DECLOUDING TIME SERIES OF LANDSAT DATA

LUIS MARCELO TAVARES DE CARVALHO

UFLA - Universidade Federal de Lavras
P.O. Box 37 - 37200-000 - Lavras - MG, Brazil
passarinho@ufla.br

Abstract. Novel schemes based on multiresolution transforms were introduced to pre-process long time series of Landsat data. Particularly, removal of clouds and their shadows was tackled. We applied the product of wavelet scales to generate binary masks of corrupted observations, the robust smoother-cleaner wavelets to remove outliers in the data, and the wavelet shrinkage to estimate new values. Cloud contamination was simulated and the missing values were estimated using five methods: 1) mean value, 2) minimum value, 3) maximum value, 4) linear regression, and 5) the wavelet-based procedure. The product of wavelet scales not only identified clouded and shadowed pixels but also other anomalies like misregistration effects and changes of short duration (*e.g.*, burn scars). The wavelet-based approach was more accurate for interpolating the missing values in clouded areas, whereas linear regression performed better in shadowed areas. The robust non-linear wavelet regression holds promise for effective time series analysis and has the potential to produce noise-reduced images at any point in the time series.

Keywords: remote sensing, image processing, time series, cloud removal.

1. Introduction

The time series provided by Earth observation systems are important sources of information. In an analogy to hyperspectral imagery, we can already use hypertemporal data sets to analyse and model the environment. The Landsat system, for instance, has been acquiring images almost weekly since 1971 and producing valuable inputs for historical characterisations, predictive modelling, decision-making etc. Other systems, like the NOAA AVHRR, acquire images on a daily basis revealing detailed temporal information about the Earth's surface. Nevertheless, it is still difficult to perform spatio-temporal analysis of long time series acquired by such systems, in special Landsat, because of cloud contamination and other distortions. Image analysis techniques to support processing and information extraction from existing temporal data must be developed. Long term studies on land surface, water, carbon and energy fluxes require corrected time series to provide more realistic temporal parameters (Los *et al.* 2000). In this context, our task is to reconstruct as close as possible an estimate of the observations, which have been corrupted (*e.g.*, obscured by clouds). Then, past records might be effectively used to provide as important information as upcoming data of new Earth observation systems.

The general problem of detecting and estimating corrupted values has been tackled before with wavelet transforms (Starck *et al.*, 1998). The technique is particularly appealing to study signals that are "smooth" in some sense and have singularities of short duration. Despite sudden changes in land cover, the smoothness requirement is normally met by remotely sensed time series, where clouds, shadows, and other anomalies appear as narrow peaks in the otherwise smooth temporal profile.

This paper describes the application of the product of wavelet scales (Sandler and Swami, 1999) to generate binary masks of corrupted observations. The robust smoother-cleaner wavelets method (Bruce *et al.*, 1994) is then applied to each temporal profile where anomalous values were detected. The interpolation step is based on non-parametric function estimation applying wavelet shrinkage (Donoho and Johnstone, 1994) to the "clean" time series. The results were compared to other methods applied to the same synthetic data set.

2. Dealing with cloud contamination in remote sensing

An operational procedure for automatic cloud detection in NOAA AVHRR imagery was developed by Saunders and Kriebel (1988) and Derrien *et al.* (1993). Their pioneering algorithm uses different threshold tests applied to various combinations of channels. Pixels are identified as cloudy if one test is successful. The development was based on scenes from Western Europe and the authors highlight the necessity of tuning the algorithm if other regions are to be processed. Wang *et al.* (1999) proposed automatic cloud detection in a set of two temporal Landsat TM images by simply thresholding the differences. For shadow detection they thresholded high frequency components as extracted by a 2D discrete wavelet transform of both images. Roerink *et al.* (2000) developed the Harmonic ANalysis of Time Series (HANTS) and reported considerable improvements over the standard Fourier transform. The algorithm is based on an iterative procedure of least squares fitting based on harmonic components. Recursively, the outliers are removed and the curve fitting recomputed until it reaches a maximum acceptable error or a minimum number of remaining points. In this way, clouds and shadows are automatically detected as outliers in the temporal profile.

Once marked, the contaminated pixels are replaced to yield cloud free products. Long *et al.* (1999) compared several methods for cloud replacement in images from the Special Sensor Microwave/Imager (SSM/I) radiometer. They found that the combination of maximum value and mean methods produces better results than any of the approaches individually. Addink and Stein (1999) proposed new approaches to replace clouded pixels based on geostatistics. They concluded that instead of conventional methods unstratified co-kriging using another temporal observation as co-variable should be used to replace clouds from NOAA AVHRR images. Nevertheless, they also pointed out that the quality of the images used as co-variables must be very good; otherwise, unstratified kriging is the better option. In the algorithm of Wang *et al.* (1999) mentioned before, they recover the missing data by fusing the two images with the inverse wavelet transform. Binary decision maps of clouds and shadows are used to mask contaminated pixels during the inverse transform. Thus, masked pixels are reconstructed using complementary information, which must be available in at least one of the images. The HANTS algorithm is the only procedure listed above which is able to deal with long time series in an automatic fashion. It replaces all instances of temporal profiles by the respective values of the fitted curves, enabling the estimation of cloud free images at any moment in the time series. Similarly, Los *et al.* (2000) developed a corrected data set (FASIR NDVI), which also includes steps of Fourier adjustments, interpolation, and reconstruction of cloud free NDVI (Normalised Difference Vegetation Index) time series derived from NOAA imagery.

3. Automatic detection of cloud contaminated pixels

Wavelet analysis transforms a given signal (*e.g.*, **Figure 1a**) to a set of resolution related views, which are by definition of zero mean with variance σ^2 (**Figure 1b**, and **1c**). The discrete wavelet transform was implemented with the 'à trous' algorithm (Holschneider *et al.*, 1989) with a linear spline as the wavelet prototype. It produces a vector of wavelet coefficients *d* at each scale *j*, with *j* = 0,…, *J*. The original function *f* was then expressed as the sum of all wavelet scales and the smoothed version a$_J$:

$$f(t) = a_J(t) + \sum_{j=1}^{J} d_j(t) \, . \qquad (1)$$

In order to further enhance singularities (*i.e.*, edges and peaks) in the signal, interscale correlation was explored by forming multiscale point-wise products (**Figure 2**). The multiscale product (Sadler and Swami, 1999) of an arbitrary set *K* of wavelet scales is given by:

$$p(t) = \prod_{j \in K} d_j(t) \,. \tag{2}$$

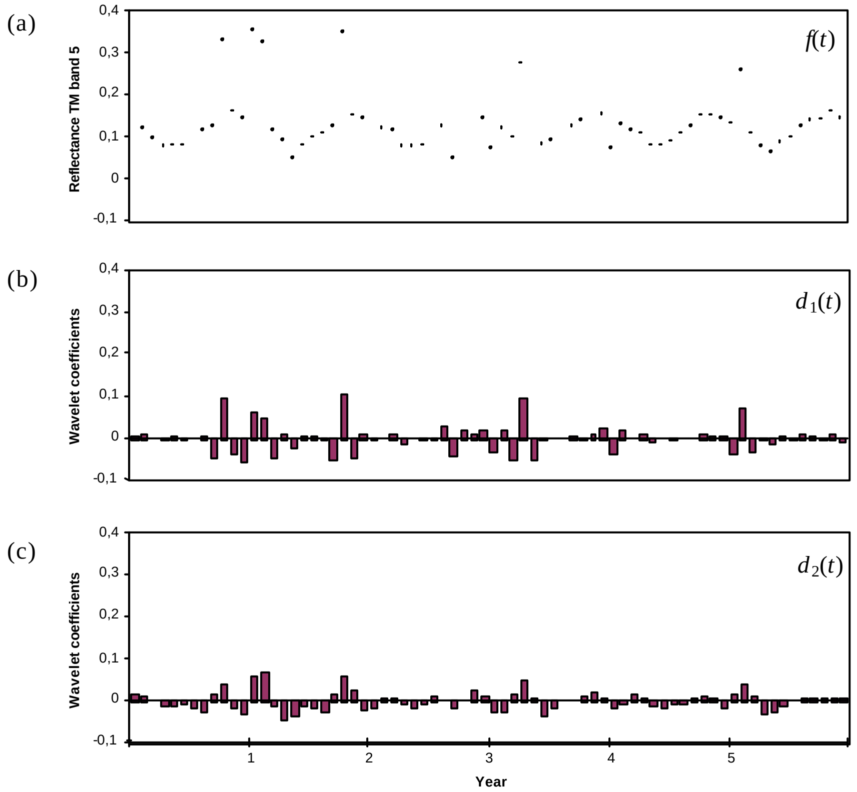

Figure 1. (a) Simulated temporal profile with cloud contamination. (b) Wavelet coefficients at first scale level. (c) Wavelet coefficients at second scale level.

The detection hypothesis is that the signal is locally constant around $p(t)$. Then, a simple approach to model noise that follows an unknown distribution is to consider it locally Gaussian. If $s_j(t)$ is the local standard deviation within the support of the wavelet template at scale j, we have a significant singularity when $p(t) > C s_j(t)$. In this study, we used wavelet scales d_1 and d_2 in the multiscale product and the constant C was empirically set to 2. Increasing the detection level might avoid false detection, but also excludes weaker singularities.

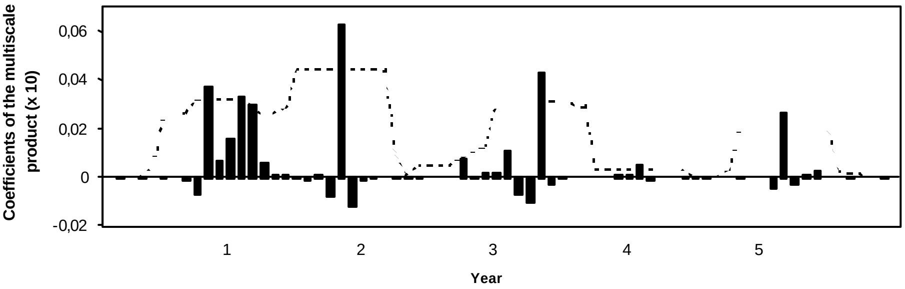

Figure 2. Product of wavelet coefficients (bars) and the detection limit (dashed line).

4. Robust non-linear wavelet regression

Recovering information obscured by clouds and shadows consists of estimating the missing data assuming that land cover varies smoothly over space and time. This is a fundamental requirement to the proper utilisation of non-parametric regression techniques. Then, wavelet methods might be applied to the problem of modelling a given digital signal *y* by estimating the unknown mean response function *f*:

$$y(t) = f(t) + \boldsymbol{e}, \tag{3}$$

where the vector $\boldsymbol{e}$ represents noise.

The wavelet approach to regression estimators brings a useful new set of basis for orthogonal series estimation, which allows characterisations of functions in terms of both time and frequency. The estimation is based on the representation of the mean response function as a linear combination of basis functions $\boldsymbol{y}_i$:

$$f(t) = \sum_i a_i \boldsymbol{y}_i(t), \tag{4}$$

where the coefficients a_i are given by:

$$a_i = \langle\, f(t), \boldsymbol{y}_i(t) \rangle. \tag{5}$$

The vector *a* of associated coefficients is called the transform of *f*. Thus, the problem of estimating *f* using the known vector *y* consists of three main steps:
(1) calculate the transform of *y*:

$$\tilde{a}_i = \langle\, y(t), \boldsymbol{y}_i(t) \rangle,$$

(2) select a subset *S* of important coefficients from $\tilde{a}$ in an attempt to remove the noise component $\boldsymbol{e}$ from the regression model (equation 3), and
(3) invert the transform using the selected coefficients to obtain the regression curve:

$$\hat{f}(t) = \hat{y}(t) - \boldsymbol{e} = \sum_{i \in S} \tilde{a}_i \boldsymbol{y}_i(t). \tag{6}$$

The subset *S* might be obtained with a suitable threshold function applied to the wavelet coefficients. In this study the so-called soft threshold (Donoho and Johnstone, 1994) was used:

$$\boldsymbol{d}_T(t) = \begin{cases} 0 & \text{if } |a_i| \le T \\ \operatorname{sign}[a_i](|a_i| - T) & \text{if } |a_i| > T \end{cases} \tag{7}$$

Note that the threshold value *T* may vary from level to level resulting in a locally adaptive function evaluation.

The robust smoother-cleaner wavelets (Bruce *et al.*, 1994) were used in the present study to reduce the sensitivity of regression smoothers to outliers. Let the input signal be $f = a_0$, which is first convolved with a median filter. The array of robust residuals *r* in every scale *j* is given by:

$$r_j = \boldsymbol{d}_T(a_j - a_j^*), \tag{8}$$

where, a^*_j is the median filtered version of a_j and $\boldsymbol{d}_T$ is a suitable threshold function, like in equation (7).

The cleaned version c_j is obtained by subtracting the residual r_j from the original vector a_j:

$$c_j = a_j - r_j . \tag{9}$$

5. Test Data and Validation

The reference cloud-free time series consisted of twenty-six co-registered subsets (256 x 256 pixels) of a Landsat TM scene (path 218, row 75). The images were acquired in varying intervals of time, from 1984 till 1999. In this study, we used TM band 5 images as input to the detection and replacement procedures. In a first simulation, real clouds and respective shadows were extracted from an image acquired in April 1989 that was not included in the time series, and placed over the same area of a cloud-free image acquired in June 1989, which was included in the time series. Thus, a simulated data set with cloud contamination was generated (**Figure 3**) in order to test the procedures for cloud detection and replacement. A second simulated time series was generated for a single forest pixel to illustrate the potential of wavelet regression. In this case, the following images were assumed to be representative of a one year cycle: Jan 1996, Mar 1988, Apr 1991, Jun 1989, Jul 1989, Aug 1991, Sep 1991, Oct 1985, and Nov 1985. This cycle was repeated four times and non-Gaussian noise was added, resulting in a 5-year time series (**Figure 1a**).

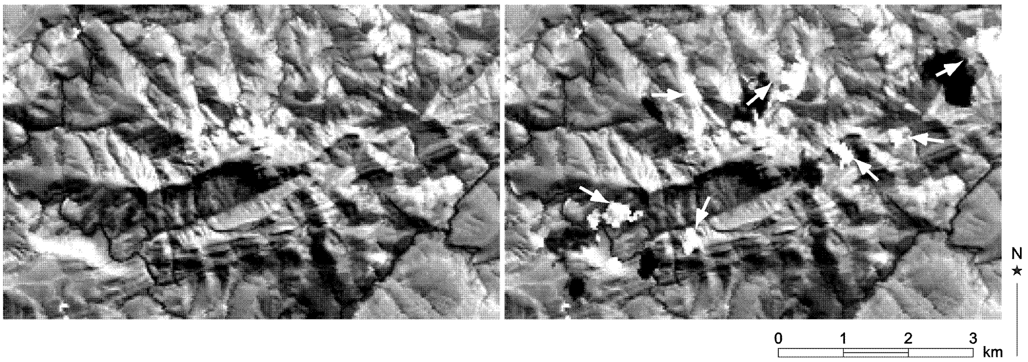

Figure 3. Band 5 of Landsat TM scene acquired in June 1989. Original (left) and simulated (right) images. White arrows indicate the locations of the added clouds.

The first simulated time series had the missing values estimated using five methods: 1) mean value, 2) minimum value, 3) maximum value, 4) linear regression, and 5) the wavelet-based procedure for non-parametric regression described above. Root mean square errors (RMSE) were calculated for the cloud-contaminated areas to evaluate the performance of each method in terms of accuracy of estimation:

$$\text{RMSE} = \frac{\sqrt{\sum_{i=1}^{N} (\hat{g}_i - g_i)^2}}{N} , \tag{10}$$

where, $\hat{g}_i$ is the ith estimated pixel, g_i is the ith original pixel, and N is the number of contaminated pixels.

6. Results and Discussion

Because of their high reflectance values, clouds have been normally detected by simply thresholding the original image or a difference image (Wang *et al.*, 1999). In the first case, other objects of high reflectance could be misdetected as clouds. In the second case, misdetection might occur in areas of significant land cover change. Moreover, the contaminated instance to be thresholded must be known in advance and the definition of proper thresholds for actual reflectance values could be difficult. Automatic cloud detection for NOAA imagery is described in Derrien *et al.* (1993). Their procedure is based on a series of tests and thresholds, which must be updated for different areas or illumination conditions. The advantage of the method described in this paper is that detection of outliers in multiscale product space is completely data driven and independent of the shape and magnitude of the original signal, avoiding false detection and aiding automation. In addition, other anomalies, like geometric misregistration, were also detected. Note in **Figure 4**, the clouds and shadows depicted with the multiscale product in 1989 and the misregistration effects depicted in 1992. Burned areas (white patch in the upper left corner of the image from 1992 in **Figure 4**) due to agricultural practices are susceptible to be misdetected as shadows and consequently removed from the time series.

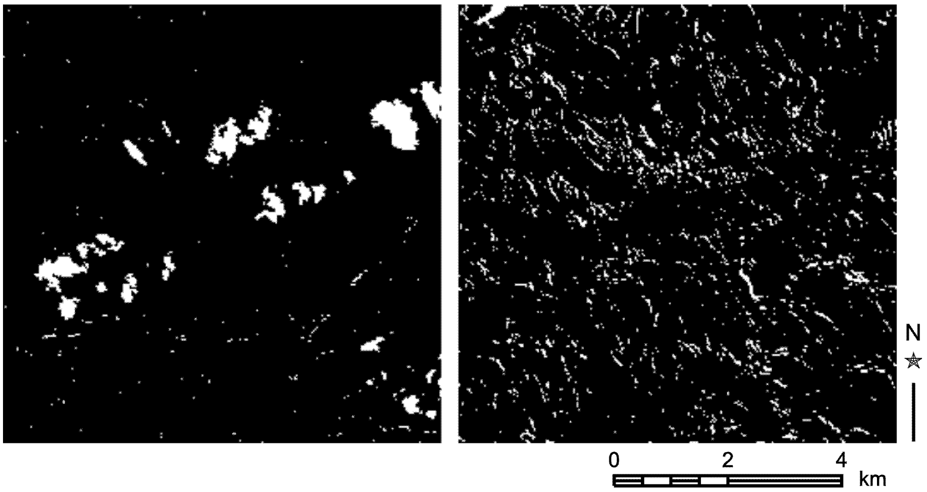

Figure 4. Binary mask of corrupted values produced by thresholding the multiscale product. Time slices are 1989 (left) and 1992 (right).

The calculated RMSE for the first simulation shows clearly that replacement methods widely used for declouding NOAA time series were very inaccurate in comparison to regression methods (table 1). Specially, the maximum value composite gave the worst results for both clouded and shadowed areas. The wavelet-based approach was more accurate for clouded areas while linear regression performed better in shadowed areas. Even then, the time series used in the first simulation represented wet and dry seasons sequentially, leading to an up and down pattern which is easy to be modelled with linear regression. More complete time series, like the second simulation, would certainly demand more elaborated regression techniques if one wants to keep track of real trends in the time series.

Table 1. Root mean square errors (x 1000) for the five interpolation methods.

Interpolation Method	Clouded Areas	Shadowed areas
Minimum value	1.8364	1.3319
Mean value	1.0286	0.6250
Maximum value	3.2227	2.0406
Linear regression	0.6699	**0.4197**
Wavelet regression	**0.5757**	0.4339

Figure 5 shows the regression curve obtained with the robust non-linear wavelet analysis applied to the simulated forest pixel. Note that the influence of outliers was completely removed and the non-linear estimation could be properly applied. The technique is also useful

to reduce geometric misregistration, radiometric noise, and other anomalies from long time series of remotely sensed data. **Figure 6** shows a small subset contaminated with the larger shadow in the upper right corner of **Figure 4**. The image estimated with non-linear wavelet regression maintained even the spatial contrast of the reference image, different from linear estimation that tends to smooth out the object's edges.

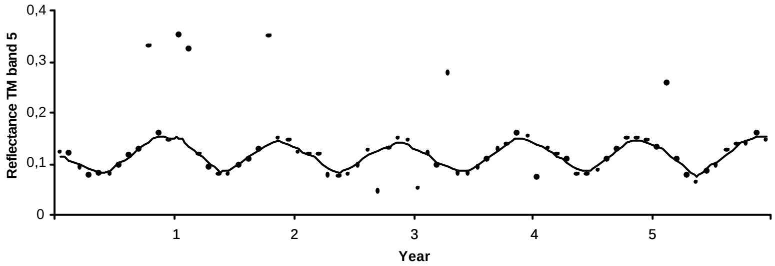

Figure 5. Regression curve obtained with the robust smoother-cleaner wavelets followed by wavelet shrinkage.

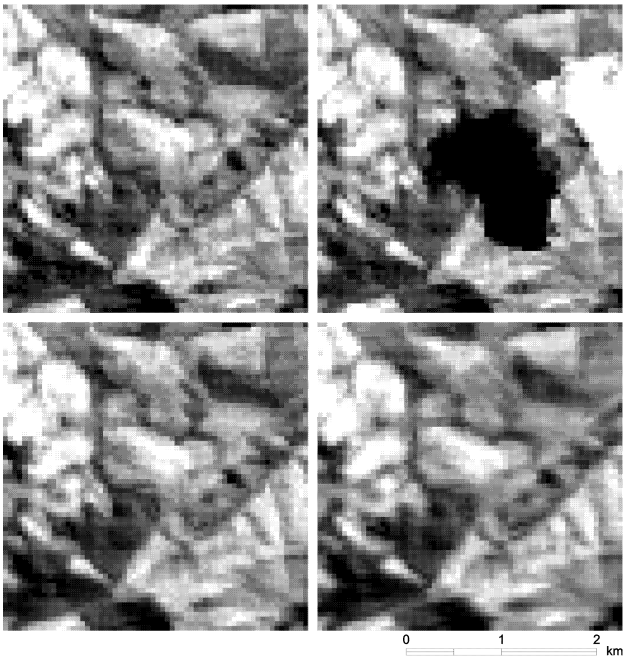

Figure 6. Reference cloud free image (upper left), simulated cloud contamination (upper right), nonlinear estimation (lower left), and linear estimation (lower right).

7. Conclusions

A framework for (non-Gaussian) noise suppression (*i.e.*, cloud removal) from remotely sensed time series was presented and demonstrated. The procedure is a step towards automation because the contaminated instances of the time series do not need to be known in advance. Multiscale products of wavelet scales might be effectively used to automatically mask corrupted values for further replacement with any desired method. The method proposed here not only identified clouded and shadowed pixels but also other anomalies like misregistration effects and changes of short duration (*e.g.*, burn scars). The robust non-linear wavelet regression can do both, detection and estimation, at the same time and produce noise reduced images at any point in the time series. Thus, the wavelet approach is promising as a pre-processing step for effective time series analysis. It can be adapted to reduce radiometric discrepancies among images in the time series, acting as a temporal smoothing operator.

Although not compared directly with Fourier-based methods (*e.g.*, HANTS, FASIR), some advantages of the wavelet-based method may be highlighted: (1) lower computational complexity, (2) simultaneous detection of positive and negative anomalies in the time series, and (3) patches of outliers (*e.g.*, cloud contamination observed sequentially for a given location) are efficiently removed from the time series with the robust smoother-cleaner wavelets. In the approach proposed by Addink and Stein (1999), the image to be declouded as well as the image used as the second variable in co-kriging must have low cloud cover because the presence of clouded pixels among the interpolators decreases the reliability of the method.

Considering the comparisons presented in this paper, wavelet regression predicted the reference values for clouded areas better than all others did, and performed almost equivalent to linear prediction in shadowed areas. Even so, more complete time series, like in the second simulation, would certainly be better modelled with non-parametric regression methods.

A similar approach might be used in the spatial domain and combined with the temporal analysis presented here. Further research on this direction and on more complete data sets (daily or monthly series) will bring insights to other possibilities and improvements of the procedure.

References

Addink, E.A., and A. Stein, 1999. A comparison of conventional and geostatistical methods to replace clouded pixels in NOAA-AVHRR images, *International Journal of Remote Sensing* 20: 961-977.

Bruce, A.G., D.L. Donoho, H.Y. Gao, and R.D. Martin, 1994. Denoising and robust nonlinear wavelet analysis. *IEEE Proceedings SPIE: Wavelet Applications*, Orlando, 2242: 325-336.

Derrien, M., B. Farki, L. Harang, H. LeGléau, A. Noyalet, D. Pochic, and A. Sairouni, 1993. Automatic cloud detection applied to NOAA-11/AVHRR imagery, *Remote Sensing of Environment* 46: 246-267.

Donoho, D.L., and I.M. Johnstone, 1994. Ideal spatial adaptation by wavelet shrinkage, *Biometrika* 81: 425-455.

Guo, L.J., and J.M. Moore, 1993. Cloud shadow suppression enhancement for geological interpretation of ATM data. *Proceedings of the 9th Thematic Conference on Geologic Remote Sensing*, Environmental Research Institute of Michigan, An Arbor, 287-297.

Holschneider, M., R. Kronland-Martinet, J. Morlet, and P. Tchmitchian, 1989. A real time algorithm for signal analysis with the help of the wavelet transform, *Wavelets: Time Frequency Methods and Phase Space*. (J.M. Combes, A. Grossman, and P. Tchmitchian, editors), Springer-Verlag, New York, 286-297.

Long, D.G., Q.P. Remund, D.L. Daum, 1999. A cloud-removal algorithm for SSM/I data, *IEEE Transactions on Geosciences and Remote Sensing* 37: 54-62.

Los, S.O., G.J. Collatz, P.J. Sellers, C.M. Malmström, N.H. Pollack, R.S. DeFries, L. Bounoua, M.T. Parris, C.J. Tucker, and D.A. Dazlich, 2000. A global 9-yr biophysical land surface dataset from NOAA AVHRR data, *Journal of Hydrometeorology* 1: 183-199.

Roerink GJ, M. Menenti, W. Verhoef, 2000. Reconstructing cloud free NDVI composites using fourier analysis of time series, *International Journal of Remote Sensing* 21: 1911-1917.

Sadler B.M., A. Swami, 1999. Analysis of multiscale products for step detection and estimation, *IEEE Transactions on Information Theory* 45: 1043-1051.

Saunders R.W., and K.T. Kriebel, 1988. An improved method for detecting clear sky and cloudy radiances from AVHRR data, *International Journal of Remote Sensing* 9: 123-150.

Starck J.L., F. Murtagh, A. Bijaoui, 1998. *Image Processing and Data Analysis*. University Press, Cambridge.

Wang B., A. Ono, K. Muramatsu, N. Fujiwara, 1999. Automatic detection and removal of clouds and their shadows from Landsat TM images, *IEICE Transactions on information and systems* E82-D: 453-460.