

GRID RAINFALL DISAGGREGATION TOWARD A PATCH-BASED ENSEMBLE KALMAN FILTER FOR SOIL MOISTURE DATA ASSIMILATION

Filip Jagodzinski, Villanova University, Bryn Mawr, PA
(filip.jagodzinski@villanova.edu)
Mentor: Dr. Xiwu Zhan, GEST / NASA GSFC, Code 974.1

***Abstract.** Data assimilation is the process by which observation data is used in model simulations and predictions. The Ensemble Kalman Filter (EnKF) is one such assimilation implementation, and efforts are underway to develop an EnKF to assimilate soil moisture data for use in the Land Data Assimilation Schemes (LDAS), developed in part at the Hydrological Sciences Branch at NASA's Goddard Space Flight Center. The EnKF propagates a set of ensemble members, the optimal characteristics and derivation of which are not yet known. Further information is sought about efficient ensemble size, the extent of ensemble member uniqueness and the physical meaning of individual ensemble members. Recent efforts have focused on the development of a rainfall disaggregation technique that aims to further increase the resolution of available precipitation data so that more efficient patch-based ensemble members can be generated. For the entire year of 1999, hourly precipitation data at 1/8th degree resolution was produced for the continuous 48 United States. Linear regression using the elevation, precipitation and land-surface type data was used to derive several elevation-precipitation relationships. These elevation-precipitation relationships, which vary on daily time-scales, can provide insight into the generation of optimal ensemble members for use with the EnKF.*

INTRODUCTION

Data Assimilation

Data assimilation is an analysis procedure by which observed values of scientific data and short-range forecasts from an earlier model run are combined to produce estimates of the initial conditions used to begin a new forecast. Observed data values may be clustered or sparse, resulting from variability in the time scales between individual measurements, variability in different rates of measurements at different locations, variability in instrument sensitivity and accuracy, and variability between different observing systems, several of which may be used to produce a complete dataset. The variability in datasets and individual data points of a single dataset is often difficult to account for; interpolation and analysis techniques are often used to help account for these discrepancies.

To ensure that the current forecast is as accurate as possible, data assimilation makes use of all available data and uses shorter-range forecasts to help "smooth" the ensuing forecast step. Because different observing systems are often used to produce a dataset—including satellites, ground-based measurements, radar, and aircraft flybys—it is necessary to incorporate these different data values into the forecast stage. Likewise, the short-range forecasts, because they tend to be more accurate because they are extrapolations into the near future only, are often used to make minute corrections to the overall, final, often long-range forecast [1].

In data assimilation, a model is used to make a series of short-range forecasts, and new observations contribute to the forecast as these observations become available. The new observations are introduced into the model in one of several ways, including by

periodic re-analysis, gradual insertion, and by mathematical blending. These insertion methods themselves tend to be complicated and are often very computationally expensive. The ultimate goal of data assimilation, thus, is to produce a model that tends to agree closely with the incoming observations, at which time the final long-range forecast is said to have a high confidence level [2].

The Kalman Filter

The Kalman Filter, introduced by Rudolf Emil Kalman in 1960, is one such example of a data assimilation procedure; it includes a set of mathematical equations that provide recursive solutions of a least-squares analysis method. The Kalman Filter supports estimations of past, present, and future states, and involves the combining of all available measurement data, plus prior knowledge about the system and measuring devices, to produce an estimate of the desired variables so that measurement and forecast errors are minimized statistically. The Kalman filter involves two main steps, an assimilation and a forecast step. The assimilation step involves the use of available new measurements and the inclusion of these measurements so that the projected estimates are improved, while the forecast step involves the projection of the current state estimate ahead in time [3,4].

The goal of the Kalman Filter is to find an equation that computes an *a posteriori* state estimate $x(k)$ as a linear combination of an *a priori* estimate $\hat{x}(k)$ and a weighted difference between an actual measurement $z(k)$ and a measurement prediction $H(\hat{x})$. Over the years, several variations of the Kalman Filter have been developed and introduced, including the Discrete Kalman Filter (KF), the Extended Kalman Filter (EKF), and the Ensemble Kalman Filter (EnKF).

Mathematically, the Kalman Filter (KF) is defined by the following 4 equations, noting that there different implementations of

the Kalman Filter may denote the required matrices and equation variables in different ways:

$$\begin{aligned} S_k &= P_k + R \\ K_k &= AP_k S_k^{-1} \\ P_{k+1} &= AP_k A^T + Q - AP_k S_k^{-1} P_k A^T \\ X(\hat{hat}) &= Ax(\hat{hat})_k + K_k (z_{k+1} - Ax(\hat{hat})_k) \end{aligned}$$

where superscript -1 indicates matrix inversion

superscript T indicates matrix transposition
subscript k indicates a current state and represents a time step

subscript $k+1$ indicates a estimate of a future state and represents a time step

R = covariance matrix

S = covariance of the innovation

K = gain matrix

Q = covariance of the white Gaussian noise

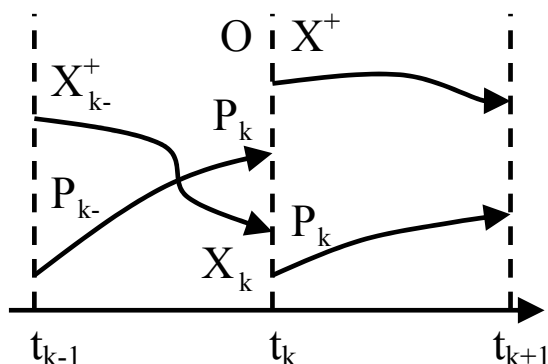
P = prediction error covariance

x = state estimate

A = is a multiplication factor.

Likewise, the EKF and the EnKF employ similar equations, but the degree of accuracy of the forecast step and the degree to which the state error covariance and measurement covariance errors are utilized increases. Depending on the required accuracy of the forecast step and the computational resources, different Kalman Filters can be applied. For example, the EnKF circumvents the expensive integration of the state error covariance matrix by propagating an ensemble of states from which the covariance is derived. Using the EnKF, the gain matrix, K , is of the order of 10-100, so the computational cost of the entire data assimilation process is increased dramatically, but this is still far less than the increase in cost of using the EKF, which requires a computation cost increase of the order of the number of degrees of freedom of the model. Graphically, the difference between the EKF and the EnKF is shown in Figure 1.

The EKF: integrate state estimate x and error covariance P



The EnKF: integrate ensemble of states and compute sample covariance P

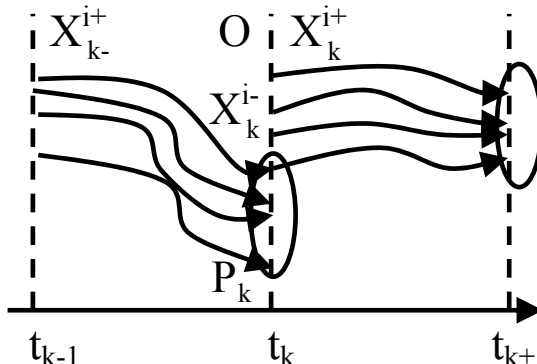


Figure 1. A comparison of the EKF and the EnKF, where t_{k-1} , t_k , t_{k+1} indicate past, present, and future states, respectively; where O indicates a new observation, where X is a state estimate, and P is the prediction error covariance. Note that at time t_k , when a new observation is attained, the state estimate is adjusted to account for the new observation data.

The Land Data Assimilation Systems (LDAS)

The Land Data Assimilation Systems, developed in part at the Hydrological Sciences Branch at NASA's Goddard Space Flight Center, aim to provide forecast simulations that will lead to more accurate reanalysis and simulations by numerical weather prediction (NWP) models. Presently, there are two main systems: North American Land Data Assimilation Schemes (NLDAS) and the Global Land Data Assimilation Schemes (GLDAS). NLDAS aggregates available observations into various land surface models to estimate land surface water and energy fluxes, covers all of the 48 contiguous United States and parts of Canada and Mexico, and utilizes several surface models, including MOSAIC, CLM, and NOAH.

The goal of the systems is to reduce the errors in the stores of soil moisture and energy which are often present in NWP models and which degrade the accuracy of forecasts. NLDAS currently runs retrospectively on a 1/8th-degree grid while

GLDAS runs at a 1/4th-degree resolution. The two systems are forced by terrestrial (NLDAS) and space-based (GLDAS) precipitation data, space-based radiation data and numerical model outputs [5].

MOTIVATION

This summer project was part of the long-term objective to develop a soil-moisture data assimilation algorithm that utilizes the EnKF so that land surface water and energy flux predictions can be more accurate. The overall goal is to use the EnKF to generate many different ensembles so that land surface models can be improved. In this project, ensemble generation is based on subgrid vegetation patches and precipitation amounts. However, the current available precipitation data is at a too-low a resolution to produce optimal ensemble members for use in the EnKF, and hence one obstacle to the development of an EnKF algorithm in use with the LDAS is that of precipitation disaggregation. Elevation and land cover data is available at a much higher resolution,

and it is hoped that the precipitation data can be disaggregated to sub-patches which correspond to the size of the elevation and land-cover grids. An elevation-precipitation relationship is desired so that the elevation data can be used to help disaggregate the precipitation data and so that ensembles for use in the EnKF can be quickly computed. The elevation-precipitation relationship can be used to disaggregate grid scale rainfall to each patch according to its cover type and average elevation. In this experiment it is assumed that the rainfall for each sub-grid tile or patch depends on its average elevation as well as dominant land cover-type.

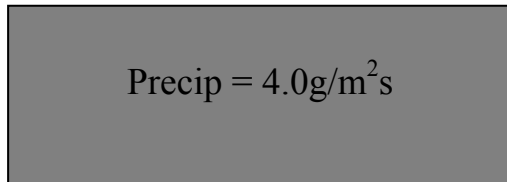
METHODS

NLDAS was run using the MOSAIC model at 1/8th degree resolution to produce grid precipitation data for each hour for all days in 1999—a total of 78 million precipitation “events”. When run at 1/8th degree resolution, the NLDAS domain is a 464 by 24 grid, comprising a total of 103,939 grid “cells”. For each hour, NLDAS was made to output a latitude longitude coordinate—down to the 1/8-degree resolution—and the corresponding precipitation amount for that location. For the sake of computational resources and output file size, when no precipitation was measured at a location for a certain hour, the particular longitude latitude location for that particular hour was left out of the output precipitation data. Fortran 90 and Fortran 95 were used to create all precipitation disaggregation routines.

Dr. Matt Rodell, also of the HSB, supplied elevation and surface type data corresponding to the NLDAS domain. Because each of the 1/8th-degree by 1/8th-degree grid precipitation cells output by NLDAS can have more than one surface-type, each of which can process precipitation in different ways, it was necessary to know how many of each surface-type was contained in each of the NLDAS precipitation output cells. The surface types used were those derived by the University of Maryland's (UMD) vegetation classification scheme (see appendix 1). Figure 2 shows graphically a representation of the resolution of the NLDAS hourly precipitation data and the corresponding resolution of the elevation and land cover data. Precipitation disaggregation involves the use of the elevation and land cover data to increase the resolution of the precipitation data.

Latitude, longitude coordinates were used to pair up the precipitation data output by NLDAS and the elevation/surface-type data provided by Dr. Rodell. It was assumed that each elevation/surface-type subpatch within each 1/8th-degree by 1/8th-degree precipitation grid cell received the same amount of precipitation as the precipitation value that was attained for the entire cell. Because the latitude and longitude coordinates were used only to match-up the elevation/surface-type data to the precipitation data, the coordinate information was left out of the final elevation/precipitation dataset. As an illustration, a section of the precipitation-elevation-surface-type output data is shown in Table 1.

NLDAS Hourly Precip. Data
At low resolution



Elev. & Land Cover Data
At high resolution

Elev=412	Elev=360	Elev=56
Elev=280	Elev=102	Elev=50
Elev=122	Elev=87	Elev=22

Figure 2. The NLDAS Hourly Precipitation Data and Elevation/Land-Cover Data, both at different resolutions. The different colors in the elevation and land-cover data represent one of 13 different land surface types, as defined by UMDs vegetation classification scheme (Appendix 1), while Elev (elevation) is in meters.

Table 1. Precipitation-elevation-surface type data for 5 January 1999, from 02:00-03:00 hours.

Land Surface Type	Elevation (m)	Precipitation (kg/m ² s)
2	3289	0.33300E-06
5	2482	0.33300E-06
8	3356	0.10500E-04
11	3442	0.17000E-04
7	3039	0.17000E-04

Note that the precipitation for the first 2 rows and the last 2 rows in Table 1 is identical, indicating that those cells are disaggregate precipitation data—the elevation/surface-type data available for that latitude, longitude NLDAS precipitation cell included several surface-types at varying elevations.

Graphically, the pairing up of the elevation/surface-type data with the precipitation data produces the results in Figure 3, which is a synthesis of the two data grids in Figure 2. Using the matched up precipitation and elevation/cover-type data, first order linear regression was performed to determine a precipitation-elevation relationship. The precipitation-elevation relationship was investigated on several time and physical scales, including relationships on the scales of hours, days, months, seasons, etc.

It was observed that on any one day there were a biased amount of data points for

certain elevation bands (see RESULTS), thus a first-order linear regression analysis was also performed using only certain elevation data points, as for example using only those precipitation events that occurred at elevations above 500m.

Elev=412m Precip = 4g/m ² s	Elev=360m Precip = 4g/m ² s	Elev=56m Precip = 4g/m ² s
Elev=280m Precip = 4g/m ² s	Elev=102m Precip = 4g/m ² s	Elev=50m Precip = 4g/m ² s
Elev=122m Precip = 4g/m ² s	Elev=87m Precip = 4g/m ² s	Elev=22m Precip = 4g/m ² s

Figure 3. Paired elevation/cover-type and precipitation data, where Elev = elevation, Precip = precipitation.

RESULTS

The results of the first-order linear regression analysis indicate that an elevation-precipitation relationship has variability on the order of hours. Tables 2-4 show the regression analysis results for the entire year

of 1999, 2 January 1999 and 3 January 1999. Results for all months, as well as for all hours on certain days were also obtained, but which are not shown here. For all results, analogous regression analysis runs were performed using only precipitation events with unique elevation points.

Table 2. Regression analysis results for the entire 1999 year.

Cover	# data	Corr Coeff	Reg. Coeff	Reg. Const
2	1468310	-0.0031	-0.139	0.75
5	596823	-0.0308	-5.494	2.242
6	7214855	-0.0974	-9.388	2.376
7	14442226	-0.068	-3.565	1.52
8	12386736	-0.0876	-4.947	1.829
9	945175	0.0236	0.996	0.688
10	2204446	-0.0475	-2.759	1.609
11	10067548	-0.0788	-3.598	1.81
12	21140163	-0.1101	-9.072	2.63
13	1290812	-0.1217	-3.367	1.492
14	1569946	-0.053	-9.623	2.544

Table 3. Regression analysis results for 2 January 1999

Cover	# data	Corr Coeff	Reg. Coeff	Reg. Const
2	2590	-0.0329	-0.302	0.259
5	3831	-0.0522	-5.874	2.111
6	29167	-0.1017	-9.863	2.535
7	56958	-0.2172	-14.774	2.305
8	81669	-0.1447	-14.566	2.702
9	4659	-0.1397	-1.816	0.285
10	1591	0.0558	0.324	0.218
11	16191	-0.2045	-7.415	2.12
12	155204	-0.1202	-17.762	3.577
13	1130	-0.5682	-11.957	2.904
14	11219	-0.0651	-12.826	2.883

Table 4. Regression analysis results for 3 January 1999

Cover	# data	Corr Coeff	Reg. Coeff	Reg. Const
2	5288	-0.1474	-2.643	0.552
5	3708	0.1034	21.32	1.677
6	29854	-0.2094	-36.888	5.278
7	52653	-0.1665	-18.552	2.697
8	64439	-0.1022	-16.48	2.962
9	2486	0.1224	3.106	0.583
10	1479	-0.1807	-4.42	1.457
11	10750	-0.2137	-12.616	3.378
12	46963	-0.2274	-33.701	4.293
13	1500	-0.301	-3.626	1.042
14	9466	-0.2051	-117.33	7.532

DISCUSSION

As can be seen from Tables 2-4, the elevation precipitation correlation varies greatly between surface land types, as different land-cover types process precipitation in unique ways. Note that not all land-cover types were used in each regression analysis, as there were instances where there were not enough precipitation events to arrive at statistically significant linear regression results. However, the elevation-precipitation relationship of the same land cover types also varies between days, and likewise similar variations are evident on the scale of hours. This variation of the elevation-precipitation relationship suggests that an elevation-precipitation relationship is very weak or possibly non-existent, or that other factors must be considered if a strong-enough elevation-precipitation relationship is to be found.

As mentioned in the methods section, analogous regression analysis was performed on the data using precipitation events with elevation points above a certain elevation. This decision was made in response to Figures 4 and 5, which help illustrate the variation in sampling.

Figure 4 illustrates a complication that arose when the regression analysis was performed. According to Figure 4, the majority of precipitation events for 2 January 1999 occur at elevation below 500m. Deriving an elevation-precipitation relationship for the entire elevation range of 0 to 3500m would bias the relationship towards precipitation points that correspond to elevations bellows 500m. For this reason, an elevation-precipitation relationship was derived for 2 separate ranges, [0-499m] and [500+m]. Unfortunately, the 2 elevation-precipitation relationships from these 2 separate elevation "regions" also demonstrated variability on an hourly scale.

Figure 5 further demonstrates the difficulty of deriving an elevation-precipitation relationship, especially in cases where sampling and data densities vary. An elevation-precipitation relationship derived from the data points in Figure 5 would be primarily based on elevation precipitation data between 100 and 600m and between 1200 and 2100m in elevation. As in Figure 4 1, it is obvious that an elevation-precipitation relationship should account for the fact that different elevation bands will have different precipitation relationships.

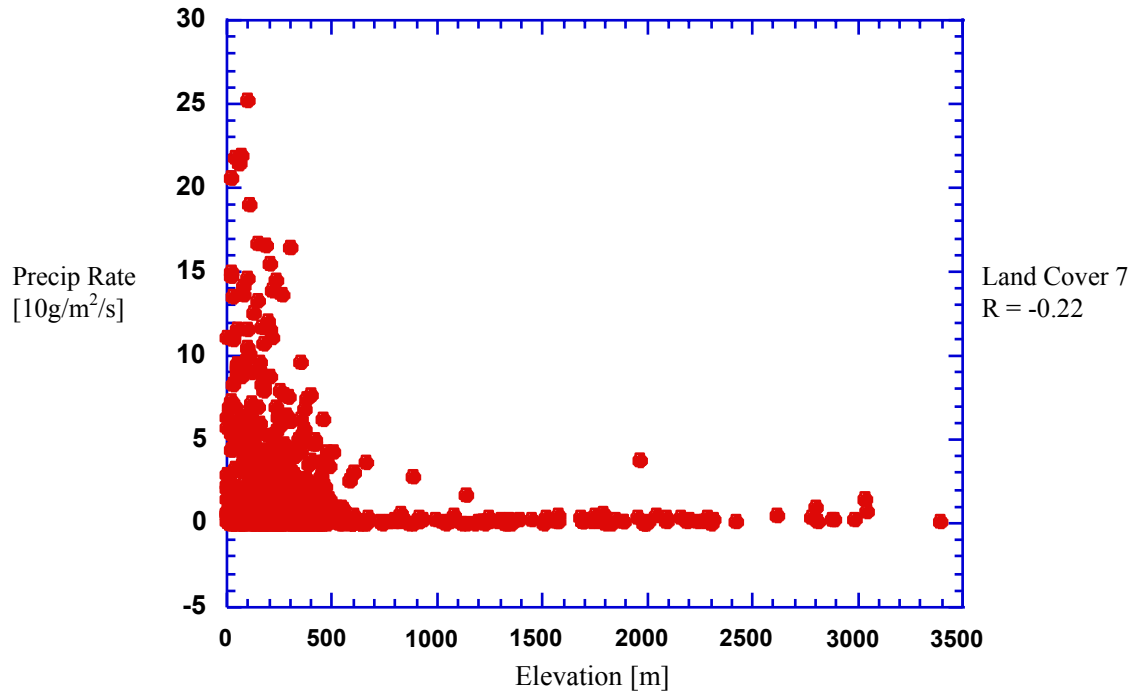


Figure 4. Elevation-precipitation data for 2 January 1999, land cover 7.

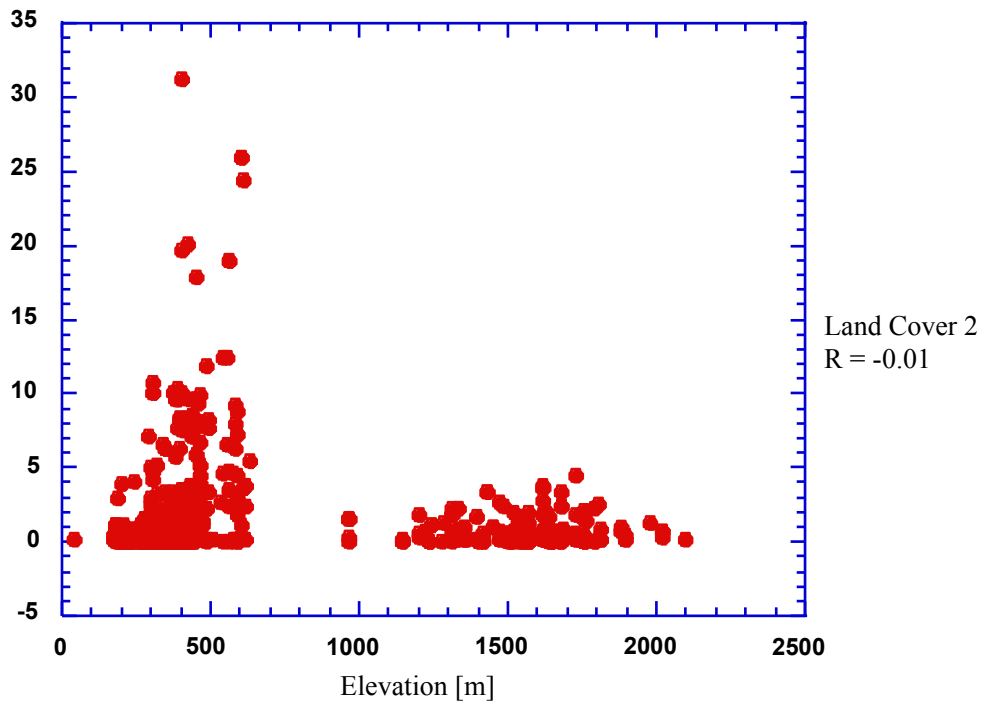


Figure 5. Elevation-precipitation data for 2 July 1999; land cover 2.

CONCLUSIONS

The elevation-precipitation relationship is thus a complicated quantity to ascertain, if at all it exists. Elevation-precipitation relationship variations exist on daily time scales, and further complications arise from the fact that different land surface types processes precipitation differently. As seen from the correlation statistics and from the plots of the elevation-precipitation data pairs, any elevation-precipitation relationship should take into account the fact that different elevation bands may manifest different precipitation relationships. This leads to the necessary consideration of further factors that may affect an elevation-precipitation relationship. Land-cover type and elevation may not be enough; for example slope and amount of sunshine received over the course of a standard time interval might also need to be taken into account. Additionally, the output of NLDAS should be more closely investigated as perhaps certain numerical assumptions were made when LDAS was being developed. Any "averaging" or "approximating" of data values that is done during the course of the running of the LDAS might affect the ensuing elevation-precipitation correlation studies.

A precise, multiple-variable elevation-precipitation relationship for use in the quick production of ensemble members for the EnKF is desired, although the choice of variables and parameters to use in deriving the relationships is not clear.

ACKNOWLEDGEMENTS

This Grid Rainfall Disaggregation research project was conducted at Goddard

Space Flight Center in Greenbelt, Maryland, from 9 June – 15 August, 2003 as part of the Graduate Student Summer Program, in part funded by The Goddard Earth Science and Technology Center (GEST) and NASA Goddard Space Flight Center. Dr. Xiwu Zhan and Dr. Paul Houser of the Hydrological Sciences Branch (HSB) were the primary mentors and provided the majority of the suggestions and project direction.

REFERENCES

1. F. Rabier, J.-N. Thepaut, P. Courtier, *Extended Assimilation and Forecast Experiments with a Four-Dimensional Variational Assimilation System*, Quarterly Journal of Research of the Meteorological Society, 124, 1861-1887, 1998.
2. Fred Carr, *Objective Analysis and Data Assimilation*, presented at the COMET COMAP Symposium, 27-31 March 2000.
3. R.G. Brown, P.Y.C. Hwang, Introduction to Random Signals and Applied Kalman Filtering, 2nd edition, John Wiley and Sons, New York, 1992.
4. Greg Welch, Gary Bishop, An Introduction to the Kalman Filter, workshop presented at the International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH), Los Angeles, CA, 12-17 August 2001.
5. <http://ldas.gsfc.nasa.gov>
6. <http://ldas.gsfc.nasa.gov/LDAS8th/MAPPED.VEG/LDASmapveg.shtml>

APPENDIX 1

UMD Vegetation Classification Scheme [6].

UMD Vegetation Category	Description
0	Water / Goode's Interrupted Space
1	Evergreen Needleleaf Forest
2	Evergreen Broadleaf Forest
3	Deciduous Needleleaf Forest
4	Deciduous Broadleaf Forest
5	Mixed Cover
6	Woodland
7	Wooded Grassland
8	Closed Shrubland
9	Open Shrubland
10	Grassland
11	Cropland
12	Bare Ground
13	Urban and Built-up