Application of Bayesian Classification to Content-based Data Management

C. Lynnes, S. Berrick, A. Gopalan, X. Hua, S. Shen, P. Smith, K-Y. Yang
NASA Goddard Space Flight Center
Code 902, Greenbelt, MD, 20771
K. Wheeler, C. Curry
NASA Ames Research Center
MS 269-2, Moffett Field, CA 94035

Abstract- **The high volume of Earth Observing System data has proven to be challenging to manage for data centers and users alike. At the Goddard Earth Sciences Distributed Active Archive Center (GES DAAC), about 1 TB of new data are archived each day. Distribution to users is also about 1 TB/day. A substantial portion of this distribution is MODIS calibrated radiance data, which has a wide variety of uses. However, much of the data is not useful for a particular user's needs: for example, ocean color users typically need oceanic pixels that are free of cloud and sun-glint. The GES DAAC is using a simple Bayesian classification scheme to rapidly classify each pixel in the scene in order to support several experimental content-based data services for near-real-time MODIS calibrated radiance products (from Direct Readout stations). Content-based subsetting would allow distribution of, say, only clear pixels to the user if desired. Content-based subscriptions would distribute data to users only when they fit the user's usability criteria in their area of interest within the scene. Content-based cache management would retain more useful data on disk for easy online access. The classification may even be exploited in an automated quality assessment of the geolocation product. Though initially to be demonstrated at the GES DAAC, these techniques have applicability in other resource-limited environments, such as spaceborne data systems.**

I. Introduction

Data system operations today are generally automated without regard to the specific content of the data; that is, operational decisions do not in general adapt in response to the content of the science data being managed. This automation could be improved dramatically by incorporating decision-making based on the data content. We are employing machine learning algorithms in this decision support role to achieve intelligent data management of data from the Moderate Resolution Imaging Spectroradiometer (MODIS) at the Goddard Earth Sciences Distributed Active Archive Center (GES DAAC).

The calibrated radiance data from MODIS is used widely, in a variety of disciplines, but users often have difficulty in obtaining scenes that are usable for their particular purposes: the volume is large, and there is no efficient way to select data that are usable from a particular point of view (say, cloud-free for an oceanographer, or cloudy for an atmospheric scientist). Furthermore, only a certain amount of data can be retained on disk for direct access; it would be useful to retain the most interesting, usable data (Table 1).

	Pixel Characteristics					
		Clear-Sky				
Study Type	Cloudy	Ocean	Sunglint	Land	Snow/Ice	Fire
Cloud Properties	X					
Aerosols		X	(X)	X		X
Ocean Color		X				
Land Vegetation				X		
Snow Cover/Sea Ice					X	
Wildfires						X

Table 1. Usefulness of various pixel types for different studies.

The characterization data needed to support the above decisions is actually output by downstream science algorithms in the oceans, land or atmospheres processing chains. However, by the time the information in these downstream products is available (from 30 minutes to several days later), the data have often already been shipped, utilized and/or deleted from cache. The goal of using machine learning algorithms to estimate the output of these science algorithms is to reduce the processing time and computing requirements to a point where the estimations can be incorporated into the data stream, thereby allowing timely content decisions.

Machine learning algorithms such as neural networks and clustering have been used for decision support in business and policy domains. These techniques have found some use in remote sensing, e.g., for cloud and land cover classification. Yet most research on remote sensing data rests on science-based algorithms, such as those based on radiative transfer equations. Machine learning for scientific applications faces challenges such as discretization constraints, non-physical basis, and the difficulty of assembling training sets. However, these difficulties may be less significant in the decision support role. For instance, it is often enough to know whether a data attribute exceeds a certain threshold when selecting it for an application, without knowing the exact value. The difficulty of obtaining training data can be surmounted by using products output by the science-based algorithms. On the other hand, an advantage of machine learning algorithms for decision support is their speed once they have been trained. Data management decisions must be made while the "fresh" data are still on disk, and in time to service near-real-time applications, i.e., within minutes.

II. Naïve Bayes Classification

Our approach uses an automatic classifier to characterize the content of MODIS calibrated radiance data. The purpose is not to create a science product, but rather to enable more effective and efficient management of the product. For these purposes, we need a classifier that is simple, non-parametric (i.e., not strongly dependent on tuning parameters), and most of all, fast in terms of execution speed. This last criterion is important for two reasons: (1) the processing budget at the GES DAAC emphasizes the production of science data over data management activities and (2) data management decisions must usually be made shortly after the data are produced in order to be useful.

A Naïve Bayesian Classifier is used to characterize the data immediately after production to support data selection and caching decisions. The data are classified into several relatively coarse categories such as: cloud, oceanic sun-glint, shallow water, deep water, snow, sea ice, fire, land and desert. The classifier is trained against the output of the downstream science algorithms, such as cloudmask and ocean color.. While these do not represent the ground truth one would need for a science product, they serve as an effective proxy for decision support.

The Naïve Bayes Classifier derives from Bayes rule of conditional probability:

$$\mathrm{prob}(X|Y,I) = \mathrm{prob}(Y|X,I) \times \mathrm{prob}(X|I) / \mathrm{prob}(Y|I),$$

or the posterior probability is proportional to the product of the likelihood function and the prior probability. In supervised Bayes classification of numeric data (such as the calibrated radiance), we begin with a training set where each evidence vector **E** (i.e., the radiance values in each channel) has been assigned to a class C. Training consists of computing the probability density function for each combination E_i and C, i.e., Pr(E_i |C), and the overall probability for each class. The forward application of the classifier computes the probability for each possible class C as:

$$\mathrm{Pr}(C|E) = \mathbf{\Pi}\, \mathrm{Pr}(E_i\,|C) \times \mathrm{Pr}(C) / \mathrm{Pr}(\mathbf{E})$$

(In practical applications, the denominator is usually bypassed by computing a normalization over all the classes.)

The training of the algorithm consists of approximating the probability distribution functions through histograms of calibrated radiance (or reflectance) for different classes of pixels. The classification of a given pixel is derived from MODIS science products. For example, pixels classified as cloud are identified through comparison with the MODIS cloudmask product MOD35[1]. Conveniently, this product also includes identification snow/ice, of land vs. sea, and of desert areas. (These are not computed by the algorithm, but rather acquired from various ancillary files.) A glint flag is also present, but the determination of glint in the cloudmask product is somewhat broader than that used by the MODIS ocean color algorithms. Since the latter glint is more germane to the usefulness of data for ocean color applications, we used the glint from Level 2 (swath-based) ocean color products instead. Fig 1 shows the histograms for Band 1 radiance for the various classes.

Forward application of the algorithm to unknown pixels begins with looking up the probability of the observed radiance in the histogram for each class. For a given class, the probability is multiplied across all bands used in the classifier, then multiplied by the prior probability of that class. The class with the highest probability is selected.

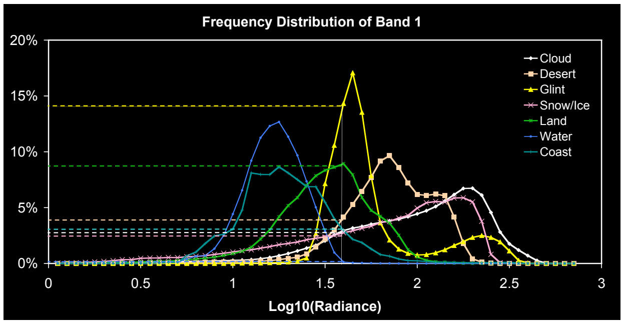

Fig 1. Frequency distribution of logarithmic calibrated radiance or Band 1 of MODIS in each class.

The prior probability is computed based on the season and geographic area (Fig 2). Twenty geographic regions were defined, based on climate distribution in latitude and either uniform or extreme distribution of categories in longitude, following Lydolph's climate distribution[2]. A set of prior probabilities was computed for these geographic regions using MODIS global gridded products for cloud, ocean-sunglint, shallow water, land/desert area and snow cover, together with the Near-real-time Snow and Ice Extent (NISE) product derived from SSMI. Different sets of priors were developed for Dec-Feb, Mar-May, Jun-Aug and Sep-Nov.

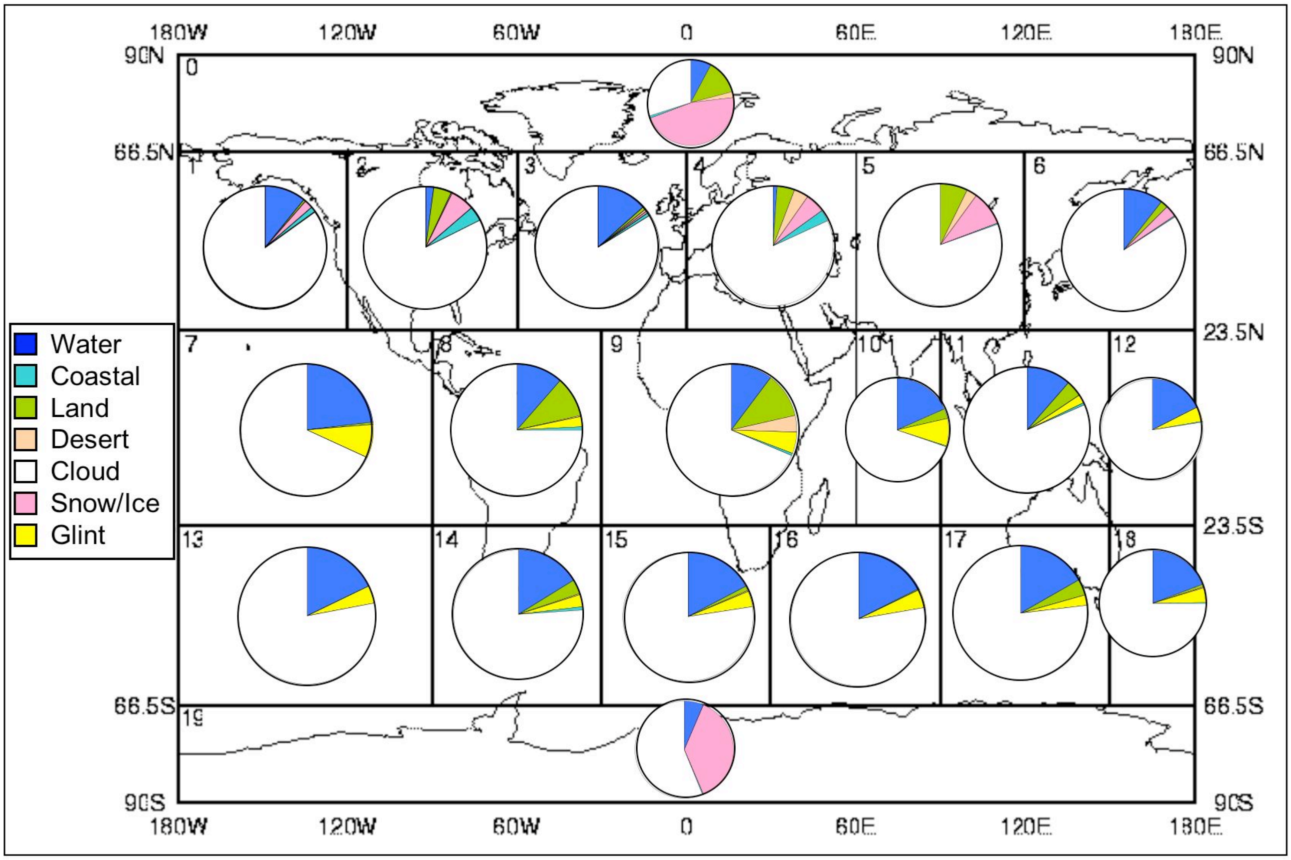

Fig 2. Regional variation of prior probabilities for December-February.

The resultant classifier is reasonably efficient at daytime classification in low to moderate latitudes. Fig 3 shows results for Terra/MODIS on 16 October 2003, 1620Z to 1625Z, covering the eastern half of North America. The leftmost picture is a true color image of the scene; the middle picture shows the classifications from the MODIS Cloudmask product; and the rightmost picture shows the classifications from the Bayesian classifier. Though there are some discrepancies, the correspondence is close enough to use for data management purposes. Just as importantly, the Bayesian classifier is extremely fast. Fig 4 shows algorithm timing results on a 250 MHz SGI Origin 2000 for the Bayesian classifier, using from 1 up to 8 bands, as well as the MODIS cloudmask science algorithm. Because the Bayesian classifier uses simple table lookups and multiplication, it consumes a negligible amount of processing resources. For data management, this speedup is the tradeoff for reduced accuracy relative to the science algorithm.

This Bayesian classifier also has the virtues of being conceptually simple and robust in practice, requiring little in the way of manual tuning. In addition, its output is not limited to a simple nominal classification, but includes a probability assigned to each class. This allows tuning to be implemented at the decision point, i.e., by adjusting the probability threshold needed to trigger a certain action.

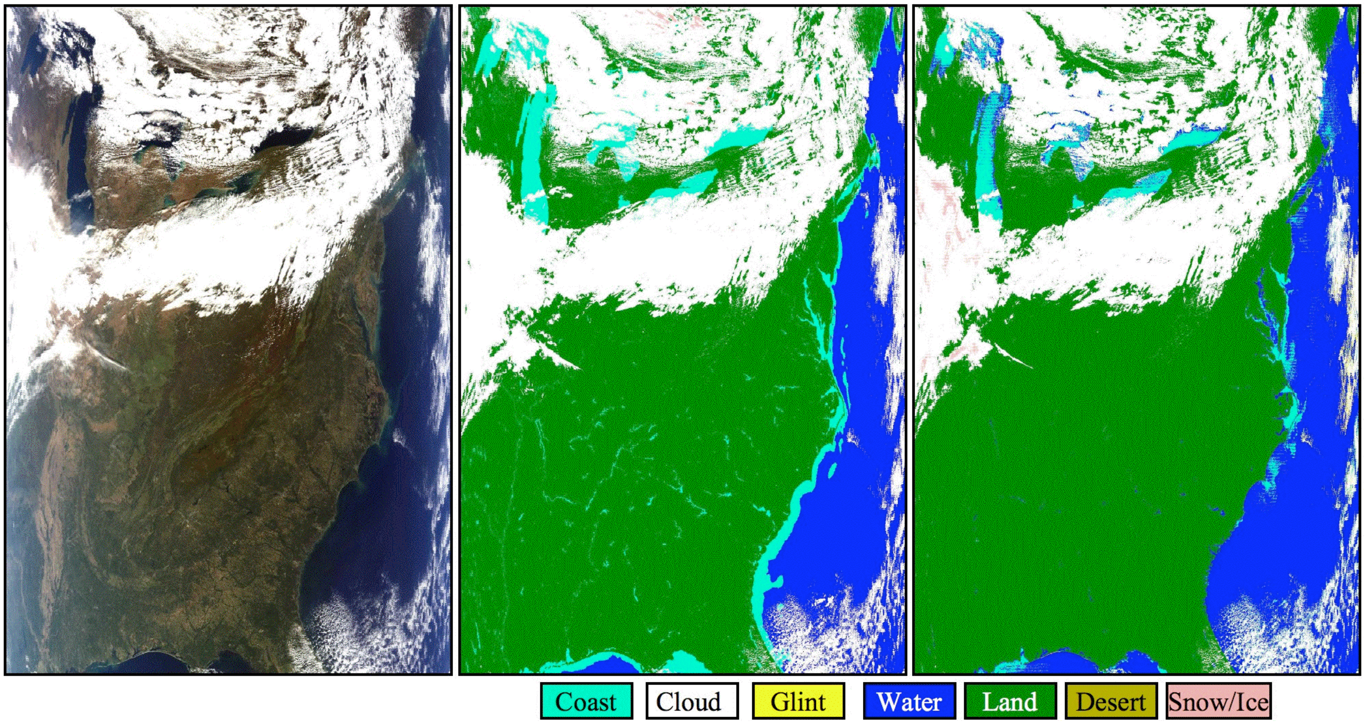

Fig 3. Terra/MODIS scene for 16:20Z-16:25Z, 16 October 2003. The left shows a true-color image, the middle is a classification derived from the MODIS cloudmask product, and the right shows the results of the Bayesian classification.

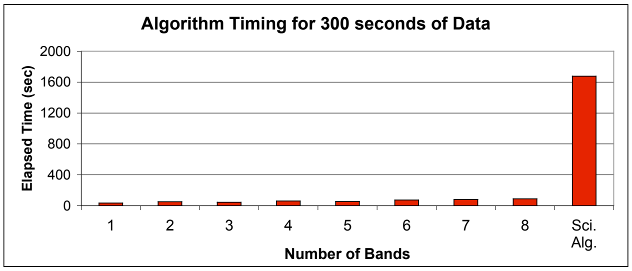

Fig 4. Timing results of Bayesian classifier using from 1 to 8 bands for a 300 second scene, alongside the MODIS cloudmask algorithm**.**

III. Content-Based Data Management

A primary goal of content-based data management is to make maximum use of available resources. These resources include storage resources, such as archive space and online cache, as well as throughput-related resources like network bandwidth. In addition, we also take into consideration the resources at the data user's end. Indeed, these resources are often more limited than those of the data provider. Fortunately, both sets of resources can be optimized using a similar criterion: the expected usefulness of the data to the end-user.

One straightforward application of this principle is to distribute only the pixels that meet certain usefulness criteria, rather than the whole scene. This content-based subsetting can be achieved for small numbers of pixels simply by extracting the individual pixels meeting the desired criteria. However, when the number of useful pixels is large, it is more efficient to mask out the unwanted pixels and then compress the data using a lossless technique such as Lempel-Ziv. In this case, we use the internal compression capabilities of the Hierarchical Data Format to construct a data file with exactly the same structure and properties, but which is much smaller than the original file. Fig 5 shows an example where the cloudy pixels have been masked out, leaving only the clear-sky pixels. Another element of selectivity based on usability addresses whether a user's area of interest within a scene contains usable data. Unfortunately, this kind of content-based selection is difficult to implement for ad hoc searches. Because we cannot predict all users' areas of interest, the content information for all pixels must be stored in a searchable catalog, a formidable technical challenge. On the other hand, if we limit content-based data selection to subscriptions, the problem becomes more tractable. Since we now examine each scene at the time of creation, the detailed content need not be stored or catalogued. Thus, a user might be able to specify: "send me data whenever the area around Lake Winnebago, Wisconsin is cloud-free". This would avoid transmissions of data where the overall cloud cover is generally low, but a given area is nevertheless covered by cloud (Fig 6).

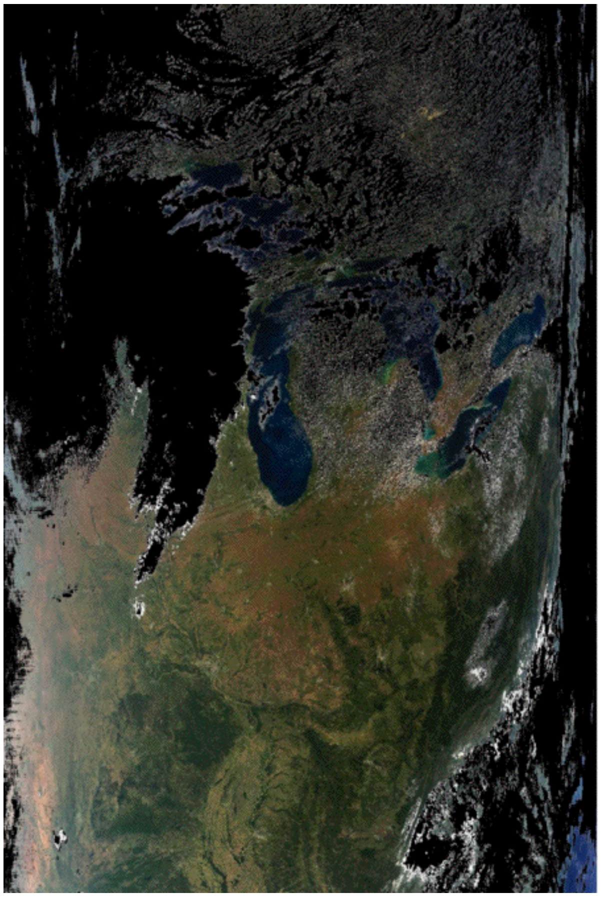

Fig 5. MODIS scene for 16:45Z, 23 September 2002, with cloud pixels masked out.

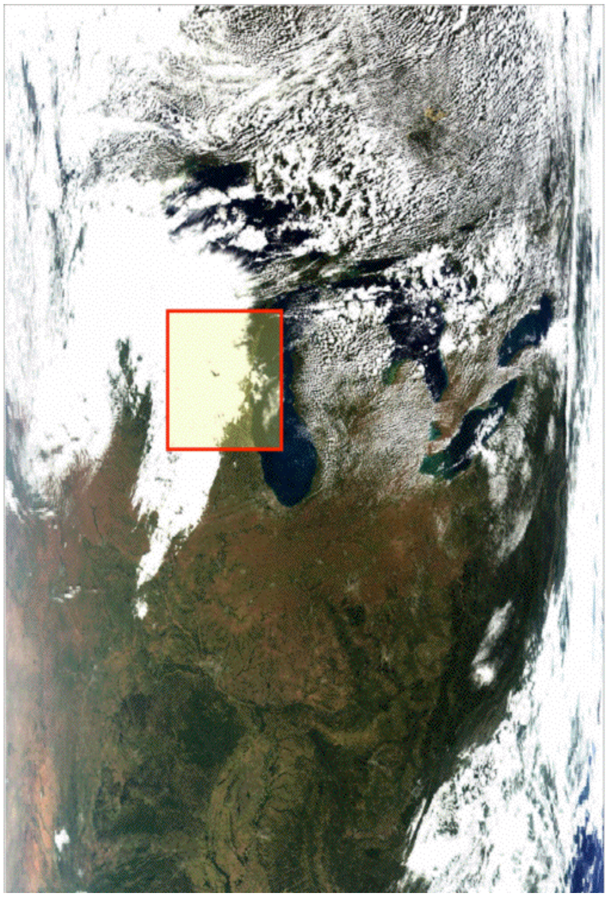

Fig 6. Conceptual example of content-based subsetting: "send me scenes only when the area around Lake Winnebago, Wisconsin is clear".

Another optimization that can be achieved is the management of precious cache resources. The GES DAAC maintains a 50 TB "Data Pool", an anonymous FTP area where data can be simply downloaded and on-the-fly services can be applied. While 50 TB seems large in 2004, it is yet a very small proportion of the GES DAAC's overall holdings of over a petabyte. As a result, even such popular products as the MODIS Level 1B may have a residence time in the Data Pool of only a few weeks, whereupon they are deleted to make room for new scenes. Accordingly, one of the aims of this project is to implement intelligent cache management in the Data Pool, purging data that are expected to be less useful to most users (in this case cloudy data for MODIS L1B), and retaining particularly useful scenes.

In addition to science and applications users, the downstream science processing algorithms in a data system constitute another important user group. For these "users", it may be pointless and wasteful to process certain data if the quality is poor. We have begun work on a proof-of-concept to automatically assess the quality of MODIS geolocation data, a key input for nearly every other MODIS product. Geolocation quality is typically assessed in an "offline" mode using control chips and island matching[4]. However, it would be useful to detect quality problems within the processing stream itself, so that they can be addressed quickly, without affecting the downstream processing. This is particularly the case for Direct Readout stations, which rely on a predicted ephemeris for processing data from the Aqua satellite, resulting in less accuracy than use of the later-arriving definitive ephemeris.

Our approach is to use the Bayesian classifier to identify land, water, cloud etc. We then use land-sea mask values for the same pixels, extracted from a digital elevation model using the latitudes and longitudes in the geolocation product. Thus, if the geolocations are incorrect in any kind of systematic fashion, the land/water pattern in the Bayesian classifier should show a systematic shift relative to that generated from the digital elevation model.

As it happens, a glitch in GES DAAC production of the Terra/MODIS Level 1 data provides an ideal test case. The data from June 19, 2002 were inadvertently processed using onboard attitude and ephemeris. Though these are generally preferred to the definitive attitude and ephemeris for accuracy reasons, the onboard attitude/ephemeris is unreliable in the general vicinity of spacecraft maneuvers, such as the drag make-up maneuver that occurred on June 19. In this case, significant errors of up to several kilometers were introduced into the geolocation product. (The error was actually first noticed by science researchers studying land cover change, and the data were subsequently reprocessed with the correct definitive attitude and ephemeris.)

Our procedure is as follows:

1. Classify the Level 1B data as to cloud, water, land, desert.
2. Extract a relatively cloud-free square from the scene.
3. Extract the land-sea mask values for the same square from the MODIS geolocation product
4. Convert the pixel classifications in each square into floating-point numbers by assigning land (or desert) the number +1.0 and water the number –1.0. Any other classification, such as ephemeral water in the land-sea mask or cloud in the classification, is assigned a random number, uniformly distributed in the interval (-1.0,+1.0).
5. Compute the cross-correlation of the two squares by applying a 2-D FFT to each, multiplying one by the complex conjugate of the other, and inversely transforming the result.

The procedure is illustrated in Fig 7 for part of the Terra/MODIS scene from 18:20Z-18:25Z, 19 June 2002. The cross-correlation in the bottom right indeed indicates a systematic shift, as does the mismatch plot in the bottom left. In order to automatically apply the technique, we still need a way of determining the significance of any shift detected in the cross-correlogram, but the technique appears to show some promise.

IV. Conclusion

The Bayesian classifier used in this study is far from sophisticated, yet it provides adequate results for the various data management uses to which we apply it. Indeed any similar classifier (such as neural nets) can serve the same purpose if only it is fast enough. After all, our goal is not so much the most accurate possible classification, but rather to utilize the classification to maximize the use of storage, throughput and processing resources within the GES DAAC. Furthermore, the applications are not limited to ground-based data archives. Content-based data management is useful anywhere that resources are limited, such as spaceborne data systems. The most dramatic impact is in actually tasking sensors based on the content of data from other sensors[5]. However, other limited resources may benefit as well, such as solid-state recorder space or instrument-to-ground communications.

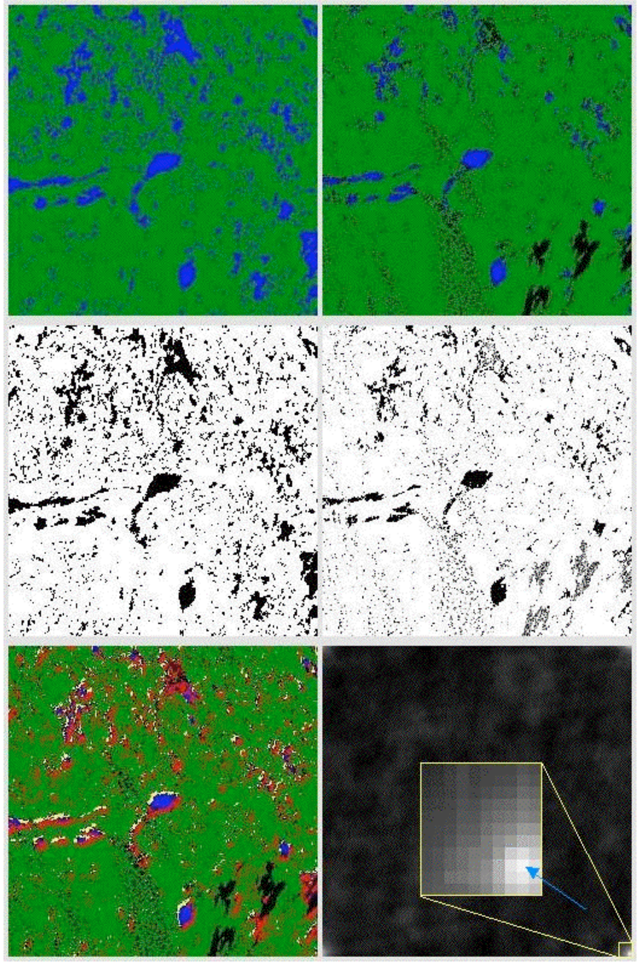

Fig 7. Automatic quality assessment of MODIS geolocation. Top: land/water/undetermined classification derived from geolocation product's land/sea mask (left) and Bayesian classification of calibrated radiance (right). Middle: conversion of land/water classifications into floating point. Bottom left: mismatch between land-sea mask and Bayesian classification of land and water. Red is water in land-sea mask and land in Bayesian classification; yellow is the reverse. Bottom right: cross-correlation of land-sea mask and Bayesian classification.

Acknowledgments

We thank NASA's Code R and T for funding the above work.

References

[1] MODIS Atmosphere: Cloud Mask Product, http://modis-atmos.gsfc.nasa.gov/MOD35_L2/.

[2] Lydolph, P.E. 1985. Weather and Climate. Rowman & Allanheld.

[3] Ackerman, S., W, Strabala, K., Menzel, P., Frey, R., Moeller, C., Gumley, L., Baum, B., Seeman, S., and Zhang, H., 2002: Discriminating Clear-Sky from Cloud with MODIS - Algorithm Theoretical Basis Document. ATBD Reference Number: ATBD-MOD-06.

[4] Nishihama, M, Wolfe, R., Solomon, D., Patt, F., Blanchette, J., Fleig, A. 1997. MODIS Level 1A Earth Location: Algorithm Theoretical Basis Document Version 3.0, http://modis.gsfc.nasa.gov/data/atbd/atbd_mod28_v3.pdf.

[5] Sohlberg, R., 2003. Use of NASA Earth Observing System data to monitor active fires and to develop SensorWeb decision support systems, 2nd International Wildland Fire Ecology and Fire Management Congress.