A pattern recognition based approach to consistency analysis of geophysical datasets

Anish C. Turlapatya, Valentine G. Anantharaj, Nicolas H. Younan

Department of Electrical and Computer Engineering, Mississippi State University, Box 9571, Mississippi State, MS 39762, USA
Geosystems Research Institute, Mississippi State University, Box 9652, Mississippi State, MS 39762, USA

ABSTRACT

Remotely sensed data from satellites are often validated by comparing them against ground-based measurements which usually are relatively sparse. Conventional consistency analysis methods provide information on each data point individually and in relation to its neighbors. In this study, a consistency analysis method based on wavelet-based feature extraction and one-class support vector machines is proposed. This method performs a consistency assessment of the entire time series in relation to others and provides a spatial distribution of consistency levels. The presented method is tested on soil moisture product from Advanced Microwave Scanning Radiometer for Earth Observing System (AMSR-E) on board Aqua satellite for the years 2005–2006. Time series of in-situ soil moisture measurements from the USDA Soil Climate Analysis Network (SCAN) are used as training data. Spatial distribution of consistency levels are presented as consistency maps for a region, including the states of Mississippi, Arkansas, and Louisiana in the USA. These results are correlated with the spatial distributions of averaged quality control information, mean soil moisture, and the cumulative counts of dense vegetation. Moreover, the methodology is tested for its robustness by examining its sensitivity to the spatial distribution of the network of training data sites. Finally, seasonal consistency maps for soil moisture data are developed. The degree to which the satellite estimates agree with the in-situ measurements has been represented seasonally as consistency maps which are helpful in interpreting the overall quality of the soil moisture product retrieved from satellite observations.

1. Introduction

Soil moisture is one of the most important environmental variables in regional weather and global climate systems. In particular, it plays an important role in modulating the energy and water cycles of the Earth’s system (Koster et al., 2004). It is also directly related to other bio- and geophysical variables, such as precipitation, vegetation characteristics, temperature, evaporation, and transpiration. It has been characterized as an “environmental descriptor that integrates much of the land surface hydrology and is a key variable linking the earth surface and the atmosphere” (Leese et al., 2001). The soil moisture near the surface determines the partitioning of latent and sensible heat fluxes, evaporation and surface runoff. Moreover, soil moisture in deeper layers also regulate how the ecosystems respond based on available water content in the soils (Robock et al., 2003). Hence, the monitoring, analysis and prediction of soil moisture is critical for: weather and climate studies of routine forecasting of weather events, including flooding; and for planting, irrigation and drought prediction and management strategies for agriculture. Despite the diverse critical application needs, accurate measurement and routine monitoring of soil moisture at global scales remains a great challenge. There is general consensus that most immediate requirement of a routine global soil moisture product at 50 km resolution could be feasible using a combination of both in-situ measurements and remotely sensed estimates, assimilated into land surface models (Leese et al., 2001).

1.1. In-situ measurements

Soil moisture variability can be monitored using in-situ measurements, from observing systems such as the Soil Climate Analysis Network (SCAN, Schaefer et al., 2007) and the Oklahoma Mesonet (McPherson et al., 2007). Robock et al. (2000) provide a survey of the history of soil moisture measurements and its importance. Soil moisture measurements on a regular basis were started in former Soviet Union in 1930s at a few agrometeorological stations. These techniques and practices were later adopted by Russia’s Asian neighbors, China, India and Mongolia.
In the United States, routine measurements started in 1980s in Illinois. Now, there are over 100 stations in Oklahoma. The global soil moisture data bank consists of data from several hundred stations spread across the globe. Nevertheless, there are difficulties in integrating these in-situ measurements into soil moisture products and comparing them to remotely sensed estimates (Robock et al., 2000; Leese et al., 2001): (i) the different observation stations have different instrumentation using various measurement techniques; (ii) the period of the records and measurement intervals vary resulting in discontinuities; (iii) there are no standardized approaches for calibrating and converting the basic observations into uniform units of volumetric soil moisture content; (iv) the measurement depths in the soil vary among stations and networks; (v) optimal network design is difficult; and (vi) the instrumentation and communication package is relatively expensive for deployment, especially for developing nations.

1.2. Remotely sensed estimates

Global estimates of soil moisture could be derived from remotely sensed observations from satellites and aircraft using a broad range of the electromagnetic spectrum, particularly in the microwave or infrared frequencies. Currently, global soil moisture products at relevant spatial scales (for hydrometeorological applications) are feasible only from microwave-based remote sensing observations. Techniques have been developed to retrieve soil moisture estimates from either passive (PMW) or active microwave (AMW) remote sensing, as well as a combination of both. Satellite platforms with active microwave payloads suitable for soil moisture include RADARSAT, Windsat, and the European Remote Sensing (ERS) satellite. Based on the concept that scattered radiation from soil moisture surface is related its dielectric constant, Alvarez et al. (2004) studied the relation between RADARSAT-1 observations and surface soil moisture. They used empirical and physical models to evaluate these observations at La Tejera watershed in Spain. The empirical models showed good agreement at large spatial aggregation scales at which influence of speckle is minimal. The physical model showed higher dispersion due to its sensitivity to surface characteristics. Lakhanakar et al. (2006) proposed a neural network and fuzzy logic based tools to retrieve soil moisture from RADARSAT (AMW) data in Oklahoma. Through combined use of vegetation information and optimized neural network classifier, an improvement in accuracy was attempted. The accuracy of the algorithm improved with inclusion of vegetation optical depth and NDVI in training data. Sanli et al. (2008) compared soil moisture content with multiple polarizations and incident angles of RADARSAT-1, ASAR on board ENVISAT and HH polarized ALOS SAR (PALSAR); and found correlations of 76%, 81% and 86%, respectively.

Dente et al. (2009) compared soil moisture datasets from AMSR-E radiometer (passive) and ERS-2 scatterometer (active) over test sites in Oklahoma Mesonet and OzNet in Australia. In both the data sets there were comparable trends, autocorrelation and temporal variations. However, the trends and autocorrelation of in-situ data at deeper levels were much longer than those of satellite time series. Dente et al. (2009) further hypothesize that there is a possibility to merge these two datasets for improved global soil moisture monitoring.

The Scanning Multichannel Microwave Radiometer (SMMR) is a PMW sensor operating at dual frequencies of 6.6 (C-band) and 10 GHz (X-band) onboard the Nimbus-7 satellite operated by the National Aeronautics and Space Administration (NASA) in the USA. It provided one of the earlier remote sensing observations of soil moisture from satellites. Vinnikov et al. (1999) have validated SMMR data against in-situ measurements in Illinois, and determined that retrieval frequencies as high as 18 GHz is are possible options for soil moisture observations in low vegetation areas. Paloscia et al. (2001) proposed a multi-frequency algorithm for retrieving soil moisture from SMMR. The algorithm was initially tested in southern France and later extended to wider spatial scales. This algorithm is capable of correcting vegetation biomass effects using polarization index in X-band. The algorithm was successful in deriving soil moisture from the C-band brightness temperature over the test sites in Russia; and the regression relationship developed had an R-square of 0.7. Guha and Lakshmi (2004) studied the soil moisture retrieval methodology from SMMR and validated this dataset in central United States. Their retrieval method is based on inversion of a radiative transfer model, and the spatial resolution of dataset is 1 x 1 deg². Monthly aggregation and averaging over larger spatial domains generally improved accuracy. One of the important lessons learned from SMMR retrievals was that the characterization of vegetation and vegetation water content is very important for soil moisture estimation.

Recently, global Level 3 surface soil moisture products are being routinely retrieved from the Advanced Microwave Scanning Radiometer for the Earth Observing System (AMSR-E), a multi-purpose instrument on-board NASA’s Aqua satellite. The sensor measures the brightness temperatures at 6.9, 10.7, 36.5 and 89 GHz channels in the microwave region. Soil moisture values are retrieved from the brightness values at 10.7 GHz by inversion of a radiative transfer model. The key physical parameters that carry soil moisture information are the surface emission and reflection coefficients in the model (Njoku, 1999). Since the signal in the C-band has unacceptable level of RFI contamination (Bindlish et al., 2005) only the measurements of the X-band have been used for the official AMSR-E soil moisture products from NASA. The spatial resolution of this dataset is 25 x 25 km². Additional description of the AMSR-E global Level 3 product is provided in Section 3.1.2.

Prigent et al. (2005) evaluated the sensitivity of several available satellite observations with respect to soil moisture on a global scale, the inter-compatibility of these datasets and possible combinations to improve soil moisture estimation. The evaluation was performed on retrievals through thermal infrared, passive and active microwave measurements against in-situ measurements from several stations in the northern hemisphere. They noted that: (i) passive MW observations above 19 GHz are sensitive to vegetation; (ii) active MW observations are more sensitive to soil moisture at lower incident angles; (iii) when evaporation controlled the surface temperature, infrared observations did not correlate well with soil moisture; and (iv) the sensitivity of each satellite instrument is very different to various soil surface characteristics such as moisture, vegetation, surface texture and roughness. These conclusions further indicate the set of challenges involved in the cross-validation of soil moisture estimates from different sensor systems and comparisons against in-situ measurements. Hence, there is a need to develop and test new and innovative techniques that will help relate soil moisture information from one source against another.

Presently, there are several new satellite missions, capable of observing soil moisture, at different phases of planning and development. The Soil Moisture and Ocean Salinity (SMOS) mission scheduled to launch in November of 2009 will provide soil moisture measurements with 4% volumetric accuracy, spatial resolution of 35–50 km and revisit times of 1–3 days. This dataset will also have global coverage and high sensitivity (Berger et al., 2002; Kerr et al., 2000). A potential soil moisture product is from L-Band Aquarius Radiometer and Scatterometer on-board Aquarius satellite, scheduled for launch in 2010. This will be the
first satellite instrument to provide simultaneous active/passive measurements (Jackson, 2007; Joseph et al., 2008; Luo et al., 2008). Though the Aquarius is designed to be a pathfinder mission for ocean salinity measurements, it has the potential to retrieve soil moisture estimates at weekly time scales which could then be further synthesized into soil moisture products via data assimilation using a land surface model (Anantharaj et al., 2009; Luo et al., 2003). NASA’s Soil Moisture Active and Passive (SMAP) mission, dedicated to observing soil moisture and land surface state, is scheduled for launch in 2012. One of the major goals of the SMAP mission is to develop new methodologies to combine radar and radiometer measurements (Entekhabi et al., 2008). Further, there are also considerations for a Level 4 soil moisture product using a land surface model that will assimilate the SMAP observations (Houser, P. R., July 2008, personal conversation).

1.3. Value-added soil moisture products from land data assimilation systems

A rather different approach to add value to the analysis (Level 4 products) is to use a land surface model (LSM) and assimilate available observations (Anantharaj et al., 2008; Houser et al., 1998; Luo et al., 2003; Rechile et al., 2002, 2007; Rodell et al., 2004; Sahoo et al., 2006). The idea is to use full physics numerical models of land surface hydrology to assimilate and downscale the observations to higher temporal and spatial resolutions. The complex numerical models using advanced data assimilation techniques were capable of generating better information but limited in terms of using all the available data. Before assimilating the soil moisture data into the land surface models, the validity of the observational data have to be verified and the statistical properties of the land surface model and the observations be reconciled, using techniques such as CDF matching. Rechile et al. (2007) assimilated multi-year observations of AMSR-E and SMMR in separate runs using the NASA Catchment LSM. They also observed that the multi-year climatologies of both, the satellite data sets differed from one another as well as from the LSM climatology. Hence, a bias correction was applied using a CDF matching technique. The final integrated product from the Catchment model was found to be better than both the individual satellite estimates as well as the model simulations without the data assimilation. However, it is not always simple to compare the satellite and model-derived soil moisture estimates to in-situ measurements. LSM model derived soil moisture observations do not compare well with in-situ measurements due to reasons such as: (i) sensitivity to parameterizations of complex processes in the model; (ii) uncertainties in atmospheric forcing; (iii) soil heterogeneity and lack of soil adequate texture information (Mostovoy and Anantharaj, 2008); and (iv) in-adequate data for satisfactory evaluation of the soil skin temperature and surface moisture.

1.4. Consistency analysis

Consistency analysis in the context of our study can be defined as a method to assess the degree of statistical agreement or adherence between an experimental dataset and a reference dataset. Conventional consistency analysis methods of satellite data include methods such as those introduced in Feng et al. (2004). They include: (i) basic plausibility check, a simple method for plausibility check is extreme value check; (ii) temporal consistency check, which consists of identifying temporal outliers in a time series; (iii) internal consistency, where the rate of change of a certain parameter would have certain limits, i.e., ground temperature cannot change more than a few degrees in a short duration of time; and (iv) spatial consistency checks, where data points are compared with the data from the surrounding spatial locations, e.g., buddy check, interpolation based methods. One of the major advantages with these methods is that they provide consistency information on individual data points. An improvement over these simple consistency analysis methods is the complex quality control (CQC), developed by Gandin (1998), where the decision on the consistency of data is made only after collecting the information from all CQC components. CQC is generally implemented in two stages. The first stage consists of the application of CQC components, and the second stage involves decision making on the acceptance or rejection of data points. These CQC components are generally individual QC methods, such as conventional consistency check methods. Furthermore, this method provides a sequence of quality flags on each data point. One of the disadvantages with these methods is they do not provide overall consistency information on large datasets. Spatio-temporal structure of consistency information of a time series of a spatial data grid cannot be computed using existing methods.

A more fundamental problem arises from the challenges of relating point measurements to areal and layer averaged estimates derived from remote sensing and inversion methods. Each cell of the passive MW (satellite) observations represents averaged soil moisture from a footprint covering a surface of several square kilometers. Besides, the satellite swath varies in each pass, based on the orbital characteristics. Famiglietti et al. (1999) addressed the problem within footprint variations during the Southern Great Plains hydrology experiment. Key findings were (i) the mean moisture variations among different study sites were in agreement with respective local characteristics such as soil types, rainfall gradients and vegetation covers and (ii) with the decrease of moisture content the statistics up to fourth order increased. In particular, the skewness changed from negative for wet soils to positive for dry soils. Also, typically the soil moisture retrievals from remotely sensed data are representative of a thinner upper layer (~1 cm for AMSR-E) where as the point measurements likely are from 2 cm or lower. Hence the dynamics of the soil moisture variations are different at lower depths.

Understanding the scales in soil moisture variation in terms of spatial grids and temporal steps is vital for determining how well a land surface model integrates soil moisture observations. Entin et al. (2000) studied the scales of temporal and spatial variations in soil moisture over extra tropical regions in United States and Eurasia. The temporal autocorrelation was modeled as an exponential with two components. First one is the red noise component corresponding to atmospheric forcing and the second component representing short term processes such as infiltration, cloud coverage, precipitation and drainage. Temporal scales varied from one month in the south to over two months in the north of China. Spatial scales were in the order of several hundred kilometers for upper 1m soil layers in Eurasian fields. Such analyses are adequate for climate scale studies but not for understanding regional processes. Due to the limitations of conventional validation methodologies which require high resolution in-situ soil moisture data, it is difficult to perform regional validations of satellite estimates of soil moisture. It is necessary to understand the soil moisture variability at smaller spatial scales for regional validation and applications.

The objective of this research was to develop a methodology to assess the level of agreement between remotely sensed data and in-situ measurements (usually sparse). The algorithm is based on wavelet based-feature extraction and one-class SVM. This pattern recognition based method can be used to develop consistency maps that provide spatial structure of consistency analysis information of a spatio-temporal dataset. The methodology operates on feature vectors in feature space instead of time series in time domain. The methodology has been applied toward assessing the soil moisture data from AMSR-E against soil
moisture data at regional scales from Soil Climate Analysis Network (SCAN) sites. However, it is generally applicable to other geophysical data sets remotely sensed from satellites and aircrafts.

2. Methodology

2.1. Motivation

Consider a spatio-temporal dataset with a spatial resolution of 0.25° × 0.25° and revisit time of 1 to 2 days. Second, consider a set of time series of measurements available at several ground points distributed over the study region. The goal is to compare the spatio-temporal dataset against the ground measurements. Traditional consistency analyses are time domain methods with simple statistical tools as discussed in the introduction. In this study we propose a consistency analysis tool that operates on feature vectors in feature space instead of time series. The time series data are transformed into feature space via feature extraction process. Feature extraction has been successfully applied in a variety of applications, such as image classification and dimensionality reduction. Feature extraction is usually used in pattern recognition for data classification, where a large dataset can be converted into a relatively smaller but transformed data, which represents the original data. Wavelet-based feature extraction has been successfully used for hyperspectral signal analysis, image classification and segmentation, and signal detection (Bruce et al., 2002). The proposed consistency analysis tool is based on one-class SVM learning machine and is more sophisticated than traditional methods. Recently, one class SVM have been successfully applied to various anomaly detection applications, such as intrusion detection in computer networks, ion etching fault detection, and windows registry access detection, just to name a few (Heller et al., 2003; Sarmiento et al., 2005). Anomaly detection using a one-class SVM works as follows: the data are collected while processing is in normal operation. Then, a binary classification is performed on this data using SVMs. New datasets are classified based on information from previously collected data. For instance, in the intrusion detection method, the statistics of the traffic in the computer network are collected initially. The SVM algorithm determines the hypothesis class based on these statistics and a particular kernel. A test set of statistics is then collected and each set is tested for its relevance to the hypothesis. If it falls outside the hypothesis, it is treated as an anomaly. In our study, instead of anomaly detection, the goal is to determine the qualitative position of each test sample with respect to the hypothesis class. The key idea is that the wavelet-based feature set corresponding to the in-situ measurements can be used to define the hypothesis class based on the position of support vectors. This hypothesis class is a sort of consistency yardstick for the satellite measurements (Fig. 1). A simplified feature space for consistency analysis of test samples against a set of training samples may appear as shown in Fig. 1. Each feature vector has two features. Any test sample that falls within the boundary of the hypothesis class can be considered consistent. All the outside samples are considered inconsistent with degree of inconsistency proportional to the statistical distance to the hypothesis class. However, in the actual implementation, the hypothesis class will be M-dimensional hypersurface, where M is the no. of features in a sample. A detailed description of the proposed methodology follows.

2.2. Method description

The proposed consistency analysis is a three step process, namely (i) feature extraction, (ii) classification, and (iii) consistency assessment. In the first step, the discrete wavelet transform is applied to the time series at each cell of the spatial grid to obtain wavelet coefficient series. The energies of the coefficients in every sub-band are treated as the features. Then, feature vectors are developed for the in-situ dataset. In the second step, the feature set from in-situ data is used to train the one-class SVM learning machine. A set of support vectors and Lagrange multipliers are deduced. These, along with a kernel function and the set of test feature vectors are used to obtain as set of distance measures. Finally in the third step, a consistency assessment is performed to generate a consistency map based on the distance measures. Mathematical details of the three steps follow.

2.2.1. Step one: feature extraction

In general, feature extraction is a two stage process, as shown in Fig. 2. The first stage is feature construction, where a feature set is extracted from the original dataset. For instance, a set of major spectral signatures can be retrieved from a time series. The second stage is feature selection, since all the components of a feature set

![Fig. 1. Feature space for consistency assessment of samples with two features.](image-url)
are not useful for classification. Feature selection can be a simple statistical method like a t-score test, where only the first few spectral signatures can be useful for classification and the remaining signatures are redundant (Guyon et al., 2006). Wavelet coefficient based energies are extracted as follows.

2.2.1.1. Features from discrete wavelet transform. Let \([n]\) represent a time series at a cell \((s_{\text{lat}}, s_{\text{lon}})\) in the satellite image grid, here, \(s_{\text{lat}}\) and \(s_{\text{lon}}\) are the latitude and longitude co-ordinates of the center of the grid cell. Here, \(n\) is the time index. In the case of training data, the point \((s_{\text{lat}}, s_{\text{lon}})\) corresponds to the ground station. Then, the wavelet coefficients corresponding to the discrete wavelet transform of \([n]\) can be computed as follows. The DWT is used to decompose the original time-series into approximations \(c_j[k]\) and details \(d_j[k]\) coefficients, where \(k\) is the index of respective sequences and \(j\) is the level of decomposition. At the

---

**Fig. 2.** Feature extraction process applied on soil moisture time series from a SCAN site.
highest level \( J \) the initial approximation is
\[
c_j[k] = f[m] \quad (1)
\]
Then the approximations \( c_j[k] \) for the successive lower levels are given by
\[
c_j[k] = \sum_{m} h[(m-2k)c_{j+1}[m] \quad (2)
\]
and the details \( d_j[k] \) for successive levels are determined by
\[
d_j[k] = \sum_{m} g[(m-2k)c_{j+1}[m] \quad (3)
\]
Here the filters \( h[k] \) and \( g[k] \) correspond to the mother wavelet under consideration (Burrows et al., 1997) and \( m \) is the index of convolution. In practice the Eqs. (2) and (3) are implemented using circular convolution. In DWT the lengths of the approximation and detail sequences reduce with the decomposition level. For level 3 decomposition the value of \( j \) varies from \( J-1 \) to \( J-3 \) and the value of \( J \) itself depends on the length of the time series \( f[n] \).

Once the wavelet coefficients are obtained, the energy for each sub-band can be computed \( e_j = \sum d_j^2[k] \). The feature vector corresponding to the time series \( f[n] \) at \( (s\text{, lat}, s\text{, lon}) \) can be defined as \( X_{\text{RWT}} = [e_1, e_2, ..., e_M] \). Here, \( e_1, e_2, ..., e_M \) are energies of the selected \( M \) sub-bands. Since there is one approximation sequence and three details sequences for a three level decomposition, the value of \( M=4 \).

### 2.2.1.2. Features from redundant discrete wavelet transform (RDWT)

For comparison purposes, a RDWT based feature extraction is also performed as follows. RDWT is a discrete approximation of continuous wavelet transform. The filtering operation is similar to the DWT except the filters are recursively upsampled at each scale, i.e.,
\[
h_j[k] = h_{j+1}[k] \uparrow 2 \quad \text{and} \quad g_j[k] = g_{j+1}[k] \uparrow 2
\]
where \( h_j[k] \) and \( g_j[k] \) are the filters at level \( j=J \), the filters are \( h_J[k] = h[k] \) and \( g_J[k] = g[k] \).

The lengths of the approximation and detail sequences are the same as the length of the signal. The sequences at level \( j \) are obtained by circular convolution as shown below:
\[
c_j[k] = h_{j+1}[k-d_{j+1}] \uparrow 2 \quad \text{and} \quad d_j[k] = g_{j+1}[k-d_{j+1}] \uparrow 2
\]
This process starts at level \( j=J-1 \) and ends at the last level \( j=1 \) (Fowler, 2005). Thus, the sequence \( c_j[k] \) is decomposed into the redundant sequence \( Y \):
\[
Y = [c_1, d_1, d_{1+1}, ..., d_{J-2}, d_{J-1}]. \quad (7)
\]
The entropy of each of the sequences in Eq. (7) constitutes the feature vector for a given time series.

The corresponding feature vector in the case of RWT is \( X_{\text{RWT}} = [e_1, e_2, ..., e_M] \), where \( e_1, e_2, ..., e_M \) are energies of the selected \( M \) sequences in \( Y \) in Eq. (7).

#### 2.2.2. Step two: classification with one-class support vector machines

Let \( X = \{x_1, x_2, ..., x_m\} \) be the \( M \)-dimensional feature vector. The feature vector \( X = X_{\text{RWT}} \) when using DWT for feature extraction and \( X = X_{\text{RWT}} \) when using RWT. If \( X \) belongs to the hypothesis class, then the classifier \( y=1 \), otherwise, \( y=0 \). Thus the function \( y \) decides the class of the feature vector. The \( M \)-dimensional weight vector \( w \) is defined as
\[
w = \Phi(X)\alpha \quad (8)
\]
where \( \Phi(X) \) is a mapping function on \( X \). If the data are separable, then the decision function is
\[
D(X) = w^T X + b
\]
\[
D(X) = 2^q \Phi^T(X) X + b
\]
where \( b \) is the margin pertaining to the hyperplane. This is also known as the hyperplane that separates the hypothesis class from other vectors. The objective in a one class SVM is to find a hyperplane which provides optimal margin and good generalization ability (Abe, 2005; Cristianini and Taylor, 2000; Friedman and Kandel, 1999).

The optimal hyperplane can be obtained by transforming the problem into a quadratic optimization problem. In the following expression, \( \zeta_i \) is the slack variable signifying the soft margin support vector machines under consideration.
\[
\min_{\alpha} \frac{1}{2} \alpha^T \alpha - \rho \sum_{i=1}^{l} \zeta_i
\]
such that
\[
w^T \Phi(X) \geq \rho - \zeta_i
\]
\[
\zeta_i \geq 0, i = 1, ..., l
\]
In the above expression \( \rho \) is the offset for parameterizing the hyperplane and thus \( (w, \rho) \) if determined would specify the hyperplane. The parameter \( \epsilon \in [0, 1] \) is the trade-off parameter. The product \( \alpha \) influences the generalization ability of the SVM classifier where, \( l \) is no. of slack variables. Define a kernel function \( Q \)
\[
Q_o = k(x_i, x_j) = \Phi(x_i)^T \Phi(x_j)
\]
By substituting Eq. (8), in Eqs (10) and (11) the dual problem can be reduced to
\[
\min_{\alpha} \frac{1}{2} \alpha^T Q \alpha
\]
such that
\[
0 \leq \alpha_i \leq 1/(l\epsilon) \quad i = 1, 2, 3, ..., L
\]
\[
\sum_{i=1}^{L} \alpha_i = 1
\]
Detailed derivation from (10) to (12) is available in Schölkopf et al. (2001). Above problem can be solved by employing a quadratic optimization approach developed by Coleman and Li (1996). The \( \alpha \) vector is obtained from the above optimization (Gill et al., 1981; Rusin, 1971).

#### 2.2.3. Step three: consistency assessment technique

There exists one \( \alpha \) for each training vector. The training vectors corresponding to non-zero \( \alpha \) values are treated as support vectors. Let \( S = \{s_1, s_2, ..., s_m\} \) be set of support vectors. The set of support vectors is a subset of the training feature vectors. The kernel function measure between test data vector \( X \) and a support vector is given by \( k(s_j, x) \). In order to compute the kernels such as Euclidean distance \( k(s_j, x) = |X - s_j|^2 \) Minkowski distance \( k(s_j, X) = (\sum |X - s_j|^p)^{1/p} \), linear kernel \( k(s_j, x) = x^T s_j \) and exponential radial basis function kernel function \( k(s_j, x) = \exp(-|X - s_j|^2) \) to mention a few can be used (Canu et al., 2005; Rakotomamonjy and Canu, 2005).

The vectors of these kernels constitute the kernel matrix \( K(S, X) \) given by
\[
K(S, X) = [k(S_1, X), k(S_2, X), ..., k(S_m, X)]
\]
A statistical distance measure \( d(X, S) \) is obtained between a test data vector and the set of support vectors using a kernel function and the \( \alpha \).
\[
d(X, S) = \alpha^T K(S, X)
\]
In this method, instead of using the classifier \( y \) for assessing a feature vector, we propose to utilize the distance \( d(X, S) \) and the vector of the distance measures of all the test feature vectors to
analyze the consistency. From this vector, the relative position of each of the test vector with respect to the hypothesis class can be determined and a consistency level can be assigned in a manner similar to Fig. 1. However, the hyperplane will be a hypersurface in this case study and the hypothesis class will be a hypersphere. The consistency assessment is shown in Table 1.

In this study, the proposed methodology is applied for consistency assessment of AMSR-E soil moisture time series in relation to SCAN soil moisture data. A block diagram of the presented methodology is presented in Fig. 3.

3. Experimental results

3.1. Time series generation

3.1.1. Soil moisture data from SCAN

Soil moisture is measured at the SCAN sites using a capacitance based instrument known as the Hydra probe. Basically this instrument generates estimates of the real dielectric constant values of soil surface. Soil moisture is computed from these dielectric constant values via a set of soil specific calibration equations. These calibration equations are third degree polynomials. The accuracy of these measurements without specific soil information is $\pm 0.03$ volume fraction of water. The robustness of the instrument and the measurement technique were well tested in different soils, temperatures, and precipitation conditions (Bosch, 2004; Kennedy et al., 2003; Schaefer et al., 2007; Seyfried and Murdock, 2004; Soil Survey Staff, 2007).

Soil moisture data sets are collected from 21 SCAN sites shown in Fig. 4(a), of which one site has no useful data (fill data) and three have insufficient data. Thus, there are 17 sets of time series which can be used for training purposes. Each data set corresponds to data collected in a given month, taken ideally 24 h a day and all the days in the month. The soil moisture field for the upper most soil layer is extracted from these data sets and a time series is generated over two years (2005 and 2006).

3.1.2. AMSR-E soil moisture data

In our study, the soil moisture data used is from AMSR-E Level 3 “AE_Land3” product, developed by NASA and distributed by the National Snow and Ice Data Center (NSIDC). The product release version is “Beta 03 release of Version 001”. This dataset had been corrected for structural errors involving the corner co-ordinates. Along with the soil moisture fields, this product also included other parameters such as brightness temperatures, vegetation water content, land surface temperature, and quality control data which presented information on land surface

### Table 1

<table>
<thead>
<tr>
<th>Quality level</th>
<th>Deviation $d$ from mean (in $\sigma$’s)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>$d &lt; = 1$</td>
<td>Good quality data</td>
</tr>
<tr>
<td>4</td>
<td>$1 &lt; d &lt; = 2$</td>
<td>Acceptable data</td>
</tr>
<tr>
<td>3</td>
<td>$2 &lt; d &lt; = 3$</td>
<td>Marginal data</td>
</tr>
<tr>
<td>2</td>
<td>$3 &lt; d &lt; = 4$</td>
<td>Poor data</td>
</tr>
<tr>
<td>1</td>
<td>$d &gt; 4$</td>
<td>Very poor or no data</td>
</tr>
</tbody>
</table>

![Fig. 3](image) A block diagram of consistency analysis method applied to AMSR-E soil moisture data.
The AMSR-E soil moisture estimates are retrieved by inversion of a land surface microwave emission model (MEM) with the support from ancillary data. The ancillary data used in Level 3 processing include properties of a grid cell such as (i) open water; (ii) surface topography; (iii) soil texture; (iv) vegetation type; (v) snow cover; and (vi) atmospheric parameters. The Level 3 products are composited using Level 2B soil moisture estimates into a global cylindrical EASE grid. The physical principle behind the MEM is that the brightness temperature sensed by the radiometer consists of three radiation components: (i) upwelling radiation from atmosphere; (ii) the surface emissions; and (iii) the down welling emission from atmosphere that is reflected back. The second component has three sub-components namely: (i) direct emission from the surface of the ground; (ii) vegetation radiation reflected by the ground surface; and (iii) vegetation radiation scattered by vegetation layer itself. The emission and reflection coefficients of the soil surface in this brightness temperature are functions of the complex soil dielectric constant which again is a function of soil moisture content along with other related variables including soil bulk density and other variables that can be obtained from ancillary data. A detailed description of this algorithm is presented in Njoku (1999) and Njoku et al. (2003). Though the AMSR-E instruments measures brightness temperatures in the C-band also, this signal is vulnerable to radio frequency interference especially near cities. So X-band brightness temperatures, which are less vulnerable, had been used in the Level 3 algorithm to retrieve soil moisture. This product has a $25 \times 25$ km$^2$ equal area scalable earth grid (EASE-grid), global cylindrical and equal area projection true at 30°N and S (Njoku, 1999; Njoku et al., 2003).

Accordingly, soil moisture data fields are extracted and data sets are generated for a region of $28 \times 23$ pixels for a period of two years. This region consists of the states of Mississippi, Arkansas, and part of Louisiana. There are 727 spatial data fields of soil moisture, from which a 3D matrix corresponding to the time series is generated. Note that the first two indices represent the spatial location and the third index specifies the day. Thus, in this study, 644 sets (one for each pixel) of time series are generated (Kawanishi et al., 2003; Njoku, 1999, 2005; Njoku et al., 2003; Paloscia et al., 2006).

The time series data thus obtained is used to generate the wavelet features. A three-level wavelet decomposition was applied to the time series (Duzenli and Monique, 1998). The energy features were calculated from the four sub-bands; the first one represents the approximation sub-band, and the others represent the three levels of detail sub-bands. One set of features developed for the soil-moisture time-series from scan sites and one for AMSR-E. SCAN features were used as training data and the AMSR-E features as test data for the validation process.

![Consistency maps and scan sites](image-url)
3.2. Implementation

From the $\mathbf{a}$ vector, the distance measure vector is computed. The mean and standard deviation of the distance distribution is measured. Based on the distance of the measure from the mean in terms of $\sigma$, a consistency level is assigned to it. Since the length of each time series is 727 days, the maximum level of wavelet decomposition is nine since 1024 is the nearest power of 2 to 727. Since the test data vectors are only 20, the number of features in each vector would be ideally $N/10=2$, where $N$ is the number of training feature vectors. In this study, level 3 decomposition is performed and feature vectors were reduced to a 20 by 3 matrix, using FLDA (Webb, 1999). The same weight vector also multiplies the test feature matrix. The resulting consistency map in this case is shown in Fig. 4(c).

The results obtained for the consistency analysis of AMSR-E soil moisture data are compared to the ones obtained from a method based on statistical properties of the time series (El-khamy et al., 2000). Statistical properties of each time series are computed and a statistical feature matrix is constructed for the entire geographic region. Similarly, a training feature matrix is constructed from SCAN data and the Mahalanobis statistical distance is measured between each test vector and the training feature matrix. The above consistency assessment method is repeated on these distance measures and a consistency map is developed based on this consistency information, which is illustrated in Fig. 4(d). These maps are compared to the results from the machine learning method.

3.3. Method validation

In order to validate the machine learning based algorithm, a conventional consistency analysis is performed on the AMSR-E data. Conventional consistency analysis is performed on each soil-moisture field value and the consistency information is stored in a binary sequence where each bit represents a flag for individual consistency check. In this experiment, the sequence has six bits or flags as described below. (i) Missing value check: if the field value is either bad retrieval or a fill value it is flagged and the remaining checks need not be performed; (ii) Range check: the possible range of a volumetric soil-moisture value is between zero and one half; (iii) Temporal consistency check: each field value is compared with the rest of the values in the time series at that pixel and if it is more than two standard deviations from the mean it is flagged; (iv) Step check: difference series of the time series is calculated and if the difference is too large a flag is set; (v) Step consistency check: temporal consistency check is performed on the difference series; and (vi) Spatial consistency check: each field value is compared with its spatial neighbors by computing a median test statistic. If the statistic is greater than two, the field value is flagged (Feng et al., 2004).

The consistency information is stored as a decimal equivalent of the binary sequence with $f_i$ as the individual bit value or consistency flag from the $i$th consistency check with $f_1$ and $f_6$ as most significant and least significant bits, respectively. In order to compare with the maps from the machine learning method, the conventional QC data is time averaged and spatial distribution of consistency is developed. Scale of Fig. 4(b) is adjusted by using the linear transformation $[6-(\text{conventional QC value})/6.4]$. This converts it to the scale 1–5. Thus, this scale is similar to scales in other consistency maps in Fig. 4(c and d). The correlation between the ML map and the conventional map is nearly 60%, which suggests the ML method is an extension of conventional methods by providing the spatio-temporal consistency of data.

Another method used for verification purposes is based on the $k$ nearest neighbor algorithm ($k$NN) (Friedman and Kandel, 1999). Treating consistency levels of time series as classes, which means there are five classes of data. This classification is verified using a leave-one-out method and the 3NN algorithm. The resulting classification is compared to the SVM consistency information and the number of exact matches gives the accuracy of the machine learning method.

3.4. Sensitivity studies

The robustness of the proposed machine learning algorithm has been tested by systematically selecting the in-situ data. The sensitivity of the algorithm is tested by dropping the individual SCAN sites from the training data. The sensitivity is presented as a distribution of average of the SVM based distance measures.

![Sensitivity plot: average SVM distance measure versus SCAN site dropped.](Fig. 5)
versus the individual site dropped. From Fig. 5 it can be inferred that average distance measure is fairly constant. To further illustrate the robustness of the algorithm, consistency map is presented for an instance in which a selected scan site is dropped and shown in Fig. 6.

3.5. Interpretation and possible applications of consistency maps

The consistency maps provide relative consistency of the measurements with respect to the reference data in both spatial and temporal dimensions simultaneously. The measurements in the regions with low consistency levels can be interpreted as data which have spectral energies far from those of reference data in the spectral energy space. A spatial coherency can be observed in these maps in areas with either high consistency or inconsistency and thus special attention can be given to these inconsistent regions for improvement of the measurements.

3.6. Performance comparison and seasonal variation

The performance of features from DWT is compared with the performance of features from RDWT. For different kernels the entropy features from RDWT give a better performance with respect to the average QC data, average soil moisture distribution, and the dense vegetation distribution. Correlations are computed between the map from DWT and RDWT features and the above mentioned distributions as shown in Table 2. Consistency maps from the RDWT features are presented in Fig. 7(a) and (b) which correspond to linear and Minkowski kernels, respectively. Since the original analysis was performed for a period of two years, the performance is also tested for individual seasons. The consistency maps are shown in Figs. 8(a) to (b) for fall of 2005 and summer of 2006.

4. Discussion and summary

The consistency maps are correlated with the spatial distribution of mean soil moisture and the cumulative dense vegetation pixel counts for the same geophysical region (Figs. 7(c) and (d)). There is significant positive correlation of over 60% between average soil moisture distribution and the consistency maps as shown in Table 2. This positive correlation suggests the measurement consistency is generally related to average soil moisture variation. Hence, we are cautiously optimistic that the soil moisture information retrieved from AMSR-E could lead to improved soil moisture analysis at higher spatial and temporal resolutions, using data assimilation techniques in land surface models, especially in areas where the land surface models perform poorly (Anantharaj et al., 2008).

In our study, the distribution of cumulative counts of dense vegetation has a negative correlation of more than 50% with the linear kernel consistency map and nearly 30% with the Minkowski distance based consistency map. This negative correlation suggests that the measurement consistency is inversely related to the density of vegetation. This interpretation agrees with the previous findings. So, the consistency maps in conjunction with
information about vegetation density at the pixel level could be used appropriately in the weighing functions of data assimilation algorithms; and thus providing intelligent means of selectively using the remotely sensed soil moisture data.

As stated earlier, soil moisture retrieval of AMSR-E observations is achieved by inversion of a radiative transfer model of soil, vegetation, and atmosphere medium in microwave region (Njoku and Li, 1999). According to the model there is a non-linear relation between vegetation density and retrieval uncertainty. The inversion algorithms are sensitive to the accuracies of vegetation optical depth, land surface temperature, and soil moisture. In order to account for this dependence, an iterative least squares minimization
approach is used. It is observed that the amount of soil emission reaching the sensor decreases with the density of vegetation. Moreover, for regions with sufficiently high vegetation density the soil emission may be completely lost. It is also understood that the vegetation influence depends on observation frequencies as well; especially, attenuation due to vegetation can be higher at higher frequencies (De Jeu et al., 2008; Hollinger et al., 1990). Njoku and Entekhabi (1996) suggested that attenuation due to vegetation is lower at low frequencies and the sensor is sensitive to sub-surface moisture. The SMAP mission, under formulation by NASA dedicated to measure soil moisture, will include an L-band active/passive instrument which is expected to perform better in areas of dense vegetation than the X-band retrievals from AMSR-E estimates used in this study. Further discussion about the different approaches of various retrieval algorithms are beyond the scope of this manuscript. They have been discussed and summarized in Wigneron et al. (2003). In general, retrieval accuracies could be improved by: (i) improving the a priori knowledge of the land surface conditions and vegetation; (ii) using lower frequencies (L-band); (iii) enhancing our understanding of the response of soil moisture and vegetation canopy to observational frequencies, polarization and look angles; and (iv) improved retrieval methods, based on i–iii above.

We further validated this methodology for different seasons and also studied the sensitivity of the algorithm to the consistency and spatial density of training data. Sensitivity of the algorithm has been observed in terms of the average SVM based statistical distance of all samples versus SCAN dropped. This distance remained approximately constant irrespective of the site dropped. The algorithm has been tested to be robust in the study region.

A pattern recognition based approach has been used to develop a new methodology for consistency assessment of large spatial temporal datasets. Features are extracted from individual time series using wavelet-based feature extraction. One-class SVM’s are applied to classify the features and thus time series into good and bad consistency data. The consistency information is shown in the form of consistency maps. The method is validated by correlating with distribution of related parameters like average soil moisture. The method can be improved via: (i) improvement of feature selection process; (ii) optimal parameter selection for classification; and (iii) optimal selection of a mother wavelet for feature extraction. The method also needs to be applied in other study areas, particularly in semi-arid regions, and the performance of the algorithm validated. Though the application of the methodology has been demonstrated using soil moisture data, it is also applicable to other geophysical data obtained from remote sensing and validated using in-situ measurements.

Acknowledgments

This research is sponsored by the NASA Applied Sciences Program via NAS13-03032 and the NOAA Office of Atmospheric Research via NA06OAR4600181. We also thank Dr. Christa Peters-Lidard for helpful suggestions on validation.

References


