Increased-specificity famine prediction using satellite observation data

John A. Quinn
Makerere University
Kampala, Uganda
jquinn@cit.mak.ac.ug

Washington Okori
Makerere University
Kampala, Uganda
wokori@cit.mak.ac.ug

Anthony Gidudu
Makerere University
Kampala, Uganda
agidudu@tech.mak.ac.ug

ABSTRACT

This paper examines the use of remote sensing satellite data to predict food shortages among different categories of households in famine-prone areas. Normalized Difference Vegetation Index (NDVI) and rainfall estimate data, which can be derived from multi-spectral satellite radiometer images, has long been used to predict crop yields and hence famine. This gives an overall prediction of food insecurity in an area, though in a heterogeneous population it does not directly predict which sectors of society or households are most at risk.

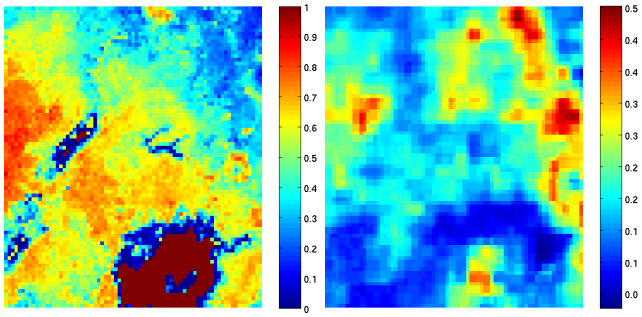
In this work we use information on 3094 households across Uganda collected between 2004-2005. We describe a method for clustering households in such a way that the cluster decision boundaries are both relevant for improved-specificity famine prediction and are easily communicated. We then give classification results for predicting food security status at a household level given different combinations of satellite data, demographic data, and household category indices found by our clustering method. The food security classification performance of this model demonstrates the potential of this approach for making predictions of famine for specific areas and demographic groups.

Categories and Subject Descriptors

J.2 [**Computer Applications**]: Physical Sciences and Engineering—*Earth and atmospheric sciences*; I.5 [**Computing Methodologies**]: Pattern Recognition

1. INTRODUCTION

Having an early warning of an impending famine increases the chance that something can be done about it. Both demographic and satellite data have been used in different ways to drive systems which predict food insecurity for this purpose. In this study, we combine satellite image data with data on specific households (for example, on the number of people in the household, the land size avaiable for farming, ownership of livestock and distance to the nearest road) in order to make more specific predictions than could be made with either source of data alone.



(a) (b)

Figure 1: Example snapshots of (a) NDVI and (b) RFE data across the extent of Uganda, both at a resolution of 8000 metres. A value of 1 in the NDVI data indicates a body of water, other values indicate the vigour of vegetation.

Examples of the type of satellite data we use are shown in Figure 1. Panel (a) depicts a Normalized Difference Vegetation Index (NDVI) map of Uganda derived from the National Oceanic and Atmospheric Administration's (NOAA) Advanced Very High Resolution Radiometer (AVHRR). The AVHRR is an example of a meteorological satellite that collects data about the earth's land cover types and conditions, cloud cover patterns, sea surface temperature etc. The AVHRR collects this data across five spectral bands of the electromagnetic spectrum, from which derivatives such as NDVI and rainfall estimates (Figure 1(b)) can be extracted. It also has a temporal resolution of 12 hours, thus making it possible to derive information about any given location on the surface of the globe at least twice a day. In spite of the coarse resolution of 8km, it has the advantage of wide area coverage.

Our goal is to be able to make predictions of famine which go beyond a blanket warning for a given region. In heterogenous populations, while some sectors of the population may be at risk of famine in a certain area, other sectors may not. We therefore try to split up the households in the country into defined categories, with a particular focus on category definitions which are easy to communicate so that the results of predictions can be easily translated into policy. Most con-

ACM DEV'10, December 17–18, 2010, London, United Kingdom.

ventional clustering methods such as k-means give cluster regions for which the decision boundaries may be somewhat complex, e.g. a set of inequalities based on some distance measure. Hand-chosen household categories may be easier to communicate, but may not be optimal with respect to the specificity of predictions based on them (e.g. if two hand-crafted categories are highly correlated). In section 4, we introduce a methodology for finding clusters of households which are informative for famine prediction, and which are also easy to communicate. Our clusters are defined by a binary tree, which we optimise using a simulated annealing strategy.

We then look at the use of supervised learning to give warnings of famine. Learning the relationship between vegetation stress and food insecurity is not straightforward, as high density of vegetation in a region does not necessarily mean there is more to eat. The relationship between the different demographic indicators such as production,household income and available labour is also not straightforward, and it is important to establish the relationship between the different variables used to detect famine [7].

Our results, in section 5, show the improvement in accuracy and specificity which can be attained by combining satellite and demographic data, and the utility of our clustering method.

2. RELATED WORK

Sub-Saharan Africa is a region heavily reliant on agriculture, and monitoring trends in agricultural production has long been imperative. Application of remote sensing in famine early warning systems has been used since the mid-1980s to monitor the crop and rangelands of semi-arid sub-Saharan Africa [4, 5, 9]. Vegetation stress which is measured as vegetation index (VI) at different wave bands (in the red and NIR portions of the spectrum) is used to show the spatial and temporal variations of biophysical vegetation parameters and this has been applied previously in operational famine early warning system [1, 5].

International organizations including the United States Agency for International Development (USAID) have worked on famine early warning systems [11]. Their methods involve the use of climate monitoring and weather forecasting results in comparison with methods of food access to determine food security in an area, although without integrating both types of data into a single model.

Attempts have been made to integrate other sources of data like prices of major food crops in an area to satellite remote sensing data to improve on food insecurity prediction. Brown et al [1] used satellite remote sensing data in a spatially explicit price model to assess food insecurity of communities and regions in less-developed parts of the world. This model created a leading indicator of potential price movements for early warning of food insecurity indicating the importance of integration of other source of date to satellite remote sensing data. Other work has taken a different approach rather than making warnings of crop failure directly; for example Khan et al [6] carried out work to predict prices of major food grains and malnutrition rates from NDVI data.

3. SATELLITE AND HOUSEHOLD DATA

The NDVI and rainfall estimate data for the three study areas of interest were obtained from the Famine Early Warning System Network[1] through the Africa Data Dissemination Service to coincide with the available demographic data for two agricultural seasons July-December 2004 and January-June 2005. The rainfall estimate is a product of an algorithm developed by NOAA at the Climate Prediction Center for Rainfall Estimate known as the CPC-RFE which has been widely tested and applied in the African region [8]. It is a technique that combines satellite and surface based rainfall estimation. The CPC-RFE uses a merging technique that increases the accuracy of the rainfall estimates by reducing significant bias and random error compared with individual precipitation data sources [12], thereby adding value to rain gauge interpolations.

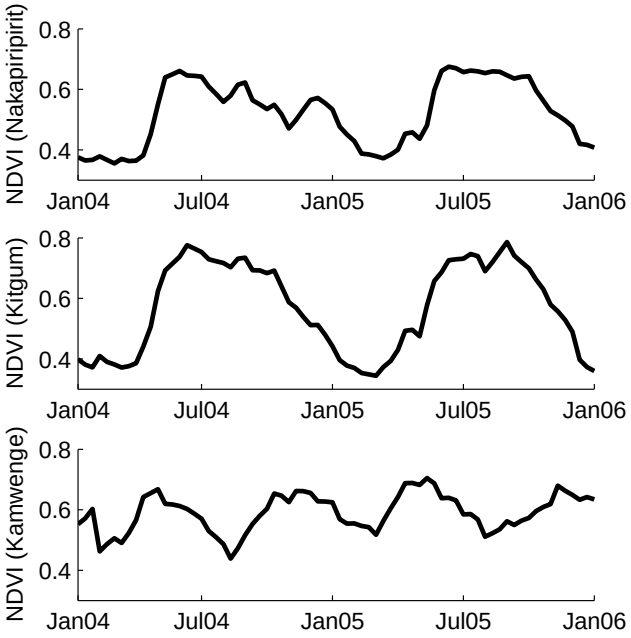


Figure 2: NDVI fluctuations according to seasons for three districts in the east, north, and south-west of Uganda respectively.

NDVI is based on the principle that actively growing green plants strongly absorb radiation in the visible region of the spectrum while strongly reflecting radiation in the near infrared region. It is calculated as

$$\text{NDVI} = \frac{\text{NIR} - \text{R}}{\text{NIR} + \text{R}} \tag{1}$$

where NIR is the intensity measured in the near infrared spectrum, and R is the intensity in the visible red spectrum. The formulation of NDVI makes it resilient to variations attributed to calibration, noise, and changing irradiance conditions that accompany changing sun angles, topography, clouds/shadow and atmospheric conditions.

Examples of NDVI fluctuations over time can be seen in Fig. 2, showing variation with wet and dry seasons in districts in different parts of Uganda. Here we use the mean NDVI across the extent of each district.

Demographic data for the same two agricultural seasons

[1] `http://www.fews.net`

Sex of the household head	male/female
Age of the household head	years
Marital status of the household head	married/ divorced/ single/ widowed
Size of household	number of people
Size of land available to the household for farming	acres
Amount of labour available for cultivation per year	person-years
Distance from household residence to the nearest main road	km
Distance from household residence to farm land	km
Total annual production of crops available for consumption by the household (excluding crops which are sold)	kg
Agricultural shock (e.g. presence of flooding, drought, market fluctuation)	true/false
Crops attacked by pests	true/false
Ownership of livestock	true/false
Household famine status (whether daily calorie intake per person in the household is above 1800 kCal)	famine/ not famine

Table 1: Variables in the famine dataset describing each household surveyed.

(July-December 2004 and January-June 2005) was collected by the Uganda Bureau of Statistics on households in 56 of the 80 districts across Uganda [10]. This data included the district of each household, the occupation, gender, marital status, education level and age of the household head, the household's exposure to agricultural shock (e.g. pest attacks), the distance from the home to a main road, and the distance from the home to their place of food cultivation. Finally the data included the agricultural production of each household and calorific consumption per person in the household. These variables are summarised in Table 1.

The raw data in the study had problems with consistency, for example with crop production for different households which was reported in varying units such as tins, kilogram and baskets. This was corrected as far as possible, and anomalous rows in the data were excluded.

4. HOUSEHOLD CLUSTERING TREES

Our goal is to make predictions of famine not just for all residents of a certain district, but for specific categories of households. In a population containing different types of households, at a specific point in time the fact that some households have a high risk does not mean this is true for all housholds.

In order to do this we have to group households together in some way. In this section we describe a novel clustering algorithm intended to produce interpretable cluster boundaries which are informative for famine prediction.

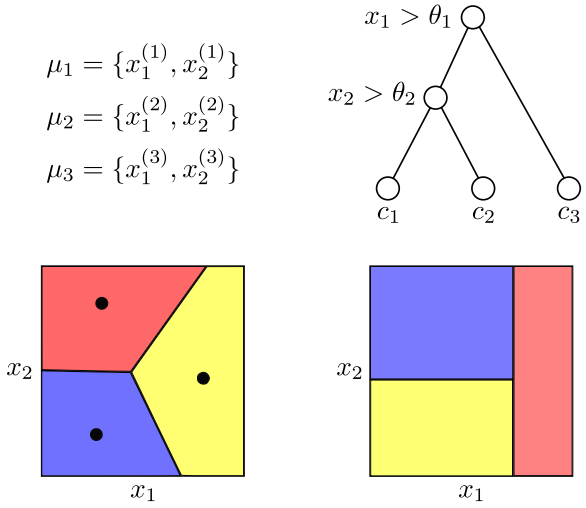


Figure 3: Prototype-based clusters (left) are defined by a set of points $\{\mu_i\}$, giving a Voronoi decision boundary in 2D. Defining clusters with a binary tree (right) gives cuboid regions with edges perpendicular to the axes.

Many well-known clustering algorithms are available, such as k-means, though we are constrained in this case by the form of the decision boundary after learning clusters. A prototype-based clustering algorithm gives clusters which may be difficult to communicate, being based on a distance measure and a set of inequalities. Clusters can alternatively be represented by a binary tree, as used in classification and regression trees (CART); this produces cuboid regions which can be expressed simply as a set of ranges on each descriptor variable. This is illustrated in Figure 3.

We now describe the structure of our clustering trees. Such a tree $\mathcal{T}$ is defined by a set of vertices V and vertex parameters Θ. The vertices are divided into leaf nodes and decision nodes, $V = V_L \cup V_D$. Each vertex has a set of parameters, $\Theta_v = \{c_v, i_v, \theta_v, e_v^-, e_v^+\}$, which are non-null under the following conditions:

$$\begin{array}{lrcll} v \in V_L: & c_v & \in & \{1,\ldots,k\} & \text{(Cluster index)} \\ v \in V_D: & i_v & \in & \{1,\ldots,V\} & \text{(Variable index)} \\ v \in V_D: & \theta_v & \in & \mathbb{R} & \text{(Threshold)} \\ v \in V_D: & e_v^- & \in & V/\{v \cup e_v^+\} & \text{(Left child)} \\ v \in V_D: & e_v^+ & \in & V/\{v \cup e_v^-\} & \text{(Right child)} \end{array}$$

Some vertex v_1 is assigned to be the root node, which must have no incoming edges (that is, e_v^- and e_v^+ are never equal to to v_1). For the clustering tree to be consistent, the graph structure defined by the edges implicit in Θ must contain a unique path from v_1 to every other vertex.

To assign a vector $\mathbf{x} = x_1, \ldots, x_V$ to a cluster index, we begin at the root vertex. If $c_1 \in \{1,\ldots,k\}$, then we assign that cluster index. Otherwise, we test if $x_{i_1} > \theta_1$. If this condition is satisfied we move to vertex e_1^+, otherwise we move to vertex e_1^-, and repeat the procedure.

4.1 Clustering evaluation

To evaluate a clustering tree $\mathcal{T}$ with respect to N rows of

input data, we need an evaluation metric $O(\mathcal{T}, \mathbf{x}_{1:N})$. It is common to evaluate clusters using some distance measure, for example preferring clusterings which minimise the average intra-cluster distance and maximise inter-cluster distance for a training dataset.

In this application, we are interested in finding clusters which allow us to make specific predictions of famine risk. Instead of applying a distance measure, we therefore look at the levels of correlation between clusters in terms of household food production in different areas.

Where each household in our training data has a district and a seasonal production, we can calculate the matrix $\mathbf{P}$, where $\mathbf{P}_{c,d}$ is the average production in the dth district for households in the cth cluster (as assigned by clustering tree $\mathcal{T}$). Our clustering metric is then calculated as follows:

$$\mathbf{C}_{ij} = \frac{\langle (\mathbf{P}_{i,:} - \langle \mathbf{P}_{j,:} \rangle)(\mathbf{P}_{i,:} - \langle \mathbf{P}_{j,:} \rangle) \rangle}{\langle \mathbf{P}_{i,:} - \langle \mathbf{P}_{i,:} \rangle \rangle \langle \mathbf{P}_{j,:} - \langle \mathbf{P}_{j,:} \rangle \rangle} \quad (2)$$

$$O(\mathcal{T}) = \frac{1}{k^2} \sum_{i=1}^{k} \sum_{j=1}^{k} \mathbf{C}_{ij} \quad (3)$$

where $\langle \cdot \rangle$ denotes an expectation. This metric gives us the average correlation between clusters in terms of production across different districts. Minimising $O(\mathcal{T})$ gives us clusters which we expect to be specific in terms of famine risk.

4.2 Stochastic search

We use a simulated annealing method to learn cluster descriptors with low values of $O(\mathcal{T})$. In this approach, we first initialise a clustering tree. We constrain the clusters so that each cluster must occupy a single cuboid region in the data space, which is easy to implement by having the number of leaf vertices equal to the number of clusters. For k clusters, we must have $|V_L| = k$ leaf vertices and therefore (in a directed tree) $|V_D| = k - 1$ decision vertices. We initialise $c_v, i_v, \theta_v, e_v^-, e_v^+$ in the decision vertices randomly, though fulfilling the tree structure constraint described above, and also with non-conflicting threshold values (that is, if vertex v_b is a descendent of v_a, and $i_a = i_b$, then $\theta_b < \theta_a$ if v_b is in the left subtree of v_a, and $\theta_b > \theta_a$ otherwise).

Candidate trees are generated by iteratively making modifications to the tree. The possible "moves" are:

- *Swap nodes*, taking any two vertices v_a, v_b not including the root vertex and swapping the parameters Θ_{v_a} and Θ_{v_b}. In this case, we have to check that the resulting graph is still a valid tree. This can be done easily, for example checking for cycles by testing that the eigenvalues of $A + I$ are all positive where A is the adjacency matrix of the tree and I is the identity matrix;
- *Change threshold*, where we alter the value of one of the thresholds θ_v, resampling according to some prior distribution and in such a way that the tree is still consistent;
- *Change variable* being considered at a certain decision vertex, i.e. altering the value of i_v for some v. Again we have to check whether the tree is consistent after making such a change.

Given a neighbouring tree $\mathcal{T}^*$ generated by one of these moves (made at random), we can evaluate the improvement in score and accept the change if:

$$O(\mathcal{T}^*, \mathbf{x}_{1:N}) - O(\mathcal{T}, \mathbf{x}_{1:N}) > r.T, \qquad r \sim U[0,1] \quad (4)$$

where $U[\cdot]$ denotes a uniform distribution. The "temperature" parameter T starts at a high level and is gradually reduced. A common cooling scheme is to use $T = \frac{c}{\iota}$ at iteration ι, which we use here with the constant $c = 1$. More complicated cooling schedules are also possible. The effect of this is to prevent a solution being trapped in a sub-optimal local minima.

Figure 5 shows an example clustering tree obtained after carrying out this cluster learning procedure for 1000 iterations. We find that distance to the road is an important factor, as it gives information on how much trade a household is engaged in. These cluster boundaries are easy to communicate. For example, if we wanted to raise the alarm for the cluster on the far right in Fig. 5, we simply communicate the alarm for "All households more than 92 percentiles away from the nearest main road, and more than 46 percentiles from their food-cultivation area".

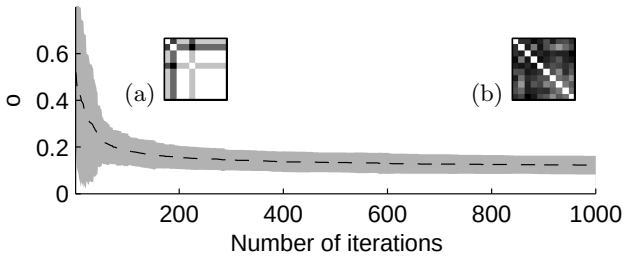


Figure 4: Objective function averaged over 25 trials, where the shaded area shows two standard deviations away from the mean. Insets show correlation matrices C (a) at initialisation and (b) after 1000 iterations, where lighter shades denote matrix entries closer to 1, and darker shades denote values closer to 0.

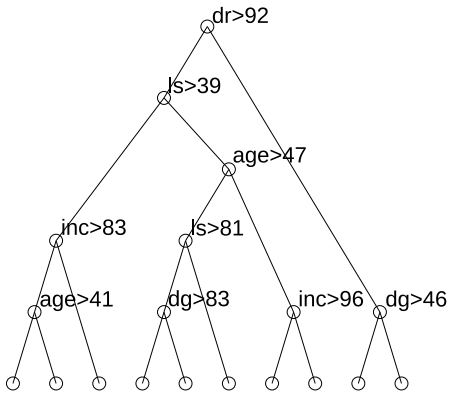


Figure 5: Example household clustering tree, corresponding to the final matrix C in Fig. 4(b). Household indicator variables are: distance to road (dr), land size (ls), age of household head (age), income (inc), distance from house to garden (dg). Figures are expressed in percentiles of the training data.

	AUC	Specificity
Satellite data only	0.632	0.417
Demographic data only	0.614	0.432
Satellite + Demographic	0.671	0.470
Satellite + Cluster IDs	0.656	0.465

Table 2: Results from using satellite and demographic data for prediction of famine. Specificity figures given for the threshold giving 50% sensitivity.

5. RESULTS

We used the preceeding data to make classifications of food security on a household level for the 3094 households in our dataset. For each household, we look at records of the calorific intake per person over two seasons, Q3/Q4 2004, and Q1/Q2 2005. We define food insecurity as a calorific intake of less than 1800 kcals/day, a level at which a family is vulnerable to famine.

To predict food insecurity, we use four sets of covariates: (a) demographic data only, (b) satellite data (NDVI and RFE) only, (c) satellite and demographic data, and (d) satellite data and household cluster IDs. We carried out training and classification with the AdaBoost algorithm [2] using decision stumps as a base classifier. Evaluation was done with 10-fold cross validation. The Weka framework [3] was used to carry out these experiments.

Table 5 gives AUC and specificity statistics (the latter corresponding to a classifier threshold giving 75% sensitivity). Figure 6 (upper panel) shows ROC curves for datasets (a-c). It can be seen that the combination of satellite and demographic data gives marginally more accuracy than either of the datasets alone.

Figure 6 (lower panel) shows the difference between ROC curves when using satellite data only for classification, and satellite data with household cluster IDs. Again, we see a marginal improvement in specificity and overall classification performance with the cluster IDs included. Although the margins of improved prediction accuracy are small, when dealing with large populations an incremental improvement may be of high significance.

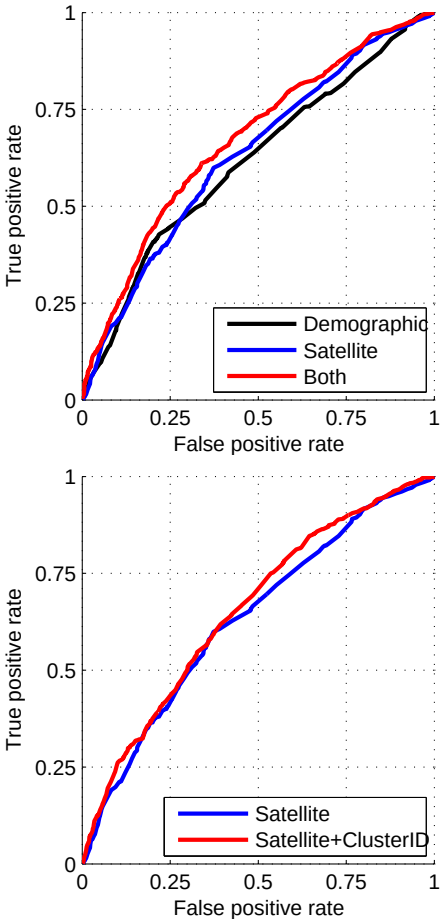


Figure 6: ROC curves for food security classification with different datasets.

6. CONCLUSION

We show that adding demographic information about households to satellite observation data gives better accuracy in making predictions at a household level. We have described a clustering method for this data which gives household categories that are easy to communicate, and which can be used as the basis of an improved satellite famine warning system to give famine risk alarms with increased specificity.

Promising ways to expand this work would include extracting features from the satellite data, i.e. learning measures other than NDVI and RFE from raw multi-spectral satellite images. The model could also be made explicitly spatial, if we assume that households which are geographically close together are likely to be correlated in terms of famine risk.

7. REFERENCES

[1] M. E. Brown, J. E. Pinzon, and S. D. Prince. Using Satellite Remote Sensing Data in a Spatially Explicit Price Model: Vegetation Dynamics and Millet Prices. *Land Economics*, 84(2):340–357, 2008.

[2] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[3] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA Data Mining Software: An Update. SIGKDD Explorations 11(1), 2009.

[4] J. U. Hielkema and F. Snijders. Operational use of environmental satellite remote sensing and satellite communications technology for global food security and locust control by FAO. *Acta Astronautica*, 32:603–616, 1994.

[5] C. F. Hutchinson. Use of satellite data for famine early warning in sub-Saharan Africa. *International Journal of Remote Sensing*, 12(6):1405–1421, 1991.

[6] M. M. Khan, N. Mock, and W. B. Bertand. Composite Indicators for Famine Early Warning Systems. *Disasters*, 16(3):195–206, 2007.

[7] E. Mwebaze, W. Okori, and J. Quinn. Causal structure learning for famine prediction. In *In*

Proceedings of the Association for the Advancement Artificial Intelligence (AAAI) Spring Symposium on AI-D, 2010.

[8] M. Shrestha, G. Artan, S. Bajracharya, and S. R.R. Using satellite-based rainfall estimates for streamflow modelling: Bagmati Basin. *Journal of Flood Risk Management*, 1:89–99, 2008.

[9] H. Silvia, R. Fensholt, K. Rasmussen, S. R. Proud, and A. Anyamba. Improving early warning systems for food security in Africa with geostationary earth observation data. *Earth and Environmental Science*, 6(47):472007, 2009.

[10] UBOS. Uganda National Household Survey 2005/2006, Report on the Agricultural Module. Technical report, Uganda Bureau of Statistics, 2007.

[11] USAID. Uganda Food Security Outlook, January to June 2010: Famine Early Warning Systems Network. Technical report, USAID, 2010.

[12] P. Xie and P. Arkin. Analyses of global monthly precipitation using gauge observations, satellite estimates and numerical model predictions. *Journal of Climate*, 9:840–858, 1996.