Virtual green band for GOES-R

Irina Gladkova[a], Fazlul Shahriar[a], Michael Grossberg[a],
George Bonev[a], Donald Hillger[b], and Steven Miller[c]

[a]NOAA-CREST, City College of New York
[b]NOAA/NESDIS/STAR/RAMMB
[c]CIRA, Colorado State University

ABSTRACT

The ABI on GOES-R will provide imagery in two narrow visible bands (red, blue), which is not sufficient to directly produce color (RGB) images. In this paper we present a method to estimate green band from a simulated ABI multi-spectral image. To address this problem we propose to use statistical learning to train and update functions that estimate the value for the 550 *nm* green channel using the values that will be present in other bands of the ABI as input parameters. One challenge is that in order to exploit as many bands as possible, we cannot use straightforward non-parametric methods such as a look-up tables because the number of entries in look-up tables grows exponentially with the number of input parameters. Other simple approaches such as simple linear regression on the multi-spectral input parameters will not produce satisfactory results due to the underlying non-linearity of the data. For instance, the relationship among different spectra for cloud footprints will be radically different from that of a desert surface. The approach we propose is to use piecewise multi-linear regression on the multi-spectral input to train the green channel predictor. Our predictor is built from the combination of a classifier followed by a multi-linear function. The classifier assigns each pixel to a class based on the array of values from the simulated (or proxy) ABI bands at that pixel. To each class is associated a set of coefficients for a multi-linear predictor for 550 *nm* green channel to be predicted. Thus, the parameters of the predictor consist of parameters of the classifier, as well as coefficients defining the approximating hyperplane for each class. To determine these classifiers we will use methods based on K-means clustering, as well as multi-variable piecewise linear approximation.

Keywords: GOES-R, ABI, green, color, regression, true color

1. INTRODUCTION

The growing wealth of remote sensing data from hundreds of space-based sensors is providing us with enormous new opportunities to better understand the earth at a time when that understanding may be critical. Yet designing, building, and launching these sensors entail enormous projects typically taking many years and costing billions of dollars. Unfortunately these missions are often delayed or canceled. Even when the sensors are deployed they may not provide data for all of the earth, or they may not measure the precise spectral bands of interest, or they may not visit sites of interest frequently enough. Even when remote sensing projects are successful, an instrument providing data may fail, or lack funding for continuing operation and thus be retired.

Fortunately, growing computational power and statistical learning techniques provide a means to leverage the hundreds of satellites already observing the earth, which produce multi-, hyper- or ultra-spectral images. Statistical analysis and machine learning techniques can be used to produce virtual integrated sensors. These integrated sensors can help address some of the issues that cannot be addressed with hardware sensors alone.

For example, current GOES is providing panchromatic visible imagery every 15 (or 30) minutes, while the future ABI on GOES-R will improve this to at least every 5 minutes (over the CONUS) and will provide 2 visible color components, but will still not provide true color imagery at that sampling rate. As a result, GOES-R will not produce color images despite the widespread demand for such images for use as decision aids by meteorologists

Further author information: (Send correspondence to I.G.)
I.G.: E-mail: gladkova@cs.ccny.cuny.edu

and for visualization by the public. To address this problem we propose to use non-parametric methods such as nearest neighbor regression, kernel methods, and cluster analysis to capture these complex relationships.

In this paper we present a method to estimate 550 nm green band from a simulated ABI constructed from similar MODIS bands as a proxy. The algorithm produces a function that estimates required color bands from the GOES-R visible and near-visible bands. The algorithm produces this function by training on data where both proxies for GOES-R and green band are available. This makes it possible to build a statistical predictor to determine the green values for GOES-R. Hence the currently planned GOES-R data acquisition is leveraged by this method to create a new virtual green band data acquisition product. This is important because natural color images are easier for both meteorologists and the public to interpret, providing better decision aids and visualizations for GOES-R visible imagery.

This paper grew out of an attempt to refine the methods for the synthesis of the green spectral data from other spectral data, as developed in.[2,3] To briefly review the techniques developed in,[2,3] an extensive collection of observational and modeled spectral data, including green, was analyzed, and the green value for each such data point was tabulated as a function of the remaining variables, which in[2,3] consisted of values for 470 nm blue band (B), 640 nm red band (R), and 860 nm near infra-red (NIR) band. In the prior work, a look up table was created for this data, which was used, in the absence of green data, to approximately predict the likely green value, based on the known values of the other variables. In the prior work, there were several variations used to produce the look-up table. In the present paper, we propose a refinement and improvement over building a lookup table on three bands to predicting green values. Our improvements involve refinements in the following three areas:

1. Our approximating function for green will depend on five spectral parameters as opposed to the three that were employed in[2,3]
2. We replace the piecewise constant green approximation by a piecewise linear approximation over the Voronoi cells associated with our sampling points.
3. We will employ variably spaced sampling points, as opposed to the lattice-spaced sampling points used in[2,3]

The first area is based on the idea that there may be different valued green pixels which have similar associated R, G and NIR triplets, but can be disambiguated by the values in other bands. Because we use a larger input space a fully non-parametric lookup table becomes infeasible, but we argue that a locally parametric method, that of piecewise multi-linear function, is able to maintain enough flexibility to handle nonlinear variations in the surface but still able to cope with the relative data sparsity due to the larger input space. The third point makes it possible for the predictor function to change rapidly in parts of the input space where there is more variation, while efficiently remaining simple in parts of the input space where a multi-linear prediction works well.

2. BACKGROUND

The issue of generating a synthetic green band has been investigated by both CIMSS and CIRA using a look-up table (LUT) method.[2,3] The starting point is a large collection of data gathered from remote sensing data and generated from models, which we refer to as "training" data because the LUT is trained using this data, before it is used. Two algorithms presented in [2] and[3] differ in how the LUT was trained, either with raw or Rayleigh-corrected reflectance values. Both LUT use values for the 470 nm blue band, 640 nm red band, and the 860 nm near infra-red band as inputs and predicts 550 nm green band value. The reflectance value in each B,R, and NIR are quantized into one of N evenly spaced bins per band. The number stored in that entry is the average reflectance value of G restricted to those pixels in the training data whose B,R, and NIR values fall into the range determined by that entry in the 3D data-tables.

Figure 1 shows how the LUT is used (once built) for the algorithm presented in.[2] In this variation, $N = 200$. The predicted green value reflectance along with reflectances in R and B bands are then Rayleigh and solar zenith corrected then stretched using a log function to enhance contrast.

Figure 2 shows a second variation of the algorithm[3] where the Rayleigh and solar zenith corrections are applied to B, R, and NIR before being quantized to find the entry into the 3D LUT. In this case, $N = 250$ quantization levels, and the G values averaged and stored in that entry of the LUT are also Rayleigh and solar zenith corrected. As a further refinement, the images can be first segmented by ground surface type. A separate LUT can then be built for pixels in that surface type. Since the surface type is assumed known, the selection of LUT is done through geo-location pre-process. A typical result of applying these algorithms to generate a true color RGB image such as is shown in Figure 3. A comprehensive analysis on which of the LUT algorithms has better performance has not been completed. There are indications that the results depend on the image. As a representative of all of these methods we compare only to the first algorithm outlined in Figure 1.

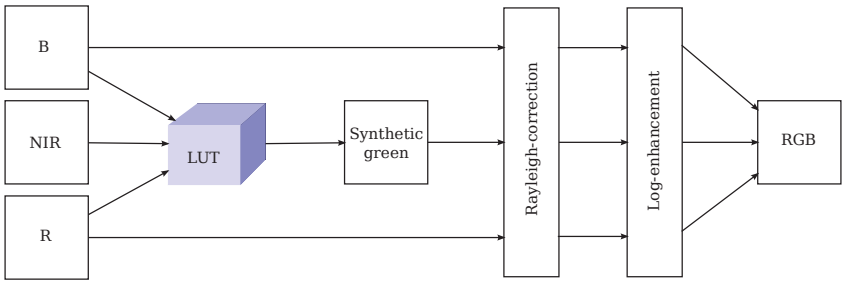

Figure 1. Block diagram for generating true-color/RGB images, with application of Green-LUT, followed by Rayleigh-correction, followed by log-enhancement.

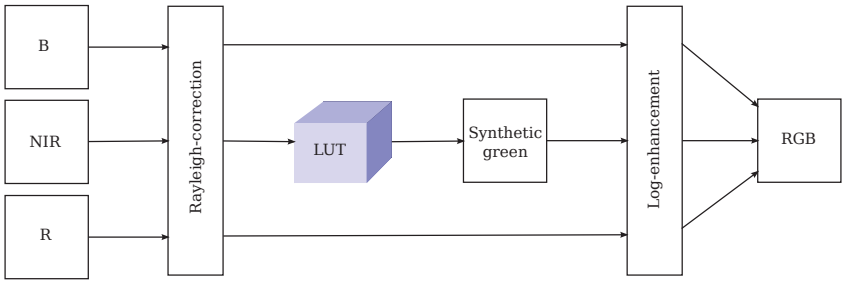

Figure 2. Block diagram for generating true-color/RGB images using Rayleigh-corrected reflectances prior consulting the table

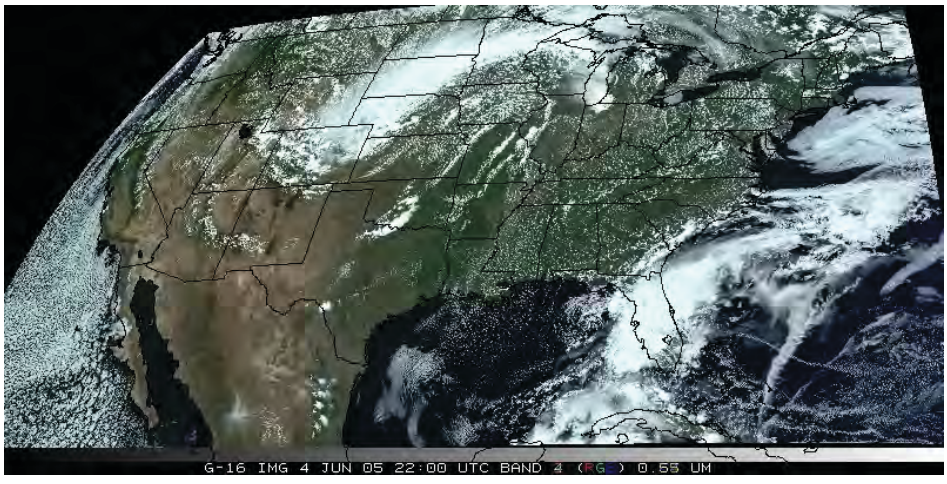

Figure 3. True color image simulated at CIMSS .

3. APPROACH

In this section we outline an approach to improving the results obtained with a lookup table. The approach we propose can be understood as three component ideas though which to improve the green prediction function: increasing input dimension, partitioning input space into homogeneous clusters, and determining a piecewise multi-linear prediction function.

Dimension: Figure 4 shows the mutual information, measured in bits, of bands 1, 2, 3, 6 and 7 with green band 4. MODIS band 5 was not considered because it does not have a direct mapping to the ABI. Clearly bands 1 (Red) and bands 3 (Blue) show the most significant mutual information with tha band 4 (Green). In addition band 2 (NIR) clearly does have some mutual information with band 4 but so do bands 6 and 7 (also NIR). Though a multivariate mutual information would be needed to separate how informative bands 6 and 7 are about band 4 independent of band 2, this at least suggests band 6 and 7 are at least as informative about band 4 as is band 2. In fact, our experiments have convincingly shown that there is more information available in five spectral channels that can improve the prediction of green than available in the three used in in.[2,3] From the point of view of building a prediction function, we can interpret this as the extra information disambiguating the prediction of green where three channels are unable to do so.

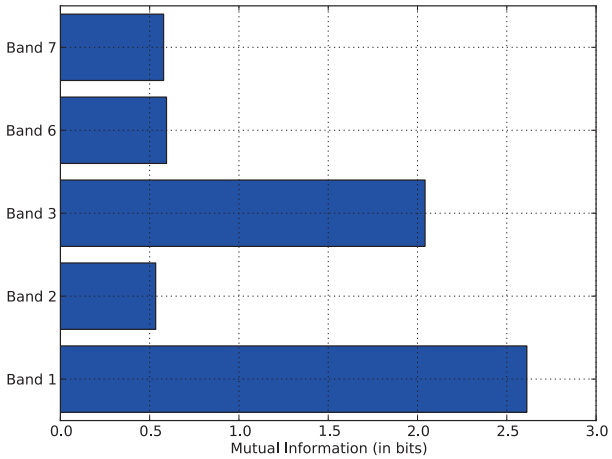

Figure 4. Mutual information of bands 1, 2, 3, 6 and 7 with green band 4 in bits.

One challenge in using more channels is that LUT become increasingly problematic with increasing the number of input dimensions. To see this assume we have d input dimension, in our case $d = 5$ vs. $d = 3$. If there we use N bins for quantization in each dimension, then the total number of entries in the lookup table is N^d. Thus the size of the LUT table is exponential in the number of input parameters. So adding two more channels assuming $N = 200$ increases the size of the LUT by a factor of $40,000$. This is not only a problem of increased space needed for memory or computational complexity. Ultimately the most significant impact is on data. For the entries of the LUT to be filled we need data with multiple instances for each bin. Again, in the case of an increase by 2 dimensions with $N = 200$ this means $40,000$ times as much data is needed.

Partitioning: In,[2,3] the regularly spaced partitioning points give rise to cubical cells within which the green prediction function is approximated by a constant. As noted above, the number of regular cells required is exponential in the number of input dimensions. However, if we create regions based on Voronoi cells (cf. Figure 5) around selected sample points in the input space where the relationship of the input values to green is relatively homogeneous, we can dramatically reduce the number of pieces, and thus the number of parameters. Again this both helps in the amount of data that needs be stored, and the amount of data needed to train the predictor function. The association of a region with an input value based on the Voronoi cell in which the input cell is located is a equivalent to a nearest neighbor classifier.

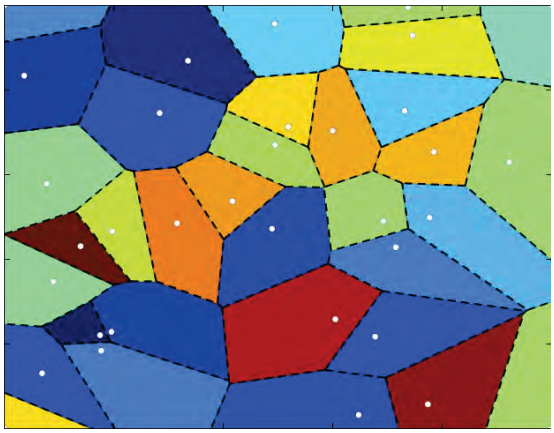

Figure 5. Voronoi regions.

Approximating functions: While an LUT is a piecewise constant function, by moving to a piecewise linear function we can reduce the number of pieces without reducing accuracy. By reducing the number of pieces we effectively reduce the number of parameters, using less memory, and requiring less data.

In summary, our predictor of the green value will consist of a classifier, which will assign an ABI pixel to a class based on the array of values from all the other bands at that pixel. Each class then has an associated normal vector that defines a hyperplane for a multi-variable piecewise linear predictor. Thus, the parameters of the predictor consist of parameters of the classifier, as well as a multidimensional vector, defining the approximating function for each class. We will use the training data to fit a classifier and then fit a hyperplane for each class. To determine these classifiers we will use methods based on K-means clustering, as well as standard least squares approximation. The resulting algorithm for computing the green values form the ABI measurements can now be quickly described as follows. When new data (a collection of points in six-dimensional space) is received, these points are input, respectively into the appropriate regions, and the corresponding piecewise linear function is computed, which yields the green value.

The general outline of the approach we propose is presented in the diagram shown in figure 6. The proposed approach falls into three components. The first component is obtaining and building ground truth training and testing data sets. The second is development of an effective parameterized predictor, along with an efficient method of estimating the predictor parameters. Finally, the third component is the development of our testing and evaluation of the predictor.

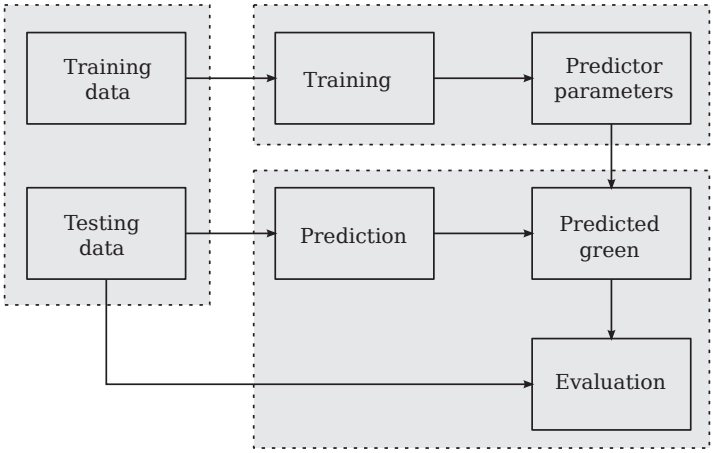

Figure 6. Block diagram of the algorithm.

The first component, data preparation, starts with selection of the MODIS bands 1, 2, 3, 6, and 7 corresponding to the ABI 470 nm, 640 nm, 865 nm, 1.64 μm, and 2.25 μm center wavelengths, followed by two preprocessing steps for each band. We start processing the data by attempting to fix out-of-valid range values in each of the considered MODIS bands. We preprocess out-of-range values using an adaptive mean value filter, which replaces isolated missing pixels with the mean value of the valid pixels in a window with an adaptive size. The adaptive window size is the minimum size such that the majority of the pixels in the window are within valid range. Note that the window is limited to a fixed maximum size. The file is abandoned if more than half of the pixels in the input bands are bad.

The next step when working with the data is destriping the radiances. As observed in Rakwatin et al.,[8] destriping can significantly improve regression. In theory, an image of properly calibrated radiances should not have stripes, nevertheless some striping artifacts remain and can be removed using histogram specification as is commonly done,[6].[7] We apply destriping to all bands, but if a band's detectors are well matched so that destriping is not required then the algorithm essentially returns the unstriped values. We will maintain an independent testing data set for evaluation.

The third component, prediction and evaluation, can be described very simply, as has been done above. We will therefore concern ourselves with a more detailed description of the second component, which is training. We begin with six preprocessed MODIS bands listed above, which can be thought of as a collection of points in six dimensional space, one of which represents green value. We regard the remaining five dimensions as the domain space and the green dimension as standing in a functional relationship to the other five. The point of the training component algorithm is to formulate an accurate model of the functional relationship, which can then be used by the predictive algorithm. The process we use to create our predictive green value is presented in the diagram shown in Figure 7.

1 For each training granule, we randomly select N points in the domain space $\mathbf{R}^5$. We use these random points to initialize the K-means clustering algorithm which we run on the input data to determine a set of clusters for a single granule.

2 The union of all the cluster centers for all granules is combined, and the aggregated list of centers is used to build Voronoi domains.

3 For each such Voronoi domain we find a least squares best linear approximation for the green value expressed as a multi-linear function of the remaining five parameters.

The prediction function, following this method, takes a tuple with values from bands 1, 2, 3, 6, and 7 as input. The first step is to use a nearest classifier to assign this pixel to a Voronoi domain by finding the closest cluster center. Since each center has an associated multi-linear function, this function is applied to the tuple of bands to predict band 4 (green). Again, for the ABI the corresponding bands would be used.

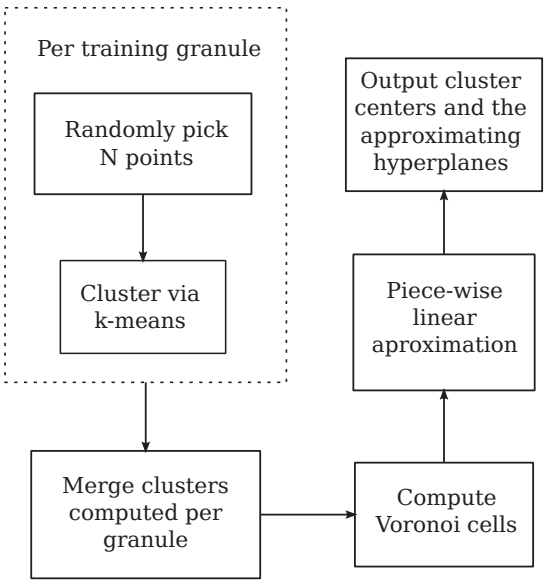

Figure 7. Block diagram of the training algorithm.

4. DISCUSSION

In this paper we present a preliminary algorithm implementing the approach outlined in the previous section. To explore how the algorithm is impacted by changing the number of Voronoi domains, we first set the per-granule number of clusters to $N = 50$ training using 40 training granules, resulting in $2,000$ clusters total. When the clusters were aggregated, very small clusters were eliminated. For illustrative purposes we picked a test granule (not part of the training granules) which illustrates smoke-over-vegetation case in the midwestern U.S., collected by Terra MODIS on 28 June 2002 at 1640 UTC. The portions where the relative error (the absolute error divided by original value) is over 15% are mapped in the figure 8. We can see in this figure that the largest errors occur on the boundaries. This is likely because on the boundaries, more Voronoi cells are needed in the partition in order to handle pixels of mixed type. If we increase the per-granule number of initial clusters to $N = 200$ the result improves dramatically, as seen in figure 9. Further increases in the number of clusters used in training do not seem to result in improvements in the testing, but perhaps increasing the number of training granules could.

The results of partitioning into Voronoi cells, followed by the multi-linear regression appears to compare quite favorably with results obtained by LUT.[2] Figure 10 shows the relative error for the first variation of the LUT described in section 2. When $N = 200$ is used in per-granule partitioning, the resulting relative error from the piecewise multi-linear prediction is shown in figure 11. The error in the upper left portion and over the lake in the middle improves dramatically. The reflectance values over the lake regions are relatively low. If we consider the scatter plot of predicted band 4 value to actual band for value using the LUT method, as shown in figure 12, the errors over the lakes result in a wider spread in the lower left of the figure. The corresponding scatter using the piecewise multi-linear method has a smaller spread in the lower left, as shown in figure 13. The figure 14

shows the RGB using the original green band 4. Visually this is nearly identical to the result obtained, shown in figure 15, using the green channel predicted from the piecewise multi-linear method.

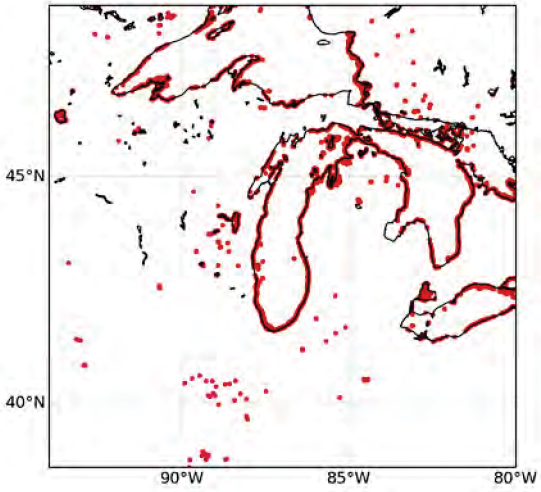

Figure 8. Relative error with $N = 50$ above 0.15 threshold on the map.

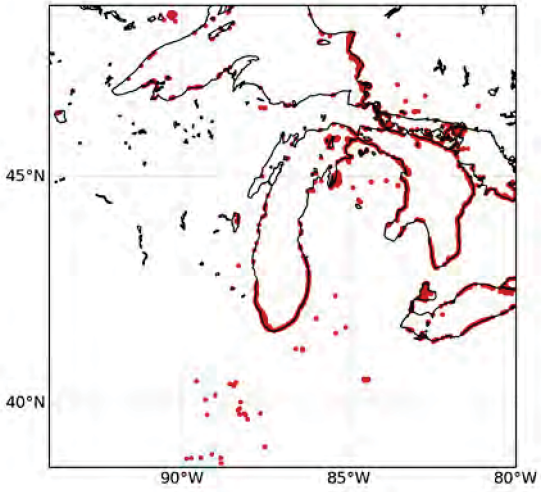

Figure 9. Relative error with $N = 200$ above 0.15 threshold on the map.

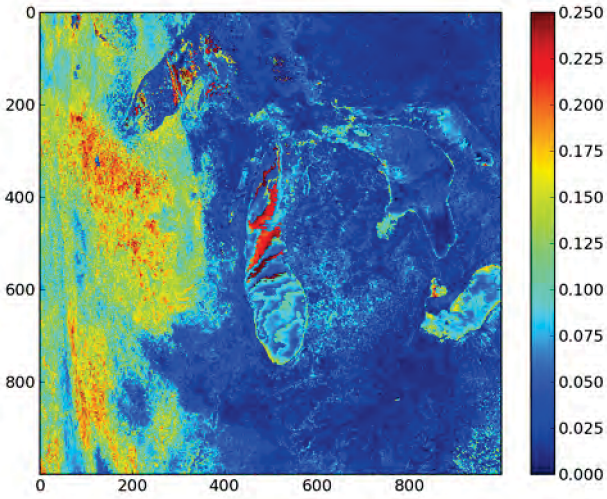

Figure 10. Relative error using LUT as in^2 .

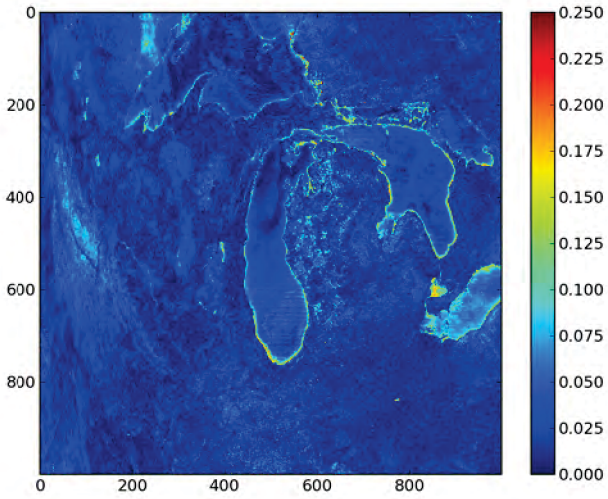

Figure 11. Relative error using multi-linear piecewise approximation.

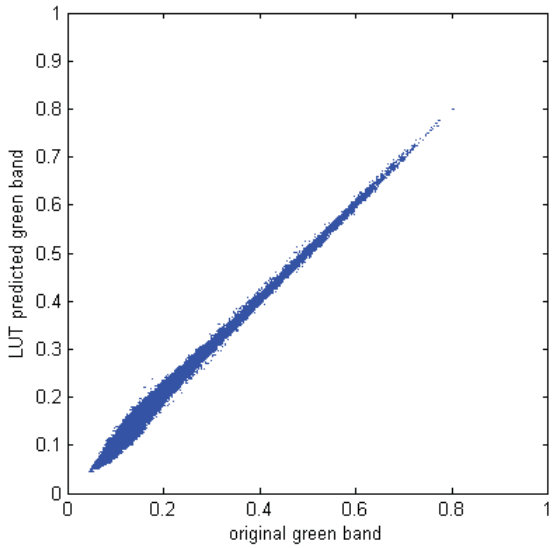

Figure 12. Scatter plot of the LUT restored green band as in^2 versus original green band.

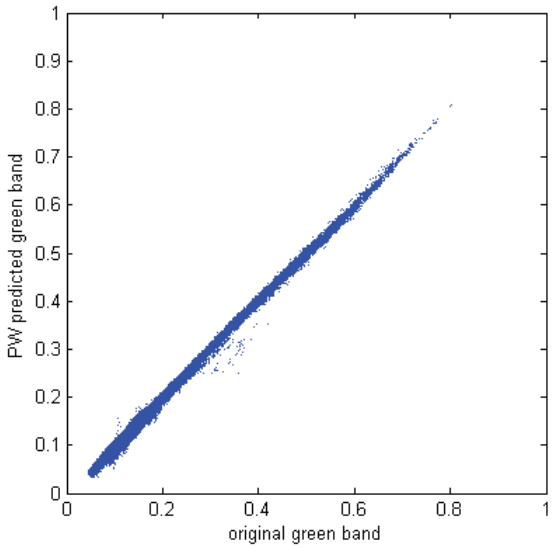

Figure 13. Scatter plot of the piece wise approximation versus original green band.

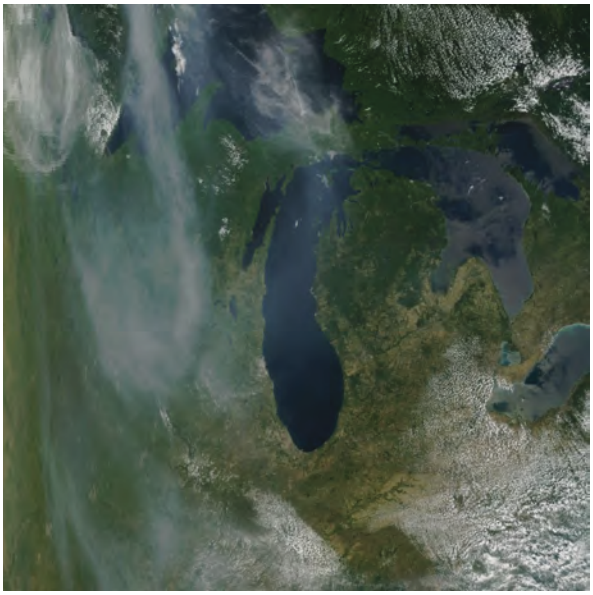

Figure 14. Natural color image using original 550 nm MODIS band.

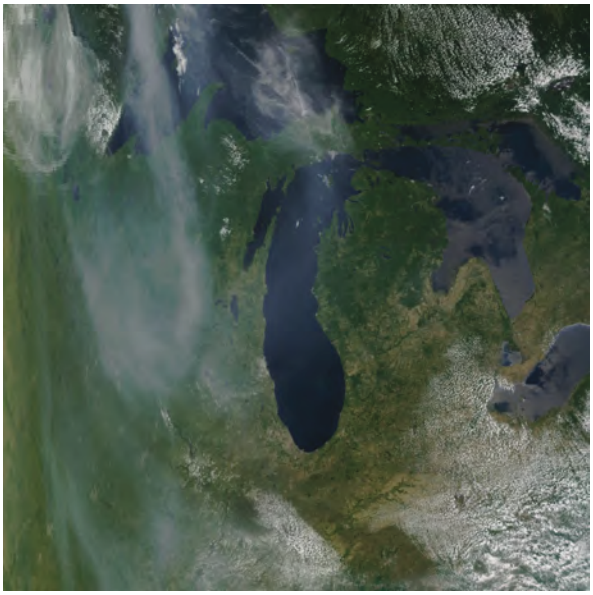

Figure 15. Natural color using predicted green band.

5. CONCLUSION

We presented an approach to the problem of estimating a green band from GOES-R using piecewise multi-linear predictor on five input bands to improve upon previous a LUT-based approach which uses a Red, Blue and single NIR band. We show that there are two additional NIR bands which have comparable mutual information to the NIR used previously to build the LUT. We outline a new approach which uses these extra bands to predict a green band. To overcome the explosion in LUT entries with increasing dimension due to more bands, we use a partition of the input space followed by a multi-linear regression. We have presented preliminary results that show that this approach can significantly reduce the error and produce a result that is visually indistinguishable from the true color RGB image with a green band. In future work we will make the partition adaptive to

further reduce the error by introducing more partitions in those regions where the error is highest. We will also systematically compare with the multiple variations on the LUT approach. In addition, we will examine whether by combining our method with the LUT approach which classifies by surface type, can further improve the prediction of the green band.

REFERENCES

[1] Grossberg, M., Shahriar, F., Gladkova, I., Alabi, P., D. Hillger, D. and Miller, S., "Estimating True Color Imagery for GOES-R," Proc. SPIE 8048(50), (2011).

[2] Hillger, D., Grasso, L., Miller, S.D., Brummer, R., and DeMaria, R., "Synthetic advanced baseline imager true-color imagery," J. Appl. Remote Sens. 5, in press (2011).

[3] Miller, S.D., Schmidt, C., Schmit, T.J. and Hillger, D.W., "A Case for Natural Color Imagery from Geostationary Satellites and an Approximation for the GOES-R ABI," Int. J. of Remote Sens., (in press) (2011)

[4] Schmit, T.J., Gunshor, M. M., Menzel, W. P., Li, J., Bachmeier, S., Gurka, J. J., "Introducing the Next-generation Advanced Baseline Imager (ABI) on GOES-R," Bull. Amer. Meteor. Soc., 8, 1079-1096 (2005)

[5] Grasso, L. D., Sengupta, M., Dostalek, J. F., Brummer, R. and DeMaria, M., "Synthetic satellite imagery for current and future environmental satellites," Int. J. of Remote Sens., 29(15), 4373-4384 (2008)

[6] Gumley, L., Frey, R. and Moeller, C., "Destriping of MODIS L1B 1KM Data for Collection 5 Atmosphere Algorithms," http://modis.gsfc.nasa.gov/sci_team/meetings/200503/poster.php (2005)

[7] Cheng, J. S., Shao, Y., Guo, H. D., Wang, W. and Zhu, B., "Destriping MODIS data by power filtering," IEEE Trans. on Geoscience and Remote Sens., 49, 2119-2124 (2003).

[8] Rakwatin, P.,Takeuchi, W. and Yasuoka, Y., "Restoration of Aqua MODIS Band 6 Using Histogram Matching and Local Least Squares Fitting," IEEE Trans. on Geoscience and Remote Sens., 47(2), 613-627 (2009).