IOPscience iopscience.iop.org

Home Search Collections Journals About Contact us My IOPscience

SEVIRI Cloud mask by Cumulative Discriminant Analysis

This content has been downloaded from IOPscience. Please scroll down to see the full text.

2015 J. Phys.: Conf. Ser. 633 012056

(http://iopscience.iop.org/1742-6596/633/1/012056)

View the table of contents for this issue, or go to the journal homepage for more

Download details:

IP Address: 193.204.22.243
This content was downloaded on 30/09/2015 at 09:49

Please note that terms and conditions apply.

4th International Conference on Mathematical Modeling in Physical Sciences (IC-MSquare2015) IOP Publishing
Journal of Physics: Conference Series **633** (2015) 012056 doi:10.1088/1742-6596/633/1/012056

SEVIRI Cloud mask by Cumulative Discriminant Analysis

M G Blasi[1], C Serio[1,2], G Masiello[1,2], S Venafra[1] and G Liuzzi[1]

[1] School of Engineering, Università della Basilicata, Via Ateneo Lucano, 85100 Potenza, Italy
[2] CNISM, Unità di Ricerca Università della Basilicata, 85100 Potenza, Italy

E-mail: maria.blasi@unibas.it

Abstract. In the context of *cloud detection* for satellite observations we want to use the method of Cumulative Discriminant Analysis (CDA) as a tool to distinguish between clear and cloudy sky applied to Spinning Enhanced Visible and Infrared Imager (SEVIRI) data. The methodology is based on the choice of several statistics related to the cloud properties, whose correlation has been analyzed by Principal Component Analysis (PCA). Results have been compared with the SEVIRI reference cloud mask provided by the European Centre for the Exploitation of Meteorological Satellite (EUMETSAT), in order to find suitable thresholds able to discriminate between clear or cloudy conditions. We trained the statistics on a selected region, the Basilicata area, located in the south of Italy, in different periods of the year 2012, in order to take into account the seasonal variability. Moreover we separated land and sea surface and distinguished between day-time or night-time. The validation of thresholds, obtained through SEVIRI observations analysis, shows a good agreement with the reference cloud mask.

1. Introduction

At present EUMETSAT geostationary meteorological programme is preparing for Meteosat Third Generation (MTG) to give continuity to the scientific community with infrared and visible/ultraviolet observations through four imaging and two sounding satellites in a twin configuration. The imaging satellites, MTG-I, comprise the Flexible Combined Imager (FCI) that will continue the very successful operation of SEVIRI on Meteosat Second Generation (MSG) satellites. The new imager will be provided of additional channels with higher spatial, temporal and radiometric resolution compared to the current MSG instruments.
Nowadays SEVIRI is the operational imager on board of the Meteosat Second Generation (MSG) geostationary satellites, then we want to exploit its high temporal resolution (one acquisition every 15 min) in order to study and construct a very accurate cloud mask scheme.
Infrared data are affected by clouds, so we aim to develop a scene analysis methodology able to discriminate between clear and cloudy areas since this is a crucial operation for many remote sensing applications over continental or oceanic surfaces. There are different methodologies employed in the cloud detection context [2, 9, 10, 13] most of which make use of statistics sensitive to spectroscopic properties of clouds. At present it is preferred to develop new methodologies that consider multiwavelength statistics [1, 6], namely physical quantities that involve two or more wavelengths that are sensitive to the surface parameters and to the cloud phase.
In this context we want to evaluate the potentiality and the efficiency of Cumulative Discriminant Analysis (CDA) as a tool to develop a cloud mask, being able to classify clear and cloudy pixels

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

in several climatic conditions (different surface type and illumination condition), and also to analyze the functionality of the selected statistics used in this method. We will not focus on cloud phase classification. The structure of the paper is organized as follows.
In Section 2 we describe the characteristics of data sets used to train and validate CDA method, and the reference cloud mask. The statistics used to discriminate between clear or cloudy areas are defined in Section 3, while in Section 4 we present the basic theory of the methodology: PCA to analyze the statistics' inter-correlation and CDA as classification procedure for the cloud detection. Then we show the results in Section 5 and Conclusions are given in Section 6.

2. Instrument and data sets

The cloud detection scheme elaborated in this work was performed on SEVIRI data. SEVIRI is the imager on board of MSG geostationary satellites ([11]), characterized by very high temporal and spatial resolution, that are respectively one acquisition every 15 min and 3 km at nadir. It is provided of 12 spectral channels: 3 in the visible band (one of them has higher spatial resolution of 1 km), 1 in NIR band and 8 in the thermal IR band (from 3.9 μm to 13.4 μm).
The analysis was performed on four data sets: two were used to train the cloud detection method, while the other two data sets have been used to validate the cloud detection scheme. The region of interest was the same for all data sets: it is a rectangular area with latitude and longitude respectively ranging from 39.25° to 42° and from 13° to 18.75°, including the entire Basilicata region, located in the south of Italy.
Data sets for the training phase are SEVJAN_1 and SEVJUL_1, comprising SEVIRI data with the related SEVIRI cloud masks respectively from 1^{st} to 10^{th} of January 2012 and from 1^{st} to 10^{th} of July 2012, each one including 240 h observation and 960 acquisitions.
Data sets considered for validation are SEVJAN_2 and SEVJUL_2, that comprise SEVIRI data with the related SEVIRI cloud masks respectively from 11^{st} to 20^{th} of January 2012 and from 11^{st} to 20^{th} of July 2012, each one including 240 h observation and 960 acquisitions.
The two validation data sets are temporally located just ten days after the training ones, because validation should be done on periods with similar climatic conditions. Each SEVIRI acquisition on the area of interest corresponds to 9643 pixels, for a total of 9257280 radiances in each dataset. We took into account two main aspects to choose these data sets. On one side we want to analyze a statistically significant sample with two well and similarly populated classes. On the other side the statistics we choose to perform CDA are season dependent, and we aim to develop a method which is fairly accurate and sensitive with respect to seasonal variability.

2.1. SEVIRI reference cloud mask

The SEVIRI reference cloud mask we used has been developed at EUMETSAT by the Satellite Application Facility in Support to Nowcasting and Very Short Range Forecasting (NWC-SAF), and it describes the scene type identifying all cloud-free pixels. Each pixel is classified in one of the following four types: clear sky over water, clear sky over land, cloudy, or not processed. The algorithm is based on a multispectral threshold technique, consisting in a series of threshold tests applied to various channels combination for each pixel. For more details see e. g. [4, 5].

3. Statistics

In this section we provide a description of the statistics tested to discriminate between clear and cloudy pixels in the CDA, and to build up the estimated cloud mask. Each statistics refers to the Brightness Temperature (BT) spectrum, computed from the SEVIRI radiance inverting the Planck function. We considered ten statistics, computed from MSG Level 1.5 Image data: the BT from all 8 IR SEVIRI channels (from 3.9 μm to 13.4 μm), and two particular differences between these values. They are listed and defined in table 1.

Table 1. List of the statistics and their spectral characteristics.

Statistic	Spectral Range (μm)	Nominal Centre Wavelength (μm)	Method
T_1	3.48 - 4.36	3.92	Brightness Temperature
T_2	5.35 - 7.15	6.25	Brightness Temperature
T_3	6.85 - 7.85	7.35	Brightness Temperature
T_4	8.30 - 9.10	8.70	Brightness Temperature
T_5	9.38 - 9.94	9.66	Brightness Temperature
T_6	9.80 - 11.80	10.80	Brightness Temperature
T_7	11.00 - 13.00	12.00	Brightness Temperature
T_8	12.40 - 14.40	13.40	Brightness Temperature
W_1	-	-	Difference T_7-T_1
W_2	-	-	Difference T_7-T_6

Infrared SEVIRI channels provide information on clouds and other important geophysical parameters, such as surface temperature, water vapor and ozone, exploiting the thermal contrast between surface and clouds, that are usually colder than terrestrial surface. Moreover, a proper combination of spectral channels provides information on atmospheric instability.
SEVIRI IR channels at 3.9, 8.7, 10.8 and 12.0 μm are window channels, where the absorption is almost negligible, while IR channels at 6.2 and 7.3 μm are centred on the water vapor absorption bands, then they are used for monitoring the water vapor content in the troposphere, but also to observe winds. The IR channel at 9.7 μm is centered on ozone absorption band, so that the evolution of total ozone profile can be monitored. Finally the IR channel at 13.4 μm is centered on CO_2 absorption channel, but it also contributes to estimate static instability because it is able to give temperature information from the lower troposphere. The only remark is about the 3.9 μm IR channel, where the observed radiance is the sum of terrestrial and solar contribution so that for our purposes it can be exploited only at night-time.
The statistic W_1 exploits the fact that the cirrus and stratus cloud types have an higher reflectance than the most of surface features at 3.9 μm, and it has been widely used in the AVHRR cloud detection. The last statistic we considered consists in the split window difference $T_{12\mu m} - T_{10.8\mu m}$, that is similar to Inoue's slope test [7, 8], used to detect cirrus clouds.

4. Methodology

In this section we present the basic theory for both PCA and CDA, paying particular attention to define the categorical variables from the statistic point of view, in order to understand how the Cumulative Discriminant theory is applied to.

4.1. PCA decomposition

In order to obtain a single cloud mask applying CDA on one parameter we want to reduce the dimensionality of the problem, exploiting information coming from each single statistic and combining them into a single one. For this purpose we used PCA and the related Singular Value Decomposition (SVD), that are appropriate tools in the field of data reduction [3, 12].
PCA is a statistical procedure that estimates the inter-correlation between variables, projecting the initial variables in an orthogonal space and defining their Principal Components (PC).
It is necessary to build up the covariance matrix of the process, that PCA decomposes through the most suitable orthogonal basis (in order to maximize the variance of the first component), yielding orthogonal functions and their related eigenvalues.
Let's call $\mathbf{s}$ a vector of $n_s \times 1$ size, where n_s is the number of the statistics, $\mathbf{s} = (s_1, ..., s_{n_s})^T$. The representative matrix of the process, $\mathbf{S}$, is an $n_s \times N$ matrix, where N is the number of samples for a suitable data set. The covariance matrix of $\mathbf{S}$, namely $\mathbf{C} = \frac{1}{N}\,\mathrm{S}\,\mathrm{S}^T$, can be SVD

decomposed, obtaining $\mathbf{C} = \mathbf{U}\Lambda\mathbf{V}^T$ where $\mathbf{U}$ and $\mathbf{V}$ are the two $n_s \times n_s$ orthogonal basis, so that $\mathbf{U}\mathbf{U}^T = \mathbf{I}$ and $\mathbf{V}\mathbf{V}^T = \mathbf{I}$, with $\mathbf{I}$ the identity matrix. It is worth to note that have to be $\mathbf{U}$=$\mathbf{V}$, because the covariance matrix is symmetric and definite positive; then Λ is a diagonal matrix whose elements are the singular values, namely the eigenvalues, positive definite of the covariance matrix: $\mathbf{\Lambda} = diag(\lambda_1^2, \ldots, \lambda_{n_s}^2)$. It shall be $\lambda_1^2 > \cdots > \lambda_{n_s}^2$, hence the PCs, are defined as $\mathbf{c} = \mathbf{U^T}\,\mathbf{s}$. The columns of matrix $\mathbf{U}$ define an orthogonal basis, hence an orthogonal space, characterized by the important property of orthogonality of PCs, that are the projection of the initial variables into this new orthogonal space. This orthogonal transformation is defined so that the first PC describes most of the variance of the initial data set and each eigenvalue is a measure of the PCs' inter-correlation. The first eigenvalue in our data sets is always several magnitude orders higher than following ones, that is the first PC has a far higher variance than that of the other PCs. Then we decided to apply CDA only on the first PC in the scene analysis.

4.2. Classification of categorical binary variable

We need to clarify the classification of a binary event: the *clear sky event* on one side and the *cloudy sky event* or *not-clear sky event* on the other side. From the probabilistic point of view the cloud detection process can be described by a 2×2 contingency table, which means to recognize four different categories comparing the prediction with the truth:

- *clear*: n. of occurrences a *clear sky* prediction agrees with the true *clear sky event*
- *I-type error*: n. of occurrences a *cloudy sky* prediction corresponds to a true *clear sky event*
- *II-type error*: n. of occurrences a *clear sky* prediction corresponds to a true *cloudy sky event*
- *cloudy*: n. of occurrences a *cloudy sky* prediction agrees with a true *cloudy sky event*

The first and last categories are also denoted as *True Positives* (TP) and *True Negatives* (TN). The efficiency of CDA in classifying the pixels of a certain scene can be evaluated computing the *score* $SC = (TP + TN)/(N_{clear} + N_{cloudy})$, where N_{clear} and N_{cloudy} are respectively the number of clear and cloudy sky events according to the reference cloud mask.

Many classification methods resort to estimate the probability density function, but in real applications they do not often fit very well real distributions. Anyway it is possible to recur to an estimator that is also convenient from the computational point of view: the Empirical Cumulative Density Function (ECDF), $F(x < \vartheta)$, that is the probability for the variable x to be below the threshold value, specified here with ϑ. An important result in asymptotic theory is that the estimator almost surely converges to the true cumulative function asymptotically [14]. The formulation related to the cloud detection problem is based on the individuation of a suitable variable, able to describe the cloudiness of the scene discriminating between clear or cloudy pixels, defining two populations. The best situation would be when the two populations do not overlap, but in practice the continuity of the variable involves a misclassifications for some pixels because there is no discontinuity in the passage from clear to cloudy scenes.

4.3. Cumulative Discriminant Analysis

In this section we describe CDA methodology in the univariate case, corresponding to a single statistic analysis. The basic assumption is that the predictor variable is of Discriminant Analysis type, where the estimator $\Gamma(\vartheta, \mathbf{x})$ is based on a threshold ϑ as follows:

$$\Gamma(x, \mathbf{x}) = \begin{cases} 1\ (Clear) & if\ x \geq \vartheta \\ 2\ (Cloudy) & if\ x < \vartheta \end{cases} \tag{1}$$

where $\mathbf{x} = x_1, \ldots, x_N$ is the statistic sample of size N defined by the data set, and each component x_i represents the predictor variable to be classified according to the threshold ϑ.

By CDA we are able to find the threshold ϑ, but we have to define a *cost function*, in order to

minimize the error due to I-Type and II-Type misclassifications. Let E^I and E^{II} be respectively the I-type error and II-type error. According to the decision rule in Eq. 1 they are determined by: $E^I = F^{Clear}(\vartheta)$ and $E^{II} = 1 - F^{Cloudy}(\vartheta)$. Two possibilities of *cost function* can be:

$$A) \qquad C(\mathbf{x}, \vartheta) = E^I + E^{II} = F^{Clear}(\vartheta) + 1 - F^{Cloudy}(\vartheta) \tag{2}$$

$$B) \qquad C(\mathbf{x}, \vartheta) = \max(E^I, E^{II}) = \max(F^{Clear}(\vartheta), 1 - F^{Cloudy}(\vartheta)) \tag{3}$$

To find the best estimate of the threshold ϑ we have to minimize the *cost function*, which means to minimize both the I-type error and II-type error, and to obtain a balance between these two categories. If the training dataset is not balanced between Clear and Cloudy categories, the threshold overweights the most populated class, that will be better described from the estimated cloud mask. This is a common problem during the cloud detection process, because of the seasonal and meteorological variability of the scene, that could involve a quite significant difference in the relative fraction of Clear and Cloudy pixels.
The classification rule given in (1) is used to classify clear and cloudy pixels but CDA methodology does not assume the hypothesis of Gaussian probability density function, that does not occur in real conditions, conversely it computes the optimal estimate of the threshold ϑ making use of ECDF. We analyzed results for both *cost functions*, with the only difference that each error type is weighted with respect to the populousness of their own class, in order to take into account the different size of the two classes. Hereafter we indicate with A or B the results obtained considering respectively equations (2) and (3) in CDA.
CDA methodology can be generalized to the multidimensional case, that exploits simultaneously more than one statistic under the assumption they have independent density functions. In this way the multidimensional probability density function can be factorized as the product of each single probability density function.

5. Results

Here we show the main results keeping in mind this is a preliminary work to inspect and probe the potentiality of CDA method, applied to PCs of the considered statistics. We work out some sample test cases to obtain the optimal thresholds that we used for validation. The methodology has been applied for a small area, centred on the Basilicata region, located in the south of Italy, for two different periods, corresponding respectively to Mid-Latitude Winter and Mid-Latitude Summer climatic zones in the North hemisphere [12]. For each dataset we distinguished several cases according to the type of surface (land or sea) and illumination conditions (day-time or night-time). We selected the first PC for the cloud detection by CDA considering that the first eigenvalue is usually three or four orders of magnitude greater than the others from the SVD, which means that it alone is able to satisfactorily describe the scene cloudiness.

5.1. CDA for the training datasets

We computed the threshold ϑ applying CDA to SEVJAN_1 and SEVJUL_1 data sets for all the mentioned cases. In Table 2 we show the percentage of I-type error, II-type error and the performance of the cloud detection process computed according A or B method. We can see that the performance is usually higher for sea surface because of minor changes in temperature with respect to land surface, while in figure 1 there are two examples of ECDF.

5.2. CDA for the validation data sets

We used the thresholds obtained during the training to validate CDA methodology on SEVJAN_2 and SEVJUL_2 data sets. We evaluated the performance of the method comparing the estimated cloud mask with the SEVIRI reference cloud mask. Results are showed in Table 3.
In figure 2 there are two comparisons between the estimated cloudy scene and the reference cloud

Table 2. Training results applying CDA to the first PC of SEVJAN_1 and SEVJUL_1, for all cases.

Surface Type	Illumination condition	I-type error (%)		II-type error (%)		Total score (%)		Total of Soundings
		A	B	A	B	A	B	
SEVJAN_1								
Land	Day-time	9.14	9.33	9.65	9.48	81.21	81.20	1298142
	Night-time	5.24	13.14	19.40	13.25	75.36	73.61	2514978
Sea	Day-time	5.17	8.90	11.36	9.09	83.47	82.01	1866948
	Night-time	11.10	10.53	9.97	10.58	78.93	78.89	3577212
SEVJUL_1								
Land	Day-time	0.79	6.99	10.48	7.04	88.72	85.97	2202531
	Night-time	0.92	6.45	8.46	6.41	90.62	87.14	1610589
Sea	Day-time	0.45	8.04	10.65	8.44	88.90	83.52	3134058
	Night-time	0.44	3.11	4.01	3.27	95.55	93.62	2310102

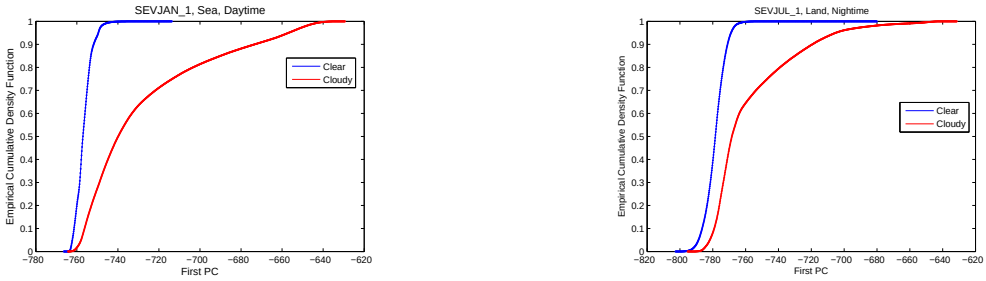

Figure 1. Two examples of ECDF: (left) ECDF applying CDA to sea surface in day-time for SEVJAN_1 dataset; (right) ECDF applying CDA to land surface in night-time for SEVJUL_1 dataset.

mask in two different cases. Comparing Table 2 and Table 3 we see that CDA methodology shows quite consistent scores both for winter and for summer data sets, meaning that the computed thresholds are suitable to differentiate between clear or cloudy areas.

Table 3. Validation results applied to the first PC, for SEVJAN_2 and SEVJUL_2 and for all cases.

Surface Type	Illumination condition	I-type error (%)		II-type error (%)		Total score (%)		Total of Soundings
		A	B	A	B	A	B	
SEVJAN_2								
Land	Day-time	10.20	10.02	11.32	11.53	78.48	78.45	1335666
	Night-time	16.54	9.85	12.65	23.57	70.81	66.59	2477454
Sea	Day-time	10.13	7.86	5.71	8.90	84.16	83.24	1918213
	Night-time	19.98	21.48	9.84	9.17	70.19	69.34	3525947
SEVJUL_2								
Land	Day-time	6.57	4.03	0.24	4.31	93.19	91.66	2172792
	Night-time	11.92	10.23	0.47	3.11	87.61	86.65	1640328
Sea	Day-time	7.12	6.31	0.05	0.81	92.82	92.88	3097970
	Night-time	9.75	9.66	0.03	0.14	90.22	90.20	2346190

6. Conclusions

In view of new missions, such as MTG, with improved satellite technologies, able to obtain enhanced quality data, we have experimented a new methodology in the field of cloud detection. We tested CDA potentiality as a new classification methodology to discriminate between clear

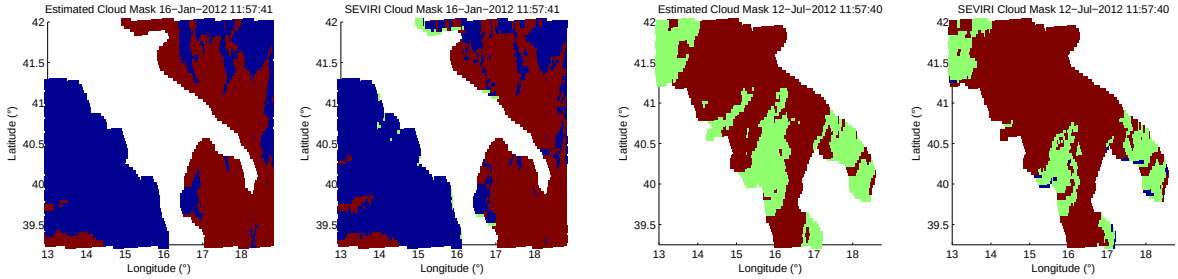

Figure 2. Comparison between CDA estimated cloudy scene and SEVIRI reference cloud mask for one SEVIRI acquisition in two cases. *Green* is for clear over land, *blue* for clear over sea and *dark red* for cloudy. (Left) Sea surface, Day-time for SEVJAN_2. (Right) Land surface, day-time for SEVJUL_2.

and cloudy pixels. We selected ten statistics, related to the cloud properties and computed from SEVIRI thermal IR data. PCA analysis has showed that most of information is contained in the first PC, that we analyzed by CDA. The method needs a reference cloud mask and a training data set in order to find the thresholds that allow us to classify clear and cloudy pixels.
We investigated the area focused on Basilicata region in two different periods aiming to evaluate the seasonal variability. For the validation we obtained quite good results: the methodology performance is higher for the summer data set, both for sea and land surface, with an agreement with respect to SEVIRI reference cloud mask that ranges from 87.6% to 93.2%, while the efficiency of the cloud detection process drops to about 70% when we applied CDA to the winter data set. We plan to improve this methodology analyzing other climatic zones and other statistics, that are descriptive of seasonal variability.

6.1. Acknowledgments

The research in this paper is a part of the corresponding author Ph.D. program, carried out in cooperation with the Italian Space Agency (ASI). The corresponding author also thanks ASI for supporting its Ph.D. scholarship.

References

[1] Ackerman S, Strabala K, Menzel W, Frey R A, Moeller C and Gumley L E, 1998 *J. Geophys. Res.*, **103**, 141
[2] Amato U, Antoniadis A, Cuomo V, Cutillo L, Franzese M, Murino L and Serio C, 2008 *Remote Sens. Envir.*, vol. **112**, 750
[3] Amato U, Lavanant L, Liuzzi G, Masiello G, Serio C, Stuhlmann R, and Tjemkes S A, 2014, *Atmospheric Measurement Techniques*, Vol. **7**, 3355
[4] Derrien M and Le Glau H, 2005, *Int. J. Remote Sens.*, **26**, 4707
[5] Derrien M, 2012, *Validation Report for Cloud Products* (CMa-PGE01 v3.2, CT PGE02 v2.2 and CTTH-PGE03 v2.2), EUMETSAT, Darmstadt.
[6] Heidinger A K, 2004, CLAVR-x Cloud Mask Algorithm Theoretical Basis Document, NOAA/NESDIS/Office of Research and Applications, Washington, D.C
[7] Inoue T, 1985 *J. Meteor. Soc. Japan*, **63**, 88
[8] Inoue T, Ackerman S A, 2002 *J. Meteor. Soc. Japan*, Vol. **80**, 138394
[9] Ricciardelli E, Romano F and Cuomo V, 2008 *Rem. Sens. Envir.*, **112**, 2741
[10] Rossow W B, 1989 *J. Climate*, **2**, 201213
[11] Schmetz J, Pili P, Tjemkes S, Just D, Kermann J, Rota S and Ratier A 2002 *Bull. Am. Met. Soc.*, **83**, 977
[12] Serio C, Masiello G, Venafra S, Amato U and De Feis I, 2013, Application Data for MTG-IRS Cloud Detection Method, *Final Report EUMETSAT*, EUMETSAT, Darmstadt
[13] Tapakis R and Charalambides A G, 2012 *Solar Energy*, **95**, 392−430
[14] Van der Vaart A W, 1998, *Asymptotic statistics*, Cambridge University Press, Cambridge