ISyE 6416 – Computational Statistics - Spring 2016
Project Final Report
Submitted: April 26th 2016

Team Member Names
Courtney Di Vittorio & Robert Woessner

Project Title
Multi-temporal Classification of a Seasonally Flooded Wetland Using Optical Remote Sensing (Satellite) Derived Indices

Problem Statement:

The Sudd Wetland in South Sudan is a vital component of the Nile River Basin. During the dry season (Dec – Mar), the permanent wetland area is estimated to be 15,000 km^2. During the height of the rainy season in September, releases from Lake Victoria upstream augment the flows through the White Nile. As the high flows enter the Sudd, water spills into the floodplains and expands outward. As a result of this combination of heavy rainfall and high flows, the flooded area of the wetland is believed to double in size, to approximately 30,000 km^2. Due to the high evapotranspiration rates in this semi-arid environment, only half of the water that enters the Sudd exits the wetland at the downstream end. Figure 1 shows the main channels of the Sudd and the estimated extents of seasonal flooding. This figure was published by Sutcliffe and Parks to illustrate a simple mass balance hydrologic model of the wetland. The volume of water within the Sudd could be determined by the difference between measured outflows at Malakal and inflows at Juba (1987).

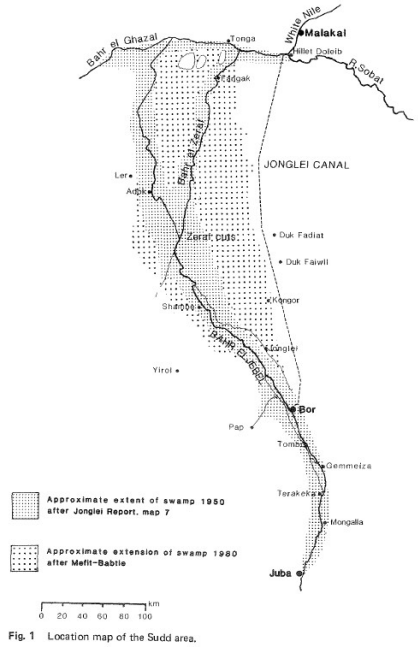


Fig. 1 Location map of the Sudd area.

Figure 1: Hydrologic Model (Sutcliffe & Parks, 1987)

The desert countries downstream of the Sudd (Sudan and Egypt) would like to channelize the wetland and convey water downstream before it overflows and evaporates. However, there are nomadic communities who live within the Sudd and rely on the seasonal flooding cycle to regenerate the grasslands which feed their cattle. In addition, flow alterations to the Sudd would disrupt the sensitive ecosystem and the services it provides. Therefore, prior to diverting water from the wetland and conveying it downstream, decision makers should consider the detrimental impacts such alterations could have on the local people and environment (Howell, Lock, & Cobb, 1988).

In order to understand and quantify these impacts, a better understanding of the Sudd hydrology is needed. A key input required for the calibration of a hydrologic model of the Sudd is flooded area, which changes with time. In the past, areal extents of flooding were roughly estimated from only a few aerial images; ground data is not available given the scale of the wetland and the country's state of political instability. However, opportunities now exist to more accurately determine this parameter using remote sensing data from earth observing satellites. For our project, we have applied classification techniques introduced in the course to this satellite imagery in an attempt to distinguish between open water, permanently flooded vegetation, seasonally flooded vegetation, and upland vegetation. The extent of flooded area could then easily be extracted from these classified images and applied to a hydrologic model of the Sudd.

Data Source

The data used in this classification is derived from multi-temporal imagery from NASA's Terra Satellite which carries the Moderate Resolution Imaging Spectroradiometer (MODIS). From this sensor, the 500m spatial resolution 8 day composite land surface reflectance product (MOD09A1) was downloaded. Each image from this product contains reflectance values from 7 bands (wavelengths), which have been atmospherically corrected (Vermote, 2015).

The Terra satellite re-visits the Sudd on a near daily basis, but due to frequent cloud cover a more complete image can be obtained by compiling the 'best' pixels over an 8 day period. However, even the 8 day composite product contains missing data from cloud cover. Each downloaded image includes a quality assurance layer which indicates whether clouds were detected or if there were any other quality issues (NASA LP DAAC, 2012). Using this data layer, pixels flagged as poor quality were converted to 'NaN' in Matlab. However, to make the data compatible with classification algorithms, the missing values were linearly interpolated temporally, using the previous and following image for each missing value. Terra MODIS data is available from Feb 2000 – present, but we decided to apply the classification methods to the year 2008 only. Based on historic river flow data of the White Nile, 2008 was a typical year without excessive or limited flooded; therefore, we believe 2008 is ideal for comparing classification techniques for the Sudd.

From the 7 spectral bands, various remote sensing indices can be derived. Because we are interested in mapping flooded area, we decided to use the Normalized Difference Vegetation Index (NDVI) which is a combination of the near infrared (NIR) and red bands and measures the amount of vegetation, and the Normalized Difference Wetness Index (NDWI) which is a combination of the green and short wave infrared (SWIR) bands and measures the amount of moisture (Xu, 2006). The equations used to derive these indices are as follows:

$$NDVI = \frac{NIR - RED}{NIR + RED} \qquad NDWI = \frac{GREEN - SWIR}{GREEN + SWIR}$$

A major challenge in detecting flooded areas of wetlands from satellite imagery is that they are often covered in dense vegetation, and it can be difficult to distinguish between vegetation where the soil is wet from rainfall and vegetation where the soil is flooded. However, we hypothesized that we could separate these two classes by using

a combination of the NDVI and NDWI indices and looking at their temporal trajectories versus individual values in time. A total of 46 NDVI and NDWI images were derived for the year 2008, and each image is a 1228x669 pixel grid. Within the grid, an asymmetrical area centered on the wetland was chosen to be considered in the classification; this subset area contains 671,057 pixels. The 46 8-day composite images were averaged by month and reduced to a set of 12 monthly images for both the NDVI and NDWI.

To provide an alternative data source instead of the 24 NDVI and NDWI images, we used principal component analysis (PCA) to reduce the dimensionality of the data. By vectorizing the 24 images, stacking them in time or space, and performing a spectral decomposition of the covariance matrices, we can reduce the full time series into a few images. We derived the PC's in 2 different 'modes': T-mode and S-mode. For T-mode, a matrix was constructed, denoted by X, where each row represents a pixel and each column represents an image. The values in the covariance matrix, calculated by $(X-u)(X-u)^T$, where u is the mean of each row, represent the magnitude of the correlation between temporal trajectories of the pixels. Therefore, the eigenvectors derived from the covariance matrix can be reconstructed into images, where pixels with similar values have similar NDWI and NDVI temporal trajectories. The principal components, or loadings, from the spectral decomposition indicate which months the corresponding images are most applicable for explaining a large amount of variance of the data. For S-mode, a matrix is constructed where each row represents an image and each column represents a pixel. The loadings from the spectral decomposition of the covariance matrix can be reconstructed into images, where pixels with similar values have similar NDWI and NDVI values in space. The corresponding eigenvectors indicate which months the images are most applicable (Wilks, 2011).

The T-mode and S-mode PCA was performed for each index's individual time series (12 images each), as well as the combined indices (24 images). The first 3 eigenvectors and loadings for each iteration were plotted and analyzed. The results from the T-mode analysis of the combined indices are shown in Figure 2 and the remaining images and loadings along with the amount of variance explained for each image are shown in Appendix A. In Figure 2, the first eigenvector highlights the permanently flooded vegetation, the second highlights the permanent open water bodies, and the third is much noisier and more difficult to interpret. By analyzing the results for each mode and each index, we decided to test our classification algorithm on two sets of PCA images. The first set (from the individual index analyses) consists of: the first T-mode component of both the NDWI and NDVI, the second S-mode component of the NDWI, and the third S-mode component of the NDVI. The second set (from the combined indices analyses) consists of: the first two T-mode components and the first S-mode components.

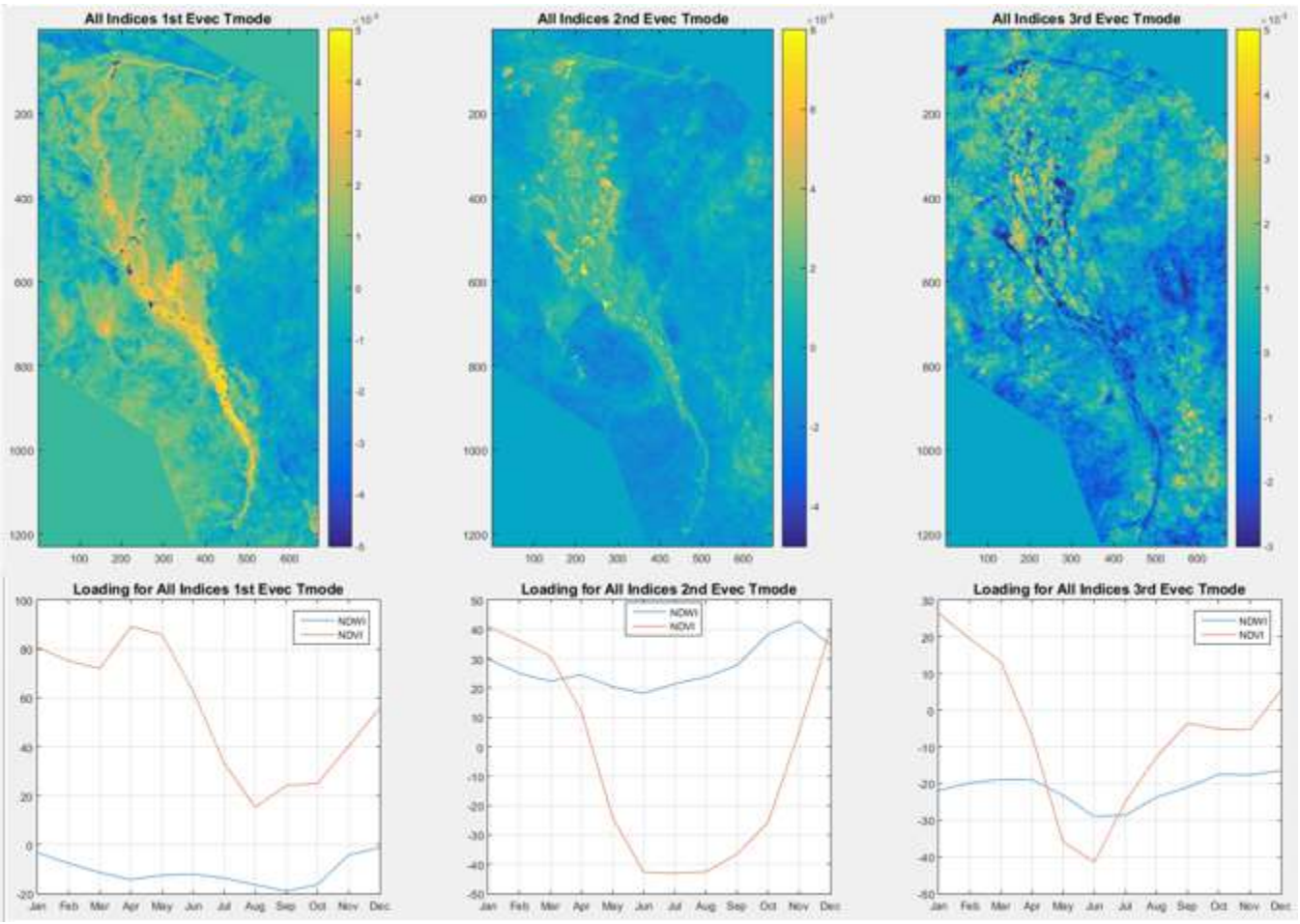


Figure 2: T-mode PCA Results from Combined Indices

In order to train the classification algorithms and measure their accuracy we needed a set of ground truth data. We accomplished this using Google Earth imagery to identify geographic coordinates of locations with open water, permanently flooded vegetation, seasonally flooded vegetation, and upland vegetation. From the permanently flooded vegetation we further distinguished papyrus, typha, and permanently flooded grasslands. From the seasonally flooded and upland vegetation, we further distinguished between grasslands and hardwoods. Examples of each class identified in Google Earth are shown in Figure 3. We identified 100 points for each class, giving us a total of 800 data points for each image and principal component.

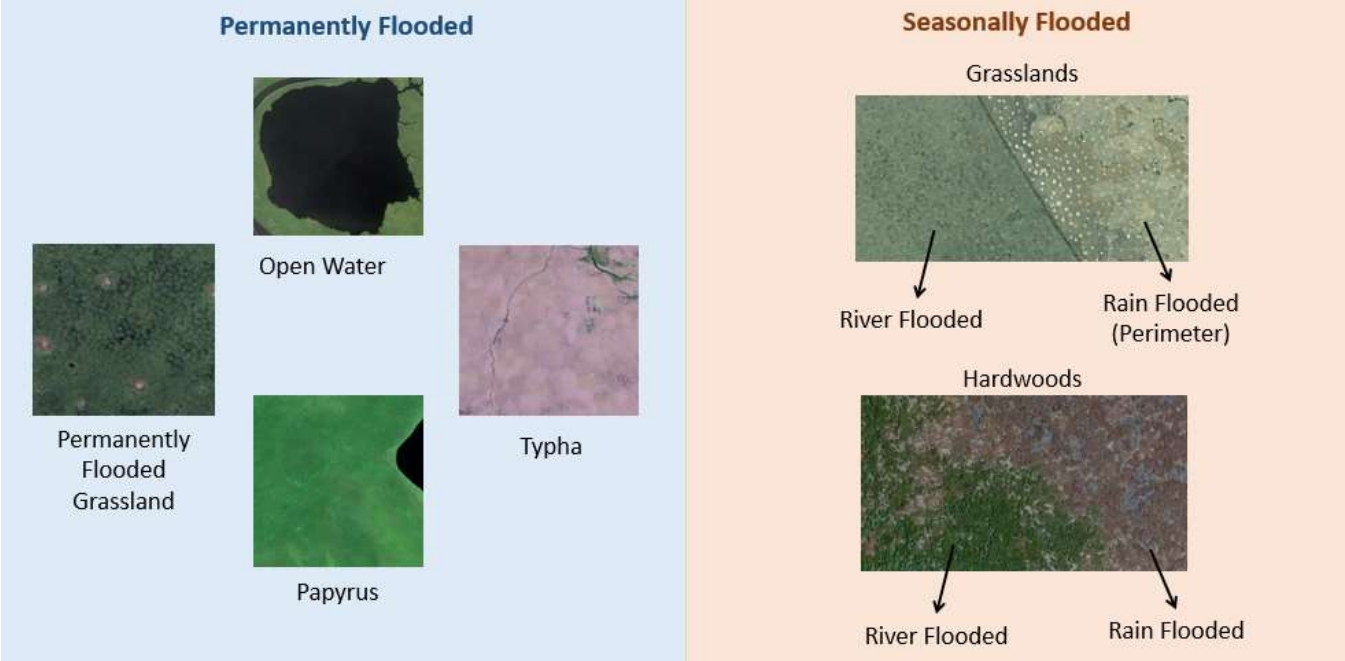


Figure 3: Google Earth Ground Truth Class Examples

Methodology

The clustering and classification methods we applied to the data were presented in the course notes and are listed and briefly described below. We decided to compare nonparametric and parametric as well as unsupervised and supervised techniques. Robert completed the nonparametric analysis and Courtney completed the parametric analysis. For unsupervised classification, we started with clustering methods to automate the identification of unique classes within the Sudd and then assigned each pixel to a cluster. To allow for comparison between the unsupervised and supervised results, we matched the unsupervised clusters to our pre-determined classes using the geographic coordinates of the ground truth points and determining which cluster aligned with each ground truth class most frequently. For supervised classification, the class statistics were determined from the ground truth data and then each pixel was assigned to one of these classes. We also experimented with a two stage classification approach that combined supervised parametric and nonparametric techniques. The accuracy of each method was determined using the true class identity of the ground truth points and counting the percentage of pixels that were correctly classified.

Nonparametric:

1. ***Unsupervised clustering with k-means followed by classification***
   The k-means algorithm iteratively calculates the mean of a pre-defined number of clusters, and groups each pixel into the cluster that is closest in terms of a Euclidean distance measure. For the full time series of NDWI, NDVI data, the distance measure is defined below. The initial means for each image were chosen by using the value of a randomly selected pixel within the image. The change in the mean values

between iterations was used as a convergence criterion. A custom code was written in Matlab to implement this procedure.

$$D_{i,k} = \sum_{j=1}^{12} \left(NDWI_{i,j} - \mu_{k,NDWI}\right)^2 + \sum_{j=1}^{12} \left(NDVI_{i,j} - \mu_{k,NDVI}\right)^2$$

$$where \quad i = pixel\ number$$
$$k = cla$$
$$j = mon$$

2. ***Supervised classification using minimum Euclidean distance from ground truth means***

The same distance measure was used as defined in the unsupervised classification, except instead of iteratively calculating the mean of each cluster, the means were calculated from the NDWI, NDVI and PC values of the ground truth points. The mean of each class was calculated from 85 of the 100 points, and the remaining 15 points were used to measure the accuracy of the classification; this ensures that the training and validation data sets are independent. A custom code was written in Matlab to implement this procedure.

Parametric:

1. ***Unsupervised clustering with expectation maximization (EM) for Gaussian mixture models (GMM) followed by classification***

Assuming the NDWI, NDVI, and PC values of each land class are normally distributed within each image, we can apply this method to determine the posterior probability that a given pixel belongs to each class. Each pixel can then be classified by finding the class that corresponds to the maximum posterior probability. The algorithm maximizes the log-likelihood function of the distribution of the observations, which is assumed to be a mixture of Gaussian distributions. The weights of the distributions that make up the mixed distribution can be thought of as the prior probabilities and represent the proportion of pixels that fall within each group. These weights, along with the mean and variance of each class for each NDWI, NDVI, and PC image are calculated iteratively. The difference in the log likelihood value between iterations was used as a convergence criterion. The iterative procedure for the NDWI, NDVI time series classification is described below. The initial weights were set to 1/8 and the initial covariance matrices were the identity matrix. The initial means for each NDWI, NDVI image were randomly selected from a normal distribution with parameters calculated from all pixels within the image. A custom code was written in Matlab to implement this procedure.

$$\boldsymbol{E\ Step}: Calculat \quad conditional\ probability\ that\ pixel\ x\ belongs\ to\ class\ k$$

$$p(x_i = k|X) = \frac{\pi_k\, p(X|\mu_k, \Sigma_k)}{\sum_{j=1}^{8} \pi_j p\left(X\middle|\mu_j, \Sigma_j\right)}$$

$$where: \quad X = \{NDVI_1, NDWI_1 \ldots NDVI_{12}, NDWI_{12}\}, observations$$
$$p(X|\mu_k, \Sigma_k) \sim Multivariate\ Normal$$
$$\pi_k = weights\ (priors) for\ class\ k$$

$$\textbf{M Step}:\ Re-estimat\ \ parameters$$

$$\mu_k = \frac{1}{N_k}\sum_{i=1}^{N} p(x_i = k|X)x_i$$

$$\Sigma_k = \frac{1}{N_k}\sum_{i=1}^{N_k} p(x_i = k|X)(x_i - u_k)(x_i - u_k)^T$$

$$\pi_k = \frac{N_k}{N}\sum_{i=1}^{N_k} p(x_i = k|X)$$

$$where \qquad N_k = number\ of\ pixels\ that\ belong\ to\ class\ k$$

$$N = tot \quad number\ of\ pixels$$

$$\textbf{Evaluate the Log} - \textbf{likelihoo}\ :$$

$$log\big(\ p(X|\mu_k, \Sigma_k, \mu_k)\big) = \sum_{i=1}^{N} log\left\{\sum_{k=1}^{8} \pi_j p\big(X|\mu_j, \Sigma_j\big)\right\}$$

2. ***Supervised classification using linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA)***

   Similar to the EM for GMM, we assume the NDWI, NDVI, and PC values of each land class are normally distributed within each image. Again, we want to maximize the joint probability of the observations (weighted by their priors); however, in this formulation we assume that the statistics for each class can be estimated from the ground truth points. The optimization problem is then to find the class that optimizes the log-likelihood of the observed joint distribution. If the discriminant functions, which are derived from the log-likelihood functions, of each class (defined below) are set equal to each other, the NDWI, NDVI, and PC values that satisfy the equations can be considered the decision boundaries. Each pixel can then be classified by which side of the boundary it falls on. If the covariance matrices are assumed to be equal for all classes, the sigma terms in the discriminant functions cancel out and the result is a linear decision boundary (hence the name linear discriminant analysis). If the covariance matrices are different for each class, then the decision boundaries are quadratic. This classification was implemented using the Matlab 'classify' function. 85 out of 100 ground truth points for each class were used in the classification, and the remaining 15 values were used for validation and measuring accuracy.

$$discriminant\ function\ for\ class\ k\text{:}\quad \delta_k(x) = x^T\widehat{\Sigma_k}^{-1}\widehat{\mu_k} - \frac{1}{2}\widehat{\mu_k}^T\widehat{\Sigma_k}^{-1}\widehat{\mu_k} + \log(\widehat{\pi_k})$$

2-Stage classification using QDA and logistic regression with lasso model selection

We came up with a final classification technique to test based on the idea that during the driest time of year, March, the permanently flooded areas are easiest to distinguish. We used QDA to obtain a classified image for the month of March, using both the NDVI and NDWI values. While the permanently flooded areas were accurately portrayed, the algorithm performed poorly when distinguishing between the river flooded and rain flooded grasslands and hardwoods. Therefore we indexed the hardwood and grassland pixels and fit a logistic regression

model to the NDWI and NDVI values for all months of each class that indicated the probability that the pixel is flooded by the River. These models are illustrated below. To train the models, the river flooded hardwoods and grasslands were set to a value of 1, and the rain flooded pixels were set to 0. We used the 'lassoglm' function in Matlab to select the best parameters for fitting the model, which were determined by the combination of parameters that resulted in the minimum cross-validation error. Each grassland and hardwood pixel in the image can then be classified as being river flooded or rain flooded by calculating the probability from the fitted model and assigning pixels with a probability greater than 0.5 to the river flooded class.

$$p(grassland\ is\ river\ flooded) = \frac{1}{1 + e^{\{\beta_0 + \beta_1 NDVI_{1,G} + \beta_2 NDW_{1,G} + \cdots + \beta_{24} NDW_{24,G}\}}}$$

$$p(hardwood\ is\ river\ flooded) = \frac{1}{1 + e^{\{\beta_0 + \beta_1 NDVI_{1,H} + \beta_2 NDWI_{1,H} + \cdots + \beta_{24} NDWI_{24,H}\}}}$$

$$where:$$

$$G: grassland\ pixel \quad \& \quad H: hardwood\ pixel$$

The set of reduced parameters resulting from the lasso procedure were then used in a 4th iteration on the LDA and QDA classification to see if there was a significant change in accuracy between the reduced time series and the full time series. If the accuracy measure is similar for both data inputs, then the dimensionality of the data can be reduced to save computation time in future classifications.

Evaluation and Final Results

The resulting accuracy of each classification technique and data set is summarized in Table 1, where the percentage of correctly classified pixels for each class as well as the average accuracy for all classes is presented. The abbreviations in the table are defined as follows: RES = reservoir (open water), PFG = permanently flooded grassland, PAP = papyrus, TYP = typha, RFG = river flooded grassland, PG = perimeter grassland (rain flooded), RFH = river flooded hardwood, and PH = perimeter hardwood (rain flooded).

		Data Used in Classification	RES	PFG	PAP	TYP	RFG	PG	RFH	PH	Average
Nonparametric	Unsupervised K-Means	NDWI, NDVI Full	0.61	0.97	0.87	0.89	0	0.62	0.78	0.21	0.62
		PCA Comb	0.84	0.82	0.93	0.37	0.38	0.42	0.57	0	0.54
		PCA Indiv	0.32	0.53	0.63	0.48	0.46	0	0.45	0.05	0.37
	Supervised Euclidean Distance	NDWI, NDVI Full	0.67	1.00	1.00	1.00	0.40	0.87	0.87	0.87	0.84
		PCA Comb	0.67	1.00	0.67	0.20	0.40	0.60	0.33	0.53	0.55
		PCA Indiv	0.07	0.53	0.73	0.33	0.33	0.00	0.60	0.07	0.33
Parametric	Unsupervised EM for GMM	NDWI, NDVI Full	0.98	0	0.91	0	0.73	0.90	0	0.89	0.55
		PCA Comb	0.97	0	1.00	0	0.81	0.48	0	1.00	0.53
		PCA Indiv	0.98	0	0.98	0.03	0.60	0.75	0	0.98	0.54
	Supervised LDA	NDWI, NDVI Full	0.93	1.00	1.00	1.00	0.60	0.87	0.93	0.67	0.88
		NDWI, NDVI Red	0.93	1.00	1.00	0.93	0.67	0.87	0.87	0.60	0.86
		PCA Comb	0.80	1.00	1.00	0.53	0.40	0.73	0.60	0.33	0.68
		PCA Indiv	0.80	1.00	1.00	0.47	0.40	0.73	0.60	0.40	0.68
	Supervised QDA	NDWI, NDVI Full	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		NDWI, NDVI Red	1.00	1.00	1.00	1.00	1.00	1.00	0.93	1.00	0.99
		PCA Comb	1.00	1.00	1.00	0.47	0.53	0.93	0.80	0.73	0.81
		PCA Indiv	1.00	1.00	1.00	0.40	0.47	0.87	0.80	0.73	0.78
Mixed	Log. Reg.	NDWI, NDVI Red	0.93	0.88	0.96	0.86	0.48	0.96	0.87	0.81	0.84

Table1: Summary of Classification Accuracy

From the logistic regression model with model selection performed with the lasso technique, the following 14 indices and months were considered to be the best reduced parameter set: NDWI values for months Feb, Mar, May, Sep, and Oct, and NDVI values for months Jan, Feb, Mar, May, Jun, Jul, Sep, Oct, and Dec. The deviance measure and the optimal selection of the regularization parameter for the grassland and hardwood models are illustrated in Figure 4. The green line shows the lambda value with that corresponds to the overall minimum cross-validation error, and the blue line corresponds to the lambda value with the minimum cross validation plus one standard deviation. Since we were interested in reducing the number of parameters in the model, we chose the lambda that intersects the blue line. The parameters associated with these optimal lambda values were combined and chosen as the reduced time series of NDWI and NDVI values.

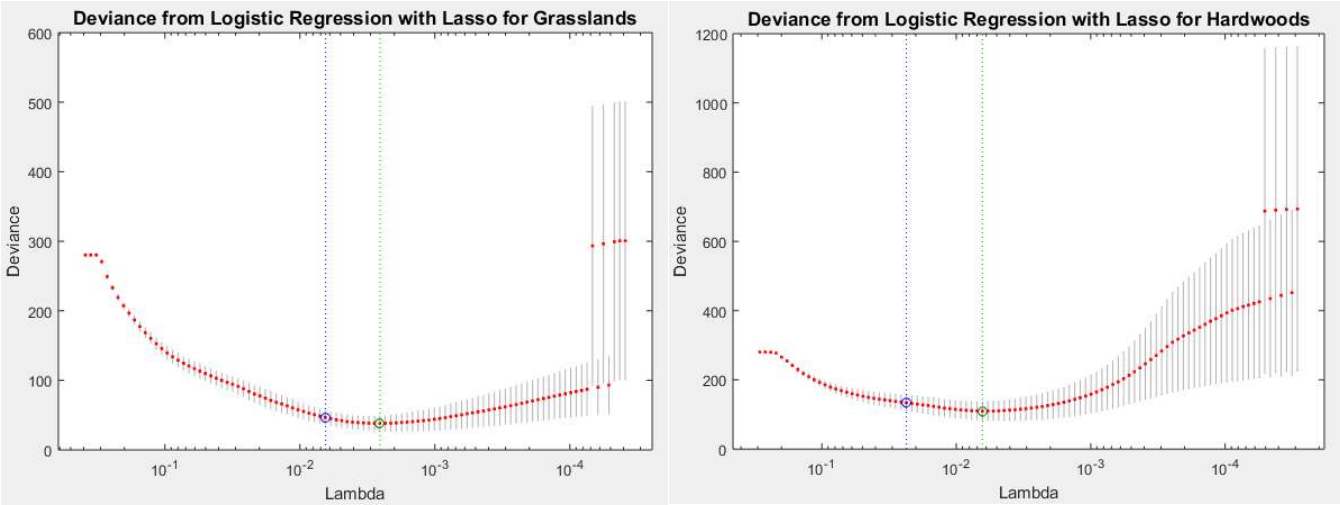


Figure 4: Optimization of regularization parameter for logistic regression models

All of the final classified images are shown in Appendix B, and a few are included here for discussion. The following is a summary of our main observations and conclusions from analyzing the classified images and the accuracy percentages listed in Table 1:

1. NDWI, NDVI time series versus principal components
   For all classification methods, the NDWI and NDVI time series as input data resulted in higher accuracy percentages as compared to the PC combinations. In addition, the PC's derived from the combined NDWI, NDVI data set performed slightly better than the PC's derived from the NDWI and NDVI individually. Since there are many ways the PC's can be combined, the classification algorithms could be performed on more combinations (than just the 2 sets presented here) and the optimal combination could be determined by the result with the highest accuracy. However, based on these results we conclude that the NDWI, NDVI time series data results in a superior classification.

2. Full NDWI, NDVI time series versus reduced NDWI, NDVI time series
   The classifications resulting from the reduced time series (14 images) had nearly identical accuracy measures as compared to the full time series (24 images). Therefore, we conclude that using the reduced

time series is a preferred alternative in order to reduce the time required for downloaded the data, flagging pixels with quality control issues, deriving the indices, and finally, applying the classification algorithms.

3. Unsupervised versus Supervised Classification

   The supervised classification performed much better than the unsupervised classification. As can be seen from Table 1, many of the classes had a classification accuracy of 0, indicating that they were missed entirely. To illustrate this result, the final cluster means determined with the EM for GMM algorithm for the NDWI, NDVI time series are compared to the means calculated from the ground truth points for the month of September in Figure 5. In addition, the resulting classified image for this iteration is shown in Figure 6. Note that these Figures are consistent with the accuracy reported in the table; none of the ground truth pixels were correctly classified for the river flooded hardwoods (RFH), typha (TYP), and permanently flooded grasslands (PFG) and these classes are not represented in Figure 6. In addition, the cluster means in Figure 5 are relatively far from the ground truth means for these classes. In order to address this issue, we could increase the number of clusters to be identified and then redistribute them to our ground truth classes; however, we believe the supervised classification method would still have a superior classification accuracy.

4. Parametric versus Nonparametric Classification

   The parametric classification procedures resulted in higher levels of accuracy. This indicates that the ground truth classes have different magnitudes of variance, which should be accounted for in the classification. Additionally the higher accuracy supports the assumption that the NDWI, NDVI and PC values are normally distributed within each image.

5. LDA versus QDA

   While the QDA classification results in a higher accuracy than the LDA classification, it may be overfitting the ground truth data. The resulting classified images are shown in Figure 7. On first notice, note that the LDA image appears more homogenous and has less speckle, which could indicate a superior classification. In order to explore this further, we calculated the total area of permanently flooded wetland and seasonally flooded wetland from the QDA and LDA results and compared them to the areas determined from the aerial imagery study completed in the early 1980's (Table 2). Based on this comparison, we believe the LDA classification is superior to the QDA, even though the accuracy measure us lower.

	1980's Aerial Imagery	QDA Results	LDA Results
Permanently Flooded Area (km^2)	15,000	12,054	16,260
Seasonally Flooded Area (km^2)	30,000	56,978	34,866

Table 2: Comparison of Flooded Areas

6. Performance of the 2-Stage classification

   This classification technique performed well, with an average accuracy of 84%; however, it severely underestimated the area of rain flooded grassland (RFG). Overall, the accuracy was very similar to that of

the supervised Euclidean distance classification. This result seems logical considering logistic regression is also a supervised nonparametric classification method in this instance.

After implementing and reviewing all of the classification methods, we believe that the LDA supervised classification technique should applied to the reduced time series of NDWI and NDVI monthly mean values to obtain the extent of seasonal flooding for the Sudd Wetland. A similar analysis to the one presented here for 2008 could be applied to all years where MODIS data is available (2000 – present) to derive a full time series of flooding extents.

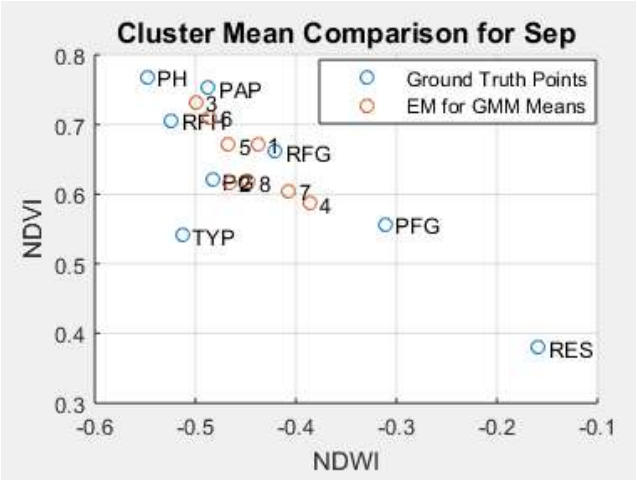


Figure 5: Comparison of clusters and ground truth means

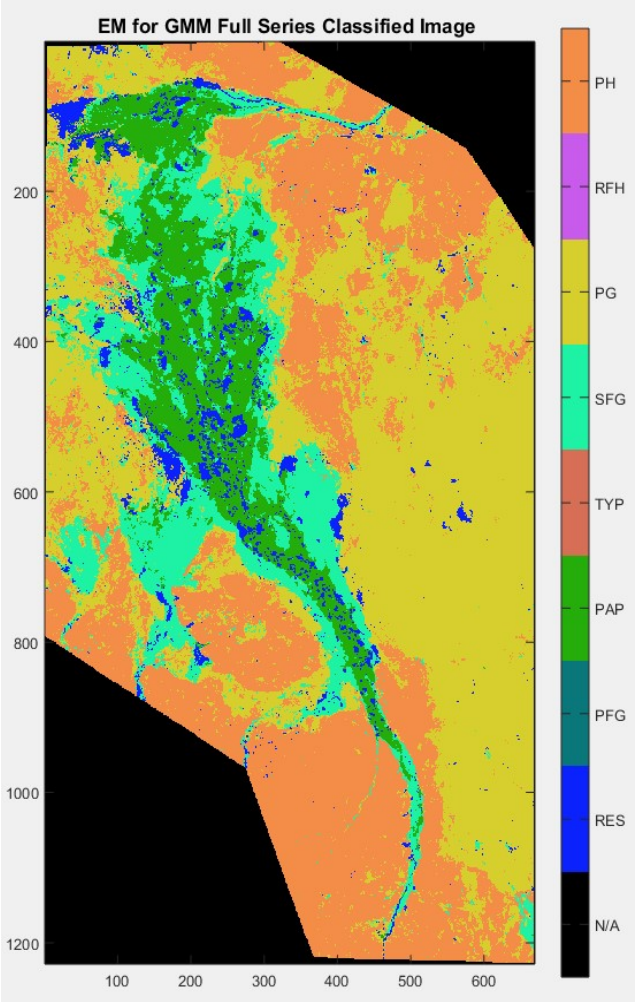


Figure 6: EM for GMM Full Series Classified Image

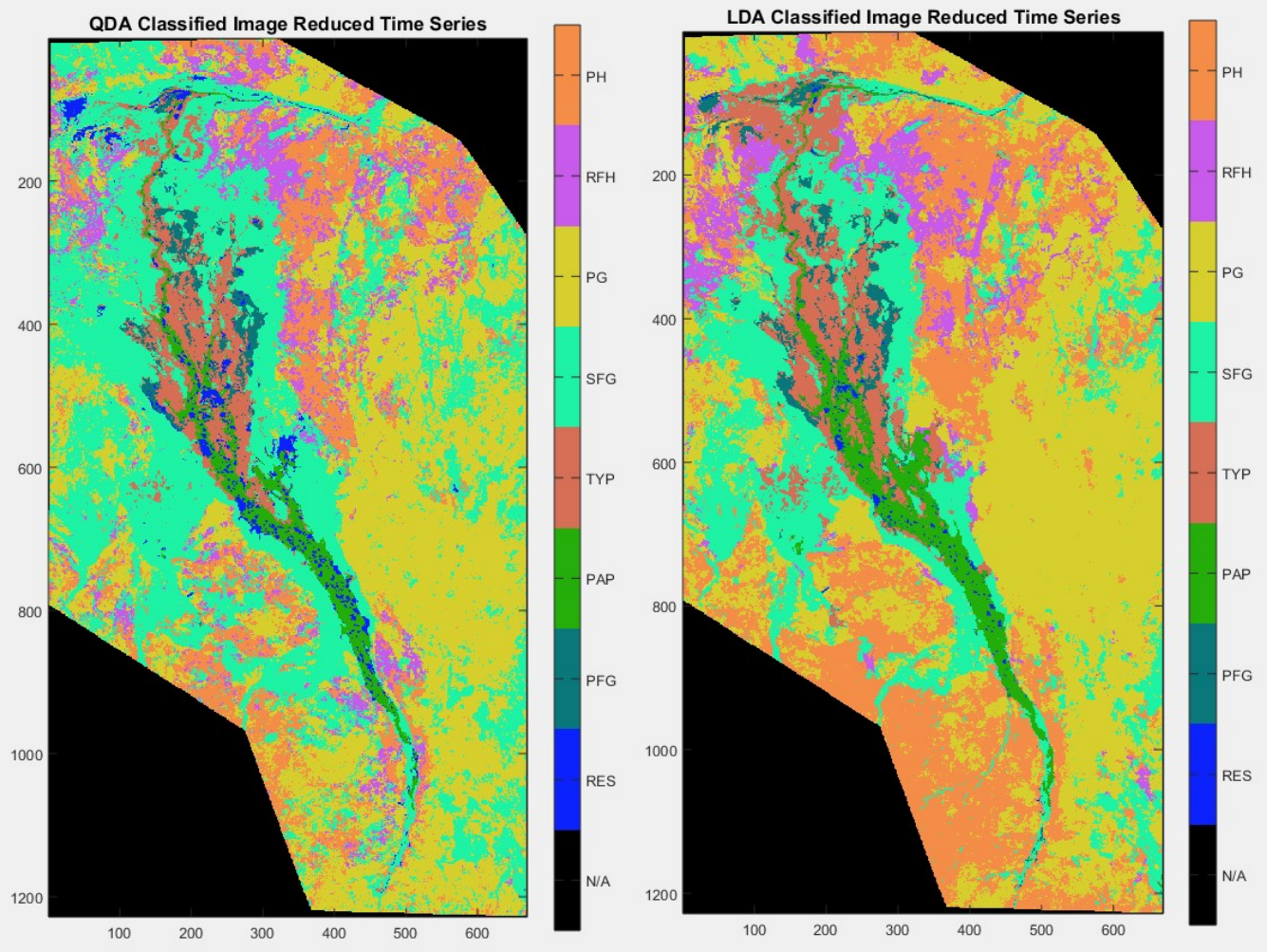


Figure 7: Comparison of QDA and LDA Classifications from the Reduced NDWI and NDVI Time Series

Bibliography

Howell, P., Lock, M., & Cobb, S. (1988). *The Jonglei Canal.* Cambridge: University Press.

NASA LP DAAC. (2012). *MODIS Land Products Quality Assurance Tutorial: Part-1.* Sioux Falls: NASA LP DAAC.

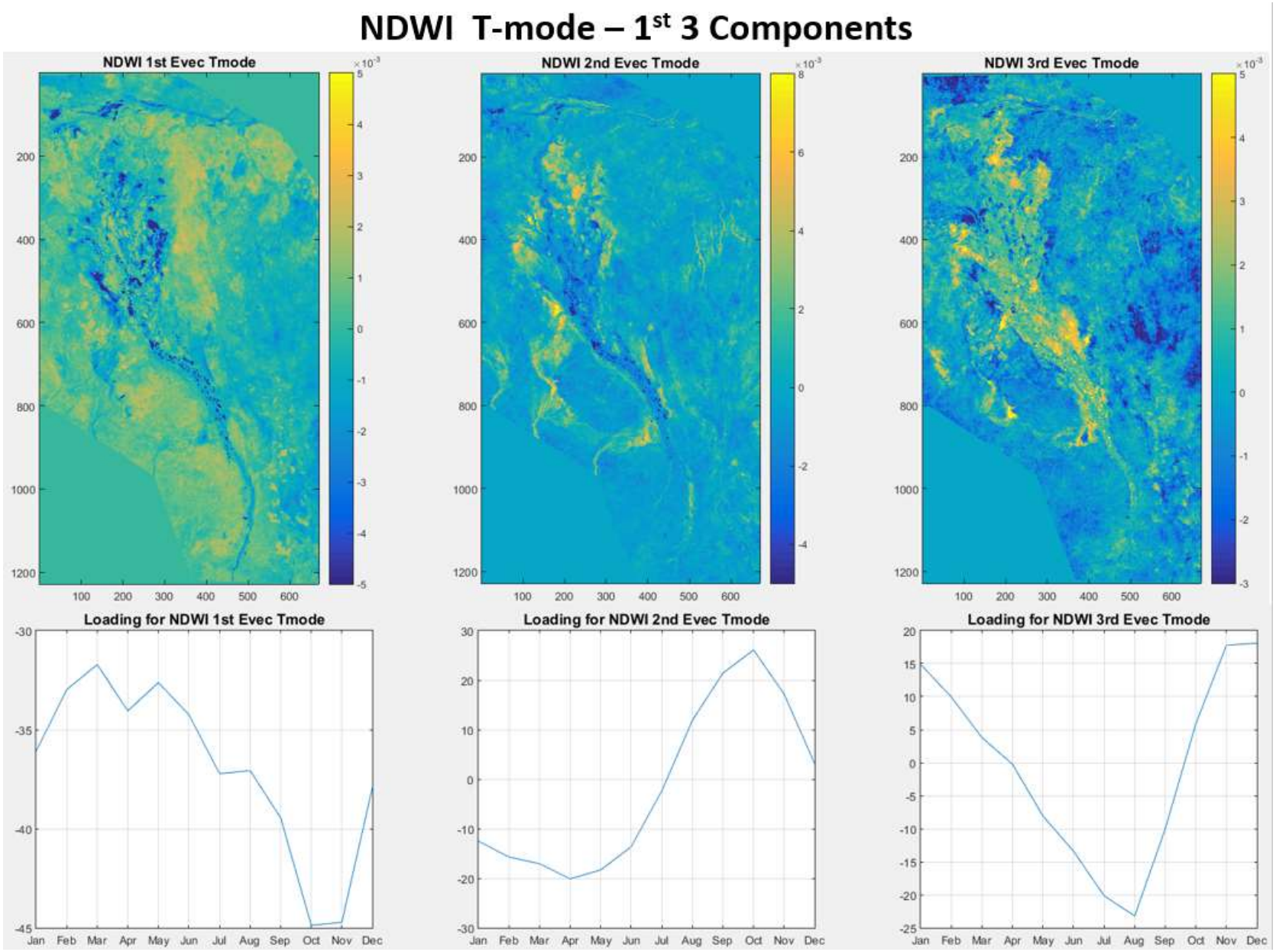
Sutcliffe, J., & Parks, Y. (1987). Hydrological modelling of the Sudd and Jonglei Canal. *Hydrological Sciences*.

Vermote, E. (2015). *MODIS Surface Reflectance User's Guide.* Silver Spring: NASA.

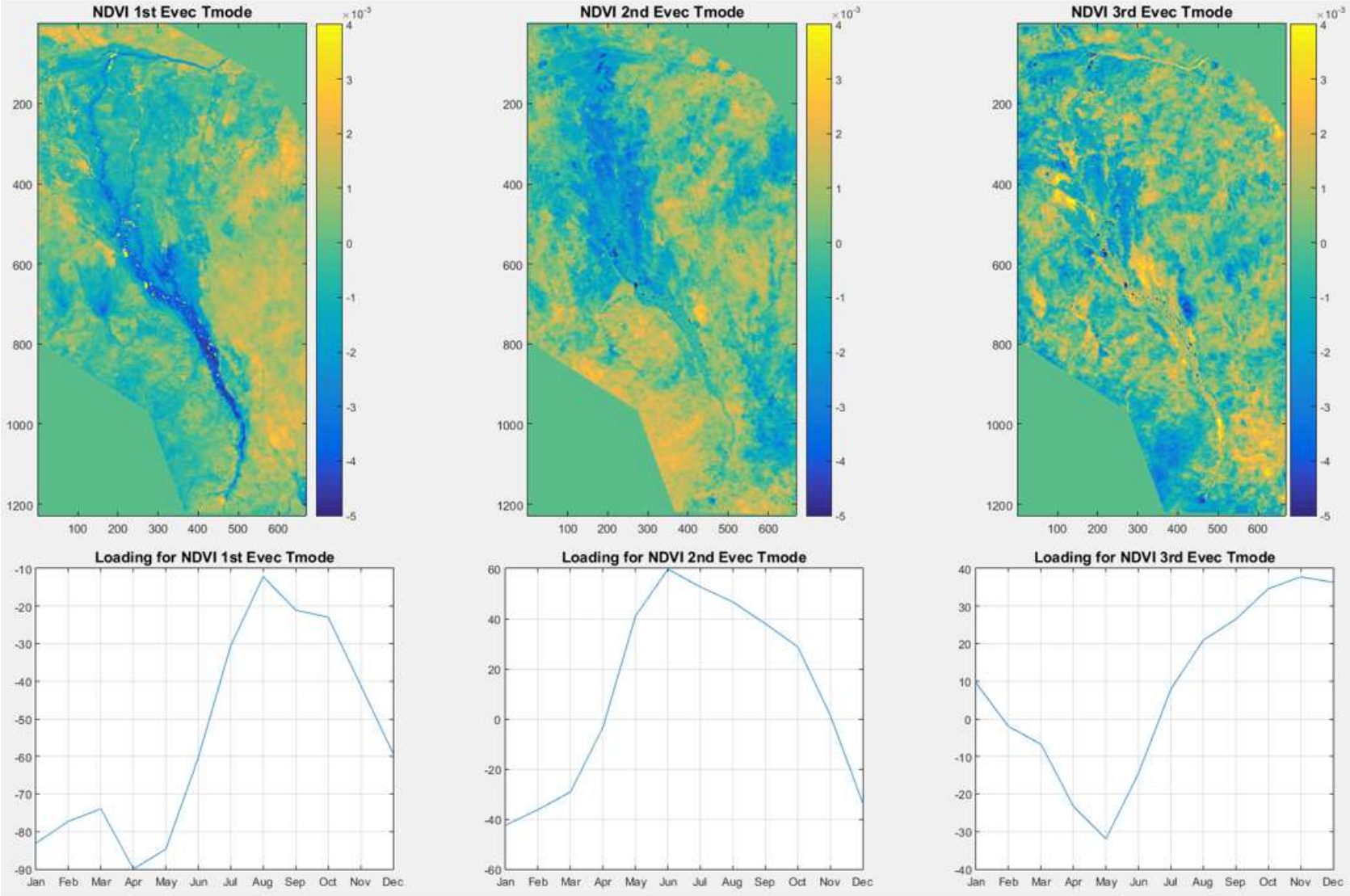
Wilks, D. (2011). *Statistical Methods in the Atmospheric Sciences.* Oxford: Academic Press.

Xu, H. (2006). Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. *International Journal of Remote Sensing, 27*(14).

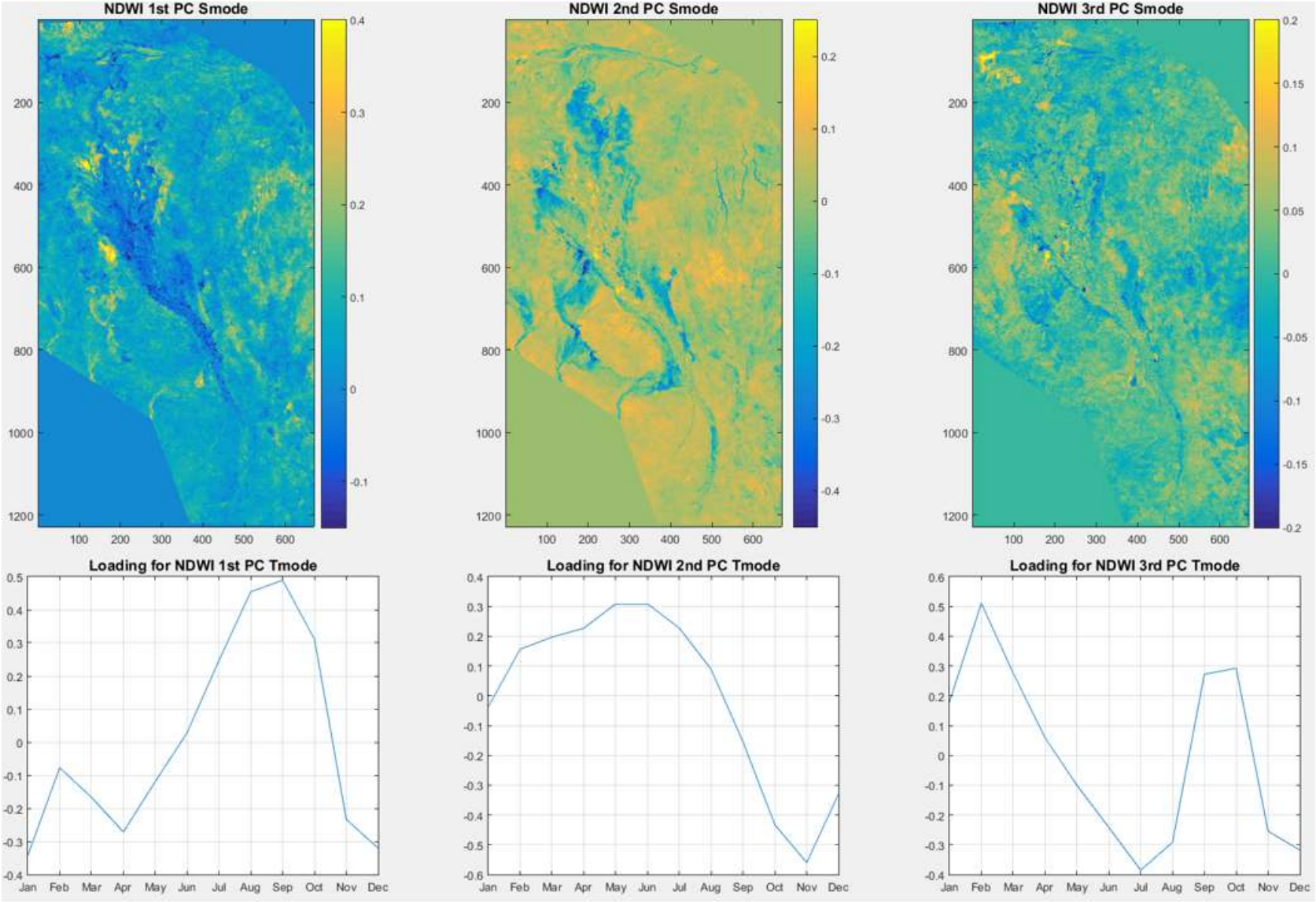
Appendix A: PCA Results



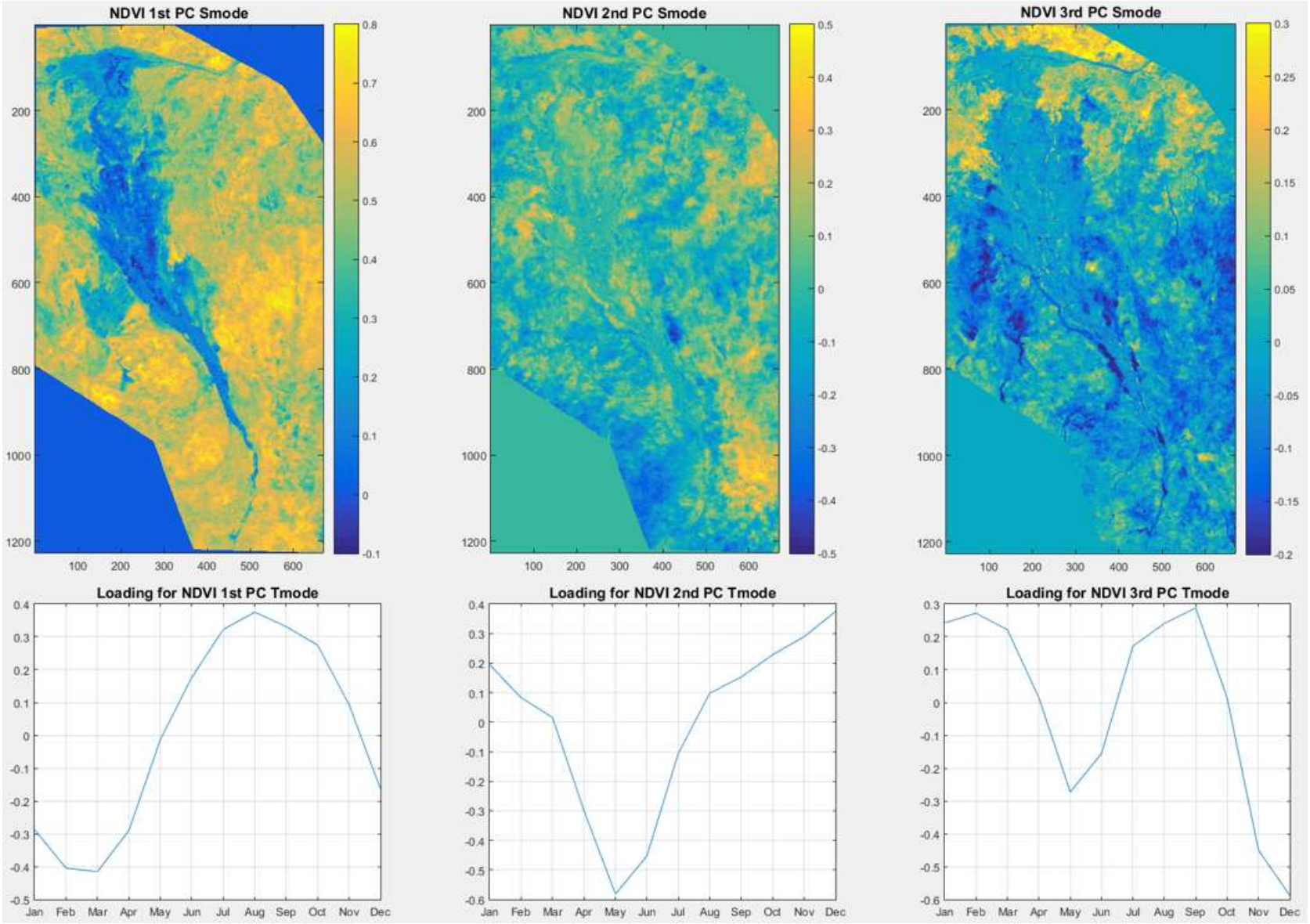
Appendix A: PCA Results



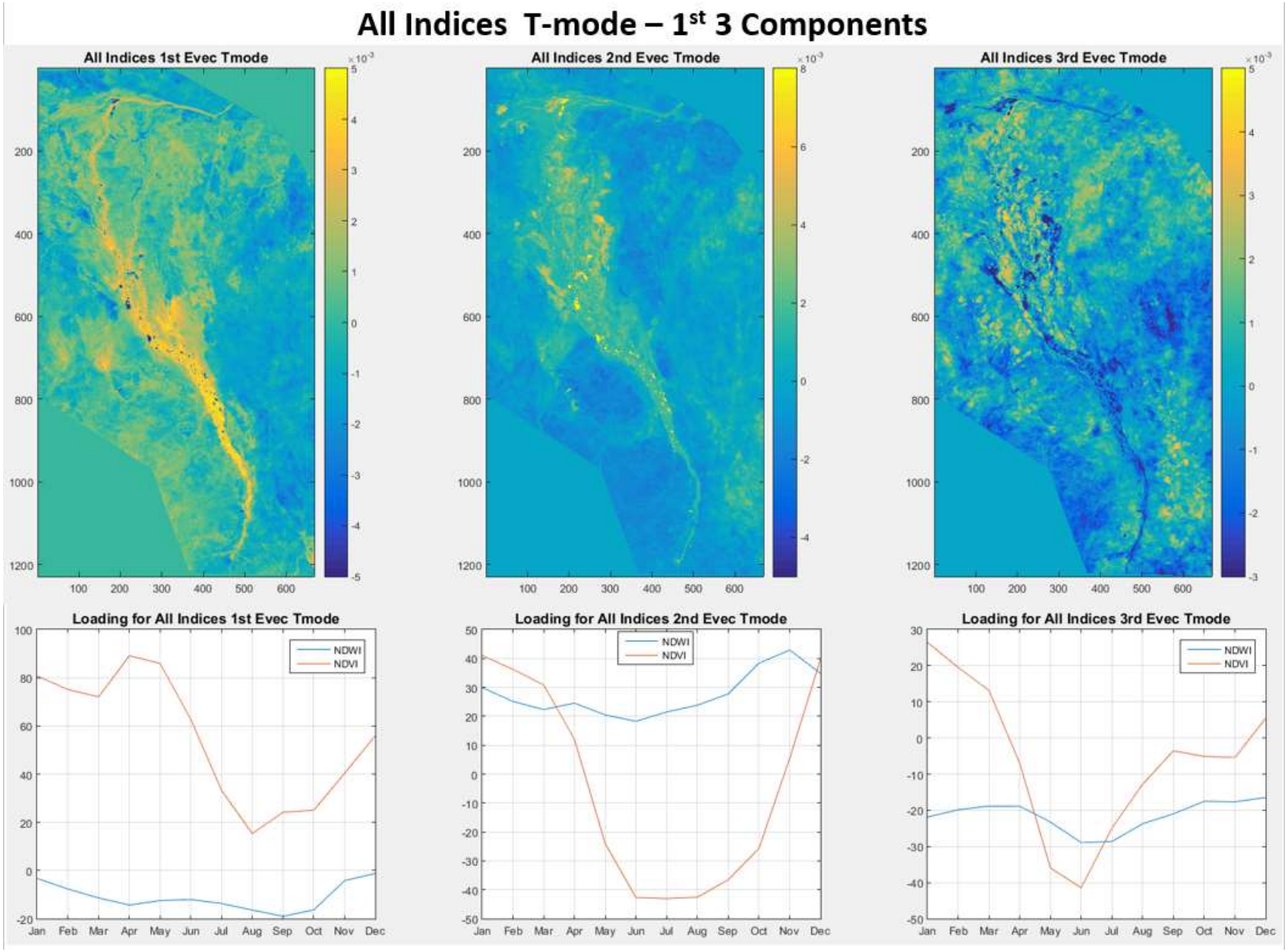
Appendix A: PCA Results



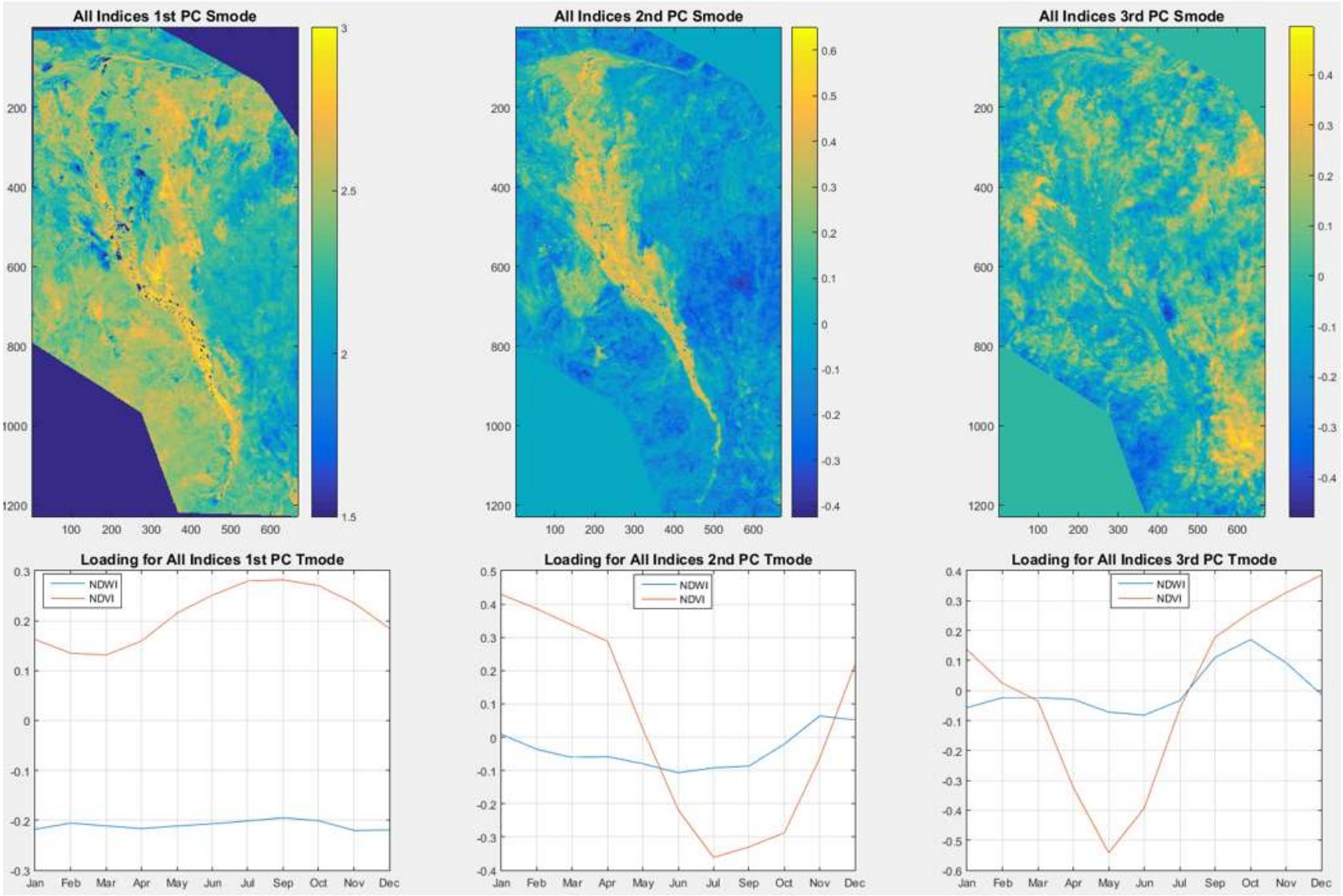
Appendix A: PCA Results



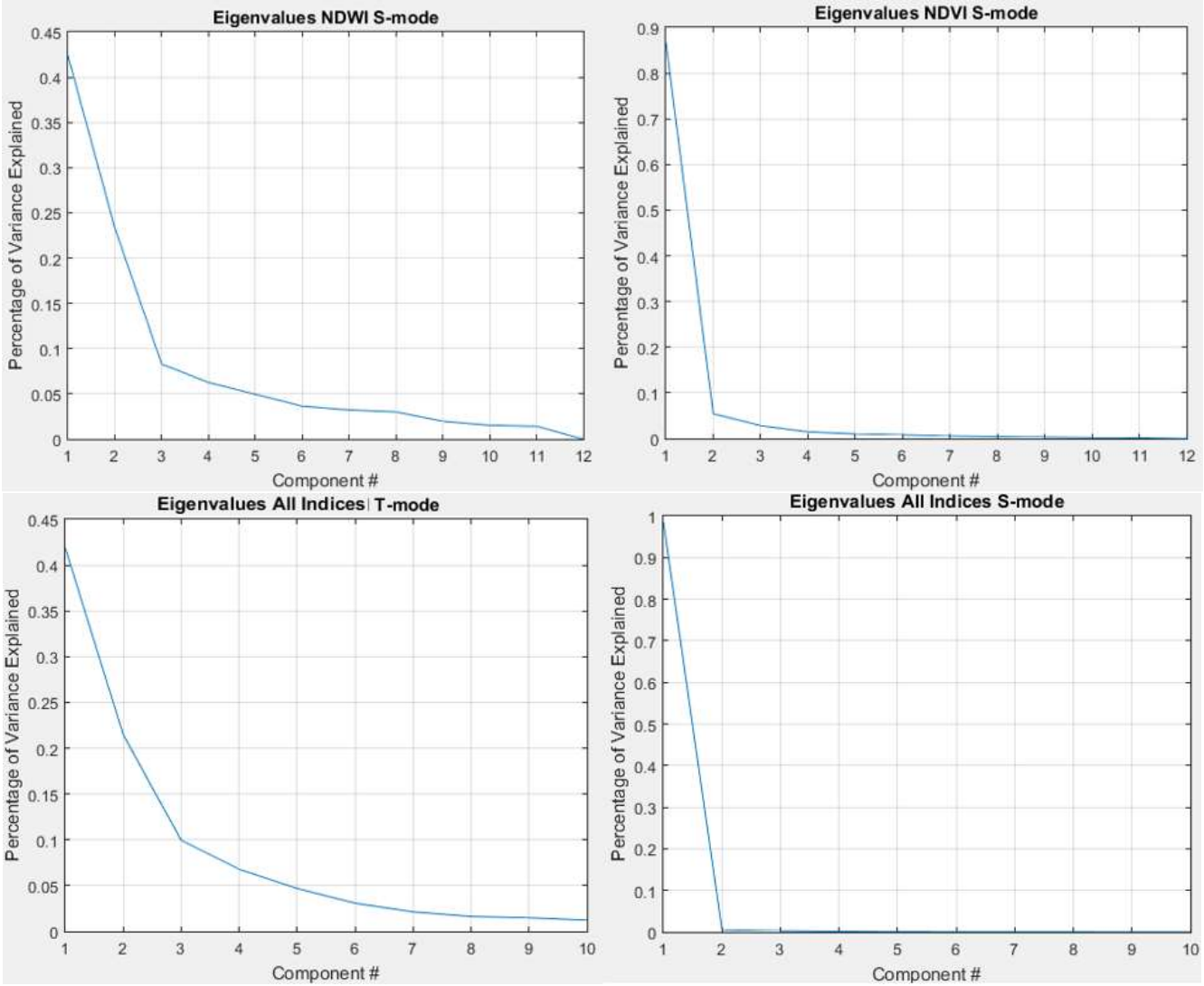
Appendix A: PCA Results



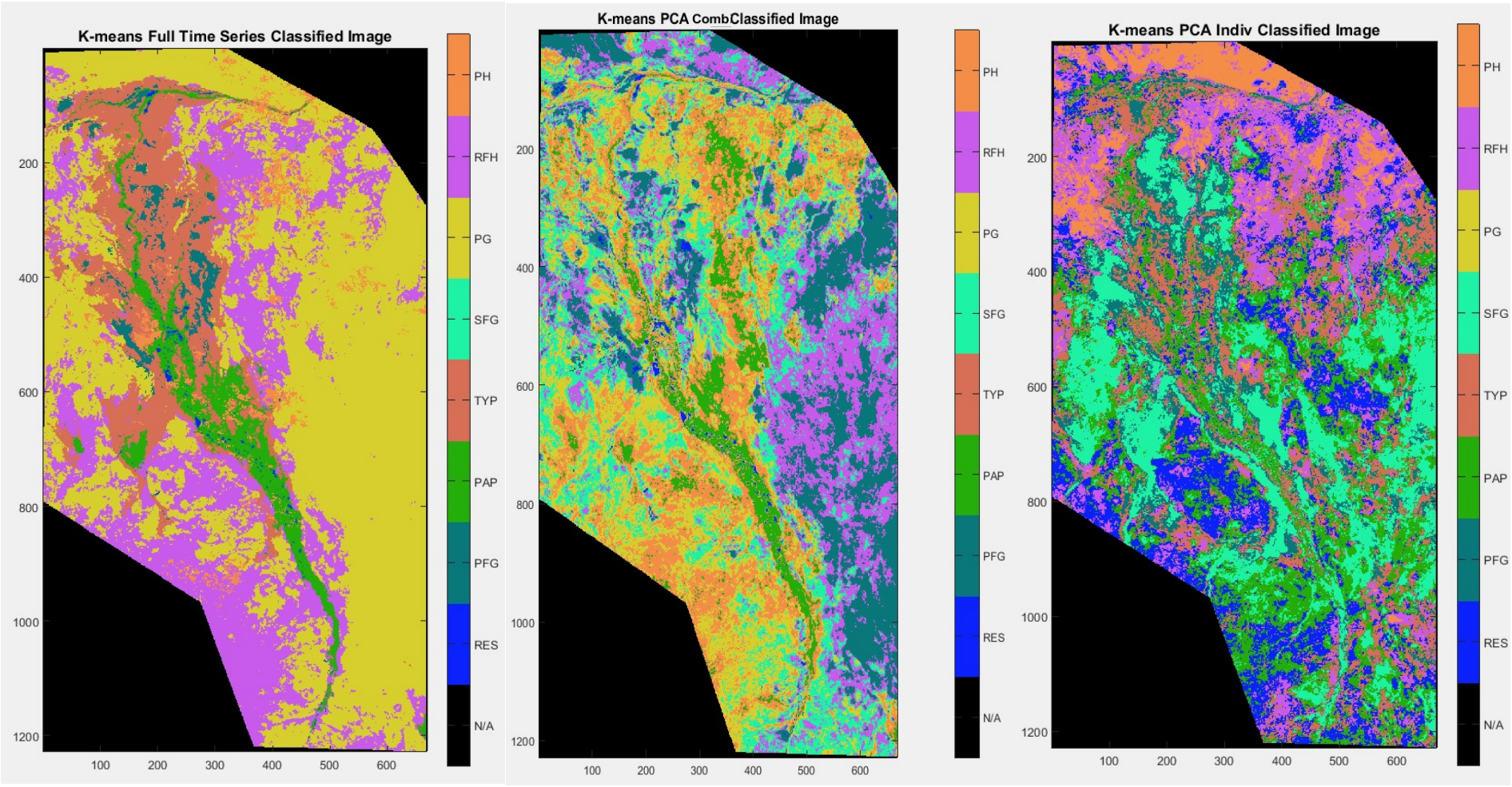
Appendix A: PCA Results



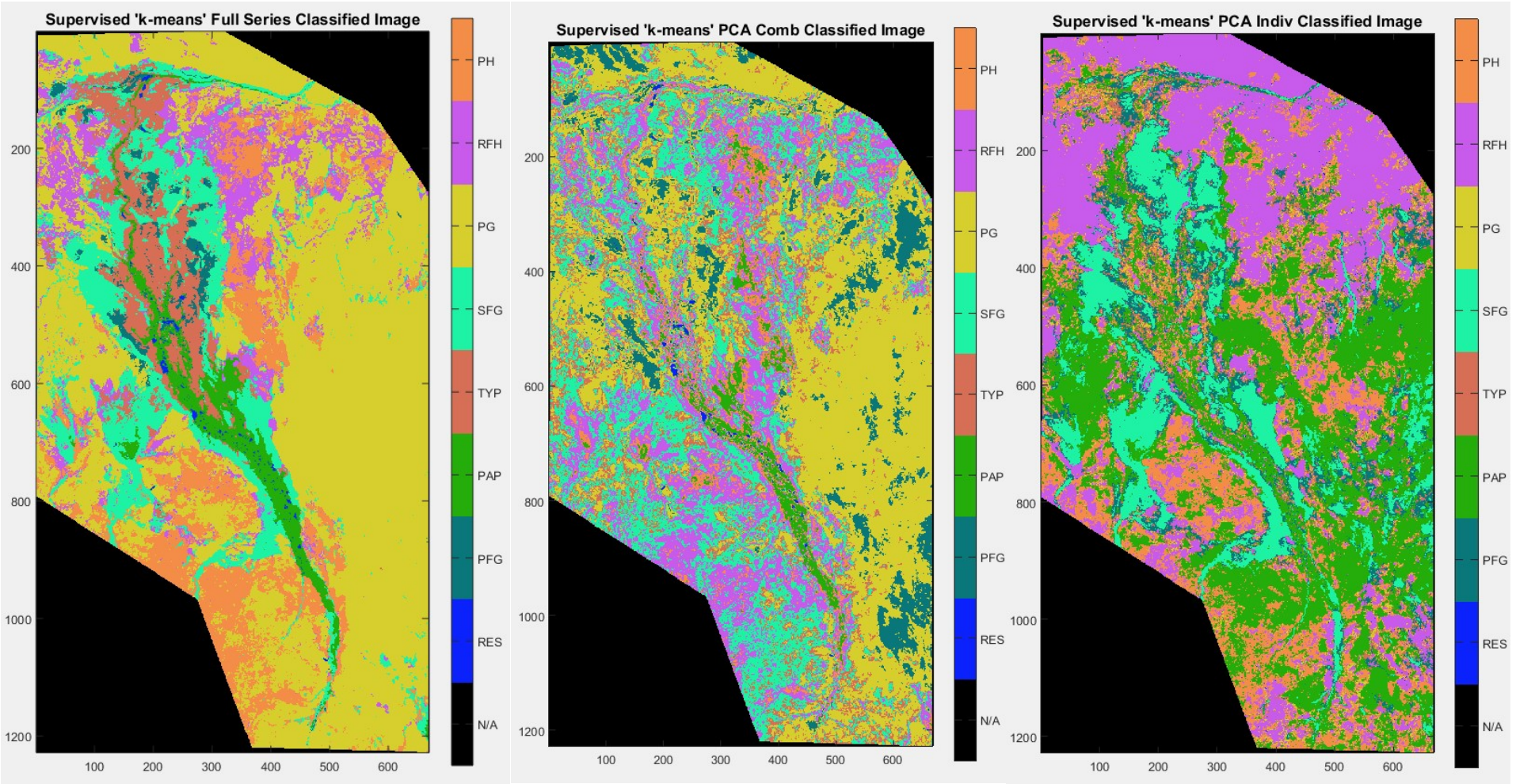
Appendix A: PCA Results



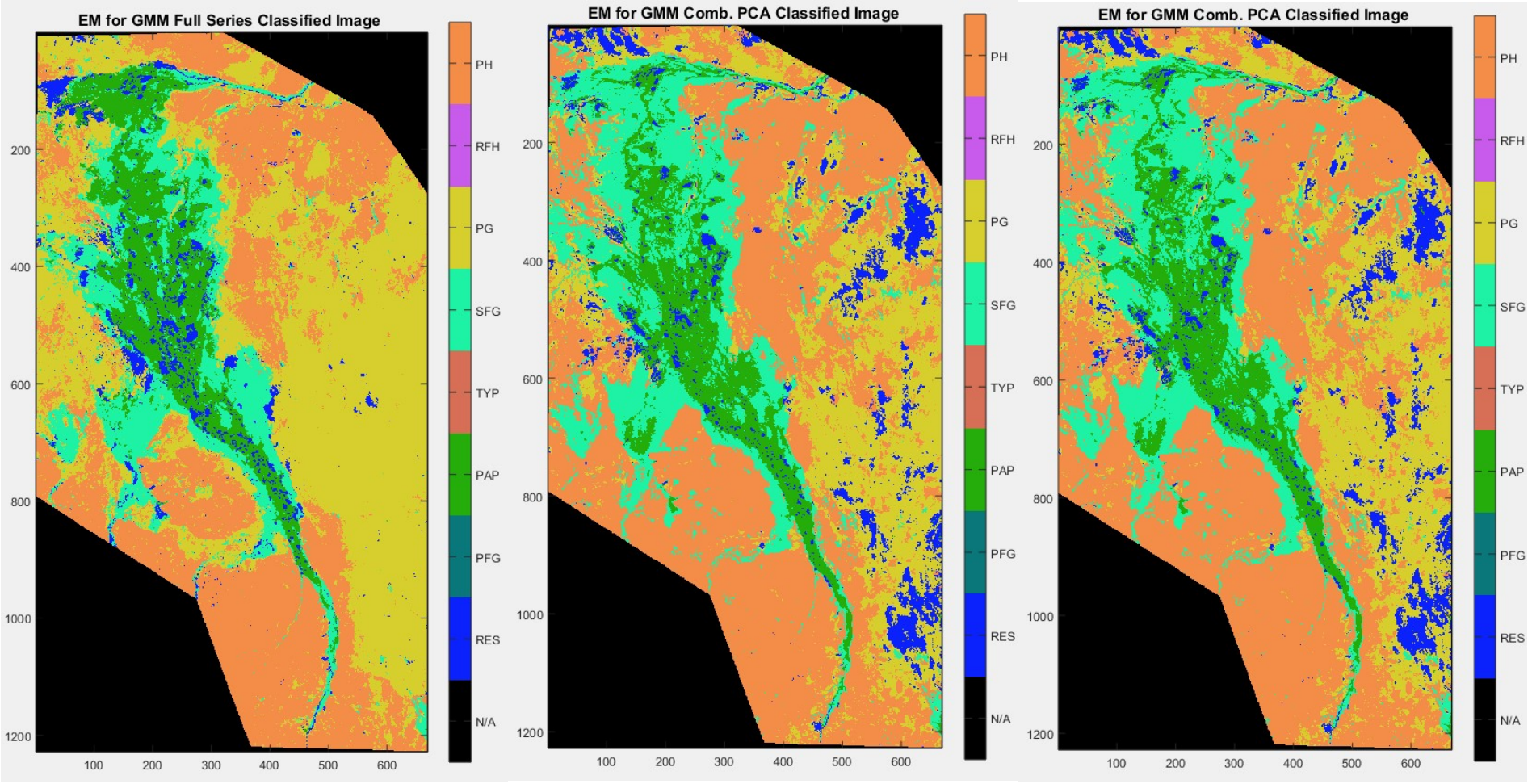
Appendix B: Classified Images



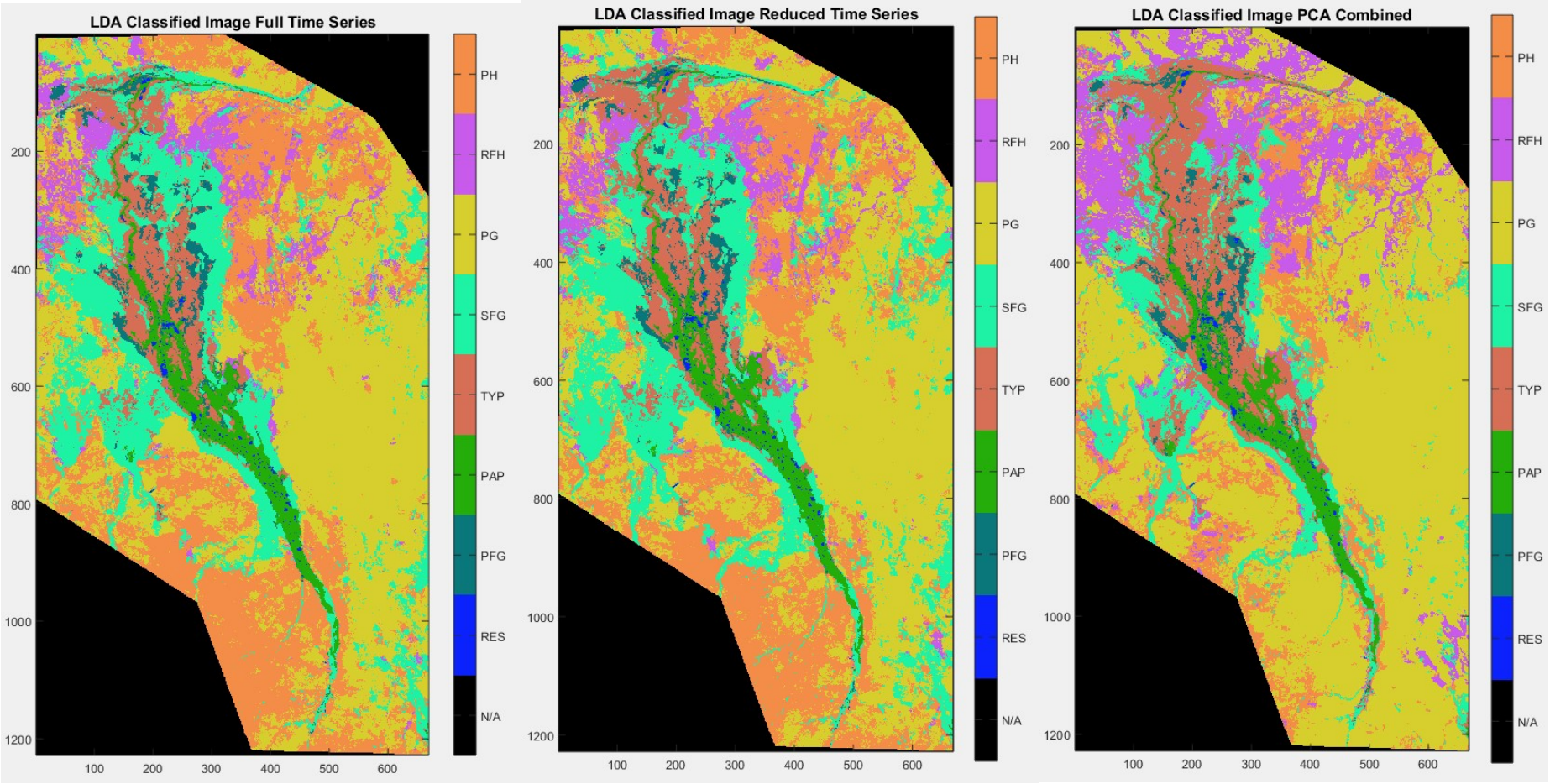
Appendix B: Classified Images



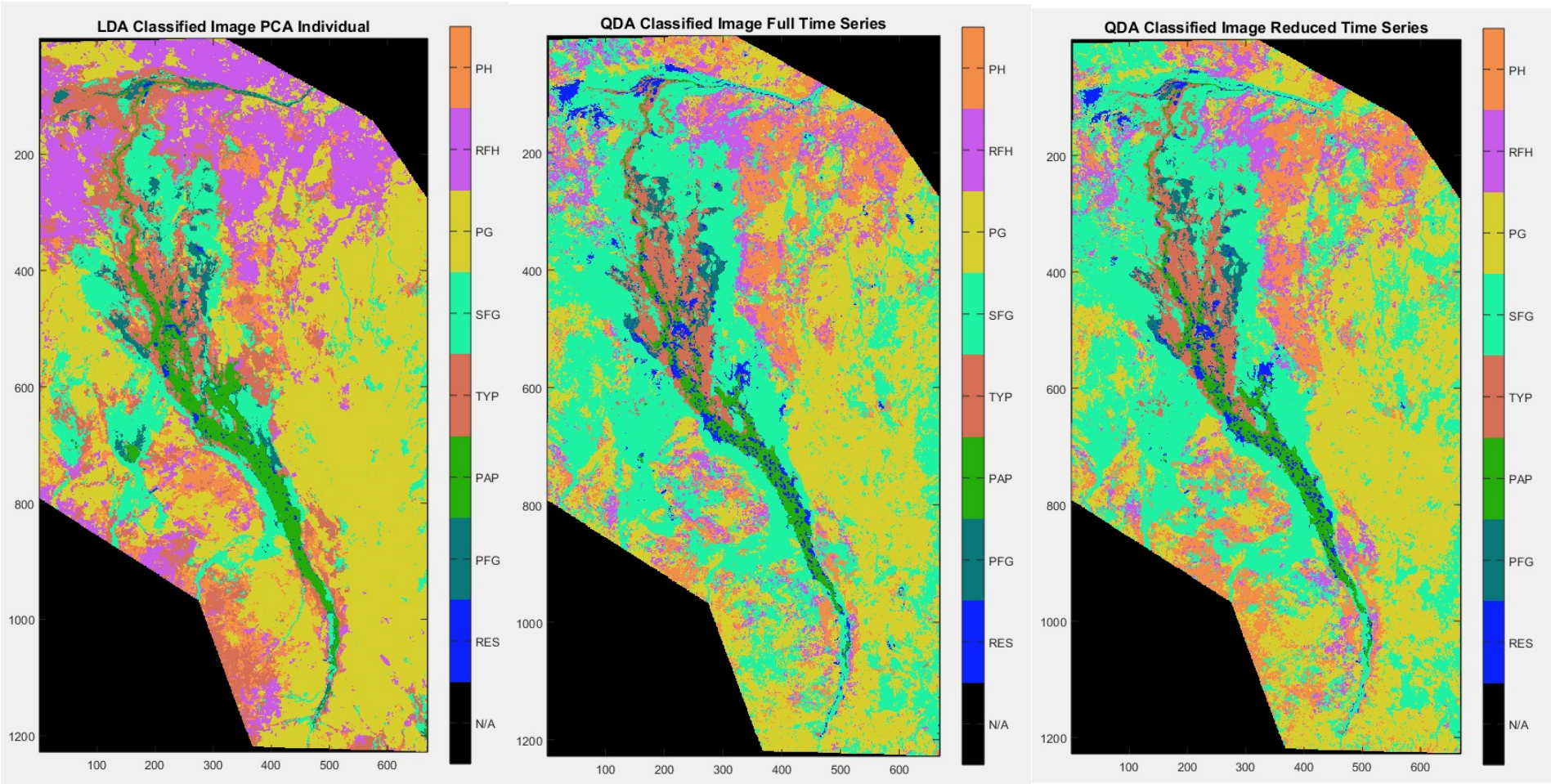
Appendix B: Classified Images



Appendix B: Classified Images



Appendix B: Classified Images



Appendix B: Classified Images