Journal of the Indian Society of Remote Sensing (October 2019) 47(10):1671–1682
https://doi.org/10.1007/s12524-019-01015-4

RESEARCH ARTICLE

Soil Moisture Retrieval Using Quad-Polarized SAR Data from Radar Imaging Satellite 1 (RISAT1) Through Artificial Intelligence-Based Soft Computing Techniques

Manali Pal[1] · Rajib Maity[1]

Received: 18 July 2018 / Accepted: 6 June 2019 / Published online: 27 June 2019

Abstract
The present study aims to explore and compare the potential of different Artificial Intelligence-based Soft Computing (AISC) techniques to prepare surface Soil Moisture Content (SMC) map using fine-resolution ($\sim$ 5 m), quad-polarized Synthetic Aperture Radar (SAR) data obtained from Radar Imaging Satellite 1 (RISAT1). Potential of three different AISC techniques, i.e. Support Vector Machine (SVM), Random Forest (RF) and Genetic Programming (GP), is explored. The estimated surface SMC is validated with the field soil moisture values in both bare and vegetated lands ($<$ 30 cm height). Different techniques have their own merits and demerits; however, we recommend GP to be most useful due to its other features. For example, GP provides the mathematical relationship, importance and sensitivity of each individual input to the surface SMC. This helps us to quantify the contribution of quad-polarized backscattering coefficients and soil texture information. It is noticed that the use of only SAR data without soil texture information may be acceptable with reasonable accuracy with an enormous benefit of its applicability to the locations without soil texture information. Using this, an exemplary fine-resolution ($\sim$ 5 m) SMC map is developed. Such high-resolution maps for large spatial extent are expected to be highly useful in many applications.

Keywords Soil moisture · Remote sensing · Synthetic Aperture Radar (SAR) · Quad-polarized data · Radar Imaging Satellite 1 · Artificial Intelligence-based Soft Computing (AISC) techniques

Introduction

Artificial Intelligence-based Soft Computing (AISC) techniques are potential due to their feasibility to address the problems having significant number of observation data without a complete knowledge of theoretical background. In AISC techniques, a comprehensive training data set of examples is constructed covering as much of the system parameter space as possible. Typically, a random subset of the data is put aside for a completely independent validation. These AISC techniques have shown promise in handling large amount of deviation and noise hidden in data sets. Utilizing such properties of AISC technique-based approaches, it may be potential to use in retrieving surface SMC from backscattering data for different Land Use Land Cover (LULC) and surface roughness conditions (Paloscia et al. 2013) and studying the physical processes influencing the SMC generally represented as nonlinear functions (Coleman and Niemann 2012; Espinoza-Dávalos et al. 2016) to improve the retrieval algorithms.

There is a plethora of space-borne active and passive microwave sensors which have been deployed to provide useful global-scale surface soil moisture estimates. Examples include Tropical Rainfall Measuring Mission (TMI), the Scanning Multichannel Microwave Radiometer (SMMR), the Special Sensor Microwave/Imager (SSM/I), the WindSAT mission, the Advanced Microwave Scanning Radiometer–Earth Observing System (AMSR-E), the

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s12524-019-01015-4) contains supplementary material, which is available to authorized users.

✉ Rajib Maity
rajib@civil.iitkgp.ac.in; rajibmaity@gmail.com

[1] Department of Civil Engineering, Indian Institute of Technology Kharagpur, Kharagpur 721302, India

Advanced Microwave Scanning Radiometer2 (AMSR2) mission, the Soil Moisture Ocean Salinity (SMOS) mission and the Soil Moisture Active Passive (SMAP) mission (Karthikeyan et al. 2017). The passive and active microwave soil moisture products have been merged under the European Space Agency (ESA) Climate Change Initiative (CCI) program to produce a long-term global-scale SMC record (Liu et al. 2012). However, these soil moisture missions provide data at very low spatial resolution ($\sim$ km range) due to sensor limitations. Additionally, most of the retrieval algorithms are influenced by the instrument characteristics; sensor frequency and spatial heterogeneity of the SMC at sub-satellite grid resolution which are poorly characterized.

Indian Space Research Organization (ISRO) has launched Radar Imaging Satellite 1 (RISAT1), on April 26, 2012. It uses a multi-mode Synthetic Aperture Radar (SAR) at C-band (5.35 GHz). Its spatial and temporal resolutions are 3–50 m and 25 days, respectively. The quad-polarized backscattering coefficients are acquired from RISAT1 (henceforth RISAT1 data). The polarizations of incident and backscattered radar signal are either horizontal (H) or vertical (V). The outcomes are co-polarized (HH or VV) and cross-polarized (HV or VH) backscattering coefficients. It is indeed true that there could be certain limitations while producing very high resolution SMC maps, and it may require high computing power along with advanced techniques. Hence, based on the problem in hand, the study attempts to explore the potential of AISC techniques to develop high-resolution soil moisture maps utilizing the available RISAT1 data. Also, in our knowledge, only a few of the existing missions produce a soil moisture map of spatial resolution in the range of a few metres using all four polarizations, i.e. HH, HV, VH and VV of SAR data (Pal et al. 2017). However, exploring the potential of the AISC techniques to develop a fine-resolution SMC map using the quad-polarized SAR data is still exceptional and is the focus of the present study.

The Artificial Neural Network (ANN), Genetic Programming (GP) and Support Vector Machines (SVMs) are some of the AISC techniques that are extensively used in different fields of hydrology (Unnikrishnan and Jothiprakash 2017; Li et al. 2018). The SVM is applied for predicting soil moisture from SAR data showing a reasonably good match with the observed soil moisture data (Gill et al. 2006; Pasolli et al. 2015). The SVM model was also found to perform better than ANN (Ahmad et al. 2010). Despite the good performances of SVM and many of the other AISC techniques, they are considered as black-box models, i.e. they are not capable of revealing physical processes. However, GP can be considered as a relatively better method to deal with this issue. It is due to the feature of GP that they can produce mathematical relationships between inputs and outputs without a need of pre-defined form of any relationship (Maity and Kashid 2011; Alavi and Gandomi 2011). Application of GP technique in soil moisture estimation from SAR data and meteorological data was found potential in earlier studies (Makkeasorn et al. 2009; Elshorbagy and El-Baroudy 2009). However, it did not provide the prediction equations which are considered as the principal advantage of GP over other AISC techniques in the current study.

Random Forest (RF) is a nonlinear modelling tool that provides estimates regarding the hierarchy of variables in the classification, and thus is able to estimate contribution of each index to the total risk (Breiman 2001). The RF algorithm has been applied to different fields of studies. A number of theoretical and empirical studies have detailed many advantages of RF, including high forecast accuracy, acceptable tolerance to outliers and noise, and easy avoidance of overfitting problems. Based on this body of knowledge, RF may be useful for surface SMC retrieval from SAR data and able to address the multi-variate and nonlinear issues.

The above discussion on the utility of different AISC techniques motivates the present study to explore their potential in preparing large-scale fine-resolution surface SMC map without the priori information of on-ground soil information (e.g. soil texture and soil roughness) and utilizing only the quad-polarized SAR data. The above discussion also explains that SVM and GP have been already used in this field of study with reasonably good performances. However, it is also discussed that the accuracies of these models highly depend on more number of supplementary information. Though the AISC techniques are generally black-box models, GP can provide the predictive equations which can be advantageous to explore the predictor–predictand relationship. In this study, the predictive equations obtained from GP are explored and analysed. Also, the sensitivity analysis and evaluation of importance (discussed in subsequent sections in detail) of input variables allow the study to quantify the contribution of each input in SMC estimation from SAR data. Finally, keeping in mind the importance of soil moisture maps, having information at a very fine resolution in climate modelling as well as hydrological and agricultural modelling, the study aims to prepare a SMC map for the study area using the best-performing AISC technique. Therefore, in a nutshell, the objective of this study is to explore the efficacies of different AISC techniques, namely SVM, RF and GP for the estimation of surface SMC using quad-polarized SAR data, and to develop a SMC map of the study area with the best-performing technique. The target is to avoid the use of ancillary information (e.g. soil texture and roughness) as much as possible. Some information might be unavoidable and some may be preferable. It also aims to explore the

mathematical relationship to associate as well as quantify the contribution of the input variables to the target variable which is one of the principal gaps in the previous studies.

Study Area

The study area of the present study is defined by boundaries of the satellite images acquired. The RISAT1 data are obtained for the study area on 25 and 27 October 2014. The image of 25th October is enclosed by the following four coordinates, 22°1′50.37″N × 87°13′51.26″E; 22°4′6.1104″N × 87°26′9.57″E; 22°21′21.94″N × 87°22 30.62″E; 22°19′6.77″N × 87°10′14.06″E, and the area is calculated to be 21.5 km × 32.6 km. Similarly, the image of 27th October is enclosed by the following four coordinates, 22°4′16.84″N × 87°29′45.13″E; 22°1′26.3892″N × 87°16′2.86″E; 22°18′39.04″N × 87°11′53.35″E; 22°21′30. 2364″N × 87°25′39.39″E and the area is calculated to be 24.1 km × 32.68 km.

The collective areas within the boundaries of the images for both the days predominantly consist of agricultural lands with a few patches of establishments and forested areas. The climatic condition of the study area is tropical where the average temperatures in summer and winter are 30 °C and 22 °C, respectively. It experiences an average rainfall of 1140 mm during the monsoon. Figure 1 shows the study area with the ground sampling points.

Data

Satellite Data

Two Single-Look Complex (SLC) satellite images (RISAT1 data) are procured for the 2 days, i.e. 25 October 2014 and 27 October 2014, in Fine Resolution Stripmap Mode 2 (FRS-2) mode from National Remote Sensing Centre (NRSC), Hyderabad, with a resolution of ∼ 20 m. Table 1 represents the specifications of the procured satellite images.

For each image, the digital numbers are provided by NRSC for which the effect of the 'speckle' due to multiple within-pixel scattering objects is removed. The speckle-filtered digital numbers DN_p are converted to backscattering coefficients expressed in decibel (dB). The conversion uses the SAR calibration coefficients for each linear polarization in the equation provided in the RISAT1 data products by Space Application Centre (SAC) (SAC 2015) which can be represented by the following form,

$$\sigma_0 = 20\log_{10}(\mathrm{DN}_p) - K_{\mathrm{dB}} + 10\log_{10}\frac{\sin i_p}{\sin i_c}, \quad (1)$$

where σ_0 is the radar backscatter coefficient in dB, DN_p is the digital number grey-level count for the pixel p, K_{dB} is the calibration constant, i_p is the incidence angle at pixel position p, i_c is the incidence angle at scene centre.

Incidence Angle Normalization

The RISAT1 data correspond to two different incidence angles for two different dates as shown in Table 1. Since the backscattering values highly vary with the incidence angles of the sensor, it is not possible to combine the backscattering values of the two dates to investigate the association with the surface SMC. Hence, to study the sensitivity of these backscattering values for the two different dates to the surface SMC, the measurements should be normalized to a reference incidence angle. The present study normalizes the incidence angles of the concerned two dates to a reference angle of 30° using the library realization prepared by Zribi et al. (2005). The library realization consists of backscattering values and incidence angles corresponding to a large range of surface roughness condition. The incidence angles at each pixel are normalized to the reference angle. This method of incidence angle normalization was successfully used by Pal et al. (2017) to estimate the probabilistic surface SMC using copula from SAR data.

Field Data

The soil samples within the top 5 cm of the surface are collected from 375 monitoring points within the areas of the acquired images. The ground data are collected from bare and vegetated lands having < 30 cm vegetation height. Out of 375 monitoring points, 320 data points are from vegetated land areas and rest (55 data points) are from bareland areas. It is worthwhile to mention two points here. Firstly, the soil samples are collected within ± 1 h of satellite visit to the study area (at 5:00 pm IST) in order to assure the accuracy of the ground data. Almost no changes in SMC owing to soil humidity change, between time of satellite passing and the time of soil sample collection, are assured through collecting the soil sample as close to the time of satellite passing as possible. Secondly, the soil samples are collected from only barelands and vegetated lands having < 30 cm vegetation height since the penetration depth of RISAT1 is not more than 30 cm of vegetation height. It restricts the model to be applied only for less than 30 cm vegetation height. Though it is not possible to cover the entire image area (∼ 32 km × 24 km) within a short period of time of satellite passing over the region, the selection of sampling points ensures a wide range of soil moisture content, soil roughness and soil texture.

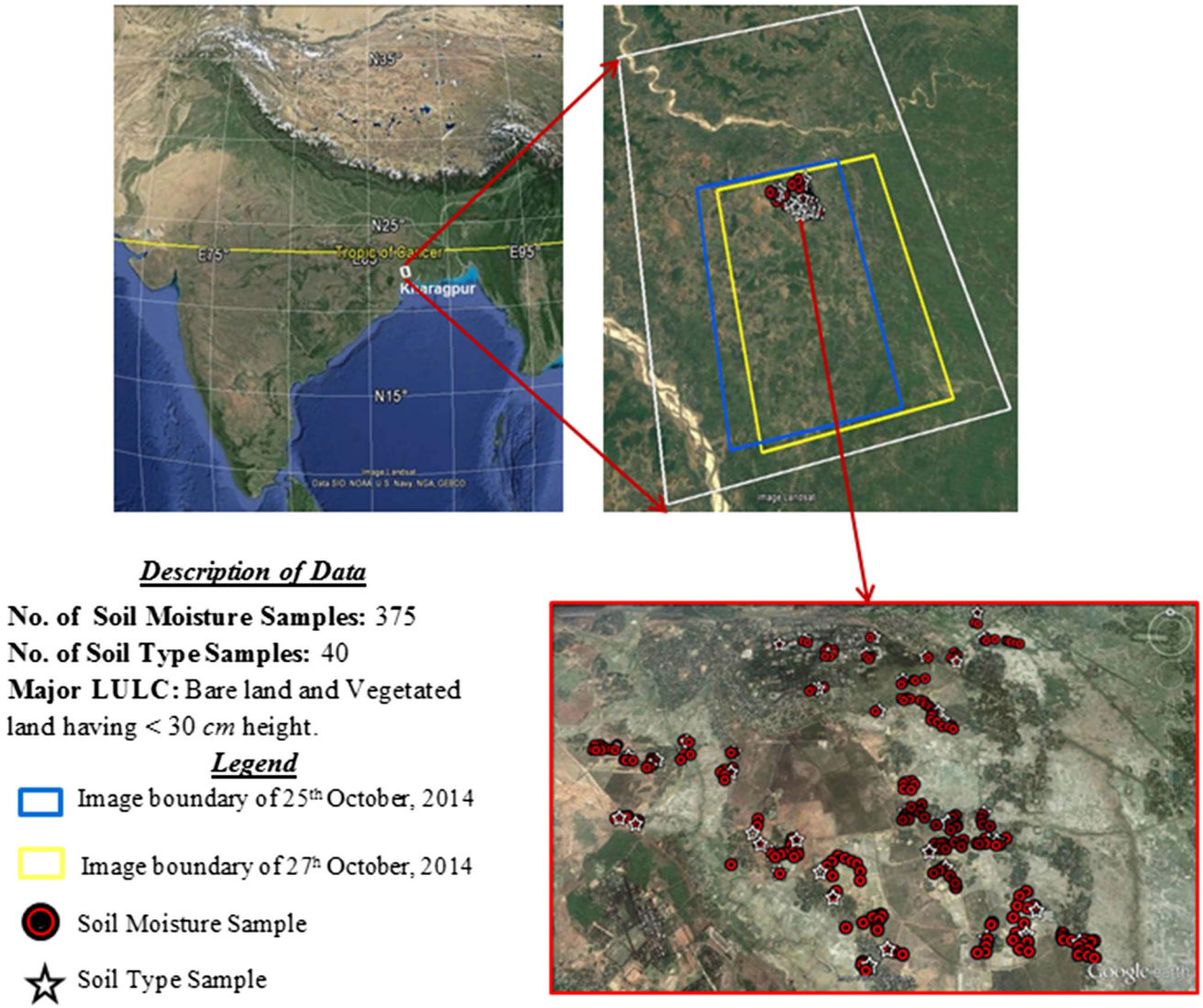

Fig. 1 The study area showing two images and sampling point distribution

Table 1 The description of incidence angles and scene centres of satellite images for the two dates

Date of passing (RISAT1)	Incidence angle	Scene centre	
		Latitude	Longitude
October 25, 2014	14.25039°	22.194216 °N	87.307328 °E
October 27, 2014	27.03315°	22.190772 °N	87.345272 °E

These ranges are expected to well represent the entire study region. The volumetric SMC and the soil texture information of the collected soil samples are analysed in the laboratory using Gravimetric Method (IS:2720, Part-2, 1973). A brief description of the experimental results is given in Table 2 where the soil texture of the study area is observed to be mainly sandy and silty.

The inability of plants to excerpt water below wilting point is an important concern in agricultural applications as well as the hydrological modelling. The present study deals with the SMC extracted by the plants above wilting point since it provides better correlation coefficient and lesser error (Pal et al. 2017). The water that is essentially available to the plant is the difference between the observed SMC and the wilting point, i.e. the moisture at 15 bar pressure. The detail description of computing the SMC at 15 bar can be found in Pal et al. (2017). Thus, the SMC mentioned in this paper refers to the SMC above the wilting point or available SMC hereafter.

Table 2 The volumetric SMC and soil texture description obtained from the experimental results of the study area for bare and vegetated land areas

Type of LULC	Sampling points	v/v SMC range	%Gravel	%Sand	%Silt	%Clay
Bare	55	1.48–39.19	0–20.34	37.16–67.07	17.54–53.12	3.87–15.74
Vegetation	320	2.14–73.59	0–20.34	7.62–73.01	17.54–65.74	0.46–26.61

Methodology

The study has explored the data-driven AISC techniques, namely SVM, RF and GP to estimate the surface SMC from RISAT1 data. In this study, the individual performances of the three approaches have been investigated with different sets of input combinations using backscattering coefficients (HH, VV, HV and VH) and the soil texture information, i.e. the experimental values of percentages of sand, silt and clay. To study the effect of vegetation on SAR data-based surface SMC estimation and to consider the effect of all backscattering coefficients, the Radar Vegetation Index (RVI) is used as one of the inputs. The RVI is expressed as the following (Kim et al. 2012),

$$\mathrm{RVI} = \frac{8\sigma_{\mathrm{HV}}}{\sigma_{\mathrm{HH}} + \sigma_{\mathrm{VV}} + 2\sigma_{\mathrm{HV}}} \tag{2}$$

The range of RVI is 0 to 1. For bareland, the value of RVI is near 0 and it increases with the crop growth (Kim et al. 2012). Table 3 illustrates the different cases using various input combinations used for the three approaches.

The following section describes the mathematical background of the three approaches. The SVM and RF are applied through available packages of R-studio. The GP is run using a software 'Eureqa' version 0.98 (Web source: https://www.eureqa.com/). The software uses the GP-based symbolic regression to provide a mathematical relationship between the input and output variables.

Table 3 Different input combinations for each model (refered as different cases in the text)

Cases	Input combination
1	HH, HV, VH, VV, sand, silt, clay
2	HV, VH, VV, sand, silt, clay
3	HH, VV, HV, sand, silt, clay
4	HH, VH, HV, sand, silt, clay
5	HH, VH, VV, sand, silt, clay
6	HH, HV, VH, VV, silt, clay
7	HH, HV, VH, VV, sand, clay
8	HH, HV, VH, VV, sand, silt
9	HH, RVI
10	HH, RVI, sand, silt, clay
11	HH, HV, VH, VV
12	HH
13	HH, sand, silt, clay

Support Vector Machine (SVM)

In SVM, the ultimate goal is to find a functional dependency, $f(x)$ between independent variables $\{x_1, x_2, \ldots x_L\}$ obtained from $x \in R^K$. In the present study, the independent variables are the eight different input variables, i.e. HH, HV, VH, VV, RVI, %sand, %silt and %clay. The output (dependent which is the volumetric SMC in the present study) $\{y_1, y_2, \ldots y_L\}$ is obtained from $y \in R$ selected from a set of L independent and identically distributed (i.i.d) observations. The functional dependency is given by $f(x) = \langle w, x \rangle + b$, where $\langle w, x \rangle$ denotes the dot product of a weighting vector w and input vector x; and b is the bias. For this purpose, the original input domain is mapped onto a higher dimensionality space, where the function underlying the data is assumed to be linear. The optimal linear function in the transformed space is identified by solving an optimization problem, which is the combination of the training error (empirical risk) and the model complexity (confidence term) through a regularization parameter C:

$$\begin{aligned} &\text{Minimize} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{L}\left(\xi_i + \xi_i^*\right) \\ &\text{Subject to} \quad \begin{cases} Y_i - \sum_{j=1}^{K}\sum_{i=1}^{L} w_j x_{ji} - b \le \varepsilon + \xi_i, \\ \sum_{j=1}^{K}\sum_{i=1}^{L} w_j x_{ji} - y_i \le \varepsilon + \xi_i^*, \\ \xi_i, \xi_i^* \ge 0, \end{cases} \end{aligned} \tag{3}$$

where ε is the Vapnik's insensitive loss function; C is the capacity parameter cost; and ξ_i and ξ_i^* are called the slack variables which measure the distance (in the target space) of the training samples lying outside the ε-insensitive tube from the tube itself (Han et al. 2007).

In a nutshell, the phases involved in SVM modelling can be described as: (1) selection of an appropriate kernel function and kernel parameter (kernel width—γ), (2) designating the 'ε' insensitive parameter and (3) defining the capacity parameter cost, 'C'.

In the present study, the Radial Basis Function (RBF) kernel is used in the soil moisture estimation from SAR data which has been proven the best kernel function for nonlinear regressions (Han et al. 2007) using the 'R' package. Mathematically, the RBF can be defined as,

$$k(x_i, x) = \exp\left(-\gamma \|x - x_i\|^2\right) \quad \gamma > 0 \tag{4}$$

Genetic Programming (GP)

Genetic Programming (GP) (Koza 1992) attempts to find some mathematical relationships among the input (the backscattering coefficients and soil texture information in this study) and target variables (the surface SMC in the study). In the GP technique, the model space is constructed over a defined set of parameters using linear, exponential, logical, trigonometric and other mathematical operators. An optimal set of covariates should be provided to the GP since the dimensionality of the model space increases with the number of covariates determined. For a specific objective function, the GP algorithm moves over the model space, selects a model, and evaluates its parameters by means of symbolic regression and variable pairing techniques (Schmidt and Lipson 2009; El-Baroudy et al. 2010).

The GP also aims to evaluate the sensitivity of each input variable on the surface SMC variation. Assume $z = f(x, y, \ldots)$, the influence metrics of x on z are defined as the sensitivity and the sensitivity (Sn) is calculated at all input data points as follows,

$$\mathrm{Sn} = \overline{\left|\frac{\partial z}{\partial x}\right|} \bullet \frac{\sigma(x)}{\sigma(z)}, \tag{5}$$

where $\frac{\partial z}{\partial x}$ is partial derivative of z with respect to x; $\sigma(x)$ is the standard deviation of x in the input data; $\sigma(z)$ is the standard deviation of z; $\left|\frac{\partial z}{\partial x}\right|$ represents the absolute value of $\frac{\partial z}{\partial x}$ and $\overline{\frac{\partial z}{\partial x}}$ is the mean of $\frac{\partial z}{\partial x}$. Sn may be positive where $\frac{\partial z}{\partial x} > 0$ and negative where $\frac{\partial z}{\partial x} < 0$. Positive and negative percentages are defined by the percentage of data points where $\frac{\partial z}{\partial x} > 0$ and $\frac{\partial z}{\partial x} < 0$, respectively (Schmidt and Lipson 2009).

Random Forest (RF)

The RF is an ensemble regression tree method which constructs each tree using a different bootstrap sample of the data (Breiman 2001). The decision (classification or regression) trees are constructed by recursively splitting the data based on covariates. The observations in a node become progressively pure (in terms of outcome) as data move from the root node to terminal nodes. Every terminal node is summarized by the average outcome value of all the observations that culminate in there. While the trees are being constructed in random forest, only random subset of covariates at each internal node is assessed to reach the best split. The averaging process of RF enhances the prediction accuracy over a single tree. Moreover, the method is generally robust against overfitting (Breiman 2001). It has three parameters to predefine:

(1) Number of predictor variables in the random subset at each node, denoted as m_{try},
(2) Number of trees to be grown in the forest, denoted as n_{tree},
(3) Minimum size of terminal nodes, i.e. the *nodesize*.

The importance of each input variable in estimating the target variable can be assessed in RF by Gini impurity index, which is a standard decision-tree splitting metric which can be mathematically described as follows (Rutkowski et al. 2015),

$$g(S) = 1 - \sum_{k=1}^{K} (p_k(S))^2, \tag{6}$$

where $g(S)$ is the impurity measure, S is the set of data elements where the number of different classes is denoted as K; and $p_k(S)$ is the ratio of number of elements present in the set and number of elements of the set from the kth class.

Results and Discussion

The k-fold cross-validation method is used for evaluating the generalized model performance to an independent data set and avoiding overfitting. This cross-validation technique randomly splits a data set into k-separate folds with equal or almost equal data length, and each fold is in turn used to test the model developed from the remaining $(k - 1)$ folds. In our study, the entire data set is split into three different folds where each fold contains 40 and 200 data points randomly selected from bare and vegetated land areas, respectively, from the entire pool of 375 data points. Remaining data points (15 from bareland and 120 from vegetated land areas) are used for the testing of the models.

The performances of the selected models during the training and testing periods are evaluated by different performance metrics, namely Correlation Coefficient (CC), Refined Degree of Agreement (D_r) and the Root Mean Square Error (RMSE) for all the 13 cases (Table 3). Initially, the models have been developed independently for bare and vegetated land areas for all the 13 cases using three different AISC methodological approaches. The independently developed models for each case are applied to bare and vegetated land areas, and the results are evaluated through performances of the models both in the

training and testing periods. The model performances are assessed in terms of the mean values of the performance metrics across the three folds.

Determination of Model Parameters

For the modelling of SVM, the kernel functions or kernel parameters, namely kernel width (γ), insensitive parameter (ε) and capacity parameter cost (C), are selected for bare and vegetated land areas for each of the 13 cases and three different folds. The SVM models are tuned by varying the values of these three kernel parameters from 0.01 to 1 with 0.05 increments. The best model parameters are selected based on the minimum Mean Squared Error (MSE) values. The values of these kernel parameters for the three folds of 13 cases for bare and vegetated land areas are provided in Table S1 and Table S2, respectively, in supplementary document.

For each case of the bare and vegetated land areas, a model is developed by using the GP technique to estimate the surface SMC from the set of input variables. The software 'Eureqa' uses the GP-based symbolic regression to provide a mathematical relationship between the input and output variables. The approach includes the basic mathematical and exponential functions in search process and minimizes the Mean Absolute Error (MAE) to define the objective function to make the search algorithm to be robust to the outliers present in the data set. This technique is robust to the multicollinearity problem. Moreover, the surface SMC is accurately modelled by an optimal number of input variables by developing parsimonious models. The solutions for the estimation of surface SMC in terms of the input variables determined by the Eureqa are given in Table S3 and Table S4, respectively, in supplementary document for bare and vegetated land areas, respectively. Eureqa generates many ($\sim 10^5$ to 10^6) predictive equations for the target variable. Among these, the best model is selected based on minimum error measures on validation data shown in Eureqa. The error measures are MSE and MAE.

In case of modelling with the RF, the parameters, namely m_{try}, n_{tree} and *nodesize*, need to be pre-defined. The study observes that the outcome of the RF model becomes stable after 500 trees. Hence, keeping the n_{tree} constant as 500 for each fold of each case both for bare and vegetated land areas, the parameters m_{try} and *nodesize* are selected for bare and vegetated land areas. The RF models are tuned by varying the values of m_{try} from 1 to 4 with 1 increment for each of the 13 cases and three folds, whereas the values of *nodesize* are varied from 5 to 40 and 5 to 200 for bareland are vegetated land areas, respectively, for all three folds of 13 cases. The best model parameters are selected based on the minimum MSE values. The values of these RF model parameters are shown in Table S5 in supplementary document.

Relative Importance and Sensitivity of Different Input Variables

Individual contributions of different input (predictor) variables are determined in RF modelling by Gini impurity index and in GP by sensitivity analysis. The importance values of each input variables obtained in RF modelling for all 13 cases and across the three folds are provided in Table S6 and Table S7 in supplementary document for bare and vegetated land areas, respectively. The comparison of importance values of all the input variables indicates the significant dominance of HH for estimating surface SMC both for bare and vegetated land areas. The ranges of importance values of HH for bareland for the three folds are 0.020–0.128, 0.048–0.151 and 0.015–0.063, respectively. The importance values corresponding to remaining variables except RVI, i.e. VV, VH, HV, sand, silt and clay, range from 0.002 to 0.066 across the three folds for all the 13 cases which are visibly lesser than HH importance value. However, it is observed in cases 9 and 10 that the RVI contribution is also significant, i.e. (0.070–0.091, 0.022–0.044 and 0.014–0.020 for three folds) when compared with HH. Additionally, it is observed that at the omission of HH, the silt, sand, HV and VV contribute significantly in estimation of surface SMC. The trend of observing visible dominance of HH also persists for the vegetated land areas where the ranges of magnitude of importance values corresponding to HH are 0.265–0.919, 0.549–1.242 and 0.924–1.477 for the three folds. Whereas the importance values of VV, VH, HV, sand, silt and clay range from 0.010 to 0.418 across all the three folds and 13 cases. The importance values of RVI range from 0.009 to 0.120, 0.005–0.181 and 0.200–0.586 across the three folds for the vegetated land areas. The sand, silt and VH are observed to have higher magnitude of importance values at the omission of HH for the vegetated land areas.

The sensitivity of the surface SMC to the different input variables for all the cases of bareland and vegetated land areas are presented in the GP analysis. The observation clearly shows that the surface SMC is most sensitive to the HH for bare land areas. The HH always shows (100% of time) positive sensitivity, i.e. the increase in the values of HH results in the increase in the target variable. The 100% positive sensitivity of HH to estimate surface SMC is observed for all the 12 cases (except case 2) for both the bare and vegetated land areas. The magnitude of positive sensitivity of HH is 1 for all the 12 cases. For case 2, where HH is emitted from the input variable combination, the %silt is showing 100% positive sensitivity to the surface SMC with a positive magnitude of 1 for all the three folds.

For the vegetated land areas, the significant association of HH to the surface SMC is confirmed for all the cases

where the HH has shown 100% positive sensitivity across all the three folds with sensitivity magnitude of 1. For the second case when the HH is eliminated from the input combination, the %sand shows the 100% positive sensitivity with the positive magnitude 1 for the first and second folds. However, for the third fold, VV shows 100% positive sensitivity with the magnitude of sensitivity of 0.853, while the %sand shows the 100% negative sensitivity with the negative magnitude of 0.474.

The prediction equations obtained from GP for bare and vegetated land areas, shown in Tables S3 and S4, provide the same conclusion as the above discussion that the HH is the most significant variable to determine the surface SMC. The occurrence of only HH across all the models with an optimum model performance and the worst model performance at the exclusion of HH establish the evident importance of HH for both the bare and vegetated land areas. For bareland areas, the occurrence of percentages of sand and silt and VV is observed in more complex predictive equations. It is also observed that among the soil texture information, silt is the most contributing variable based on its occurrence in more number of predictive equations for all the cases, better model performance at its presence and occurrence along with HH in most cases at higher model complexity. As mentioned earlier, the study area mainly consists of silt and sand content with a very less amount of clay content. Hence, it can be concluded that for bareland areas, the soil texture information plays an important role in surface SMC retrieval from SAR data with higher model complexity. For the vegetated land areas, the major soil components of the study area play a crucial role to estimate the surface SMC only at the omission of HH. However, it is also observed that the presence of soil texture information does not show a significant improvement in model performance which is almost comparable with the performance with only backscattering coefficients. Consequently, to select the parsimonious models, the predictive equations consisting of only HH (except case 2) are chosen to estimate the surface SMC using the GP technique.

Similar to the bareland areas, the dominance of HH in determining the surface SMC is evident for vegetated land areas. It has been observed that the predictive equations selected for all 12 cases except case 2 (where the HH is omitted from the input combination) consist of only HH component. This finding of major contribution of HH, for both the bare and vegetated land areas, can be validated with the findings of some previous established studies stating the more dominance of HH in surface soil moisture estimation from SAR data than the other polarizations (Kornelsen and Coulibaly 2013) making the present study more reliable. Moreover, the ability to quantify the contribution of each input variable makes the present study more relevant than the previous studies. Thus, the results indicate the applicability and reliability of the models in the ungauged locations where soil texture information is unavailable but only SAR data are available. Along with HH and soil texture information, the contributions of HV and VV are observed in the predictive equations with higher complexity for estimation of surface SMC for the vegetated land areas. The study also uses RVI along with HH and the soil texture information to study the effect of vegetation (< 30 cm in height) on SMC retrieval. It can be seen from Tables S3 and S4 that the RVI does not occur at all for both the bare and vegetated land areas although, as discussed before, the RVI occurs in the predictive equations for both the bare and vegetated land areas with increasing model complexity without substantial improvement in model performances. In a nutshell, the study of importance values and sensitivity magnitudes of all the input variables suggest the significant contribution of HH to SMC estimation with and without soil texture information for both the bare and vegetated land areas. Also, the influence of predominant soil texture type, when HH is not in use, is demonstrated for both the bare and vegetated land areas.

Model Performances

The comparison of mean values across the three folds of the performance metrics of three methods used, for training and testing periods for all the 13 cases, is shown in supplementary document in Figures S1 and S2 combined for bare and vegetated land areas due to less umber of data points for bareland area.

The ranges of CC, Dr and RMSE are 0.669–0.723, 0.470–0.491 and 0.114–0.120 for SVM; 0.417–0.672, 0.570–0.659 and 0.091–0.112 for GP; and 0.568–0.612, 0.611–0.635 and 0.096–0.100 for RF across all the 13 cases during the training period. Thus, the values of performance metrics show comparable model performances for all three models used although GP is observed to perform better in terms of higher Dr and lower RMSE almost for all the case except input variable combination case 2 (HH omitted) and case 5 (HV omitted). The performance metrics range for SVM, GP and RF during the testing period are as follows—CC: 0.688–0.747, Dr: 0.495–0.515, RMSE: 0.109–0.116; CC: 0.488–0.687, Dr: 0.590–0.669, RMSE: 0.086–0.103, and CC: 0.624–0.669, Dr: 0.646–0.666 and RMSE: 0.088–0.092, respectively. Here also, the GP shows a better model performance in terms of Dr and RMSE for all the cases except case 2 and case 5.

The comparison of model performances across the 13 cases shows that worst model performances are obtained for case 2 when HH is omitted from the input variable combination. It is true for GP and RF during both training

and testing periods. The deduction applies both for bare and vegetated land areas. However, for SVM, performance is nearest to worst for case 2 (HH is excluded), whereas it shows the worst performance for case 12 where only HH is used as the input variable during training and testing period. The contribution of the soil texture information is assessed by comparing the model performances between the case where soil texture is used with all four backscattering coefficients, i.e. case 1 and the case where only backscattering coefficients are used, i.e. case 11, for all the three approaches. The observation suggests that the omission of soil texture information marginally decreases the model performance of SVM, remains same in GP and increases in RF modelling. Hence, it can be said that the use of only SAR backscattering coefficients without the information of soil texture information can provide almost accurate SMC estimate by GP and RF and thus would be highly beneficial for ungauged locations. And, for SVM the performances of the cases without soil texture information and using HH can be considered acceptable since the soil texture information is unavailable/unreliable in many regions and the contribution of HH is well established assuring the applicability of the model in different regions. Although some studies have stated that the sensitivity of HH and VV is identical, especially for bare land areas, the fact that at C-band HH is more sensitive to soil moisture than VV due to relatively smaller attenuation by vegetation stand (Ulaby et al. 1982) allows the models to be applied more proficiently with the dominance of HH observed irrespective of LULC condition.

The model performances obtained in the present study show better model performance when compared with relatively well-established data-driven methods such as SVM and ANN in the similar field of study to verify the beneficial findings of the study (Notarnicola et al. 2008; Ahmad et al. 2010; Rodríguez-Fernández et al. 2015). The results of these established studies, using data-driven methods such as ANN and SVM, indicate to obtain a good performance with defined soil roughness condition or LULC using many other satellite data information and ground-truth information of the study area. While the present study emphasizes on using only the satellite data information to estimate the surface SMC, the comparison of model performances for all the methods with and without the soil texture information shows the comparable model performances for both the cases sustaining the novelty of the study. Otherwise, in the studies using the backscattering coefficient data, the model performance is found to be lesser than obtained in the present study, e.g. Notarnicola et al. (2008) and Ahmad et al. (2010). Therefore, the comparison leads to the pledge of enhanced and acceptable model training and performance of the present study.

The comparison of the mean bias is performed by computing the difference between the mean of the observed SMC and the mean of the estimated SMC during the training and testing period for combine bare and vegetated land areas. The averages of the mean bias across the three folds for each method and each of the 13 cases are presented in supplementary document in Figure S3 for combined bare and vegetated land areas. It can be observed that the mean bias is insignificant in RF for both training and testing periods where the ranges are 0.000067–0.00463 and 0.0078–0.0102, respectively, for all the 13 cases. The highest mean bias values are observed for SVM where the ranges are 0.075–0.079 and 0.081–0.088 during training and testing periods, respectively, across all the cases. In a nutshell, the performances from all the three approaches, i.e. SVM, GP and RF models, are reasonably acceptable during both the training and testing periods for all the 13 cases for combined bare and vegetated land areas.

Figure 2 shows scatter plots between observed and estimated surface SMC data for GP of case 9 for all the three folds combined for bare and vegetated land areas. The similar scatter plots using SVM and RF are shown in supplementary document in Figure S4 and Figure S5. The HH and RVI are shown as the input combination as it provides a good model performance and it allows the study to validate its novelty to be applicable for the ungauged locations with the unavailability of soil texture information. From these figures, it can be observed that the scatter plots fit well for all the three methods within the SMC range of 10–40% which confirms the applicability of proposed approaches for this soil moisture range.

The indicated SMC range of model applicability is strongly supported by the physical phenomenon involved in the SMC estimation from SAR data. The backscattering coefficient increases with an increase in volumetric SMC up to 35% after which any change in dielectric properties in the soil does not influence the radar signal and it becomes insensitive to soil moisture. This phenomenon strongly favours the SMC range where all the three models are observed to perform the best.

Soil Moisture Maps

Using the three different AISC techniques, we are able to apply the supplementary data sets of RISAT1 for each date on a pixel-by-pixel basis to develop a soil moisture map of the study area. It has been observed that the results from all the three methods and for all the cases except case 2 and 12 are comparable. The soil moisture maps of our study area have been prepared using all the three methods discussed in the study and as observed in the model performance, the maps are also comparable for all the three methods. However, the map prepared by the methodology based on

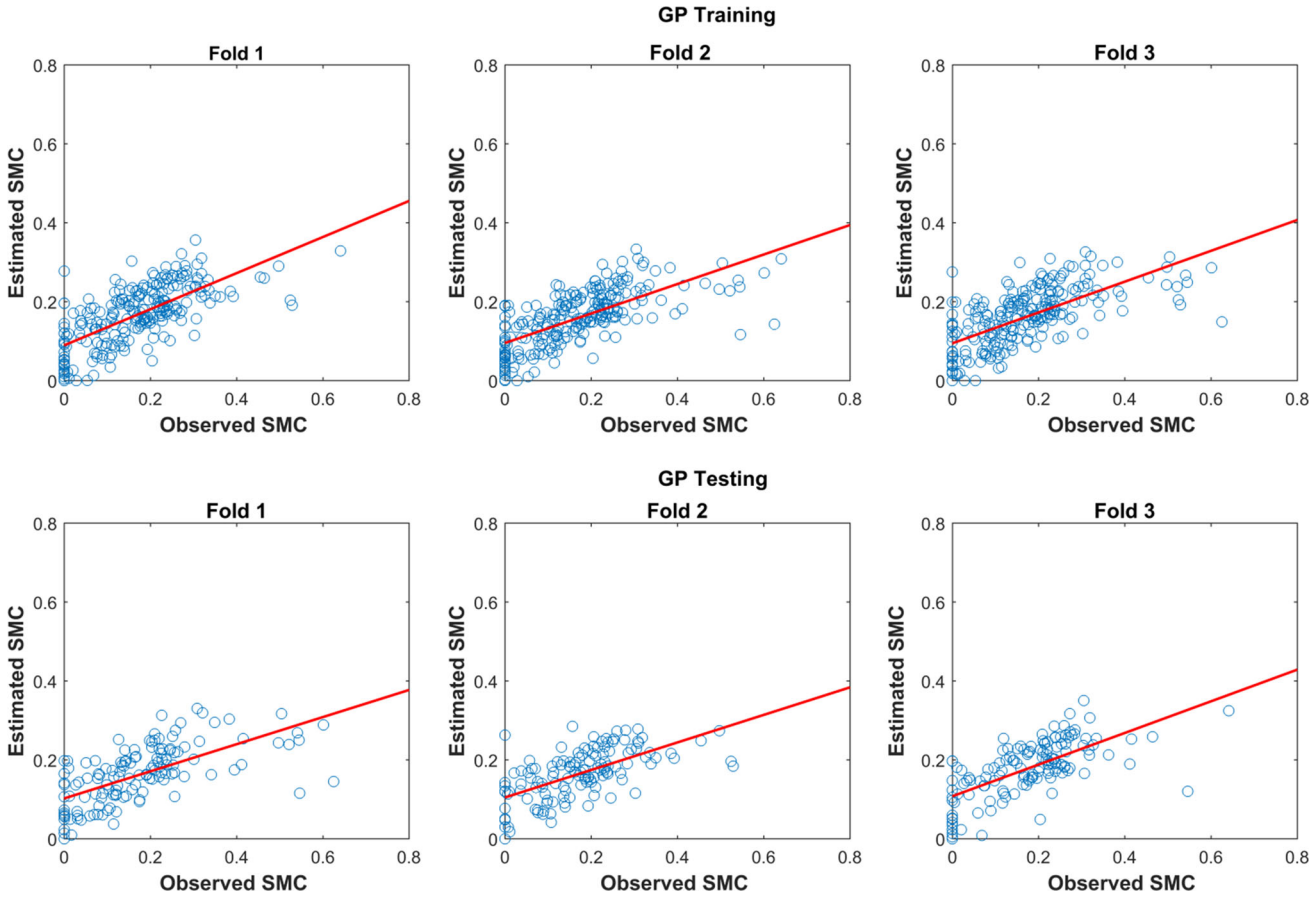

Fig. 2 The scatter plots between the observed and estimated surface SMC of the case 9 combined for bare and vegetated land areas for folds 1, 2 and 3 for GP. Best fit and 45° lines are also shown in the plots

GP technique is finalized due to its least overfitting and consistent performance for all the cases. The soil moisture maps developed with GP-based methodology for both the dates are shown in Fig. 3. The soil moisture maps are obtained by using the HH and RVI as the input combination. The selection of input variable combination is based on the substantially good model performance and applicability to the ungauged locations without soil texture information as discussed before. In the SMC maps for the two dates (Fig. 3), the blue regions indicate higher soil moisture and the yellow indicates the areas having lower soil moisture values. The white areas in the SMC maps mostly represent the masked regions consisting of establishments, roads, rivers and few patches of tall trees/forest where the models are not applicable. However, it also represents the SMC of 0 m^3/m^3. Since the study area mostly consists of agricultural lands, there is less chance of having an area with SMC value of 0 except some patches of barelands with very low SMC values which are shown in very light yellow colour in the SMC maps. Disparities in SMC due to drainage after a rainfall event are evident which are easily detectable due to homogeneous cover conditions for two different dates. As observed from the SMC maps of two different dates, the deeper blues representing higher soil moisture indicate the rainfall occurred on 27 October 2014.

A spatial validation is also attempted to compare the model output with the well-established, latest global atmospheric reanalysis ERA-Interim soil moisture product from European Centre for Medium-Range Weather Forecasts (ECMWF), which is based on the ECMWF Integrated Forecast System model using tiled ECMWF Scheme (Dee et al. 2011). However, the spatial resolution of ERA-Interim data ($0.125° \times 0.125°$ or $\sim$ 13 km $\times$ 13 km) is much coarser than our output ($\sim$ 20 m). Thus, the data are collected and initially compared with the observed SMC data of the study area for the two dates. The upscaled field-observed SMC values are compared with the SMC value of nearest four ERA-Interim grid points. The upscaling is carried out by taking the average of the volumetric SMC through all the monitoring points falling within a pixel of ERA-Interim data. The observed SMC values are found to be 0.245 and 0.233 for 25 October and 27 October 2014, respectively, and respective SMC values of ERA-Interim data are found to be 0.299 and 0.304. Hence, it can be noticed that there is some obvious uncertainty in the ERA-

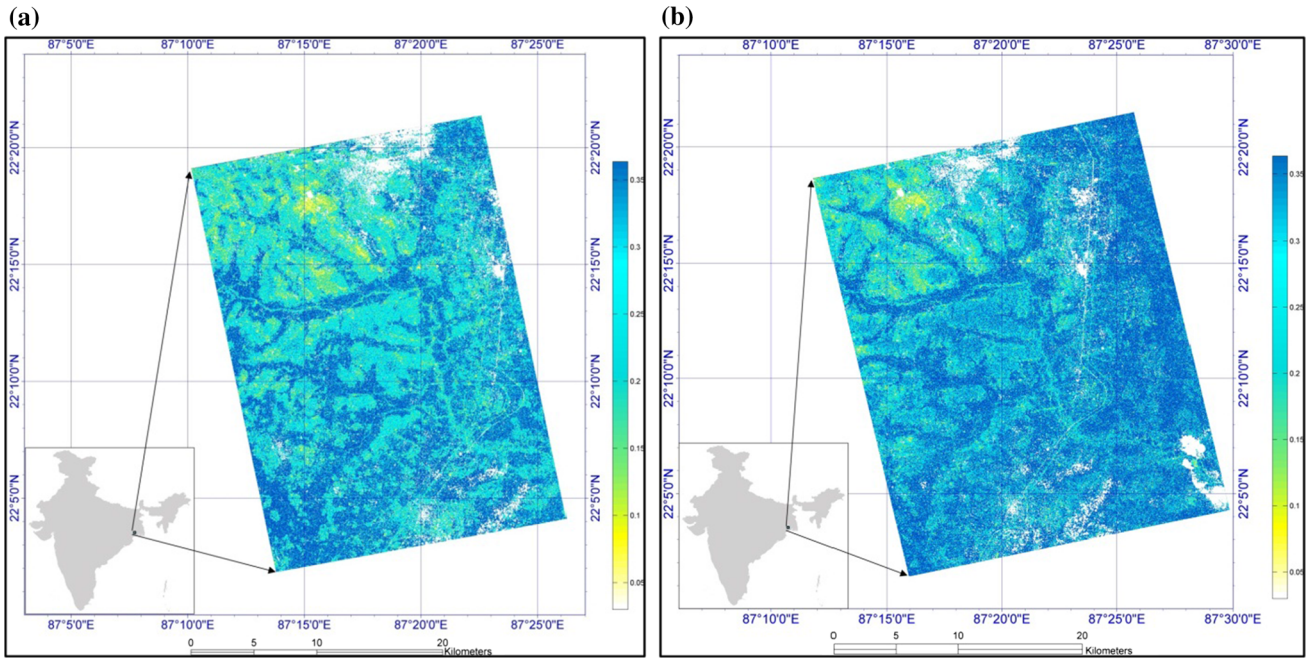

Fig. 3 Soil moisture map of the study area on **a** 25 October 2014 and **b** 27 October 2014 using the GP model. The unit of volumetric SMC shown in maps is m^3/m^3

Interim data that does not match fully with the observed field data. Still the correspondence is acceptable. However, as mentioned before, the resolution of ERA-Interim is very coarse as compared to our model output using RISAT1 data. This wide gap between spatial resolutions restricts a true comparison and indicates the benefit of very high resolution soil moisture map that is apparent (Figures S6 and S7 in the supplementary document). The presence of higher soil moisture in the low-lying areas and in the vicinity of water paths/river networks is very clearly visible in the maps developed in this study, which is highly masked in the SMC map using ERA-Interim values.

Conclusions

In this paper, different Artificial Intelligence-based Soft Computing (AISC) techniques are explored for their potential to develop soil moisture maps using the quad-polarized SAR data obtained from RISAT1. Different combinations of inputs consisting of radar backscattering data and soil texture information (HH, HV, VH, VV, RVI and percentages of sand, silt and clay) are used to compare the contribution of each input variable.

Three AISC techniques, namely SVM, GP and RF, can be beneficially utilized for surface SMC retrieval from SAR data with their relative merits and demerits. The comparison of the performance statistics obtained from these techniques suggests that they perform reasonably well for both training and testing data. However, the importance analysis in RF and sensitivity analysis in GP technique helps to quantify the contribution of each input variables and, in turn, it helps to precisely select the substantial subset of input variables.

The substantial sensitivity of HH to the surface SMC distribution is authenticated by the deteriorated model performances for all the three approaches when HH is dropped from the input set. The dominance of HH in soil moisture estimation is also quantified in the importance analysis in RF and sensitivity analysis in GP technique as well. The soil texture information (silt and sand in the present study) also plays an important role in determining the surface SMC for both bare and vegetated land areas with more complex models. However, model performance is still reasonably good even without the use of this information. In case the information of soil texture is not available, it is recommended to use only the HH and RVI. It increases the applicability of the AISC techniques to the areas without soil texture information. The validation of the model estimated SMC data with the ERA-Interim SMC values also indicates the beneficial use of quad-polarized SAR data in AISC-based techniques to obtain a more accurate high-relation SMC estimate. Overall, the AISC techniques are highly potential and we recommend the use of GP due to its ability to provide mathematical relationship between satellite-based SAR data and in situ soil texture information with surface SMC. It also provides importance and sensitivity of each individual input. The typical fine-resolution ($\sim$ 5 m) surface SMC map for a large area is extremely useful, and high-resolution, quad-

polarized SAR data can be beneficially used through the potential of AISC techniques, especially GP, without using the any priori ground information to develop the surface SMC map.

Acknowledgements This study is partially supported by a research project sponsored by the Space Application Centre (SAC), Indian Space Research Organization (ISRO), Govt. of India (Ref No.: IIT/SRIC/CE/VIR/2016-17/88).

References

Ahmad, S., Kalra, A., & Stephen, H. (2010). Estimating soil moisture using remote sensing data: A machine learning approach. *Advances in Water Resources, 33*(2010), 69–80.

Alavi, A. H., & Gandomi, A. H. (2011). A robust data mining approach for formulation of geotechnical engineering systems. *Engineering Computations, 28*(3), 242–274.

Breiman, L. (2001). Random forests. *Machine Learning, 45,* 5–32.

Coleman, M. L., & Niemann, J. D. (2012). An evaluation of nonlinear methods for estimating catchment-scale soil moisture patterns based on topographic attributes. *Journal of Hydroinfomatics, 14*(3), 800–814. https://doi.org/10.2166/hydro.2012.145.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., et al. (2011). The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society, 137,* 553–597.

El-Baroudy, I., Elshorbagy, A., Carey, S. K., Giustolisi, O., & Savic, D. (2010). Comparison of three data-driven techniques in modelling the evapotranspiration process. *Journal of Hydroinformatics, 12*(4), 365. https://doi.org/10.2166/hydro.2010.029.

Elshorbagy, A., & El-Baroudy, I. (2009). Investigating the capabilities of evolutionary data-driven techniques using the challenging estimation of soil moisture content. *Journal of Hydroinformatics, 11*(3–4), 237–251.

Espinoza-Dávalos, G. E., Arctur, D. K., Teng, W., Maidment, D. R., García-Martí, I., & Comair, G. (2016). Studying soil moisture at a national level through statistical analysis of NASA NLDAS data. *Journal of Hydroinformatics, 18*(2), 277–287. https://doi.org/10.2166/hydro.2015.231.

Gill, M. K., Asefa, T., Kemblowski, M., & McKee, M. (2006). Soil moisture prediction using support vector machines. *Journal of the American Water Resources Association, 42*(4), 1033–1046.

Han, D., Chan, L., & Zhu, N. (2007). Flood forecasting using support vector machines. *Journal of Hydroinformatics, 9*(4), 267–276. https://doi.org/10.2166/hydro.2007.027.

Karthikeyan, L., Pan, M., Wanders, N., Kumar, D. N., & Wood, E. (2017). Four decades of microwave satellite soil moisture observations: Part 2. Product validation and inter-satellite comparisons. *Advances in Water Resources, 109,* 236–252. https://doi.org/10.1016/j.advwatres.2017.09.010.

Kim, Y., Jackson, T., Bindlish, R., Lee, H., & Hong, S. (2012). Radar vegetation index for estimating the vegetation water content of rice and soybean. *IEEE Geoscience and Remote Sensing Letters, 9*(4), 564–568.

Kornelsen, K. C., & Coulibaly, P. (2013). Advances in soil moisture retrieval from synthetic aperture radar and hydrological applications. *Journal of Hydrology, 476,* 460–489.

Koza, J. (1992). *Genetic programming, on the programming of computers by means of natural selection*. Cambridge, MA: MIT Press.

Li, X., Sha, J., Li, Y., & Wang, Z. L. (2018). Comparison of hybrid models for daily streamflow prediction in a forested basin. *Journal of Hydroinformatics, 20*(1), 191–205. https://doi.org/10.2166/hydro.2017.189.

Liu, Y. Y., Dorigo, W. A., Parinussa, R. M., De Jeu, R. A. M., Wagner, W., McCabe, M. F., et al. (2012). Trend-preserving blending of passive and active microwave soil moisture retrievals. *Remote Sensing of Environment, 123,* 280–297. https://doi.org/10.1016/j.rse.2012.03.014.

Maity, R., & Kashid, S. S. (2011). Importance analysis of local and global climate inputs for basin-scale streamflow prediction. *Water Resources Research, 47*(11), W11504. https://doi.org/10.1029/2010WR009742.

Makkeasorn, A., Chang, N. B., & Li, J. (2009). Seasonal change detection of riparian zones with remote sensing images and genetic programming in a semi-arid watershed. *Journal of Environmental Management, 90*(2), 1069–1080.

Notarnicola, C., Angiulli, M., & Posa, F. (2008). Soil moisture retrieval from remotely sensed data: Neural network approach versus Bayesian method. *IEEE Transactions on Geoscience and Remote Sensing, 46*(2), 547–557.

Pal, M., Maity, R., Suman, M., Das, S. K., Patel, P., & Srivastava, H. S. (2017). Satellite based probabilistic assessment of soil moisture using C-band Quad-polarized RISAT 1 data. *IEEE Transactions on Geoscience and Remote Sensing, 55*(3), 1351–1362. https://doi.org/10.1109/TGRS.2016.2623378.

Paloscia, S., Pettinato, S., Santi, E., Notarnicola, C., Pasolli, L., & Reppucci, A. (2013). Soil moisture mapping using Sentinel-1 images: Algorithm and preliminary validation. *Remote Sensing of Environment*, *134*, 234–248.

Pasolli, L., Notarnicola, C., Bertoldi, G., Bruzzone, L., Remelgado, R., Greifeneder, F., et al. (2015). Estimation of soil moisture in mountain areas using SVR technique applied to multiscale active radar images at C-band. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 8*(1), 262–283.

Rodríguez-Fernández, N. J., Aires, F., Richaume, P., Kerr, Y. H., Prigent, C., Kolassa, J., et al. (2015). Soil moisture retrieval using neural networks: Application to SMOS. *IEEE Transactions on Geoscience and Remote Sensing, 53*(11), 5991–6007.

Rutkowski, L., Jaworski, M., Pietruczuk, L., & Duda, P. (2015). A new method for data stream mining based on the misclassification error. *IEEE Transactions on Neural Networks and Learning Systems, 26*(5), 1048–1059.

SAC. (2015). RISAT-1 Data products format—version 1.4, Space Applications Centre, Indian Space 627 Research Organization, Ahmedabad, India, September, 2015. https://nrsc.gov.in/sites/all/pdf/format3.pdf. Accessed March 2018.

Schmidt, M., & Lipson, H. (2009). Distilling free-form natural laws from experimental data. *Science, 324*(5923), 81–85.

Ulaby, F., Moore, R., & Fung, A. (1982). *Microwave remote sensing: active and passive. Volume III, from theory to application*. Dedham, MA: Artech House.

Unnikrishnan, P., & Jothiprakash, V. (2017). Data-driven multi-time-step ahead daily rainfall forecasting using singular spectrum analysis-based data pre-processing. *Journal of Hydroinformatics*, jh2017029. https://doi.org/10.2166/hydro.2017.029.

Zribi, M., Baghdadi, N., Holah, N., & Fafin, O. (2005). New methodology for soil surface moisture estimation and its application to ENVISAT-ASAR multi-incidence data inversion. *Remote Sensing of Environment, 96*(3), 485–496.