# Testing near real-time detection of contaminated pixels in AVHRR composites

Josef Cihlar[1], Rasim Latifovic[2], Jing Chen[1], Zhanqing Li[1]
1 Canada Centre for Remote Sensing, Ottawa, Ontario, K1A 0Y7
2) Intermap Technologies, Ottawa, Ontario K2E 1A2

Corresponding author: J. Cihlar, tel. 613 947 1265, fax 613 947 1406
josef.cihlar@ccrs.nrcan.gc.ca

**Abstract**

When using composite optical satellite images for land studies, an accurate and sensitive method is needed to detect pixels contaminated by unwanted atmospheric and surface effects. In this paper, we have examined the feasibility of using an algorithm previously developed for post-season analysis but in a 'forward' mode, i.e. for identifying contaminated pixels in current-season data. The CECANT algorithm (Cloud Elimination from Composites using Albedo and NDVI Trend; Cihlar, 1996) uses AVHRR channel 1 reflectance to detect strongly contaminated pixels (bright clouds, snow), and the normalized difference vegetation index (NDVI) to identify partially contaminated pixels over land. However, this approach needs a complete NDVI seasonal trajectory to derive the adaptive thresholds. Since some important applications imply near-real time processing we have examined the possibility of deriving the thresholds from previous (historical) AVHRR data. Using four years of AVHRR data of Canada we found the accuracy to vary (55-100%), with thresholds from a single year yielding anomalous results in some compositing periods. On the other hand, the use of thresholds derived from averaged data (over 3 years in this case) produced more accurate and consistent results. Assuming that contamination masks derived from each year's data are 100% accurate, the masks based on the three years of AVHRR data were 75-95% correct (4-year average), depending on the compositing period. Two alternative adjustments were attempted to improve the accuracy and consistency of the identification of contaminated pixels. By relaxing one of the CECANT thresholds the errors of omission and commission could be more equally balanced (~10 and 20%, respectively). Further improvement was obtained by adjusting the NDVI data to the data set initially used in deriving the threshold coefficients, and then deriving a new set of coefficients for that period. In this case, the omission and commission errors were nearly equal and, importantly, the performance was consistent among all the years. The performance of an

alternative, simplified cloud screening method (based on fixed reflectance and temperature thresholds) was also examined for a comparison, and was found to have lower accuracy (58% compared to 87%, 4-year average). It is concluded that CECANT may be effectively applied in near-real time processing of satellite optical data. However, the residual mismatches indicate that a reprocessing after the end of the growing season may be desirable for applications requiring high radiometric accuracy.

## Introduction

Numerous studies have demonstrated the importance of optical satellite data for studies of vegetation dynamics (e.g., Townshend, 1994). Although the NOAA Advanced Very High Resolution Radiometer (AVHRR) is currently the primary data source for such studies, new sensors will be launched in 1998 and subsequently with significant improvements for biospheric monitoring. Since these data are subject to strong atmospheric interference as well as other types of noise, corrections are essential prior to the extraction of biophysical information. The necessity of detecting fully as well as partly contaminated pixels in composite satellite images intended for land biosphere studies has been previously demonstrated (Gutman, 1992; Gutman et al., 1995; Cihlar et al., 1998).

Various techniques have been developed to identify cloudy pixels over ocean and land surfaces to exclude the pixels from further analysis. In most cases, the techniques have been developed for use with single-date images. Several such schemes have been previously developed for land applications (Gutman, 1992; Gutman et al., 1987; Saunders and Kriebel, 1988; Stowe et al., 1991; Derrien et al., 1993; Loudjani et al., 1994; Simpson, 1994; Simpson and Gobat, 1996). In the case of composite images, two options exist. One is to detect clouds in individual input scenes, and then retain all clear pixels for further analysis; the second is to retain the least contaminated pixel, selected by means of a compositing criterion such as the maximum normalized difference vegetation index (NDVI; Holben, 1986). The former is a highly desirable approach but it requires a very accurate cloud detection method. Ultimately, this will be the preferred approach after the variety of measurements increases to the point where small atmospheric effects (and surface, such as snow) can be reliably detected. The MODIS data (Salomonson, 1988) will make major strides in this regard. Unfortunately, due to the temporal variability and spatial heterogeneity of land cover, it is difficult to define accurate and universally applicable thresholds that would differentiate between clear and contaminated pixels. So far, accurate detection of subpixel clouds and other similar contamination has not been successfully demonstrated for single-date AVHRR and similar data because of insufficient spectral information about the atmospheric conditions represented in the satellite images. For example, a cloud mask is produced in the AVHRR Pathfinder for all input scenes (James and Kalluri, 1994) but the final composites are created using maximum NDVI as a selection criterion. An important practical consideration in near-real time operation is data processing burden. For composites containing images from d days, the cloud screening effort is reduced by a factor of (d-1)/d.

Recently, Cihlar (1996) described a technique for detecting contaminated pixels in image composites. CECANT (Cloud Elimination from Composites using Albedo and NDVI Trend) detects fully cloudy pixels as well as those contaminated by subpixel clouds, fog, smoke, and snow. While this method performs well with composite images (Cihlar, 1996) the derivation of its thresholds requires knowledge of the seasonal NDVI trajectory. The method can thus be used for post-seasonal analysis of the surface dynamics (Cihlar et al., 1997, 1998) but its usefulness for near-real time processing, e.g. crop and range condition monitoring (Allison et al., 1989; Smith, 1994), has not been established.

In this paper we explore the impact of the changes in CECANT thresholds on the identification of the contaminated pixels. The objective is to determine if this approach can be used for near-real time processing of AVHRR data. After briefly reviewing the principal requirements of the CECANT method we describe the data set and the analysis procedure employed to evaluate the effect of interannual changes in the thresholds. For comparison, we also evaluate the performance of a physically-based cloud detection algorithm of Wu et al. (1995).

## Methodology

### *1. Principles*

The CECANT method is based on the following assumptions (Cihlar, 1996):

1. For a given pixel, at least some compositing periods during the growing season contain uncontaminated pixels.
2. The first peak in the AVHRR channel 1 histogram (corresponding to surface reflectance in the red spectral region of about ≤0.20) contains clear-sky, snow/ice-free, atmospherically uncontaminated land pixels.
3. NDVI values for a given pixel increase monotonically from the beginning of the growing season until they reach the seasonal peak green (maximum NDVI) for that pixel.

CECANT uses the AVHRR channel 1 for coarse filtering and the seasonal trajectory of the NDVI curve as fine filter for partly contaminated pixels. Its application requires the computation of two values for each pixel, R and Z:

$$R(i,j,t) = \frac{NDVI(i,j,t) - NDVI_a(i,j,t)}{M(i,j)}, \qquad [1]$$

$$Z(i,j,t) = \frac{NDVI_{\max}(i,j,t) - NDVI(i,j,t)}{NDVI_{\max}(i,j,t)}, \qquad [2]$$

where:

NDVI = measured value;
$NDVI_a$ = estimated average value obtained as a best fit to the NDVI seasonal curve;
$NDVI_{max}$ = estimated maximum value obtained by fitting an upper envelope to the NDVI seasonal curve;
M = median value of the absolute difference between NDVI and $NDVI_a$ for each pixel (i,j)
i,j = line and pixel coordinates;
t = compositing period.

Thus in essence, CECANT flags a pixel as contaminated if its NDVI is much lower than would be expected for that pixel and period. Four thresholds are required for each period to identify partially contaminated pixels:

C1: the maximum channel 1 reflectance of a clear-sky, snow- or ice-free land pixel in the data set (defined as C1≤0.3).

$R_{min}(t)$: the maximum acceptable deviation of the measured value NDVI(i,j,t) *below* the estimated $NDVI_a(i,j,t)$. Lower NDVI(i,j,t) values are considered contaminated.

$R_{max}(t)$: the maximum acceptable deviation of the measured value NDVI(i,j,t) *above* the estimated $NDVI_a(i,j,t)$. Pixels with R values higher than $R_{max}$ represent anomalously high NDVI, e.g. due to local misregistration.

$Z_{max}(t)$: the maximum acceptable deviation of the measured value NDVI(i,j,t) *below* the estimated $NDVI_{max}(i,j,t)$. Pixels with lower NDVI are considered unacceptable even if $NDVI>NDVI_a$.

The three thresholds ($R_{min}$, $R_{max}$, $Z_{max}$) are obtained as follows (Cihlar, 1996):

$$R_{\min}(t) = R_{mean}(t) - 1, \qquad [3]$$

$$R_{\max}(t) = R_{mean}(t) + 4, \qquad [4]$$

$$Z_{\max}(t) = Z_{mean}(t) + 2*|Z_{mean}(t)|, \qquad [5]$$

where $R_{mean}$ and $Z_{mean}$ are computed from R, Z values of all pixels (i,j,t) for which C1≤0.3. These thresholds are thus determined separately for each compositing period. A given pixel is then considered to be clear if {C1(i,j,t)≤0.3 and {$R_{min}(t)<R(i,j,t) \leq R_{max}(t)$} and $Z(i,j,t) \leq Z_{max}(t)$}, and contaminated otherwise. Eq. (3)-(5) were derived empirically after inspecting a seasonal RZ plots for the Canadian landmass (Cihlar, 1996).

From the above description it is evident that CECANT requires for each pixel (i,j): M(i,j), seasonal trajectories of $NDVI_{max}(i,j)$ and $NDVI_a(i,j)$, and thresholds R(t) and Z(t). If these values did not vary among years CECANT could be used for near-real time processing provided that at least one year of data is available to derive the thresholds. In this paper we test the effect of using $NDVI_{max}$, $NDVI_a$, $R_{mean}$ and $Z_{mean}$ thresholds from year k in year l, k≠l.

### *2. Data and methods*

We have employed AVHRR data over the landmass of Canada from the 1993 to 1996 period, NOAA-11 for the first two years and NOAA-14 for the last two. In each case except 1994 (last period 1-10 September) data from the 11 April to 31 October were used. The satellite data were processed using the CCRS Geocoding and Compositing system (GEOCOMP; Robertson et al., 1992). Briefly, GEOCOMP performs sensor calibration with time degradation (Cihlar and Teillet, 1995); registration to the Lambert Conformal Conic projection ($49^oN$ and $77^oN$ parallels, reference meridian $95^oW$) and resampling to 1km pixel size; and compositing for 10-day periods by retaining data with the maximum NDVI value. Atmospheric correction of AVHRR channels 1 and 2 was made using the SMAC algorithm of Rahman and Dedieu (1994). Cihlar et al. (1997) provide more detailed information on data processing. The data were then subsampled by retaining every 6th line and 6th pixel to make analysis easier; total number of land pixels was 246521.

For each year, CECANT ancillary data were computed: M(i,j), $NDVI_{max}(i,j,t)$, $NDVI_a(i,j,t)$, $R_{mean}(t)$ and $Z_{mean}(t)$. From these data, the three thresholds ($R_{min}(t)$, $R_{max}(t)$, $Z_{max}(t)$) were computed using Eq. [3]-[5]. In addition, a combined set was computed from the three full years (‘93, ‘95, ‘96). To this end, the NDVI(i,j,t) values were first averaged for each pixel and compositing period over the 3 years and using this data set, R(t) and Z(t) were computed as if the averaged values originated in one year.

With the above thresholds five sets of cloud masks were computed for each year: one using thresholds for the same year ($RZ_{sy}$), three with RZ thresholds derived in the other years ($RZ_{oy}$), and one using the 3-year average thresholds ($RZ_m$). To evaluate these various combinations we

assumed that the mask derived from same-year data ($RZ_{sy}$) was 100% accurate and could thus serve as a reference. We then logically compared the pairs of masks to count the pixels that were:

$a_1$ = contaminated in both masks;

$a_2$ = clear in both masks;

$a_3$ = contaminated in the test mask but clear in the reference mask; and

$a_4$ = clear in the test mask but contaminated in the reference mask.

Based on these pixel counts we computed three measures to characterize the relation between the reference and the test set for each period t:

$$MCF_{CO}(k,l,t) = \frac{100 * NP(a_1(k,l),t)}{NP(a_1(k,k),t)}, \quad [6]$$

$$OCF_{CL}(k,l,t) = \frac{100 * NP(a_4(k,l),t)}{NP(a_2(k,k),t)}, \quad [7]$$

$$CCF_{CL}(k,l,t) = \frac{100 * NP(a_3(k,l),t)}{NP(a_2(k,k),t)}, \quad [8]$$

where:

- $MCF_{CO}$ is matching contaminated pixels fraction, or the number of contaminated pixels in the reference set which were also identified as contaminated in the test set, expressed in percent of the number of contaminated pixels in the reference set;
- $OCF_{CL}$ is the fraction of undetected (omitted) contaminated pixels, or pixels identified as clear in the test set but contaminated in the reference set, expressed as proportion of clear-sky pixels in the reference set (in %);
- $CCF_{CL}$ is the fraction of wrongly included (committed) contaminated pixels, i.e. pixels identified as contaminated in the test set but clear in the reference set, expressed as proportion of clear-sky pixels in the reference set (in %).
- $NP(a_i(k,l),t)$ is the number of pixels in the period t for the combination $a_i$ from the two years (i refers to the above combinations, i=1 to 4);
- k is the year of the reference set;
- l is the year of the test set;

The three measures were chosen to quantify various aspects of the CECANT performance. $MCF_{CO}$ describes its ability to identify exactly the same pixels with different sets of coefficients.

$OCF_{CL}$ and $CCF_{CL}$ are expressed in relation to clear-sky pixels and thus represent relative errors remaining in the data set, since clear-sky pixels are the focus of the data processing and analysis. By definition, the range of $MCF_{CO}$ is (0,100) for any period t. The lower bound for $OCF_{CL}$ and $CCF_{CL}$ is zero but the upper bound does not have a fixed value; it could exceed 100 if the number of differently labeled pixels exceeds clear pixels (Eq. [7], [8]).

**Results and discussion**

***1. Effects of $RZ_{oy}$***

Figure 1a-1b shows $MCF_{CO}$ values for the 12 combinations of reference year $RZ_{oy}$ (e.g., 93RZ) vs. test year (e.g., 1994 data). The curves span an overall $MCF_{CO}$ range of about 40%, and they also have a definite seasonal trend with a broad minimum in the summer. In the first half of the season (prior to Period 11) the combinations involving 1994 tended to have the lowest $MCF_{CO}$ values but in the second half others occupied the extreme positions. Figure 1a-1b also shows examples of a sudden fluctuation in $MCF_{CO}$ values, e.g. for periods 11 and 16 in the combination 96data/95RZ. The performance was most consistent early and late in the season where the albedo (C1 threshold) was an important criterion.

The consistently lower $MCF_{CO}$ values involving 1994 data in the first half of the season could be partly due to imperfect corrections for atmospheric and

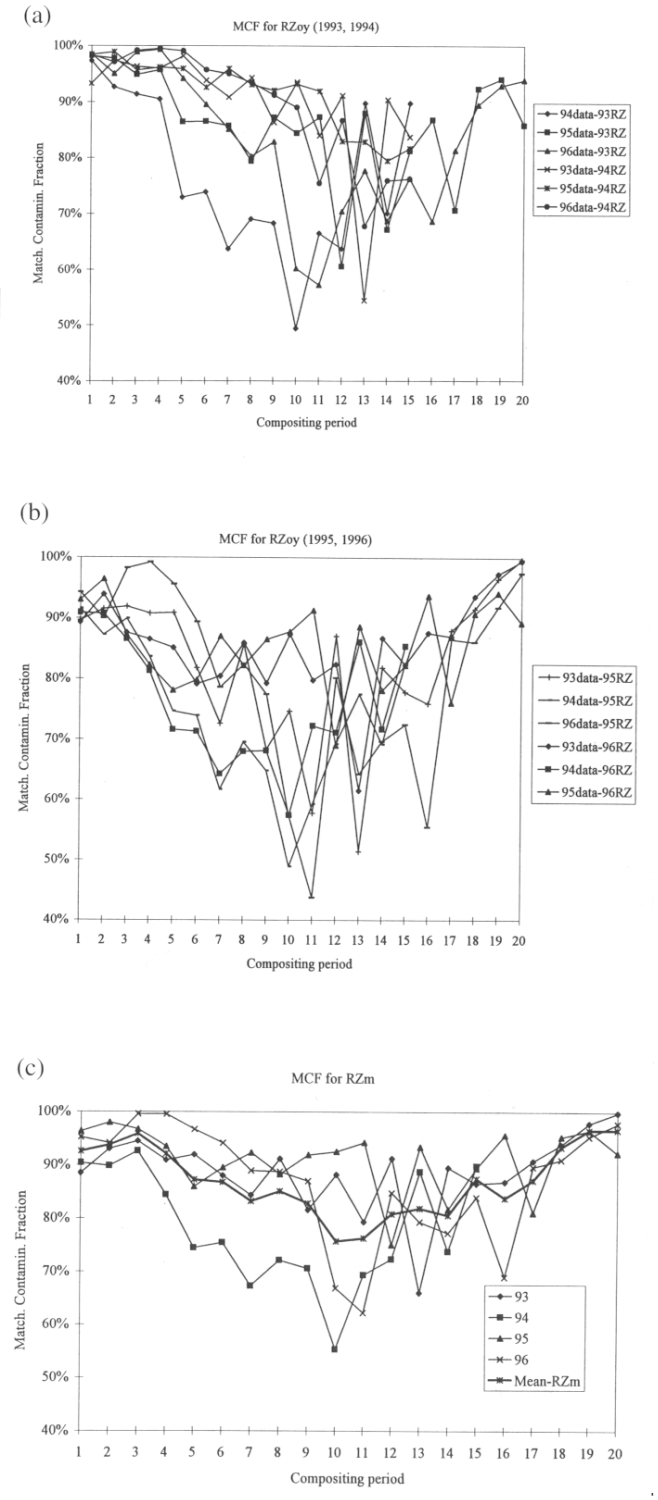



**Figure 1.** Comparison of the contaminated pixels detected by RZ thresholds derived from the original data and those from another year, shown as mean values for Canada for 20 compositing periods and four years. Figure 1a-1b: Matched Contaminated Fraction $MCF_{CO}$ for RZ thresholds from year x applied to data from year y, $y \neq x$. Figure 1c: Average $MCF_{CO}$ accuracy by the data set year.

bidirectional effects. Since the data acquisition geometry was very different for the NOAA-11 AVHRR data compared to other years (with mean solar zenith angle SZA>55$^o$ for all compositing periods), both bidirectional corrections (i.e., normalization to SZA=45$^o$) and atmospheric corrections (which loose accuracy above solar zenith of 60$^o$ and view zenith of 50$^o$; Rahman and Dedieu, 1994) could introduce bias into the corrected data. Cihlar et al. (1998) found that the mean NDVI values for vegetated pixels had consistently higher NDVI values in 1994 compared to other years. This explanation is supported by the higher $MCF_{CO}$ results for the 93data/94RZ combination. However, the effect could not be traced directly to a particular threshold because of the self-normalization nature of CECANT (Eq.[1]-[3]). In fact, the 1994 thresholds did not consistently deviate from the other years. The difference in the contamination masks is thus introduced mostly through the $NDVI_a$ and possibly through M (Eq.[1],[2]). - On the other hand, the random fluctuations in the summer (e.g., Period 13) are probably due to the differences between the respective RZ threshold values and the relatively small number of contaminated pixels involved.

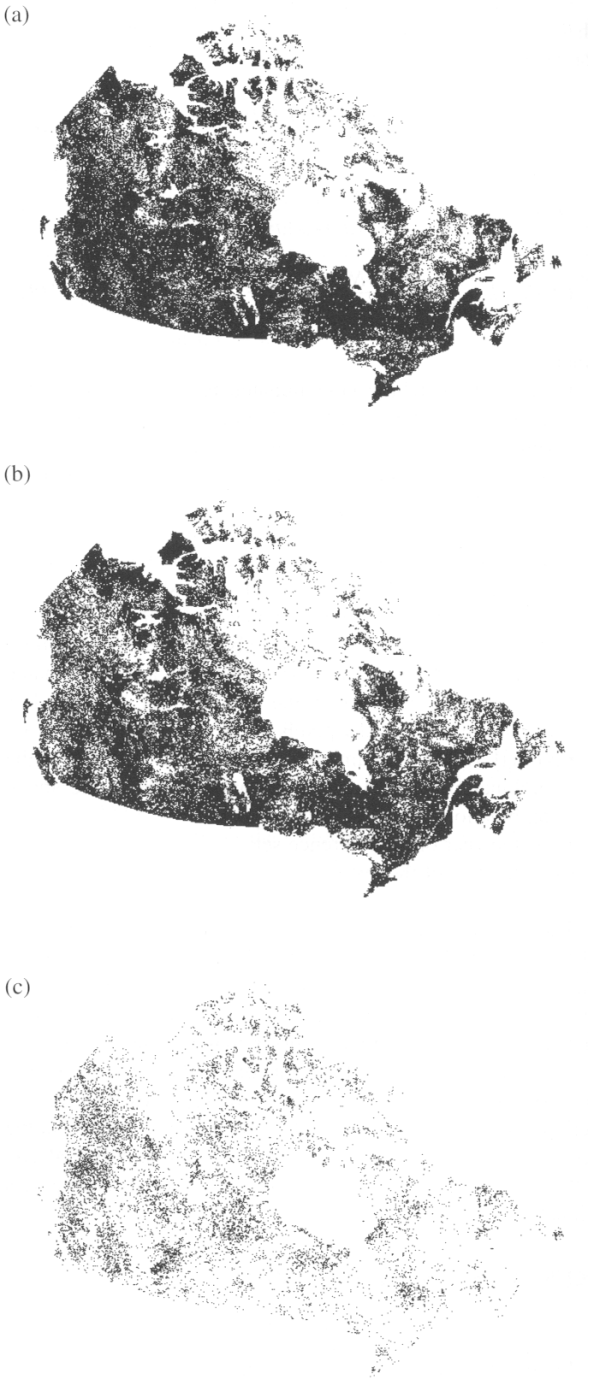



**Figure 2**. The distribution of contaminated pixels (in white) for Period 10 (1-10 July 1995). Figure 2a: contaminated pixels were identified using RZ for 1995 (29.1% of all pixels). Figure 2b: contaminated pixels were identified using $RZ_m$ (34.2% of all pixels). The values for $MCF_{CO}$, $OCF_{CL}$, and $CCF_{CL}$ were respectively: 92.5%, 3.1%, 27.1% when using Eq.[3] for $R_{min}$; and 85.0%, 6.1%, and 16.2% with Eq. [9]. Figure 2c: Difference between Figure 2a and 2b (shown as black, 18.9% of all pixels).

When the mean $RZ_m$ set was applied to each of the four years the range of values among the years at any period t was reduced substantially (Figure 1c). The deviation was largest for 1994 which was not included in computing the $RZ_m$. The direct identification accuracy also

increased. The average $MCF_{CO}$ value for the four years (heavy line in Figure 1c) varied between 76% and 97% depending on the period, the range being slightly higher (79%-97%) if only the 3 years used in deriving the $RZ_m$ values were considered. The annual average $MCF_{CO}$ ranged from 78% (1994) to 91% (1995). The increase between Figure 1a-1b and 1c indicates that when data from several years are combined the thresholds and representative limiting values (Eq.[1],[2]) are less susceptible to specific effects that may play a role in individual years.

Figure 2 shows an example of the masks produced by $RZ_{sy}$ and $RZ_m$ thresholds for Period 10 in 1995 when most land pixels (70.9%) were clear. The $MCF_{CO}$ was 92.5% for $RZ_m$, i.e. only 7.5% pixels identified as contaminated using $RZ_{sy}$ were missed by $RZ_m$. Figure 2c shows the difference between the two masks. The variations in the spatial distribution are typical of other cases found in the study. They tended to occur in patterns, indicating regional concentrations of mismatches. Since three thresholds, derived with the assistance of three NDVI trajectory-related parameters ($NDVI_{max}$, $NDVI_a$, M), are involved in the determination of a contaminated pixel there is no single explanation for these patterns. In general, as the reference set changes it brings about differences due to the changing pixel-specific trajectory (Eq. [1],[2]) as well as the period-specific thresholds (Eq. [3]-[5]). Small shifts in these could result in appreciable differences in contaminated pixels identification, especially for $R_{min}$ and $Z_{max}$ because of the skewness of the distribution near the thresholds (Cihlar, 1996).

The fraction of undetected contaminated pixels $OCF_{CL}$ is illustrated in Figure 3a-3b and 3c. In most cases, $OCF_{CL}$ was <25% (Figure 3a-3b) with $RZ_{oy}$ although occasional larger anomalies occurred. They showed a tendency to increase near the end-periods but since these are outside the growing season the increase is not a serious problem. The $OCF_{CL}$ errors were much reduced with $RZ_m$ (Figure 3c), averaging about 10% among the four years. Consistently with the $MCF_{CO}$ results, the largest $OCF_{CL}$ values were found for 1994.

The proportion of erroneously included pixels ($CCF_{CL}$) was consistently greater than the $OCF_{CL}$ (Figure 3c, 3f). Although most were <40%, significant outliers occurred during the growing season for some combinations when most data year/RZ year combinations had high $CCF_{CL}$ errors. The three worst combinations involve RZ94, supporting the explanation that the higher overall NDVI in 1994 was the main reason since this would increase $NDVI_a$, thus causing $R_{min}$ values to be too high. The values were also high for Period 1 but since this occurs in the dormant stage, the errors are not critical. The average $CCF_{CL}$ values for the combined $RZ_m$ were also high due to the 1994 data, about 30% (Figure 3f). Without the 1994 data the $CCF_{CL}$ would be around 20% in the May - August period, i.e. most of the growing season.

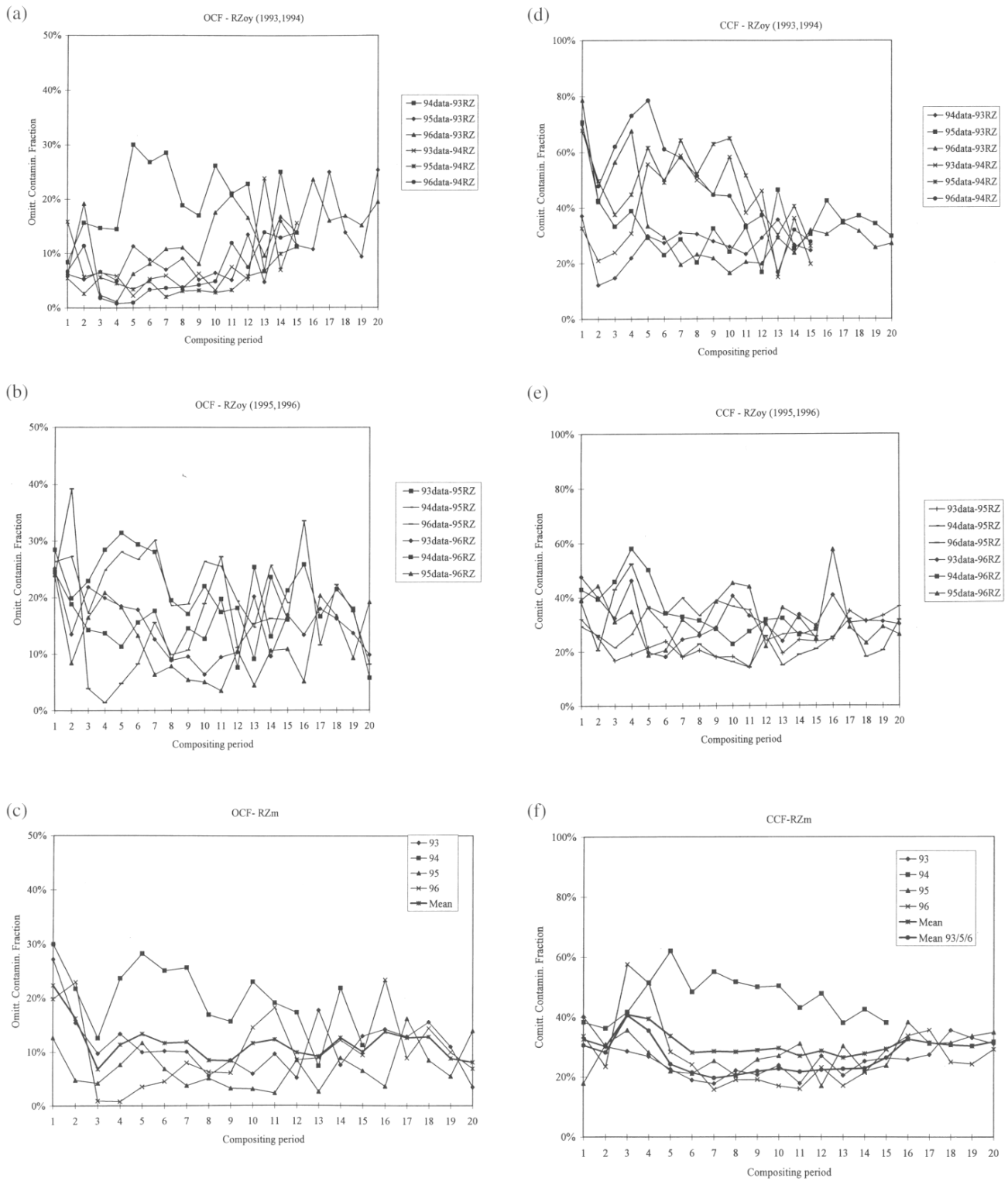



**Figure 3.** Incorrectly identified pixels for two error measures and different combinations of reference data and test data. Figure 3a-3b: Undetected contaminated fraction $OCF_{CL}$ for RZ thresholds from year x applied to data from year y, y≠x. Figure 3c: $OCF_{CL}$ for RZ thresholds from three years of combined data (93,95,96) applied to individual years. Figure 3d-3e: Committed Contaminated Fraction $CCF_{CL}$ for RZ thresholds from year x applied to data from year y, y≠x. Figure 3f: $CCF_{CL}$ for RZ thresholds from three years of combined data (93,95,96) applied to individual years.

The varying accuracy in Fig. 1-3 is related to differences in the data sets among years. The accuracy can be expected to be high when the data set used to derive the coefficients has 'similar characteristics' as the one to be screened. 'Similar' means that the clear-sky NDVI(i,j,t) values have not changed significantly enough to modify $NDVI_{a,}$, $NDVI_{max}$ or M, and secondly that the overall pixel contamination in the region of interest has not changed. While the second assumption is sound for a large region such as Canada, the changes in the measured NDVI can be caused by various factors. The main one is interannual land cover change but this occurs only in a small proportion of the total area. Surface changes which reduce NDVI (e.g., drought) are also likely to cause errors since they are spectrally indistinguishable from the effect of atmospheric contamination. Other causes are measurement effects: changes in sensor calibration, viewing geometry, or data processing (atmospheric and SZA corrections). Careful processing can mitigate the impact of these effects but as the 1994 data shows they cannot be assumed to removed entirely, given the present state of knowledge. It should be noted that none of these factors cause problems when the CECANT coefficients are derived from the data set to be screened.

### *2. Effect of relaxing $R_{min}$*

Among the R and Z thresholds $R_{min}$ has the strongest effect on the selection of contaminated pixels. A test was therefore undertaken to determine if a relaxation of $R_{min}$ would improve the identification of contaminated pixels. Using the 3-year averaged data set, the new $R_{min}$ was computed as

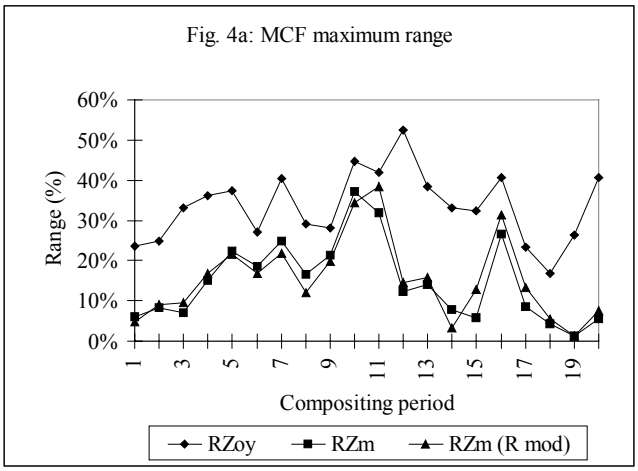
$$R_{\min} = R_{mean} - 2.0. \qquad [9]$$

C1, $R_{max}$ and $Z_{max}$ were not changed.

Figure 4 compares the accuracies for the two sets of $RZ_m$, as an average for the four years. Each curve represents the overall range of values obtained. For example, the top curve in Figure 4a shows the difference between the highest and the lowest $MCF_{CO}$ value obtained among the various $RZ_{oy}$ values over the four years. As noted above, the fluctuation was quite high for individual $RZ_{oy}$ vs. data year combinations, the difference in $MCF_{CO}$ ranging from 17 to 53% depending on the compositing period. With $RZ_m$ the fluctuation decreased significantly, to 1-30% in most cases. After relaxing $R_{min}$ the overall range did not change substantially (Figure 4a). A similar trend was found for the omitted and committed contaminated fractions (not shown) where the range also decreased significantly only between $RZ_{oy}$ and $RZ_m$.

While the modified $R_{min}$ did not change the range of differences in $MCF_{CO}$, it improved the balance between omitted and committed pixels. Figure 4b shows these for the two $R_{min}$ values

and the difference between them. It is evident that the $OCF_{CL}$ and $CCF_{CL}$ became more nearly balanced, the average difference decreasing from 19.2% ($RZ_m$ with $R_{min}$) to 9.1% ($RZ_m$ with modified $R_{min}$). However, the improvement came partly at the expense of the direct identification of the contaminated pixels. A comparsion the mean $MFC_{CO}$ values for the two $R_{min}$ values and 4 years (Figure 6a) shows that $MCF_{CO}$ decreased somewhat for the modified set, from 87% to 81.1%. This also points to the interannual and spatial variability of quantities used to compute R and Z as the reasons for the residual errors in detecting contaminated pixels.



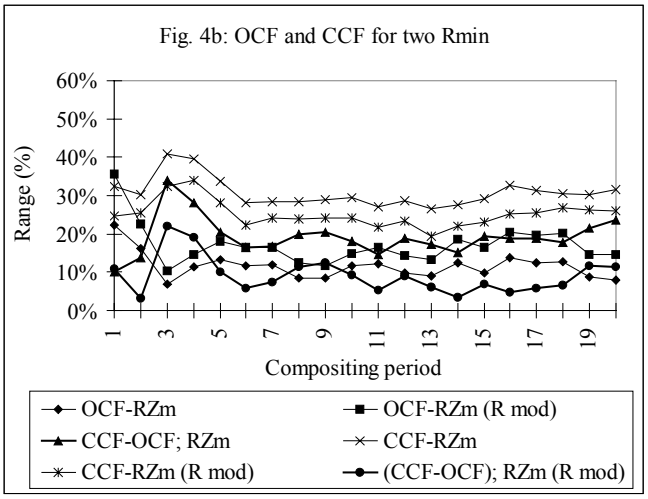




**Figure 4**. The overall range of accuracy values for three sets of RZ thresholds: from individual other years $RZ_{oy}$, as mean over three years $RZ_m$ (93, 95, 96), and modified from data over three years ($RZ_m$ mod). Figure 4a: Matching Contaminated Fraction $MCF_{CO}$. Figure 4b: Comparison of $OCF_{CL}$ and $CCF_{CL}$ for two sets of $R_{min}$ coefficients based on Eq.[3] and [9], respectively.

### *3. Adjusted CECANT*

The above results indicate that changing $R_{min}$ had some effect but the imbalance of commissions and omissions persisted. After examining reasons for these inclusions we concluded that the interannual differences in the NDVI trajectories play an important role. We therefore devised a way of adjusting the target year statistics to the multiyear ones. This is accomplished by modifying Eq. (1) and (2) as follows:

$$R(i,j,t) = \frac{NDVI(i,j,t) - A(t) * NDVI_a(i,j,t)}{A(t) * M(i,j)}, \qquad [10]$$

$$Z(i,j,t) = \frac{B(t) * NDVI_{\max}(i,j,t) - NDVI(i,j,t)}{B(t) * NDVI_{\max}(i,j,t)}, \qquad [11]$$

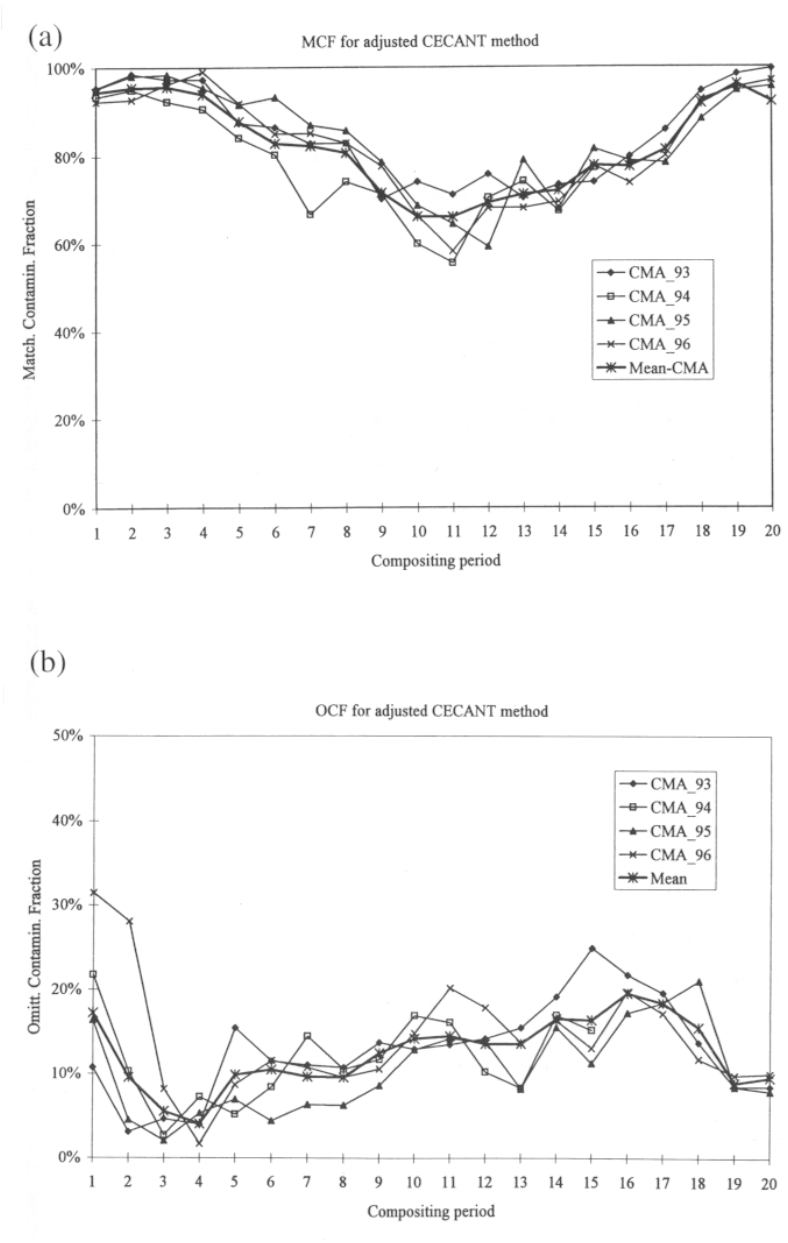
$$A(t) = \frac{\sum_{i,j} NDVI(i,j,t)}{\sum_{i,j} NDVI_a(i,j,t_m)},$$

$$B(t) = \frac{\sum_{i,j} NDVI(i,j,t)}{\sum NDVI_{\max}(i,j,t_m)}.$$

The impact of A and B is to reconcile the difference (if any) between the two sets, the multiyear reference data set (data for all compositing periods available) and the current year (only data for periods <1,t> available). Once the R and Z values are recomputed for all pixels, the adjusted thresholds are calculated from Eq. [3]-[5]. Only pixels with NDVI>0 and C1<0.3 are included in calculating the thresholds.

Using data from all four years, Figure 5 compares the performance of the modified approach (CECANT Mean-Adjusted, CMA) with $RZ_m$ and the modified $R_{min}$. The $MCF_{CO}$ values (Figure 5a) decreased compared to $RZ_m$ but are somewhat higher (overall mean value 82.3%) than for modified $R_{min}$ (81.1%). The reason for the lower $MCF_{CO}$ (compared to $RZ_m$) are the larger commission errors in the latter. That is, too many pixels were labeled as contaminated, thus increasing the $MCF_{CO}$ values. The CMA errors of omission and commission again varied between 5-15% ($OCF_{CL}$, Figure 5b) and 10-30% ($CCF_{CL}$, 5c) in most cases. However, overall the mean errors decreased, by 27% for $OCF_{CL}$ (17.1 to 12.4%) and by 16% for $CCF_{CL}$ (25.1 to 21.4%). The improved stability of the estimates in individual years and reduced interannual variability indicated by all three measures in Figure 5 is important. Based on the four years of

data, Figure 6a shows that the mean values of $MCF_{CO}$ varied much less in case of CMA than for either $RZ_m$ or modified $R_{min}$. Similarly, the $OCF_{CL}$ and $CCF_{CL}$ values fluctuated much less for CMA than for the other two methods (Figure 6b). The interannual stability is an important consideration because it ensures more consistent performance when applied to a new data set. The comparable performance in $MCF_{CO}$ and large improvements in $OCF_{CL}$ and $CCF_{CL}$ show that the CMA approach performs the best among the options tested.



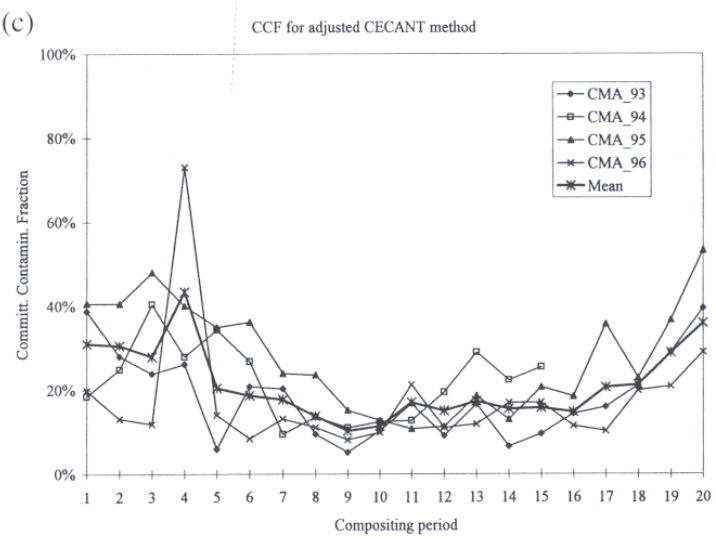



**Figure 5.** The accuracy of correct identification of contaminated pixels using RZ coefficients adjusted for the difference between reference and current year data sets (CMA approach) and all four years. Figure 5a: Matching Contaminated Fraction $MCF_{CO}$. Figure 5b: Omitted Contaminated Fraction $OCF_{CL}$. Figure 5c: Committed Contaminated Fraction $CCF_{CL}$.

### *3. Comparison with a threshold-based cloud screening scheme*

To obtain an independent assessment of the CECANT results when applied to other years we have used part of a fixed threshold-based cloud detection scheme of Wu et al. (1995), hereafter FTM. They developed a method for single-date images of sites with homogenous land cover using: a) thresholds applied to several channels following the work of Gutman et al. (1987) and Gutman (1992); and b) context analysis based on the assumption of local surface homogeneity. Since no contextual information is available in composite images and since land surfaces in general are not homogenous we only employed the initial, pixel-based thresholds. As stipulated by Wu et al. (1995) these were applied to top-of-the-atmosphere reflectance and temperature data (before atmospheric corrections). The method uses three thresholds to identify clear pixels as those for which {(R1+R2)/2<0.35 and R2/R1>1.3 and T4>280}, where R1 and R2 are top-of-the-atmosphere (TOA) reflectances (dimensionless) in AVHRR channels 1 and 2, and T4 is brightness temperature in AVHRR channel 4 ($^{o}$K). The same thresholds were employed for all compositing periods over the four years. The resulting masks were then compared to those produced by CECANT for the same period, again taking the latter as an accurate reference because of its sensitivity to subpixel contamination (Cihlar, 1996).

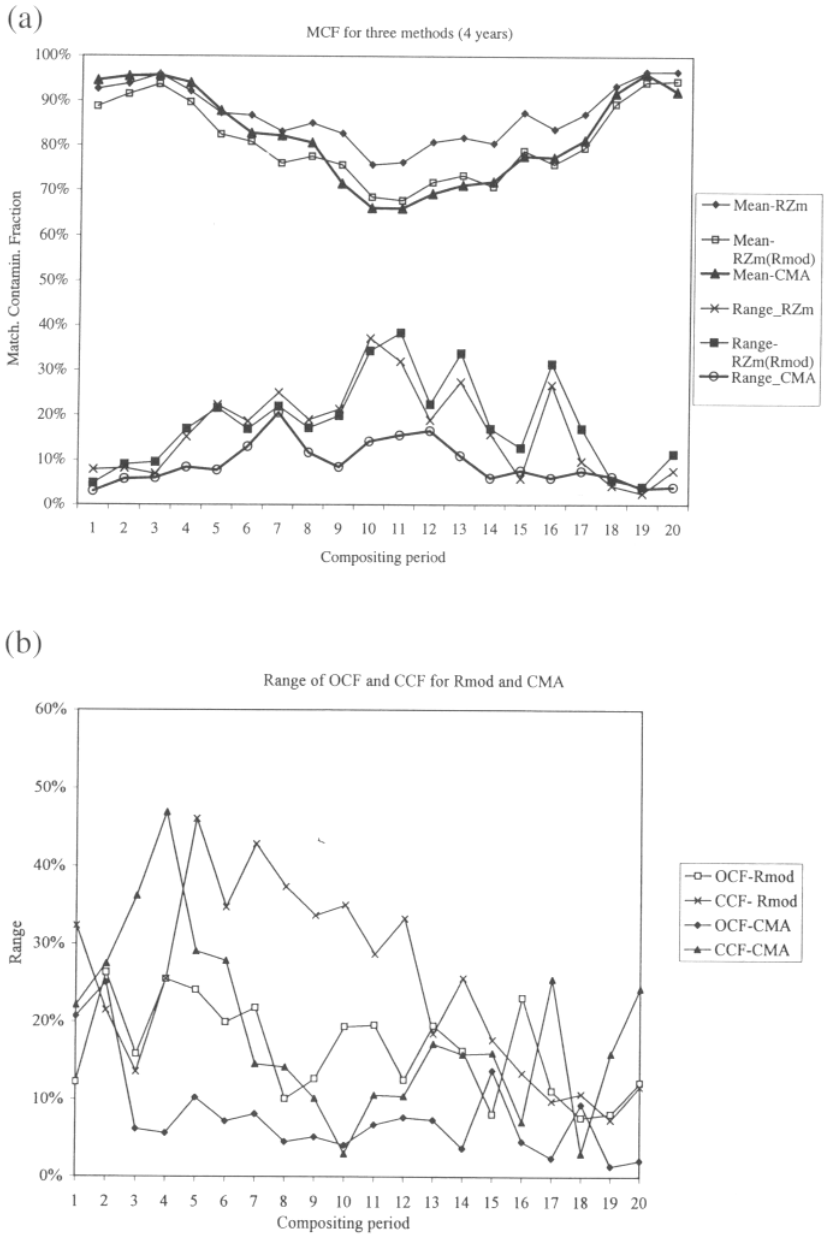



**Figure 6**. A comparison of the $MCF_{CO}$ and the range of $MCF_{CO}$ values for the four years, using three approaches: average RZ ($RZ_m$), modified $RZ_m$ (Eq. [9]), and CMA (Eq. [10]-[11]). Figure 6a: Matching Contaminated Fraction $MCF_{CO}$ values and their range over the four years. Figure 6b: Omitted and committed contaminated fractions ($OCF_{CL}$, $CCF_{CL}$).

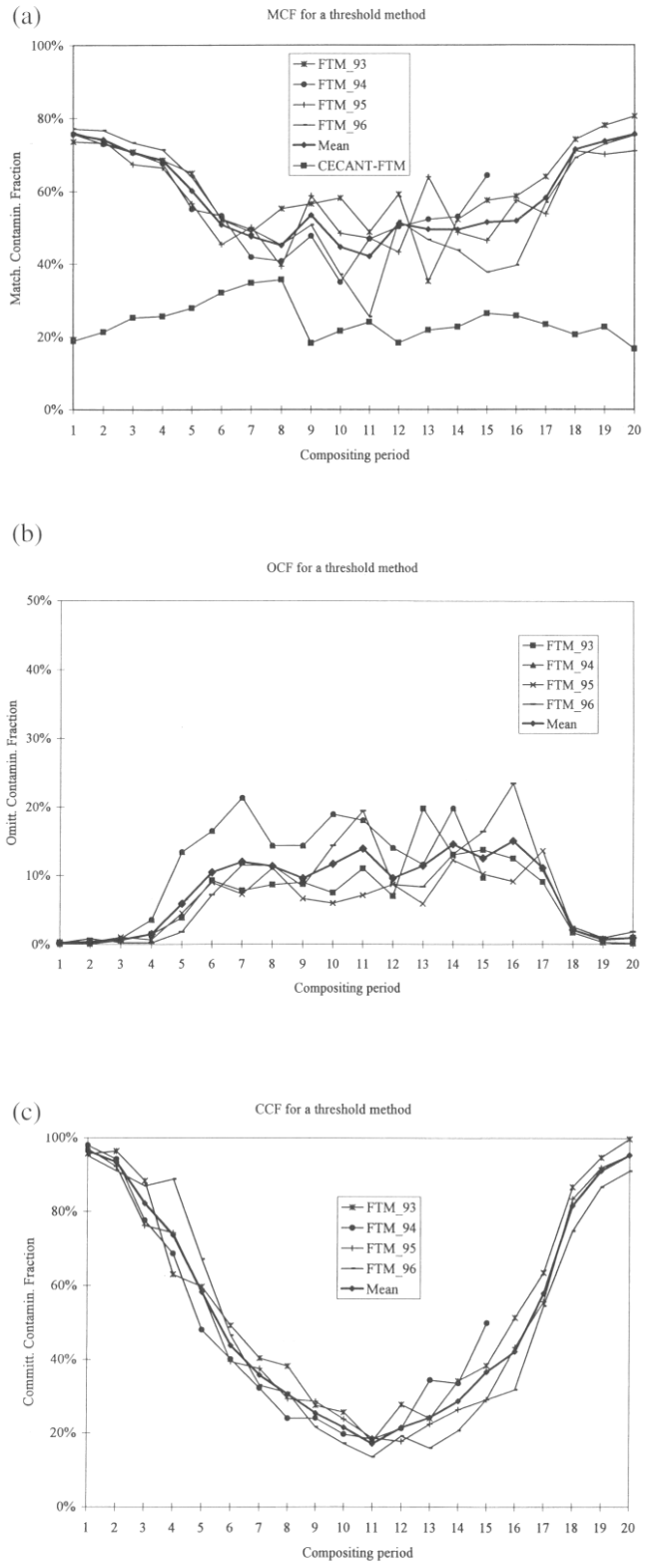



**Figure 7**. Comparison of the contaminated pixels detected by a fixed threshold method (FTM) and by CECANT. Figure 7a: Matching Contaminated Fraction $MCF_{CO}$; Figure 7b: Undetected contaminated fraction, $OCF_{CL}$; Figure 7c: Committed Contaminated Fraction $CCF_{CL}$. The curve CECANT-FTM shows the difference between the two masks (CMA version).

Figure 7 shows the result of this comparison. Overall, the direct identification accuracy ($MCF_{CO}$) of the FTM was below that of CECANT, by 28.7% on average. The two approaches yielded closer results early and late in the season, when the identification is based mostly on surface albedo. During the growing season the discrepancy was above 25%. A relatively small proportion of the contaminated pixels ($OCF_{CL}$) were missed, between 10 and 15% on average during the growing season (Figure 7b) and less outside of it. The proportion of clear-sky pixels erroneously included in the contaminated mask ($CCF_{CL}$) was much higher, from 20-30% in the summer to >70% outside the growing season. The seasonal trend in $CCF_{CL}$ is mostly due to the varying number of clear pixels; the actual number of $CCF_{CL}$ pixels fluctuated much less (between 30000 and 80000) with a minimum in mid-summer. The lower $MFC_{CO}$ and higher $CCF_{CL}$ values suggest that compared with CECANT the FTM tended to identify as contaminated pixels other than those identified by CECANT. Since the FTM thresholds correspond to AVHRR channel 1 of about 0.3 (at TOA) part of the explanation is in the difference between the channel 1 thresholds. In addition, the condition that channel 2 be 1.3 times higher than channel 1 is not met for many northern barren areas, thus labeling as contaminated clear-sky pixels in the north. Furthermore, FTM relies on the context criteria b) to detect partly contaminated pixels. The results do show, however, better performance by CECANT when applied to composite images. While it should be possible to adjust the FTM thresholds the latitudinal variability in vegetation suggests that the thresholds might have to depend on land cover.

It is worth noting that the $OCF_{CL}$ and $CCF_{CL}$ values for the FTM were as low or lower than those for CECANT (Figure 3c vs. 7b, 3f vs. 7c). Also, the year-to-year variability was significantly reduced, especially for $CCF_{CL}$ (Figure 7c). This indicates that the FTM method is less sensitive to the changes in the characteristics of the satellite data set in a particular year, but at the expense of lower $MCF_{CO}$ values.

**Summary**

When processing optical satellite data over land, the ultimate approach will likely consist of screening out fully or partly contaminated pixels from single-date input images, and then retaining the remaining pixels for a comprehensive analysis of the clear-sky measurements. However, most satellite sensors do not provide sufficient spectral and radiometric information to perform such screening accurately and reliably. This is why maximum-value compositing has been widely used. Nevertheless, even with this approach some contamination remains since only the least contaminated pixel is retained for each compositing period. Yet, the detection of such pixels is essential for quantitative applications.

In this paper, we have examined the feasibility of using the CECANT method, previously developed for post-season analysis, in a ‘forward’ mode. Since the method is applied to

composites, it reduces the data processing burden compared to single image-based procedures, a particular advantage for near-real time operations. However, its use implies that the thresholds characterizing contaminated pixels must be based on previous data over a reference period. Thus, they may not be representative of the pixel in the current year. We have examined the impact of using thresholds derived for other periods and, using data sets over Canadian landmass from four years and contamination masks obtained with optimized coefficients, we found that:

1. The absolute accuracy of identifying contaminated pixels ranged from 60 to 100% (Figure 1a-1b), depending on the time in the growing season and the reference year used. When the RZ thresholds were derived from a data set averaged over several years, the accuracy increased to >76% on the average per period and 87% for the year (Figure 1c). Values for individual years/compositing periods could be higher or lower. The relatively high values are partly due to commission errors.

2. When RZ thresholds are applied to another year, the commission errors exceeded errors of omission by about 20% on the average compared to the number of clear-sky pixels (Figure 3). This imbalance can be adjusted by modifying the calculation of $R_{min}$, so that the two errors are more nearly equal, within approximately 10% on average (Figure 4b), partly at the expense of reduced direct identification accuracy (from 87% to 81.1% on average).

3. The best results were achieved using a modified approach (CMA, Figure 5,6) in which the thresholds are computed after the current year data are adjusted with respect to the seasonal trajectories of the reference data set. As a result, the commission and omission errors were more nearly equal and more consistent among years, while the matching accuracy was comparable to or better than for the other approaches.

4. Based on four years of data (including 1994 with different data set characteristics), the mean contaminated fraction could be estimated using CMA with an accuracy of 82.3%, with errors of omission (commission) 12.4% (21.4) on the average. These values are based on the use of Eq.(10) and (11). The interannual stability was very good, with a range of <20% in most cases. The discrepancies in identifying contaminated pixels result from differences between the reference and current year data sets, particularly as regards the NDVI values. Both surface changes and measurement/data processing factors are potential reasons for incorrect identification of contaminated pixels.

5. Using a same-year mask as a reference, the results with the CECANT approach were better by about 20-30% than those obtained with fixed thresholds based on AVHRR channels 1, 2

and 4. The misidentification errors were comparable, especially during the growing season, and they were as or more stable between years than for CECANT.

We conclude that the modified CECANT approach (Eq. (10, (11)) is suitable for the identification of contaminated pixels over land in near-real time processing of AVHRR composite data. However, the identification errors appear sufficiently high to justify reprocessing after the growing season, using the entire data set to optimize the threshold coefficients.

**Acknowledgments**

The data used in this study were composited by the staff of the Manitoba Remote Sensing Centre: Part Hurlburt, Roy Dixon, and Hartley Pokrant. Fenting Huang from Environment Canada and Goran Pavlic from the Canada Centre for Remote Sensing assisted with data processing.

**References**

Allison, E.W., R.J. Brown, H.E Press, and J.G. Gairns. 1989. Monitoring drought affected vegetation with AVHRR. Proceedings of IGARSS'89/12th Canadian Symposium on Remote Sensing, Vancouver, B.C.: 1965-1967.

Cihlar, J., 1996. Identification of contaminated pixels in AVHRR composite images for studies of land biosphere. Remote Senisng of Environment 56: 149-163.

Cihlar, J., and P.M. Teillet. 1995. Forward piecewise linear model for quasi-real time processing of AVHRR data. Canadian Journal for Remote Sensing 21:22-27.

Cihlar, J., J.M. Chen, Z. Li, F. Huang, R. Latifovic, and R. Dixon. 1998. Can interannual land surface signal be discerned in composite AVHRR data? Journal of Geophysical Research (accepted).

Cihlar, J., H. Ly, and Q. Xiao. 1996. Land cover classification with AVHRR multichannel composites in northern environments. Remote Sensing of Environment 58: 36-51.

Cihlar, J., H. Ly, Z. Li, J. Chen, H. Pokrant, and F. Huang. 1997. Multitemporal, multichannel AVHRR data sets for land biosphere studies: artifacts and corrections. Remote Sensing of Environment 60: 35-57.

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

Gutman, G.G. 1992. Satellite daytime image classification for global studies of Earth's surface parameters from polar orbiters. International Journal of Remote Sensing 13: 209-234.

Gutman, G,. D. Tarpley, A. Ignatov, and S. Olson. 1995. The enhanced NOAA global land dataset from the Advanced Very High Resolution Radiometer. Bulletin of the American Meteorological Society 76 (7): 1141-1156.

Gutman, G., D. Tarpley, and G. Ohring. 1987. Cloud screening for determination of land cover characteristics in a reduced resolution satellite data set. International Journal of Remote Sensing 8: 859-870.

Holben, B.N. 1986. Characteristics of maximum-value composite images from temporal AVHRR data. International Journal for Remote Sensing 7: 1417-1434.

James, M.E., and S.N.V. Kalluri. 1994. The Pathfinder AVHRR land data set: an improved coarse resolution data set for terrestrial monitoring. International Journal for Remote Sensing 15: 3347-3363.

Loudjani, P., F. Cabot, V. Gond, and N. Viovy. 1994. Improving NDVI time-series using imposed thresholds on irt, ir and visible values (INTUITIV): a method for reducing cloud contamination and noise in NDVI time-series over tropical and subtropical regions. Proceedings of the International Symposium on "Physical Measurements and Signatures in Remote Sensing, ISPRS Commission VII, January 1994, Val d'Isere, France : 93-102.

Rahman, H., and G. Dedieu. 1994. A SMAC: a simplified method for the atmospheric correction of satellite measurements in the solar spectrum. International Journal of Remote Sensing 15 (1): 123-143.

Robertson, B., A. Erickson, J. Friedel, B. Guindon, T. Fisher, R. Brown, P. Teillet, M. D'Iorio, J. Cihlar, and A. Sancz. 1992. GEOCOMP, a NOAA AVHRR geocoding and compositing system. Proceedings of the ISPRS Conference, Commission 2, Washington, D.C.: 223-228.

Salomonson, V. V. 1988. The Moderate Resolution Imaging Spectrometer (MODIS). IEEE Geoscience and Remote Sensing Newsletter, August 1988: 11-15.

Simpson, J.J. 1994. Cloud detection in AVHRR scenes over land and ocean: a status report. Chapter 5 in: AVHRR Workshop Papers, CSIRO Division of Atmospheric Research, Melbourne, Australia: 41-55.

Simpson, J.J., and J.I. Gobat. 1996. Improved cloud detection for daytime AVHRR scenes over land. Remote Sensing of Environment 55: 21-49.

Smith, R.C.G. 1994. Vegetation watch. Australian Vegetation Watch, Final Report, CSIRO Division of Exploration and Mining. 111p.

Stowe, L.L., R. McClain, R. Carey, P. Pellegrino, G.G. Gutman, P. Davis, C.Long, and S. Hart. 1991. Global distribution of cloud cover derived from NOAA/AVHRR operational satellite data. Advances in Space Research 3: 51-54.

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.

Wu, A., Z. Li, and J. Cihlar. 1995. Effects of land cover type and greenness on AVHRR bidirectional reflectances: analysis and removal. Journal of Geophysical Research 100:9179-9192.