A hydrometeorological approach for probabilistic simulation of monthly soil moisture under bare and crop land conditions

Sarit Kumar Das¹ and Rajib Maity¹

¹PhD Student, Department of Civil Engineering, Indian Institute of Technology, Kharagpur, India, ²Associate Professor, Department of Civil Engineering, Indian Institute of Technology, Kharagpur, India

Abstract This study focuses on the probabilistic estimation of monthly soil moisture variation by considering (a) the influence of hydrometeorological forcing to model the temporal variation and (b) the information of Hydrological Soil Groups (HSGs) and Agro-Climatic Zones (ACZs) to capture the spatial variation. The innovative contributions of this study are: (i) development of a Combined Hydro-Meteorological (CHM) index to extract the information of different influencing hydrometeorological variables, (ii) consideration of soil-hydrologic characteristics (through HSGs) and climate regime-based zoning for agriculture (through ACZs), and (iii) quantification of uncertainty range of the estimated soil moisture. Usage of Supervised Principal Component Analysis (SPCA) in the development of the CHM index helps to eliminate the “curse of dimensionality.” Typically arises in the multivariate analysis. The usage of SPCA also ensures the maximum possible association between the developed CHM index and soil moisture variation. The association between these variables is modeled through their joint distribution which is obtained by using the theory of copula. The proposed approach is also spatially transferable, since the information on HSGs and ACZs is considered. The “leave-one-out” cross-validation (LOO-CV) approach is adopted for stations belong to a particular HSG to examine the spatial transferability. The simulated soil moisture values are also compared with a few existing soil moisture data sets, derived from different Land Surface Models (LSMs) or retrieved from different satellite-based missions. The potential of the proposed approach is found to be promising and even applicable to crop land also, though with a lesser degree of efficiency as compared to bare land conditions.

1. Introduction

The United Nations Framework Convention on Climate Change (UNFCCC) has declared soil moisture as the Essential Climate Variable (ECV) [GCOS, 2010]. In this paper, a hydrometeorological approach is proposed for probabilistic simulation of the monthly variation of this ECV. The monthly assessment of soil moisture is useful for various scientific studies/applications, such as, climate change impact assessment, variability in hydrological extremes (floods and droughts), and policy decision in agricultural practices.

The spatiotemporal variation of soil moisture depends on various factors. The spatial variation of soil moisture is controlled by precipitation and runoff [Arora and Boer, 2006], antecedent moisture content [Ivanov et al., 2010], meteorological forcing [Rosenbaum et al., 2012], vegetation and evapotranspiration [Grayson et al., 1997], and soil-hydrologic properties [Vachaud et al., 1985; Vereecken et al., 2007; Gaur and Mohanty, 2013]. On the other hand, the temporal variation of soil moisture is influenced by the variation in hydrometeorological variables, such as, potential evaporation and precipitation [Manabe and Delworth, 1990; Huang et al., 1996; Teuling et al., 2007; Lakhankar et al., 2010; Oyedele and Tijani, 2010; Daly and Porporato, 2010]. In brief, the influencing factors for monthly variation of soil moisture at a location are found to be jointly influenced by the soil-hydrologic properties and hydrometeorological forcings [Coopersmith et al., 2014].

Majority of the physically based and conceptual hydrological models for soil moisture simulation incorporate precipitation as the primary input [Burnash et al., 1973; Liang et al., 1996; Todini, 1996; van Dam et al., 1997; Arnold et al., 1998; Wigneron et al., 1999; Pan et al., 2003; Das and Maity, 2014]. However, the temperature and potential evaporation/evapotranspiration are also considered as other influential hydrometeorological variables for soil moisture variation [Burnash et al., 1973; van Dam et al., 1997; Arnold et al., 1998]. A few simulation approaches also consider solar radiation [van Dam et al., 1997; Arnold et al., 1998], wind speed, and relative humidity [Arnold et al., 1998] as the driving hydrometeorological factors for soil moisture variation.
dynamics. However, recent findings indicate that the consideration of hydrometeorological factors and soil characteristics information may improve the soil moisture estimation for ungauged locations [Coopersmith et al., 2014].

The classification of various soil types based on these soil-hydrologic factors is defined in the US National Engineering Handbook as four major Hydrological Soil Groups (HSGs), viz., A, B, C, and D. The rate of infiltration decreases from HSG-A to HSG-D [USDA, 2009]. In addition, the long-term statistical characteristics of soil moisture series can be classified based on Agro-Climatic Zones (ACZs). The ACZs can be delineated based on soil moisture index (level of available soil-water) calculated by following Thornthwaite and Mather [1957]. Further details on HSGs and ACZs are provided later in section 3.

Major objective of this study is to develop a probabilistic hydrometeorological approach for monthly soil moisture simulation by utilizing possibly influencing hydrometeorological variables as well as the information on local soil-hydrologic factors and agro-climatic information. In the proposed approach, all the possibly influencing factors are segregated into two major groups—spatially changing (but temporally invariant) soil characteristics and time varying hydrometeorological factors. The information on local soil properties and climate regimes at any location can be considered through HSGs and ACZs as mentioned before. Recognizing the numerous hydrometeorological factors that may have possible influences on the spatiotemporal variation, the “curse of dimensionality” is a major challenge toward the development of a multivariate probabilistic model [Bellman, 1961; Donoho, 2000]. Though copulas are excellent tools for multivariate analysis, the computational complexity increases exponentially as dimensionality increases [Kao and Govindaraju, 2008; Maity et al., 2013]. Thus, a dimensionality reduction technique is required to reduce the dimension of the input space. A generalized form of Principal Component Analysis is adopted to reduce the dimension of hydrometeorological input variables and to obtain an index, named as Combined Hydro-Meteorological (CHM) index. The association between the developed CHM index and soil moisture is modeled through the joint probability distribution obtained from best fitted copula. The proposed model output is compared against the observed in situ data to determine the performance of the proposed hydrometeorological approach. The proposed model output is also compared with the output of a few existing simulated soil moisture data sets derived from different LSMs or retrieved from different satellite missions. Spatial transferability of the proposed model, developed using HSG and ACZ information, is also tested.

2. Methodology

A flowchart, showing the entire methodology along with its performance evaluation (temporal and spatial validation), is presented in Figure 1. In the following subsections, major steps are discussed in detail.

2.1. Data Preprocessing

Anomaly values of all the data series are obtained by subtracting the long-term mean series of the corresponding variable. The anomaly series of each variable is fitted to the best-suitable probability distribution function and the reduced variate of the anomaly data series are derived using the best probability distribution function. Thus, the biases, arising from the range of individual variable, are eliminated, and the probabilistic features of the hydrometeorological variables are considered. The best-suitable probability distribution function is identified through statistical tests. These reduced variates are used in developing the CHM index through dimensionality reduction.

2.2. Dimensionality Reduction and Development of CHM Index

Among the various statistical techniques employed for dimensionality reduction, the Principal Component Analysis (PCA) is the most popular and commonly used in many studies [Keyantash and Dracup, 2004; Pan et al., 2012; Maity et al., 2013; Das and Maity, 2013]. Based on the eigenvalue decomposition, the PCA technique uses an orthogonal transformation to convert a set of data series into a set of linearly uncorrelated variables (known as principal components) in a decreasing order of variance [Haan, 2002]. However, by design, these components do not consider their association with the response variable. Maity et al. [2013] applied such an unsupervised PCA technique for dimensionality reduction to a drought related study. It was shown that the first few principle components comprised most of the variability of the original data set. As a consequence, n-dimensional, multivariate modeling is essential, where n is the required number of principal components to be considered. Maity et al. [2013] adopted three copula (copula that combines

Water Resources Research
marginal distribution of three different random variables to obtain a trivariate joint distribution function of the random variables) to consider first two principal components, which is still manageable. If the considerable amount of variation is shared by more than two principal components, such modeling approaches would be much more computationally challenging, if not impossible. Therefore, in this study, an approach based on the Supervised Principal Component Analysis (SPCA) is adapted for dimensionality reduction of the set of all the hydrometeorological variables to develop the CHM index.

The approach of SPCA is able to derive the principal component having the highest degree of association with the target variable [Barshan et al., 2011]. Several approaches are proposed by researches for the SPCA [Barshan et al., 2011; Fisher, 1936; Fukumizu et al., 2004; Globerson and Roweis, 2006; Yu et al., 2006]. The methodology developed by Barshan et al. [2011] is adopted here. This approach has an advantage of a closed-form and quantitative solution of the SPCA. In this technique, the Hilbert-Schmidt Independence Criterion (HSIC), proposed by Gretton et al. [2005], is utilized to derive the principle components based on an orthogonized transformation of the input matrix. Methodology is explained here in the context of the set of various hydrometeorological variables as input matrix and soil moisture series as target variable.
Suppose, $X \in \mathbb{R}^{p \times n}$ is the $p \times n$ dimensional matrix ($[x_{1,1}, x_{1,2}, \ldots, x_{1,n}]$) of $p$ different hydrometeorological variables expressed in terms of their reduced variates. The expanded form of $x_i$ can be shown as $x_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,n}]^T$ ($i = 1, \ldots, n$), where $n$ is the number of observations. Also, assume that $Y \in \mathbb{R}^{1 \times n}$ is the one-dimensional vector of the length $n([y_1, y_2, \ldots, y_n])$ of the reduced variate of soil moisture anomaly. Assuming $Z = ([x_1, y_1], [x_2, y_2], \ldots, [x_6, y_6])$ as independent and identically distributed (i.i.d.) samples, there exists an orthogonal projection matrix, $U$ of size $p \times 1$, such that $Y$ is directly dependent on the projected input matrix, $U^T X = [U^T \times [x_1, x_2, \ldots, x_n]]$. The methodology to obtain $U$ is presented in Appendix A.

The elements of $U$ can be directly used as multiplying coefficients for different hydrometeorological variables (reduced anomaly values) to obtain a combined index that ensures the maximum association between the combined variable and the target variable. However, it does not guarantee the in-phase association between them, i.e., two variables may or may not attain same or related stages (high values, low values, etc.) at the same instance of time. The in-phase association is generally reflected through a positive correlation coefficient. In such case, elements of $U$ are directly used as coefficients for CHM index. On the other hand, if the correlation between the derived variable and the target variable is found to be negative, out-of-phase association is indicated. In such situation, the elements of $U$ are multiplied by $(-1)$ in order to obtain in-phase association. Thus, in such case, the coefficients for CHM index are obtained by $(-1) \times [U]$. The coefficients of CHM index are multiplied with the reduced anomaly values of the hydrometeorological inputs to obtain the CHM index, denoted as $M$. The CHM index, thus developed, stores the combined information on different hydrometeorological variables responsible for the temporal variation of soil moisture. Physically, the coefficients of the CHM index provide information on the weightages for each of the considered hydrometeorological variables. It should be noted that the sum of the square of the CHM coefficients equals to 1.

### 2.3. Probabilistic Soil Moisture Simulation Model Using CHM Index

To obtain the joint distribution between any two or more random variables, copulas are the most suitable choice. Copulas are the functions that join the individual marginal distribution of any two or more variables [Nelsen, 2006]. Recently, these functions are used to obtain joint distribution among hydrological or climatological variables like precipitation, runoff, storm duration, flood frequency, soil moisture, etc. [e.g., Salvadori and De Michele, 2007; Kao and Govindaraju, 2007; Maity and Nagesh Kumar, 2008; Maity et al., 2013, and references therein].

#### 2.3.1. Development of Copula-Based Probabilistic Simulation Model

The theory of copulas is adopted to formulate joint probability distribution between the CHM index and the soil moisture anomaly. First, the association between the CHM index and the reduced variate of soil moisture anomaly is measured using Kendall’s Tau ($\tau$). The sample estimate of $\tau$ is obtained as the difference between the probability of concordance and discordance of two random variables. Let, $V$ and $Y$ are the reduced variate of the CHM index and soil moisture anomaly, respectively. Thus,

$$\tau = P([V_i - V_j] (Y_i - Y_j) > 0) - P([V_i - V_j] (Y_i - Y_j) < 0)$$

where the subscripts $i, j$ indicate any two time steps with $i \neq j$. The estimated $\tau$ is used to obtain the dependence parameter of a copula. It may be noted that the reduced variates of the random variables are not mandatory for the computation of $\tau$. It can be computed either with actual values or the quantiles of the random variables, and the value of $\tau$ will be same in either case.

The commonly used bivariate Archimedean copulas are Ali-Mikhail-Haq (AMH), Clayton, Frank, and Gumbel-Hougaard (GH). However, the AMH copula is not considered in the tentative pool of copulas since the range of dependence for this copula is very narrow ($\tau = -0.182$ to 0.333). The details of selected copulas are provided in Table 1. First, the tentatively eligible copulas are picked out to model the association between reduced variate of soil moisture anomaly and CHM index based on the Kendall’s Tau ($\tau$) value. If there are more than one feasible copulas, the final selection of a suitable copula is based on the best fit to the observations.

#### 2.3.2. Selection of Best Fit Copula

The suitability of a particular copula is assessed statistically by goodness of fit (GOF) tests using—(a) empirical copula, (b) Kendall’s transform, and (c) Rosenblatt’s transform. The test using empirical copula is based on the null hypothesis $H_0 : C \in \mathcal{C}$ for a particular copula $C$ against $H_1 : C \notin \mathcal{C}$. This approach was
developed by Genest et al. [2009] and applied to other hydrologic studies as well [e.g., Maity et al., 2013]. These tests compare the distance between the empirical copula ($C_n$) and parametric estimate ($C^0_n$) of $C$, obtained under $H_0$. Formally, the goodness of fit tests are based on the statistic $\sqrt{n} \{ C_n(v, y) - C^0_n(v, y) \}$, where $v$ is the reduced variate of CHM index; $y$ and $n$, as explained before, are the reduced variate of soil moisture anomaly and the number of observations, respectively. The empirical copula ($C_n$) is defined as [Deheuvels, 1981] in equation (2)

$$C_n(v, y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(V \leq v, Y \leq y), \quad v, y \in [0, 1]$$ (2)

where $\mathbb{I}(\cdot)$ is the indicator function that takes a value 1 if the argument $\cdot$ is true and 0 if it is false. The Cramér-von Mises and Kolmogorov-Smirnov (KS) statistics are based on the distance explained above. The Cramér-von Mises statistic ($S_n$) is a popular goodness of fit test procedure for copula models [Genest et al., 2009]. The statistic, $S_n$, is expressed as:

$$S_n = \sqrt{n} \{ C_n(v, y) - C^0_n(v, y) \}^2$$ (3)

where $C_n$ and $C^0_n$ are as explained before. The KS statistic ($T_n$) is based on the absolute maximum distance between $C_n$ and $C^0_n$. It is expressed as:

$$T_n = \max_{v \in [0, 1]} | \sqrt{n} \{ C_n(v, y) - C^0_n(v, y) \} |$$ (4)

For the copula selection procedure using Kendall’s transform, $\kappa$, the univariate distribution function of joint distribution is derived parametrically using a particular copula, $C^\kappa_n$. The $\kappa$ is determined either parametrically ($\kappa^0_n$) or nonparametrically ($\kappa_n$). The null hypothesis, $H_0$: $\kappa \in \kappa_0$ for a particular distribution of $C^\kappa_n$ is checked against the alternate hypothesis $H_1$: $\kappa \notin \kappa_0$. $\kappa_n$ is derived using the empirical distribution function of the rescaled form of $C^\kappa_n$ which is the Kendall’s transform [Genest and Rivest, 1993; Genest et al., 2009] as given below,

$$\kappa_n(c^\kappa_n) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(C^\kappa_n \leq c^\kappa_n)$$ (5)

The test statistics ($S_n^{(\kappa)}$ and $T_n^{(\kappa)}$) are basically the rank-based analogs of the Cramér-von Mises and KS statistics [Genest et al., 2009]. The test statistic $S_n^{(\kappa)}$ and $T_n^{(\kappa)}$ are expressed as:

$$S_n^{(\kappa)} = \sum_{i=1}^{n} K_n(c^\kappa_n)^2 (\kappa_n - \kappa_n)$$ (6)

$$T_n^{(\kappa)} = \sup_{c^\kappa_n} | K_n(c^\kappa_n) |$$ (7)

where $K_n = \sqrt{n}(\kappa_n - \kappa_n)$ and $n$ is the number of observations.

The Rosenblatt probability integral transformation of the copula is defined as $R(v, y) = (e_1, e_2)$, where $e_1 = \frac{\partial C_n}{\partial v}$, $e_2 = \frac{\partial C_n}{\partial y}$. Based on the properties of Rosenblatt’s transform, $(v, y)$ is approximately distributed as $C^0_n$ if

<table>
<thead>
<tr>
<th>Name</th>
<th>Copula Function</th>
<th>Generator Function, $\phi_n(t)$</th>
<th>Parameter of Joint Distribution ($\theta$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clayton</td>
<td>$\max { (u^\nu + v^{-\nu} - 1, 0) }^{1/\nu}$</td>
<td>$\frac{1}{\nu}(u^\nu - 1)$</td>
<td>$[-1, \infty)$ excluding 0</td>
</tr>
<tr>
<td>Frank</td>
<td>$-\frac{1}{\nu} \ln \left( 1 + \left( \frac{1}{\nu} - 1 \right) \left( \frac{u}{v} - 1 \right) \right)$</td>
<td>$-\ln \left( \frac{u^\nu - 1}{v^\nu - 1} \right)$</td>
<td>$(-\infty, \infty)$</td>
</tr>
<tr>
<td>Gumbel-Hougaard</td>
<td>$\exp \left( -\left( -\ln u^\nu + (-\ln v^\nu) \right)^{1/\nu} \right)$</td>
<td>$(-\ln t)^{\nu}$</td>
<td>$[1, \infty)$</td>
</tr>
</tbody>
</table>

*First-order Debye function, $D_1(\theta) = \frac{1}{\nu} \frac{\partial}{\partial \nu} \frac{\partial}{\partial t} d$ for $\theta > 0; D_1(\theta) = D_1(0) + \frac{1}{\nu}$. 

Table 1. Properties of the Preliminary Selected Bivariate Archimedean Copula Functions
and only if the \( R(v, y) \) is a bivariate independence copula, i.e., \( C_{\perp}(e_1, e_2)=e_1 \times e_2 \), where \( e_1 , e_2 \in [0, 1] \). Thus, the null hypothesis is expressed as: \( H_0: \ R \in R_0 \) and checked against the alternate hypothesis, \( H_0^R: \ R \notin R_0 \). The \( R \) is estimated either parametrically (\( R_0^P \)) or nonparametrically (\( R_0^N \)). The \( R \) is derived following Genest et al. [2009] as given in equation (8) whereas, \( R_0^N \) is equivalent to \( C_{\perp} \)

\[
R_n(e_1 , e_2)= \frac{1}{n} \sum_{i=1}^{n} 1(E_1 \leq e_1, E_2 \leq e_2) \quad e_1 , e_2 \in [0, 1]
\]  

Further, two Cramér-von Mises statistics, \( S_n^{(p)} \) and \( S_n^{(c)} \) are estimated to check the distance between \( R_n \) and \( R_0^{(p)}, c^{(p)} \) can be calculated as:

\[
S_n^{(p)}=n \sum_{i=1}^{n} (R_n(e_1, e_2) − C_{\perp}(e_1, e_2))^2
\]

and \( S_n^{(c)} \) can be estimated as:

\[
S_n^{(c)}=n \sum_{i=1}^{n} (R_n(e_1, e_2) − C_{\perp}(e_1, e_2))^2R_n(e_1, e_2)
\]

For all the measures \( (S_n, T_n, S_n^{(c)}, T_n^{(c)}, S_n^{(p)} \), and \( S_n^{(c)} \)), lower the value better is the fit. Thus, the copula function with the lowest value of these statistics indicates the best fit copula. Further, when the best fit copula is found different for different statistics, the more preferable statistic is honored while selecting the best fit copula. The preference order is \( S_n^{(p)}>S_n>S_n^{(c)}>T_n^{(c)}>T_n^{(c)}>T_n^{(c)} \) based on their power [Genest et al., 2009]. The copula showing best fit based on these criteria is selected for further analysis and denoted as \( C(v, y) \).

### 2.3.3. Derivation of Probabilistic Conditional Distribution of Soil Moisture Anomaly

The best fit copula is employed to obtain the joint distribution of soil moisture anomaly and CHM index. The probabilistic estimation of the former is carried out by employing the conditional distribution (conditioned on the latter), which is derived from their joint distribution. The conditional (cumulative) distribution of the soil moisture anomaly \( (S) \) conditioned on the CHM index \( (M) \) is expressed as:

\[
F_{S|M}(s|M=m) = \frac{\partial C(v, y)}{\partial v}
\]

Using this conditional distribution, the 50th quantile value is used as the Expected Value (EV) of the soil moisture anomaly. Range of estimated value is quantified at 5th quantile (used as Lower Limit, LL) and 95th quantile (used as Upper Limit, UL). The estimated quantiles of soil moisture anomaly are added to the long-term mean to obtain the estimated soil moisture and the range of estimation.

### 2.4. Assessment of Model Performance

The model performance is assessed during the model development period. The approach being probabilistic, LL, EV, and UL of the soil moisture are estimated as mentioned before. In fact, any desired range of estimation can be made available from the proposed model output (shown later). However, different statistical measures, such as, correlation coefficient \( (r) \), normalized Root Mean Square of Error (nRMSE), and refined index of agreement \( (d_r) \) proposed by Willmott et al. [2012] are computed between a specific estimated value and the observed value. Thus, these statistics are computed between in situ data and EV of the estimates. The \( d_r \) is adopted over the commonly used Nash-Sutcliffe Coefficient of Model Efficiency (NSE) because \( d_r \) provides more logical lower boundary of “–1” instead of –\( \infty \) in case of NSE, and it overcomes the bias of NSE on high magnitude of error values as it takes the absolute value of error rather than the square of as in NSE. However, there is a caution needed in explaining the \( d_r \) value close to “–1” as it indicates low variability in the observed values rather than the low efficiency of the simulation model [Willmott et al., 2012].

### 2.5. Model Performance for Temporal Consistency and Spatial Transferability

The model performances are checked for both temporal consistency and spatial transferability. The temporal consistency of the models ensures its applicability to a time period other than the model development period and the spatial transferability ensures its applicability to other (ungauged) locations. However, soil/
climate properties, in terms of HSG and ACZ classification, and information on input variables should be known.

To check the temporal consistency of the proposed approach, first, the simulation models are developed for a location using the data from the development period. The performance of the developed models is then assessed with the data from some other time period (testing period) at the same location. If the performance during the testing period is reasonable and comparable to that during the model development period, the developed model can be considered temporally consistent.

To develop spatially transferable models, HSG-based models (specific to a particular HSG) are developed. The HSG-based models mean that the coefficients of CHM index and values of Kendall’s τ are obtained for each HSG individually using the combined data from all the stations belongs to a particular HSG. The methodology remains same, as explained in sections 2.1–2.3, except that the data from all the stations within the HSG considered altogether. The station-based anomalies for the hydrometeorological input variables are computed using their individual long-term means derived from the station data. However, for soil moisture, the anomaly values are obtained by using the long-term mean, derived for an ACZ in which the particular station belongs.

The spatial transferability of the HSG-based model is checked by adopting “leave-one-out” cross-validation (LOO-CV) scheme [Hubert and Engelen, 2007]. LOO-CV is implemented in the following stepwise way:

Step 1: A particular HSG is considered.

Step 2: A particular station is selected as a target station within the considered HSG. The “target station” (or location) is considered as ungauged, i.e., historical records of soil moisture data are not available. The ACZ is identified in which the target station belongs.

Step 3: The statistical characteristics (the long-term mean and parameters of probability distribution) for the soil moisture data at the target station are considered to be the average of all the stations belong to the same ACZ, identified in Step 2. This is because, as mentioned before, the zoning according to ACZ is based on the similar soil moisture characteristics.

Step 4: The simulation model is developed based on the CHM index and Kendall’s τ, derived from the combined data set of all stations belong to the considered HSG, excluding the target station.

Step 5: The performance of the developed model (Step 4) is checked at the selected target station.

Step 6: The Steps 2–5 are repeated to consider another station within the HSG, considered in Step 1. This is repeated till all the stations are considered.

Step 7: Steps 1–6 are repeated to consider other HSGs. This step is repeated till all the HSGs are considered.

Now, following this approach, if the performance at the target station is more or less similar to the station-based model (developed using data from that station itself), the proposed approach may be considered to be spatially transferable.

Next, to check the temporal consistency of the spatially transferable model, the performance of the developed HSG-based model is tested with the data from some other time period (testing period) at all the monitoring stations within the HSG. The potential of the proposed approach with respect to its spatial transferability can be identified by the performance of the developed model at the target station. This is because the target station is treated as ungauged with respect to soil moisture data, which is not used in the model development process as explained above.

3. Study Area and Data

3.1. Study Area

The Agricultural Meteorology Division (AgriMet) under India Meteorological Division (IMD) has classified entire India into different Agro-Climatic Zones (ACZs) [Rao et al., 1972; Agrimet, 2012]. The Central Ground Water Board [Central Ground Water Board (CGWB), 2007] has classified entire India into four major HSGs, discussed earlier. The properties of the different HSGs and hydrometeorological stations belonging to each HSG are presented in Table 2. In this study, 24 monitoring stations in India, located across various ACZs and
HSGs are selected for the analysis. The locations of the monitoring stations are shown on the combined ACZ and HSG map (Figure 2) of India. The altitude of these stations varies between 10 and 874 m. All the soil moisture monitoring stations have a facility to collect meteorological data, which is applicable for nearby bare as well as crop land. However, simulation models are developed for bare land and crop land locations separately.

### 3.2. Data Used

The weekly in situ surface (0–7.5 cm) soil moisture for the bare land as well as crop land, and daily observed data for hydrometeorological inputs (evaporation, precipitation, relative humidity, sunshine hours, and temperature) measured at 24 monitoring stations across different HSG and ACZs in India during 1991–2006 are used. All these data sets are procured from National Data Centre (NDC) of IMD, Pune, India. Evaporation, precipitation, relative humidity, sunshine hours temperature, and soil moisture are measured using standard USA Open Pan-Evaporimeter, standard Rain-guage, hygrograph, standard sunshine recorder, thermometer in Celsius scale, and by Gravimetric method, respectively [IMD, 2009]. For a few time steps, the in situ soil moisture data and/or the data for the hydrometeorological input variables are missing. These time instances are ignored from the analysis.

All the data are converted to monthly scale by taking the time-weighted average. The monthly anomaly values of all the variables are obtained by subtracting the respective long-term means. The data available during 1991–2002 are utilized for model development and the same during 2003–2006 are reserved for testing. The proposed framework of developing soil moisture simulation model is applied for bare land and crop land areas. However, crop land areas in those monitoring stations, where perennial deep-rooted crops like, coconut, areca nut, etc. are grown, are applied with the model developed for bare land as these deep-rooted crops have minimum impact on the moisture content in top soil layers. Out of the 24 monitoring stations used for bare land, only 12 stations have adequate data points for crop land, and used for developing models of crop land. The model performance for temporal consistency as well as spatial transferability for models developed for crop land is checked as similar to the bare land.

The simulated soil moisture values by the proposed hydrometeorological approach are also compared with four existing soil moisture data sets obtained from various Land Surface Models (LSMs) and different satellite-based missions. These are obtained from (i) Climate Prediction Center (CPC H96), (ii) European Centre for Medium-Range Weather Forecasts (ECMWF) (ERA-Interim/Land), (iii) Global Land Data

### Table 2. Properties of Different Hydrological Soil Groups (HSG)

<table>
<thead>
<tr>
<th>HSG</th>
<th>Description</th>
<th>Infiltration Rates (mm/h)</th>
<th>Name of the Stations Belonging to</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Soils having high infiltration rates even when thoroughly wetted and consisting chiefly of deep, well to excessively drained sands or gravels. These soils have a high rate of water transmission</td>
<td>Above 25.0</td>
<td>Bellary, Bhubaneshwar, Hebbal, Hisar, Kovilpatti, Pillamedu</td>
</tr>
<tr>
<td>B</td>
<td>Soils having moderate infiltration rates when thoroughly wetted and consisting chiefly of moderately deep to deep, moderately well to well-drained soils with moderately fine to moderately coarse textures. These soils have a moderate rate of water transmission</td>
<td>12.5–25.0</td>
<td>Agra, Durgapur, Kanke, Karnal, New Delhi, Sabour, Udaipur</td>
</tr>
<tr>
<td>C</td>
<td>Soils having slow infiltration rates when thoroughly wetted and consisting chiefly of moderately deep to deep, moderately well to well-drained soils with moderately fine to moderately coarse textures. These soils have a moderate rate of water transmission</td>
<td>2.5–12.5</td>
<td>Kalyani, S K Nagar, Vellanikara, Vellayani, Vittal</td>
</tr>
<tr>
<td>D</td>
<td>Soils having very slow infiltration rates when thoroughly wetted and consisting chiefly of clay soils with a high swelling potential, soils with a permanent high water table, soils with a clay pan or clay layer at or near the surface, and shallow soils over nearly impervious material</td>
<td>Below 2.5</td>
<td>Adhartal, Bhopal, Nagpur, Niphad, Pune</td>
</tr>
</tbody>
</table>
Assimilation System with Noah Land Surface Model (GLDAS-Noah), and (iv) Modern-Era Retrospective analysis for Research and Applications with land model component (MERRA-Land). Details of these models are provided in Table 3. Data from all these models are gridded. However, the available grid intersections do not always match with the station location. Thus, the data from four grid intersections around the station location are used, and the value at the station location is computed by Inverse Distance Weighting (IDW) method.

4. Results and Discussions

4.1. Computation of CHM Index for Station-Wise Models

Having tested different feasible probability distribution functions (e.g., Normal, Logistic, and Pearson Type III), Logistic distribution is adopted to obtain the reduced variates of all the input variables and soil moisture anomalies based on results of the Kolmogorov-Smirnov Test (see Table S1 in the supporting information). Next, following the methodology explained in section 2.2, the coefficients of the CHM index are computed for all the stations considered in the analysis. Results are shown in Table S2 (first seven columns) in the supporting information. As stated before, while ensuring the maximum possible association between the developed CHM index and soil moisture anomaly, the values of CHM coefficients also represent the weightages of different hydrometeorological variables in the CHM index. It is noticed that there is a similarity in the coefficients for a particular input across different stations. For instance, evaporation, sunshine hours, and temperature are found to have negative coefficients; whereas precipitation and relative humidity are found to have positive coefficients. These observations also support the expected nature of influence of these hydrometeorological variables on the variation of soil moisture. Thus, as desired, extraction of target-specific hydrometeorological forcing (CHM index) following the proposed approach is found to be effective in suitably blending the different hydrometeorological variables. The CHM index is also analyzed to obtain its best fit probability distribution function. Unlike the individual input variables, the marginal distribution of the CHM index is found to follow the Normal distribution at 5% significance level in the KS Test. Before proceeding further, a sensitivity analysis is carried out in the next section to investigate the sensitivity of individual hydrometeorological inputs.

4.2. Sensitivity Analysis for Hydrometeorological Inputs

As explained before, the five hydrometeorological inputs are selected for soil moisture simulation based on their possible role in soil moisture dynamics. However, all these inputs may or may not be equally sensitive for simulation across all ACZs and HSGs. This is because the spatiotemporal variation of soil moisture changes with climate and soil regimes. Thus, the sensitivity analyses are carried out, by eliminating one input variable at a time, and by determining the model performance measures. These analyses are performed during the model development period for station-based models at all the stations. The model performance with or without a particular input variable is assessed both at actual and anomaly scale. An input variable is classified as “more sensitive” if elimination of that particular variable results in deterioration in the model performance. On the other hand, the elimination of “less sensitive” input variable does not change the model performance much. This sensitivity analysis is executed both bare land and crop land models using monthly mean and monthly anomaly values of the variables. Results for bare land and crop land models at monthly anomaly scale are grouped into different ACZs and presented in Figure 3. The results indicate that all the variables are more or less sensitive. Though the results do distinguish the variables in order of their insensitivity, in general, it is noticed that the precipitation, evaporation, and relative humidity are “more sensitive” for bare land cases, whereas, precipitation and sunshine hours for crop land cases for most of the ACZs. At many stations, sunshine hours and temperature for bare land, and relative humidity and temperature for crop land are found to be “less sensitive” across different ACZs. In wetter zones (Per-humid, Humid, Moist subhumid) as well as in dry zones (semiarid and Arid), precipitation is found to be more sensitive as compared to evaporation. However, for moderately dry zone (Dry subhumid), evaporation and relative humidity are found to be more sensitive than precipitation.

4.3. Selection of Best Fit Copula and Simulation of Soil Moisture

As mentioned in section 2, the joint distribution between the soil moisture anomaly and the CHM index is modeled using three preliminarily selected Archimedean copulas—Clayton, Frank and Gumbel-Hougaard. Goodness of fit tests are carried out to select the best fit copula using the statistics explained in the
methodology (section 2.3.2), and the results are presented in Table S3 (in the supporting information). It is noticed that the Clayton copula is the most preferred at the majority of stations based on the measures $S_B$ and $S_n$. Moreover, the measure $S_j$ indicates Clayton to be the best choice among all other copulas for all the stations. Other statistics also indicate the preference of the Clayton copula over others, at most of the stations. Thus, honoring the preference ranking of different measures (i.e., $S_B$/$S_n$, $S_j$/$S_C$, $S_T$, as mentioned in section 2), Clayton is selected as the best fit copula. Thus, the joint distribution between the soil moisture anomaly and the CHM index is modeled using the Clayton copula. The dependence parameter ($\theta$) for the joint distribution is obtained from the Kendall's tau ($\tau$) between soil moisture anomaly and CHM index. The values of $\theta$ are shown in the last column of Table S2 (supporting information) for all the stations spreading over different HSGs. The developed joint distribution is used to obtain the conditional distribution of soil moisture anomaly, conditioned on CHM index using equation (11). The conditional (cumulative) distribution is used to compute the quantile values at 5%, 50%, and 95% probability and

![Figure 2. Locations of monitoring stations on combined map of Agro-Climatic Zone (ACZ) and Soil-Hydrologic Group (HSG) of India.](image)
### Table 3: Details of Existing Modeled Data Sets Used to Assess Model Performance

<table>
<thead>
<tr>
<th>Existing Data Set Name</th>
<th>Soil Layer Depth</th>
<th>Spatial Resolution</th>
<th>Temporal Resolution</th>
<th>Available Time Period</th>
<th>Model Reference</th>
<th>Source</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>Climate Prediction Center (CPC H96) Version 2</td>
<td>1.6 m</td>
<td>0.5° x 0.5° (latitude x longitude)</td>
<td>Monthly</td>
<td>1948 to present</td>
<td>CPC H96 [Huang et al., 1996; Fan and van der Dool, 2004]</td>
<td>National Oceanic and Atmospheric Administration (NOAA), USA</td>
<td><a href="http://www.esrl.noaa.gov/pds/">http://www.esrl.noaa.gov/pds/</a></td>
</tr>
<tr>
<td>European Centre for Medium-Range Weather Forecasts (ECMWF) ERA-Interim/Land</td>
<td>0–7 cm</td>
<td>80 km by 80 km</td>
<td>6 h</td>
<td>1979–2010</td>
<td>ECMWF Scheme for Surface Exchanges Over Land (<a href="http://apps.ecmwf.int/datasets/">HTESSEL</a> land-surface model [Alberge et al., 2013])</td>
<td>European Centre for Medium-Range Weather Forecasts (ECMWF)</td>
<td><a href="http://apps.ecmwf.int/datasets/data/ensembles-ecmwf-data/">http://apps.ecmwf.int/datasets/data/ensembles-ecmwf-data/</a></td>
</tr>
<tr>
<td>Global Land Data Assimilation and Modeling System (GLDAS-Noah)</td>
<td>0–10 cm</td>
<td>0.25° x 0.25° (latitude x longitude)</td>
<td>Monthly</td>
<td>1948–2010</td>
<td>GLDAS-Noah Land Surface Model (<a href="http://disc.sci.gsfc.nasa.gov/hydrology/data-holdings">Rodell et al., 2004</a>)</td>
<td>Goddard Earth Sciences Data and Information Services Center (GES DISC)</td>
<td><a href="http://disc.sci.gsfc.nasa.gov/hydrology/data-holdings">http://disc.sci.gsfc.nasa.gov/hydrology/data-holdings</a></td>
</tr>
<tr>
<td>Modern-Era Retrospective analysis for Research and Applications with test model component (MERRA-Land)</td>
<td>0–2 cm</td>
<td>0.5° x 0.3° (latitude x longitude)</td>
<td>Daily</td>
<td>1979 to present</td>
<td>Modern-Era Retrospective analysis for Research and Application (MERRA) <a href="http://disc.sci.gsfc.nasa.gov/hydrology/data-holdings">Rienecker et al., 2011</a></td>
<td>Goddard Earth Sciences Data and Information Services Center (GES DISC)</td>
<td><a href="http://disc.sci.gsfc.nasa.gov/hydrology/data-holdings">http://disc.sci.gsfc.nasa.gov/hydrology/data-holdings</a></td>
</tr>
</tbody>
</table>

References:
- Albergel, Y., et al., 2013: [HTESSEL](http://apps.ecmwf.int/datasets/) land-surface model. [http://apps.ecmwf.int/datasets/data/ensembles-ecmwf-data/](http://apps.ecmwf.int/datasets/data/ensembles-ecmwf-data/)
used as the Lower Limit (LL), Expected Value (EV), and Upper Limit (UL) of the estimated soil moisture anomaly, respectively. These values are then added to the station-wise long-term mean to obtain the simulated soil moisture and the uncertainty range of estimation.

4.4. Model Performance
Aforementioned statistical measures are also calculated between in situ soil moisture and all the existing soil moisture data sets. Results during the model development period for both bare land and crop land are presented in the Figures S1–S3 (in the supporting information). It is noticed that the proposed model outputs, in terms of $r$, are either the best or second best at 10 out of 24 stations for bare land, and 9 out of 12 for crop land. On the other hand, proposed model outputs, in terms of $nRMSE$ and $d_I$, are found to be the

---

**Figure 3.** Results of sensitivity analysis at monthly anomaly scale for (left) bare land and (right) crop land.
best at 22 and 21 out of 24 stations for bare land, and all 12 stations for crop land, respectively. At stations Bhubaneshwar, Durgapura, and S K Nagar for bare land, the $d_r$ values are found to be negative. The values of $d_r$ are also found to be negative for all the existing soil moisture data sets as well. This is perhaps due to very low variance in the observed (in situ) values of soil moisture data itself. Apart from this, most of the statistical measures for the proposed model are found to be either comparable or better than those for the existing data sets.

The performance of the developed model is tested for temporal consistency during the testing period. However, a few stations, e.g., Kovilpatti, S K Nagar, Tirupati, and Vellayani, are excluded since sufficient data is not available at these stations during this period. The results for both bare land and crop land are presented in the Figures S4–S6 (in the supporting information). Performance measures are obtained for the existing soil moisture data sets as well and also reported in the supporting information Figures S1–S3. For bare land models, correlation coefficient ranges from 0.428 (Bellary) to 0.944 (Vittal), nRMSE from 0.886 (Bellary), 0.121 (Vittal), and $d_r$ from −0.290 (Bellary) to 0.830 (Vittal). Although the values of $r$ between the output of the proposed model and in situ data are found to be better compared to the existing models at six stations, the nRMSE and $d_r$ values are found superior at 16 and 17 out of 20 stations, respectively. For crop land models, correlation coefficient ranges from 0.338 (New Delhi) to 0.829 (Udaipur), nRMSE from 0.345 (New Delhi), 0.177 (Agra), and $d_r$ from −0.463 (Vittal) to 0.737 (Vellanikara). Crop land models are better than existing models at six out of nine stations. Comparing these results, during testing period (shown in the supporting information Figures S4–S6), with the same during the model development period (shown in the supporting information Figures S1–S3), it is found that the temporal consistency of the developed models is assured by the comparable model performance measures during both the periods. At some stations, the performance during the testing period was found to be even better than that during the model development period. This is perhaps due to the presence of lesser number of extreme values during the testing period as compared to those during the model development period. The comparison between the performance of the bare land models and crop land models (top and bottom plots of supporting information Figures S4–S6) reveals that the crop land models perform reasonably, however lose about 25% efficiency on average across different statistical performance index.

Overall, the potential and efficiency of the proposed approach may be appreciated both for bare land and crop land. However, thus far the proposed approach might have been considered to be highly tuned to the in situ data. Thus, spatial transferability of the proposed approach is also checked and presented in the next section.

5. Spatial Transferability of Proposed Approach

As mentioned in section 2.5, the spatial transferability of the proposed approach for soil moisture simulation is assessed through LOO-CV scheme by developing HSG-based models. The ACZ-wise variations of long-
term mean soil moisture for different ACZs for bare land and crop separately are shown in Figure 4. A wide variation is noticed in the mean soil moisture values across all the ACZs—higher during the monsoon months (June–September) and lower during the nonmonsoon months. However, the range is broader for the wetter zones (e.g., Per-humid, Dry subhumid) than the drier zones (e.g., Semiarid, Arid).

The \( p \) values for the KS test (at 5% significance level) are shown from top to bottom in each cell in the supporting information Table S4 for the Logistic, Normal, and Pearson Type III distribution, respectively. According to these results, the Logistic distribution is found to be the best fit probability distribution function for the soil moisture anomaly for all the ACZs. The parameters of the Logistic distribution for soil moisture anomaly are presented in Table 4 for all stations belong to different ACZs. However, in case of other hydro-meteorological inputs, the anomaly values are calculated based on the long-term mean of that variable, observed at the target station only.

Next, the CHM indexes for different are derived using the SPCA technique. The probability distribution function for the HSGs specific CHM index is also found to follow the Normal distribution, similar to station-wise CHM index. Thus, the reduced variates of CHM index are derived using the Normal distribution. This is followed by the calculation of the parameter of the joint distribution between CHM index and soil moisture anomaly using the best fit copula. Following the methodology explained earlier, the Clayton copula is found to be the best fit copula for all the LOO-CV cases (see Table S5 in the supporting information).

The results of HSG-based model performance during the development period (1991–2002) and testing period (2003–2006) are presented in Table 5. Results are expected to be little inferior as compared to the station-wise models since the target station is “ungauged” with respect to the soil moisture data. Performances of the models are found as expected; however, it is still significant for almost all the stations for both bare land and crop land models. The output (50th quantile, i.e., EV), along with the LL and UL (i.e., 90% CI) of the spatially transferable models, is compared against the observed in situ soil moisture. In most of the cases, the observed in situ soil moisture is found to be captured reasonably well by the estimated range. In very few cases (e.g., Bhubaneshwar) estimated range is found to over-estimate the soil moisture for bare land case. However, this is true for all the existing soil moisture data sets (mentioned before) as well.

It is worthwhile to note here that the in situ observed soil moisture value from the previous time step is not considered as an input in the proposed approach. Only the hydrometeorological variables are considered as input. Thus, the efficacy of the proposed approach to extract the hydrometeorological information to model the soil moisture variation shows some merit.

One representative plot (testing period) from each HSG—Bhubaneshwar (HSG-A), Agra (HSG-B), Kalyani (HSG-C), and Niphad (HSG-D) during testing period (2003–2006) for bare land and crop land are separately presented in Figures 5–8, respectively. The performances of HSG-based spatially transferable model and station-wise model are compared during both development and testing period (Table 5 and supporting information Figures S1–S6, respectively). For bare land, the correlation coefficient \( r \) varies between 0.425 (Bellary) and 0.943 (Vittal), \( nRMSE \) varies between 0.551 (Bellary) and 0.119 (Vittal), and \( d \), varies between −0.188 (Pillamedu) and 0.928 (Agra). For crop land, the correlation coefficient \( r \) varies between 0.149 (Bhopal) and 0.851 (Vellanikara), \( nRMSE \) varies between 1.898 (Bhubaneshwar) and 0.185 (Vellanikara), and \( d \),

### Table 4. The Parameters of Best Fitted Probability Distribution (Logistic Distribution) of Monthly Soil Moisture Anomaly for Different Agro-Climatic Classes in Bare Land and Crop Land Areaa

<table>
<thead>
<tr>
<th>Agro-Climatic Class</th>
<th>Parameter “( \alpha )”</th>
<th>Parameter “( \beta )”</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bare Land</td>
<td>Crop Land</td>
</tr>
<tr>
<td>Per-humid</td>
<td>−0.715</td>
<td>−0.715</td>
</tr>
<tr>
<td>Humid</td>
<td>1.300</td>
<td>−0.202</td>
</tr>
<tr>
<td>Moist subhumid</td>
<td>1.262</td>
<td>−0.076</td>
</tr>
<tr>
<td>Dry subhumid</td>
<td>−1.803</td>
<td>−0.180</td>
</tr>
<tr>
<td>Semiarid</td>
<td>−2.159</td>
<td>0.042</td>
</tr>
<tr>
<td>Arid</td>
<td>2.062</td>
<td>−0.145</td>
</tr>
</tbody>
</table>

*The values in bold fonts represent the information for “crop land”.*
A few stations (e.g., S K Nagar, Tirupati, and Vellayani) still reasonably well in terms of wise model. At Bhubaneshwar and Pillamedu, the efficiency of the HSG-based model is inferior, however, Kanke, and Vittal), the performance of the HSG-based model is found to be even better than the station-wise model. At Bhubaneshwar and Pillamedu, the efficiency of the HSG-based model is inferior, however, still reasonably well in terms of $r$ and $nRMSE$. It is further noticed (supporting information Figure S3) that for a few stations (e.g., S K Nagar, Tirupati, and Vellayani) $d$ is found to be negative, despite of having reasonably high $r$ and $nRMSE$. Possible reasons are already explained before in case of station-wise models. Thus, at the very least, it can be concluded that the spatial transferability of the proposed approach for bare land is acceptable for surface soil moisture modeling using the information of HSG, ACZ, and

Table 5. Assessment of Spatial Transferability of HSG-Based Model Using LOO-CV Scheme During Model Development Period (1991–2002) and Testing Period (2003–2006) for Bare Land and Crop Land (Bold Font)

<table>
<thead>
<tr>
<th>HSG Station</th>
<th>Correlation Coefficient ($r$)</th>
<th>Normalized Root Mean Squared Error ($nRMSE$)</th>
<th>Refined Index of Agreement ($d$) in Terms of In Situ Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Development Period</td>
<td>Testing Period</td>
<td>Development Period</td>
<td>Testing Period</td>
</tr>
<tr>
<td><strong>A Bellary</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bhubaneshwar</td>
<td>0.878</td>
<td>0.781</td>
<td>0.168</td>
</tr>
<tr>
<td>Hebbal</td>
<td>0.856</td>
<td>0.558</td>
<td>0.111</td>
</tr>
<tr>
<td><strong>B Agra</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tirupati</td>
<td>0.666</td>
<td>D.N.A.</td>
<td>0.196</td>
</tr>
<tr>
<td>0.273</td>
<td>D.N.A.</td>
<td>0.337</td>
<td>D.N.A.</td>
</tr>
<tr>
<td>0.407</td>
<td>0.834</td>
<td>0.293</td>
<td>0.169</td>
</tr>
<tr>
<td>Durgapur</td>
<td>0.448</td>
<td>0.540</td>
<td>0.256</td>
</tr>
<tr>
<td><strong>C Kalyani</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.439</td>
<td>0.690</td>
<td>0.402</td>
<td>0.549</td>
</tr>
<tr>
<td>Karnal</td>
<td>0.659</td>
<td>0.818</td>
<td>0.248</td>
</tr>
<tr>
<td>New Delhi</td>
<td>0.787</td>
<td>0.761</td>
<td>0.168</td>
</tr>
<tr>
<td>0.354</td>
<td>0.161</td>
<td>0.267</td>
<td>0.387</td>
</tr>
<tr>
<td>Sabour</td>
<td>0.861</td>
<td>0.788</td>
<td>0.193</td>
</tr>
<tr>
<td><strong>D Udaipur</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.677</td>
<td>0.905</td>
<td>0.187</td>
<td>0.189</td>
</tr>
<tr>
<td>0.420</td>
<td>0.652</td>
<td>0.231</td>
<td>0.378</td>
</tr>
<tr>
<td><strong>S K Nagar</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.404</td>
<td>0.719</td>
<td>0.284</td>
<td>0.294</td>
</tr>
<tr>
<td>Vellaniakara</td>
<td>0.313</td>
<td>D.N.A.</td>
<td>0.230</td>
</tr>
<tr>
<td>0.801</td>
<td>0.818</td>
<td>0.164</td>
<td>0.177</td>
</tr>
<tr>
<td>0.873</td>
<td>0.851</td>
<td>0.180</td>
<td>0.185</td>
</tr>
<tr>
<td>0.928</td>
<td>D.N.A.</td>
<td>0.132</td>
<td>D.N.A.</td>
</tr>
<tr>
<td>Vittal</td>
<td>0.930</td>
<td>D.N.A.</td>
<td>0.122</td>
</tr>
<tr>
<td>0.931</td>
<td>0.943</td>
<td>0.130</td>
<td>0.119</td>
</tr>
<tr>
<td>0.950</td>
<td>0.775</td>
<td>0.219</td>
<td>0.281</td>
</tr>
<tr>
<td><strong>D Adhartal</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bhopal</td>
<td>0.722</td>
<td>0.697</td>
<td>0.177</td>
</tr>
<tr>
<td>Nagpur</td>
<td>0.574</td>
<td>0.908</td>
<td>0.334</td>
</tr>
<tr>
<td>Niphad</td>
<td>0.609</td>
<td>0.737</td>
<td>0.207</td>
</tr>
<tr>
<td>0.185</td>
<td>0.511</td>
<td>0.321</td>
<td>0.390</td>
</tr>
<tr>
<td>Pune</td>
<td>0.798</td>
<td>0.840</td>
<td>0.158</td>
</tr>
</tbody>
</table>

*Note: D.N.A. = All Data at same time step is Not Available. The values in italics fonts represent the information for “Testing Period”.

Varies between $-0.643$ (Bhubaneshwar) and $0.781$ (Vellaniakara). It is found that the performance of the HSG-based model for bare land is reasonably well, but not so for crop land. At few stations (e.g., Agra, Kanke, and Vittal), the performance of the HSG-based model is found to be even better than the station-wise model. At Bhubaneshwar and Pillamedu, the efficiency of the HSG-based model is inferior, however, still reasonably well in terms of $r$ and $nRMSE$. It is further noticed (supporting information Figure S3) that for a few stations (e.g., S K Nagar, Tirupati, and Vellayani) $d$ is found to be negative, despite of having reasonably high $r$ and $nRMSE$. Possible reasons are already explained before in case of station-wise models. Thus, at the very least, it can be concluded that the spatial transferability of the proposed approach for bare land is acceptable for surface soil moisture modeling using the information of HSG, ACZ, and
hydrometeorological inputs. However, spatial transferability of the proposed approach for crop land is relatively inferior, possibly due to nonuniform cropping patterns across the stations.

This analysis indicates that the coefficients of CHM index and parameter for the joint distribution (between soil moisture anomaly and CHM index) for any particular HSG can be utilized to any new location within that HSG. The coefficient of CHM index and the parameter for joint distribution in bare land and crop land models for different HSGs are presented in Table 6. If the hydrometeorological input data are available, the developed approach can be applied to simulate the soil moisture at that new location. The HSG will guide to identify the coefficients of CHM index, and ACZ will help to adopt the long-term mean series of soil moisture as well as the probability distribution parameters for the anomalies.

6. Summary and Conclusions

In this study, a probabilistic hydrometeorological approach to model the spatiotemporal variation of monthly soil moisture is proposed. Five different hydrometeorological variables, i.e., evaporation, precipitation, sunshine hours, relative humidity, and temperature are used as inputs, primarily based on their possible physical interaction with soil moisture dynamics. In addition, the information on HSGs and ACZs are utilized to represent the soil-hydrologic properties and the climate regimes, respectively.

Figure 5. Time series plots of volumetric soil moisture (vol %) as output for the proposed HSG-based model with uncertainty range (90% confidence interval) and observed in situ data at Bhubaneswar (HSG-A) during 2003–2006 in bare land and crop land.

Figure 6. Time series plots of volumetric soil moisture (vol %) as output for the proposed HSG-based model with uncertainty range (90% confidence interval) and observed in situ data at Agra (HSG-B) during 2003–2006 in bare land and crop land.
The proposed approach utilizes the strength of the recently developed Supervised PCA (SPCA) technique to develop the CHM index. The developed CHM index represents the collective information of the hydrometeorological forcing. The SPCA ensures the maximum possible in-phase association between the developed CHM index and the soil moisture. This approach also reduces the dimensionality of the set of multiple inputs. The “curse of dimensionality” issue generally hinders the application of multivariate probabilistic model using copulas. Thus, the development of CHM index also eases the utilization of the potential of copulas by reducing the problem to a bivariate case.

The potential of the proposed approach is investigated both for station-wise models and spatially transferable HSG-based models both for bare land and crop land areas. For station-wise models, the anomaly of soil moisture and other hydrometeorological inputs can be obtained by subtracting the long-term mean series observed at that station. The long-term mean series is obtained from historical records available at the station. The proposed model output, for station-wise models, have shown reasonably strong correspondence between model output and in situ observed soil moisture values during model development as well as testing period. The output of the proposed model is also compared to four different existing soil moisture data sets, such as, CPC H96 soil moisture data, ERA-Interim/Land soil moisture data, GLDAS-Noah soil moisture data, and MERRA-Land soil moisture data. It

![Figure 7. Time series plots of volumetric soil moisture (vol %) as output for the proposed HSG-based model with uncertainty range (90% confidence interval) and observed in situ data at Kalyani (HSG-C) during 2003–2006 in bare land and crop land.](image)

Figure 7. Time series plots of volumetric soil moisture (vol %) as output for the proposed HSG-based model with uncertainty range (90% confidence interval) and observed in situ data at Niphad (HSG-D) during 2003–2006 in bare land and crop land.

![Figure 8. Time series plots of volumetric soil moisture (vol %) as output for the proposed HSG-based model with uncertainty range (90% confidence interval) and observed in situ data at Niphad (HSG-D) during 2003–2006 in bare land and crop land.](image)
is noticed that the output of the proposed approach better corresponds to the in situ data as compared to the existing soil moisture data sets. Moreover, the proposed model provides a range of estimation, which is not available with the majority of the existing models.

A sensitivity analysis is also carried out to investigate the sensitivity of the selected inputs. It is found that all the considered hydrometeorological variables are more or less sensitive. Still, in general, it is noticed that the precipitation, evaporation and relative humidity for bare land, and precipitation, sunshine hours for crop land are found to be “more sensitive” than other inputs.

To check the spatial transferability of the proposed approach, the HSG-based models are developed. The HSG specific coefficients of CHM indices are also developed. The HSG-based models are applicable to any target (ungauged) locations, where the historical records of soil moisture data are not available. However, the required hydrometeorological inputs and the information of HSG and ACZ should be known at these target locations. The performance of HSG-based models are found to be little inferior as compared to that of station-wise models. However, it is still significant and indicates the applicability of the proposed approach to a new location. The temporal consistency of the spatially transferable HSG-based models is also found to be reasonably good.

In brief, the potential of the proposed approach is promising to model the spatiotemporal variation of soil moisture. The spatial transferability of the proposed approach ensures its applicability to an ungauged location (“ungauged” with respect to soil measure measurement), knowing the information of HSG, ACZ, and hydrometeorological inputs at that location. However, there are a few limitations of the proposed approach as well. For instance, the accuracy of the model primarily depends on the hydrometeorological inputs. Thus, the quality of these inputs should be checked before using. This is particularly true while using the future modeled climate data to assess the future projection of soil moisture variation using the proposed hydrometeorological approach. Second, the information on HSG and ACZ and the hydrometeorological data should be available at the new “ungauged” station to utilize the potential of spatial transferability of the proposed model. In addition, the model provides the estimate of monthly soil moisture along with a probabilistic range of uncertainty. This might be extremely useful in many application fields. The approach for simulating soil moisture can be adopted in other parts of the globe. Moreover, as stated before, this simulation approach can also be used for the future projection using modeled values of hydrometeorological inputs derived from climate models. This can be considered as the future scope of this study.

### Appendix A : Derivation of Coefficients for Supervised Principal Component Analysis (SPCA)

The Reproducing Kernel Hilbert Space (RKHS) is the point-wise evaluation of a complete, (possibly) infinite-dimensional linear space endowed with an inner product [Aronszjan, 1950]. According to HSIC, the measure of dependence between two random variables, $X$ and $Y$ are associated with their RKHS [Gretton et al., 2005].

As per HSIC, $X$ and $Y$ are independent if and only if all possible bounded continuous functions of these two random variables are independent [Barshan et al., 2011]. For the series of $n$ independent observations, the empirical form of HSIC can be written as:

$$ D^2_{X,Y} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu_X)(Y_i - \mu_Y) $$

where $\mu_X$ and $\mu_Y$ are the means of $X$ and $Y$, respectively.

$$ C^2_{X,Y} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu_X)^2(Y_i - \mu_Y)^2 $$

$$ B^2_{X,Y} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu_X)(Y_i - \mu_Y)^2 $$

$$ A^2_{X,Y} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu_X)^2(Y_i - \mu_Y) $$

### Table 6. The Values of CHM Index Coefficients and Parameter for Joint Distribution for HSG-Based Models for Bare Land and Crop Land (Bold Font)

| CHM Index Coefficients | Parameter for Joint Distribution (|i|) |
|------------------------|-------------------------------|
|                        | Bare Land | Crop Land | Bare Land | Crop Land | Bare Land | Crop Land | Bare Land | Crop Land |
| HSG                    |           |           |           |           |           |           |           |           |
| A                      | −0.496    | −0.486    | 0.394     | 0.391     | 0.516     | 0.514     | −0.413     | −0.459     |
| B                      | −0.526    | −0.496    | 0.434     | 0.375     | 0.499     | 0.517     | −0.102     | −0.468     |
| C                      | −0.454    | −0.484    | 0.367     | 0.360     | 0.532     | 0.533     | −0.476     | −0.458     |
| D                      | −0.516    | −0.471    | 0.370     | 0.455     | 0.502     | 0.540     | −0.492     | −0.518     |
HSIC(Σ, F, G) := (n−1)^−2 tr(KHLH)

(A1)

where F and G are the separable RKHS containing all possible continuous bounded real-valued functions of X and Y, respectively, and kernels H, K, L ∈ R^{n×n}, K_{j} := k(U^j, Y), L_{j} := l(y, j), and H_{j} := I − n^−1ee^T with i is the unit vector of size n×1 and e is one-dimensional vector of all ones having length n. Then the trace of the reduced matrix [KHLH] is maximized to get the principal component having maximum association with the target output (Y). Now, using the property of cyclic permutation of trace of a product, tr(KHLH) can be rewritten as tr(HKHL). Again, the kernels K and L can be expressed as [X^TUU^TX] and [Y^TY], respectively. Thus, tr(HKHL) can be expanded as:

tr(HKHL) = tr((X^TU)^2HH^T)

(A2)

Now, tr((X^TU)^2HH^T) can be rearranged as tr(U^2XHLHX^T) and to solve the orthogonal transformation matrix U, the optimization problem is defined as,

\[
\arg \max_U \quad \text{tr}(U^2XHLHX^T), \quad \text{subject to} : \quad U^TU = I
\]

Further, the symmetric real matrix, Q = XHLHX^T of size p×p, have p eigenvalues (λ₁ ≤ ⋯ ≤ λₚ) and corresponding eigenvectors [v₁, ⋯, vₚ], each having p elements. In general, the maximum value of the cost function is λ_p + λ_p−1 + ⋯ + λ_p−d+1 and the optimum solution is \(U = [v₁, v_p−1, ⋯, v_p−d+1]\), where d denotes the dimension of \([X^TU]\). In our case, soil moisture being the only target variable, d = 1, i.e., \([U^TX]\) is a one-dimensional vector. Thus, \(U = [v₁]\), which yields the coefficients for p different hydrometeorological variables ensuring the product to be best associated with the target variable.

Acknowledgments
This work was partially supported by the Ministry of Earth Science (MoES), Government of India, through sponsored Project No.: MoES/PAMC/HSC/30/2013-PC-II. Authors also acknowledge the thoughtful comments of Heye R. Bogena that helped a lot to improve the manuscript. The weekly in situ soil moisture data and daily observed hydrometeorological data (evaporation, precipitation, relative humidity, sunshine hours, and temperature) are purchased from the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, through their web site at http://www.esrl.noaa.gov/psd/. The ERA-Interim soil moisture data are obtained from the ECMWF, Reading, London, UK, through their website at http://apps.ecmwf.int/data set/). The GLDAS_NOAH025_M soil moisture data are obtained from the Goddard Earth Sciences Data and Information Services Center (GES DISC) through their website at http://disc.sci.gsfc.nasa.gov/hydrology/data/hold ings. The MERIS soil moisture data are obtained from the GES DISC through their website at http://disc.sci.gsfc.nasa.gov/daac-bin/FTPSubset.pl.

References
Agrawat (2012), Agroclimatic Classification: 1. Based on Moisture Index (Thronthwaite), Pune, India. [Available at http://www.imdagrimet.gov.in/nioin287/]


IMD (2009), Specifications for Surface Meteorological Instruments, Pune, India. [Available at http://www.impdpune.gov.in/surface_instruments/Specs%20of%20SurfaceInsrtuments.pdf.]


