decline than the 8.0% average decline observed for the 10 pairs for which soil/terrain information suggests dissimilarity. These results are significant with a p-value of approximately 0.02. Additionally, the upper-most two green dots in Figure 10, where calibrated parameters at one location perform poorly at another of similar hydro-climatic class, fall within these 10 cases.

These observations show the importance of soil information, or edaphic similarity. While pairs of calibration/validation locations with similar hydro-climates, but dissimilar soils, show a decline in performance as compared with pairs of locations where both are similar, so too do locations with similar soils, but dissimilar hydro-climates. The shaded circles in Figure 12 illustrate groups of sensors that are quite similar in terms of soil textures. However, despite their soil similarities, differences in hydro-climates hinder cross-application, showing a decline in performance of 10.9% for all (x, y) pairs within the shaded regions of Figure 12 for which x and y are not from the same hydro-climatic class.

[2] http://websoilsurvey.nrcs.usda.gov/app/WebSoilSurvey.aspx

As summarized in Figure 13, these results suggest that in cases where both soil type and hydro-climate align, very little performance is lost when parameters are re-applied (1.0%), moderate declines in performance are observed when one of these two factors are aligned (8.0% if hydro-climates align and soil types do not; 10.9% if soil types align, but hydro-climates do not), and large declines in performance appear when neither align (20.5%). Clearly both types of attributes are important and should be considered in future modeling work in which the relative importance of hydroclimates and soil textures can be examined in greater detail.

4. Discussion: Future Work to Improve Predictions

This section discusses other approaches that could be used in the future to improve and broaden the applicability of the methods developed in this work. First, we will consider micro-topographic effects on soil moisture, as local peaks and valleys can cause soils to dry more or less rapidly. Second, we will discuss a conceptual omission within the diagnostic soil moisture equation – infiltration excess. Finally, we will discuss the role of future satellite data on soil moisture modeling.

4.1 Estimates Enhanced By Topographic Classification

Ultimately, the combination of a hydro-climatic classification system and the diagnostic soil moisture equation demonstrates a generalization of calibrations, facilitating predictions at any location where a viable sensor exists within a similar hydro-climatic class and soil type. However, the lumped, bucket model is not ideally-suited for landscapes with complex topography. Conveniently, the majority of SCAN sites are placed on relatively flat surfaces. Integration of topographic insights is a fertile area for future research. One possible approach to further improving predictive accuracy is to disaggregate the soil moisture estimates as a function of local topography. While SCAN sites used for soil moisture data are generally located on flat surfaces, predictions may be needed at locations located on ridges or in valleys where the soils are likely to be wetter or drier than their surroundings. This requires the notion of regional topological classification. In this manner, the notion of similarity is extended to include hydro-climatology, soil characteristics, and topographic designation (ridge, slope, valley, etc). Preliminary analyses suggest that small-scale topography does play a meaningful role in the wetting/drying process. Future research with more extensive datasets in locations with more complex topological contours could improve soil moisture predictions by enabling the models developed in this work to be adjusted as a function of local topographic classification.

4.2 An Enhanced Diagnostic Soil Moisture Equation

The diagnostic soil moisture equation could also be improved in future modeling efforts by considering overland and subsurface flows, specifically in areas characterized by more complex topography. Currently, the model assumes that, in the absence of saturation, all rainfall will ultimately infiltrate, as the porosity parameter serves as an upper bound on soil moisture levels. The diagnostic soil moisture equation was designed originally as a daily model, and it is probably rare that on any given day, a significant fraction of precipitation does not infiltrate. However, at the hourly scale it is quite possible that the water from an intense rainfall event will not make its way into the soil at the location of the sensor. To address this lateral transfer phenomenon, additional parameters can be introduced into the diagnostic soil moisture equation that place an upper bound on the quantity of rainfall that can be infiltrated during any hour (or other interval) of the convolution calculation for any particular soil type. Agricultural decision-support includes trafficability when wet and irrigation support when dry. While overland flow is perhaps an unneeded component in water-limited catchments where irrigation schemes represent the most significant soil-moisture-related decision, in wetter catchments, in which trafficability is a real concern, such an addition could improve the model. While this approach would require the fitting of additional parameters, it is likely that predictions would be improved. These additional parameters could also be considered in assessing cross-site edaphic similarity using the methods described above, although they may be highly correlated with existing parameters such as porosity, residual soil moisture, and drainage.

4.3 NASA's Soil Moisture Active Passive (SMAP) Mission

With NASA satellite data for soil moisture available at the 36 km, 9 km, and 3 km scales throughout the United States, and with the SMAP satellite scheduled to launch during 2014

(O'Neill et al, 2011), the models developed in this work will have ample measurements against which to test and improve their results, and can be used to help check the accuracy of satellite measurements. Future research in LiDAR-driven disaggregation, proposed above, could also be used to improve satellite soil moisture estimates by accounting for smaller-scale topography.

4.4 Water Balance Models and Up-Scaling

The diagnostic soil moisture equation used in this paper (Pan et al, 2003; Pan, 2012) was an appropriate choice due to its ability to generate soil moisture estimates without the need for knowledge of antecedent soil moisture conditions. Koster and Mahanama (2012) and Orth et al. (2013) have developed approaches to estimate soil moisture at the watershed scale by leveraging hydroclimatic variability and long-term streamflow measurements in a water-balance model – also without employing previous soil moisture conditions. If the parameters calibrated and then generalized in this work produce point estimates of soil moisture at a diversity of locations, integration with a water balance approach could help with the up-scaling process.

5. Conclusions

This work has demonstrated the feasibility of estimating soil moisture at locations where soil moisture sensors are unavailable for calibration, provided they fall within hydro-climatically and edaphically similar areas to gauged locations. By calibrating the diagnostic soil moisture equation via a two-part genetic algorithm, improving its performance via a machine learning algorithm for error correction, then validating that algorithm at the same location in subsequent years, a baseline level of predictive performance is established at fifteen locations. Next, these results are cross-validated – deploying parameters calibrated at a given site at sites of similar and different hydro-climatic classes, demonstrating that parameters can be re-applied elsewhere within the same class, but not without. Finally, by incorporating edaphic information, we observe the strongest cross-validation results when hydro-climatic and edaphic characteristics align. As only 24 hydro-climatic classes describe the entire nation (and only 6 describe a significant majority), it is entirely possible that a couple dozen well-placed soil moisture sensors can enable reasonably accurate soil moisture modeling at any location within the continental United States.

It is likely that the types of errors made when parameters are cross-applied between sites of different hydroclimates will differ from the types of errors that appear when the sites differ edaphically. Further research extending beyond model performance into the specific conditions under which models perform less effectively along with the magnitude and bias of those errors would be highly illustrative for future researchers.

This analysis can improve agricultural decision-support by offering insight into locations that can benefit from targeted irrigation in drier conditions, or conversely, by minimizing risks of ruts and damaged equipment when fields are no longer trafficable during wetter conditions. Scaling the results of these models upward can assist with larger-scale assessments of flood risks or as calibration/validation tools for satellite estimates of soil moisture. Scaling these results downward can help maximize yields. Given the ubiquity of precipitation data, which are the only inputs these models require, better understanding of the transferability of modeled parameters is a step towards far wider availability of soil moisture estimates.

Leveraging these findings, the discussion section also presented the results of preliminary analysis that illustrates how further improvements in soil moisture predictions could be gained by disaggregating based on local topography. This would enable more accurate predictions at sites characterized by peaks and valleys that dry faster or slower than the relatively flat locations at which soil moisture algorithms are generally calibrated. Incorporating overland flow into the diagnostic soil moisture equation and integrating satellite data into the approach could also improve predictions in the future.

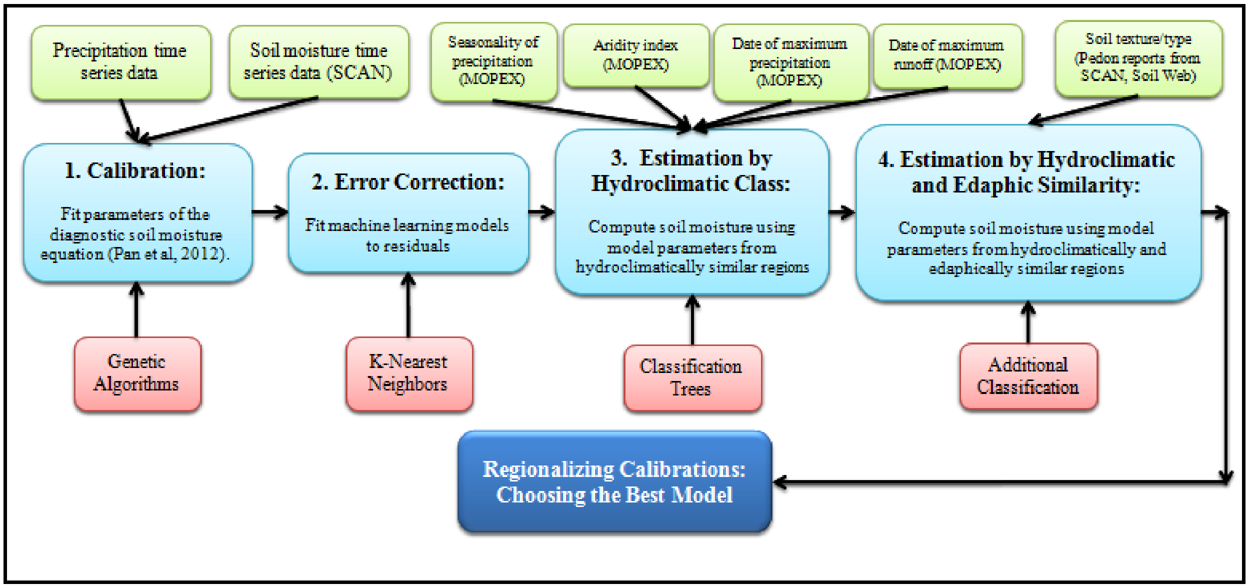

Figure 1, Methodological flow chart

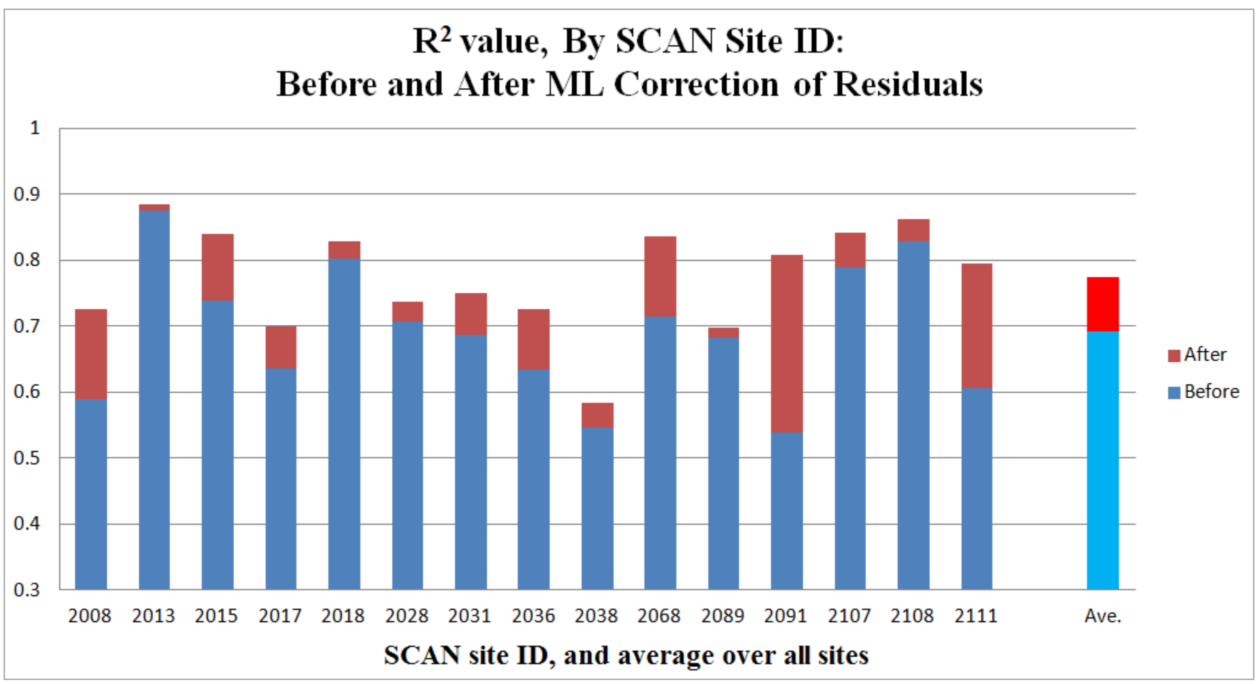

Figure 2, Improvements from machine learning (KNN) models of residuals.

Figure 2, Improvements from machine learning (KNN) models of residuals.

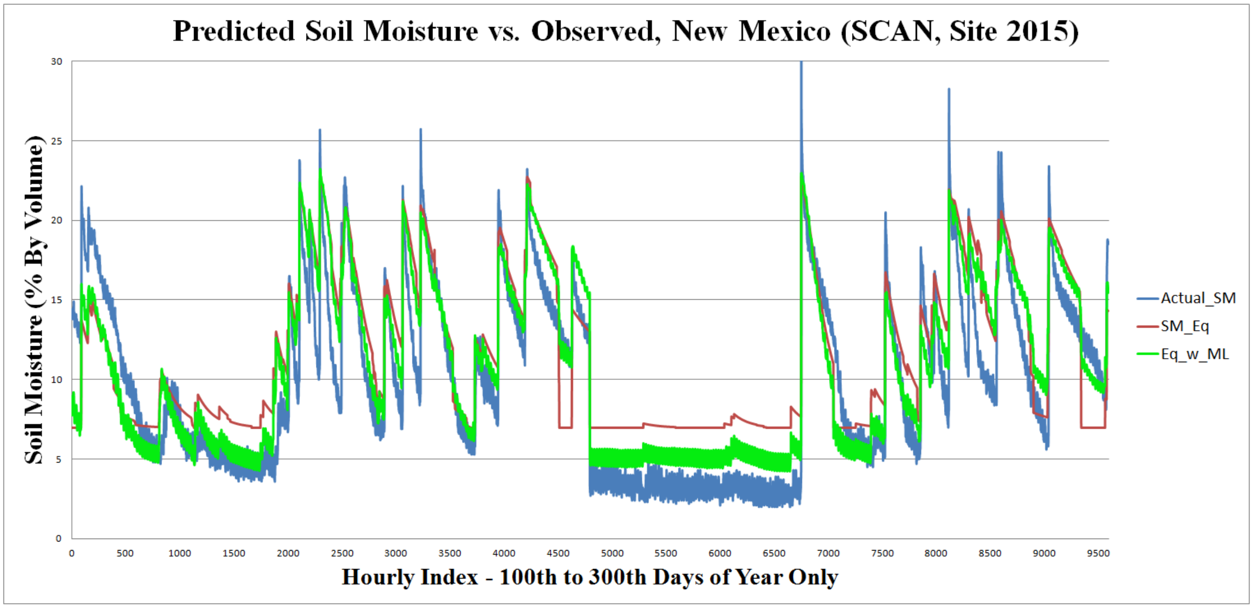

Figure 3, Soil Moisture Time Series, SCAN Site 2015, New Mexico (USA), Actual Soil Moisture (Blue Line), Diagnostic Soil Moisture Equation Estimate (Red Line), and Diagnostic Soil Moisture Equation with Machine Learning Error Correction (Green Line). Hydroclimate: IAQ (Intermediate Seasonality, Arid, Summer Peak Runoff) Soil Texture: Loamy Sand

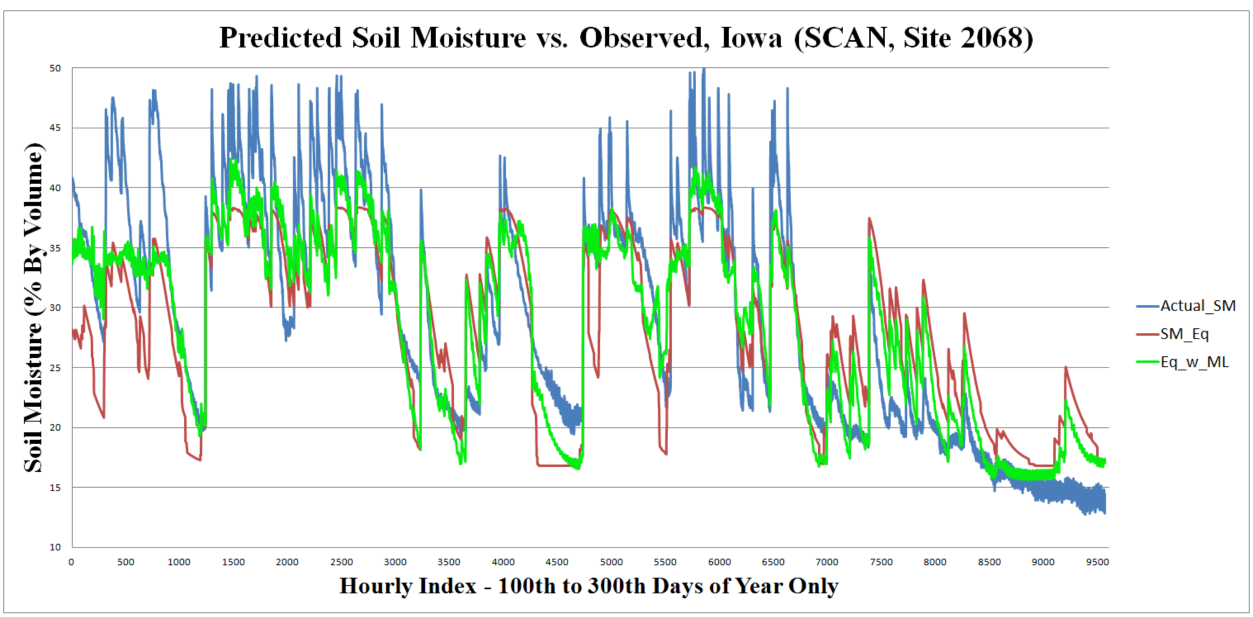

Figure 4, SM Time Series, SCAN Site 2068, Iowa (USA), line colors from Fig. 3 Hydroclimate: ISCJ (Intermediate Seasonality, Semi-Arid, Winter Peak Runoff, Summer Peak Precipitation) Soil Texture: Silty Clay Loam

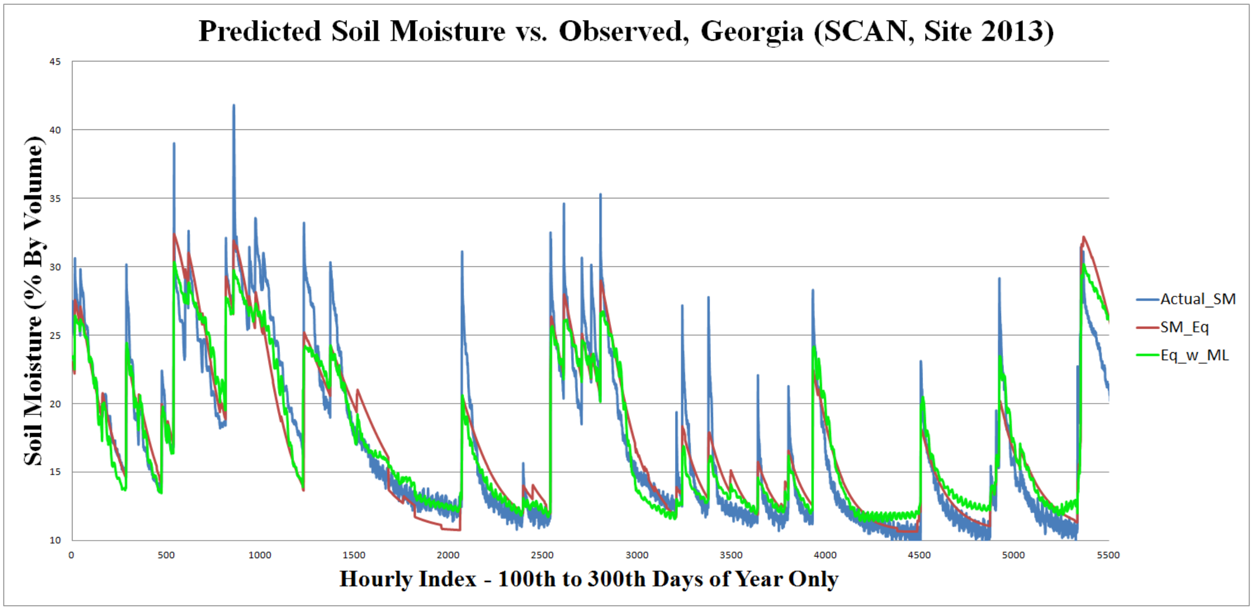

Figure 5, SM Time Series, SCAN Site 2013, Georgia (USA), line colors from Fig. 3
Actual Soil Moisture (Blue Line), Diagnostic Soil Moisture Equation Estimate (Red Line), and Diagnostic Soil Moisture Equation with Machine Learning Error Correction (Green Line)
Hydroclimate: LWC (Low Seasonality, Winter Peak Precipitation, Winter Peak Runoff)
Soil Texture: Sandy Loam

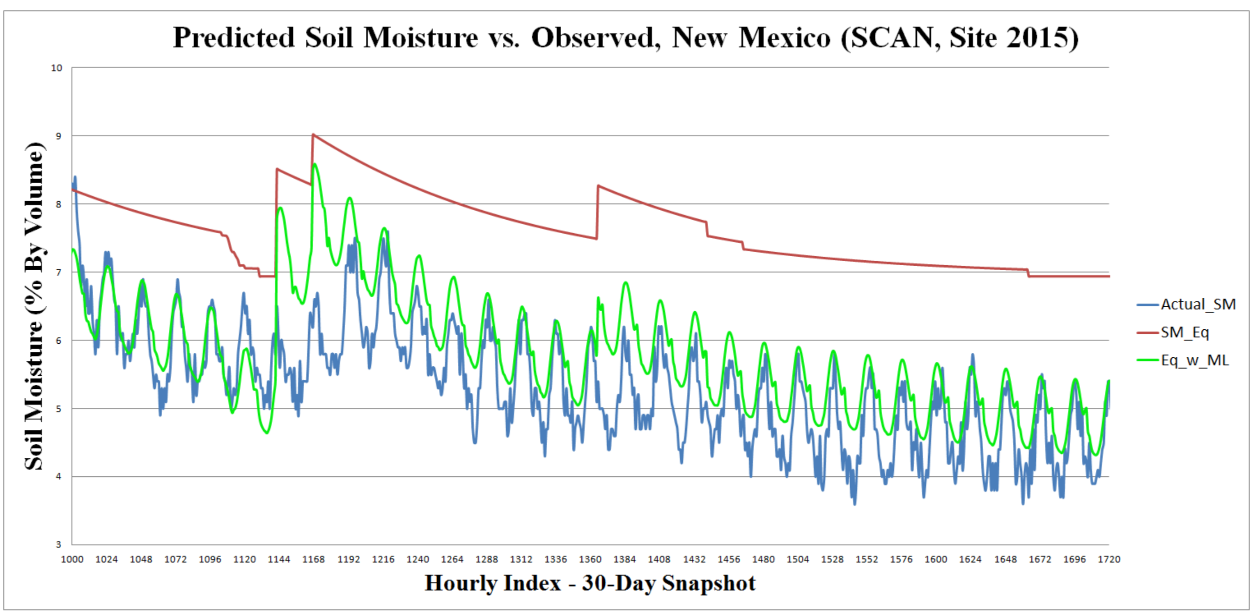

Figure 6, Soil Moisture Time Series, SCAN Site 2015, New Mexico (USA), Actual Soil Moisture (Blue Line), Diagnostic Soil Moisture Equation Estimate (Red Line), and Diagnostic Soil Moisture Equation with Machine Learning Error Correction (Green Line)

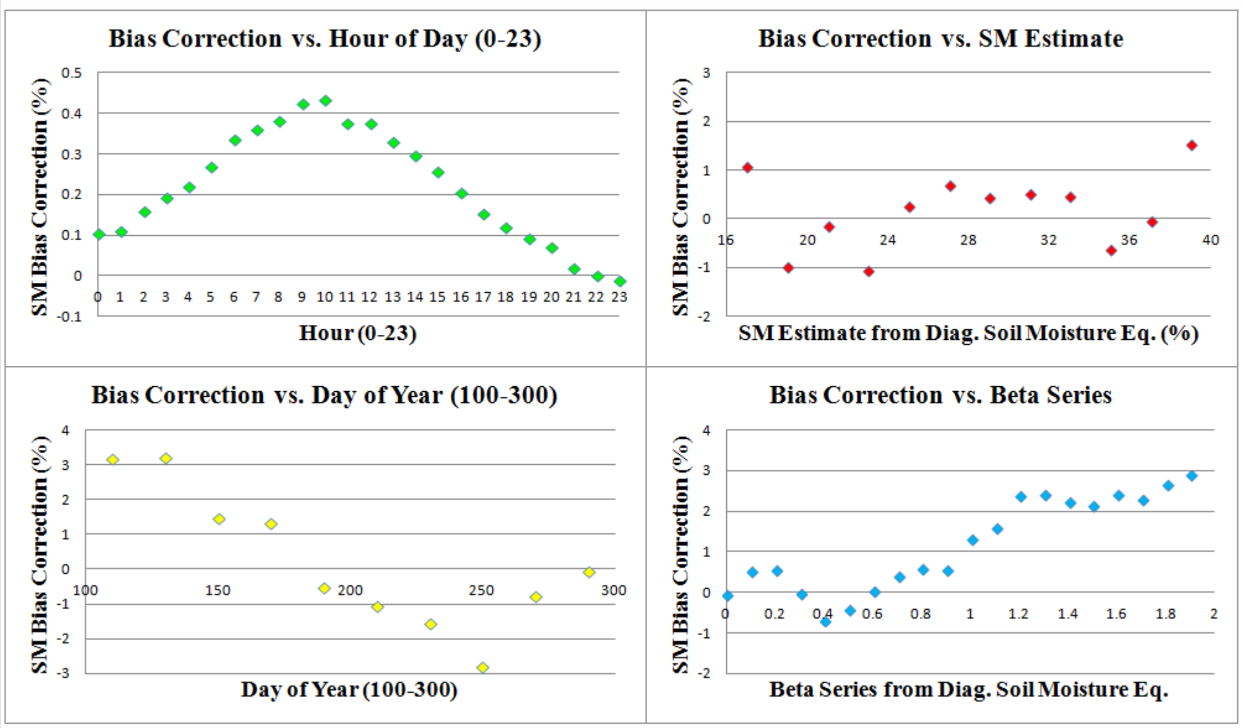

Figure 7, Bias Correction Analysis, SCAN Site 2015 (IAQ, Desert, Loamy Sand)

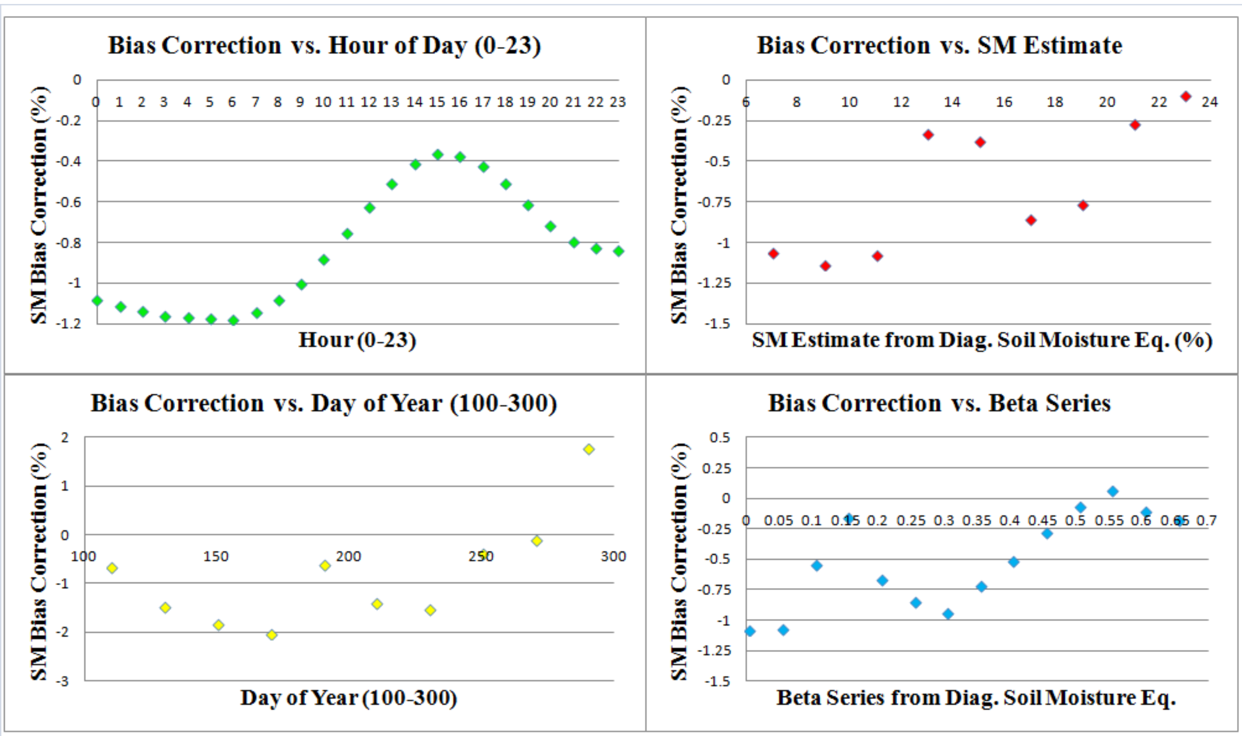

Figure 8, Bias Correction Analysis, SCAN Site 2068 (ISCJ, Plains, Silty Clay Loam)

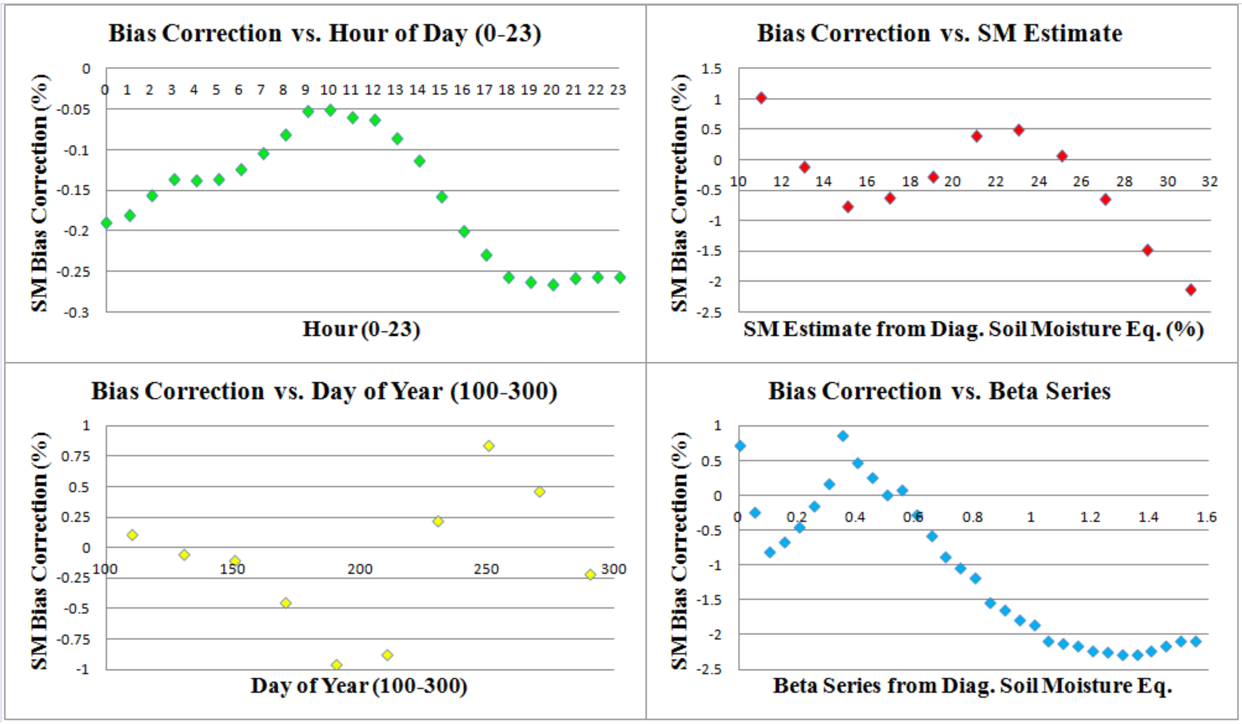

Figure 9, Bias Correction Analysis, SCAN Site 2013 (LWC, Woods, Sandy Loam)

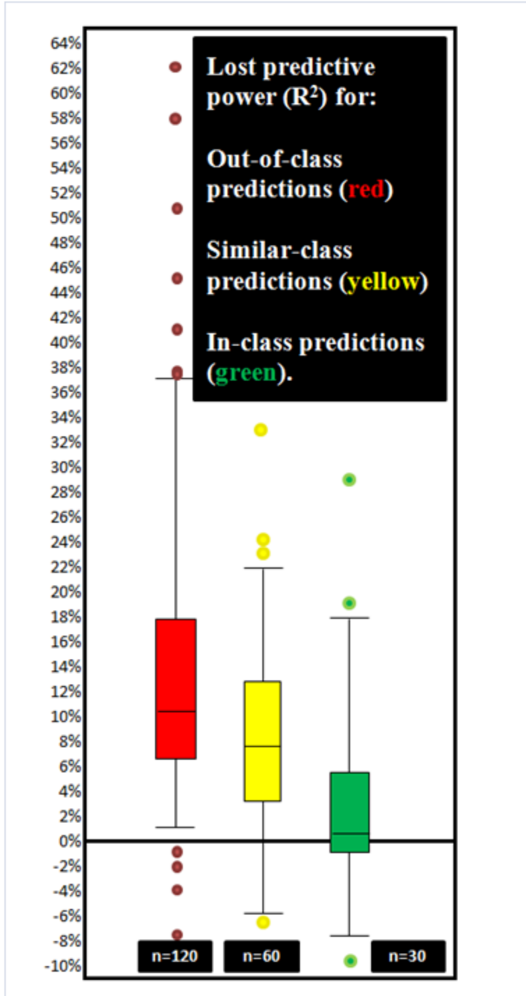

Figure 10, Loss of Predictive Power (R^2) (y-axis) Between Baseline Predictions (model calibrated in the same watershed) and Cross-Validation Predictions (model calibrated in other watersheds)

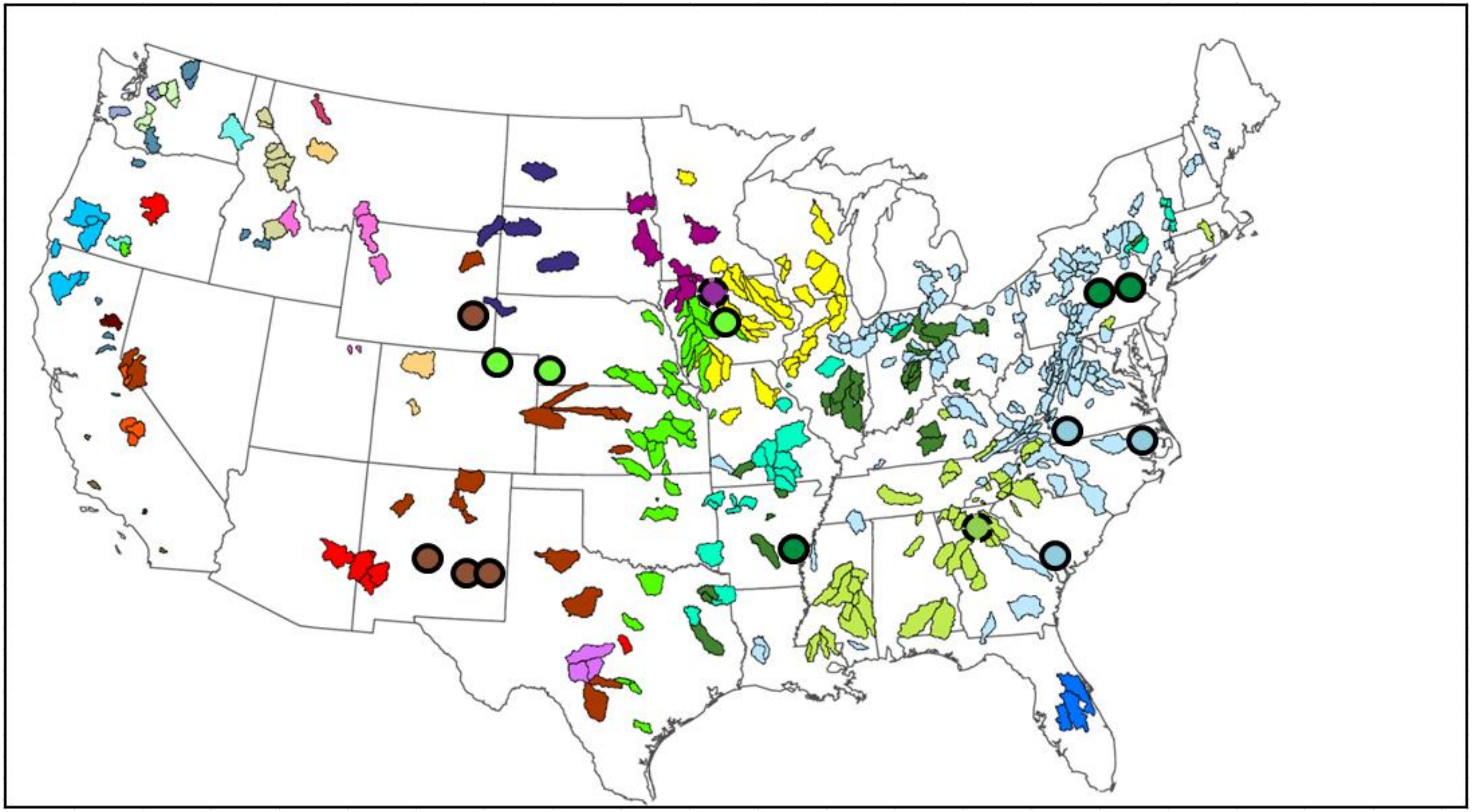

Figure 11, 428 MOPEX catchments colored by hydro-climatic class (Coopersmith et al, 2012).
15 SCAN sensors (for which the Diagnostic Soil Moisture Equation is calibrated) are shown as colored circles.
Circle colors correspond to the hydro-climatic class of the point in question.
Circles with dotted borders are unique (no other sensor for calibration is available within that class

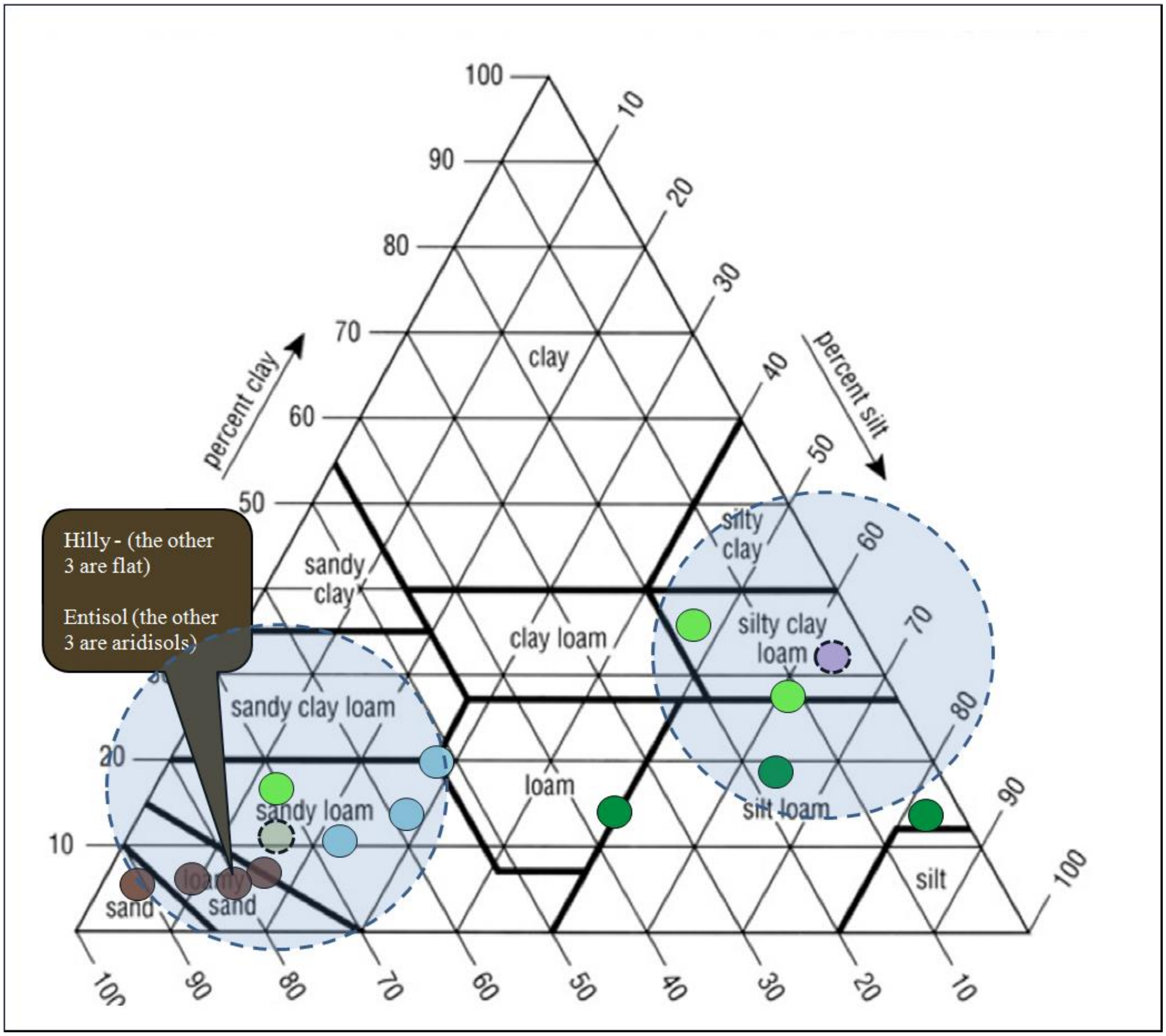

Figure 12, The 15 SCAN sensors, color-coded to match their hydro-climatic class, with similar soil textures shaded.

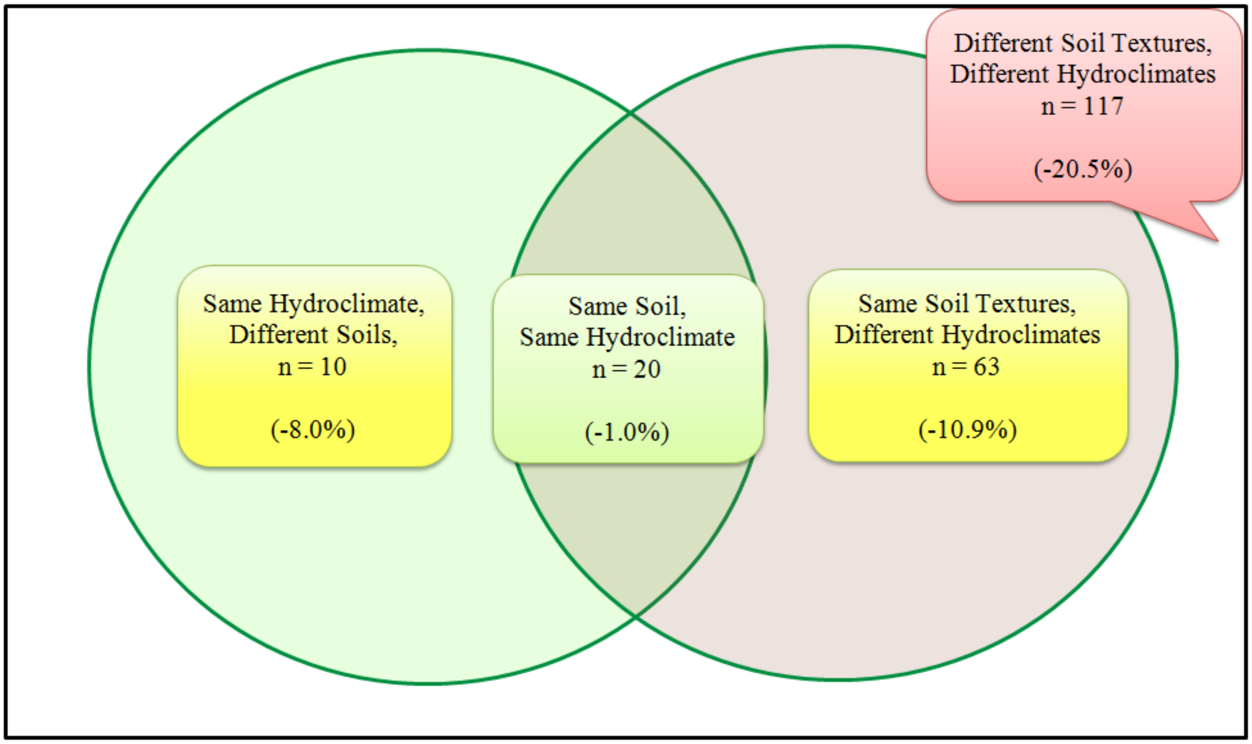

Figure 13 Venn-Diagram of Modeling Errors with Similar and Different Soils and Hydro-climates

SiteID	Hydro-climate	Soil Information	RMSE	RMSE w/ KNN	R^2	R^2 w/ KNN
2008	LJ	Sandy Loam	8.38	7.69	0.590	0.726
2013	LWC	Sandy Loam	2.16	2.06	0.876	0.885
2015	IAQ	Loamy Sand	3.29	2.37	0.740	0.841
2017	ISQJ	Sandy Loam	3.62	3.27	0.637	0.701
2018	IAQ	Loamy Sand*	2.23	2.16	0.803	0.828
2028	LPC	Loam	4.89	4.71	0.707	0.738
2031	ISQJ	Silty Clay Loam	5.46	6.00	0.687	0.750
2036	LPC	Silt Loam	4.61	3.95	0.635	0.726
2038	LJ	Sandy Loam	4.81	4.51	0.546	0.584
2068	ISCJ	Silty Clay Loam	5.28	4.03	0.716	0.837
2089	LJ	Sandy Loam	6.7	6.31	0.682	0.697
2091	LPC	Silt	8.12	6.89	0.539	0.808
2107	IAQ	Loamy Sand	1.98	1.85	0.790	0.843
2108	IAQ	Loamy Sand/Sand	1.26	1.12	0.828	0.863
2111	ISQJ	Silty Clay Loam	5.38	5.01	0.607	0.796

*Not similar to other sandy soils, see Figure 9.

Table 1, The Fifteen SCAN Sites: Class & Soil Information and Performance

	Unrelated Class	Similar Class	Same Class
Median	-10.5%	-7.3%	-0.8%
Mean	-13.7%	-7.7%	-3.4%
Standard Deviation	1.0%	1.1%	1.4%

Table 2, Cross-Validation Results

References

Barricelli, N.A. "Numerical testing of evolution theories. Part II. Preliminary tests of performance, symbiogenesis and terrestrial life". *Acta Biotheoretica* (16): 99–126. 1963.

Capehart, W. J., and Carlson, T.N. “Estimating near-surface soil moisture availability using a meteorologically driven soil water profile model,” J. Hydrol., 160, 1– 20, 1994.

Cheng, C.T., Zhao, M.Y., Chau, K.W., and Wu, X.Y. “Using Genetic Algorithm and TOPSIS for Xinanjiang model calibration with a single procedure.” Journal of Hydrology. Vol. 316, Issue 1-4, pp. 129-140, 2006.

Chico-Santamarta, L., Richards, T., and Godwin, R.J. “A laboratory study into the mobility of travelling irrigators in air dry, field capacity and saturated sandy soils.” American Society of Agricultural and Biological Engineers Annual International Meeting 2009. Volume 4, 2009, Pages 2629-2646.

Choudhury, B. J. and Blanchard, B.J., “Simulating soil water recession coefficients for agricultural watersheds,” Water Resour. Bull., 19, 241–247, 1983.

Coopersmith, E., Yaeger, M., Ye, S., Cheng, L., and Sivapalan, M. “Exploring the physical controls of regional patters of flow duration curves – Part 3: A catchment classification system based on regime curve indicators.” Hydrol. Earth Syst. Sci. doi:10.5194/hess-16-1-2012.

Entekhabi, D. and Rodriguez-Iturbe, I., “Analytical framework for the characterization of the space-time variability of soil moisture,” Adv. Water Resour., 17, 35–45, 1994.

Farago, T., “Soil moisture content: Statistical estimation of its probability distribution,” J. Clim. Appl. Meteorol., 24(4), 371– 376, 1985.

Fix, E., Hodges, J.L. Discriminatory analysis, nonparametric discrimination: Consistency properties. Technical Report 4, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.

Gamache, R.W., Kianirad, E., and Alshawabkeh, A.N. "An automatic portable near surface soil characterization system." Geotechnical Special Publication, Issue 192, 2009, Pages 89-94.

Goldberg, D.E. Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley Professional. 1989.

Jackson, T.J., Bindlish, R., Cosh, M.H., Zhao, T., Starks, P.J., Bosch, D.D., Seyfried, M., Moran, M.S., Goodrich, D.C., Kerr, Y.H., Leroux, D., "Validation of soil moisture and ocean salinity (SMOS) soil moisture over watershed networks in the U.S." IEEE Transactions on Geoscience and Remote Sensing, Volume 50, Issue 5, Part 1, 1530-1543. May 2012.

Jones, H. G. (2004). "Irrigation scheduling: advantages and pitfalls of plant-based methods."Journal of Experimental Botany, 55, 2427-2436

Koster, R.D. and Mahanama, S.P.P. 2012: "Land Surface Controls on Hydroclimatic Means and Variability." J. Hydrometeor, 13, 1604–1620. 2012. doi: http://dx.doi.org/10.1175/JHM-D-12-050.1

O'Neill, P., Entekhabi, D., Njoku, E, and Kellogg, K. "The NASA Soil Moisture Active Passive (SMAP) Mission: Overview". *NASA*. Goddard Space Flight Center, Jet Propulsion Laboratory. http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20110015242_2011016052.pdf (Retrieved 14 May 2013.)

Orth, R.A , Koster, R.D.B, Seneviratne, S.I.A "Inferring soil moisture memory from streamflow observations using a simple water balance model." J. Hydrometeor. Vol.14, Iss. 6, 1773-1790, 2013.

Pan, F., "Estimating daily surface soil moisture using a daily diagnostic soil moisture equation. Journal of Irrigation and Drainage Engineering, 138(7), 625-631, 2012.

Pan, F., and Peters-Lidard, C. D. (2008). "On the relationship between the mean and variance of soil moisture fields." Journal of the American Water Resources Association. 44(1), 235-242.

Pan, F., Peters-Lidard, C. D., and Sale, M. J. (2003). "An analytical method for predicting surface soil moisture from rainfall observations." Water Resources Research, 39(11), Art.

Rosenbaum, U., J.A. Huisman, J. Vrba, H. Vereecken and H.R. Bogena. 2011. Correction of temperature and electrical conductivity effects on dielectric permittivity measurements with ECH2O Sensors. Vadose Zone J. 10: 582-593. doi:10.2136/vzj2010.0083.Saxton, K. E., and Lenz, A.T, "Antecedent retention indexes predict soil moisture," J. Hydraul. Div. Proc. Am. Soc. Civ. Eng., 93, 223– 241, 1967.

Sharifat, K., and Kushwaha, R.L., "Sinkage simulation model for vehicles on soft soil." 2000 ASAE Annual Intenational Meeting, Technical Papers: Engineering Solutions for a New Century 1, pp. 2549-2553

Simunek J., Sejna, M., van Genuchten, M. "The HYDRUS-1D software package for simulating water flow and solute transport in two-dimensional variably saturated media. Version 2.0, IGWMC –TPS –70, International Ground Water Modeling Center, Colorado School of Mines, Golden, CO, 1998.

Singh, A., and Minsker, B.S., "Uncertainty-based multiobjective optimization of groundwater remediation design," Water Resources Research, 44, W02404, doi:10.1029/2005WR004436.

Silva, R.B., Lancas, K.P., Miranda, E.E.V., Silva, F.A.M., and Baio, F.H.R. "Estimation and evaluation of dynamic properties as indicators of changes on soil structure in sugarcane fields of Sao Paulo State – Brazil." Soil and Tillage Research 103 (2), pp. 265-270. 2009.

Steel, R.G.D, and Torrie, J. H., Principles and Procedures of Statistics with Special Reference to the Biological Sciences., McGraw Hill, 1960, pp. 187, 287

Wetzel, P. J., and Chang, J. T., "Evapotranspiration from nonuniform surfaces— A 1st approach for short-term numerical weather prediction," Mon. Weather Rev., 116, 600–621, 1988.

Zhang, X., Srinivasan, R., and Bosch, D. "Calibration and uncertainty analysis of the SWAT model using Genetic Algorithms and Bayesian Model Averaging." Journal of Hydrology, 374, pp. 307-317 (2009).