5-1-2010

Application of pattern recognition and adaptive DSP methods for spatio-temporal analysis of satellite based hydrological datasets

Anish Chand Turlapaty

APPLICATION OF PATTERN RECOGNITION AND ADAPTIVE DSP METHODS FOR SPATIO-TEMPORAL ANALYSIS OF SATELLITE BASED HYDROLOGICAL DATASETS

By

Anish Chand Turlapaty

A Dissertation
Submitted to the Faculty of
Mississippi State University
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
in Electrical Engineering
in the Department of Electrical and Computer Engineering

Mississippi State, Mississippi

May 2010

APPLICATION OF PATTERN RECOGNITION AND ADAPTIVE DSP METHODS

FOR SPATIO-TEMPORAL ANALYSIS OF SATELLITE

BASED HYDROLOGICAL DATASETS

By

Anish Chand Turlapaty

Approved:

Nicolas H. Younan
Professor and Department Head of
Electrical and Computer Engineering
James Worth Bagley Chair
(Major Advisor and Dissertation Director)

Lori M. Bruce
Associate Dean for Research & Graduate Studies
Bagley College of Engineering
(Committee Member)

James Fowler
Professor and
Graduate Program Director
Electrical and Computer Engineering
(Committee Member)

Jenny Q. Du
Associate Professor
Electrical and Computer Engineering
(Committee Member)

Sarah Rajala
Dean, Bagley College of Engineering

Name: Anish Chand Turlapaty

Date of Degree: May 1, 2010

Institution: Mississippi State University

Major Field: Electrical Engineering

Major Professor: Dr. Nicolas H. Younan

Title of Study: APPLICATION OF PATTERN RECOGNITION AND ADAPTIVE DSP METHODS FOR SPATIO-TEMPORAL ANALYSIS OF SATELLITE BASED HYDROLOGICAL DATASETS

Pages in Study: 131

Candidate for Degree of Doctor of Philosophy

Data assimilation of satellite-based observations of hydrological variables with full numerical physics models can be used to downscale these observations from coarse to high resolution to improve microwave sensor-based soil moisture observations. Moreover, assimilation can also be used to predict related hydrological variables, e.g., precipitation products can be assimilated in a land information system to estimate soil moisture. High quality spatio-temporal observations of these processes are vital for a successful assimilation which in turn needs a detailed analysis and improvement. In this research, pattern recognition and adaptive signal processing methods are developed for the spatio-temporal analysis and enhancement of soil moisture and precipitation datasets. These methods are applied to accomplish the following tasks: (i) a consistency analysis of level-3 soil moisture data from the Advanced Microwave Scanning Radiometer – EOS (AMSR-E) against in-situ soil moisture measurements from the USDA Soil Climate Analysis Network (SCAN). This method performs a consistency assessment of the entire

time series in relation to others and provides a spatial distribution of consistency levels. The methodology is based on a combination of wavelet-based feature extraction and one-class support vector machines (SVM) classifier. Spatial distribution of consistency levels are presented as consistency maps for a region, including the states of Mississippi, Arkansas, and Louisiana. These results are well correlated with the spatial distributions of average soil moisture, and the cumulative counts of dense vegetation; (ii) a modified singular spectral analysis based interpolation scheme is developed and validated on a few geophysical data products including GODAE's high resolution sea surface temperature (GHRSST). This method is later employed to fill the systematic gaps in level-3 AMSR-E soil moisture dataset; (iii) a combination of artificial neural networks and vector space transformation function is used to fuse several high resolution precipitation products (HRPP). The final merged product is statistically superior to any of the individual datasets over a seasonal period. The results have been tested against ground based measurements of rainfall over our study area and average accuracies obtained are 85% in the summer and 55% in the winter 2007.

DEDICATION

I would like to dedicate this dissertation to my parents Saiprasad and Vijaya

ACKNOWLEDGEMENTS

I would like to express gratitude to my advisor Dr. Nicolas H. Younan for facilitating this opportunity and also for his technical guidance throughout this research. I am also grateful to my supervisor Dr. Valentine Anantharaj for his excellent scientific guidance and support. I would also like to thank my committee members Dr. Lori M. Bruce, Dr. James Fowler, and Dr. Qian Du for serving on my committee and providing valuable feedback and advice. I would also like to extend my sincere thanks to Dr. F. Joseph Turk of the Naval Research Laboratory for providing us research data and insightful discussions. I also appreciate the suggestions provided by Dr. Christa Peters Lidard from NASA, Goddard Space Flight Center. Finally, I would like to acknowledge both NASA and NOAA for providing financial support throughout the research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1 Background

Satellite-based sensors are used to obtain information with large coverage pertaining to applications such as land classification, ocean surface properties, and climate processes. For instance, climate phenomena, such as precipitation, soil moisture, and temperature, are remotely sensed and spatio-temporal data of their approximate states are obtained. The advancement of the remote sensing technology has improved the spatial and temporal resolutions of these datasets. Spatio-temporal analysis techniques include analytical model-based methods, exploratory analysis of geo-spatial patterns in epidemics, and data mining methods for knowledge extraction from large scale geophysical data [1]. Spatio-temporal analysis methods have been successfully used in understanding phenomena such as wildfire events in Florida [2], and land cover/ land use change in the yellow river delta in China [3]. Spatio-temporal analyses of geophysical data include i) recognition of hidden structures in data, ii) anomaly detection in large datasets, and iii) regression to discover temporal trends. These techniques were originally developed for temporal data and are recently extended to study spatio-temporal aspect of data [4].

Pattern recognition and signal processing are emerging tools for spatio-temporal analysis of geophysical data obtained from satellites observations. Broad arrays of methods are available for these applications. Pattern recognition can be defined as an application of machine learning to engineering problems. Some examples of these engineering problems include anomaly detection, data fusion, object tracking and identification, and land surface classification, just to mention a few. In this area, the main objective is to learn hidden structures or processes from a large set of examples and apply that knowledge to analyze new unseen observations. In the context of remote sensing, pattern recognition is mainly used in applications such as image and data classification. Supervised and unsupervised classification of land surface images is a popular application of pattern recognition in remote sensing. For example, a satellite image of an urban area can be classified into different classes based on land use by using a simple classification algorithm.

Two major signal processing tools used for spatio-temporal data analysis are digital spectral analysis and digital filters. Traditionally, spectral analysis tools, such as the discrete Fourier transform (DFT) and short-time Fourier transform (STFT), were developed only for one-dimensional data. These methods have fixed basis functions, for instance, complex exponential for DFT. Later, more sophisticated spectral analysis tools with adaptive basis functions were developed. Some examples include the discrete wavelet transform (DWT), singular value decomposition (SVD) analysis, and Huang Hilbert transform (HHT). An interesting application of wavelets for satellite images is image fusion. In image fusion, several images with varying spatial resolutions are merged

together into a final image which inherits superior qualities of all its contributors [5]. Recently, tools, such as multivariate spectral analysis, have been developed for spatio-temporal signal detection. An example of multivariate spectral analysis is the inquiry of interactions between several climate processes. Most well known global signals include the El-Niño Southern oscillation and North Atlantic oscillation. The ensemble Kalman filter-based methods are most widely used in data assimilation. L-band microwave soil moisture observations from the southern Great Plains hydrology experiment were assimilated into a soil-vegetation-atmosphere model. An optimal ensemble size for robust assimilation performance has been determined for the experiment [6].

The area of focus in this research is spatio-temporal analysis of two key hydrological variables surface soil moisture and precipitation. 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 [7]. 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" [8]. 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 regulates how the ecosystems respond based on available water content in the soils [9]. 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. The other hydrological phenomenon, precipitation, is also an important component of the global energy and water cycle; it is one of the main variables predicted in weather forecast models. Moreover, it is a key process in short-term meteorological and long-term climatological studies. Precipitation events are a driving force behind the hydrological phenomenon, such as floods and storms [10, 11]. These two variables are highly interdependent, for instance, spatio-temporal structure of soil moisture is dependent on long-term variability in precipitation [12 -14]. Based on this fact, the soil moisture observations can be used to estimate errors in precipitation retrievals, and those errors can be corrected by using assimilation with physics based water-balance models [15] . Before this type of assimilation, it is necessary to analyze and improve the consistency and accuracy of respective satellite based retrievals.

In this research, we propose spatio-temporal analysis methods to accomplish the following tasks: (i) consistency analysis of satellite-based soil moisture data, (ii) interpolation of missing data in soil moisture datasets, and (iii) merging of satellite-based precipitation observations. Novel pattern recognition approaches are developed in the first and the third tasks. Existing signal processing methodologies are used and modified in the second task. This dissertation is structured as follows: (i) motivation behind each individual task, (ii) contribution for each application, (iii) discussion of related work in respective fields, (iv) methodologies to achieve the objectives, and (v) implementation, results, and discussion.

1.2 Motivation

1.2.1 Consistency analysis of soil moisture data

The soil moisture dynamics at the surface layer (Figure 1) is highly inter-related to hydrometeorological forcing fields (precipitation, air temperature, incident shortwave and longwave radiation) and other bio- and geophysical parameters, such as vegetation (type, fraction, leaf and stem area indices), topography and soil parameters (type, texture and hydraulic properties). Soil moisture budget can be modeled as a difference between accumulated precipitation and various forms of water distribution such as evaporation, transpiration, runoff and groundwater losses Huang et al. [16]. The spatial scale of the soil moisture is also characterized by the spatial heterogeneity of the vegetation and soil parameters (Figures 2 and 3). The response of the soil moisture is a complex physical process that is determined by both the external hydrometeorological processes as well as the soil hydraulic properties. In the Lower Mississippi River Valley (aka. The Mississippi Delta), the soil moisture depends primarily on the soil texture which is used to determine the soil hydraulic properties [17]. Further, evapotranspiration also exerts a controlling influence on the variability of the soil moisture in this region during most of the year, except during the summer [18], (Anantharaj, V., 2010 – personal communication). Hence, sophisticated signal processing and pattern recognition techniques are necessary to extract and analyze the information content from soil moisture fields at multiple and spatial scales.

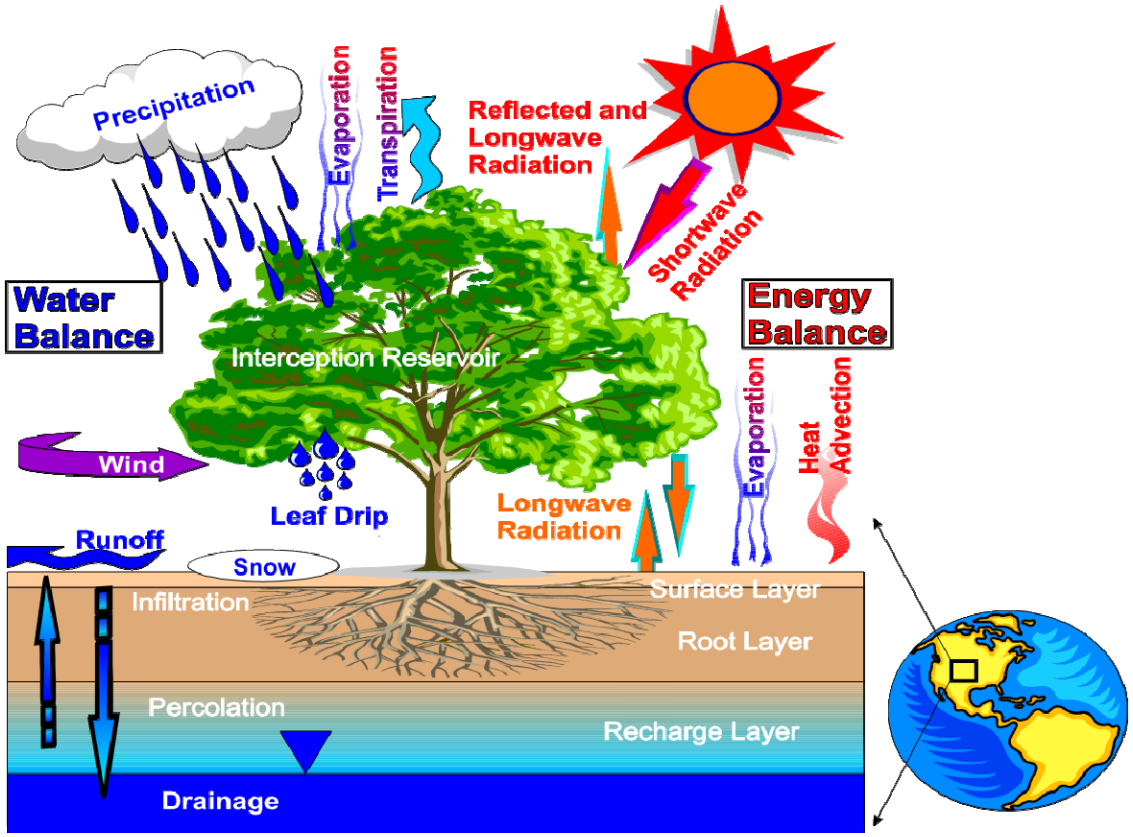


Figure 1. Concept of a Land Data Assimilation System (LDAS) illustrating the partitioning of the energy and moisture fluxes, including response of the surface soil moisture to external forcings (precipitation, temperature and radiation) and the vegetation and soil parameters.

Figure Courtesy: Paul Houser, George Mason University.

A comparative time series analysis of soil moisture data with these land surface processes would provide insight into the above mentioned relationship. Traditional time series analysis tools such as windowed Fourier transform has many limitations such as aliasing for high-low frequencies and determination of optimal window length that make the process of time-frequency localization inefficient. Wavelet analysis is a very popular alternative for such analysis of geophysical time series data. An important objective of wavelet analysis is to understand localized variations of frequency components in the data. Thus, wavelet analysis decomposes the soil moisture time series into sequences at multiple temporal resolutions. These separate sequences in the wavelet decomposition should show the significant signals and their variations with correspondence to the contributions from the individual physical components in the soil moisture model. Parent et al, [19] studied the temporal variability (using wavelet analysis) in soil moisture time series at very short time scales from 1h to 2 weeks. It was found that for scales less than 48h soil moisture is directly related to precipitation events, but for longer scales upto 1week it depends on frequency of precipitation and for even larger scales 1 to 2 weeks it is linked to dry spells. An easy to follow wavelet analysis toolbox for analysis of meteorological time series was developed by Torrence and Compo [20]. A similar wavelet analysis between the soil moisture data and other related land surface processes would provide a better understanding of such physical significance of these wavelet based features. Thus, energy and entropy features constructed from wavelet analysis would be very useful for analyzing the statistical agreement (consistency) between ground based and remotely sensed soil moisture data.

Moreover, soil moisture for a given grid cell is basically an average for a heterogeneous area with different possible land classes. For in-situ measurements, soil moisture budget also depends on the specific soil type (affects ground water loss and evaporation) and land cover (affects transpiration and runoff) (Figure 1). A wavelet analysis of spatio-temporal soil moisture data would address the relation between the soil moisture variations and the corresponding land classes.

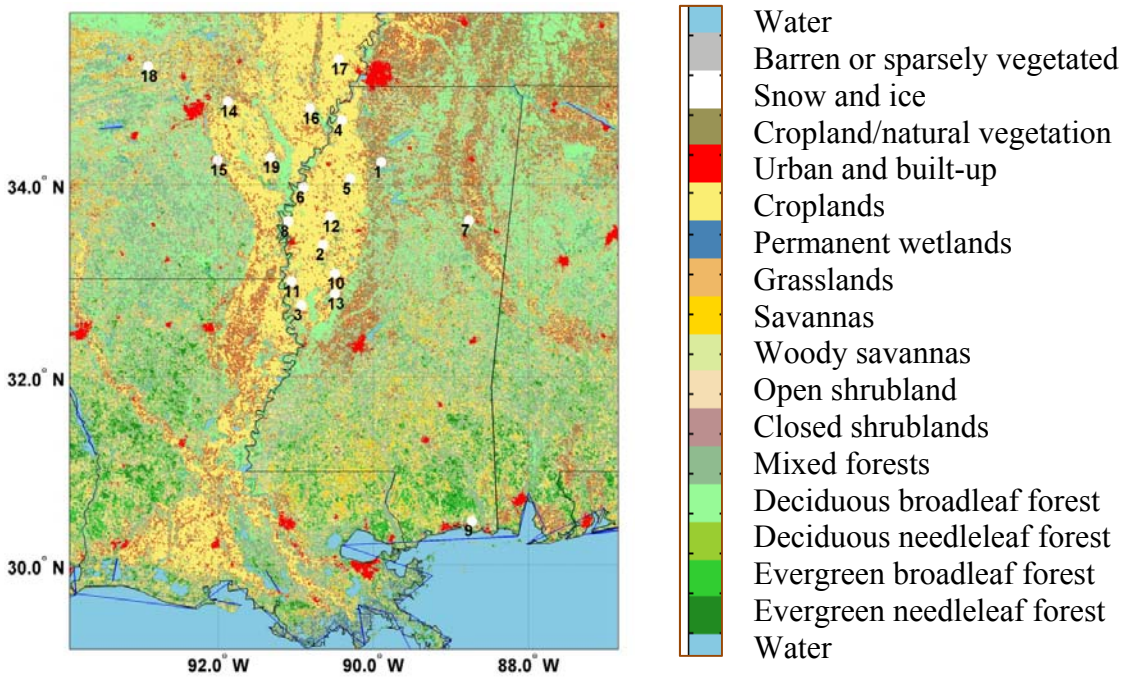


Figure 2. Map of Land cover for the study region. Scan sites are marked with Arabic numerals and site names are given in Table 1.

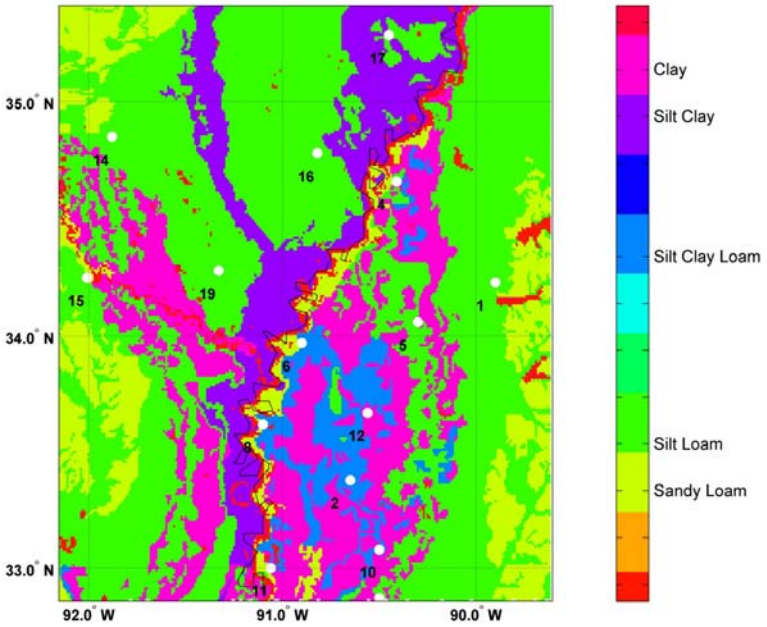


Figure 3. Soil texture map for a part of our study region. Scan sites are marked with Arabic numerals (Figure Courtesy: Mostovoy and Anantharaj [17])

Table 1. List of SCAN sites

No.	Site Name
1	Goodwin Ck Timber
2	Beasley Lake
3	Onward
4	Tunica
5	Vance
6	Perthshire
7	Starkville
8	Scott
9	TNC Fort Bayou
10	Silver City
11	North Issaquena
12	Sandy Ridge
13	Mayday
14	UAPB-Lonoke Farm
15	UAPB Campus-PB
16	UAPB-Marianna
17	UAPB-Earle
18	UAPB Point Remove
19	UAPB Dewitt

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 [8]. An approach to deal with this problem is the use of the Noah land surface model of NASA Land Information System (LIS) [21]. The idea is to downscale the data to higher temporal and spatial resolutions. Before assimilation of soil moisture data into the LIS, the validity of the data has to be verified. In this context, consistency analysis can be defined as an attempt to understand the spatio-temporal quality of satellite-based data with respect to in-situ data obtained at certain stations within the study region for the same temporal duration.

1.2.2 Interpolation of geophysical datasets

The time scales of interactions of the Earth's subsystems are usually in the order of years or longer. These complex interactions result in quasi-periodic and low frequency fluctuations in the climate. A couple of advantages for studying these interactions are a better understanding of the climate and a possible improvement in the forecast of future climate. The complex nature of climatic interactions does not support any single methodology. Periodic components can be best understood using frequency domain methods. However, episodic events, such as volcanic eruptions, can be best studied using time domain methods. There are some phenomena in climate structure which exhibit both oscillatory and episodic behavior, for instance, the El-Niño southern oscillation.

Mann and Park [22] developed the multi-taper multivariate singular value decomposition (MTM-SVD) method, an improvement over the existing spectral analysis techniques, to study couplings between various climatic processes. In a study on a synthetic data set, the MTM-SVD method has detected a spatio-temporal signal that is statistically significant over the underlying noise in the data. Los *et al*. [23] employed the MTM-SVD method on the datasets such as adjusted NDVI from the Advanced Very High Resolution Radiometer (AVHRR), precipitation and land surface temperature from the National Oceanic and Atmospheric Administration's (NOAA) and the National Climate Data Center (NCDC), and sea surface temperature from the National Center for Atmospheric Research (NCAR). A principal mode, strong in sea surface temperature, was found corresponding to a 2.6 year period and related to the El-Niño southern oscillation index. Wu *et al*., [13] applied a SVD-based method to analyze the spatio-temporal relationship between spring soil moisture and summer precipitation in the United States. The NCAR community climate model coupled with multilayer land model (CLM) was analyzed while simulating the US land-atmospheric system. The first SVD mode accounted for 27% of the covariance between soil moisture and precipitation, while the second mode has accounted for 16% of the variance. In a recent work, Kim and Wang [24] studied the influence of soil moisture on precipitation in North America and found that there was a considerable time lag for the soil moisture impact on precipitation. Overall, the SVD analysis has been a successful method for the analysis of the interactions between different phenomena and their overall influence on global climate.

In general, satellite-based sensors provide the most common types of large scale geophysical datasets. Depending on the orbital location of the satellite and various other factors, the resolution and the extent of the satellite image may change. Moreover, the revisit time of the satellite depends on its location and it can vary for a given satellite. For instance, the revisit time of the NASA's Aqua satellite varies from fraction of a day to several days and depends on the latitude of the region. Thus, gaps occur in spatio-temporal datasets due to unsuccessful retrievals, which can reduce the statistical significance of inferences in spatio-temporal inter-relation studies; especially, if there is an important temporal event or a significant signal in the missing portion. These gaps are usually either random or systematic. Random gaps occur due to noise or some unknown reasons and can be relatively easily filled using statistical imputation methods. However, the systematic gaps occur at fairly regular intervals and are related to observation and retrieval methodologies. Thus, these gaps are more complicated and require a sophisticated analysis [25]. For example, systematic gaps are common in satellite-based geophysical data due to a systematic instrument failure. The goal of this task is to improve on existing interpolation methodologies to fill such systematic gaps.

1.2.3 Rainfall measurements and fusion

The amount of rainfall in a given location can be measured by rain gages and estimated over a given area by remote sensing techniques, both from ground-based and space-borne platforms. Rainfall estimates from land-based radars are usually limited to continental land areas and the coastal zone. Even then, there is no uniform radar coverage across the global land areas. For example, the continental United States and

Western Europe have a much better coverage than the African continent. Similarly, the *in-situ* measurements are also not uniformly distributed. For practical reasons, most of the *in-situ* measurements are reported only on a cumulative daily basis whereas radar estimates are routinely available every hour and estimates of rain intensities even more frequently as necessary. Space-borne remote sensing platforms collectively provide nearly global coverage. Most of the low Earth orbiting (LEO) satellites have better coverage near the poles than the tropical regions, and any given geostationary (GEO) satellite observes continuously the region in view and has the capability to make measurements more frequently. Hence, the LEO and GEO satellites have different spatio-temporal sampling patterns. Besides, the LEO and GEO platforms have different kinds of instruments on-board for precipitation estimates. The rainfall estimates from geostationary satellites are typically based on infra-red (IR) measurements of cloud top temperatures whereas most of the LEO satellites make passive microwave (PMW) measurements that can be used to make more accurate estimates of rainfall. The Tropical Rainfall Measurement Mission (TRMM) has active precipitation radar (PR) on-board, capable of making high quality sensing of precipitation ([26, 27].

A number of high resolution precipitation products (HRPP) from satellite observations are routinely produced by various research and operational agencies across the world. In addition, short-term precipitation forecasts are also available from global numerical weather prediction (NWP) models. However, every one of these HRPP products has inherent advantages and limitations, and their performance varies across the seasons. Ebert *et al* [28] verified twelve sets of rainfall products (including products from

numerical weather prediction models) against ground data in the United States, Australia, and Western Europe. The comparisons performed in the Australian region showed that satellite-based rainfall estimation algorithms had greater skill during the summer in tropical regions. However, models were effective during the winter in mid-latitude regions. Both types of products were not very skillful with heavy rainfall. In the continental United States (CONUS), the agreement between the combined IR-PMW rainfall and ground data varied with geographic locations; with a closest agreement in the central states. Also, the products based on PMW only data had similar characteristics as the PMW+IR combined products. However, the rain rates derived from IR only data were severely underestimated in many locations. In the Western Europe studies, the most important finding was that the climate prediction center morphing (CMORPH) algorithm's dataset outperformed all other satellite-based rainfall estimation algorithms, thus, suggesting that CMORH, which is a combination of several PMW estimates and finally merged with IR data, is an effective technique [28].

The Global Precipitation Measurement (GPM) is a flagship mission, involving a group of international partners including the National Aeronautics and Space Administration (NASA) and the Japanese Aerospace Exploration Agency (JAXA). It consists of a constellation of satellites along with a "GPM Core" satellite (to be launched in 2013). The "GPM Core" satellite will have dual-frequency precipitation radar (PR) and a PMW radiometer called the GPM Microwave Imager (GMI). The GPM PR will be used to calibrate the measurements from the rest of the GPM constellation of satellites planned to be launched by the GPM mission partners. The goal is to obtain reliable

observations of rainfall data on a global scale with high spatial and temporal resolutions [29]. These measurements will benefit researchers studying both short-term and long-term meteorological phenomena, specifically over oceans and in areas with little ground based measurements. The GPM is also conceived to be a "science mission with broad societal applications." Besides supporting weather and climate research, the high resolution precipitation products based on GPM are envisioned to benefit a number of applications including human health, disaster management, and agriculture. For instance, the GPM measurements will be useful for some of the Southeast Asian countries, where floods and storms are major issues; as they do not have required ground based measurement infrastructure to observe and respond to these contingencies in a timely manner. Since the GPM-era precipitation products will be based on measurements from a constellation of satellites, it is necessary to develop novel data fusion techniques to merge observations from satellite instruments, with different technical characteristics, capable of monitoring different physical characteristics of the precipitation process.

Satellite-based estimation of precipitation is usually not a direct measurement of rainfall, but it is based on the observation of a closely related physical entity. For example, in the case of microwave-based observations, the scattering properties of water drops in the atmosphere are measured; which are related to the amount of rainfall if there are multiple sets of data available. Thus, the accuracy, coverage, resolution, and consistency of any single precipitation product may not be the best compared to the corresponding properties of any other product at every point in space and time. The idea of multi-sensor data fusion is to combine the information from all the available

measurements and synthesize a new product, which is comparatively better than any given instance of any of the individual dataset over a period of time.

1.3 Contributions

1.3.1 Consistency analysis of AMSR-E soil moisture data

The objective of this task is to compare the spatio-temporal characteristics of the remotely sensed soil moisture estimates from the Advanced Microwave Scanning Radiometer – EOS (AMSR-E) against in-situ soil moisture measurements from the USDA Soil Climate Analysis Network (SCAN). We have developed a consistency assessment method based on wavelet-based feature extraction and one-class support vector machines (SVM). This method performs a consistency assessment of the entire time series in relation to others and provides a spatial distribution of consistency levels whereas conventional approaches typically provide information on every data point individually in relation to its neighbors only. We have applied this new methodology to assess the spatio-temporal characteristics of the soil moisture products from AMSR-E. The in-situ SCAN measurements have been 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 for the years 2005 and 2006. To verify this methodology, the results obtained from this study are correlated with the spatial distributions of the averaged consistency information, mean soil moisture, and the cumulative counts of dense vegetation.

1.3.2 Interpolation of gaps in AMSR-E soil moisture product using modified SSA

Soil moisture data available from the Advanced Microwave Scanning Radiometer-Earth Observation System (AMSR-E) onboard the National Aeronautic and Space Administration's (NASA) AQUA satellite has many inherent gaps. For a region in the Southeast United States, data is collected for years 2005 and 2006. This dataset has nearly 30% missing data due to radio interference, instrument errors, just to mention a few. To address this issue, an adaptive singular spectral analysis (SSA) -based interpolation scheme is presented. For the validation of the interpolation scheme, subsets of NDVI and LST products from moderate resolution imaging spectroradiometer (MODIS) onboard the NASA's TERRA satellite, and SST from GODAE's high resolution sea surface temperature pilot project (GHRSST-PP) are considered. Finally, the presented scheme is tested on satellite soil moisture retrievals from AMSR-E. Optimization of the method is based on minimizing the mean square error (MSE) and it is found to be dependent on the nature of the data. The top two to three dominant SSA modes are usually sufficient for interpolation of missing values.

1.3.3 Precipitation data fusion

In order to evaluate the value added by the GPM-era precipitation products, a Rapid Prototyping Capability (RPC) project has been sponsored by NASA. One of the objectives of this project is to develop and test a data fusion methodology to merge the satellite precipitation products available from different rainfall estimation algorithms. Data fusion is generally used in fusing information from a set of sensors with a common final goal, for example, target identification. In the past, data fusion has been used to

merge satellite and ground based rainfall data. In this study, a fusion method is developed to merge precipitation data available from four different products. The final objective is to develop a product which is better than any individual product at any given spatial or temporal location. The precipitation data from the Arkansas Red Basin River Forecast Center (ABRFC) region is used as reference data. The fusion method is based on binary classification of the input data. A combination of vector transfer function and a two-layer neural network is used as classifier. Initially, the input rainfall data are arranged into vectors with each four precipitation values. Then, these vectors are transformed and scaled into a new vector space using a scaled exponential transfer function. These new vectors are used as inputs to the neural network.

The rainfall information from a small portion of the reference data from the summer of 2007 is used as a target vector in the training process. The trained neural network is used to classify the input vector data and the resulting binary classification data is multiplied with the average dataset of all individual products to produce a final merged product. In order to validate the merged product, it is compared with the reference data using statistical skill scores, such as the Heidke skill score (HSS), critical success score, and bias score. At a given spatial location, if the rainfall time series has a better HSS compared to all other products, then it is counted as a success. Based on this criterion, the merged product has a maximum success rate of 90% during the summer, a minimum rate of 60% during the winter, and 80% during the fall and the spring seasons.

CHAPTER II

LITERATURE REVIEW

2.1 Soil moisture measurements and consistency analysis

2.1.1 In-situ measurements

Soil moisture variability can be monitored using *in-situ* measurements, from observing systems such as the Soil Climate Analysis Network (SCAN) [30] and the Oklahoma Mesonet [31]. Robock *et al.* [9] provide a survey of the history of soil moisture measurements and its importance. Soil moisture measurements on a regular basis started in the former Soviet Union in 1930s at a few agrometeorological stations. These techniques and practices were later adopted by Russia's Asian neighbors, China, and India. 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 [8, 9]. These include: (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.

2.1.2 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 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.* [32] 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. Lakhankar *et al.* [33] proposed a neural network and fuzzy logic based tools to retrieve soil moisture from RADARSAT 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.* [34] 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.* [35] 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 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. They 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 GHz (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.* [36] 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.* [37] proposed a multi-frequency method for retrieving soil moisture from SMMR. This method is capable of correcting vegetation biomass effects using polarization index in X-band. The method was initially tested in southern France and later extended to wider

spatial scales and 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 [38] 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 the dataset is 1deg x 1deg. 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.

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 [39]. Since the signal in the C-band has unacceptable level of RFI contamination [40], 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 25x25 km^2. Additional description of the AMSR-E global Level 3 product is provided in Section 4.1.1.2.

Prigent *et al.* [41] 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 19GHz 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 to 50 km and revisit times of 1 to 3 days. This dataset will also have global coverage and high sensitivity [42, 43]. 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 [44-46]. 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 [47, 48]. NASA's Soil Moisture Active and Passive (SMAP) mission, dedicated to observer 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 [49]. Further, there are also considerations for a Level 4 soil moisture product using a land surface model that will assimilate the SMAP observations (P.R. Houser, July 2008 - personal conversation).

2.1.3 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 land surface model (LSM) and assimilate available observations [6, 21, 48, 50 -53]. 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. Reichle *et al.* [52] 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. The 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 [17]; and (iv) in-adequate data for satisfactory evaluation of the soil skin temperature and surface moisture.

2.1.4 Review of 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*. [54]. 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, ex: 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, [55], 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.* [56] 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 (~ 1cm for AMSR-E) where the point measurements are likely 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.* [57] studied the scales of temporal and spatial variations in soil moisture over extra tropical regions in the United States and Eurasia. The temporal autocorrelation was modeled as an exponential with two components. The 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 the upper 1m soil layers in the 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 task is to develop a methodology to assess the level of agreement between remotely sensed data and *in-situ* measurements (usually sparse). The method 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 the feature space instead of time series in the 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 remotely sensed data sets from satellites and aircrafts.

2.2 Importance of interpolation techniques for geophysical datasets

2.2.1 Spectral analysis of geophysical variables

Standard spectral analysis techniques, such as Fourier transform and wavelet analysis, require complete data for optimal representation in the frequency domain. Several spectral analysis techniques have been developed to compute the spectral properties of incomplete data without explicitly solving for missing data. The CLEAN algorithm, for example, is an iterative process in which the initial dirty spectrum gradually evolves into a clean spectrum through successive subtraction of signal peaks from the residue spectrum [58]. They applied a modified CLEAN algorithm to analyze the spectral properties of an incomplete time series of elastic propagation velocities of a

seismological array in Germany. This analysis excluded a signal corresponding to solid Earth tides as a source of periodicity in the elastic velocity changes.

A singular spectral analysis method is often used to analyze an incomplete time series of suspended sediment concentration (SSC) data. This method allows for extracting spectral information from incomplete data without filling data gaps. Schoellhamer [59] successfully applied this method for extracting the top 10 modes in the SSA spectrum of the SSC data, which correspond to several related physical processes. Smith *et al*. [60] analyzed incomplete $PM_{2.5}$ (particulate matter under aerodynamic diameter 2.5μm) by decomposing the data into a deterministic non-parametric spatio-temporal signal and a spatially correlated random component. A modified expectation maximization (EM) algorithm was used to determine weekly aggregations of the random component by taking the data gaps into account. A wavelet-based analysis can also be used for computing the time variant spectra of a time series with missing data. Moreover, a cross wavelet analysis can be applied to compute the cross-variance between two related time series with gaps [61].

2.2.2 Data gap filling methods

Traditionally, interpolation techniques address one-dimensional data and later extended to two-dimensional data, for example, interpolating missing pixels of spatial images from satellite sensors. The advancement of computational power has facilitated the extension of 2D interpolation methods to address three-dimensional data. Spatio-temporal interpolation, which uses relations in both space and time, has some advantages over traditional interpolation techniques, which work only in either space or time. Gap

filling methods for geophysical datasets can be categorized into four groups: (i) linear regression and non-linear polynomial fitting based interpolation; (ii) parametric model-based interpolation, for instance, fitting a deterministic covariance or a physical model to data; (iii) pattern recognition approaches, such as support vector machines and artificial neural networks (ANN); and (iv) spectral analysis approaches, such as a periodogram technique.

Examples of non-linear interpolation methods include methods such as spline interpolation and nearest neighbor method [62]. According to Li and Revesz [63], parametric interpolation methods for geophysical data are of two types: (i) a reduction approach where the 3D interpolation first interpolates in time and then reduced to spatial interpolation and (ii) an extension approach that treats time as another spatial dimension and the whole problem as a 3D spatial interpolation. Gorban *et al*. [64] developed a gap recovery approach by modeling the data in terms of small-dimensional manifolds. These models can be linear, quasi-linear, and non-linear. A self organizing curve paradigm determines the best non-linear manifold model. This method had successfully worked on an incomplete time series of Carbon-14 (^{14}C) concentrations in the atmosphere. Moreover, a multi-fractal spectrum of reconstructed ^{14}C series had been successfully computed and was applied for understanding the distribution of the concentration structure. An example of an interpolation method based on a kernel technique is the application of support vector machines (SVMs). SVMs have been successfully applied for non-uniformly sampled signal interpolation in the presence of noise for a single-channel time series. Rojo-Alvarez *et al,* [65] developed and compared primer and dual

SVMs for signal interpolation in the presence of Gaussian and impulse noise. The results from tests on a radial basis function (RBF) and sinc kernels were in good agreement with Yen's optimal interpolation method [66]. Moffat *et al.* [67] performed a detailed survey of existing gap filling methods to fill the missing values of CO_2 in an eddy covariance time series. A set of non-linear regression-based techniques, parametric model-based approaches, and artificial neural network (ANN)-based methods were compared. Each technique had its own advantages; however, the ANN-based methods showed a slightly better performance.

The main idea in spectral analysis is to extract the signal spectrum from the existing incomplete dataset and use the most significant modes of this spectrum for the interpolation of missing values. Hocke and Kampfer's [68] Lomb-scargle periodogram-based interpolation scheme illustrates this concept. In the initial step, complex spectral information, i.e., both the amplitude and the phase, is extracted from non-uniformly sampled observations. In the next step, dominant modes are determined and used to modify the Fourier transform with complex spectral information. Finally, an inverse transform of this modified Fourier transform generates uniformly sampled data; thus, filling the gaps. This method was successfully applied in filling gaps in lower mesospheric ozone concentration time series [68]. Beckers and Rixen [69] were among the earliest researchers to apply empirical orthogonal function (EOF) analysis for data filling in oceanographic satellite images. The method relies on EOF analysis of the image itself. A partial reconstruction of the first few EOFs approximates the missing values. The error estimate is the metric for the selection of the optimal number of spatial EOFs.

The tests on advanced very high resolution radiometer (AVHRR) cloud cover images with the first nine EOFs selected for reconstruction were successful [70].

In this task, a modified approach to SSA-based interpolation is developed. A covariance matrix is recursively computed for a spatio-temporal data block instead of a lag-covariance matrix at a single or multiple channels. Several sample datasets of geophysical variables, such as NDVI, precipitation, and land and sea surface temperatures, are used to validate the applicability of the presented method. The mean square error and correlation are used as statistical measures to tune the algorithm parameters.

2.3 Multi-sensor data fusion techniques

Multi-sensor data fusion has recently emerged as an established engineering discipline mainly due to research done and funded by the Department of Defense (DoD). Data fusion has been successfully used in military applications, such as target tracking, identification, and battle assessment. Moreover, fusion has been applied in non-military applications, such as remote sensing, medical diagnosis, overseeing machine building, and robotics. In simple words, a fusion process for target identification consists of the following: (i) a set of homogenous or heterogeneous sensors that observes a phenomenon generating a set of observations [71]; (ii) a set of attributes or features that is extracted from each series of observations to develop a feature set; and (iii) using a suitable classification method to extract a feature set, which, in turn, helps in identifying the target. One of the emerging fields of data fusion is in the area of remote sensing for applications such as image fusion and rainfall data merging.

Recently, fusion of precipitation data is successfully employed for improving several precipitation and related products [72-74]. Nirala [75] proposed a merging method to use rainfall datasets from multiple satellite-based sensors to improve the quality of precipitation estimation. Two of these satellite-based products are from the Advanced Microwave Sounding Unit-B (AMSU-B) and TRMM Microwave Imager (TMI). The importance of high resolution satellite-based rainfall data is recognized as most of the physical models do not perform well in some regions of the world, e.g. Indonesia [73]. Rainfall estimates from IR data have reasonable spatio-temporal coverage but have limited accuracy. The IR measurements, representing the cloud top temperatures, are fitted to a model relating it to rain rates. Hence, the actual rain rates may be different from that estimated ones by taking into consideration the relationship between the rain rate and cloud top temperatures. Generally, more accurate rainfall estimates are derived from microwave observations, which are available from PMW sensors from the LEO satellite but with limited spatial and temporal sampling. Many of the HRPP algorithms have adopted innovative techniques to merge PMW data from different satellites and, in some cases, IR data from geostationary as well. The traditional HRPP blending methods are generally based on either (i) adjustment-based techniques, where instantaneous PMW data from Special Sensor Microwave Imagers (SSMI) are merged with IR data from a geostationary satellite or (ii) motion-based techniques, where IR data at higher spatial and temporal resolutions are used to estimate propagation vectors for microwave data. This approach does not require the IR temperature-rainfall relationship assumption [74]. An example of a motion-based method is CMORPH, a

morphing method developed to merge these two sets of observations using propagating vector matrices. The resulting merged rainfall product was better than the products derived solely from PMW or IR data [73]. Recently, a study was done in the Wu-Tu region of Taiwan to merge satellite rainfall data with gauge observations and flash flood forecasting as the final goal [72]. The method consisted of a linear model for precipitation merging and a hydrological model for flood forecasting. The hydrological model was implemented using recurrent neural networks (RNN). Calibration of the model was done using a set of historical stream flow events. For flood forecasting applications, Chiang *et al,* [72] found that the satellite data contributed only around 4% - 5% toward the merged precipitation. Nevertheless, high resolution precipitation products derived from satellite remote sensing have the potential for improving several other hydro-meteorological and water management applications, such as monitoring water availability.

The objective of this study is to develop an intelligent methodology for merging different observation sets. The proposed approach is different from the traditional HRPP merging methods as it does not require any assumptions in terms of cloud-actual rainfall relations. The fusion tool tries to learn the underlying patterns in the rainfall information from different HRPPs, relating them to actual rainfall over a relatively short period of time, for example, a single season in the year, and use that knowledge to merge precipitation over a longer period, for instance, the entire year. The precipitation observations available from several satellites and ground-based sensors are used to develop downscaled spatio-temporal data on a uniform grid with a spatial resolution of

0.1º by 0.1º and a time resolution of one hour or less. The goal is to merge these different HRPP datasets into an improved product that is better than any individual dataset at any time or location. Our approach is also fundamentally different from the methodologies adopted by other HRPP algorithms; our technique uses a set of estimated HRPP rather than directly using the data obtained from the instruments onboard the satellites. In the subsequent sections corresponding to this task, a detailed analysis of the proposed two-step fusion process followed by a description of the experiments performed on the precipitation data, illustrating the advantages of the proposed approach, are provided.

CHAPTER III

METHODOLOGY

3.1 Pattern recognition based consistency analysis

3.1.1 Foundation

First, consider a spatio-temporal dataset with a spatial resolution of $0.25^{\circ} \times 0.25^{\circ}$ 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 based on time domain methods with simple statistical tools as discussed in the introduction section. In this study, we develop a consistency analysis tool that operates on feature vectors in the feature space instead of time series. The time series data is transformed into a feature space via a 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 methods have been successfully used for hyperspectral signal analysis, image classification and segmentation, and signal detection [76]. 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 [77, 78]. Anomaly detection using a one-class SVM works as follows: The data is 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 method 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 (Figure 4). A simplified feature space for consistency analysis of test samples against a set of training samples may appear as shown in Figure 4. 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 a degree of inconsistency proportional to the statistical distance to the hypothesis class. However, in the actual implementation, the hypothesis class will be the M-dimensional hyper-surface, where M is the number of features in a sample.

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 the *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 a set of distance measures. Finally, in the third step, a consistency assessment is performed to generate a consistency map based on the distance measures. The mathematical details of the three steps follow.

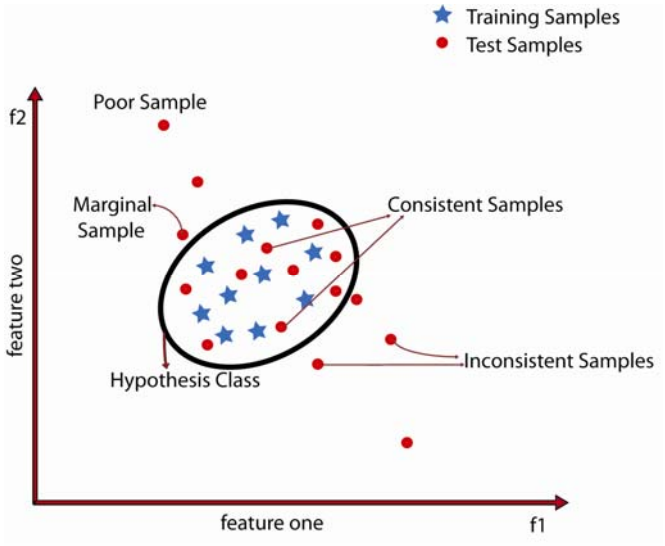


Figure 4. Feature space for consistency assessment of samples with two features

3.1.2 Step one: feature extraction

In general, feature extraction is a two stage process, as shown in Figure 5. 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 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 [79]. The wavelet coefficient based energies are extracted as follows.

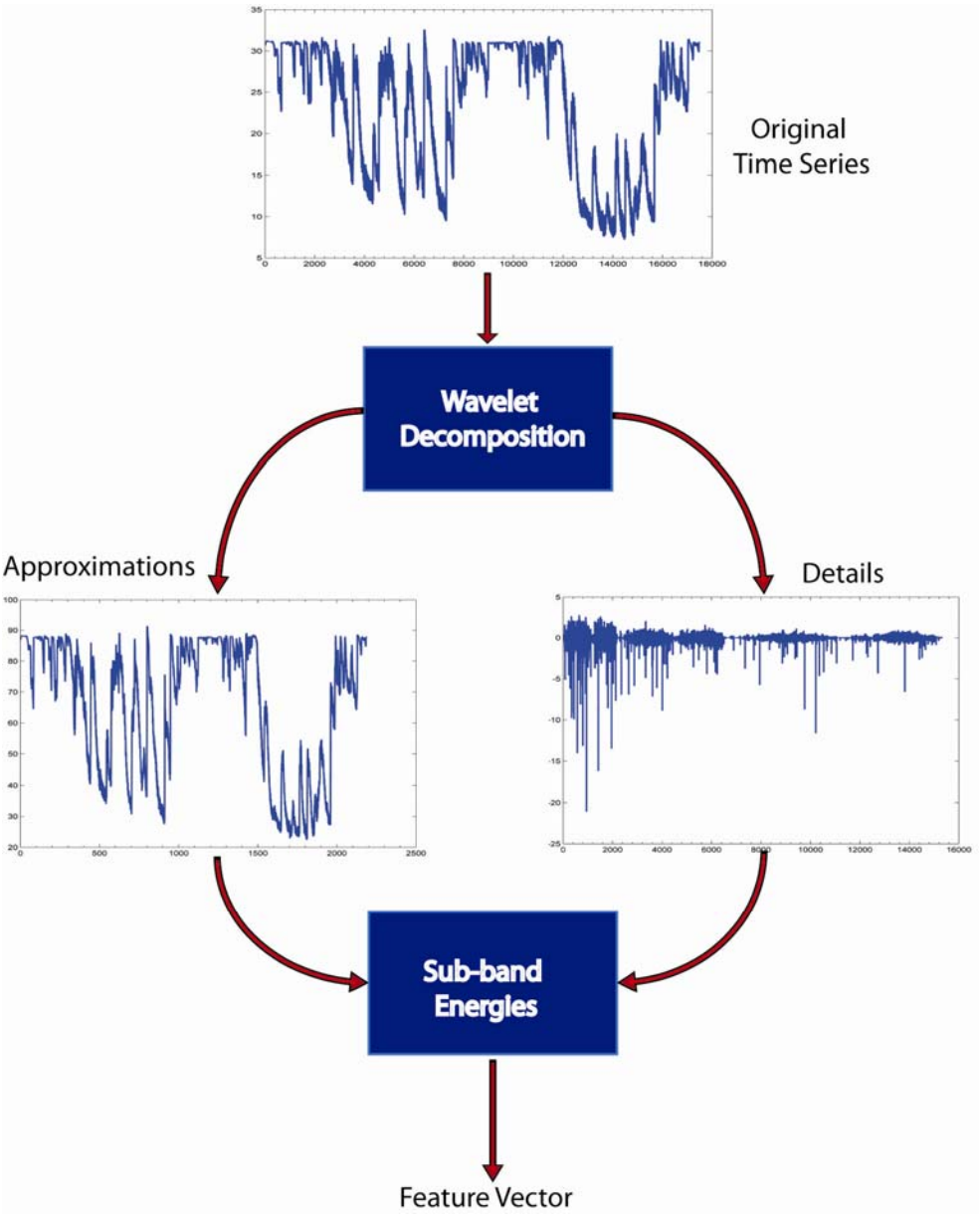


Figure 5. Feature extraction process applied on soil moisture time series from a SCAN site

3.1.2.1 Features from the discrete wavelet transform (DWT)

Let $f[n]$ represent a time series at a cell (s_lat, s_lon) in the satellite image grid, where s_lat and s_lon are the latitude and longitude coordinates of the center of the grid cell and n is the time index. In the case of training data, the point (s_lat, s_lon) corresponds to the ground station. Then, the wavelet coefficients corresponding to the discrete wavelet transform of $f[n]$ 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 highest level J, the initial approximation is

$$c_J[k] = f[n]. \tag{1}$$

Then, the approximations $c_j[k]$ for the successive lower levels are given by

$$c_j[k] = \sum_m h[m-2k]c_{j+1}[m] \tag{2}$$

and the details $d_j[k]$ for successive levels are determined by

$$d_j[k] = \sum_m g[m-2k]c_{j+1}[m]. \tag{3}$$

Here, the filters $h[k]$ and $g[k]$ correspond to the mother wavelet under consideration [80] and m is the index of the convolution process. In practice, Equations (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, $e_j = \sum_k d_j^2[k]$. The feature vector corresponding to the time series $f[n]$ at (s_lat, s_lon) can be defined as $\mathbf{X_{DWT}} = [e_1, e_2, ..., e_M]$. 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.

3.1.2.2 Features from the redundant discrete wavelet transform (RDWT)

For comparison purposes, a RDWT-based feature extraction is also performed. The RDWT is a discrete approximation of the continuous wavelet transform. The filtering operation is similar to the DWT except the filters are recursively up-sampled at each scale, i.e.,

$$h_j[k] = h_{j+1}[k] \uparrow 2, \text{ and } g_j[k] = g_{j+1}[k] \uparrow 2 \quad (4)$$

where $h_j[k]$ and $g_j[k]$ are the filters at level $j = J_1$, the filters are

$$h_{J_1}[k] = h[k] \text{ and } g_{J_1}[k] = g[k]. \quad (5)$$

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 a circular convolution as shown below

$$\begin{aligned} c_j[k] &= h_{j+1}[-k] * c_{j+1}[k] \\ d_j[k] &= g_{j+1}[-k] * c_{j+1}[k] \end{aligned} \quad (6)$$

This process starts at level $j = J_1 - 1$ and ends at the last level $j = J_0$ [81]. Thus, the sequence $c_{J_1}[k]$ is decomposed into the redundant sequence Y as

$$\mathbf{Y} = [c_{J_0} \quad d_{J_0} \quad d_{J_0+1} \quad ... \quad d_{J_1-2} \quad d_{J_1-1}]. \quad (7)$$

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 RDWT is $\mathbf{X}_{\mathbf{RDWT}} = [en_1, en_2, ..., en_M]$, where $en_1, en_2, ..., en_M$ are entropies of the selected M sequences in Y in Eq. (7).

3.1.3 Step two: classification with one-class support vector machines

Let $\mathbf{X} = \{x_i\}_{i=1,2,...M}$ be the M-dimensional feature vector. The feature vector X = X_{DWT} when using DWT for feature extraction and X = X_{RDWT} when using RDWT. 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

$$\mathbf{w} = \Phi(\mathbf{X})\boldsymbol{\alpha} \tag{8}$$

where $\Phi(\mathbf{X})$ is a mapping function on X. If the data are separable, then the decision function is

$$\begin{aligned} D(\mathbf{X}) &= \mathbf{w}^{\mathbf{T}}\mathbf{X} + b \\ D(\mathbf{X}) &= \boldsymbol{\alpha}^{\mathbf{T}}\boldsymbol{\Phi}^{\mathbf{T}}(\mathbf{X})\mathbf{X} + b \end{aligned} \tag{9}$$

where b is the margin pertaining to the hyper-plane. This is also known as the hyper-plane that separates the hypothesis class from other vectors. The objective in a one class SVM is to find a hyper-plane which provides optimal margin and good generalization ability [82-84]. The optimal hyper-plane can be obtained by transforming the problem into a quadratic optimization problem. In the following expression, ξ_i is the slack variable signifying the soft margin support vector machines under consideration.

$$\min_{\alpha} (1/2)w^T w - \rho + (1/(\upsilon l))\sum_{i=1}^{l} \xi_i \tag{10}$$

such that,

$$\begin{aligned} & w^T \Phi(x) \geq \rho - \xi_i \\ & \xi_i \geq 0, i = 1,...,l \end{aligned} \tag{11}$$

In the above expression, ρ is the offset for parameterizing the hyper-plane and thus (w, ρ) if determined would specify the hyper-plane. The parameter $\upsilon \in [0,1]$ is the trade-off parameter. The product υl influences the generalization ability of the SVM classifier where, l is the number of slack variables. If we define a kernel function Q

$$Q_{ij} = k(x_i, x_j) = \Phi(x_i)^T \Phi(x_j), \tag{12}$$

then, by substituting Eq. (8) into Eqs (10) and (11), the dual problem can be reduced to

$$\begin{aligned} & \min_{\alpha} (1/2)\alpha^T Q \alpha \\ & \textit{subject to} \\ & 0 \leq \alpha_i \leq (1/(\upsilon l)) \quad \text{i} = 1,2,3,...,\text{L} \\ & \sum_i \alpha_i = 1 \end{aligned} \tag{13}$$

More details of the derivation from (10) to (12) is available in Schölkopf *et al.* [85]. The above problem can be solved by employing a quadratic optimization approach developed by Coleman and Li [86]. The α vector is obtained from the above optimization [87, 88].

3.1.4 Step three: consistency assessment technique

There exists one α for each training vector. The training vectors corresponding to non-zero α values are treated as support vectors. Let $\mathbf{S} = [s_1, s_2, ..., s_L]$ be a set of support vectors. The set of support vectors is a subset of the training feature vectors. The kernel function measure between the test data vector X and a support vector is given by

$k(s_j, \mathbf{X})$. In this study, kernels such as the Euclidean distance $k(s_j, \mathbf{X}) = \left\|\mathbf{X} - s_j\right\|^2$, Minkowski distance $k(s_j, \mathbf{X}) = (\sum_j \left|\mathbf{X} - s_j\right|^p)^{1/p}$, linear kernel $k(s_j, \mathbf{X}) = \mathbf{X}.s_j$, and exponential radial basis function kernel function $k(s_j, \mathbf{X}) = \exp(-\gamma\left\|\mathbf{X} - s_j\right\|^2)$ to mention a few were used [89, 90].

The vectors of these kernels constitute the kernel matrix K(S,X); given by

$$\mathbf{K(S, X)} = [k(s_1, \mathbf{X}), k(s_2, \mathbf{X}), ..., k(s_L, \mathbf{X})] \quad (14)$$

A statistical distance measure $d(\mathbf{X,S})$ is obtained between a test data vector and the set of support vectors using a kernel function and α, i.e.,

$$d(\mathbf{X,S}) = \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{K}(\mathbf{S}, \mathbf{X}) \quad (15)$$

In this method, instead of using the classifier *y* for assessing a feature vector, we propose to utilize the distance $d(\mathbf{X,S})$ 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 Figure 4. However, the hyper-plane will be a hyper-surface in this case study and the hypothesis class will be a hyper-sphere. The consistency assessment is shown in Table 2.

Table 2. Consistency analysis method

Quality Level	Deviation d from mean (in σ's)	Comments
5	d<= 1	Good quality data
4	1< d <= 2	Acceptable data
3	2< d <= 3	Marginal data
2	3< d <= 4	Poor data
1	d > 4	Very poor or no data

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 methodology is presented in Figure 6.

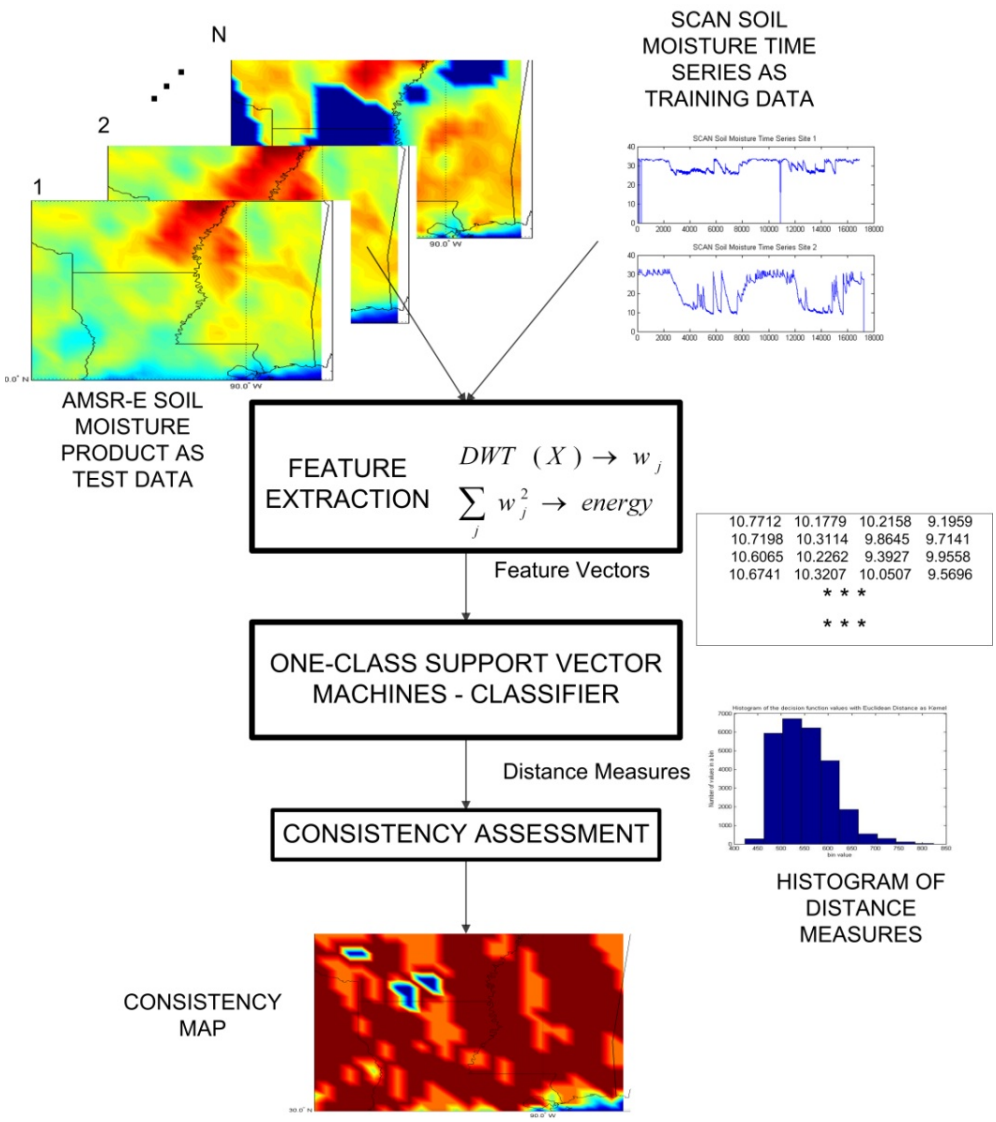


Figure 6. A block diagram of consistency analysis methodology

3.2 Modified SSA-based interpolation

3.2.1 Foundation

Consider a spatio-temporal dataset where several missing points are prevalent in both space and time. The objective is to estimate the missing values from the covariance of the available data. The datasets under consideration are observations of geophysical processes such as land surface temperature, sea surface temperature, normalized difference vegetation index, and surface soil moisture. An important characteristic of these processes is that there are significant spatio-temporal signals embedded in the datasets which can be detected and used for estimating the missing values. Recently, Kondrashov and Ghil [91] developed a multivariate singular spectral analysis (SSA) method to interpolate the gaps by detecting these spatio-temporal signals. The lag-covariance matrix of the time series, instead of the data itself, was used for the singular spectral analysis. The key idea of this approach, known as SSA gap filling, is that the covariance depends only on the lag; thus, the missing data does not pollute the covariance structure. The spatio-temporal EOFs are computed instead of the spatial EOFs and the optimal SSA parameters are determined using a calibration dataset. The SSA method was successfully applied to a global monthly dataset of sea surface temperature. The SSA method is based on computing a lag-covariance c_{ij} for a time series and interpolate each time series separately, i.e., $c_{i,j} = [1/(N-i-j)] \sum_{t=1}^{N-|i-j|} X(t)X(t+|i-j|)$, where X(t) is a data point, |i-j| is the lag, and N is the total number of data points in the time series. The gaps are then filled using the most significant principal components of the lag-covariance

matrix; which is computed for a time series from one or more channels. The SSA method with the reconstruction process is explained in details in [91].

In this task, a modified approach to SSA-based interpolation is developed where the covariance matrix is computed for a smaller spatial-temporal subset. In the multi-channel SSA [91] and the spatial EOF [69] methods, the entire spatial grid of size LxL is considered for the computation of the lag covariance structure. The proposed method optimizes the size of 2D spatial block for the computation of the local covariance (Figure 7). The advantages of the proposed approach are: (i) local spatio-temporal variations instead of distant lag correlations can be exploited to estimate the missing values and (ii) the number of computations required for building the covariance matrix and the corresponding eigenanalyis is reduced significantly without any loss of performance.

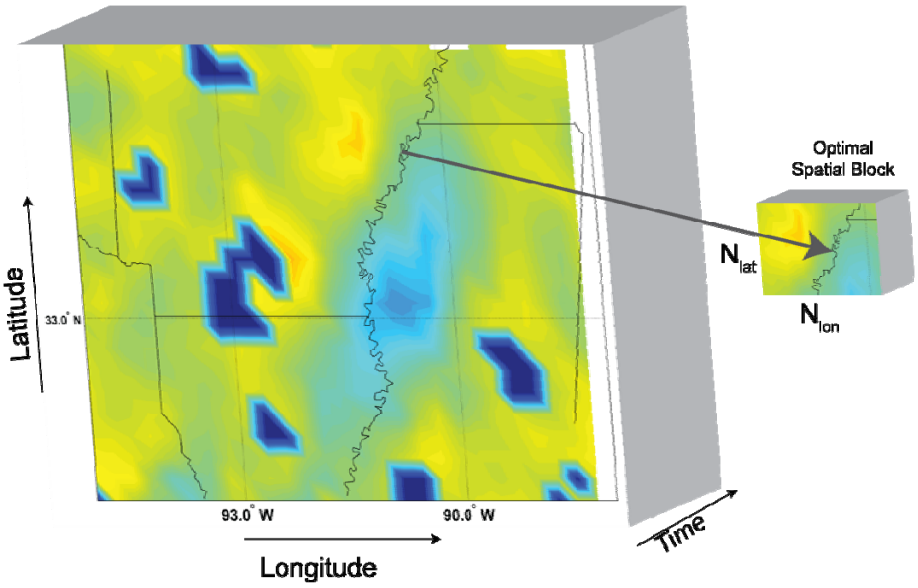


Figure 7. An illustration of spatial grid with missing values and determination of the optimal subset size

The modified SSA-based interpolation scheme consists of two stages: decomposition and reconstruction. In step one, SSA analysis is performed on the covariance matrix of a spatio-temporal subset to generate a set of eigenvectors and eigenvalues. From the set of eigenvectors, the dominant modes that contribute to a major portion of the variance of the dataset are selected. Then, projections of the dataset on each of the mode or eigenvector are computed. In step two, using these projections and the selected SSA modes, the dataset is partially reconstructed. This method is similar to the reconstruction of a time series data from Fourier coefficients or any other types of spectral coefficients [68]. The missing values of the original dataset are filled with the values from the reconstructed dataset.

These two steps are repeated recursively until the mean square error between two consecutive datasets converges to a minimum possible value. Once the gaps in the initial data block are filled, the process is repeated on the remaining blocks in a raster scan fashion. A block diagram illustrating this process is shown in Figure 8. A detailed mathematical analysis of the interpolation process follows.

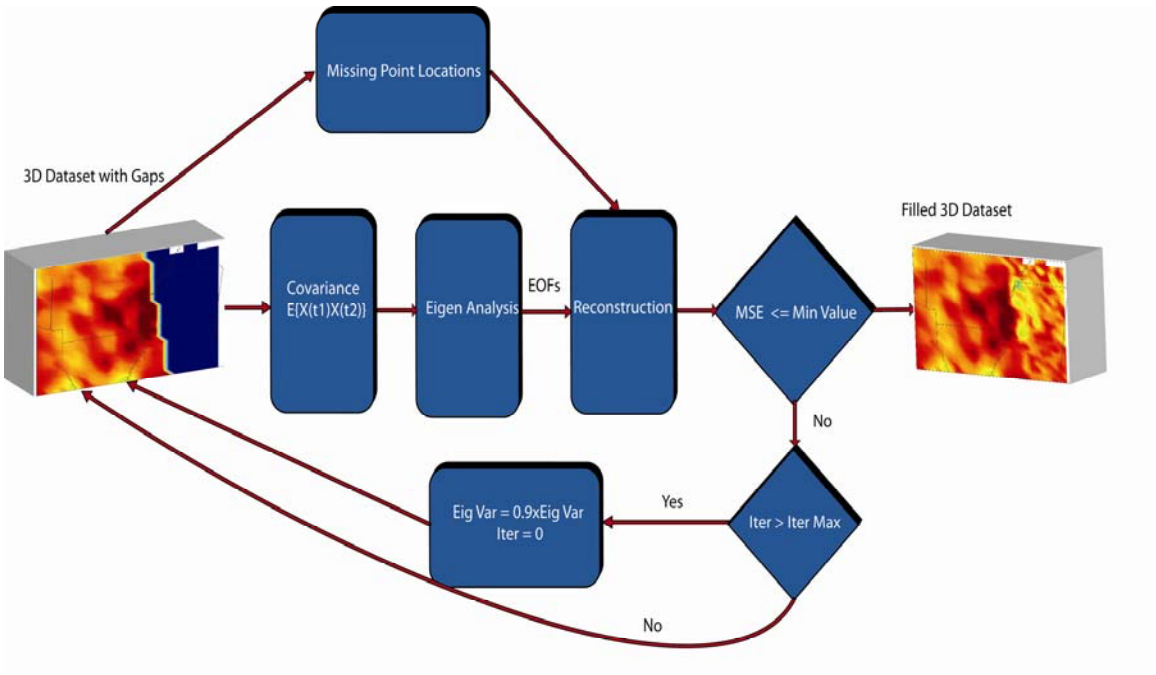


Figure 8. A block diagram of the modified SSA interpolation scheme

3.2.2 Method description

Consider a spatio-temporal sub-dataset of the form $D(s_i, t_j)$, where s_i is a location on the surface of the Earth and t_j is a time step; e.g., hours or days. The optimal size for $s_i = (N_{lat}, N_{lon})$ is shown in Figure 7. Examples of this datasets may include but not limited to observations of geophysical variables. An important assumption with this type of analysis is that there exists an underlying structure within the data which can be determined and used to explain the data properties. Moreover, the gaps in this dataset are assumed to be occurring at fairly regular intervals in contrast to large contiguous chunks of missing data. The missing points $\mathbf{S}_{\mathbf{missing}}$ in the dataset are noted in the set as

$$\mathbf{S}_{\mathbf{missing}} = [(s_1, t_1), (s_2, t_2), ..., (s_M, t_M)] \quad (16)$$

To compute the covariance $C(s_i, s_j)$ of the dataset $D(s_i, t_j)$, temporal means $\bar{D}(s_i)$ are subtracted at a location s_i, and the corresponding covariance of this dataset is computed,

i.e., $$C(s_i, s_j) = \sum_t (D(s_i, t) - \bar{D}(s_i))(D(s_j, t) - \bar{D}(s_j)), \quad (17)$$

with 1 < i, j < *Nblock*. Here *t* is the time index and *Nblock* is the size of a data block [92]. In this computation, it is assumed that enough data points are available in each data block to compute its covariance structure. The size of the data block is important for extracting a suitable covariance structure. For instance, a very small data block may not have enough covariance to approximate the gaps, and the covariance computed from a very large data block may not be very reliable as the number of gaps is also very large. Thus, determining an optimal size for a data block is very important for obtaining a reliable covariance structure.

The orthogonal basis vectors needed for the process are determined from the data. This process is different from other schemes, such as Fourier or wavelet decomposition schemes, where the basis vectors are assumed in advance. The eigendecomposition is applied to the covariance matrix of Eq. (17), and the corresponding eigenvalues and eigenvectors are extracted. In matrix form, this can be written as $C = V*K*V^T$ where V contains the eigenvectors corresponding to the eigenvalues κ_r in the diagonal matrix K. The singular values of D are actually the square roots of the corresponding positive eigenvalues in K. These eigenvalues are arranged in a descending order in K [93-97]. Thus, in analytical form, the decomposition can be expressed as

$$C(s_j, s_k) = \sum_r \kappa_r V(s_j, s_r) V(s_r, s_k) \tag{18}$$

From these basis vectors and the original data, the projections *A(r, t)* in the new vector space are computed as

$$A(r,t) = \sum_{k=1}^{N1} D(s_k, t) V(r, s_k), \tag{19}$$

where N1 is the total number of SSA modes in this decomposition. In matrix form, this projection is expressed as A = DU. The SSA mode corresponds basically to the projection of the data on a single orthogonal basis vector and A(r, t) corresponds to the data in the new transformed space, which is similar to the Karhunen-Loeve transformed data.

The original data matrix can be obtained by projecting the matrix A back to the original data space, i.e., $\mathbf{D} = \mathbf{A}\mathbf{V}^{\mathbf{T}}$. However, the objective is to determine the missing values in the original data. This goal can be achieved by computing the partial reconstruction $R(s_i, t_j)$ of the original data. The dataset R without the missing points is obtained by projecting the first few modes of the transformed data *A(r,t)*, at location r and time t, into the original data space, i.e.,

$$R(s_i, t_j) = \frac{1}{M1} \sum_{k=1}^{M1} V(s_i, s_k) A(s_k, t_j), \tag{20}$$

where M1 is the selected number of SSA modes for reconstruction [98]. In matrix form, this reconstruction is expressed as $R = A_{M1} V_{M1}^T$. The values in R at the locations $\mathbf{S}_{\mathbf{missing}}$ are then inserted in the dataset D. This yields the following:

$$R(\mathbf{S}_{\mathbf{missing}}) \rightarrow D(\mathbf{S}_{\mathbf{missing}}) \tag{21}$$

However, the reconstruction by a single iteration of decomposition and partial projection rarely estimates the missing values accurately. In order to improve the approximation of the missing values, the process is iterated using the interpolated data from the current iteration as the input data for the next iteration. An important feature of this iterative process is that the covariance structure and the interpolated values improve each other recursively. One of the parameters that need to be adjusted is the number of significant modes M. This can be accomplished by using the k^{th} partial variance as the tuning parameter. The k^{th} partial variance can be expressed as

$$ev(M1) = \sum_{i=1}^{M1} \lambda_i \Big/ \sum_{j=1}^{N1} \lambda_j \ , \tag{22}$$

where ev(M1) is the partial variance up to the $M1^{th}$ eigenvalue. The k^{th} partial variance is defined as the amount of variance the first k singular values account for in the total data variance. For instance, the first singular value or the projection of the data onto the corresponding eigenvector or basis vector accounts for the highest fraction of the variance of the original data. As more singular values are taken into account, more variance can be accounted for.

This process, as defined by Eqs. (16) to (22), is repeated on each reconstructed data set and the adjustment of the value for M is carried out at each iteration. This process is repeated until convergence occurs, i.e., the mean square error between the datasets of two successive iterations is minimized [69, 91, and 99]. If this iterative process does not result in a reasonable mean square error after several iterations, say Max_iter, then, the partial variance for the next loop of iterations is reduced by a small fraction. Initially, the interpolation process is performed with a very high partial variance, such as 95%, which

results in the selection of several SSA modes that may include noise components of the data. The partial variance parameter is reduced by 10% for the next outer loop. This outer loop is iterated until either convergence is achieved or the value of M reaches unity. If convergence cannot be reached for the last iteration of the outer loop (i.e. for M =1), then it can be inferred that the original dataset does not possess enough covariance information for interpolating the missing values. However, if convergence is successfully reached, then the reconstructed dataset with gaps filled can be used. The next and important step in the algorithm is optimization of the spatial block size $(N_{lat}, N_{\log})$. This can be achieved by cross-validation as follows: The optimal parameters $\{N_b, M_{SSA}\}$ can be determined by performing a grid search on the mean square error surface. Here $N_b = N_{lat} = N_{long}$ and M_{SSA} is the optimal number of the SSA modes required.

3.3 **Data fusion: a two-step process**

3.3.1 Pattern recognition-based fusion

Consider a set S consisting of M sensors providing simultaneous observations of a physical phenomenon, say rainfall occurrence ξ, at a grid cell in the observation space. The observations from the set S are $\mathbf{X} = \{x_1\ x_2 \ldots x_M\}$, where each x_i is related to ξ. For instance, if ξ is a screening parameter for rain detection in a grid cell, then the problem is a binary classification problem with $\xi \in \{0, 1\}$ and $x_i \in \{0, 1\}$. Here, ξ =1 indicates rain occurred in a pixel and ξ =0 indicates no rain occurred. The objective of multi-sensor data fusion is to find a function, $fuser(\bullet)$, such that $\hat{\xi} = fuser(x_1, x_2, \ldots, x_M)$ is the best estimate of ξ. Assuming $P(x_i, \xi)$ is the probability distribution function of x_i, then ξ, the

best fuser, can be determined by minimizing the expectation error *I(S)*, given by, $I(s) = \int C(\xi, x_i)\ dP(x_i, \xi)$, where C(ξ, x) is the cost function. If the distributions $P(x_1, \xi)$ are known, that is the error characteristics of each sensor are completely understood, then, the fuser is uniquely determined. However, the distributions $P(x_i, \xi)$ are generally unknown in many practical applications. A feasible approach to solve this problem is to minimize the empirical risk, $I_{emp}(f)$ [100], defined as

$$I_{emp}(\hat{\xi}) = \frac{1}{l}\sum_{j=1}^{l}[(\xi_i - \hat{\xi})^2], \tag{23}$$

where *l* is the sample size, i.e., the total number of available observations.

With this objective in mind, we propose a function $fuser(\bullet)$ as a feed forward neural network augmented by a vector space transformation. Thus, the fuser function has two key components and is defined as $fuser(\mathbf{X}) = f_{ANN}(f_{VT}(\mathbf{X}))$, where $f_{ANN}(\bullet)$ is the feed forward neural network classifier and $f_{VT}(\bullet)$ is the proposed vector transformation function. The block diagram of Figure 9 shows the individual steps used in the fusion methodology.

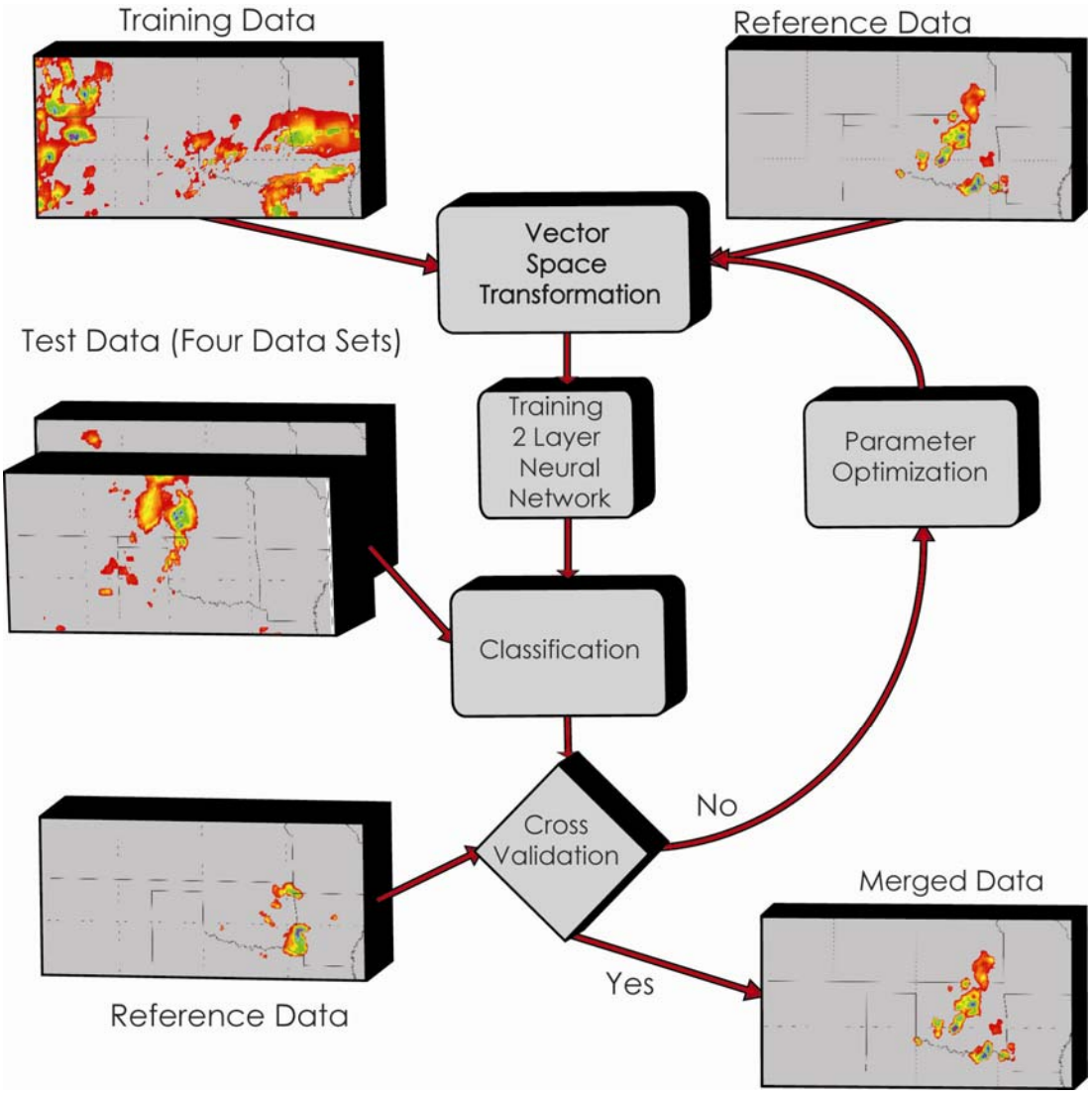


Figure 9. A block diagram of the fusion process

3.3.1.1 Vector space transformation

In general, a transfer function is used as an activation function in a neural network to obtain the output in a predefined range. For instance, a sigmoid transfer function would make sure that the output is in the range of [-a, a], where *a* is a positive number. In this study, a new approach is developed, where a non-linear function is used to transform the input into a new feature space before entering the network. The range of this new feature space is artificially limited; thus, the vector transformation controls the range in the input feature space and as a result influences the output space of the network.

Consider the set of observations $\mathrm{X} = \{x_1,\ x_2,\ ...,\ x_M\}$; clearly these observations can be viewed as a vector in the P dimensional vector space V, where V ϵ R^P. If the observations are of a positive physical quantity, then the range of any component of X is [0, ∞). Consider the vector transformation function $f_{VT}(\bullet)$, which transforms the vector X from the input vector space into a new vector space T, where the range of any component is [0, ρ) with $\rho << \max(\mathbf{V})$. Let the transformed observation vector be $\mathbf{Y} = [y_1\ y_2\ \ ...\ y_M]^T \in \mathbf{T}$, where *T* denotes transpose, $\mathbf{Y} = [\,f_{VT}(x_1)\ f_{VT}(x_2)\ \ ...\ f_{VT}(x_M)]^T$, then, the range of the space T can be controlled by an appropriate selection of the function $f_{VT}(\bullet)$ and thus influencing the fusion rule estimation. For example, consider a set of rainfall observations from five different sensors, $\mathrm{X_{rf}} = \{0, 0.5, 2.3, 0, 7\}$ ranging over [0, 7], and let the transformation function be $f_{VT}(\mathbf{X}_{rf}) = \exp[-(X_{rf} - 0.01)^2]$ then, the transformed vector is Y = [0.99 0.78 0.005 0.99 0]^T. Accordingly, the range is [0 1). Thus, the use of an exponential function controls the range in the case of this sample rainfall data. The selection of the non-linear transformation function is application

dependent. For instance, if the objective is a binary classification of the input features, then, a suitable transformation function will be either a Gaussian or an exponential function of the form of $\exp(-\mathbf{X}^{\gamma})$. This type of transformation will emphasize the separation between features of different classes.

3.3.1.2 Artificial neural networks

Artificial neural networks (ANNs) are well known for their success in the field of pattern recognition. Recently, ANNs are also applied in the field of data fusion. ANNs play different roles in data merging methods. For instance, they can be used to search for optimal merging parameters in the case of a linear merging methodology [72], or they can be used to classify features from multi-sensor data in the field of target recognition [101, 102].

The architecture of a two layer feed forward neural network used in this study is illustrated in Figure 10. Once the transformed feature vector is obtained, this feature vector can be classified by the feed forward neural network to obtain an output $\hat{\xi} = f_{ANN}(\mathbf{Y})$, which estimates the rain occurrence ξ. As shown in Figure 10, $f_1(\bullet)$ and $f_2(\bullet)$ are activation functions in the hidden and output layers, respectively. Table 3 shows the corresponding parameters used in a feed forward artificial neural network. Here i *and* j are indices corresponding to the input and hidden neurons, respectively. Let y_i be the input to the *jth* hidden neuron with a bias b_j; the output of this neuron is given by $p_j = b_j + \sum_{i=1}^{Nin} w_{ji} y_i$, where j = 1, 2, …, Nh and i = 1, 2, …, Nin with Nh and Nin being

the total number of the input and hidden neurons. The output of the same neuron is $q_j = f_1(p_j) = f_1(b_j + \sum_{i=1}^{Nin} w_{ji} y_i)$. In the output layer, the input to the neuron is $z_{in} = \sum_{j=1}^{Nh} v_j q_j$. Thus, the final output to the neural net is $\hat{\xi} = f_2(z_{in})$ and, upon expansion, the final output of the ANN can be expressed as

$$\hat{\xi} = f_2(\sum_{j=1}^{Nh} v_j f_1(b_j + \sum_{i=1}^{Nin} w_{ji} y_i)) . \tag{24}$$

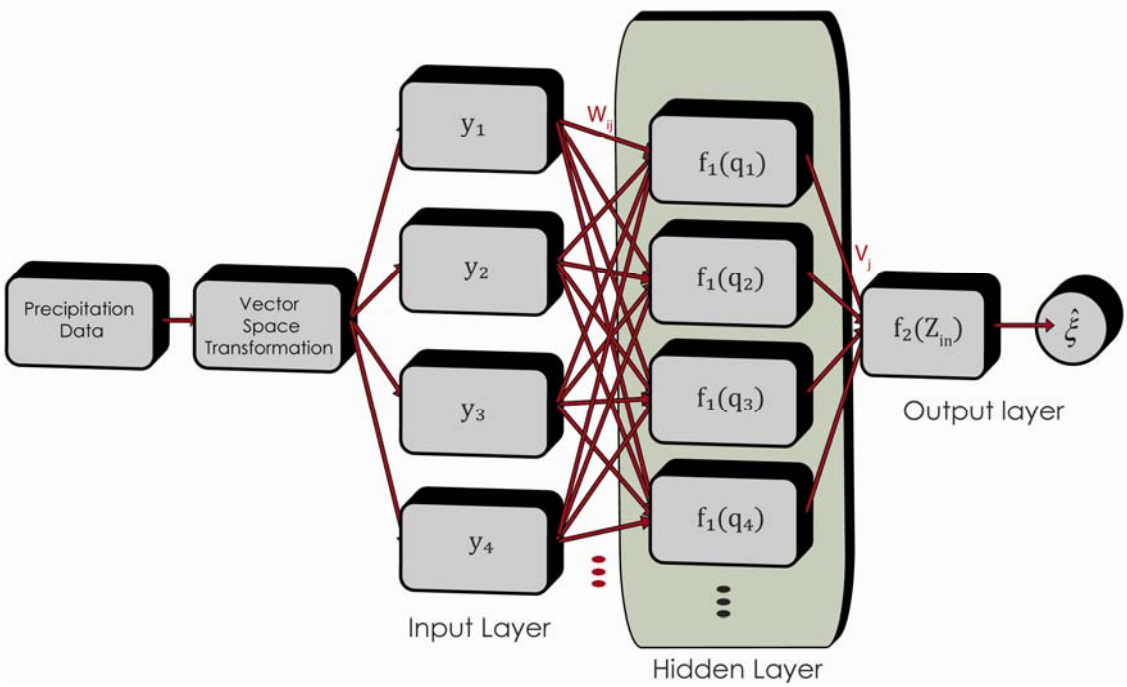


Figure 10. A two layer artificial neural network architecture with vector transformation function

Table 3. Parameter list for the feed forward artificial neural network

layer	index	weights on connections	no. of neurons	activation function	weighted sum	neuron output
hidden	i	w_{ji}	Nin	logsigmoidal	p_j	q_j
output	j	v_j	Nh	threshold function	z_{in}	$\hat{\xi}$

In this study, if the weighted sum input to the neuron is a, then, the activation functions are $f_1(a) = 1/[1+\exp(-a)]$ and $f_2(a) = a$. The inputs are mapped onto a new feature space using the following transformation:

$$y_i = f_{VT}(x_i) = \rho * \exp(-\beta(x_i + \alpha)^{\gamma})/(x_i + \kappa) \ , \tag{25}$$

where the parameters $\alpha, \beta, \gamma, \kappa,$ and ρ determine the nature of feature transformation and these parameters need to be optimized. Thus, the relation between the actual observations and the fused output can be expressed as:

$$\hat{\xi} = f_2(\sum_{j=1}^{Nh} v_j f_1(b_j + \sum_{i=1}^{Nin} w_{ji} f_{VT}(x_i))). \tag{26}$$

the fuser learning process is done in two stages. In stage one, back-propagation with a momentum term and adaptive learning rate is employed for learning the optimal parameters of the feed forward neural network. In stage two, the parameters of the vector transformation function in Eq. (25) are determined using an evolutionary approach based on ant colony optimization.

3.3.1.3 Stage one learning

The weight parameters in the output layer are updated by a three step back-propagation algorithm and can be expressed as:

$$v_j(k+1) = v_j(k) + mc * \Delta v_j(k) + \eta(k) * mc * \delta_{out}(k) * f_2'(z_{in}) \ . \quad (27)$$

Here, k is the iteration index, mc is a momentum term, $\eta(k)$ is the adaptive learning rate, and $\delta_{out}(k)$ is the error between the network output and the desired output. For the pure linear function $f_2(a) = a$, the derivative is one, i.e., $f_2'(a) = 1$, which simplifies the computation of the third term in Eq. (27). The weight parameters in the hidden layer are also similarly updated by back-propagation, i.e.,

$$w_{ji}(k+1) = w_{ji}(k) + mc * \Delta w_{ji}(k) + \eta(k) * mc * \delta h_j(k) * f_1'(p_j) \ . \quad (28)$$

For the log sigmoid function $f_1(\bullet)$, the derivative is a function of $f_1(\bullet)$, i.e., $f_2'(a) = f_1(a)(1 - f_1(a))$. The error term for the hidden layer is usually computed by back-propagation as a weighted sum of all the output errors. However, in this algorithm, only one output neuron is used. Accordingly, the error is given by $\delta h_j(k) = v_j \delta_{out}(k)$. The bias parameters for the hidden neurons can be computed along similar lines. For further details of the back-propagation algorithm, refer to [103-105].

3.3.1.4 Stage two learning

A parameter search method, analogous to ant colony optimization, as described by Dorigo *et al*, [106], is used to determine the best set of various parameters, such as (1) the parameters of the vector transformation function and (2) the optimal parameters for the feed forward neural network corresponding to the best fuser performance. The

optimization process is an evolutionary approach in a cross-validation setting. These parameters of the transformation function would be useful in characterizing the methodologies used for obtaining the observations. For instance, consider β in Eq. (25) whose value can determine the influence of the corresponding input feature on the fuser output. A detailed list of parameters considered are: (a) the coefficients α and β in the vector transformation function in Eq. (25), (b) the exponent γ, the scaling factor ρ, and zero correction for the base κ in the transformation function, (c) the minimum error for training the neural network, (d) the number of neurons in the hidden layer, and (e) the momentum term. The evolutionary optimization process can be described as follows. The objective of the optimization process is to determine the global optima of the performance hyper-surface. The number of dimensions of the hyper-surface is $D_{hyper} = N(f_{VT}) + N(f_{ANN})$, where $N(f_{VT})$ is the number of parameters in the transformation function. For instance, the term $N(f_{ANN})$ represents the parameters related to the neural network, including (a) the minimum training error, (b) the number of neurons in the hidden layer, and (c) the momentum term. The correction factor κ in eq. (25) is assumed to be equal to β. A set of agents is defined such that each agent basically holds a vector *pvec*, expressed as

$$pvec = [\beta_1, \beta_2, \ ..., \beta_M, \rho, \gamma, \alpha, \min err, Nh, mc]. \tag{29}$$

Let the agents in the evolution process be termed as ants *ant{j}*, for better analogy. The objective is to move these ants on the hyper-surface, such that, as a group, they have to determine the global optimum. The number of ants is set equal to the dimensionality of the hyper-surface, i.e., $N_{ants} = D_{hyper}$. Initially, this set of ants starts

with randomly assigned parameter vectors. However, any two ants in the set will only differ in one dimension. In the first step of the evolution process and for the j^{th} ant, using the parameters in its vector, the data is transformed into the new feature space and the neural network is trained on a suitable training set. The training data selection is described in section 4.3.1.3. This trained network is used to fuse a season of data and a performance metric, *acc,* corresponding to this ant, is computed. The same process is repeated for all the ants in the set. The ant with the best performance metric $best\_acc$ is selected as the $best\_ant$ and all the other ants are assigned to the same vector for the next step in the evolution process. In the subsequent steps, if the current best performance is better than the previous best metric, then, the overall best metric and the best ant are updated. This process is repeated until a global maximum is reached by the ant group. A step-by-step procedure is presented below. Assume that the range R for each *pvec* is known, and let the maximum possible accuracy be *Amax*, then, the steps involved in the optimization process are:

Step 1: Initialize the N_{ants} ants

Step 2: For each ant{j}, randomly select a value in R for *pvec*

Step 3: For each ant{j}, train and test the neural network and record the accuracy in acc{j}

Step 4: Find the maximum among all the acc{j}, i.e.,

if max(acc) > best_acc

then best_acc = max(acc)

best_ant = ant{k}; k – value of j for which acc{j} = max(acc)

Step 5: If best_acc is updated

Then for all ant{j} = best_ant

Step 6: Repeat steps 2 to 5 until the number of iterations or *Amax* is reached

3.3.2 Cross-validation

In neural network research, a methodology is evaluated using cross-validation methods, such as leave-one-out validation, hold-out method, or k-fold cross-validation as discussed in [107]. In this study, a repeated hold-out cross-validation methodology is employed. The optimal values for the parameters of Eq. (29) are determined in a ¼ split hold-out cross-validation setting. In order to perform cross-validation of the fusion methodology, a metric is required to assess the performance of each parameter set in step 4 of the optimization process. The evaluation methodology used in this study is described in section 4.3.1.4.

CHAPTER IV

RESULTS AND DISCUSSION

4.1 Implementation of pattern recognition based consistency analysis

4.1.1 Time series generation

4.1.1.1 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 +/-0.03 volume fraction of water. The robustness of the instrument and the measurement technique were well tested for different soils, temperatures, and precipitation conditions [30, 108-111].

Soil moisture data sets are collected from 21 SCAN sites, shown in Figure 11(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 hours 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).

4.1.1.2 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 coordinates. 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 classification for each grid cell. 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 [39] and Njoku *et al,* [112]. 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 25x25 km^2 equal area scalable earth grid (EASE-grid), global cylindrical and equal area projection true at 30° N and S [39, 112].

Accordingly, soil moisture data fields are extracted and data sets are generated for a region of 28x23 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 [39, 112-115].

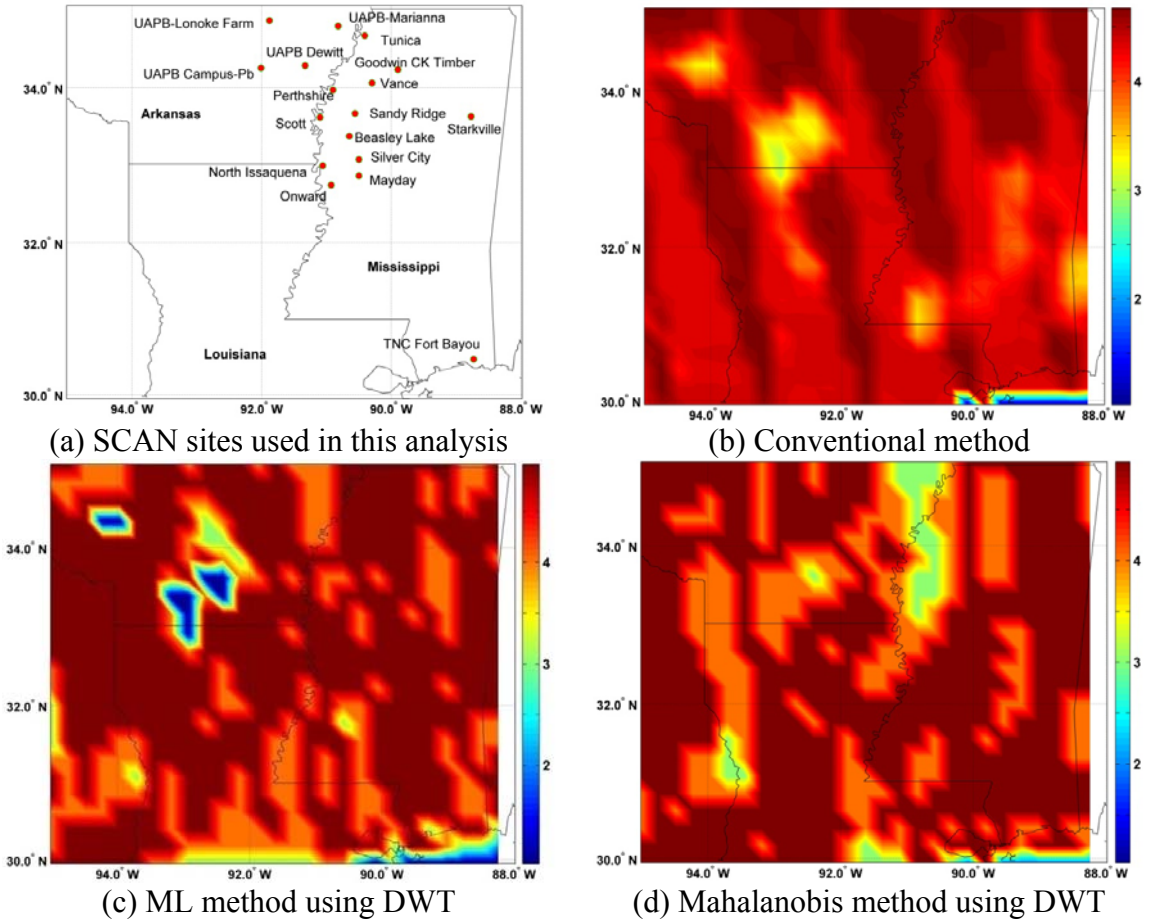
The time series data thus obtained is used to generate the wavelet features. Three-level wavelet decomposition was applied to the time series [116]. The energy features were calculated from the four sub-bands; the first one represents the approximation sub-band, and the others represent the three levels 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.

4.1.2 Implementation

From the α 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 σ, 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 [117]. The same weight vector also multiplies the test feature matrix. The resulting consistency map in this case is shown in Figure 11(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 [118]. 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 Figure 11(d). These maps are compared to the results from the machine learning (ML) method.



(a) SCAN sites used in this analysis (b) Conventional method

(c) ML method using DWT (d) Mahalanobis method using DWT

Figure 11. Scan sites and consistency maps

4.1.3 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 [54].

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 quality control (QC) data is time averaged and spatial distribution of consistency is developed. The scale of Figure 11(b) is adjusted using linear transformation [6-(conventional QC value)/6.4]. This converts it to the scale 1 to 5. Thus, this scale is similar to scales in other consistency maps in Figures 11(c and d). The correlation between the machine learning (ML) map and the conventional map is nearly 60%, which suggests that 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 (kNN) [84]. Treating consistency levels of time series as classes implies that 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.

4.1.3.1 Sensitivity studies

The robustness of the proposed machine learning algorithm has been verified 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 versus the individual site dropped. From Figure 12, it can be inferred that the average distance measure is fairly constant.

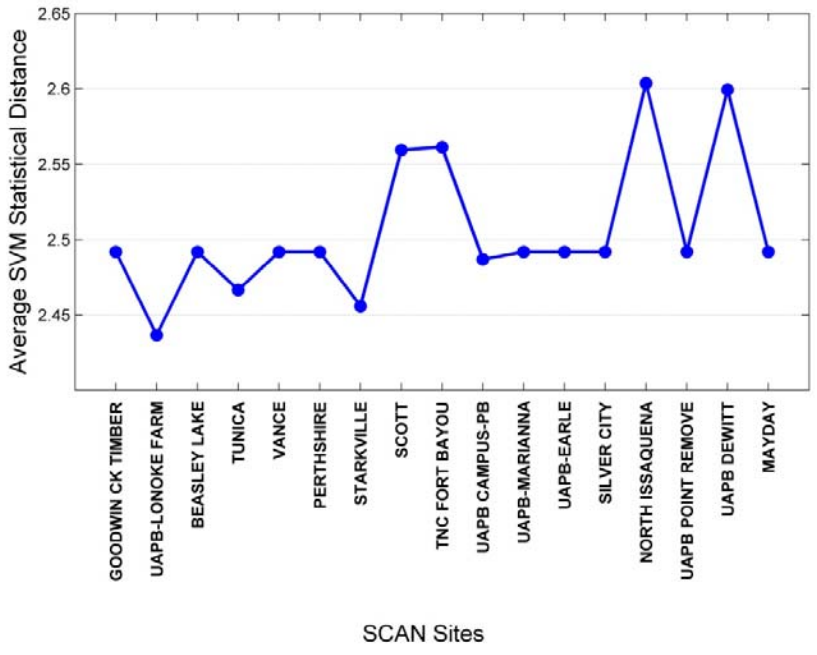
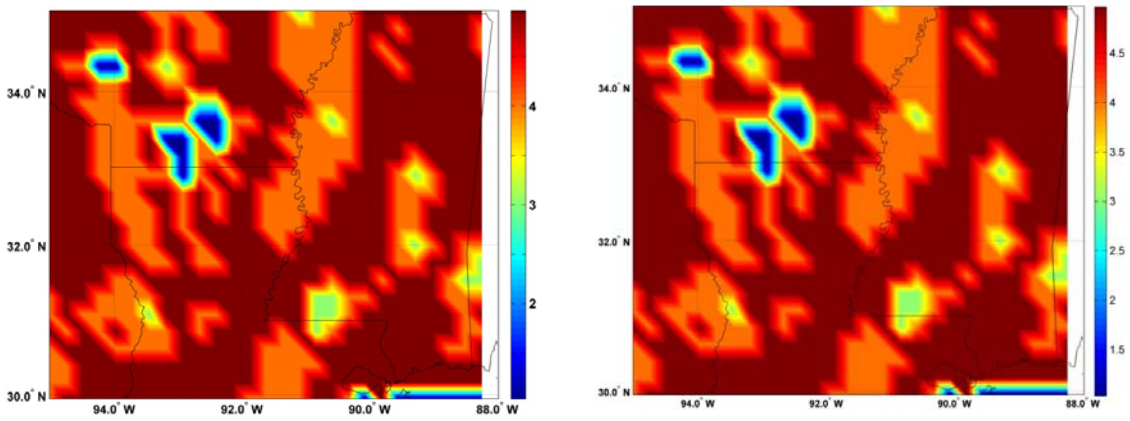


Figure 12. Sensitivity plot: average SVM distance measure versus SCAN site dropped

To further illustrate the robustness of the algorithm, a consistency map is presented for an instance in which selected scan sites are dropped and shown in Figure 13.



Site 1 Site 5

Figure 13. Consistency maps with SCAN sites dropped

4.1.3.2 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 the 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.

4.1.3.3 Performance comparison and seasonal variation

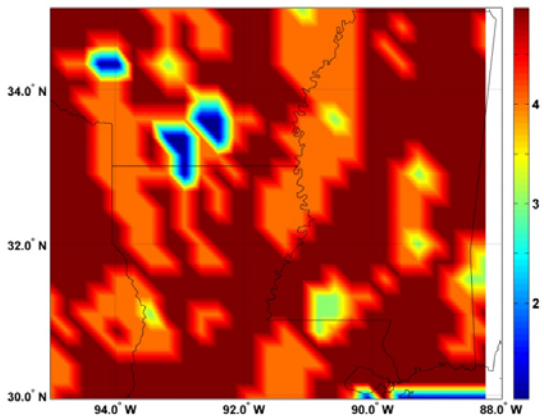
The performance of features from the DWT is compared with the performance of features from the RDWT. For different kernels, the entropy features from the RDWT

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 the DWT and RDWT features and the above mentioned distributions, as shown in Table 4.

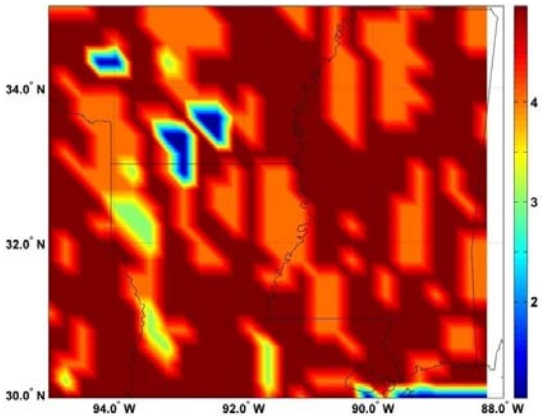
Table 4. Performance comparisons

Kernel	WT type	Features	Correlation with Mean Consistency Distribution	Correlation with Mean SM	Correlation with DVEG Dist
Minkowski	RDWT	Entropy	-0.51	0.44	-0.27
Linear	RDWT	Entropy	-0.66	0.56	-0.53
Minkowski	DWT	Energy	-0.56	0.55	-0.27
Mahalanobis	--	Statistical	-0.41	0.51	-0.16

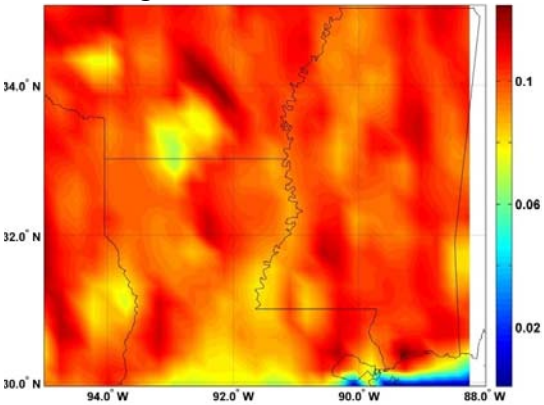
Consistency maps from the RDWT features are presented in Figure 14(a) and Figure 14(b), which correspond to linear and Minkowski kernels, respectively. Since the original analysis was performed for a period of two years, the performance is also tested for individual seasons. The consistency maps are illustrated in Figures 15(a) to 15(d) for fall 2005 to fall 2006.



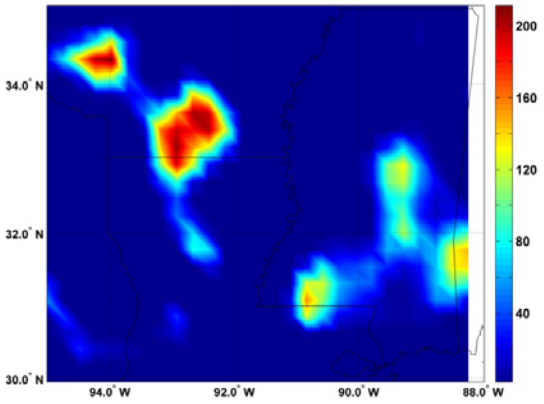
(a) Consistency map from the ML method using RWT and linear kernel



(b) Consistency map from the ML method using RWT and Minkowski kernel



(c) Spatial distribution of mean soil moisture



(d) Spatial distribution of dense vegetation counts from AMSR-E.

Figure 14. Consistency maps comparison

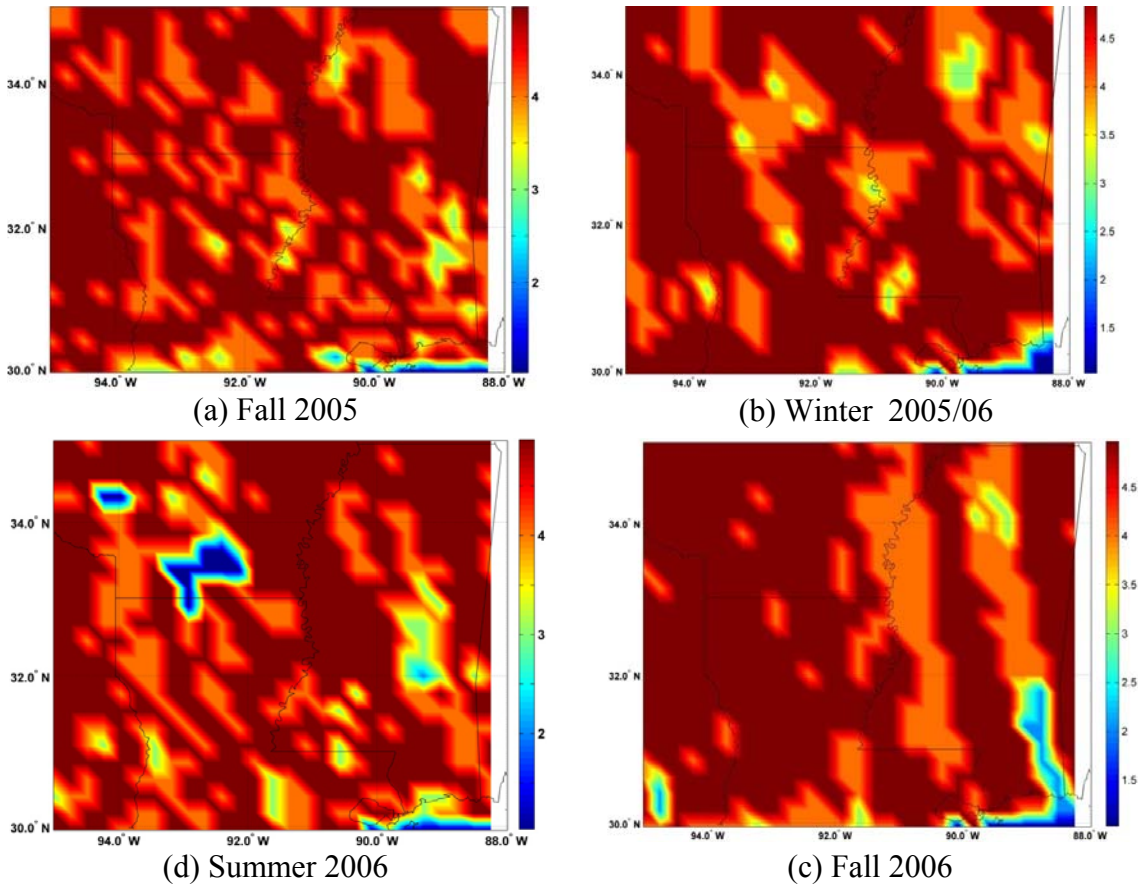


Figure 15. Consistency maps of AMSR-E soil moisture data for seasons

4.1.4 Discussion

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, as shown in Figures 14(c) and 14(d). There is a significant positive correlation of over 60% between the average soil moisture distribution and the consistency maps as shown in Table 4. This positive correlation suggests that the consistency measurement is generally related to the 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 [50].

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 [119]. 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 was 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 the observation frequencies as well; especially, attenuation due to vegetation can be higher at higher frequencies [120, 121]. Njoku and Entekhabi [122]

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 has been discussed and summarized in Wigneron et al. [123]. 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. The 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. Thus, the algorithm has been found to be robust in the study region.

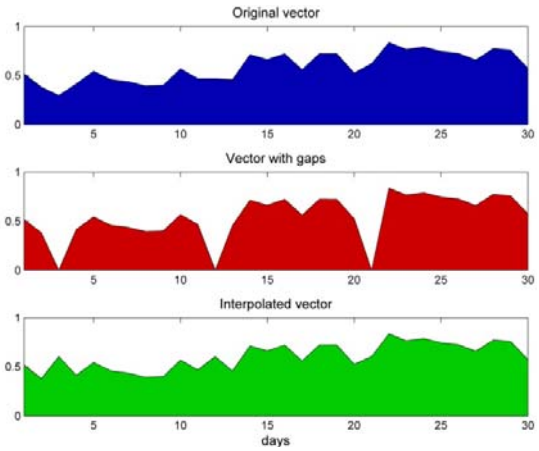
4.2 Implementation of the modified SSA interpolation

4.2.1 Validation with sample sets

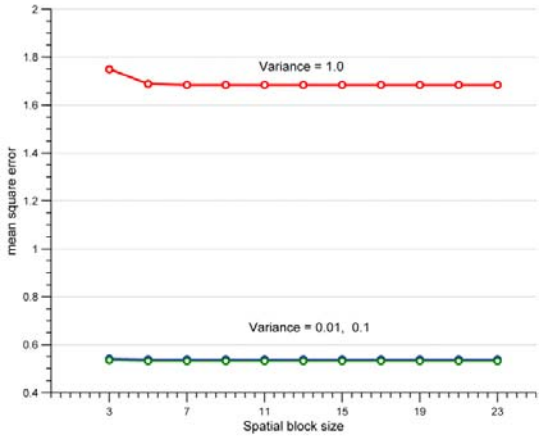
4.2.1.1 Synthetic spatio-temporal dataset

A synthetic spatio-temporal dataset, mvd, is generated by adding noise to two spatio-temporal sinusoidal signals, i.e.,

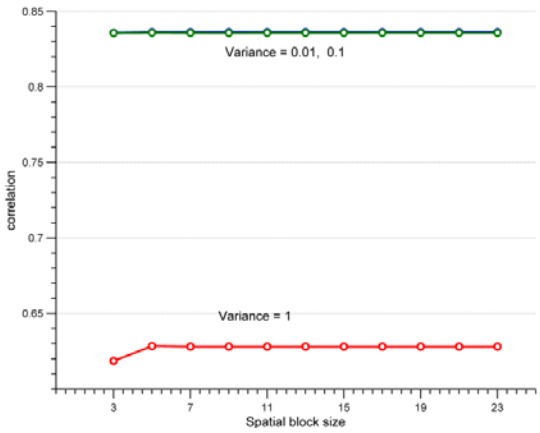
$mvd = \sin(x + 2y + t) + \cos(y - x - t) + noise(x, y, t)$, where x, y, and t are spatio-temporal coordinates. Ten percent synthetic gaps are introduced in this dataset and the performance of the modified SSA algorithm is studied for different noise variances 0.01, 0.1, and 1. As the dataset clearly consists of two dominant signals, two SSA modes were used in the reconstruction of the missing values. Figure 16(a) shows a vector from this dataset when the noise variance is 0.1. It can be seen that the missing values were estimated accurately. Figure 16 illustrates the performance of the interpolation scheme. Figure 16(a) shows a visual comparison of the original and interpolated data, while Figures 16(b) and (c) show the MSE and correlation of the interpolated values with the original values in the dataset. It can be observed that the performance of the algorithm deteriorates with an increase in the noise variance. Finally, it can be seen that a block size of 5 is sufficient for satisfactory performance.



(a) A vector before and after interpolation



(b) MSE vs. spatial block size for different noise variances



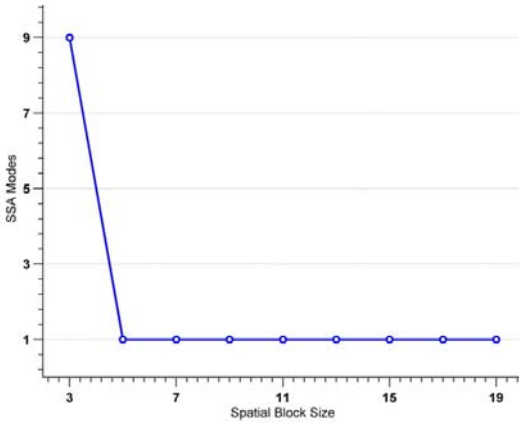
(c) Correlation vs. spatial block size for different noise variances

Figure 16. Performance of the interpolation algorithm on a synthetic dataset with two multivariate signals

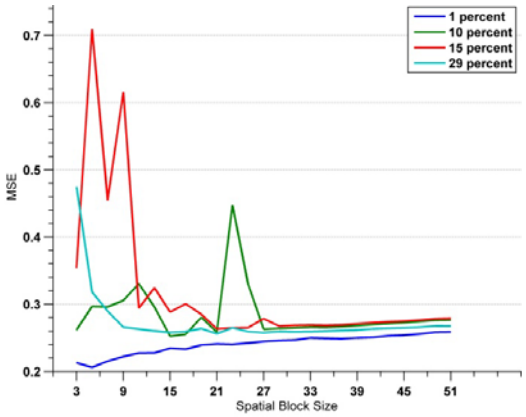
4.2.1.2 Sea surface temperature

The SST data used here is level-4 analysis from the Global Ocean Data Assimilation Experiment (GODAE) high resolution SST pilot project (GHRSST-PP). The level-4 data is a fusion of four sets of microwave observations retrieved from the following instruments: AMSR-E, U.S. geological survey's AVHHR onboard the TIROS-N satellite, Meteosat Second Generation (MSG) satellite, and Advanced Along Track

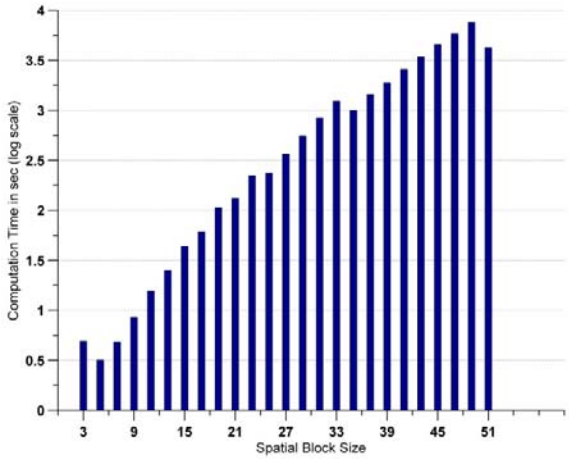
Scanning Radiometer (AATSR) onboard the European space agency's ENVISAT satellite. The final SST analysis is a bias corrected data and the correction is based on *in-situ* observations. This dataset is a daily SST with a 25km spatial resolution and global coverage [124, 125]. Three different cases are analyzed to assess the algorithm performance by synthetically introducing gaps in the dataset. The percentages of missing data in each case are 1 %, 10%, 15%, and 29%, respectively. As a first step, the number of SSA modes needed for the reconstruction of the missing values is determined as a function of the spatial block size. Figure 17 illustrates the effect of the spatial block size on the algorithm performance. From Figure 17(a), it can be observed that the first dominant SSA mode is sufficient for obtaining a minimum mean square error between the actual SST values and the reconstructed values. Then, using only the first SSA mode in the reconstruction process, the performance of the interpolation algorithm is evaluated for all the four cases as a function of the spatial block size. It can be seen from Figure 17(b) that a block size of 21 or less usually delivers a satisfactory performance. Next, the computational times taken for the interpolation algorithm for different block sizes are studied. It can be clearly seen that as the block size increases the computation time increases exponentially. For instance, for block sizes of 21 or less, the computation time is less than 100s while for the largest block size of 51, the time lapse is over 4200s. This is illustrated in Figure 17(c). A map of the SST for the study region with 1% gaps is shown in Figure 18(a) and a reconstructed map with gaps filled based on one SSA mode is shown in Figure 18(b). From this Figure, it is clearly seen the effectiveness of the presented interpolation method.



(a) No. of SSA modes used in the reconstruction with minimum MSE



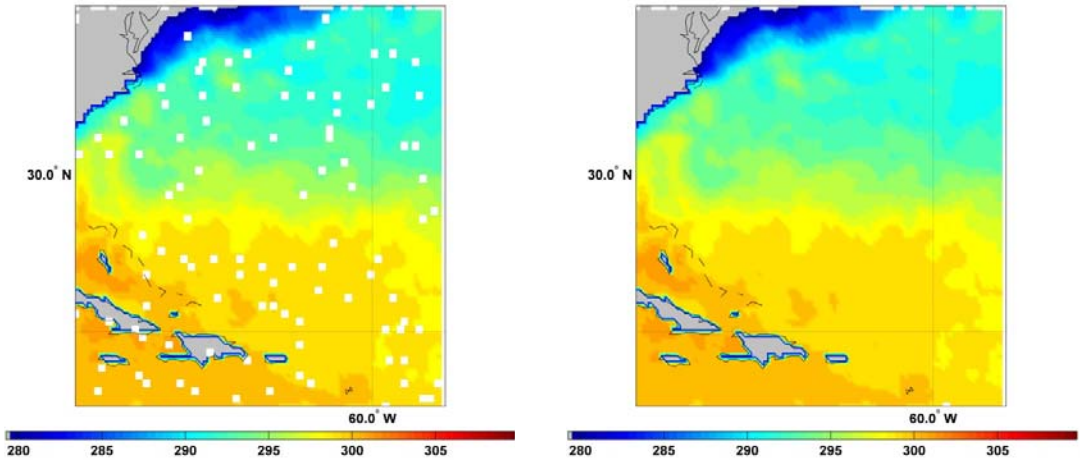
(b) Mean square error for different amounts of missing data



(c) Computation time on a logarithmic scale (base 10)

Figure 17. Algorithm performance vs. spatial block size

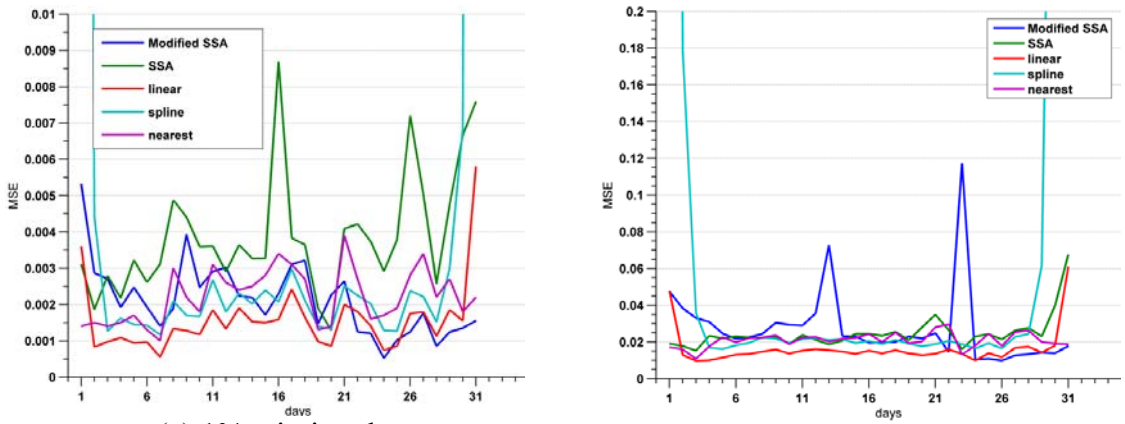
When only 1% of the data is missing, (Figure 19(a)), the MSE of the modified SSA method is always close to that of the SSA method and lesser than any other non-SSA methods after 21 days. As the percent of missing data is increased to 10%, the performance still compares well with other methods. Moreover, the simpler methods do not perform well at the temporal boundaries. However, the spectral methods do a much better interpolation. This is illustrated in Figures 19(b).



(a) Original image with random gaps (b) Image with gaps filled

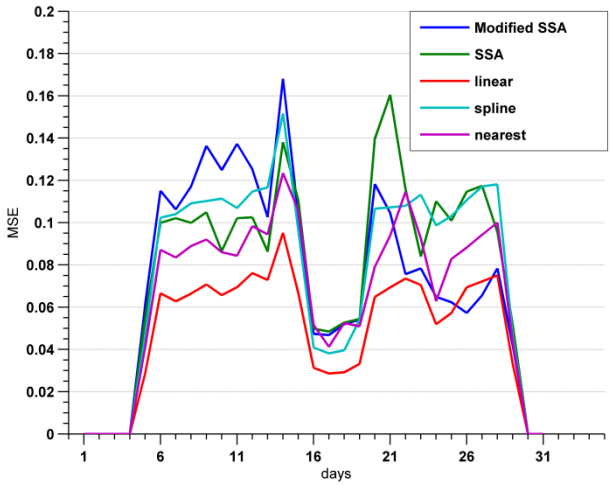
Figure 18. SST from GHRSST-PP for a 25° x 25° region centered at (27.5°N, 67.5°W)

For the first two experiments presented (1% and 10% gaps), the synthetic gaps introduced in the datasets are randomly distributed. However, in this case, a set of systematic gaps are introduced in the SST dataset; as a result 29% of the data is dropped. The resulting MSE comparison is illustrated in Figure 19(c). In this case, the performance of the algorithm is similar to that of the standard interpolation schemes. This result supports the idea that if there is a consistent covariance structure available from the data points, it is possible to interpolate the intermediate missing points. In summary, when the amount of missing data is increased, the modified SSA method has shown similar performance (Figures 19b and c). This observation shows the importance of using spatio-temporal signals for interpolation.



(a) 1% missing data

(b) 10% missing data



(c) 29% missing data

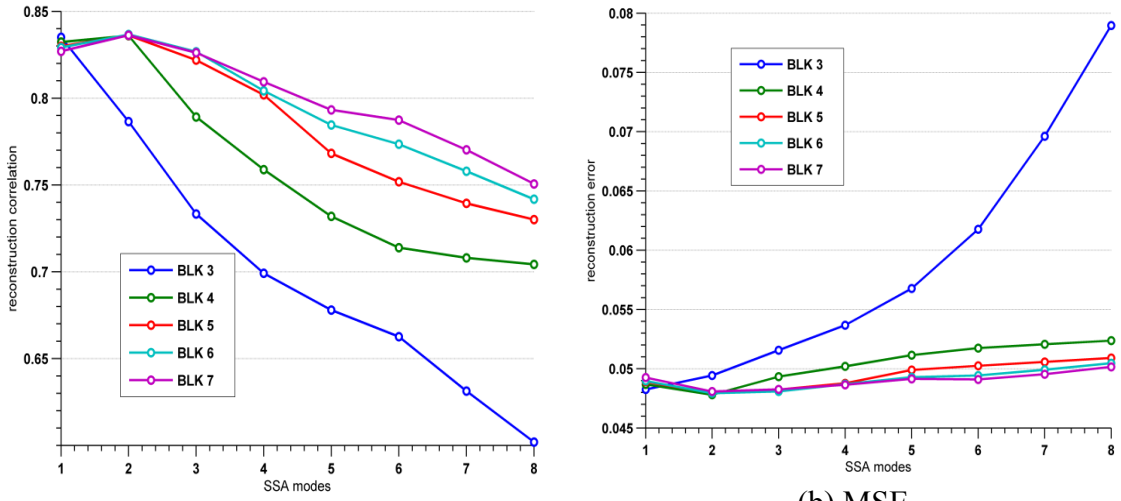
Figure 19. MSE Comparison between the actual SST versus interpolated SST, on a daily basis, computed from different interpolation algorithms

4.2.1.3 Normalized difference vegetation index

The NDVI data used here is "monthly level 3 global vegetation indices" from the Moderate-resolution Imaging Spectroradiometer (MODIS) onboard the Terra satellite. This instrument facilitates the retrieval of a suite of land surface related parameters including but not limited to vegetation indices, surface temperature, reflectance, and albedo. The NDVI data is derived from blue, red, and near-infrared reflectance, centered

at 470, 648, and 848-nanometers, respectively. This dataset is a monthly product with a spatial resolution of 0.05°. The exact version is a global version 5 data, at validation stage 2 (MOD13C2) that is obtained from composites of 16 day and 1km MOD13A2 on a 5.6km climate modeling grid (CMG). This product was validated for a wide range of spatio-temporal locations [126, 127].

In our experiment, a 5° x 5° subset centered at $(40^{\circ} N, 107^{\circ} \text{ W})$ is taken from Jan 2003 to Dec 2004 (24 time steps). Normally distributed synthetic gaps are introduced in the dataset. The amount of gaps introduced amounts to 10% of the total dataset. The performance of our method is tested for different block sizes and number of modes. This is illustrated in Figure 20.



(a) Correlation (b) MSE

Figure 20. Performance of the modified SSA algorithm, on MODIS NDVI dataset, based on different spatial block sizes

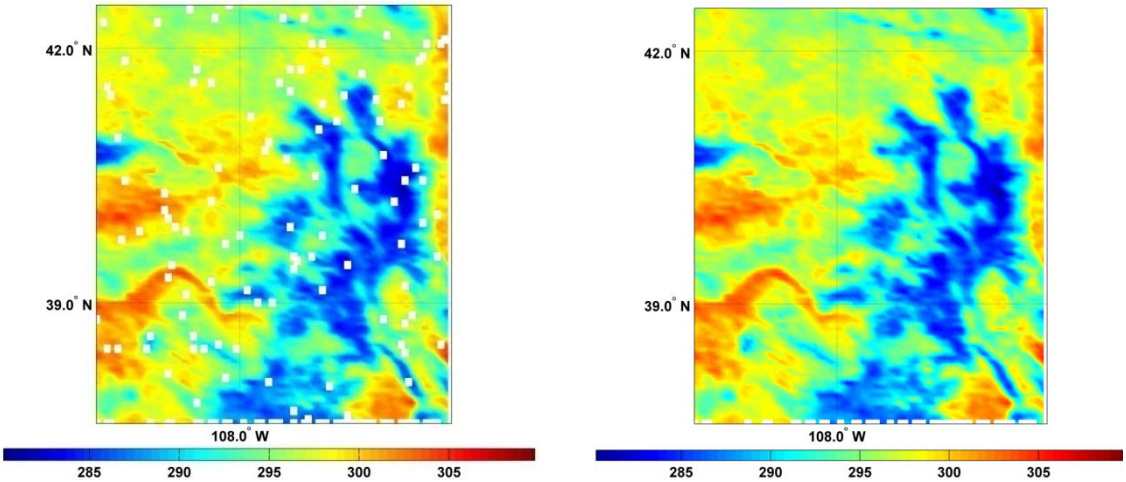
Note that in the MSE and correlation plots, the symbols SSA N represent the simulations with the SSA algorithm with a block size of NxN. From Figure 20(a), it can

be observed that the correlation increases as the block size increases, especially for higher modes. However, the maximum correlation can be seen when two dominant modes are selected. A similar observation can be made for the MSE plots; except, the MSE values decrease for larger blocks (Figure 20(b)). These observations suggest that the optimal performance can be achieved when the top two dominant modes are used in the reconstruction process.

4.2.1.4 Land surface temperature

Land surface temperature (LST) used here is the level 3 and version 5 dataset from MODIS data product suite (MOD11C3). The LST retrievals are based on the application of Wan and Li's [128] LST algorithm on a pair of day and night observations from MODIS. The version 5 data is a composite version of daily LST which is a re-gridded version the level 2 data onto a 5km sinusoidal grid and validated up to stage 1. The dataset represents monthly averages with a spatial resolution of 0.05^{o} [129, 130].

For the LST subset, a spatio-temporal region is chosen similar to the one selected for the NDVI data case. A similar interpolation experiment is then conducted and the corresponding results are illustrated in Figure 21. Figure 21(a) shows the LST image for this region with synthetic gaps and Figure 21(b) shows the same region with gaps filled. It can be inferred that a small percentage of normally distributed gaps can be efficiently filled.



(a) Original image with random gaps (b) Image with gaps filled

Figure 21. MODIS LST for a 5^o x 5^o region centered at (40^oN, 109^oW)

4.2.2 Interpolation of incomplete AMSR-E soil moisture data

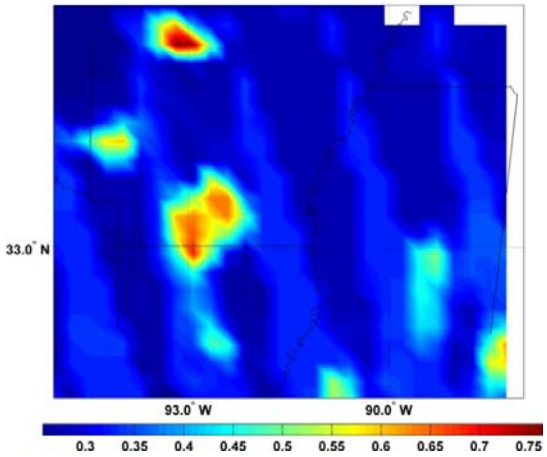
In this experiment, the soil moisture dataset used is from the NASA'S AMSR-E Level 3 "AE_Land3" product, distributed by the National Snow and Ice Data Center (NSIDC). The dataset has a 25km spatial resolution [115, 119, 122, 131-133]. Soil moisture data fields are extracted and data sets are generated for a region consisting of the states of Mississippi, Arkansas, and a part of Louisiana [39, 112, and 114]. The dataset is collected over a period of two years from Jan 2005 to Dec 2006. The dataset studied in this experiment is same as the region considered in the consistency analysis task. The level 3 soil moisture product has many intermittent gaps due to various reasons including varying revisit times, interference, and dense vegetation.

The percentage of missing data points in AMSR-E soil moisture retrievals is usually under 35 with some exceptions. For fall 2005, a spatial distribution of the number of missing data points as a fraction of the total number of data points is shown in Figure

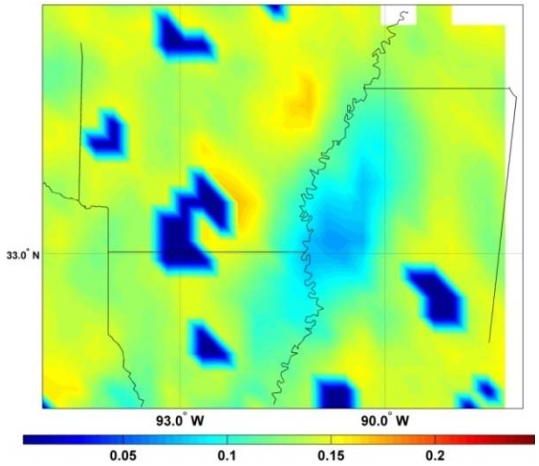
22(a). First, significant features of this map are reddish yellow patches which signify a large amount of missing data, generally more than 70%. This type of structure reveals a systematic error with this season's retrievals. Second, the light bluish wavy pattern, extending throughout the map, corresponds to missing data because of the lack of coverage by the satellite due to its orbit. The amount of missing data points in the second case is around 30%. An image of soil moisture for October 16, 2005 is shown in Figure 22(b). The blue patches representing missing data agrees well with the trend shown in Figure 22(a). The algorithm with these parameters was applied to seven seasons of AMSR-E soil moisture for the regions shown in Figure 22(a). The method is tested on data from spring 2005 to fall 2006. For comparison purposes, the SSA algorithm is also applied to the same seasonal data sets. The soil moisture image for October 16, 2005 with data filled using the modified SSA method is shown in Figure 22(c) and the result from the original SSA method is shown in Figure 22(d). It is evident that both methods yield similar spatial structures. For instance, the missing region shown by the larger blue patch in Figure 22(b) is surrounded by higher soil moisture values. The same regions in the interpolated images are filled with slightly higher values, thus, in agreement with the spatial continuity with its surrounding pixels. Moreover, the spatial continuity in the image from the presented method is better compared to the image from the SSA method. This difference is illustrated by the regions marked A and B in each image respectively. Note that, in region B, there is a sudden change in the spatial structure whereas in region A the structure has a much smoother transition. Based on the method described in [91]

the optimal time lag of 20 and 10 SSA modes are used for SSA interpolation method on soil moisture dataset.

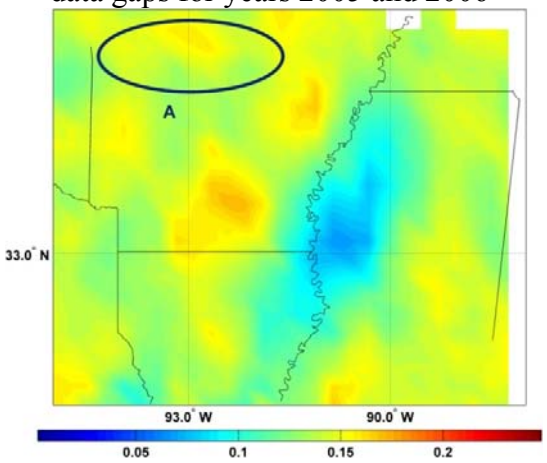
It is also of interest to compare the interpolated data based on the modified SSA and SSA on the AMSR-E data. The resulting performance comparison is illustrated in Figure 23. A mean square difference between the modified SSA and SSA generated values for fall 2005 is presented in Figure 23(a). It can be observed that these methods agree well with each other. An illustration of correlation between these two methods is presented in Figure 23(b). As shown in Figure 23(c), the general trend in missing points of this time series is that they occur in pairs with at least one day separation. The plots in blue and red are the time series obtained from modified and the original SSA methods. It is evident that the interpolation process is successful in retaining the temporal structure of the original data as there are no sudden jumps in the time series. The overall comparison between the modified SSA and SSA algorithms for the whole study period is shown in Figure 23(d). From this plot, it can be observed that the two methods are in good agreement for most of the seasons, with $mse = 1 \times 10^{-4}$ in summer 2005 and $mse = 4 \times 10^{-4}$ in spring 2005.



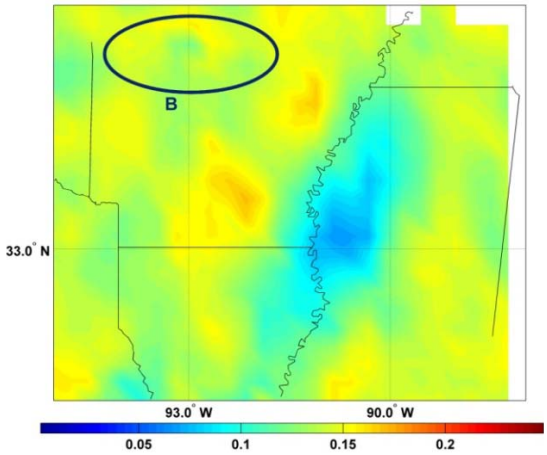
(a) Study region showing the fraction of data gaps for years 2005 and 2006



(b) Soil moisture map on October 16, 2005

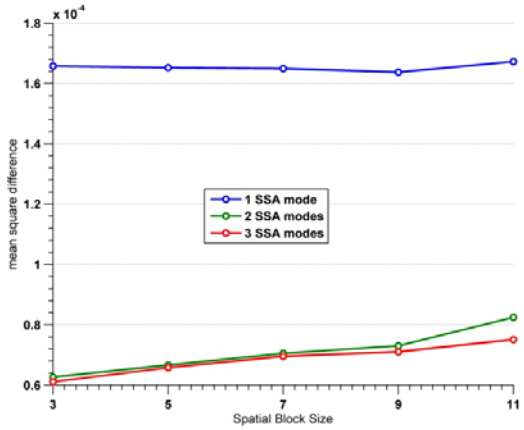


(c) Soil moisture map on October 16, 2005 with gaps filled from modified SSA algorithm

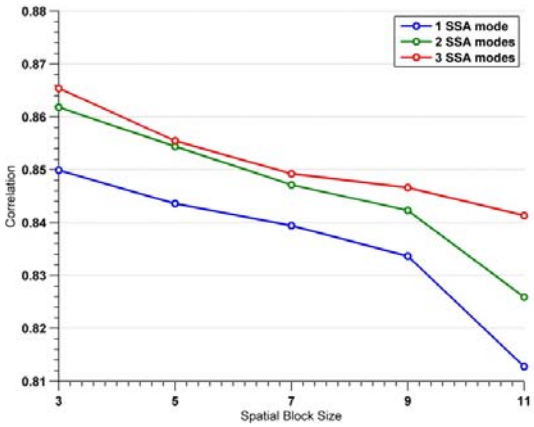


(d) Soil moisture map on October 16, 2005 with gaps filled from SSA algorithm

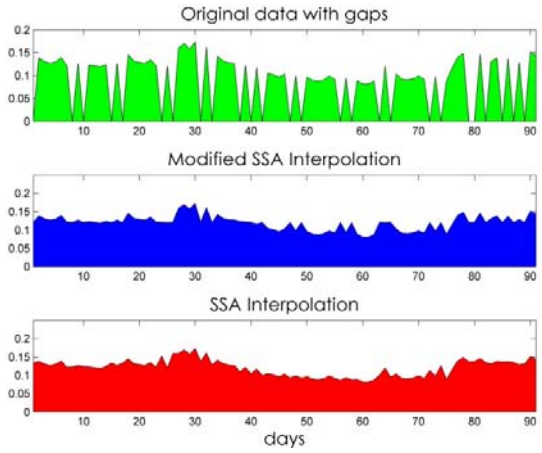
Figure 22. AMSR-E soil moisture maps before and after interpolation comparisons



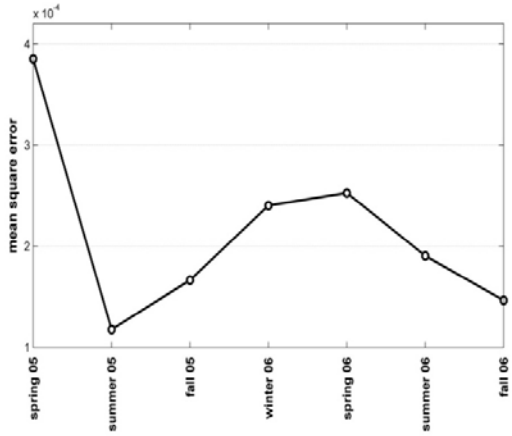
(a) The mean square difference vs. block size



(b) Correlation vs. block size



c) Time series comparison from at a grid cell



(d) Mean square difference for different seasons

Figure 23. Performance comparison of interpolated data on AMSR-E data: modified SSA vs. SSA

4.2.3 Discussion

The following parameters are of significant importance in developing the modified SSA algorithm: (i) Dimensions of the spatial data block in the modified SSA algorithm: Depending on the type of geophysical variable interpolated, the optimal spatial block size varied as follows: 5 to 21 for SST, 5 for NDVI and LST, and 3 for soil moisture. This explains the importance of local covariance in interpolating the intermittent gaps. (ii) SSA modes: The number of modes in the modified SSA algorithm depends on the dimensions of the data-block. Optimal SSA modes used in reconstruction were 1 for SST, 2 for NDVI and LST and 3 for soil moisture.

(iii) Iterations: The number of iterations in the reconstruction process used is based on the convergence error. The process is stopped when an error of less than $1x10^{-12}$ is reached. The average number of iterations for the modified SSA algorithm was 20 with a standard deviation of 16.

4.3 Fusion of HRPPs case study

4.3.1 Fusion process for rainfall data

Our case study on the set of HRPPs can be summarized as follows. Our merging method is implemented on rainfall estimates from a collection of four different satellite precipitation products. These precipitation sets are extracted from various HRPPs for a region around the Arkansas Red Basin River Forecast Center's (ABRFC) study area. The common temporal and spatial resolutions of all the products are one hour and *10km by 10km*, respectively. The rainfall data from each dataset at a given location is arranged into

a vector X, where $\mathbf{X} = [x_1 \; x_2 \; \ldots x_M]^T$ with M being the number of HRPPs considered at any given location. For instance, in this study, M = 4 at any given time and location. Thus, the entire region would have a time series representative of these vectors for four datasets. This vector data is transformed into another vector space using a transformation function as defined by Eq. (25), where α, β, γ, and κ are transformation parameters to be optimized. The result of the neural network classification is basically a binary array, where '1' corresponds to a *rain* pixel and '0' corresponds to a *no-rain* pixel. A precipitation estimate is generated by multiplying this binary array with the average dataset of all the individual rainfall products. A detailed description of the case study follows.

4.3.1.1 Input data description

Precipitation datasets are collected from the following: (1) Climate Prediction Center morphing method (CMORPH); (2) Auto Estimator algorithm for Geostationary Operational Environmental Satellite data (GOES AE); (3) Hydro Estimator algorithm for Geostationary Operational Environmental Satellite data (GOES HE); (4) Naval Research Laboratory blended satellite HRPP (NRL-BLEND); and (5) A Self Calibrating Real-Time GOES Rainfall Algorithm (SCAMPR). The CMORPH product is a blend of rainfall data from a passive microwave sensor and rainfall product from an infra-red based sensor [73]. For the GOES AE, an algorithm, called Auto Estimator, produces rainfall data in real-time with the purpose to generate quality product for hydrology research applications. It uses IR data from sensors onboard the GOES satellite. The final data is constrained by a mask based on cloud growth rate measured in terms of temperature

change and spatial gradient of cloud top temperature. The AE algorithm was improved by screening cold clouds by a mask developed using Doppler reflectivity data from a weather satellite radar-1988 (WSR-88D) [134]. The HE is developed by applying the following major modifications to the AE algorithm: (i) the definition of a raining pixel is modified to those pixels with a brightness temperature (at $10.7\mu m$ band) less than the average temperature of a predetermined region. This modification helps in reducing the rain areas overestimated by the AE; (ii) the rain rate curve is also modified using the same principle; and (iii) the influence of the multiplicative moisture adjustment, which is a product of precipitable water (PW) and relative humidity (RH), is changed by separating the components [135]. The NRL-BLEND product is basically a HRPP developed from blending of IR data from GEO satellites and PMW data from LEO satellites through a sophisticated, real time spatio-temporal collocation scheme. In this blending scheme, several cases of datasets were developed through systematic selection and omission of satellites from the blending scheme. In this work, we used the blended product in which all the satellites were considered [136]. Since the temporal resolution of the NRL-BLEND data is three hours, it was disaggregated to one hour data to match other datasets in the fusion process. Finally, SCAMPR is a screening technique that is used to separate *rain* and *no-rain* pixels and then, a linear regression-based rain rate predictor is developed and calibrated against a microwave sensor-based rain rate [137]. The first four datasets were used in the seasons of summer 2007, fall 2007, and spring 2008. However, in winter 2007/08, the NRL-BLEND is replaced with SCAMPR due to lack of availability.

4.3.1.2 Reference data

The Arkansas Basin River Forecast Center of the U.S. National Weather Service develops the reference rainfall data used in this study. This region is composed of the state of Oklahoma and small portions of all its neighboring states. This reference data (hereafter referred to as ABRFC) is a multi-sensor precipitation estimate from the combination of hourly radar estimates and hourly rain gage measurements. The incoming hourly data from rain gages is mapped on an irregular triangulated grid on which the radar mosaic is overlaid. At the overlapping points, the average is considered and, at missing locations, the value is estimated from the neighbors. The ABRFC data is quality controlled on an hourly basis for the following errors: (1) radar-based errors, such as hail contamination and beam blockage, (2) gage-based errors, such as sampling errors and mechanical problems with gages, and (3) software errors. As a last step, the data is quality controlled against a manually observed daily rainfall totals [138, 139]. Finally, because of the incorporation of the rain gage data, this product can be considered as the closest estimate to the ground truth and thus adopted as reference data for this analysis.

4.3.1.3 Training

From each dataset, a uniformly gridded dataset with a cell size of 0.1° by 0.1° was developed. The grid is composed of 80 by 200 cells, surrounding the ABRFC region, and is collected over the summer 2007 - spring 2008 period. The rainfall values from each method are arranged such that there is a four-element vector at every grid cell. The training data is selected as follows: The precipitation values in the ABRFC grid are classified as *rain* or *no-rain* based on a threshold. From the rearranged data vectors, the

energy $E = \sum_{i=1}^{M} x_i^2$ of each vector is computed. If this energy E is large and the corresponding ABRFC value is larger than the threshold, then, this vector is selected as a training vector for class *rain*. Similarly, if the energy is close to zero and the corresponding ABRFC is smaller than the threshold value, then, the vector is selected as a training vector for class *no-rain.* If the number of vectors *Ntrain* is selected in this screening, a training dataset of size (*Ntrain* by M) is developed with a corresponding class vector (*rain or no-rain*). Using a trial and error method, several non-linear functions, such as exponential, logarithmic, and trigonometric, are tested for feature transformation. A scaled exponential function, with the corrected observation as the scaling function, as defined in Eq. 25, is selected as a suitable transformation function. The selection criterion is based on higher success rates. Moreover, higher success rates suggest that feature transformation improves the suitability of the observation for classification by a neural network classifier. This function is used to transform the training dataset into a new vector space, which is used along with the class vector to train a 2-layer neural network. The criteria for stopping the training process before reaching a zero classification error are when the error gradient reaches a very small value and the error evolution just crosses a steep gradient. Figure 24 illustrates the convergence of the neural network based on the foregoing training. This early stopping helps in improving the network classification's performance and making sure the network is not over-trained.

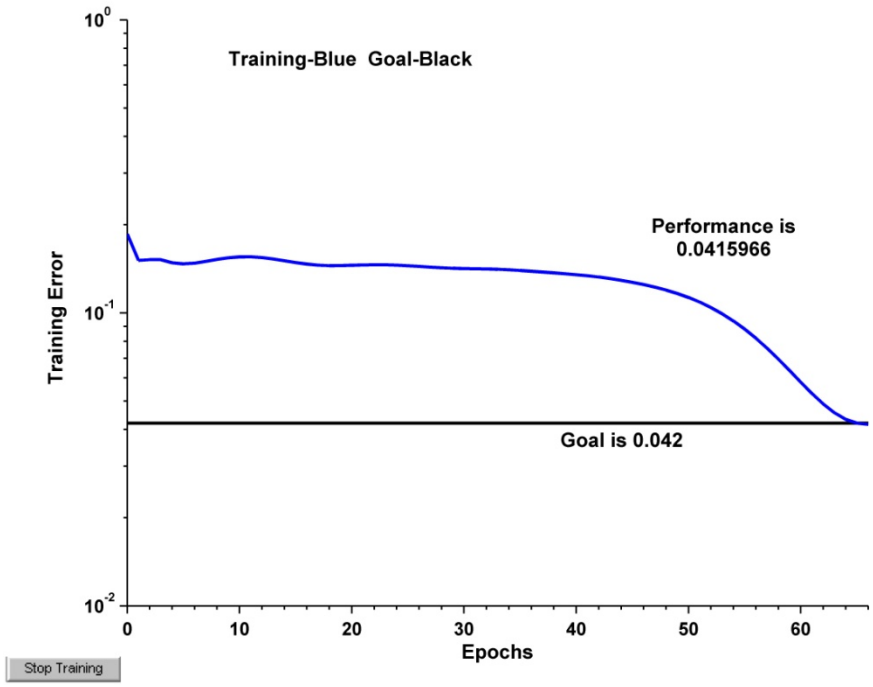


Figure 24. Convergence of neural network training

4.3.1.4 Evaluation

The screening of satellite images into *rain* or *no-rain* pixels is an integral part of many precipitation methods. In particular, screening out the *no-rain* pixels is important for the performance of many rainfall products. In other words, one of the goals of a screening method is to reduce the false alarm rate. Traditionally, screening methods are divided into four types: (1) quality control, where the implausible brightness values are screened out, (2) special data corrections based on spatial correlations, (3) coastline corrections, and (4) geographic corrections based on the terrain [140]. Thus, the goal of the traditional screening techniques is to reduce false alarms. However, in this study, screening is performed not only to reduce false alarms, but also to improve successful detections. Threshold-based scores, such as bias score, correlation, critical success score,

and Heidke skill score (HSS), are computed. Usually, metrics, such as successful detection and false alarm rate, can be misleading if used individually as it is possible to improve one metric by ignoring others. However, HSS is a better measure of agreement between the estimation and the ground truth. HSS is very similar to the Kappa coefficient used in remote sensing classification studies and the critical skill score used in meteorological studies. In general, HSS is defined in terms of contingency table elements as $HSS = 2(ad - bc) / [(a + c)(c + d) + (a + b)(b + d)]$, where a is the number of successful rain detections, b is the number of false alarms, c is the number of misses, and d is the number of successful *no-rain* detections. Thus, HSS measures the skill of the predictor against the agreements by random chance [141]. For a given grid cell, if the merged time series is better than the input time series in terms of HSS, it is counted as a success. The success rate SR is defined as the percentage of the number of grid cells in the study region in which the resulting HSS is better than the HSS of every individual HRPP. HSS is computed for cells where the reference data is available. For instance, if there are 1000 cells in the study region's grid and the HSS is improved in 900 of them, then, the success rate is 90%. Using the HSS-based success rate as the performance metric, a cross-validation of the fusion method is performed as follows. Precipitation data is collected for four seasons: (a) summer 07, (b) fall 07, (c) winter 07/08, and (d) spring 08. The parameters in $pvec$ are optimized for the summer season and tested on the fall, winter, and spring seasons, with a success rate as the performance criterion. This process is repeated in hold-out cross-validation setting by optimizing $pvec$ on the fall, winter, and spring seasons separately and testing on the remaining three seasons. The results

from cross-validation experiments are discussed in section 4.3.2.1. In order to better understand the performance of this fusion methodology, HSS scores are computed for the merged data over the whole study region. A discussion of the HSS score obtained for the merged data is presented in section 4.3.2.2. Comparisons with HSS scores of individual HRPPs are discussed in section 4.3.3

4.3.2 Results

4.3.2.1 Cross-validation results

The success rates obtained from hold-out cross-validation performed on the rainfall data are presented in Table 5. It can be seen that the average success rates for individual seasons are 85 for summer 2007, 68 for fall 2007, 55 for winter 2007/08, and 77 for spring 2008, respectively.

Table 5. Success rates from the cross-validation experiments

		Test Seasons			
		Summer	Fall	Winter	Spring
Training Seasons	Summer	88.21	67.00	56.31	81.14
	Fall	86.31	70.10	46.45	63.95
	Winter	84.69	66.45	58.12	82.04
	Spring	80.99	68.76	58.30	82.10
	Average	85.05	68.08	54.80	77.31

In order to emphasize the importance of the vector transformation in this fusion process, cross-validation experiments are repeated on the data sets without using vector space transformation. The respective average success rates are shown in Figure 25. For

comparison purposes, the improvement due to feature transformation is also presented in Figure 25. It can be observed that the fuser with and without vector transformation performed well during the warmer seasons. The improvement is maximum in the summer at a rate of 37% and minimum in the spring at a rate of 10%. Note that the blue portions of the bar plots in Figure 25 indicate success rate with the neural network only fuser and the red portions indicate the improvement in the success rate due to vector space transformation.

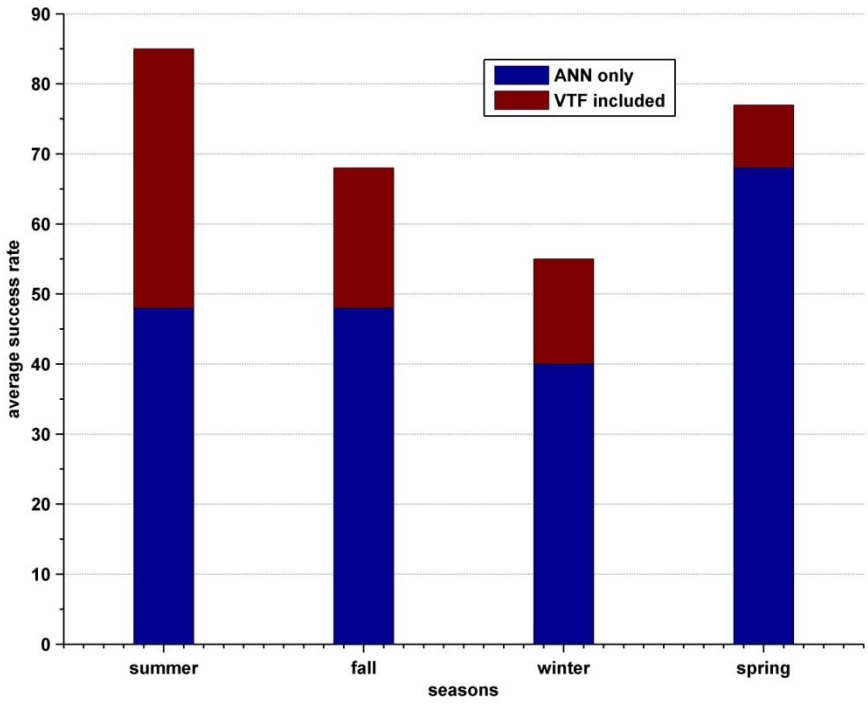
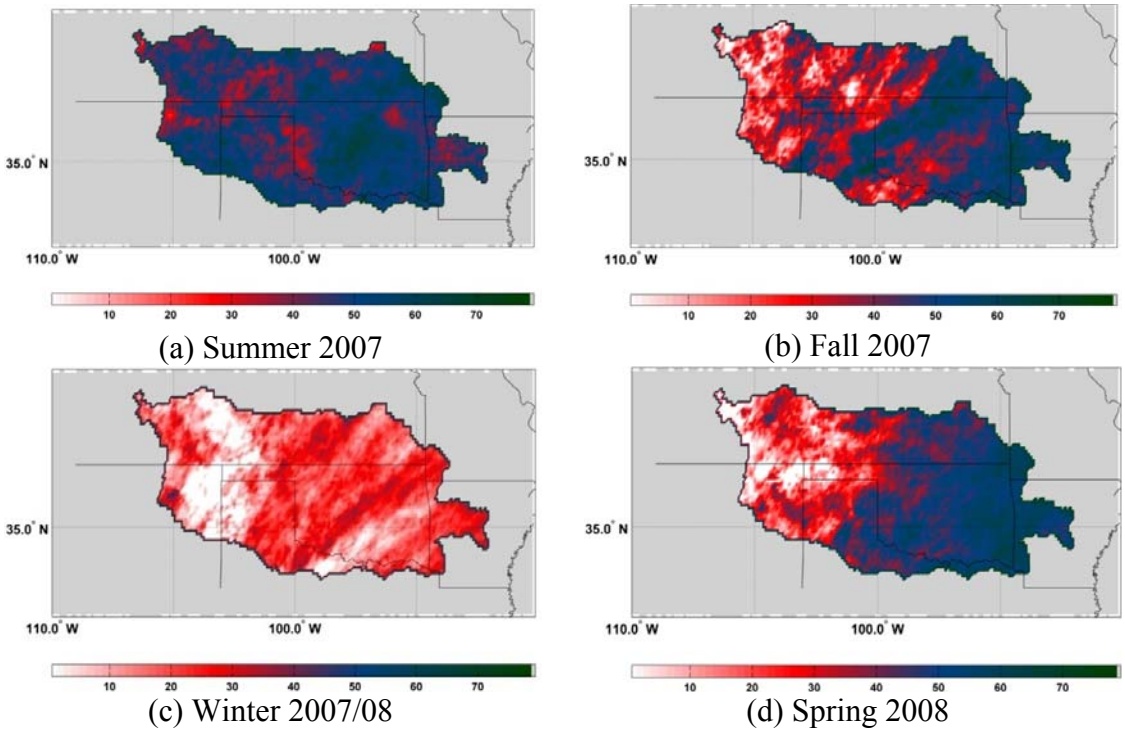


Figure 25. Improvement in success rate due to vector space transformation

4.3.2.2 Performance of the merged data against the ABRFC reference data

Figure 26 illustrates the Heidke skill score maps of the merged data compared with the ABRFC data for four seasons. During the summer season, the merged data

agrees with the reference data in most of the ABRFC region (Figure 26a). This can be attributed to the high amount of the rainfall occurring during the summer. Thus, the satellite-based rainfall data is closer to the ground-based measurements.



(a) Summer 2007 (b) Fall 2007

(c) Winter 2007/08 (d) Spring 2008

Figure 26. Heidke skill score maps and skill score distributions of the merged data compared with the ABRFC data for four seasons

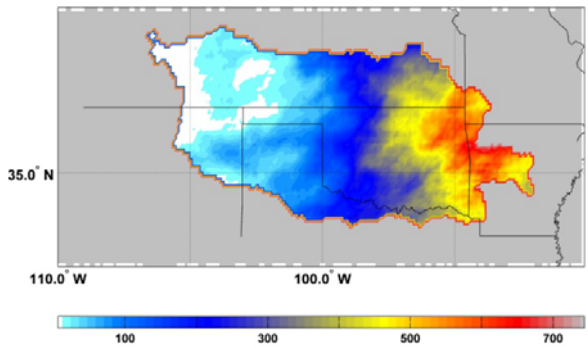
During the fall season, the merged data is close to the reference data, but the performance is not as good as in the summer. The agreement is higher in the eastern section of the ABRFC region (Figure 26b). Similar performance is observed for spring 2008. The amount of the rainfall during the fall and spring seasons is, in general, lower compared to the summer rainfall. Thus, there is a lower degree of agreement between the merged satellite data and the ground-based data (Figure 26d). During the winter, the performance is the lowest. The amount of the rainfall is also the lowest in the winter,

thus, leading to greater disagreement between the ground data and the satellite-based data (Figure 26c).

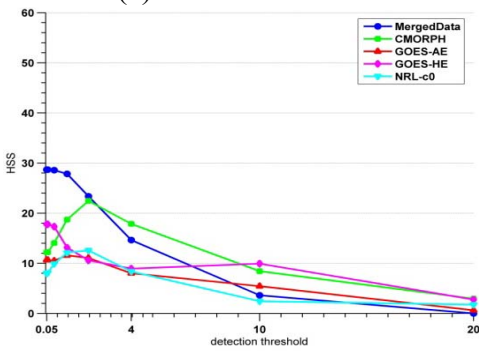
There are two distinct areas of precipitation regimes in our study area covering the ABRFC domain. The western half of the domain generally has less precipitation than the eastern half, with the normal precipitation gradually increasing eastward (Figure 27a). Thus, it is helpful to investigate the skill of the HRPP products separately at the eastern and western portions of the domain. Using the HSS as the verification metric, the skill of the individual HRPP data products (including the data merged using our method) can be determined at different levels of detection thresholds. Here, for the analysis of the hourly precipitation data, we have adopted threshold levels of 0.05, 0.1, 0.4, 1, 2, 4, 10, and 20 in units of mm. The seasonal accumulation of the merged data is highly correlated (0.87) to the skill. This level of agreement, also seen in other seasons, supports the idea that the skill of the merged data may be related to the amount of precipitation. From Figures 27b and 24c, it may seem initially that the apparent performance of the merging method depends on the detection threshold (amount of rainfall).

Furthermore, a comparison of the line plots shows that, for any given threshold, every HRPP has a better skill score in the eastern section. A similar partition in the HSS distribution is seen for each HRPP for all the four seasons. With careful inspection, it can be seen that the contributing HRPPs are simply transferring their individual skills to the merged dataset. The individual HRPPs perform well for thresholds less than 4. Then, the skill gradually decreases for thresholds greater than 4. This pattern transfers from the individual HRPPs to their merged product. An observation of the probability of detection

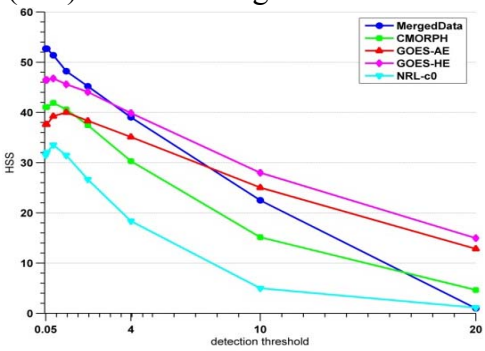
(POD) and false alarm rate (FAR) plots for the two areas also indicate a very similar partition of the skills (Figures 27d – 27g).



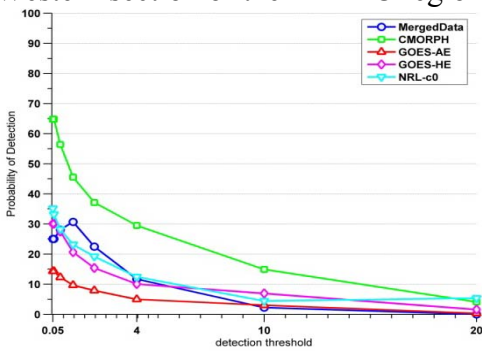
(a) Seasonal rainfall accumulations (mm) from the merged dataset



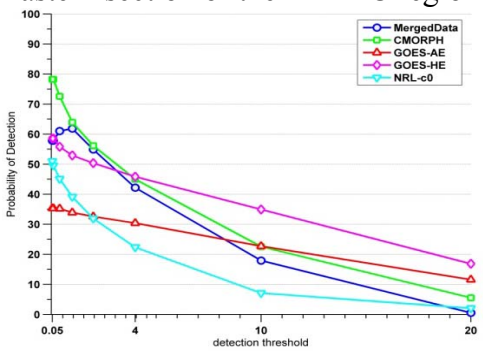
(b) HSS vs. detection threshold plot for hourly rainfall accumulations for the Western section of the ABRFC region



(c) HSS vs. detection threshold plot for hourly rainfall accumulations for the Eastern section of the ABRFC region

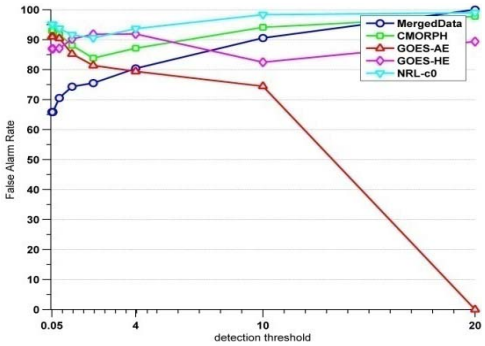


(d) Probability of detection vs. detection threshold plot for hourly rainfall accumulations for the Western section of the ABRFC region.

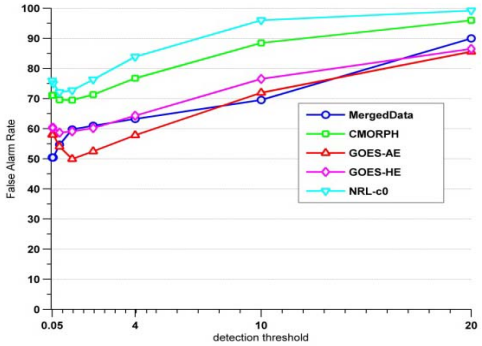


(e) Probability of detection vs. detection threshold plot for hourly rainfall accumulations for the Eastern section of the ABRFC region

Figure 27. Algorithm performance for different sections of the study region (Spring 2008)



(f) False alarm rate vs. detection threshold plot for hourly rainfall accumulations for the western section of the ABRFC region



(g) False alarm rate vs. detection threshold plot for hourly rainfall accumulations for the Eastern section of the ABRFC region.

Figure 27 (Continued)

Thus, for all four seasons, the skill of the merged data in the eastern section of the ABRFC region is, generally, better than the skill in the western section because the individual HRPPs have better skill in the higher rainfall regime in the eastern sector. Current observations of remotely-sensed precipitation, such as from passive microwave (PMW)-based satellite sensors, are far less sensitive to light rainfall and it is likely that they underestimate light events or miss them completely; a situation which is made worse by the lack of adequate validation data [142]. The problem is especially acute at middle latitudes and higher, where frontal-based precipitation, winter/cold season, and the variable land background surface are all factors that inhibit the performance of current PMW-based precipitation algorithms. Since these same PMW precipitation datasets are the same datasets that get "fused" into the multi-satellite blended precipitation products, the individual HRPPs themselves do not represent the light end of the precipitation distribution well. Hence, in our domain, the satellite precipitation products as well as our merged data have poorer skill at lower thresholds in the western half of the study area

with lower rainfall rates. The skills tend to improve for thresholds up to 2 mm/hr and then continue to drop off for higher thresholds where the quality of the HRPPs degrade due to sampling errors (spatial and temporal) and the delineation and estimation of the areal extent of rain.

It can also be noted that, in our analysis, the NRL HRPP data has the lowest skill of all HRPPs. Originally; this dataset consists of 3-hour accumulations, which were then disaggregated to 1-hour estimates, which accounts for its poor skill. The NRL-Blend product has much better skill when longer accumulation periods are considered. This also illustrates the limitations of interpolating or disaggregating rainfall to finer resolutions. The percentage of the study region in each HSS skill quartiles is shown in Table 6. For instance, in the summer season, 31% of the total study region has a HSSS greater than 3/4 of the maximum HSS for the summer. In all seasons, a large fraction of the study region belongs to quartiles 2 and 3.

Table 6. HSS skill quartile percentages of the area in the study region

skill quartile	summer	fall	Winter	spring
1	31.6267	9.3867	0.26	30.3
2	65.0311	49.7956	17.1	36.8
3	3.342	30.0444	49.55	17.76
4	0	8.6578	22.2	9.97
No skill	0	2.1156	10.9	5.15

4.3.3 HSS difference maps

Maps representing the difference in Heidke skill scores between the merged data and seasonal CMORPH, GOES AE, GOES HE, and SCAMPR data are generated for comparison purposes to illustrate the applicability of the foregoing fusion technique. Note

that skill scores for all cases considered in this study are computed against the ABRFC data for four seasons.

4.3.3.1 Comparison with CMORPH

Figure 28 represents maps of the difference in the Heidke skill score between the merged data and the seasonal CMORPH data. During the summer season, the difference in HSS percentage varied from -10% to 30% and the mean difference is 4% for the whole rectangular region. From Figure 28a, it is evident that there is an improvement of at least 10% in most of the region.

In addition, there are few cells in which the merged data is not as good as the CMORPH data. During the fall season, the difference varied from -25% to 30%. However, the difference is, in general, positive and above 15% as indicated by the green region in Figure 28b. The negative regions are comparatively very small and are located near the western border of the ABRFC region. During the winter season, the difference varied from -15% to 30% with most of the region being above 15% (Figure 28c). The negative regions are in the eastern section. In the spring season, the merged data outperforms the CMORPH data in most of the cells (Figure 28d). The similarity between the spring and fall maps indicates that the merging method works similarly in these seasons.

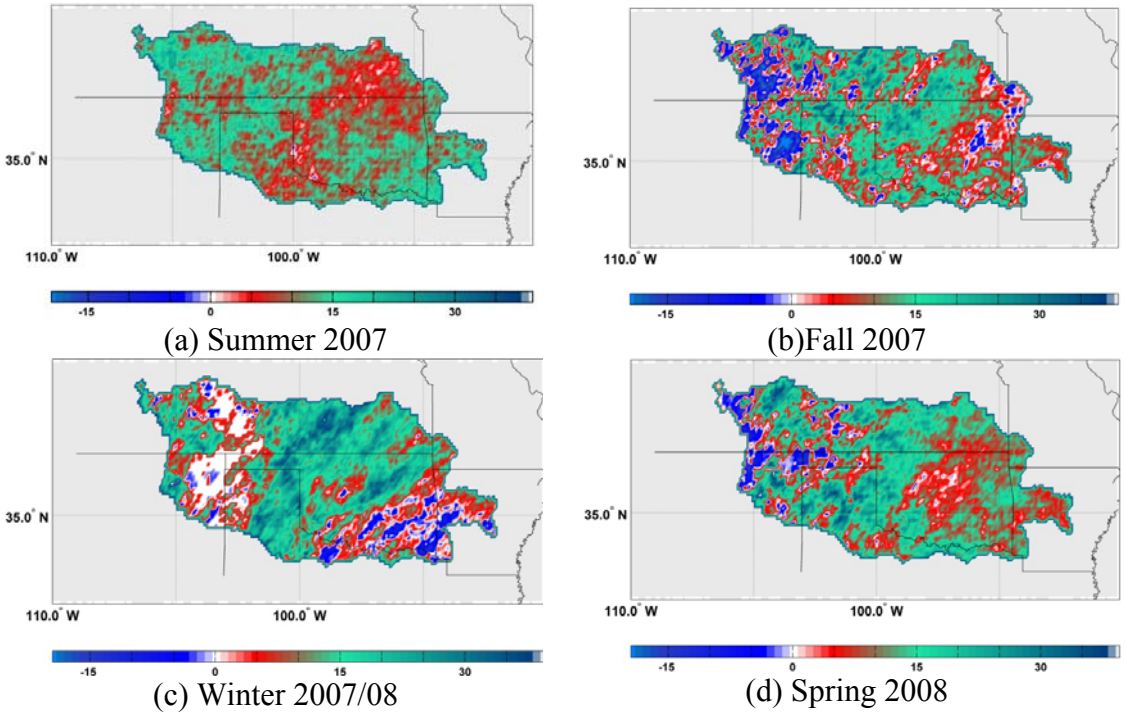


Figure 28. Maps and distributions of the difference in Heidke skill score between the merged data and the seasonal CMORPH data for four seasons

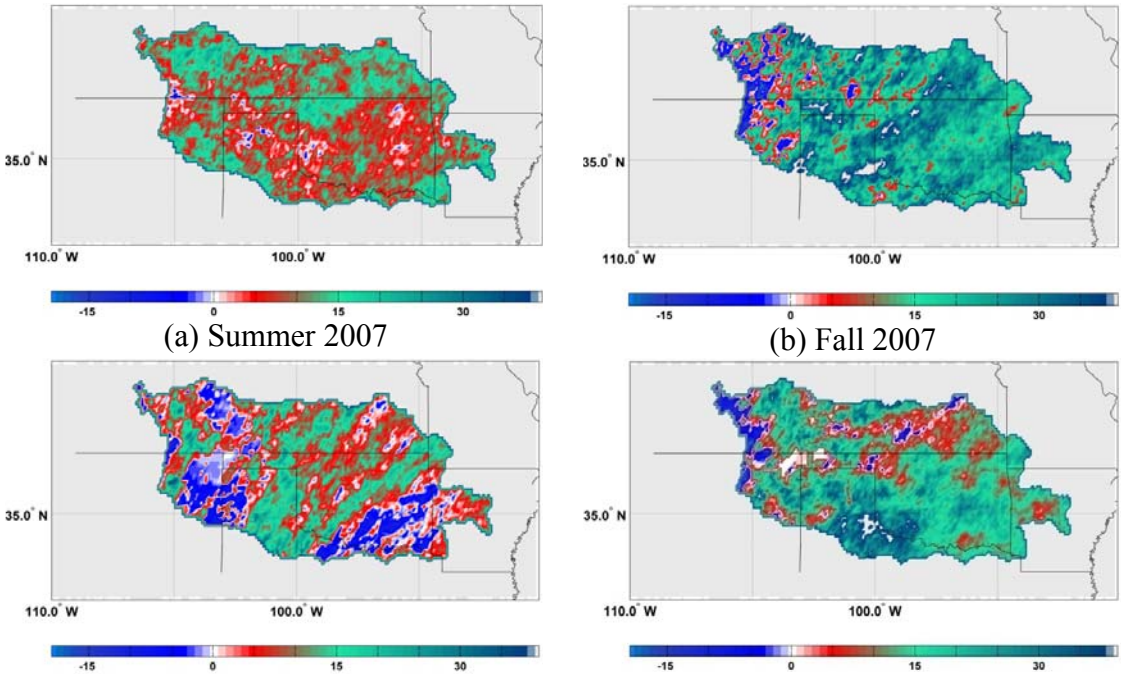
To quantify the improvement of the merged data over existing HRPPs, a similar quartile statistics for the fractional regions is presented in Table 7. Here, the quartiles reported correspond to the difference between skill scores of the merged data and the CMORPH data. From this table, it can be clearly observed that the improvement obtained from the fusion algorithm is mainly in the 3rd and 4th quartiles. Note that the skill difference is a measure between Heidke skill scores of the merged and CMORPH datasets.

Table 7. Skill difference quartile percentages of the area in the study region.

skill quartile	summer	fall	winter	spring
1	0.3556	0.6933	0.4444	0.4978
2	13.0133	7.6622	12.0178	7.0756
3	62.2044	34.2044	35.9822	39.1644
4	23.8933	41.0667	39.8756	44.7111
No improvement	0.5333	16.3733	11.6800	8.5511

4.3.3.2 Comparison with GOES AE

Figure 29 illustrates maps corresponding to the difference in the Heidke skill score between the merged data and the auto estimator (AE) data. It is seen that, during the summer and winter seasons, the difference is non-uniform throughout the region. However, in the summer, the difference is mostly positive with at least 5% improvement with a few patches of cells with a negative difference (Figure 29a). In the winter, there are grid cells with a negative difference in a large region in the eastern section and smaller region in the southwest section (blue in Figure 29c). The HSS difference is comparatively uniform during the fall and spring seasons and is, in general, positive; especially during the fall season where the improvement is over 20% in most of the region (Figure 29b). In the spring, there are few yellow regions showing only 5% improvement in HSS (Figure 29d). Both maps have a blue negative region near the western border.



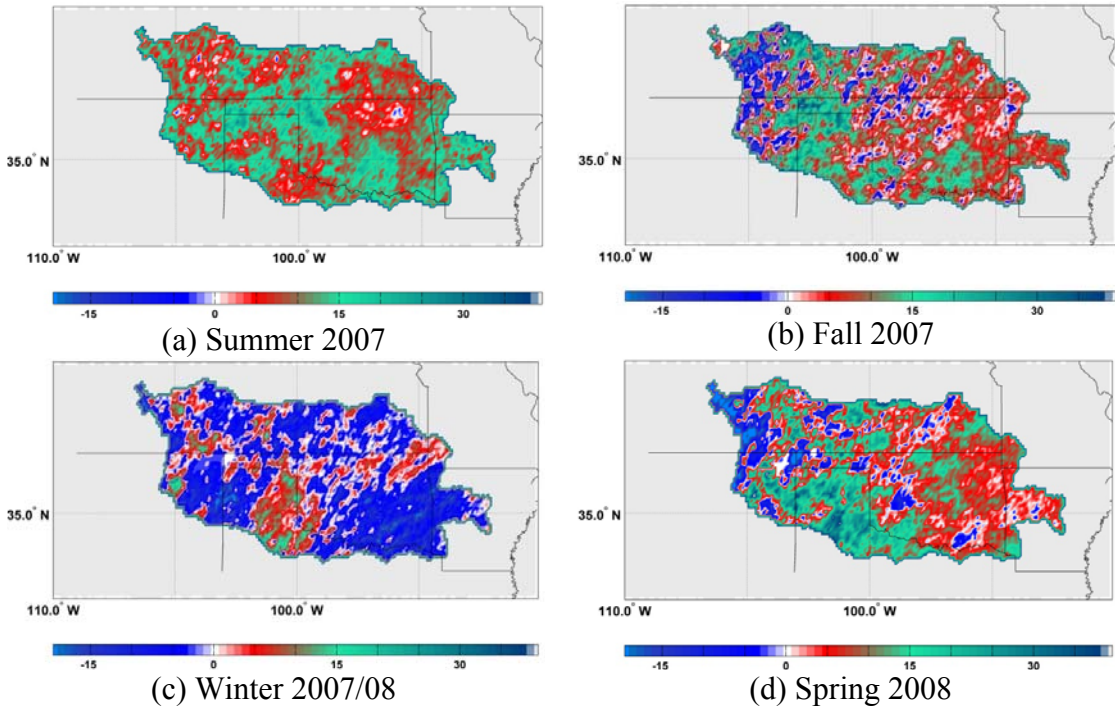
(a) Summer 2007 (b) Fall 2007

(c) Winter 2007/08 (d) Spring 2008

Figure 29. Maps and distributions of the difference in Heidke skill score between the merged data and the auto estimator data for four seasons

4.3.3.3 Comparison with GOES HE

Similarly, Figure 30 represents maps corresponding to the difference in the Heidke skill score between the merged data and the hydro estimator data. Except for the winter, the merged data performance against the hydro estimator data is similar to that against the auto estimator data (Figures 30a, 30b, and 30d). However, during the winter, the merged data is better than the GOES HE data in only 50% of the region, as shown in large blue regions in the map (Figure 30c).



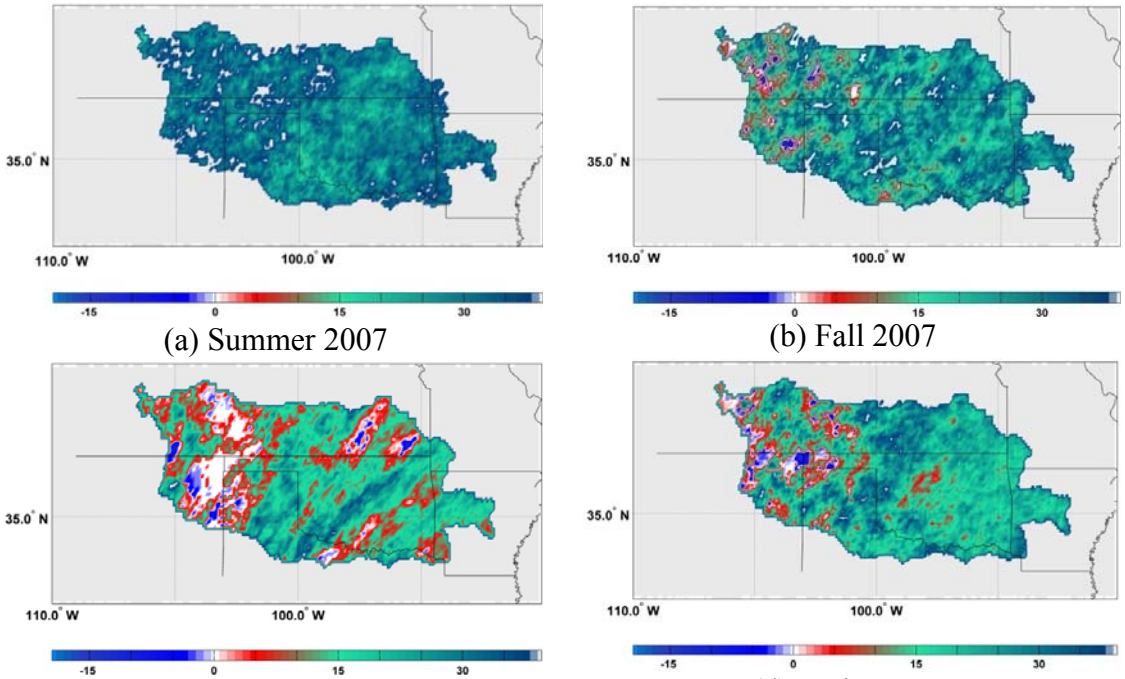
(a) Summer 2007 (b) Fall 2007

(c) Winter 2007/08 (d) Spring 2008

Figure 30. Maps and distributions of difference in Heidke skill score between the merged data and the hydro estimator data for four seasons

4.3.3.4 Comparison with NRL BLEND and SCAMPR

Finally, Figure 31 corresponds to maps generated from the difference in the Heidke skill score between the merged data and the NRL data for the summer 2007, fall 2007, and spring 08 seasons. However, during the winter, a comparison is made with the SCAMPR data. During the summer, the difference is always positive and above 25% (Figure 31a). During the fall and spring seasons, the performance is very similar to that against the auto and hydro estimators (Figures 31b and 31d). However, during the winter, the merged data has improved for most of the cells with a few small negative (blue) regions, which is not the case with other datasets (Figure 31c).



(a) Summer 2007 (b) Fall 2007

(c) SCAMPR for Winter 2007/08 (d) Spring 2008

Figure 31. Maps and distributions of the difference in Heidke skill score between the merged data and the NRL-BLEND data for different seasons

4.4 Additional discussion

The merged data is also compared with the individual datasets for all cases in terms of a success rate. This is illustrated in Figure 32. As mentioned previously, halting the training process at a non-zero error has some advantages. Furthermore, there is a correlation between the value of the minimum error and the success rate of the trained network. This is illustrated in Figure 33. Note that the error is compared with the overall success rate for the summer 2007 and fall 2007 seasons. From this figure, it is evident that, at a minimum error of 0.042, the network performs at its best in classification.

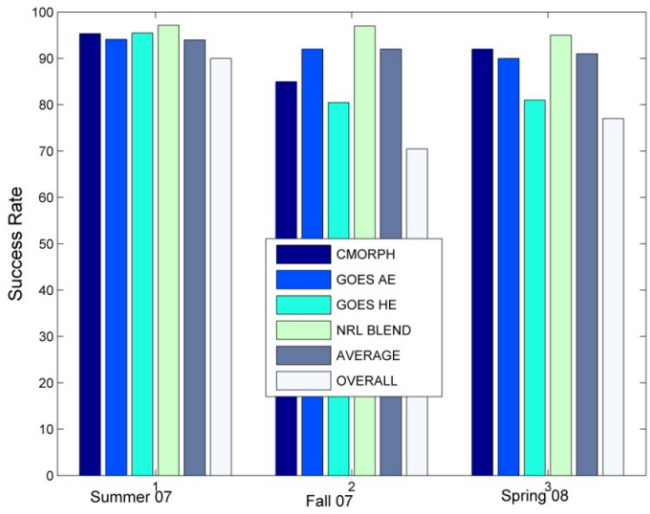


Figure 32. Comparison between the success rates of the merged data and those of the individual datasets

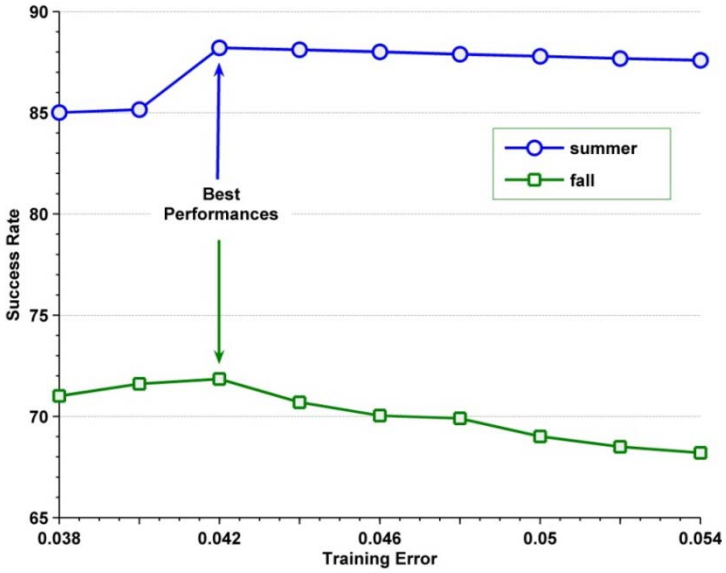


Figure 33. Comparison between the success rates and the minimum error used for early stopping of network training

In this study, several multi-layer neural networks have been tested and it is found that a two layer feed forward neural network results in the optimal mean square error during the network training. A major advantage of this approach is to improve the

accuracy of the rainfall occurrence estimation and a disadvantage is the loss of accuracy of precipitation intensity. From the distributions of HSS and difference HSS (Figures 26 - 31), it is evident that the improvement is higher during warm seasons and lower in cold seasons. This suggests that the algorithm performance is higher in a warm season and not as good in a cold season. This observation agrees with the conclusion that HRPPs from satellite-based rainfall estimation algorithms perform better during warm seasons [28]. Finally, as can be seen from Figures 27 - 31, the performance improvement of the merged product as compared to the performance of each HRPP (HSS difference) is randomly distributed throughout the study region. A potential application of the current method can be useful in inter comparisons of different HRPPs based on the success rate. This can be accomplished by selecting different combinations of HRPPs to develop a merged product and comparing the different sets of results.

CHAPTER V

CONCLUSION AND FUTURE WORK

In this research, new methods are developed for analysis and improvement of hydrological datasets derived through remote sensing. First, 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 methodology is then 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 to other study areas, particularly in semi-arid regions, and validate its performance. 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. The soil moisture data from the SCAN network is collected at hourly intervals, thus, it includes diurnal variations in soil moisture. Since the soil moisture data from AMSR-E is only a snapshot of soil water content at a particular time in any given day, a better comparison is possible between these two data-sets if we consider only the

in-situ data corresponding to a few hours centered at the acquisition time of satellite based data. As a result there will be better temporal correspondence between the two datasets. This improved training process will increase the confidence on the results from consistency analysis.

As a second task, a modified approach to SSA-based interpolation is presented. Important characteristics of this method are that spatio-temporal neighborhood of a missing value can provide useful information in estimating the missing value itself and thus reducing the computational time for the interpolation. The method is validated on three sets of geophysical data of particular importance for understanding interactions among climate processes. The interpolated data matches well with the actual data. The method is later applied to soil moisture dataset with many intermittent gaps. The interpolated time series is compared with the results obtained from the SSA gap filling method. The two approaches agreed well even when a large percentage of data was missing.

Finally, a data fusion method based on artificial neural networks augmented by a vector transformation function is developed and tested on rainfall products from several satellite-based rainfall estimation algorithms. The merged data is statistically superior to any of the individual data sets for all the seasons except the data from hydro estimator during the winter season. The next step in this work is to analyze the robustness of the algorithm by performing a sensitivity study with respect to the training data. Our methodology has the potential to enhance and add value to current as well as GPM-era precipitation estimates, especially for climate studies and other research and applications

involving retrospective studies. This method can be improved by a better selection of a vector transformation function. The current training methodology is based on using the information from the reference data. The method can be improved by developing a unsupervised approach that does not require a reference data for training.

REFERENCES

[1] X. Tang, Y. Liu, J. Zhang, "Advances in spatio-temporal analysis: an introduction," In: Advances in spatio-temporal analysis, X. Tang et.al. Ed. ISPRS Book series, Routledge, pp. 1-8, 2007.

[2] M.G. Genton, D.T. Butry, M.L. Gumpertz, and J.P. Prestemon, "Spatio-temporal analysis of wildfire ignitions in the St Johns River Water Management District, Florida," *International Journal of Wildland Fire*, vol. 15, pp. 87 – 97, 2006.

[3] Y. Qinghua, T. Guoliang, L. Xiaowen, C. Shupeng, " Tupu Methods of spatio-temporal analysis on land use/land cover change," *IEEE Proceedings of International Geoscience and Remote Sensing Symposium*,. vol. 2, pp.749 – 752, September 2004.

[4] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, "From data mining to knowledge discovery: An Overview," *In Advances in Knowledge Discovery and Data Mining,* Fayyad *et.al,* Cambridge, MA, MIT Press, pp.1-34, 1996.

[5] G.P. Lemeshewsky, "Multispectral multisensor image fusion using wavelet transforms," in: *8th conference on Visual information processing*, Orlando FL, , vol. 3716, pp: 214 – 222, April 1999.

[6] R.H. Reichle, D.H. McLaughlin, D. Entekhabi, "Hydrologic data assimilation using the Ensemble Kalman filter," *Monthly Weather Review* vol.130, pp.103-114. 2002.

[7] R.D. Koster, P.A. Dirmeyer, Z. Guo, G. Bonan, E. Chan, P. Cox, C.T. Gordon, S. Kanae, E. Kowalczyk, D. Lawrence, P. Liu, Ch-H. Lu, S. Malyshev, B. McAvaney, K. Mitchell, D. Mocko, T. Oki, K. Oleson, A.J. Pitman, Y.C. Sud, C.M. Taylor, D. Verseghy, R. Vasic, Y. Xue, and T. Yamada, "Regions of strong coupling between soil moisture and precipitation," *Science* vol. 305, pp.1138–40, 2004

[8] J. Leese, T. Jackson, A. Pitman, and P. Dirmeyer, "meeting summary: GEWEX/BAHC International Workshop on Soil Moisture Monitoring, Analysis, and Prediction for Hydrometeorological and Hydroclimatological Applications," *Bulletin of American Meteorological Society,* vol. 82, pp. 1423–1430, 2001.

[9] A. Robock, K.Y. Vinnikov, G. Srinivasan, J.K. Entin, S.E. Hollinger, N.A. Speranskaya, S. Liu, and A. Namkhai, "The Global Soil Moisture Data Bank," *Bulletin of the American Meteorological Society*, vol. 81, no. 6, pp. 1281 – 1299, 2000.

[10] K. Addison, P. Smithson, and K. Atkinson, *Fundamentals of Physical Environment*, third ed. Routledge, 2002.

[11] M. Pidwirny, *Fundamentals of Physical Geography*, 2006. http://www.physicalgeography.net/fundamentals/contents.html.

[12] W. Wu, M.A. Geller, and R.E. Dickinson, "The response of soil moisture to long-term variability of precipitation," *Journal of Hydrometeorology*, vol. 3, pp. 604-613, 2002.

[13] W. Wu, R.E. Dickinson, H. Wang, L. Yongqiang, and S. Muhammad, "Covariabilities of spring soil moisture and summertime United States precipitation in a climate simulation," *International Journal of Climatology*, vol. 27, no. 4, pp. 429-438, 2007.

[14] Y. Liu, "spatial patterns of soil moisture connected to monthly-seasonal precipitation variability in a monsoon region," *Journal of Geophysical Research*, vol. 108, no. D22, 8856, Nov. 2003. doi: 10.1029/2002JDO03124.

[15] W.T. Crow and J.D. Bolten, "Estimating precipitation errors using space-borne surface soil moisture retrievals," *Geophysical Research Letters*, vol. 34, L08403, Apr. 2007, doi: 10.1029/2007GL029450.

[16] J. Huang, H. van den Dool, and K. P. Georgakakos, "Analysis of Model-Calculated Soil Moisture over the United States (1931-93) and Application to Long-Range Temperature Forecasts," *Journal of Climate,* vol. 9, No.6, pp. 1350-1362, June 1996.

[17] G. Mostovoy and V.G. Anantharaj, "Observed and Simulated Soil Moisture Variability over the Lower Mississippi Delta Region," *Journal of Hydrometeorology*, vol. 9, pp. 1125-1150, December 2008.

[18] P.A. Dirmeyer, C.A. Schlosser, and K.L. Brubaker, "Precipitation, recycling, and land memory: An integrated analysis," *Journal of Hydrometeorology*, vol. 10, pp.278-288, July 2008. http://dx.doi.org/10.1175/2008JHM1016.1

[19] A.C. Parent, F. Anctil, and L. E. Parent, "Characterization of temporal variability in near-surface soil moisture at scales from 1h to 2 weeks," *Journal of Hydrology*, vol. 325, no. 1-4, pp. 56-66, June 2006.

[20] C. Torrence and G. P. Compo, "A practical guide to wavelet analysis," *Bulletin of the American Meteorological Society*, vol. 79, no. 1, pp. 61–78, January 1998. DOI: 10.1175/1520-0477(1998)079.

[21] A.K. Sahoo, K.X. Zhan, K. Arsenault, and M. Kafatos, "Cross-Validation of Soil Moisture Data from AMSR-E Using Field Observations and NASA's Land Data Assimilation System Simulations," *AMS 86th Annual Meeting*, Atlanta, GA, 2006.

[22] M.E. Mann and J. Park, Oscillatory spatiotemporal signal detection in climate studies: A multiple taper spectral domain approach. In: Dmowska, R., Saltzman, B., (Eds.) *Advances in Geophysics,* vol. 41, Academic Press, San Diego, CA, pp. 1-131, 1999.

[23] S.O. Los, G.J. Collatz, L. Bounoua, P.J. Sellers, and C.J. Tucker, "Global Inter-annual variations in sea surface temperature, land surface vegetation, air temperature and precipitation," *Journal of Climate*, vol. 14, no.7, pp.1535-1549, 2001.

[24] Y. Kim, and G.Wang, "Impact of initial soil moisture anomalies on subsequent precipitation over North American in the coupled land-atmosphere model CAM3-CLM3," *Journal of Hydrology*, vol. 8, no. 3, pp. 513-533, 2007.

[25] R.J. Bennet, R.P. Haining, D.A. Griffith, "The problem of missing data on spatial surfaces," *Annals of Association of American Geographers,* vol. 74, no.1, pp. 138 – 156, Mar. 1984.

[26] V. Levizzani, P. Bauer, and F.J. Turk, "Measuring Precipitation from Space: EURAINSAT and the Future," illustrated ed., Springer, Secaucus, NJ, 2007.

[27] V. Levizzani, "Satellite Clouds and Precipitation Observations for Meteorology and Climate," S. Sorooshian and F. Todini, (Eds.), *Hydrological Modeling and the Water Cycle*, Springer, Berlin, Germany, pp 49-68, 2008.

[28] E.E. Ebert, J.E. Janowiak, and C. Kidd, "Comparison of Near-Real-Time Precipitation Estimates from Sateillite Observations and Numerical Models," *Bulletin of American Meteorological Society*, vol. 88, no. 1, pp. 47-64, 2007.

[29] NASA Global Precipitation Measurement Project, Global Precipitation Measurement Mission, http://gpm.gsfc.nasa.gov, 2009.

[30] G. Schaefer, M. Cosh, and T. Jackson, "The USDA Natural Resource Conservation Service Soil Analysis Network (SCAN)," *Journal of Atmospheric and Oceanic Technology*, vol. 24, pp. 2073-2077, 2007.

[31] R. A. McPherson, C. Fiebrich, K.C. Crawford, R.L. Elliott, J.R. Kilby, D.L. Grimsley, J.E. Martinez, J.B. Basara, B.G. Illston, D.A. Morris, K.A. Kloesel, S.J. Stadler, A.D. Melvin, A.J. Sutherland, and H. Shrivastava, "Statewide monitoring of the mesoscale environment: A technical update on the Oklahoma Mesonet," *Journal of Atmospheric and Oceanic Technology*, vol. 24, pp. 301-321, 2007.

[32] J. Alvarez, N.E.C. Verhoest, J. Casali, M. Gonzalez-Audicana, and J.J. Lopez, "ADARSAT Based Surface Soil Moisture Retrieval on Agricultural Catchments of Navarre (Spain)," In: Proceedings of IEEE International Geoscience and Remote sensing symposium, Anchorage, Alaska, pp3507 – 3510, 2004.

[33] T. Lakhankar, H. Ghedira, and R. Khanbilvardi, "Soil Moisture Retrieval from Radarsat Data: A Neuron-Fuzzy Approach," *IEEE Proceedings of international Geoscience and Remote sensing symposium*, Denver, CO, pp.2328 – 2331, 2006.

[34] F.B. Sanli, Y. Kurucu, M.T. Esetlili, and S. Abdikan, "Soil moisture estimation from RADARSAT -1, ASAR and PALSAR data in agricultural fields of Menemen plane of western Turkey," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XXXVII, Part B7. Beijing, China, 2008.

[35] L. Dente, Z. Vekerdy, R. de Jeu, "Comparison of soil moisture products from active and passive microwave data," 6th European Geosciences Union General Assembly, *Geophysical Research Abstracts*, vol. 11, EGU2009-13040, http://meetingorganizer.copernicus.org/EGU2009/EGU2009-13040.pdf , 2009.

[36] K.Y. Vinnikov, A. Robock, S. Qiu, J.K. Entin, M. Owe, B.J. Choudhury, S.E. Hollinger, E.G. Njoku, "Satellite remote sensing of soil moisture in Illinois, United States," *Journal of Geophysical Research*, vol. 104, pp. 4145-4168, 1999.

[37] S. Paloscia, G. Macelloni, E. Santi, and T. Koike, "A Multifrequency algorithm for the retrieval of soil moisture on a large scale using microwave data from SMMR and SSM/I satellites," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 8, pp. 1655 – 1661, 2001.

[38] A. Guha and V. Lakshmi, "Use of the Scanning Multichannel Microwave Radiometer (SMMR) to Retrieve Soil Moisture and Surface Temperature Over the Central United States," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 7, 1482-1494, 2004.

[39] E.G. Njoku, "AMSR Land Surface Parameters Algorithm Theoretical Basis Document," Version 3.0., NASA Jet Propulsion Laboratory, Pasadena, CA, 47 pp.,http://krsc.kari.re.kr/satellite/dwd/satellite_04/AMSR/atbd-amsr-land.pdf [Accessed March 15, 2007], 1999.

[40] R. Bindlish, T.J. Jacson, A.J. Gasiewski, M. Klein, and E.G. Njoku, "Soil moisture mapping and AMSR-E validation using the PSR in SMEX02," *Remote Sensing of Environment*, vol. 103, pp. 127 -139, 2005.

[41] C. Prigent, F. Aires, W.B. Rossow, and A. Robock, "Sensitivity of satellite microwave and infrared observations to soil moisture at a global scale: Relationship of satellite observations to *in-situ* soil moisture measurements," *Journal of Geophysical Research*, vol. 110, no. D07110, 2005. doi:10.1029/2004JD005087.

[42] M. Berger, A. Camps, J. Font, Y. Kerr, J. Miller, J. Johannessen, J. Boutin, M.R. Drinkwater, N. Skou, N. Floury, M. Rast, H. Rebhan, and E. Attema, "Measuring Ocean Salinity with ESA's SMOS Mission," *ESA Bulletin*, vol. 111, no. 113f, 2002.

[43] Y.H. Kerr, J. Font, P. Waldteufel, and M. Berger, "The Second of ESA's Opportunity Missions: The Soil Moisture and Ocean Salinity Mission – SMOS," *ESA Earth Observation Quarterly*, vol. 66, 18f, 2000.

[44] T. Jackson, "Large scale field campaign contributions to soil moisture remote sensing," [abstract]. *EOS Transactions*, American Geophysical Union, Fall Supplements, vol. 88, no. 52, pp. H21J-01, 2007.

[45] A.T. Joseph, R. Van Der Velde, P.E. O'Neill, B.J. Choudhury, S. Liang, R.H. Lang, E. Kim, T.J. Gish, and P.R. Houser, "L Band observations over a corn canopy during the entire growing season," *International Journal of Remote Sensing*, vol. 46, no. 8, pp. 1-11, 2008.

[46] Y. Luo, P.R. Houser, V. Anantharaj, X. Fan, G.J. de Lannoy, L. Dabbiru, A.C. Turlapaty, and J. Aanstoos, "Potential L-Band Aquarius Radiometer and Scatterometer Soil Moisture Products from an Observing System Simulation Experiment," *American Geophysical Union Fall Meeting*, San Francisco, CA, Abstract #H23F-1034, 2008.

[47] V.G. Anantharaj and co-authors, "Potential Soil Moisture Products from the Aquarius Radiometer and Scatterometer," NASA RPC Report, GRI-TR-2009-3004, Geosystems Research Institute, Mississippi State, MS, USA, 53pp, 2009.

[48] L. Luo, A. Robock, K.E. Mitchell, P.R. Houser, E.F. Wood, J.C. Schaake, D. Lohmann, B. Cosgrove, F. Wen, J. Sheffield, Q. Duan, R.W. Higgins, R.T Pinker, and J.D. Tarpley, Validation of the North American Land Data Assimilation System (NLDAS) retrospective forcing over the southern Great Plains, *Journal of Geophysical Research*, vol. 108, no. D22, 8843, 2003. doi:10.1029/2002JD003246.

[49] D. Entekhabi, E. Njoku,P. O'Neill, M. Spencer, T. Jackson, J. Entin, I. Eastwood, and K. Kellogg, "The soil moisture active/passive mission (SMAP)," *IEEE Proceedings of International Geoscience and Remote Sensing Symposium,* vol. 3, Boston, MA, pp. III-1 - III- 4, 2008. doi:10.1109/IGARSS.2008.4779267.

[50] V.G. Anantharaj, P.R. Houser, C. Peters-Lidard, G. Mostovoy, Y. Luo, B. Li, S. Kumar, and R.J. Moorhead, "Enhancement of USDA SCAN using NASA LIS and AMSR-E," NASA RPC Report, GRI-TR-2008-3001, Geosystems Research Institute, Mississippi State, MS, USA, 39pp, 2008.

[51] P.R. Houser, W.J. Shuttleworth, S. Famiglietti, H.V. Gupta, K.H. Syed, and D.C. Goodrich, "Integration of soil moisture remote sensing and hydrologic modeling using data assimilation," *Water Resources Research*, vol. 34, no. 12, pp. 3405-3420, 1998.

[52] R. H. Reichle, R.D. Koster, P. Liu, S.P.P. Mahanama, E.G. Njoku, and M. Owe, "Comparison and assimilation of global soil moisture retrievals from the Advanced Microwave Scanning Radiometer for the Earth Observing System (AMSR-E) and the Scanning Multichannel Microwave Radiometer (SMMR)," *Journal of Geophysical Research*, vol. 112, no. D09108, 2007. doi:10.1029/2006JD008033.

[53] M. Rodell, P.R. Houser, U. Jambor, J. Gottschalck, K. Mitchell,C.J. Meng, K. Arsenault, B. Cosgrove, J. Radakovich, M. Bosilovich, J.K. Entin, J.P. Walker, D. Lohmann, D. Toll, The Global Land Data Assimilation System. *Bulletin of American Meteorological Society,* vol. 85, no. 3, pp.381–394, 2004.

[54] S.Feng, A. Qihua, and W. Qianb, "Quality Control of Daily Meteorological Data in China, 1951–2000: A New Dataset," *International Journal of Climatology*, vol. 24, pp. 853-870, 2004.

[55] L.S. Gandin, "Complex Quality Control of Meteorological Observations," *Monthly Weather Review*, vol. 116, no. 5, pp. 1137–1156. 1998.

[56] J.S. Famiglietti, J.A. Devereaux, C.A. Laymon, T. Tsegaye, P.R. Houser, T.J. Jackson, S.T. Graham, M. Rodell, and P.J. van Oevelen, "Ground based investigation of soil moisture variability within remote sensing footprints during

the southern great plains 1997 (SGP97) Hydrology Experiment," *Water Resources Research*, vol. 35, no. 6, pp. 1839-1851. 1999.

[57] J.K. Entin, A. Robock, K.Y. Vinnikov, S.E. Hollinger, S. Liu, and A. Namkhai, "Temporal and spatial scales of observed soil moisture variations in extratropics," *Journal of Geophysical Research*, vol. 105 no.9, pp. 865-877, 2000.

[58] S. Baisch and G.H.R. Bokelmann, "Spectral analysis with incomplete time series: an example from seismology," *Computers & Geosciences*, vol. 25, pp. 739-750. 1999.

[59] D.H. Schoellhamer, "Singular spectrum analysis for time series with missing data," *Geophysics Research Letters*, vol. 28, no. 16, pp. 3187-3190, http://www.ghrsst.org/L4-Gridded-SST.html , Silver Spring, MD, 2001.

[60] R. L. Smith, S. Kolenikov, and L.H. Cox, "Spatiotemporal modeling of PM2.5 data with missing values," *Journal of Geophysical Research*, vol. 108, no. D24, 9004, 2003, doi:10.1029/2002JD002914.

[61] D. Mondal, and D.B. Percival, "Wavelet variance analysis for gappy time series," *Annals of the Institute of Statistical Mathematics*, 2008. doi:10.1007/s10463-008-0195-z. http://www.springerlink.com/content/fl1503t6p52gtw14

[62] C. De Boor, *A Practical Guide to Splines*, Springer, New York, pp. 31-64, 2001.

[63] L. Li and P. Revesz, "Interpolation methods for spatio-temporal geographic data," *Computers, Environment and Urban Systems*, vol. 28, no. 3, pp. 201-227, 2004.

[64] A. Gorban, A. Rossiev, N. Makarenko, Y. Kuandykov, and V. Dergachev, "Recovering data gaps through neural network methods," *International Journal of Geomagnetism and Aeronautics*, vol. 3, no. 2, pp. 191-197, 2002.

[65] J.L. Rojo-Alvarez, C. Figuera-Pozuelo, C.E. Martinez-Cruz, G. Camps-Valls, F. Alonso-Atienza, and M. Martinez-Ramon, "Nonuniform interpolation of noisy signals using support vector machines," *IEEE Transactions on Signal Processing*, vol. 55, no. 8, pp. 4116-4126. 2007.

[66] J. Yen, "On non-uniform sampling of bandwidth-limited signals," *IRE Transactions Circuit Theory*, vol. 3, no. 4, pp. 251–257, 1956.

[67] A. M. Moffat, D. Papale, M. Reichstein, D.Y. Hollinger, A.D. Richardson, A.G. Barr, C. Beckstein, B.H. Braswell, G. Churkina, A.R. Desai, E. Falge, E., J.H. Gove, M. Heimann, D. Hui, A.J. Jarvis, J. Kattge, A. Noormets, and V.J. Stauch, "Comprehensive comparison of gap-filling techniques for eddy covariance net

carbon fluxes," *Agricultural and Forest Meteorology*, vol. 147, no. 3-4, pp. 209-232, 2007. doi:10.1016/j.agrformet.2007.08.011.

[68] K. Hocke and N. Kampfer, "Gap filling and noise reduction of unevenly sampled data by means of the Lomb-Scargle periodogram," *Atmospheric Chemistry and Physics*, vol. 9, 4197-4206, 2009.

[69] J.M. Beckers and M. Rixen, "EOF calculations and Data Filling from Incomplete Oceanographic Datasets," *Journal of Atmospheric and Oceanic Technology*, vol. 20, no. 12, 1839-1856, 2003.

[70] A. Alvera-Azcarate, A. Barth, M. Rixen, J.M. Beckers, "Reconstruction of incomplete oceanographic data sets using empirical orthogonal functions: application to the Adriatic Sea surface temperature," *Ocean Modeling*, vol. 9, no. 4, pp. 325 – 346, 2005, DOI: 10.1016/j.ocemod.2004.08.001.

[71] D.L. Hall and J. Llinas, "An introduction to multi-sensor data fusion," *IEEE Proceedings*, vol. 5, no. 1, pp.6-23, 1998.

[72] Y.M. Chiang K.L. Hsu, F.J. Chang, Y. Hong, and S. Sorooshian, "Merging multiple precipitation sources for flash flood forecasting," *Journal of Hydrology*, vol. 340, pp. 183-196. 2007.

[73] R.J.Joyce, J. E. Janowiak, P. A. Arkin, and P. Xie, "CMORPH: A method that produces global precipitation estimates from passive microwave and infrared data at high spatial and temporal resolution," *Journal of Hydrometeorology*, vol. 5, pp. 487-503, 2004.

[74] F.J. Turk, P. Arkin, E.E. Ebert, and M.R.P. Sapiano, "Evaluating High-Resolution Precipitation Products," *Bulletin of American Meteorological Society,* vol. 89, pp. 1911–1916, 2008.

[75] M.L. Nirala, "Optimal precipitation estimation using multisensor microwave datasets," *IEEE Proceedings on International Geosciences and Remote Sensing symposium,* vol. 2, pp. 875-877, 2003.

[76] L.M. Bruce, C.H. Koger, and L. Jiang, "Dimensionality reduction of hyperspectral data usingdiscrete wavelet transform feature extraction," IEEE Transactions on Geosciences and Remote Sensing, vol. 40, no. 10, pp. 2331 – 2338, 2002.

[77] K.A. Heller, K.M. Svore, A.D. Keromytis, and S.J. Stolfo, "One class support vector machines for detecting anomalous windows registry accesses,"

Proceedings of the workshop on Data Mining for Computer Security, Melbourne, FL, Nov. 2003. http://www.gatsby.ucl.ac.uk/~heller/ocsvmpr.pdf.

[78] T. Sarmiento, S.J. Hong, and G.S. May, "Fault Detection in Reactive Ion Etching Systems Using One-Class Support Vector Machines," *IEEE/SEMI Advanced Semiconductor Manufacturing Conference and Workshop*, Munich, Germany, pp.139-142, 2005.

[79] I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh, *Feature Extraction, Foundations and Applications*, Series Studies in Fuzziness and Soft Computing, Physica-Verlag, Springer, 778pp., 2006.

[80] C.S. Burrus, R.A. Gopinath, H. Guo, *Introduction to Wavelets and Wavelet Transforms: A Primer*, US edn., Prentice Hall, Upper Saddle River, N J, 268pp., August, 1997.

[81] J.E. Fowler, "The redundant discrete wavelet transform and additive noise," *IEEE Signal Processing Letters*, vol. 12, no. 9, pp. 629-632, 2005.

[82] S. Abe, *Support Vector Machines for Pattern Classification*, Series: Advances in Pattern Recognition, Springer, London, 343pp. 2005

[83] N. Cristianini and J.S. Taylor, *An Introduction to Support Vector Machines: and other Kernel-Based Learning Methods,* 1st edn., Cambridge University Press, New York, NY, 189pp., 2000.

[84] M. Friedman and A. Kandel, *Introduction to pattern recognition Statistical, structural, neural and fuzzy logic approaches*, Series in Machine perception artificial Intelligence, vol. 32, Imperial college press, London, UK., 329pp., 1999.

[85] B. Schölkopf, J. Platt, J. Shawe-Taylor A.J. Smola, and R.C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, 1443-1471. 2001.

[86] T.F. Coleman and Y. Li, "A Reflective Newton Method for Minimizing a Quadratic Function Subject to Bounds on some of the Variables," *SIAM Journal on Optimization*, vol. 6, no. 4, pp. 1040-1058, 1996.

[87] P.E. Gill, W. Murray, and M.H. Wright, *Practical Optimization*, Academic Press, London, UK., 418pp., 1981.

[88] M.H. Rusin, "A Revised Simplex Method for Quadratic Programming," *SIAM Journal on Applied Mathematics*, vol. 20, no. 2, pp. 143-160, 1971.

[89] S. Canu, Y. Grandvalet, V. Guigue, and A. Rakotomamonjy, SVM and Kernel Methods Matlab Toolbox, Perception Systèmes et Information, INSA de Rouen, Rouen, France, 2005.

[90] A. Rakotomamonjy and S. Canu, "Frames, Reproducing Kernels, Regularization and Learning," *The Journal of Machine Learning Research*, vol. 6, pp. 1485-1515, 2005.

[91] D. Kondrashov, M. Ghil, " Spatio-temporal filling of missing points in geophysical data sets," *Nonlinear Process in Geophysics*, vol. 13, pp.151-159, 2006.

[92] R.M. Gray, "Toeplitz and Circulant Matrices: A review," *Foundations and Trends in Communications and Information Theory*, vol. 2, no.2, pp. 155-239, 2006.

[93] G.H. Golub, C.F. Van Loan, *Matrix Computations, 3rd ed., Johns Hopkins, 1996.*

[94] G.H. Golub and W. Kaha, "Calculating the singular values and pseudo-inverse of a matrix," *Journal of Society of Industrial Applied Mathematics*, Series B, Numerical Analysis, vol. 2, no. 2, pp. 205-224, 1965.

[95] G.W. Stewart, "On the Early History of the Singular Value Decomposition," *SIAM Review*, vol. 35, no. 4, pp.551-566, 1993. doi:10.1137/1035134, http://citeseer.ist.psu.edu/stewart92early.html.

[96] G. Strang, "Eigenvalues and Eigenvectors," *Introduction to Linear Algebra*, 3rd edn., Wellesley-Cambridge Press, pp. 274-352, 1998.

[97] M.E. Wall, A. Rechtsteiner, and L.M. Rocha, "Singular value decomposition and principal component analysis," Berrar, D.P., Dubitzky, W., Granzow, M. (Eds.) *A Practical Approach to Microarray Data Analysis*, Kluwer, Norwell, MA, pp. 91-109, 2003.

[98] S. Esakkirajan, T. Veerakumar, and P. Navaneethan, "Best Basis Selection Using Singular Value Decomposition," *Proceedings of Seventh International Conference on Advances in Pattern Recognition (ICAPR)*, Kolkota, India, pp.65-68., 2009.

[99] M. Ghil, M.R. Allen, M.D. Dettinger, K. Ide, D. Kondrashov, M.E. Mann, A.W. Robertson, A. Saunders, Y. Tian, F. Varadi, and P. Yiou, "Advanced spectral methods for climate time series," *Reviews of Geophysics*, vol. 40, pp. 1- 41, 2001.

[100] N.S.V. Rao, "Multisensor Fusion Under Unknown Distributions: Finite Sample Performance Guarantees," A.K. Hyder (Ed.), *Multisensor fusion*, NATO Science series, Kluwer Academic Publishers, pp295-329. 2002.

[101] D. Xiao, Y. Zhou, and M. Wei, "Neural Network-based Multi-Sensor Fusion for Security Management." *1st IEEE Conference on Industrial Electronics and Applications*, pp. 5, 2006.

[102] N. Yadaiah, L. Singh, L., R.S. Bapi, V.S. Rao, B.L. Deekshatulu, and A. Negi, "Multisensor Data fusion using Neural Networks." *IEEE proceedings on International Joint conference on Neural Networks*, pp. 875-881, 2006.

[103] L. Fausett, *Fundamentals of Neural Networks Architectures, Algorithms, and applications*, Prentice Hall, New Jersey, 461 pp., Dec .1993.

[104] A.S. Pandya, and R.B. Macy, *Pattern Recognition with Neural Networks in C++*, CRC press, 1996.

[105] V.V. Phansalkar and P.S. Sastry, "Analysis of the Back-propagation Algorithm with Momentum." *IEEE Transactions on Neural Networks*, vol. 5, no. 3, pp.505 – 506, 1994.

[106] M. Dorigo, V. Maniezzo, and A. Colorni, "The Ant system: Optimization by a colony of cooperating agents," *IEEE Transactions on Systems, Man and Cybernetics*, *part B Cybernetics*, vol. 26, no. 1, pp. 1-13, 1996.

[107] T. Kwok, and D.Y. Yeung, "Efficient cross-validation for feedforward neural networks," *IEEE proceedings of International conference on Neural Networks*, vol. 5, pp. 2789-2794, Dec. 1995.

[108] D.D. Bosch, "Comparison of Capacitance-Based Soil Water Probes in Coastal Plain Soils," *Vadose Zone Journal*, vol. 3, 1380-1389, 2004.

[109] J.R. Kennedy, T.O. Keefer, G.B. Paige, and E. Barnes, "Evaluation of dielectric constant-based soil moisture sensors in a semiarid rangeland," *Proceedings of First Interagency Conference on Research in the Watersheds*, Benson, AZ, pp. 503 – 508, 2003.

[110] M.S. Seyfried, and M.D. Murdock, "Measurement of Soil Water Content with a 50-MHz Soil Dielectric Sensor," *Soil Science Society of America*, vol. 68, pp. 394-403, 2004.

[111] Soil Survey Staff, "National Soil Survey Characterization Data," Soil Survey Laboratory, National Soil Survey Center, USDA-NRCS - Lincoln, NE, 2007.

[112] E.G. Njoku, T.J. Jackson, L. Venkataraman, T.K. Chan, and S.V. Nghiem, "Soil Moisture Retrieval from AMSR-E," *IEEE Transactions on Geosciences and Remote Sensing*, vol. 41, no. 2, pp. 215-229, 2003.

[113] T. Kawanishi, T. Sezai, Y. Ito, K. Imaoka, T. Takeshima, Y. Ishido, A. Shibata, M. Miura, H. Inahata, and R.W. Spencer, "The advanced microwave scanning radiometer for the Earth observing system (AMSR-E), NASA's Contribution to the EOS for global energy and water cycle studies," *IEEE Transactions on Geosciences and Remote Sensing*, vol. 41, no. 2, pp. 184 – 194, 2003.

[114] E.G. Njoku, AMSR-E/Aqua daily L3 surface soil moisture, interpretive parms, & QC EASE-Grids, Jan 2005 to Dec 2006, Boulder, CO, USA: National Snow and Ice Data Center, Digital media, 2005. https://wist.echo.nasa.gov/api/, [Accessed March 15, 2007].

[115] S. Paloscia, G. Macelloni, and E. Santi, "Soil moisture estimates from AMSR-E brightness temperatures by using a dual-frequency algorithm," *IEEE Transactions on Geosciences and Remote Sensing*, vol. 44, no. 11, pp. 3135-3144, 2006.

[116] O. Duzenli and F.P. Monique, "Wavelet-based feature extraction methods for classification Applications," *IEEE Proceedings of 9th SP Workshop on Statistical Signal and Array Processing*, Portland, OR, pp.176 – 179, 1998.

[117] A.Webb, *Statistical Pattern Recognition*, Arnold, London, UK, 454pp, 1999.

[118] S.E. El-khamy, A.A. Alim, and M. Saii, "Neural network face recognition using statistical feature extraction," *17th National Radio Science Conference*, Minufiya, Egypt, pp.1-8. 2000.

[119] E.G. Njoku and L. Li, "Retrieval of Land Surface Parameters Using Passive Microwave Measurements at 6–18 GHz," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, no. 1, pp. 79-93, 1999.

[120] R.A.M. De Jeu, W. Wagner, T.R.H. Holmes, A.J. Dolman, N.C. van de Giesen, and J. Friesen, "Global Soil Moisture Pattern Observed by Space Borne Microwave Radiometers and Scatterometers," *Surveys in Geophysics*, vol. 29, pp. 399 - 420, 2008.

[121] J.P. Hollinger, J. L. Pierce, and G.A. Poe, "SSM/I instrument evaluation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 28, pp. 781-790, 1990.

[122] E.G. Njoku and D. Entekhabi, "Passive microwave remote sensing of soil moisture," *Journal of Hydrology*, vol. 184, pp. 101–129, 1996.

[123] J.P. Wigneron, J.C. Calvetb, T. Pellarinb, A.A. Van de Griendc, M. Bergerd, and P. Ferrazzolie, "Retrieving near-surface soil moisture from microwave radiometric observations: current status and future plans," *Remote Sensing of Environment*, vol. 85, pp. 489-506, 2003.

[124] H. Beggs, "A high-resolution blended sea surface temperature analysis over the Australian region," *BMRC Research Report No. 130*, Bureau of Meteorology Research Centre, Melbourne, Australia, 43pp. 2007. http://cawcr.gov.au/bmrc/pubs/researchreports/RR130.pdf [accessed 10 October 2009]

[125] K.S. Casey, "GODAE high resolution SST products from the GDAC and LTSRF", Long term stewardship and reanalysis facility NOAA national oceanographic data center, Silver Spring, MD, USA, 2007. http://www.ghrsst.org/L4-Gridded-SST.html

[126] C.O. Justice, J.R.G. Townshend, E.F. Vermote, E. Masouka, R.E. Wolfe, N. Saleous, D.P. Roy, and J.T. Morisette, "An overview of MODIS Land data processing and product status," *Remote Sensing of Environment*, vol. 83, no.1-2, pp. 3-15, 2002.

[127] A. Savtchenko, D. Ouzounov, S. Ahmad, J. Acker, G. Leptoukh, J. Koziana, and D. Nickless, "Terra and Aqua MODIS products available from NASA GES DAAC," *Advanced Space Research*, vol. 34, no. 4, pp. 710-714, 2004.

[128] Z. Wan and L. Li, "Radiance-based validation of the V5 MODIS land-surface temperature product. International," *Journal of Remote Sensing*, vol. 29, no. 17-18, pp. 5373-5395, 2008.

[129] G.C. Hulley and S.J. Hook, "Intercomparison of versions 4, 4.1 and 5 of the MODIS Land Surface Temperature and Emissivity products and validation with laboratory measurements of sand samples from the Namib desert, Namibia," *Remote Sensing of Environment*, vol. 113, pp.1313-1318, 2009.

[130] Z. Wan, "Collection -5 MODIS Land Surface Temperature Products Users' Guide," ICESS, University of California, Santa Barbara, 2009. http://www.icess.ucsb.edu/modis/LstUsrGuide/MODIS_LST_products_Users_guide_C5.pdf

[131] D. Conway, "Advanced Microwave Scanning Radiometer - EOS Quality Assurance Plan," Global Hydrology and Climate Center , Huntsville, AL, 2002.

[132] E.G. Njoku, and S.K. Chan, "Vegetation and surface roughness effects on AMSR-E land observations," *Remote Sensing of Environment*, vol. 100, no. 2, pp.. 190-199, 2006.

[133] E.G. Njoku, T. Chan, W. Crosson, and A. Limaye, "Evaluation of the AMSR-E Data Calibration Over Land," *Italian Journal of Remote Sensing*, vol. 29, no. 4, pp. 19-37, 2004.

[134] G.A. Vincente, R.A. Scofield, and W.P. Menzel, "The Operational GOES Infrared Rainfall Estimation Technique," *Bulletin of American Meteorological Society, vol.* 79, no. 9, pp. 1883–1898, 1998.

[135] R.A. Scofield, and R.J. Kuligowski, "Status and Outlook of Operational Satellite Precipitation Algorithms for Extreme-Precipitation Events," *Weather Forecast*, vol. 18, no. 6, pp. 1037-1051, 2003.

[136] F.J. Turk, G. Mostovoy, and V.G. Anantharaj, "The NRL-Blend High Resolution Precipitation Product and its Application to Land Surface Hydrology," F. Hossain and Gebremichael, M., (Eds), *Satellite Application for surface Hydrology,* Springer Verlag, 2009.

[137] R.J. Kuligowski, "A Self Calibrating Real-Time GOES Rainfall Algorithm for Short-Term Rainfall Estimates," *Journal of Hydrometeorology*, vol. 3, no. 2, pp. 112-130, 2002.

[138] J. Schmidt, B. Lawrence, and B. Olsen, "A Comparison of Operational Precipitation Processing Methodologies," Arkansas-Red Basin River Forecast Center, pp. 5, 2000. http://www.srh.noaa.gov/abrfc/p1vol.html

[139] C.B. Young, A.A. Bradley, W.F. Krajewski, A. Kruger, and M.L. Morrissey, "Evaluating NEXRAD Multisensor Precipitation Estimates for Operational Hydrologic Forecasting," *Journal of Hydrometeorology*, vol. 1, no. 3, pp. 241-254, 2000.

[140] R.R. Ferraro, E.A. Smith, W. Berg, and G.J. Huffman, "A screening methodology for passive microwave precipitation algorithms," *Journal of Atmospheric Sciences*, vol. 55, pp. 1583-1600, 1998.

[141] F. Woodcock, "The evaluation of yes/no forecasts for scientific and administrative purposes," *Monthly Weather Review*, vol. 108, pp. 292-297, 1976.

[142] C. Kidd, and P. Joe, "Importance, Identification and measurement of light precipitation at mid to high latitudes," *Proceedings of Joint EUMETSAT Meteorological Satellite Conference and the 15th Satellite Meteorology and oceanic Conference*, American Meteorological Society, Amsterdam, pp.24-28, September 2007.