DERIVING WATER FRACTION AND FLOOD MAP WITH THE EOS/MODIS DATA USING REGRESSION TREE APPROACH

D. L. Sun[a], Y. Y. Yu[b]

[a]Department of Geography and Geoinformation Sciences, George Mason University
Fairfax, VA 22030,USA (dsun@gmu.edu)
[b]NOAA/NESDIS, Center for Satellite Applications and Research, Camp Spring, MD 20746, USA -
yunyue.yu@noaa.gov

KEY WORDS: Regression Tree, Flood, MODIS, Water fraction

ABSTRACT:

This study investigates how to derive water fraction and flood map from the Moderate-Resolution Imaging Spectroradiometer (MODIS) onboard the Earth Observing System (EOS) using a Regression Tree (RT) approach. The RT approach can integrate all the possible candidate predictors, such as the MODIS channel 2 reflectance (CH2), reflectance ratio (CH2/CH1), reflectance difference (CH2-CH1) between MODIS channels 2 and 1, vegetation and water indices. Meanwhile, it provides accuracy estimates of the derivation. The recent floods in New Orleans area in August 2005 were selected for the study. MODIS surface reflectance with the matched surface water fraction data were used for the RT training. From the training set, 60% were used for training, and the remaining 40% for test. Rules and regression models from the RT training were applied for real applications to New Orleans flooding in 2005 to calculate water fraction values. Flood distributions in both space and time domains were generated using the differences in water fraction values after and before the flooding. The derived water fraction maps were evaluated using higher resolution Thematic Mapper (TM) data from the Landsat observations. It shows that correlation between the water fractions derived from the MODIS and TM data is 0.97, with difference or "bias" of 2.16%, standard deviation of 3.89%, and root mean square error (rmse) of 4.45%. The results show that the RT approach in dynamic monitoring of floods is promising.

1. INTRODUCTION

Satellite-derived flood maps in near-real time are vital to stake holders and policy makers for disaster monitoring and relief efforts. Precise mapping of floods and standing water is also required for detecting deficiencies in existing flood control and for damage claims.

Satellite sensors used in river and flood studies may be classified into two types: (1) passive, in which the sensor receives energy naturally reflected by or/and emitted from the earth's surface; and (2) active, in which the sensor provides illumination and records the amount of incident energy returned from the sensed surface (Smith, 1997). Sample passive sensors in visible and infrared spectrums are the Thematic Mapper (TM) and Multi-Spectral Scanner (MSS) onboard the Landsat satellites, the Advanced Very High Resolution Radiometer (AVHRR) onboard NOAA polar-orbiting meteorological satellites, Visible High Resolution (HRV) sensor onboard the Satellite Pour l'Observation de la Terre (SPOT), the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) and the Moderate-Resolution Imaging Spectroradiometer (MODIS) onboard the Earth Observation System (EOS) satellites. Passive microwave radiometers, such as the Special Sensor Microwave/Imager (SSM/I) on board the defense meteorological satellites, can transpire clouds and measure the microwave energy naturally emitted from the Earth's surface. The coarse spatial resolution of these microwave sensors (ca. 27 km at 37 GHz) has been mitigated through the combined use with the visible and infrared sensors for the flood detection (Hallberg et al., 1973; Sipple et al., 1992; Toyra et al., 2001; 2002).

Much of the pioneering work on the remote sensing of floods was accomplished using the MSS sensor on the First Earth Resources Technology Satellite, later renamed Landsat-1.

With a spatial resolution of about 80 m, MSS data was used to mapping the extent of flooding in Iowa (Hallberg et al., 1973; Rango and Salomonson, 1974), Arizona (Morrison and Cooley, 1973), Virginia (Rango and Salomonson, 1974) and along the Mississippi River (Deutsch et al., 1973; Deutsch, and Ruggles, 1974; Rango and Anderson, 1974; McGinnis and Rango, 1975; Deutsch, 1976; Morrison and White, 1976) . All of these studies show that MSS band 7 (0.8-1.1 µm) was the most useful for separating water from dry soil or vegetated surfaces due to the strong absorption of water in the near-infrared range. This feature was further confirmed by analyzing MSS band 5 (0.6-0.7 µm), band 7 and field spectral radiometer data along shoreline water-wet soil-dry soil transitions by Gupta and Banerji (Gupta and Banerji 1985). Flooded areas were delineated based on the sharp contrast between water spread and adjacent areas. The standing water areas appeared as dark blue to light blue depending upon the depth of water, while the receded water and wet areas appeared as dark to light gray.

Other studies have continued the methodology developed with the MSS, using Landsat TM and SPOT data (France and Hedges, 1986; Jensen et al., 1986; Watson, 1991; Blasco et al., 1992; Pope et al., 1992; da Silva, 1992). The coarser spatial resolution (ca. 1 km) sensors, such as the AVHRR, have been successfully used for studying large river floods (Ali et al., 1989; Gale and Bainbridge, 1990; Rasid and Pramanik, 1993).

Sheng et al. (2001) summarized the spectral characteristics of the main features (i.e. water, vegetation, soil, and clouds) during floods at the observation scale of NOAA satellites. Although AVHRR data can be displayed in 3-channel color composites for visual analysis (flood and standing water absorbs infrared wavelengths of energy and appears as blue/black in the RGB composite imagery), water body identification in AVHRR imagery evolved from qualitative visual interpretation to automatic quantitative extraction. The reflectance of AVHRR channel 2 (0.73-1.1µm, similar to MSS band 7), the reflectance difference (CH2-CH1) and ratio (CH2/CH1) between channel 2 and 1 (0.58-0.68 µm, similar to MSS band 5) are used to discriminate water from land if these parameters are less than the threshold values.

Domenikiotis *et al*. (2003) tried to use surface temperature to discriminate water from land surfaces. However, the temperature model may not work well with the flood caused by heavy rainfall during rainy seasons in the summer when there is relatively low or no temperature difference between land and water. Domenikiotis *et al*. (2003) also used Normalized Difference Vegetation Index (NDVI) to identify water from land considering that water covered surfaces usually have very small or even negative NDVI values. It can be seen from its mathematical definition that the NDVI of an area containing a dense vegetation canopy will tend to have positive values (say 0.3 to 0.8), while standing water (e.g., oceans, seas, lakes and rivers), which have a rather low reflectance in both visible (VIS: from 0.4 to 0.7 µm) and near-infrared (NIR: from 0.7 to 1.1 µm) spectral bands, result in very low positive or even slightly negative NDVI values.

Regression trees have been used with remote sensing observations (DeFries et al., 1997; Mchaelson, Schimel, Friedl, Davis and Dubayah, 1994; Prince and Steninger, 1999; Hansen et al., 2002, Solomatine and Xue, 2004). They provide a robust tool to handle nonlinear relationship within large data sets.

As described above, in previous studies, several parameters, including the reflectance of near infrared (NIR) channel, the reflectance ratio and difference between NIR and visible (VIS) channels, NDVI, brightness temperature at 11 or 12 µm, and surface temperature, might be used to identify water from land. Linear mixture model has been used by Sheng et al. (2001) to derive water fraction. However, it has not yet been shown which parameter or combination of several parameters is the most effective?

This paper explores how to derive water fraction and flood map from the MODIS data using regression tree (RT) method. Section 2 introduces the dataset used. The physics of the problem and decision algorithms are described in Section 3. Section 4 presents the results and Section 5 gives a summary and discussion.

2. DATA USED

- Surface water percentage data derived from derived from the 1km land/water map supplied by the USGS Global Land Cover Characterization Project. The percentage water was created by simply determining the percentage of 1km pixels designated as water in each 10' region. This data can be obtained from the Surface and Atmospheric Radiation Budget (SARB) working group, part of NASA Langley Research Center's Clouds and the Earth's Radiant Energy System (CERES) mission
- MODIS L3 8-day composite surface reflectance product (MYD09A1) that is computed from the MODIS Level 1B land bands 1, 2, 3, 4, 5, 6, 7, which are centered at 0.648 µm, 0.858 µm, 0.470 µm, 0.555 µm, 1.24 µm, 1.64 µm, and 2.13 µm, respectively. The product is an estimate of the surface reflectance for each band as it would have been measured at ground level after removing the atmospheric scattering and absorption.
- MODIS L1B calibrated reflectance at the Top of Atmosphere (TOA) with 1 km resolution (MOD021KM).
- MODIS geolocation fields (MOD03).
- MODIS cloud mask (MOD35) data.
- TM (Thematic Mapper) data from the Landsat observations at 30-meter spatial resolution is used to evaluate water fraction derived from MODIS.

3. METHODOLOGY

The RT, such as the M5P, is a powerful tool for generating rule-based models that balance the need for accurate prediction against the requirements of intelligibility. RT models generally give better results than those produced by simple techniques such as multivariate linear regression, while also being easier to understand than neural networks. Unlike neural networks, the RT program generates a model with rules that describe the relationships between the independent and dependent parameters in the data set. Instead of simple regression analysis techniques, RT uses a piecewise regression technique. The piecewise regression analysis (classifying the data into different subsets) will yield different regression fits for different meteorological conditions, unlike a simple regression analysis. The RT program constructs an unconventional type of tree structure, with the leaves containing linear models instead of discrete classes by DT. A decision tree would categorize the predictions into discrete classes, but the regression tree predicts actual continuous values.

Since RT integrates DT with traditional regression analysis. Like DT algorithm, RT algorithm can integrate all the possible candidate predictors, such as the MODIS channel 2 reflectance (CH2) and channel 1 reflectance CH1, the reflectance ratio (CH2/CH1) and difference (CH2-CH1) between MODIS channel 2 and channel 1, NDVI, Normalized Water Difference Index (NDWI), etc., meanwhile it can determine continuous values, in this case water fraction, and giving accuracy estimates. The NDWI [45], a satellite-derived index from the Near-Infrared (NIR) and Short Wave Infrared (SWIR) channels, is also included as one input attribute. According to Gao [45], NDWI is a good indicator for vegetation liquid water content and is less sensitive to atmospheric scattering effects than NDVI. The MODIS 8-day composite data at 500-m resolution is aggregated to the same 1/6 degree resolution of the surface water percentage map.

In this study, the M5P (Wang and Witten, 1997), a reconstruction of Quinlan's M5 algorithm (Quinlan, 1992) for inducing trees of regression models, is used to derive water fraction from MODIS observations. The M5P combines a conventional decision tree with the possibility of linear regression functions at the nodes. Techniques devised by Breiman et al. (1984) for their CART (Classification and Regression Trees) system are adapted in order to deal with enumerated attributes and missing values. Uses features from the well-known CART system and reimplements Quinlan"s well-known M5 algorithm with modifications and seems to outperform it. M5P can deal effectively with enumerated attributes and missing values.

4. RESULTS

4.1 Results from the RT training

Training set is critical to RT. MODIS 8-day composite surface reflectance and Surface water percentage data derived from derived from the 1km USGS land/water map, as shown in Figure 1, are used as the training datasets.

Figure 2 shows an example of the output regression tree structure with the M5P algorithm. The tree employs a case's attribute values to map it to a *leaf* designating one of the regression models (Figure 3). The first number in brackets following each leaf is the number of training instances falling into this leaf and the second number is the root mean squared error of the linear model on these training examples divided by the global absolute deviation.

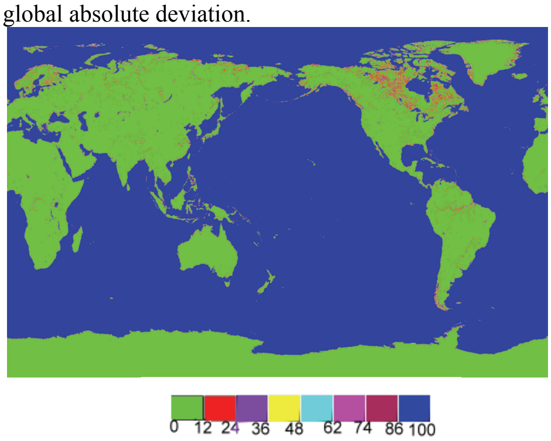

Figure 1. Water percentage map derived from the 1km USGS land/water map.

Regression models generated from the M5P regression tree algorithm are shown in the following:

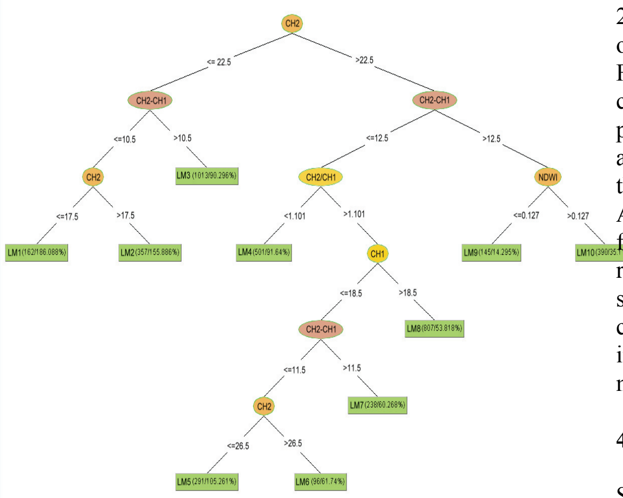

Figure 2. An example of regression tree structure derived from the M5P algorithm.

LM1: WF = -162.6945*CH1 - 0.583*CH2 - 320.8135*(CH2-CH1) + 0.3253*CH2/CH1 + 39.981*NDWI + 39.7147
LM2: WF = -0.0152*CH1 - 0.583*CH2 - 11.6055*(CH2-CH1) + 0.1568*CH2/CH1 + 0.9568*NDWI + 3.9752
LM3: WF = -0.0152*CH1 - 0.336*CH2 - 71.4079*(CH2-CH1) - 0.7859*CH2/CH1 + 19.7358*NDWI + 8.5573
LM4: WF = 0.2117*CH1 - 9.0327*CH2 - 0.9369*(CH2-CH1) - 33.0853*CH2/CH1 - 7.0218*NDWI + 43.6346
LM5: WF = -2.223*CH1 + 0.5925*CH2 - 5.2925*(CH2-CH1) + 0.0591*CH2/CH1 + 0.4396*NDWI + 2.2389
LM6: WF = -4.0638*CH1 + 0.5925*CH2 - 8.7526*(CH2-CH1) + 0.0591*CH2/CH1 + 20.3801*DWI - 0.7155
LM7: WF = -1.7928*CH1 + 0.9452*CH2 - 4.5942*(CH2-CH1) + 0.0591*CH2/CH1 + 0.2035*NDWI + 1.2859
LM8: WF = -0.0818*CH1 - 0.0378 *CH2 - 1.0327*(CH2-CH1) + 0.0524*CH2/CH1 - 1.7571*NDWI + 1.0939
LM9: WF = -0.6528*CH1 + 0.0744*CH2 - 1.2964*(CH2-CH1) - 0.0338*CH2/CH1 + 0.064*NDWI + 0.4347
LM10: WF = -0.2448*CH1 - 4.3881*CH2 - 0.9276*(CH2-CH1) - 0.5367*CH2/CH1 + 0.009*NDWI + 2.8895

Each regression model consists of:

- A linear model (LM) number -- this serves to identify the regression model.
- Every enumerated model is composed of a regression equation.

Figure 2 just shows an example of the output regression tree structure. The actual tree structure is too complicated to be shown in a figure.

4.2 Results from the tests with real applications

Since we wish to get the geolocation information, instead of using surface reflectance data (MOD09), we chose to use the MODIS L1B calibrated TOA reflectance data (MOD021KM) in conjunction with the MODIS geolocation fields (MOD03). An accurate cloud filter for the Imager data is critical for reliable results. Since our method uses satellite visible and infrared observations, the water detection will be limited to clear conditions. MODIS cloud mask (MOD35) data is used to filter the cloudy conditions. The rules and threshold values obtained from the training with surface reflectance data are applied to “re-predict” the New Orleans flooding at the end of August in 2005 due to the landfall of Hurricane Katrina, which caused over 1500 deaths and total damage costs exceeding $50 billion. Figure 3 shows the water fraction map on these three days, calculated by using the CH2 reflectance and (CH2-CH1) predictors. From these images, we can clearly detect flooded areas by comparing water fraction maps after flooding with those before flooding. Figure 4 presents the flood maps on August 31 and 30 as the difference in water fraction values after flooding with those before flooding on August 27. The flooded regions are identified in red, the original water bodies are shown in blue, while clouds are marked in grey. We can see clearly that New Orleans and its surrounded areas were inundated on August 30 and 31, 2005 after Hurricane Katrina made landfall on August 29, 2005.

4.3 Evaluations

Since there are no direct ground measurements of water fraction as the truth data, quantitative evaluations of water fraction derived from satellite observations are challenging. The use of higher resolution satellite data is a feasible way to solve this problem. In this study, TM data with 30-m spatial resolution are used to evaluate water fraction estimates from MODIS observations.

The Landsat TM pixels can be assumed to be a pure pixel composed of land or water. Using a decision tree method to perform classification to TM data, the fraction of water in a MODIS grid (1 km×1 km) can be calculated. The water fractions at the MODIS resolution aggregated from the TM observations are then used to evaluate water fractions derived from MODIS observations. The scatter plot is shown in Figure 5, the evaluation results show correlation between MODIS and TM water fractions is 0.966 with bias of 2.16%, standard deviation of 3.89%, and rms of 4.45%, for total sample number of 50423.

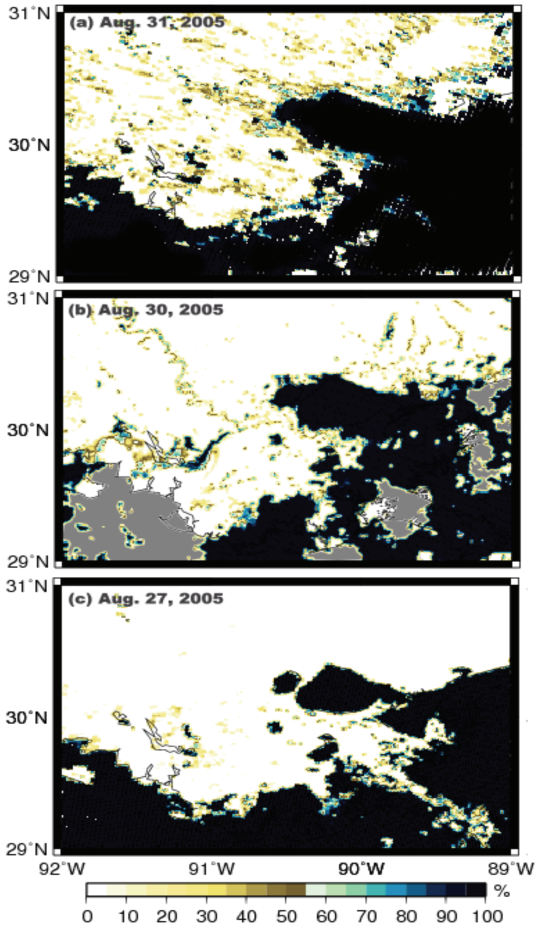

Figure 3. Water fraction map on August 31 (a), 30 (b), and 27 (c), 2008.Figure 3. Water fraction map on August 31 (a), 30 (b), and 27 (c), 2008.

5. SUMMARY

In this study, the Regression Tree technique is applied to water body and flood identification with the EOS MODIS data. MODIS data has the advantage of global coverage, and so can be available worldwide. MODIS surface reflectance with the matched surface percent water data before flooding are used for training with the RT method. MODIS surface reflectance data at 500m resolution are aggregated to the same 1/6 degree resolution as the percent water data. When we test the rules and regression models obtained from the training to "predict" or model future flood, in order to get the geolocation information, we use the Level 1B swath 1km calibrated reflectances at the TOA with the matched geolocation fields.

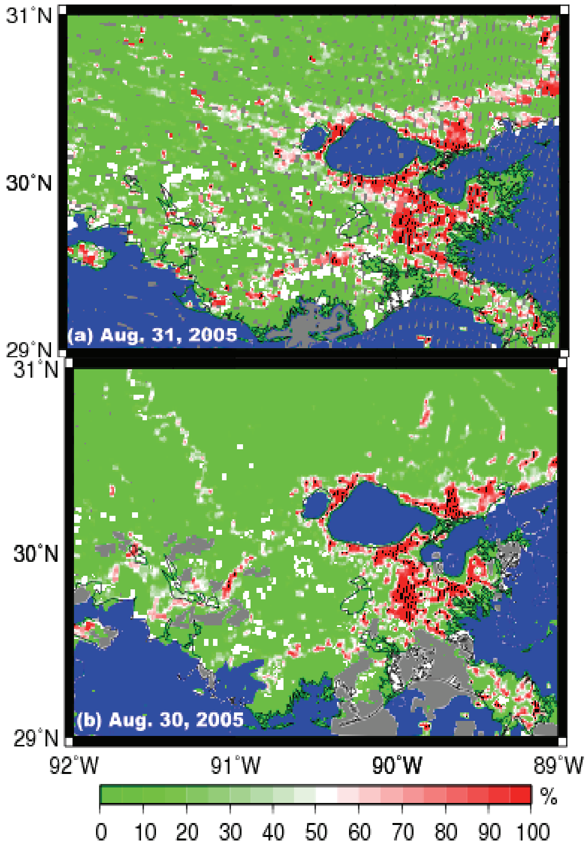

Figure 4. Flood map on August 31 (a) and 30 (b), 2005 shown as the water fraction difference after and before flooding (August 27).

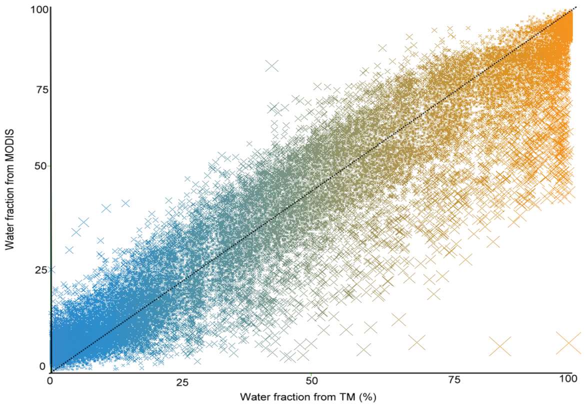

Figure 5. Scatter plot of water fractions of MODIS and TM on 08/27/2005 using regression-tree algorithm.

The time series of water fraction maps are generated, monitoring area changes of inundation. The flood maps are derived by calculating the difference in water fraction before and after flooding, and show promising results. The successful applications of MODIS observations to water body and flood identification demonstrate the effectiveness of the RT approach.

REFERENCES

Ali, A., D. A. Quadir, O. K. Huh, 1989. Study of river flood hydrology in Bangladesh with AVHRR data. *International Journal of Remote Sensing*, 10 (12), pp.1873-1891.

Blasco, F., M. F. Bellan, and M. U. Chaudhury, 1992. Estimating the extent of floods in Bangladesh using SPOT data. *Remote Sensing Environment*. 39, pp.167-178.

Breiman, L., 2004. Random Forests. *Machine Learning*, 45 (1), pp. 5-32.

Breiman, L., J. H. Friedman, R. A. Olshen, C. J. Stone, 1984. Classification and Regression Trees. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1984.

Deutsch, M., F. Ruggles, 1974. Optical data processing and projected applications of the ERTS-1 imagery covering the 1973 Mississippi River Valley floods. *Water Resource Bulletin,* 10 (5), pp. 1023-1039.

Domenikiotis, C., A. Loukas, N. R. Dalezios, 2003. The use of NOAA/AVHRR satellite data for monitoring and assessment of forest fires and floods. *Natural Hazards Earth System Science*, 3, pp. 115-128.

France, M. J., P. D. Hedges, 1986. A hydrological comparison of Landsat TM, Landsat MSS, and black and white aerial photography in Damen. Smit, and Verstappen (Eds)," Proceedings 7th International Symposium, ISPRS (International Society Photogrammetry and Remote Sensing) Commission VII, Enschede, pp. 717-720.

Gale, S. J., S. Bainbridge, 1990: The floods in eastern Australia. *Nature*, 345, pp. 767.

Gao, B.-C., 1996. NDWI - A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sensing of Environment*, 58, pp. 257-266.

Gupta R. P., S. Banerji, 1985. Monitoring of reservoir volume using Landsat data. *Journal of Hydrology*, 77, pp. 159-170.

Hallberg, G. R., B. E. Hoyer, A. Rango, 1973. Application of ERTS-1 imagery to flood inundation mapping. NASA Special Pub. No. 327, Symposium on significant results obtained from the Earth Resources Satellite-1, Vol. 1, Technical presentations, section A, 745-753.

Hansen, M.C., R.S. DeFries, J.R.G. Townshend,R. Sohlberg, C. Dimiceli , M. Carroll, 2002. Towards an operational MODIS continuous field of percent tree cover algorithm: examples using AVHRR and MODIS data. Remote Sensing of Environment 83, 303–319.

Han, J.W., Kamber, M., 2001. Data Mining: Concept and Techniques. Morgan Kaufmann Publishers, San Francisco, CA, 550pp.

Kohavi, R., 1996. Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996.

Jensen, J. R., M. E. Hodgson, E. Christensen, H. E., Jr. Mackey, L. R. Tinney, R. Sharitz, 1986. Remote sensing inland wetlands: a multispectral approach, *Photogrammetric Engineering Remote Sensing*, 52, pp. 87-100.

McGinnis, D. F. and A. Rango, 1975. Earth Resources Satellite systems for flood monitoring. *Geophysical Research Letters*, 2 (4), pp. 132-135.

Morrison, R. B., P. G. White, 1976. Monitoring flood inundation. U.S. Geol. Surv. Prof. Pap. 929, ERTS-1, A New Window on Our Planet, pp. 196-208.

Pope, K. O., E. J. Sheffner, K. J. Linthicum, C. L. Bailey, T. M. Logan, E. S. Kasichke, K. Birney, A. R. Njogu, C. R. Roberts, 1992. Identification of Central Kenyan Rift Valley Fever virus vector habitats with Landsat TM and evaluation of their flooding status with airbourne imaging radar. *Remote Sensing of Environment,* 40, 185-196.

Quinlan, R. J., 1992. Learning with Continuous Classes. In: 5th Australian Joint Conference on Artificial Intelligence, Singapore, pp. 343-348.

Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Francisco, CA, 316pp.

Rasid, H., M. A. H. Pramanik, 1993. Areal extent of the 1988 flood in Bangladesh: how much did the satellite imagery show? *Natural Hazards,* 8, pp. 189-200.

Running, S., T. R. Loveland, L. Pierce, 1995. A Remote Sensing Based Vegetation Classification Logic for Global Land Cover Analysis. *Remote Sensing Environment*, 51, pp. 39-48.

Sheng, Y., P. Gong, Q. Xiao, 2001. Quantitative dynamic flood monitoring with NOAA AVHRR. *International Journal of remote sensing,* 22 (9), pp. 1709–1724.

Smith, L. C., 1997. Satellite remote sensing of river inundation area, stage, and discharge: A review. *Hydrological Processes*, 11 (10), pp. 1427-1439.

Toyra, J., A. Pietroniro, L. W. Martz LW, et al., 2002. A multi-sensor approach to wetland flood monitoring. *Hydrological Processes*, 16 (8), pp. 1569-1581.

Toyra, J., A. Pietroniro, L. W. Martz, 2001. Multisensor hydrologic assessment of a freshwater wetland. *Remote Sensing of Environment,* 75 (2), pp. 162-173.

Vila da Silva, J. S., H. J. H. Kux, 1992. Thematic mapper and GIS data integration toevaluate the flooding dynamics within the Panatal, Mato Grosso do Sul State, Brazil, Proceedings, 1992. *Int. Geosci. Remote Sens. Symp.* (IGARSS '92), pp. 1478-1480.

Wang ,Y., Witten, I. H., 1997. Induction of model trees for predicting continuous classes. Poster papers of the 9th European Conference on Machine Learning, 1997.

Watson, J. P., 1991. A visual interpretation of a Landsat mosaic of the Okavango Delta and surrounding area. *Remote Sensing of Environ*ment 35, 1-9.

ACKNOWLEDGEMENTS

This study was supported by the GOES-R program through the GOES-R Algorithm Working Group. The manuscript contents are solely the opinions of the authors and do not constitute a statement of policy, decision, or position on behalf of NOAA or the U. S. Government.