



Cloud mask via  
cumulative  
discriminant analysis

U. Amato et al.

# Cloud mask via cumulative discriminant analysis applied to satellite infrared observations: scientific basis and initial evaluation

U. Amato<sup>1</sup>, L. Lavanant<sup>2</sup>, G. Liuzzi<sup>3</sup>, G. Masiello<sup>3,4</sup>, C. Serio<sup>3,4</sup>, R. Stuhlmann<sup>5</sup>,  
and S. A. Tjemkes<sup>5</sup>

<sup>1</sup>IAC/CNR, Napoli, Italy

<sup>2</sup>Meteo-France/DP/Centre de Meteorologie Spatiale BP 50747 22307 Lannion, France

<sup>3</sup>School of Engineering, University of Basilicata, Potenza, Italy

<sup>4</sup>CNISM, Research Unit of Potenza, University of Basilicata, Potenza, Italy

<sup>5</sup>EUMETSAT, Darmstadt, Germany

Received: 20 April 2014 – Accepted: 23 May 2014 – Published: 6 June 2014

Correspondence to: C. Serio (carmineserio@unibas.it)

Published by Copernicus Publications on behalf of the European Geosciences Union.

Title Page

Abstract

Introduction

Conclusions

References

Tables

Figures



Back

Close

Full Screen / Esc

Printer-friendly Version

Interactive Discussion



## Abstract

We introduce a classification method (Cumulative Discriminant Analysis) of the Discriminant Analysis type to discriminate between cloudy and clear sky satellite observations in the thermal infrared. The tool is intended for the high spectral resolution infrared sounder (IRS) planned for the geostationary METEOSAT (Meteorological Satellite) Third Generation platform and uses IASI (Infrared Atmospheric Sounding Interferometer) data as a proxy. The Cumulative Discriminant Analysis does not introduce biases intrinsic with the approximation of the probability density functions and is flexible enough to adapt to different strategies to optimize the cloud mask. The methodology is based on nine statistics computed from IASI spectral radiances, which exploit the high spectral resolution of the instrument and which effectively summarize information contained within the IASI spectrum. A Principal Component Analysis prior step is also introduced which makes the problem more consistent with the statistical assumptions of the methodology. An initial assessment of the scheme is performed based on global and regional IASI real data sets and cloud masks obtained from AVHRR (Advanced Very High Resolution Radiometer) and SEVIRI (Spinning Enhanced Visible and Infrared Imager) imagers. The agreement with these independent cloud masks is generally well above 80 %, except at high latitudes in their winter seasons.

## 1 Introduction

Modern meteorological satellites carry infrared sensors onboard able to sense the Earth emission spectrum at very high spectral resolution. These include, e.g., AIRS (Atmospheric Infrared Sounder), IASI (Infrared Atmospheric Sounding Interferometer), CrIS (Cross-track Infrared Sounder). All of these spectrometers are characterized by broadband spectral coverage (3.7–15.5  $\mu\text{m}$ ) and a spectral sampling rate in the range 0.25–2  $\text{cm}^{-1}$ .

AMTD

7, 5601–5650, 2014

## Cloud mask via cumulative discriminant analysis

U. Amato et al.

Title Page

Abstract

Introduction

Conclusions

References

Tables

Figures



Back

Close

Full Screen / Esc

Printer-friendly Version

Interactive Discussion







## Cloud mask via cumulative discriminant analysis

U. Amato et al.

Title Page

Abstract

Introduction

Conclusions

References

Tables

Figures



Back

Close

Full Screen / Esc

Printer-friendly Version

Interactive Discussion



of surface and, in some cases, on the time of the day. MAIA cloud mask (Lavanant et al., 2007) developed for the AVHRR onboard Metop uses a similar approach. For MODIS (Moderate Resolution Imaging Spectroradiometer) (Ackerman et al., 1998, 2008) nine statistics and corresponding tests are introduced each of them involving two or three wavelengths at most; in the case of SEVIRI (Spinning Enhanced Visible and Infrared Imager) (EUMETSAT, 2009) about ten statistics and corresponding tests are defined depending on the time of the day (daytime, sun glint, twilight, nighttime) and on the surface type (land or sea). Often tests based on spatial coherence statistics are also introduced, that rely on the different spatial correlation in images in clear and cloudy conditions (Sandhya et al., 2004). Besides statistics strictly based on radiance (thresholding, comparison of radiances or their ratios), many others have been developed aimed at picking particular features that are different in clear and cloudy conditions. For example the  $h_s$ -index is introduced in Serio et al. (2000) based on correlation and cross-correlation of spectral radiances in the atmospheric infrared window. Many other methodologies for cloud detection have been developed in the last decade in the frameworks of statistics and machine learning. It is not the purpose of the present paper to discuss them here (we defer to Tapakis and Charalambides, 2012, for a recent review) because they will not be used for the reasons that are going to be discussed. In addition, we will focus mostly on the problem of discriminating the presence of cloud contamination in the radiances, but not on cloud classification in terms of phase and top pressure and temperature.

One of the most appropriate approaches for cloud detection within the footprint or field of view of a sounding instrument is its co-location with a suitable imager radiometer, or even better to design and develop the sounder with a built-in imager. Both approaches have been developed for IASI. However, the built-in imager has only two channels and is not appropriate for cloud detection. Co-location of IASI footprint with AVHRR imagery has been used, e.g., by Lavanant et al. (2007). However, technological constraints, such as envisaged for MTG-IRS, or simply the need of real-time processing, as required, e.g., in NWP applications, can severely limit the synergy between



The present paper is mostly concerned with the scientific basis of CDA. An initial evaluation of the cloud mask is performed with global and regional IASI data complemented with cloud masks derived from AVHRR and SEVIRI.

The paper is organized as follows. Section 2 deals with the data used for developing statistics, training and validating the cloud detection scheme. The series of statistics defined and used in the paper are described in Sect. 3. Section 4 includes the full statistical development of CDA. Experiments are shown in Sect. 5, while Conclusions are drawn in Sect. 6.

## 2 Data sets

We have individuated and developed two data sets of IASI spectra to train, cross-check and validate the cloud detection scheme. These are referred to as IASI1 and IASI2 and their characteristics are listed below:

- IASI1. It comprises 888 380 IASI spectra corresponding to a 12 h global acquisition on 17–18 November 2009 Lavanant et al. (2011). The set is qualified for sky-type (clear/cloudy) through the CMS (Centre de Meteorologie Spatiale, Lannion, France) cloud mask Lavanant et al. (2011), which is based on the co-location of the IASI footprint with AVHRR imagery.
- DIFA2. It comprises 1 072 050 spectra corresponding to a 15 h global acquisition on 22–23 July 2007. Also in this case the set is qualified for sky-type (clear/cloudy) through the CMS cloud mask.

Both sets have been endowed with skin temperature fields from the ECMWF (European Center for Medium range Weather Forecasts) analysis, co-located in space and time with the IASI footprints.

To exemplify the quantity and quality of the data set sea and land surface, Fig. 1 shows the AVHRR cloud mask (co-located at the level of the IASI pixel) for sea and

## Cloud mask via cumulative discriminant analysis

U. Amato et al.

Title Page

Abstract

Introduction

Conclusions

References

Tables

Figures



Back

Close

Full Screen / Esc

Printer-friendly Version

Interactive Discussion



land, respectively. The Figure refers to the data set IASI2. For sea surface we have approximately 12 % of clear sky IASI fields of view, whereas for land this figure grows up to about 38 %.

To take into account the possible dependence of the test statistics and of ancillary data (reference spectra and ECMWF skin temperature) on season and air mass type, we have defined 10 climatic zones. To increase statistics homogeneity, some models are also split in submodels according to the hemisphere (northern, NH, or southern, SH) or to the time of the day. The climate zones are listed in Table 1.

It is important to stress that the definition of the climatic zones 1 to 5 does not include sea ice or snow/ice in case of land. For these models we assume the skin temperature to be above the freezing point of 273 K. The cases of sea ice and land ice/snow are also split with latitude. Also note that we have included Antarctica as a special climate zone, because this region is covered with permanent ice.

Only for the tropical zone the two data sets play a complementary role (one for training and the other one for validation). In the other climatologies some data sets are populated with only few IASI spectra, which makes it impossible to use them as training and/or validation data sets; therefore we had to merge them with homogeneous zones in order to reach a statistically significant size. In some case, we had no data at all and the cloud mask thresholds were not produced.

This is indeed a limitation for operational uses. However, as already mentioned the main goal of this paper is to show the scientific basis of the methodology and to outline its initial evaluation.

## 2.1 Silver standard

Choice of the reference training data set is an important step in classification, because properties of the statistics in clear and cloudy sky conditions are estimated from it. Due to its importance, this data set is also called “gold” standard, for which the class label of each sample of the data set is assumed to be exactly known. Often, this is obtained by an expert who manually trains each sample of the data set and assigns

## Cloud mask via cumulative discriminant analysis

U. Amato et al.

Title Page

Abstract

Introduction

Conclusions

References

Tables

Figures



Back

Close

Full Screen / Esc

Printer-friendly Version

Interactive Discussion



## Cloud mask via cumulative discriminant analysis

U. Amato et al.

Title Page

Abstract

Introduction

Conclusions

References

Tables

Figures



Back

Close

Full Screen / Esc

Printer-friendly Version

Interactive Discussion



the class. In practice in many applications this is not possible, or it is possible only for a limited sample; as a consequence the sample is not fully representative of the full population and the statistical properties estimated from the data set are not accurate. In the case of cloud detection there is a further problem coming from the fact that to recognize clouds from satellite imagery can be difficult even for a well-trained expert in particular conditions like nighttime or cirrus. When a “gold standard” is not available or not fully representative, it is usual to choose a training data set starting from the results of another classification algorithm that is proved to be very reliable. In this case this data set is called “silver” standard. The silver standard relies on the CMS cloud mask which has been used as the IASI reference cloud mask.

### 2.2 The CMS reference cloud mask

The IASI reference cloud mask we have used in our analysis has been developed at CMS (Lavanant et al., 2011); in turn, it is based on the cloud mask derived from the AVHRR data at full resolution (Lavanant et al., 2007).

AVHRR pixels are collocated with IASI footprints. Within each IASI footprint, the cloud fraction,  $C_f$ , is determined according to the ratio

$$C_f = 100 \frac{N_{\text{cloudy}}}{N_{\text{total}}} \times 100,$$

with  $N_{\text{cloudy}}$  and  $N_{\text{total}}$  being the number of AVHRR cloudy and total pixels, respectively.  $C_f$  ranges from 0 to 100, with 0 corresponding to a totally clear IASI observation. For the present analysis, the training of the IASI cloud mask has been performed considering a threshold  $C_f = 5$ , that is the IASI footprint is considered clear if the AVHRR-based cloud fraction is below or equal to 5 %.

The CMS-AVHRR cloud detection scheme is based on a series of tests, which use the AVHRR observations in its six channels, and take advantage of a-priori or background information for the surface and atmospheric parameters. The a-priori state vector is based on the ECMWF 12–18 h forecast for the total water vapor content and land

## Cloud mask via cumulative discriminant analysis

U. Amato et al.

Title Page

Abstract

Introduction

Conclusions

References

Tables

Figures



Back

Close

Full Screen / Esc

Printer-friendly Version

Interactive Discussion



surface temperature and on climatologies for the other parameters (Lavanant et al., 2007, 2011). In addition to the tests based on the visible channels, the AVHRR cloud detection scheme also uses a series of thermal infrared Brightness Temperature tests and a series of local uniformity tests, which are designed to detect cloud edges, thin cirrus and small cumulus, by using their high spatial variations in the visible, near infrared or infrared channels.

Comparisons of CMS with CLAVR-x show an agreement which is close to 90 % (Lavanant et al., 2007). To have an idea of the quality and accuracy of AVHRR based cloud masks, we quote the validation of CLAVR-x (Heidinger et al., 2012) on the basis of a data set that coincides with the training one. The validation indicates that the cloud mask is estimated with an accuracy ranging from 78 % to 94 % according to the land type (best for deep ocean, worst for Antarctica). Because of the good agreement between CLAVR-x and CMS, we can conclude that these figures are representative of the accuracy of the CMS cloud mask as well.

For the specific case of IASI, the CMS cloud mask has been intercompared with other IASI cloud detection schemes in Lavanant et al. (2011). The best agreement is achieved with schemes relying themselves on the AVHRR imager. The agreement ranges from 78 % to 85 %.

### 2.3 Validation data set based on the SEVIRI cloud mask

To have an independent validation data set, we have also used the SEVIRI (Spinning Enhanced Visible and Infrared Imager) cloud mask (Derrien and Le Gléau, 2005). A portion of the data set IASI2, covering the Africa Continent, has been co-located with SEVIRI imagery and the SEVIRI cloud mask has been used to identify clear and cloudy IASI pixels.

A second, smaller, independent validation data set makes use of observations collected over Europe/Africa from 25 September to 4 October 2012 in the framework of the Inter-calibration IASI/SEVIRI experiment, performed within the activities of EUMETSAT GSICS (Global Space-based Inter-Calibration System).



The two BT  $T_{790.5}$  and  $T_{791.75}$  are defined at specific wave numbers. The other temperatures are computed as averages over the corresponding spectral ranges listed in Table 2:

$$\bar{T}_\sigma = \frac{1}{N_\sigma} \sum_{\tau=\sigma_1}^{\sigma_2} T_\tau,$$

where  $\sigma$  is the central wave number of the temperatures (832, 874, 900, 1168, 2003 and 2700  $\text{cm}^{-1}$ , orderly),  $\sigma_1$  and  $\sigma_2$  are the range of the spectral interval for each central wavenumber (see Table 2), and  $N_\sigma$  is the number of spectral ordinates within the spectral range.

Table 3 summarizes the nine statistics computed from the IASI radiances. It should be stressed that at the present stage of defining statistics we do not discuss the estimate of the thresholds for discriminating clear and cloudy pixels, because this task will be accomplished with the CDA methodology (Sect. 4). CDA objectively determines the appropriate threshold for each statistic.

The four statistics,  $W_1$ ,  $W_2$ ,  $W_3$ ,  $W_4$  are Inoue-like window slope statistics (Inoue, 1985; Inoue and Ackerman, 2002).  $W_1$  is the classical Inoue slope test, which is highly sensitive to cirrus clouds.  $W_2$  is a variant of  $W_1$  and is mostly effective in case of surface features rapidly changing with the wave number, such as desert sand. The statistic  $W_4$  has been widely used in AVHRR-based cloud detection (e.g., Lavanant et al., 2007) and it is motivated by the fact that the cirrus and stratus cloud types have a reflectance at 3.9  $\mu\text{m}$  which is higher than that of most surface features. The channel at 3.9  $\mu\text{m}$  has been used both in nighttime and daytime. Among the four  $W$  statistics,  $W_3$  is the most original and has been defined to be sensitive to low and thin water clouds. The water droplet mode radius of most cloud types is 5  $\mu\text{m}$  (e.g., Liou, 1992), which means that at 5  $\mu\text{m}$  (2000  $\text{cm}^{-1}$ ) scattering effects dominate over absorption. The reverse happens at 12  $\mu\text{m}$  (833  $\text{cm}^{-1}$ ) where the absorption dominates over scattering. Thus, in the presence of a semi-transparent water cloud a strong contrast has to be expected between the BT at 5  $\mu\text{m}$  and that at 12  $\mu\text{m}$ .  $W_3$  is expected to play a significant role over land

**Cloud mask via cumulative discriminant analysis**

U. Amato et al.

Title Page	
Abstract	Introduction
Conclusions	References
Tables	Figures
◀	▶
◀	▶
Back	Close
Full Screen / Esc	
Printer-friendly Version	
Interactive Discussion	



## Cloud mask via cumulative discriminant analysis

U. Amato et al.

Title Page

Abstract

Introduction

Conclusions

References

Tables

Figures



Back

Close

Full Screen / Esc

Printer-friendly Version

Interactive Discussion



in detecting hot cirrus clouds at nighttime. Over land, because of emissivity, the two equivalent, long-wave, statistics,  $W_1$  and  $W_2$  can have a variability in clear sky, which is even much larger than that expected for cloudy sky. In contrast, we expect that  $W_3$  can assume only positive values in clear sky and negative ones in case of cloudy scenes.

Slope window test would be almost useless without suitable driver thermal contrast tests which capture the physical evidence that, normally, the land is warmer than cloud top. We have basically two thermal contrast statistics. The first is just the brightness temperature at  $833\text{ cm}^{-1}$ , that is  $T_0$ . Sea surface emissivity has a peak at  $833\text{ cm}^{-1}$ , and at this wave number natural land features have the smallest emissivity variability. In clear sky conditions, independently of the kind of surface, the channel at  $833\text{ cm}^{-1}$  is the brightest point in the spectrum.

The spatial homogeneity statistic, sh, is similar to those used with AVHRR-based cloud detection. It uses the standard deviation of  $T_0$  corresponding to a cluster of  $2 \times 2$  IASI pixels. This statistic is normally very well suited to detect cloud edges, however for IASI it has not proved to be effective because of the discontinuous IASI scan pattern geometry and the relatively large field of view. By the way, in contrast it is expected to be very effective for MTG-IRS because of the imaging capability of the instrument and the smaller field of view.

The  $\text{CO}_2$  in-band out-band statistic,  $\Delta T_{\text{CO}_2}$  (Masiello et al., 2003), exploits the strong thermal contrast which is present in clear sky conditions between the lower and upper troposphere. It is based on the absorption feature of the weak  $\text{CO}_2$  Q-branch at  $791\text{ cm}^{-1}$ .  $\text{CO}_2$  absorption yields a very well defined and sharp spectral feature centered at  $791.75\text{ cm}^{-1}$  in between of a window region with weak water vapor absorption.

The  $\chi_s^2$  statistic makes use of information provided by NWP forecasts and/or analysis for sea and surface temperature,  $T_s$ . This temperature is contrasted with that directly estimated by IASI time-space co-located observations.

For the case of sea surface, the IASI estimate for  $T_s$  relies on the classical split window algorithm for the estimation of skin temperature. It is based on the two temperatures  $\bar{T}_{874}$  and  $\bar{T}_{900}$ . The algorithm we have developed for IASI is dependent on the

## Cloud mask via cumulative discriminant analysis

U. Amato et al.

Title Page

Abstract

Introduction

Conclusions

References

Tables

Figures



Back

Close

Full Screen / Esc

Printer-friendly Version

Interactive Discussion



field of view angle and the type of air mass. For the calculation of the regression coefficients involved in the algorithm we have assumed Northern and Southern Hemisphere to be climatically equivalent. For practical computations, one needs a suitable data base of atmospheric and surface parameters. To this end, we have used the ECMWF Chevalier data base (Chevalier, 2001). The accuracy of the split window technique ranges in

between 0.3–1 K, depending on the columnar load of water vapour. The statistic  $\chi_s^2$  for sea surface compares  $\hat{T}_s$ , estimated by the IASI split window technique, against a suitable reference,  $T_{s,R}$ . We have

$$\chi_s^2 = \frac{(\hat{T}_s - T_{s,R})^2}{v_I^2 + v_R^2},$$

where  $v_I^2$ ,  $v_R^2$  are the squared uncertainties (variances) of  $\hat{T}_s$  and  $T_{s,R}$ , respectively.

If appropriately built up, and in absence of biases affecting estimated and reference  $T_s$ , the statistic  $\chi_s^2$  is distributed according to a  $\chi^2$  variable with one degree of freedom. As far as  $v_R^2$  is concerned, its value depends on the quality of the reference. In the present implementation,  $T_{s,R}$  is obtained from the ECMWF analysis, which is generally recognized to be accurate within 1 K. Again, a detailed account of the IASI split window and  $\chi^2$ -test for sea surface can be found in Serio et al. (2013).

For sea surface, the  $\chi^2$  test is mostly powerful; the quality of the split window and, the overall reliability of the statistic have been variously assessed through IASI spectra co-located in time and space with ECMWF analysis. Figure 2 shows an example obtained over the Mediterranean area with a sample of spectra covering the period March–October 2010 and qualified for clear sky according to the methodology developed in this work.

For land, mostly motivated by the lack of a robust and fast method to estimate the surface temperature, the  $\chi^2$  test is substituted by a statistic which considers the difference

$$|\bar{T}_{832} - T_{s,R}|,$$

## Cloud mask via cumulative discriminant analysis

U. Amato et al.

Title Page

Abstract

Introduction

Conclusions

References

Tables

Figures



Back

Close

Full Screen / Esc

Printer-friendly Version

Interactive Discussion



where again the reference temperature is obtained from the ECMWF analysis. In other words,  $\bar{T}_{832}$  is considered as a surrogate of the surface temperature. The fact that  $\bar{T}_{832}$  could be affected by water vapour is not of any concern here, because our aim is not to estimate the surface temperature, but to have a statistic which is as much as different in clear and cloudy conditions.

Unlike the statistics so far described, spatial homogeneity statistic  $h_s$  (Maseillo et al., 2002; Serio et al., 2000; Masiello et al., 2003) fully exploits the hyperspectral capabilities of IASI and in perspective of MTG-IRS.

Basically, it is designed to exploit the unique spectral signature of sea/land surface in the atmospheric window region  $800\text{--}950\text{ cm}^{-1}$ . It uses the observed spectrum and a clear-sky reference one,  $T_\sigma^O$  and  $T_\sigma^R$ , respectively, both converted to BT, according to the formula

$$h_s = \frac{\sum_{j=1}^{N_L} |r_j^O - r_j^R|}{\sum_{j=1}^{N_L} |r_j^{OR}|},$$

where  $r_j^O$ ,  $r_j^R$  and  $r_j^{OR}$  are correlation and cross-correlation coefficients of  $(T_\sigma^O, T_\sigma^R)$  and  $N_L$  is the number of considered wavenumber lags (e.g., Maseillo et al., 2002; Serio et al., 2000). In the present paper the reference spectra for sea surface are obtained by  $\sigma$ -IASI (Amato et al., 2002) with a set of atmospheric profiles derived from the Chevalier data set (Chevalier, 2001) using Masuda's emissivity model for sea surface (Masuda et al., 1988). In the case of land surface, uncertainty associated with surface emissivity could make the method less effective, mostly due to troubles in defining a suitable reference spectrum. Therefore we have developed an approach that is as independent of surface emissivity as possible: instead of the radiance spectrum, we consider its first difference

$$\Delta R_\sigma = R_{\sigma+\Delta\sigma} - R_\sigma,$$

which is an approximation of the differential of the spectrum with respect to the wavenumber,  $\sigma$ . This is a high-pass filter, which removes the relatively smooth, hence, slow component introduced from surface emissivity.

Figure 3 shows a demonstration of the procedure, comparing  $R_\sigma$  and  $\Delta R_\sigma$  within the spectral range  $[800, 950] \text{ cm}^{-1}$  for very different surface emissivities. It is seen that in the most transparent regions the filter largely suppresses the surface emission component and leaves the atmospheric line component unaffected. This atmospheric component is largely affected by the presence of clouds.

In perspective for MTG-IRS the use of computed reference spectra to calculate  $h_s$  is envisaged at the very beginning its operational life. Once we have collected enough observations, it is desirable to use directly MTG-IRS observations, suitably screened for clear sky.

To summarize, among the nine statistics we consider for the problem of cloud detection,  $W_1, W_2, W_4, T_0, \chi_s^2$ , sh are heritage of AVHRR/MODIS cloud detection algorithms;  $W_3$  is quite new since the channel at  $2003 \text{ cm}^{-1}$  is not available with current satellite radiometer imagers; the thermal contrast statistic  $\Delta T_{\text{CO}_2}$  relies on the high spectral resolution of IASI and, in perspective, of MTG-IRS and cannot be designed, e.g., for imagers like AVHRR. Finally,  $h_s$  is specialized for hyperspectral instrumentation and heavily relies on the concept of spectrum and/or spectral radiance, as opposite to the coarse spectral resolution of radiance which up to now has characterized coarse and moderate spectral resolution satellite imagers, such as AVHRR, MODIS, SEVIRI.

## 4 Cumulative discriminant analysis

Estimate of suitable thresholds for each statistic is required to produce a cloud mask. This is accomplished by (a) choosing the most effective statistics among the considered ones; (b) actually determining the corresponding thresholds. Towards this objective, this section introduces the Cumulative Discriminant Analysis, that is the engine that will drive the discrimination/classification methodology for producing the final cloud mask.

### Cloud mask via cumulative discriminant analysis

U. Amato et al.

Title Page

Abstract

Introduction

Conclusions

References

Tables

Figures



Back

Close

Full Screen / Esc

Printer-friendly Version

Interactive Discussion



It is important to stress that we look at discrimination/classification in a probabilistic fashion. Also the CDA engine has to be run only once to discriminate the most effective statistics and to generate thresholds. This is done on the basis of a suitable training data set. Once we get the thresholds, we do not need to run CDA again to apply the scheme to a given IASI spectrum. Of course, the full procedure is needed in case the training data set is changed.

## 4.1 Notations

We begin with some notations to make clear conventions that will be used in the paper. We assume *clear* as the target sky condition. Then Type I error is defined as the fraction of pixels being clear and classified as cloudy; this fraction is also called False Positives, FP, or miss. Analogously we define as Type II error the fraction of pixels being cloudy and classified as clear (also called False Negatives, FN, or false alarm). The fraction of pixels exactly classified as Clear or Cloudy will be denoted by True Positives, TP, and True Negatives, TN, respectively. The performance of the classification scheme can be summarized through the fraction,  $S$ , of successful classifications,

$$S = \frac{TPN_{\text{clear}} + TN_{\text{cloudy}}}{N_{\text{clear}} + N_{\text{cloudy}}}, \quad (1)$$

with  $N_{\text{clear}}$  and  $N_{\text{cloudy}}$  being the number of true clear and cloudy pixels, respectively.

As a graphical representation (see Fig. 4, left) we plot the probability density function of two classes, with  $x$  being a generic statistic. We also assume that the classifier acts by means of a threshold  $\theta$  that gives back Class 1 if  $x \leq \theta$  and Class 2 if  $x > \theta$ . Therefore in the plot True Negatives are given by the area below the higher peaked function (continuous line) up to the threshold  $\theta$  (the black and light gray areas); True Positives are given by the area below the lower peaked function (dashed line) from the threshold upwards (dark gray and white areas). Analogously, False Positives (Type I error) are given by the dark gray area and False Negatives (Type II error) by the

## Cloud mask via cumulative discriminant analysis

U. Amato et al.

Title Page

Abstract

Introduction

Conclusions

References

Tables

Figures



Back

Close

Full Screen / Esc

Printer-friendly Version

Interactive Discussion



light gray area. An alternative graphical representation can be given in terms of the cumulative density functions (Fig. 4, right).

Intercept of the threshold line with the two cumulative functions for the two classes locates the True Negatives (below the intercept) and the False Positives (above) from the curve of the first class and the True Positives (above the intercept) and False Negatives (below) from the curve of the second class.

## 4.2 Nonparametric estimates

Many classification methodologies require estimate of the probability density function of the classes. While this can be accomplished relying on parametric methods, however most density functions coming from real applications are not well approximated by such methods for their poor ability to fit actual distributions. For this reason nonparametric density functions are resorted where the shape of the density functions is not assigned from the beginning through a family of functions but it is the result of the approximation. The most widespread of such methods is Kernel density estimation (Silverman, 1986). Despite of its popularity, however nonparametric density estimation requires tuning of one or more parameters (choice of the Kernel function, bandwidth). In particular, bandwidth heavily affects accuracy of the nonparametric density estimate through its smoothness: the larger the bandwidth (i.e., the width of the Kernel function), the smoother the density estimate. Some criteria, both based on asymptotic arguments and on the actual data, have been developed (Wand and Jones, 1995) for its estimation, however choice of the bandwidth keeps on being the crux of the methodology.

On the other hand if we consider the cumulative density function, an estimator exists that enjoys many optimal theoretical properties and is very attractive from the computational point of view: the empirical cumulative density function  $F(\vartheta)$ . Given the sample  $X_i, i = 1, \dots, N$ , it is easily defined as

## Cloud mask via cumulative discriminant analysis

U. Amato et al.

Title Page

Abstract

Introduction

Conclusions

References

Tables

Figures



Back

Close

Full Screen / Esc

Printer-friendly Version

Interactive Discussion



## Cloud mask via cumulative discriminant analysis

U. Amato et al.

Title Page

Abstract

Introduction

Conclusions

References

Tables

Figures



Back

Close

Full Screen / Esc

Printer-friendly Version

Interactive Discussion



$$F(\vartheta) = \frac{\text{Number of elements } \leq \vartheta}{N} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}\{x_i \leq \vartheta\}.$$

The most important result is that the estimator almost surely converges to the true cumulative function asymptotically; therefore the estimator is consistent (van der Vaart, 1998). In addition convergence holds uniformly over  $\vartheta$ . From the computational point of view the estimator is simply computed by sorting the sample  $x_i$ ,  $i = 1, \dots, N$ , in  $O(N \log N)$  operations. We observe that no parameter exists to be tuned.

### 4.3 Cumulative discriminant analysis

Basing on the theoretical properties of the cumulative density functions, a CDA methodology is proposed for the classification.

We first consider the univariate case corresponding to one statistic. We suppose that the classifier is of the Discriminant Analysis type where the decision rule  $\Gamma(x, \mathbf{x})$  is based on a threshold  $\vartheta$  as

$$\Gamma(x; \mathbf{x}) = \begin{cases} 1 \text{ (Clear)} & \text{if } x \leq \vartheta \\ 2 \text{ (Cloudy)} & \text{otherwise,} \end{cases} \quad (2)$$

where  $\mathbf{x} \equiv \{x_1, \dots, x_N\}$  is the sample statistic of size  $N$  that defines the training data set and  $x$  is the actual statistic to be classified basing on the training data set.

Classical Discriminant Analysis would yield a threshold rule similar to Eq. (2) under the assumption of Gaussianity of the density function of the statistic for both Clear and Cloudy conditions, which does not occur in practice. By relying on the CDA we refrain from assuming this hypothesis and on the contrary we use an optimal estimate of the cumulative density function.

Now an estimate of the threshold has to be given in order that the method is fully operative. To this purpose we define a Cost function  $\mathcal{C}(x, \vartheta)$  whose minimization gives



## Cloud mask via cumulative discriminant analysis

U. Amato et al.

Title Page

Abstract

Introduction

Conclusions

References

Tables

Figures

◀

▶

◀

▶

Back

Close

Full Screen / Esc

Printer-friendly Version

Interactive Discussion



In the present project we use the Cost function (Eq. 4), since the objective is to estimate the rate of misclassifications for both Clear and Cloudy conditions. This choice is the least noncommittal with respect to the average conditions of the sky, cloudy or clear. We observe that it corresponds to assume that Clear and Cloudy classes have equal size, which has a counterpart with uniform prior of Discriminant Analysis. In the case one would take account of real or different relative size of the Clear and Cloudy sample, a weighted CDA could be devised by introducing proper weights to the Type I and II errors, in analogy with Classical Discriminant Analysis.

Note that at the minimum of the Cost function (4) it is  $E^I = E^{II}$ , that is Type I and II errors coincide.

Finally we mention that in minimizing the Cost function with respect to the threshold  $\vartheta$  we have to include in the decision rule also the sign direction of the rule, that is  $x \leq \vartheta$  or  $x \geq \vartheta$  for Clear conditions. In the latter case the decision rule becomes

$$\Gamma(x; \mathbf{x}) = \begin{cases} 1 \text{ (Clear)} & \text{if } x \geq \vartheta \\ 2 \text{ (Cloudy)} & \text{otherwise,} \end{cases} \quad (5)$$

and the Type I and II errors are written as

$$E^I = F^{\text{Clear}}(\vartheta), \quad E^{II} = 1 - F^{\text{Cloudy}}(\vartheta).$$

#### 4.4 Multivariate CDA

Formulas of the Cost function can be generalized to the case of  $D$  statistics giving rise to a procedure that can be implemented effectively.

Generalization of the decision rule (Eq. 2) results in

$$\Gamma(x^1, x^2, \dots, x^D; \mathbf{x}) = \begin{cases} 1 \text{ (Clear),} & \text{if } x^d \leq \vartheta_d, \quad 1 \leq d \leq D; \\ 2 \text{ (Cloudy),} & \text{otherwise;} \end{cases} \quad (6)$$

analogously, generalization of the rule (Eq. 5) becomes

$$\Gamma(x^1, x^2, \dots, x^D; \mathbf{x}) = \begin{cases} 1 \text{ (Clear)}, & \text{if } x^d \geq \vartheta_d, 1 \leq d \leq D; \\ 2 \text{ (Cloudy)}, & \text{otherwise.} \end{cases} \quad (7)$$

In a similar way all  $2^D$  decision rules with mixed inequality signs can be built.

We assume that the theory underlying the computation of the Type I and II Errors falls in the framework of the probability of the union of 2 or more events that are not mutually exclusive. For such events in the case of decision rule (Eq. 7) we have that the generalization of the Type I Error to the two-dimensional case is

$$E^I = F_1^{\text{Clear}}(\vartheta_1) + F_2^{\text{Clear}}(\vartheta_2) - F_1^{\text{Clear}}(\vartheta_1)F_2^{\text{Clear}}(\vartheta_2).$$

Generalization to the three-dimensional case is

$$\begin{aligned} E^I = & F_1^{\text{Clear}}(\vartheta_1) + F_2^{\text{Clear}}(\vartheta_2) + F_3^{\text{Clear}}(\vartheta_3) \\ & - F_1^{\text{Clear}}(\vartheta_1)F_2^{\text{Clear}}(\vartheta_2) - F_1^{\text{Clear}}(\vartheta_1)F_3^{\text{Clear}}(\vartheta_3) - F_2^{\text{Clear}}(\vartheta_2)F_3^{\text{Clear}}(\vartheta_3) \\ & + F_1^{\text{Clear}}(\vartheta_1)F_2^{\text{Clear}}(\vartheta_2)F_3^{\text{Clear}}(\vartheta_3); \end{aligned} \quad (8)$$

and so on for higher  $D$ .

Similar formulas, not written here for the sake of brevity, can be devised for  $E^{II}$  and other decision rules. For example for the decision rule (Eq. 6) and  $D = 3$ , the Type I and II Errors of (Eq. 3) generalize to

$$\begin{aligned} E^I &= 1 - F_1^{\text{Clear}}(\vartheta_1)F_2^{\text{Clear}}(\vartheta_2)F_3^{\text{Clear}}(\vartheta_3) \\ E^{II} &= F_1^{\text{Cloudy}}(\vartheta_1)F_2^{\text{Cloudy}}(\vartheta_2)F_3^{\text{Cloudy}}(\vartheta_3) \end{aligned} \quad (9)$$

## Cloud mask via cumulative discriminant analysis

U. Amato et al.

Title Page

Abstract

Introduction

Conclusions

References

Tables

Figures



Back

Close

Full Screen / Esc

Printer-friendly Version

Interactive Discussion



By the way, Eq. (9) can be obtained from Eq. (8) and vice-versa with abuse of notation via

$$F_1^{\text{Clear}}(\vartheta_1) = 1 - F_1^{\text{Clear}}(\vartheta_1), F_1^{\text{Cloudy}}(\vartheta_1) = 1 - F_1^{\text{Cloudy}}(\vartheta_1);$$

$$F_2^{\text{Clear}}(\vartheta_2) = 1 - F_2^{\text{Clear}}(\vartheta_2), F_2^{\text{Cloudy}}(\vartheta_2) = 1 - F_2^{\text{Cloudy}}(\vartheta_2);$$

$$F_3^{\text{Clear}}(\vartheta_3) = 1 - F_3^{\text{Clear}}(\vartheta_3), F_3^{\text{Cloudy}}(\vartheta_2) = 1 - F_3^{\text{Cloudy}}(\vartheta_3).$$

## 4.5 Limitations of CDA

CDA is based on the assumption that the decision rule involves one threshold for each statistic. If we recall how a pdf-based Discriminant Analysis works, this means that the threshold has to split the domain of the statistic in two parts and in each of them one of the pdf in turn has to be higher than the other one in general. Visually, this occurs when the two pdf's look to be translated between each other. From the cumulative function point of view, we require that the two functions do not cross (or cross at very far tails). This condition strictly occurs with parametric Discriminant Analysis (LDA, QDA), due to the hypothesis of Gaussianity of the density functions. For the IASI data this assumption is verified for most statistics. Actually CDA could be similarly generalized to eventually 2 and more thresholds.

We already mentioned the assumption of independence of statistics; for the general  $D$ -dimensional case this means factorization of the multivariate density functions into univariate functions. In addition, the multidimensional decision rule is chosen as the cartesian product of unidimensional rules based on thresholds.

We noticed that in case of multiple statistics a pixel is labeled as Clear if all corresponding tests are satisfied with the right directions (e.g., Eq. 6). This procedure is common to several operational cloud mask algorithms. However when the number of statistics and corresponding tests increases, the probability that a pixel is labeled as cloudy just due to the intrinsic randomness of the statistic, while the pixel was actually Clear, increases. As a consequence there is an increase of the Type I Error. To

## Cloud mask via cumulative discriminant analysis

U. Amato et al.

Title Page

Abstract

Introduction

Conclusions

References

Tables

Figures

⏪

⏩

◀

▶

Back

Close

Full Screen / Esc

Printer-friendly Version

Interactive Discussion



overcome this effect it is advisable to adopt a strategy of Bonferroni or False Discovery Rate (FDR, Benjamini et al., 1995) correction.

Finally, CDA also assumes a binary (2-class) problem; generalization to more classes could be accomplished, e.g., with the same techniques that apply to SVM (e.g., dealing with all couples of bands and then adopting a voting strategy).

## 4.6 Computational aspects

The code for the computation of the multidimensional CDA has been developed in Matlab. In particular, an optimized one-dimensional version has been implemented where a direct minimization of the Cost function is performed by means of a direct exhaustive search over the step-wise estimated cumulative density function.

Its generalization to the fully general  $D$ -dimensional case is not computationally possible due to the exponential increase of the computational cost with  $D$ . Therefore independence hypothesis had been resorted and minimization of the Cost function is accomplished with an optimization method that is derivative-free (Nelder-Mead, based on simplex).

We observe by simple arguments that increase of the number of involved statistics can not degrade the Cost function. In fact, by setting the threshold of the new statistic to  $+\infty$  or  $-\infty$  according to the sign of the inequality,  $\leq$  or  $\geq$ , respectively, makes the decision rule not depending on the new threshold. As a consequence the Cost function is the same and therefore it can only improve with the number of dimensions by properly choosing all thresholds.

We also observe that of course when the number of statistics increases, all the thresholds will vary as well in general, even though we should not expect dramatic variations. For this reason the initialization of the Nelder-Mead optimization is performed starting from  $D$  independent unidimensional optimizations of the thresholds (indeed also of the corresponding inequality signs). Then the threshold of the statistic that, singularly taken, yields the minimum value of the Cost function is used as initialization; for

## Cloud mask via cumulative discriminant analysis

U. Amato et al.

Title Page

Abstract

Introduction

Conclusions

References

Tables

Figures



Back

Close

Full Screen / Esc

Printer-friendly Version

Interactive Discussion



the remaining thresholds an intermediate value between the optimal one-dimensional case and  $\pm\infty$  (according to the estimated sign of the inequality) is set.

Finally we do not perform a full optimization of the sign of threshold inequalities in the  $D$ -dimensional case because the number of cases to be considered would explode exponentially with  $D$  as  $2^D$ . We assume that the sign of the threshold inequality is preserved when moving from one to more dimensions.

Starting from the formulation (Eq. 9) the algorithm for the computation of the Cost function for  $D$  statistics goes through the following steps:

- Compute unidimensional cumulative functions  $F_d^{\text{Clear}}, F_d^{\text{Cloudy}}, d = 1, \dots, D$ , for each statistic  $d$  starting from the training set
- Compute directional cumulative functions,  $\tilde{F}_d^{\text{Clear}}, \tilde{F}_d^{\text{Cloudy}}, d = 1, \dots, D$ , according to the selected direction of threshold inequalities:

$$\tilde{F}_d^{\text{Clear}}(x^d) = \begin{cases} F_d^{\text{Clear}}(x^d), & \text{if } x^d \leq \vartheta_d \Rightarrow \text{Clear} \\ 1 - F_d^{\text{Clear}}(x^d), & \text{if } x^d \geq \vartheta_d \Rightarrow \text{Clear} \end{cases}$$

$$d = 1, \dots, D.$$

$$\tilde{F}_d^{\text{Cloudy}}(x^d) = \begin{cases} F_d^{\text{Cloudy}}(x^d), & \text{if } x^d \leq \vartheta_d \Rightarrow \text{Clear} \\ 1 - F_d^{\text{Cloudy}}(x^d), & \text{if } x^d \geq \vartheta_d \Rightarrow \text{Clear}, \end{cases}$$

- Compute Type I and II errors as

$$E^{\text{I}} = 1 - \prod_{d=1}^D \tilde{F}_f^{\text{Clear}}(x^d)$$

$$E^{\text{II}} = \prod_{d=1}^D \tilde{F}_f^{\text{Cloudy}}(x^d)$$

## Cloud mask via cumulative discriminant analysis

U. Amato et al.

Title Page	
Abstract	Introduction
Conclusions	References
Tables	Figures
◀	▶
◀	▶
Back	Close
Full Screen / Esc	
Printer-friendly Version	
Interactive Discussion	



5

10

15

20



Use of Principal Components is easily accomplished by performing a transform of the original radiance data ( $x$ ) into Principal Components ( $c$ ) by PCA, then retaining the first  $\bar{D} \leq D$  Principal Components  $c^1, \dots, c^{\bar{D}}$ , and finally straightforwardly applying the algorithms described in the present section to them.

## 5 Results

This Section will show the main results of the numerical experiments worked out on the data bases IASI1 and IASI2 and with the additional validation data set qualified for clear sky with the SEVIRI cloud mask. Finally, a comparison with the IASI level 2 cloud mask implemented at EUMETSAT will be shown.

The present work is not intended to yield a fully operative cloud detection algorithm, which requires extensive and independent training and validation data sets covering all climatologies. Rather, we work out some sample test cases to prove feasibility of the full methodology.

We will present the results obtained by a transform to PC's. We will not show the optimal thresholds because they make sense only if endowed with the prior PC transform that, as well known, is unique only apart of rotations. All thresholds and transform matrices are anyway available upon request. For the present work, the statistics sh (see Table 3) was not considered, mostly because it added poor or no value to the remaining eight statistics.

Let us define the vector of statistics,  $\mathbf{x}$ , as

$$\mathbf{x} = \left( h_s, \chi_s^2, T_0, \Delta T_{CO_2}, W_1, W_2, W_3, W_4 \right)^T, \quad (10)$$

where the superscript T stands for transpose. For the case of tropical climate zone and sea surface, Fig. 5 (left) shows the covariance matrix of  $\mathbf{x}$  used to build up the PC basis along with the eigenvectors. The covariance matrix was obtained on the basis of the training data set (IASI2). It is seen that the matrix is not diagonal, which supports

## Cloud mask via cumulative discriminant analysis

U. Amato et al.

Title Page

Abstract

Introduction

Conclusions

References

Tables

Figures



Back

Close

Full Screen / Esc

Printer-friendly Version

Interactive Discussion



## Cloud mask via cumulative discriminant analysis

U. Amato et al.

Title Page

Abstract

Introduction

Conclusions

References

Tables

Figures



Back

Close

Full Screen / Esc

Printer-friendly Version

Interactive Discussion



the transformation of  $x$  to Principal Components. The plot of eigenvalues, shown in the same figure (right) indicates that the first component exhibits most variance of the statistics. It is also noteworthy that the vector  $x$  is made up of unlike quantities, some dimensionless and others in units of K. In principle we could standardized it before the transformation; however we did not get better results. We also have to keep in mind that in case the units are changed (e.g., Kelvin to centigrade degrees for temperature) than the thresholds have to be estimated again.

### 5.1 Validation–coinciding training and validation data sets

To summarize the performance of the cloud mask we introduce the Merit function

$$\mathcal{M} = 100(1 - C)$$

This is, indeed, the fraction of successful classifications  $S$  (see Eq. 1).

The CDA methodology has been applied to select the thresholds for the cloud detection scheme. In doing so, the statistics introduced in Sect. 3 have been transformed to the PC space as in Sect. 4.7. The thresholds have been estimated on the basis of IASI1 or IASI2 depending on which statistical population is the larger size for the climate zones at hand. For example we used IASI2 as a training data set and IASI1 as a validation one for the mid-latitude climatic zone in Winter (96 522 and 40 941 samples, respectively); we made the reverse choice in Summer (81 729 samples for IASI1 and 40 177 for IASI2). Both data sets are quite equivalent for the tropical climatology (186 787 samples for IASI1 and 230 267 for IASI2); they could have been jointly used as a training data set, however we kept their distinction to have different training and validation data sets.

Thresholds have been defined and calculated as a function of all the climatic zones of Table 1.

### 5.1.1 Sea surface

As a general result, we have found that CDA methodology normally selects the first Principal Component as the most effective to discriminate clear from cloudy sky. The performance for the relevant climatic zones is summarized in Table 4. The CDA methodology always selects the first Principal Component but High-Lat-Summer NH, High-Lat-Winter SH and two Sea ice climatologies, which seem to yield a better score with the first and second Principal Components. For sea ice, the basic CDA assumption of cumulative functions not crossing each other was systematically not met. Therefore we empirically fixed the thresholds by visual inspection of the cumulative functions. Occasionally, this situation also occurred for ice/snow over land (as will be shown in next section) and is also the result of inaccurate reference cloud masks over ice/snow regions.

### 5.1.2 Land surface

We recall that for land surface we distinguish between day and night for each climatic zone of Table 1. Tables 5 and 6 summarize the results of CDA for day and night, respectively. Missing values mean that there were not available enough data to perform the CDA analysis. We see that the performance of the cloud detection, according to the thresholds selected by CDA, is quite close to 90 %, with the Type II Error close to 10 % for most cases. As expected, the performance of the cloud detection tends to become less effective in case of snow/ice coverage. Nevertheless, also in this case performance is higher than 60 %. As for sea ice, also for snow/ice over land the results have been tuned in some cases to compensate for the driving effect of the cloudy population.

## 5.2 Validation—different training and validation data sets

Apart from the tropics, the data sets IASI1 and IASI2 cover different climatic zones. Therefore we present results only for the tropical climatic zone.

## Cloud mask via cumulative discriminant analysis

U. Amato et al.

Title Page

Abstract

Introduction

Conclusions

References

Tables

Figures



Back

Close

Full Screen / Esc

Printer-friendly Version

Interactive Discussion



For the case of sea surface we have used the data set IASI2 for the training and the data set IASI1 for validation. Performance of the methodology is shown in Table 7.

The same experiment has been carried out for land, where we distinguish between day and night (results also shown in Table 7). During daytime the data set used for training is IASI1, whereas validation is performed on the data set IASI2. The reverse occurs in nighttime because of the better coverage of the latter.

Comparing Table 7 with Table 4 it is seen that the performance obtained on the validation data set is largely consistent with the case of coinciding training and validation data sets. This occurs even when the two data sets are largely nonhomogeneous, showing the robustness of the methodology.

### 5.3 Validation with SEVIRI

Validation has been also performed with a cloud mask independent of CMS on which the training data set relies. For the Southern Europe and Africa continent three consecutive, daytime, IASI orbits belonging to the set IASI2 have been co-located with SEVIRI imagery.

The IASI footprints considered in this experiment are shown in Fig. 6. The first IASI footprint observation was acquired on 22 July 2007 at 06:42:12 UTC, while the last observations at 10:14:18 UTC. To have an idea of the cloud coverage and type over the target area, Fig. 6 shows a RGB composite SEVIRI full-disk image starting at 09:27:43 UTC of the same day.

Since the orbits are in daytime, the SEVIRI cloud mask also benefits from visible channels. In contrast only information in the infrared spectral region are used with IASI. The IASI footprints have been co-located with the SEVIRI operational cloud mask and flagged clear sky in case 100% SEVIRI pixels falling within the IASI field of view are themselves clear sky. In this way we have a SEVIRI-based cloud mask for IASI which can be compared with that obtained with our procedure.

## Cloud mask via cumulative discriminant analysis

U. Amato et al.

[Title Page](#)

[Abstract](#)

[Introduction](#)

[Conclusions](#)

[References](#)

[Tables](#)

[Figures](#)



[Back](#)

[Close](#)

[Full Screen / Esc](#)

[Printer-friendly Version](#)

[Interactive Discussion](#)



---

## Cloud mask via cumulative discriminant analysis

U. Amato et al.

---

Title Page

Abstract

Introduction

Conclusions

References

Tables

Figures

⏪

⏩

◀

▶

Back

Close

Full Screen / Esc

Printer-friendly Version

Interactive Discussion



The results for this exercise are shown in Table 8. It is seen that the agreement with the SEVIRI-based cloud mask is excellent, which further testifies the very good performance of the cloud detection over land in the case of the tropical latitudinal belt.

The SEVIRI cloud mask has been also used to classify a data set independent of IASI1 and IASI2. We have used a set of IASI data co-located with SEVIRI imagery from 25 September to 4 October 2010.

The data set was acquired over Europe/Africa in the framework of the IASI/SEVIRI Inter-calibration experiment, performed within the activities of EUMETSAT GSICS. We have a total of 4277 IASI spectra whose footprints are shown in Fig. 7.

As for the previous exercise, the IASI footprints have been co-located with the SEVIRI operational cloud mask and flagged clear sky in case 100 % SEVIRI pixels falling within the IASI field of view are themselves clear sky. The results are shown in the second row of Table 8.

Again we have a very good comparison with a total score of 86.7 %. In this case we have an excess of false alarms with respect to the Africa case study also shown in Table 8. However, we have checked that the problem is likely due to the IASI/SEVIRI co-location used in the GSICS system.

## 6 Conclusions

In perspective of developing a cloud mask for MTG-IRS we have designed and implemented a stand alone cloud detection tool expected to run only using MTG-IRS spectral radiances.

The strategy has been implemented and tested using IASI as a proxy for MTG-IRS. The scheme relies on a set of nine statistics. Six of them are heritage of the AVHRR-based cloud classification and discrimination, whereas three more have been specifically designed to take into account the hyperspectral characteristics of MTG-IRS.

Furthermore, a methodology (Cumulative Discriminant Analysis) has been developed in the context of classification of binary variables which using all the statistics

## Cloud mask via cumulative discriminant analysis

U. Amato et al.

Title Page

Abstract

Introduction

Conclusions

References

Tables

Figures



Back

Close

Full Screen / Esc

Printer-friendly Version

Interactive Discussion



and optimally computes thresholds needed to classify a given pixel as clear or cloudy. The system needs a training data set and a reference cloud mask, as well. Once the training has been performed the algorithms only needs the thresholds and therefore is very fast. The probabilistic approach allows us to quantify the Type I (miss) and Type II (false alarm) Errors. Within the methodology, the two types of error can have the role of design parameters, so that the cloud detection scheme can have a prescribed rate of false alarms. Alternatively, we can design and develop a cloud detection scheme in which both types of error are simultaneously minimized.

The methodology we have developed is much more than a classification tool. In fact, the nine statistics are intended to summarize the information content of the IASI spectra into a limited number of variables as an input to the classification methodology. Once a transform of the statistics into Principal Components has been performed, CDA discriminates how many and which principal components are needed for an effective cloud detection. Nevertheless, CDA is general enough to be coupled with different other statistics.

Apart from a few situations, we have found that the first Principal Component of the statistics is enough to build up a first efficient cloud screening. This largely simplifies the scene analysis algorithm because we need to compute thresholds just for one Principal Component; moreover it circumvents assumption of independence of statistics; finally the overall algorithm is very efficient.

Basically, the methodology relies on a training data set and a reference cloud mask. It is recommended that the training data set is made up of real observations, having a global coverage and a full coverage of the possible real clear and cloudy scenes.

The reference cloud mask is needed to extract from the training data set the two classes of spectra, one for clear sky and the second one for cloudy sky. At present, we rely on the AVHRR cloud mask, which is operationally produced by CMS.

The algorithm has been tested against independent data sets and cloud masks, as well. As a result, we have that the agreement is normally well above 80 %, apart from regions covered by sea ice or land snow/ice, where also the reference cloud masks

tend to have a relatively high rate of failure. We think that once applied to real MTG-IRS scenes the cloud detection should even improve performance because MTG-IRS is an imager and its pixel size is expected to be 4 times smaller than that of IASI.

The paper has been mostly dealing with the scientific basis of CDA and on an initial evaluation of the scheme. Currently we are developing appropriate seasonal and global IASI/AVHRR/SEVIRI data sets to be used to train and validate the methodology for operational applications.

*Acknowledgements.* Work performed under the EUMETSAT EUM/CO/12/4600001033/SAT contract.

## References

- Ackerman, S., Strabala, K., Menzel, W., Frey, R., Moeller, and C., Gumley, L.: Discriminating clear-sky from clouds with MODIS, *J. Geophys. Res.*, 103, 32–141, 1998. 5605
- Ackerman, S., Holz, R., Frey, R., Eloranta, E., and McGill, B. M.: Cloud detection with MODIS, Part II: Validation, *J. Atmos. Ocean. Tech.*, 25, 1073–1086, 2008. 5605
- Amato, U., Masiello, G., Serio, C., and Viggiano, M.: The  $\sigma$ -IASI code for the calculation of infrared atmospheric radiance and its derivatives, *Environ. Modell. Softw.*, 17, 651–667, doi:10.1016/S1364-8152(02)00027-0, 2002. 5615
- Amato, U., Antoniadis, A., Cuomo, V., Cutillo, L., Franzese, M. Murino, L., and Serio, C.: Statistical cloud detection from SEVIRI multispectral images, *Remote Sens. Environ.*, 112, 750–766, doi:10.1016/j.rse.2007.06.004, 2008. 5606
- Benjamini, Y. and Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing, *J. Roy. Stat. Soc. B Met.*, 57, 289–300, 1995. 5624
- Chevalier, F.: Sampled Database of 60 Levels Atmospheric Profiles from the ECMWF Analysis. Technical Report ECMWF EUMETSAT SAF programme Research, 4, ECMWF, Reading, 2001. 5614, 5615
- Derrien, M.: Validation Report for “Cloud Products” (CMa-PGE01 v3.2, CT-PGE02 v2.2 &CTTH-PGE03 v2.2), EUMETSAT, Darmstadt, 2012. 5611
- Derrien, M. and Le Gléau, H.: MSG/SEVIRI cloud mask and type from SAFNWC, *Int. J. Remote Sens.*, 26, 4707–4732, 2005. 5610, 5611

## Cloud mask via cumulative discriminant analysis

U. Amato et al.

Title Page

Abstract

Introduction

Conclusions

References

Tables

Figures



Back

Close

Full Screen / Esc

Printer-friendly Version

Interactive Discussion



## Cloud mask via cumulative discriminant analysis

U. Amato et al.

Title Page

Abstract

Introduction

Conclusions

References

Tables

Figures



Back

Close

Full Screen / Esc

Printer-friendly Version

Interactive Discussion



- EUMETSAT: CM-SAF Product CM-02, CM-08 and CM-14 Cloud Fraction, Cloud Type and Cloud Top Parameter Retrieval from SEVIRI, EUMETSAT, Darmstadt, 2009. 5605
- Fawcett, T.: ROC graphs: notes and practical considerations for researchers, ReCALL, 31, HPL-2003-4, 1–38, 2004. 5606
- 5 Heidinger, A. K.: CLAVR-x Cloud Mask Algorithm Theoretical Basis Document, NOAA/NESDIS/Office of Research and Applications, Washington, DC, 2004. 5604
- Heidinger, A. K., Evan, A. T., Foster, M. J., and Walther, A.: A naive Bayesian cloud-detection scheme derived from CALIPSO and applied within PATMOS-x, *J. Appl. Meteorol. Clim.*, 51, 1129–1144, 2012. 5610
- 10 Hilton, F., Armante, R., August, T., Barnet, C., Bouchard, A., Camy-Peyret, C., Capelle, V., Clarisse, L., Clerbaux, C., Coheur, P. F., Collard, A., Crevoisier, C., Dufour, G., Edwards, D., Fajjan, F., Fourrié, N., Gambacorta, A., Goldberg, M., Guidard, V., Hurtmans, D., Illingworth, S., Jacquinet-Husson, N., Kerzenmacher, T., Klaes, D., Lavanant, L., Masiello, G., Matricardi, M., McNally, A., Newman, S., Pavelin, E., Payan, S., Péquignot, E., Peyridieu, S., Phulpin, T., Remedios, J., Schlüssel, P., Serio, C., Strow, L., Stubenrauch, C., Taylor, J., Tobin, D., Wolf, W., and Zhou, D.: Hyperspectral Earth Observation from IASI: four years of accomplishments, *B. Am. Meteorol. Soc.*, 93, 347–370, doi:10.1175/BAMS-D-11-00027.1, 2012. 5604
- 15 Inoue, T.: On the temperature and effective emissivity determination of semi-transparent cirrus clouds by bi-spectral measurements in the 10 mm window region, *J. Meteorol. Soc. Jpn.*, 63, 88–99, 1985. 5612
- Inoue, T. and Ackerman, S. A.: Radiative effects of various cloud types as classified by the split window technique over the eastern sub-tropical Pacific derived from co-located ERBE and AVHRR data, *J. Meteorol. Soc. Jpn.*, 80, 1383–1394, 2002. 5612
- 25 Lavanant, L., Marguinaud, P., Harang, L., Lelay, J., Péré, S., and Philippe, S.: Operational cloud masking for the OSI SAF global METOP/AVHRR SST product, Joint 2007 EUMETSAT Meteorological Satellite Conference and the 15th Satellite Meteorology & Oceanography Conference of the American Meteorological Society, EUMETSATP. 50, available at: <http://www.eumetsat.int> (last access: 5 June 2014), 2007. 5605, 5609, 5610, 5612
- 30 Lavanant, L., Fourrie, N., Gambacorta, A., Grieco, G., Heilliette, S., Hilton, F. I., Kim, M.-J., McNally, A. P., Nishihata, H., Pavelin, E. G., and Rabier, F.: Comparison of cloud products within IASI footprints for the assimilation of cloudy radiances, *Q. J. Roy. Meteorol. Soc.*, 137, 1988–2003, 2011. 5607, 5609, 5610







## Cloud mask via cumulative discriminant analysis

U. Amato et al.

Title Page

Abstract

Introduction

Conclusions

References

Tables

Figures

◀

▶

◀

▶

Back

Close

Full Screen / Esc

Printer-friendly Version

Interactive Discussion



**Table 3.** Statistics derived from the IASI radiances. Temperatures are defined in Table 2.

Statistic	Method
$W_1$	Based on the difference $\bar{T}_{900} - \bar{T}_{832}$
$W_2$	Based on the difference $\bar{T}_{900} - \bar{T}_{1168}$
$W_3$	Based on the difference $\bar{T}_{832} - \bar{T}_{2003}$
$W_4$	Based on the difference $\bar{T}_{832} - \bar{T}_{2700}$
$T_0$	Based on the BT $\bar{T}_{832}$
$\Delta T_{\text{CO}_2}$	Split window test based on the $\text{CO}_2$ Q-branch at $791 \text{ cm}^{-1}$ , $\Delta T_{\text{CO}_2} = T_{790.5} - T_{791.75}$
$\chi_s^2$	Based on a $\chi^2$ -like variable defined on a couple of skin temperature values, $(\hat{T}_s, T_s^R)$ , with $\hat{T}_s$ estimated from the spectrum and $T_s^R$ a suitable reference.
$h_s$	Based on shape similarity between a couple of spectra (Observation, Reference)
sh	Spatial homogeneity statistic based on the standard deviation of $T_0$ corresponding to a cluster of $n \times n$ nearby pixels. For IASI we consider the cluster of $2 \times 2$ pixels within a given field of view.



## Cloud mask via cumulative discriminant analysis

U. Amato et al.

Title Page

Abstract

Introduction

Conclusions

References

Tables

Figures



Back

Close

Full Screen / Esc

Printer-friendly Version

Interactive Discussion



**Table 5.** CDA results based on coinciding training and validation data sets for daytime soundings over land. The percentage of *hit or success* (TP) and of *correct nonevent* (TN) can be obtained by  $100 - \text{FP}$  and  $100 - \text{FN}$ , respectively.

Climate zone	PC	FN (miss)	FP (false alarm)	Merit function
Tropical	1	8.4	8.4	91.6
Mid-Lat-Summer NH	1	17.6	17.6	82.4
Mid-Lat-Summer SH	1	6.9	6.9	93.1
Mid-Lat-Winter NH	1	11.1	11.1	89.9
Mid-Lat-Winter SH	1	10.4	10.4	89.6
High-Lat-Summer NH	1	11.8	11.8	88.2
High-Lat-Winter NH	1	10.3	10.3	89.7
Land below 1 km with snow/ice NH	1	17.6	23	77.3
Land below 1 km with snow/ice SH	1	16.3	14.1	85.3
Land above 1 km with snow/ice NH	1	7.5	7.5	92.5
Land above 1 km with snow/ice SH	1	16.2	12.4	86.4
Antarctica below 1 km with snow/ice	1	7.4	7.4	82.6
Antarctica above 1 km with snow/ice	1	34.9	34.9	65.1

## Cloud mask via cumulative discriminant analysis

U. Amato et al.

**Table 6.** CDA results based on coinciding training and validation data sets for nighttime soundings over land. The percentage of *hit or success* (TP) and of *correct nonevent* (TN) can be obtained by  $100 - \text{FP}$  and  $100 - \text{FN}$ , respectively.

Climate zone	PC	FN (miss)	FP (false alarm)	Merit function
Tropical	1	17.8	17.8	82.2
Mid-Lat-Summer NH	1	12.2	12.2	87.8
Mid-Lat-Winter NH	1	10.1	10.1	89.9
Mid-Lat-Winter SH	1	12	12	88.0
High-Lat-Summer NH	1	8.5	8.5	91.5
High-Lat-Winter NH	1	13.7	13.7	86.3
Land below 1 km with snow/ice NH	1	35.7	10	81.2
Land below 1 km with snow/ice SH	1	19.5	18.5	81.3
Land above 1 km with snow/ice NH	2	12	53	69.0
Land above 1 km with snow/ice SH	1	10	45	64.4
Antarctica below 1 km with snow/ice	2	35.2	35.2	64.8
Antarctica above 1 km with snow/ice	2	26.3	26.3	73.7

Title Page

Abstract

Introduction

Conclusions

References

Tables

Figures



Back

Close

Full Screen / Esc

Printer-friendly Version

Interactive Discussion



## Cloud mask via cumulative discriminant analysis

U. Amato et al.

**Table 7.** Validation results using IASI1 and IASI2 data sets. The total score ( $S$ ) or equivalently Merit function, false alarms and miss are shown in percentage.

Area	Training set size	FN (miss)	FP (false alarm)	Merit function
Tropics, sea surface	186 787	15.2	16.7	83.3
Tropics, land daytime	172 216	15.5	6.7	93.3
Tropics, land nighttime	85 882	5.4	20.0	80.0

[Title Page](#)
[Abstract](#)
[Introduction](#)
[Conclusions](#)
[References](#)
[Tables](#)
[Figures](#)

[Back](#)
[Close](#)
[Full Screen / Esc](#)
[Printer-friendly Version](#)
[Interactive Discussion](#)


## Cloud mask via cumulative discriminant analysis

U. Amato et al.

Title Page

Abstract

Introduction

Conclusions

References

Tables

Figures



Back

Close

Full Screen / Esc

Printer-friendly Version

Interactive Discussion



**Table 8.** Validation results based on the co-location IASI/SEVIRI. The first row refers to the data set IASI2 data set. The second row to the GSICS data set.

Area	Training set size	FN (miss)	FP (false alarm)	Merit function
IASI2 set (Europe/Africa)	47 160	7.9	5.7	96.7
GSICS set (Europe/Africa)	4277	11.3	16.7	86.7

## Cloud mask via cumulative discriminant analysis

U. Amato et al.

Title Page

Abstract

Introduction

Conclusions

References

Tables

Figures



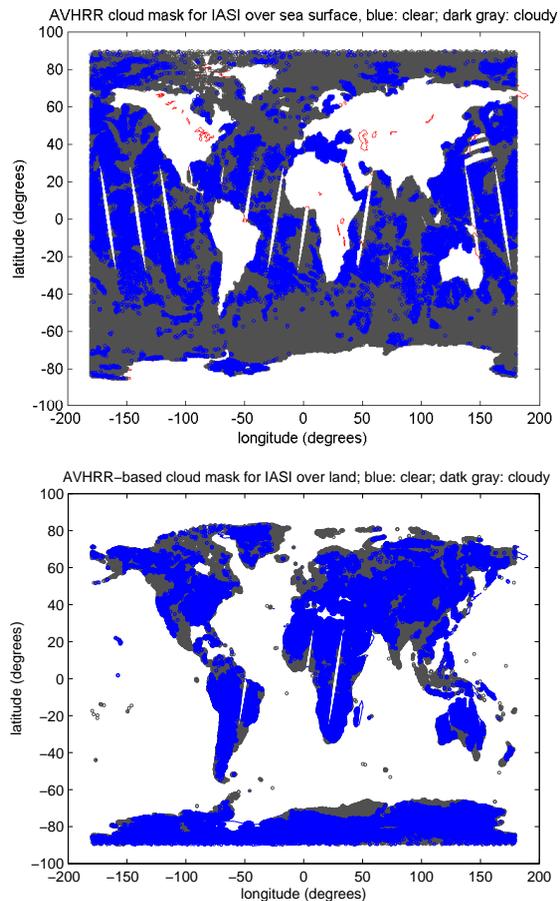
Back

Close

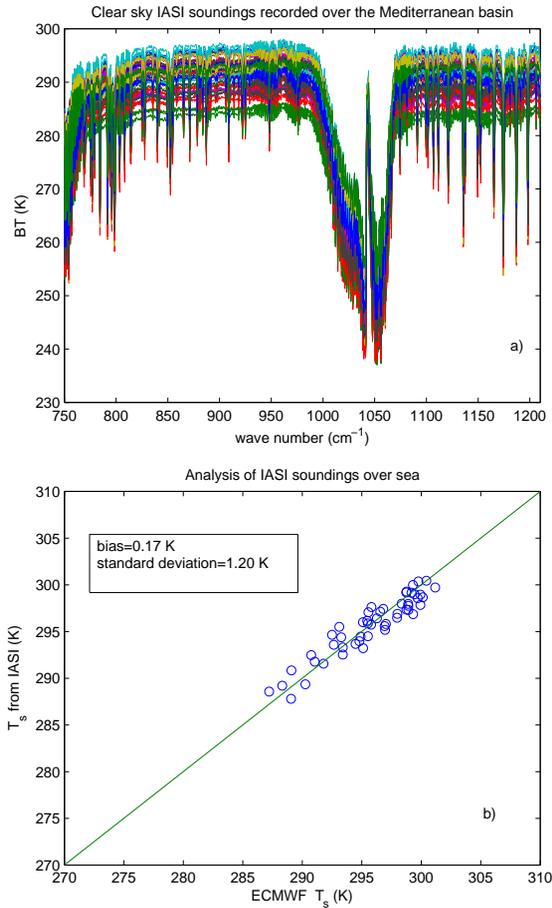
Full Screen / Esc

Printer-friendly Version

Interactive Discussion



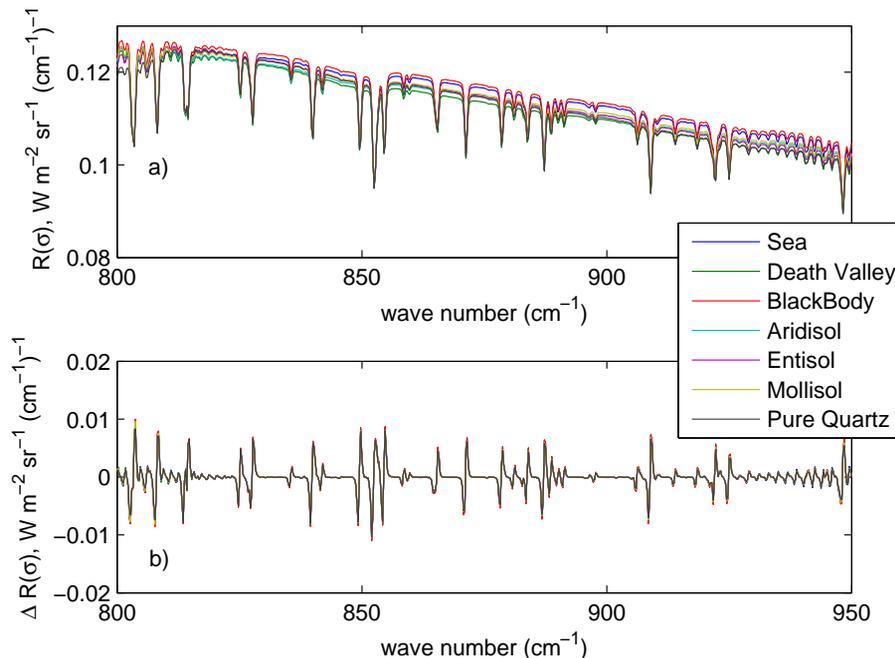
**Figure 1.** IASI cloud mask derived from AVHRR for the set IASI2; top panel: sea surface; bottom panel: land surface. Blue color refers to clear sky and dark gray to cloudy one.



**Figure 2.** Example of the performance of the split window technique for the estimation of the skin temperature. **(b)** compares the IASI  $T_s$  obtained from the spectra shown in **(a)** with the time-space co-located ECMWF analysis.

Cloud mask via  
cumulative  
discriminant analysis

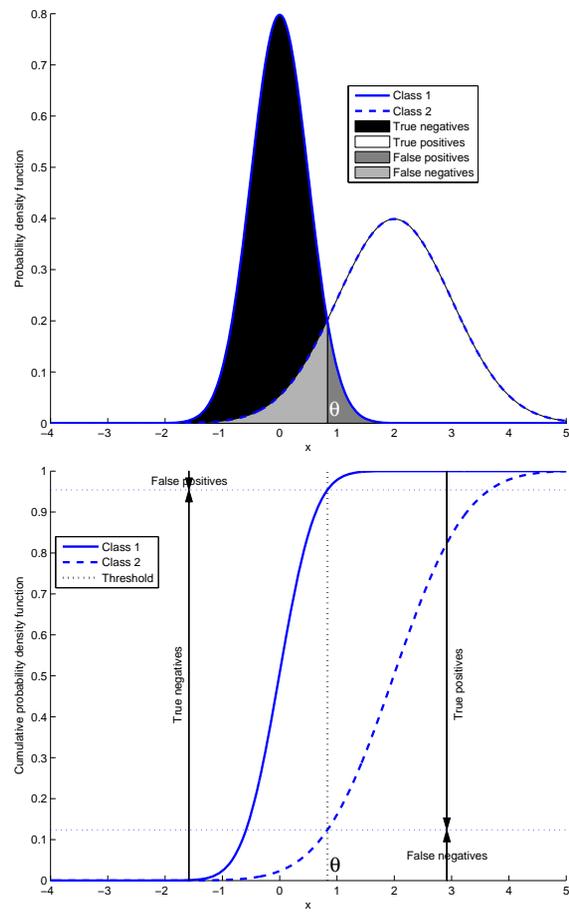
U. Amato et al.



**Figure 3.** Top panel: radiance spectra in the atmospheric window segment  $[800\text{--}950] \text{ cm}^{-1}$  for very different surface emissivities. Bottom panel: the same radiance spectra after first order differentiation. The synthetic spectra have been obtained with the same state vector, but emissivity.

## Cloud mask via cumulative discriminant analysis

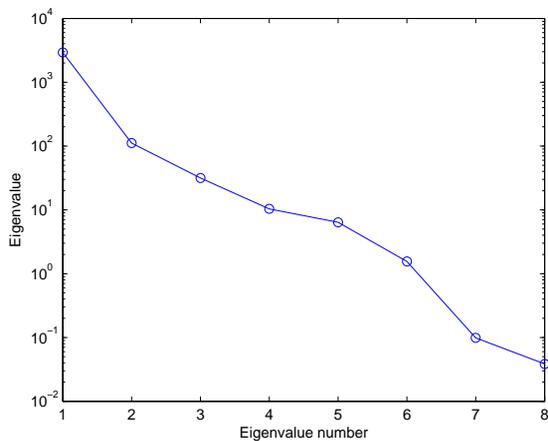
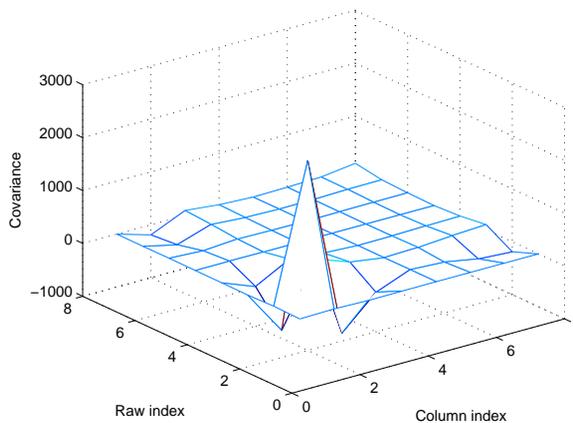
U. Amato et al.



**Figure 4.** Visual representation of Type I and II errors by density (top panel) and cumulative (bottom panel) probability functions.

Title Page	
Abstract	Introduction
Conclusions	References
Tables	Figures
◀	▶
◀	▶
Back	Close
Full Screen / Esc	
Printer-friendly Version	
Interactive Discussion	





**Figure 5.** Example of covariance matrix of the vector  $x$  for the case of the tropical climate zone and sea surface; top panel, covariance matrix; bottom panel, eigenvalues.

**Cloud mask via cumulative discriminant analysis**

U. Amato et al.

Title Page

Abstract Introduction

Conclusions References

Tables Figures

◀ ▶

◀ ▶

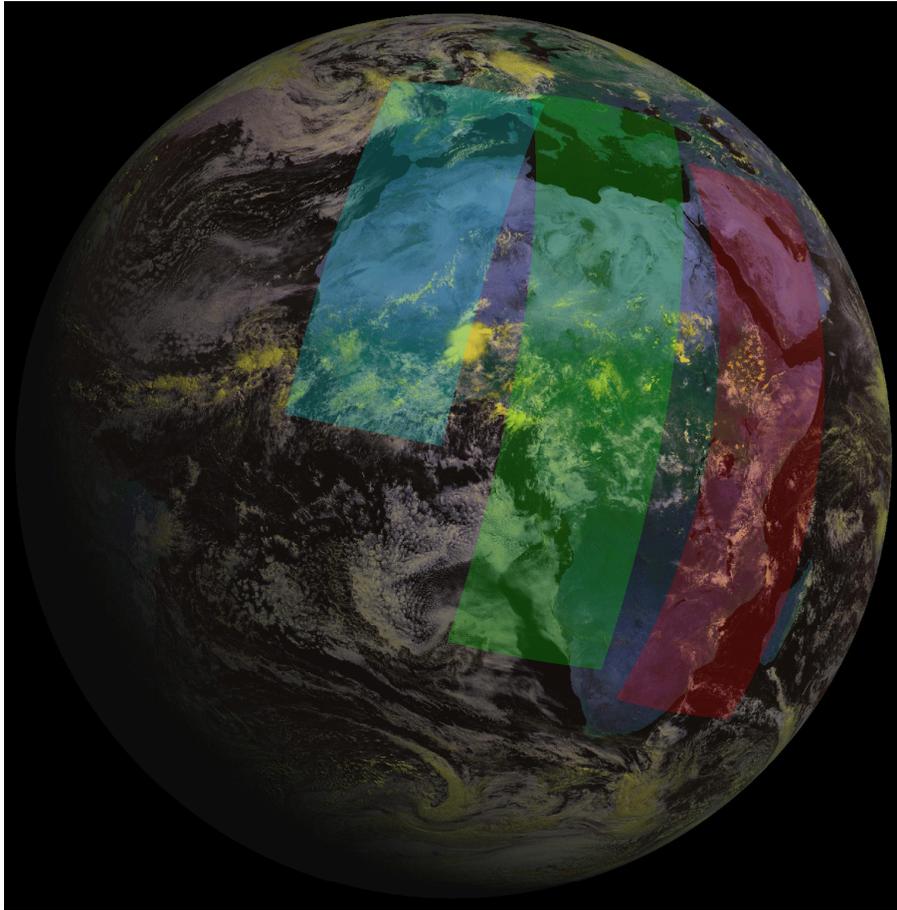
Back Close

Full Screen / Esc

Printer-friendly Version

Interactive Discussion





**Figure 6.** SEVIRI full disk imagery starting at 09:27:43 UTC for the day 22 July 2007. The target area used to check the cloud detection scheme is that encircled within red lines.

# AMTD

7, 5601–5650, 2014

## Cloud mask via cumulative discriminant analysis

U. Amato et al.

Title Page

Abstract

Introduction

Conclusions

References

Tables

Figures



Back

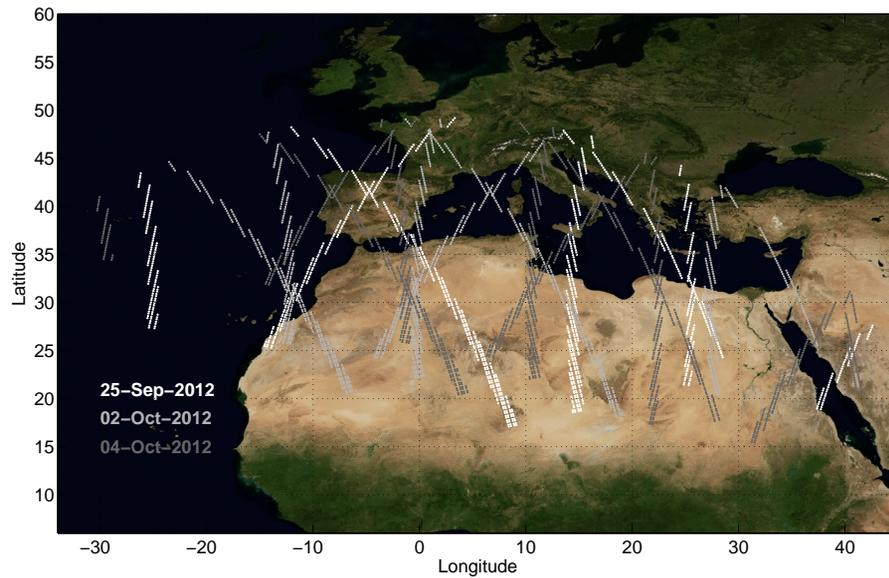
Close

Full Screen / Esc

Printer-friendly Version

Interactive Discussion





**Figure 7.** Footprint positions of IASI observations for the GSICS IASI/SEVIRI Inter-calibration experiment.

## Cloud mask via cumulative discriminant analysis

U. Amato et al.

Title Page	
Abstract	Introduction
Conclusions	References
Tables	Figures
◀	▶
◀	▶
Back	Close
Full Screen / Esc	
Printer-friendly Version	
Interactive Discussion	

