Cloud identification and classification applied to satellite Far InfraRed Observations (FORUM mission)

William Cossich[1], Iacopo Sbrolli[1], Tiziano Maestri[1]

William.cossich2@unibo.it

1- Dipartimento di Fisica e Astronomia, Università di Bologna, Bologna

Introduction

ICECAP is an innovative machine learning algorithm based on the Principal Components Analysis – PCA performing fast multi-class classifications. ICECAP is applied to a synthetic dataset consisting of up-welling radiances in the 100-1600 cm^{-1} spectral interval at 0.3 cm^{-1} resolution simulating the Sounding Instrument FSI of the FORUM, Far-infrared-Outgoing-Radiation Understanding and Monitoring, mission which has recently been selected by ESA as one of the two candidates for the Earth Explorer 9. FORUM main goal is the study of water vapor and ice clouds by filling the long-standing gap in Far InfraRed (FIR) spectral observations from space.

Synthetic Dataset

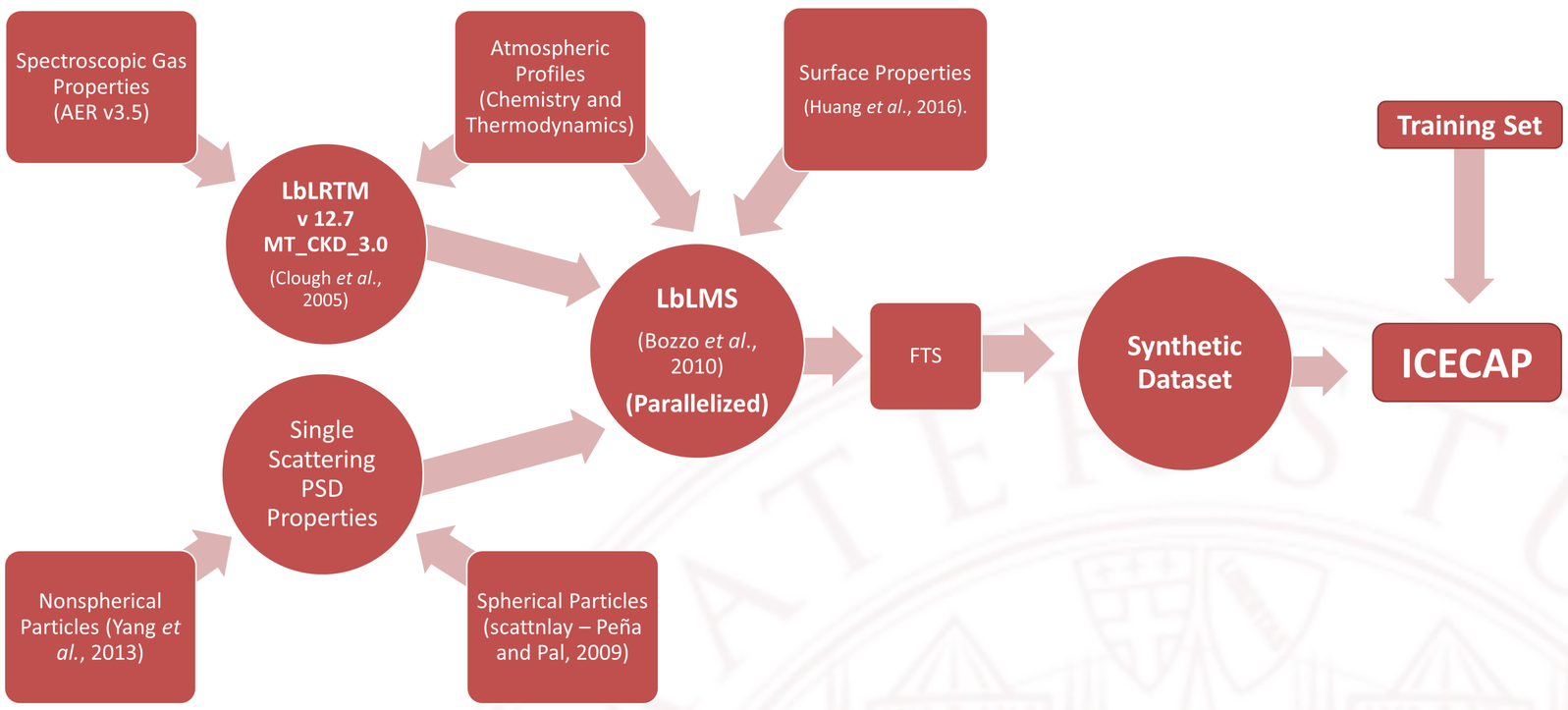

Table 1: Main cloud properties

Cloud Properties	Interval Data
Optical Depth	0.02 – 30.0
Geometrical Thickness (km)	1.0 – 5.0
Top Height (km)	1.0 – 17.0
Ice Cloud Particles Shapes	Hexagonal Plates, Solid Columns, Bullet Rosettes, Aggregates
Ice Particle Effective Dimension (µm)	4 – 100
Water Droplet Effective Radius (µm)	3 – 15
Mixed Phase Particle Effective Radius (µm)	3 – 15

Table 2: Number of simulations for latitudinal belt.

Latitude Zone	# Clear Sky	# Cloudy Sky [Liq/Mixed/Ice (sub visible cirrus)]
Tropics	706	986 [16/61/909 (212)]
Mid Latitudes	615	765 [96/48/621 (173))]
Polar Region	492	532 [0/48/484 (132)]

ICECAP

Pre-Proccessing

- Radiances from synthetic dataset;
- Selection of number of classes (2 or 3);
- Radiance to brightness temperature conversion

Algorithm

- Define the training set (*i*) and built the training set matrices - $\mathbf{TR_i}$ (i.e. cloud and clear sky)
- Calculate the PCA for each TR_i (Equation 1);
- Define the number of principal components associate with real signal variability (Turner, 2006). These are called Information-BEaring principal Components – IBEC's;
- Create the extended training set matrices – $\mathbf{ETR_{i,j}}$, by concatenating TR_i with each spectrum (*j*) of the test set – $\mathbf{TS_j}$ (Equation 2);
- Calculate the PCA for each $ETR_{i,j}$ (Equation 3);
- Calculate Similarity Indices ($\mathbf{SI_{i,j}}$) for each spectrum with each training set (Equation 4);
- Calculate SI difference – SID (Equation 5) for each spectrum
- Compute the Corrected Similarity Index Difference ($\mathbf{CSID_{i,j}}$ – Equation 6) for each spectrum . The $\mathbf{SID(f_{min})}$ is the relative frequency minimum of SID;
- Classify the spectrum by CSID and user-defined thresholds ($\boldsymbol{\theta_1}$ and $\boldsymbol{\theta_2}$), which represent the inner limits of a confidence interval. If $CSID_j < \theta_1$ the spectrum is classified as clear sky, if $CSID_j > \theta_2$ the spectrum is classified as cloudy sky . Spectra whose CSID is that $\theta_2 < CSID_j < \theta_1$ are not classified.

Outputs

- Classification labels for each spectrum of test set (correct, incorrect, uncertain).
- Classification performance (Equations 7 and 8)

Optimization and Analysis

- Optimization of training set, using Ashman D statistic (Equation 9 – Vilasa *et al.*, 2017) and detection performance (Equation 8), to obtain a bimodal SID distribution with the lowest possible mixing. The larger is the D value and the larger is the separation between the 2 distributions;
- Evaluation of the classification performance, using data considering Mid Infrared (MIR), and the full spectrum (MIR and FIR), to quantify the information content in the FIR

$$TREM_i(v,p) = eig(cov(TR_i(v,t)) \quad (1)$$

$$ETR_{i,j}(v,t') = [TR_i(v,t)||row_j(TS(v,j))] \quad (2)$$

$$ETREM_{i,j}(v,p) = eig(cov(ETR_{i,j}(v,t)) \quad (3)$$

$$SI(i,j) = 1 - \frac{1}{2K}\sum_{v=1}^{v_{tot}}\sum_{p=1}^{K}\left|\left(ETREM_{i,j}(v,p)\right)^2 - \left(TREM_i(v,p)\right)^2\right| \quad (4)$$

$$SID(j) = SI(cloudy,j) - SI(clear,j) \quad (5)$$

$$CSID(j) = SI(cloudy,j) - SI(clear,j) - SID(fmin) \quad (6)$$

$$Accuracy_i = \frac{\#\ spectra\ correctly\ classified\ in\ class\ i}{\#\ spectra\ classified\ in\ class\ i} \quad (7)$$

$$Detection\ Performance(DP) = Accuracy_{clear} + Accuracy_{cloudy} \quad (8)$$

$$Ashman\ D = \sqrt{2}\frac{|\mu_1 - \mu_2|}{\sqrt{\sigma_1^2 + \sigma_2^2}} \quad (9)$$

Results

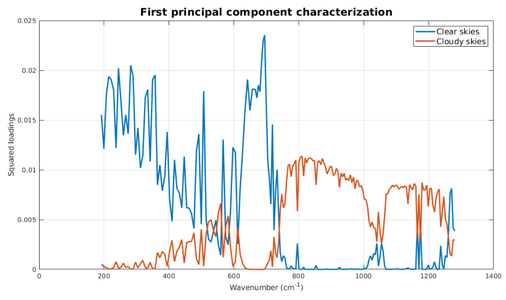

Figure 1. First principal component squared loadings of the clear sky and cloudy sky dataset.

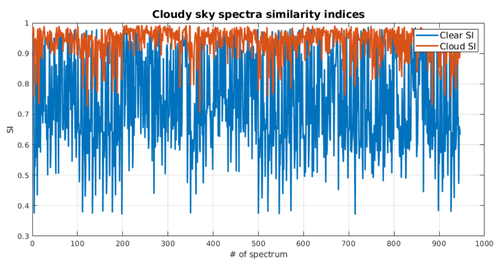

Figure 2. Similarity indices (SI) of cloudy sky spectra for the tropical case

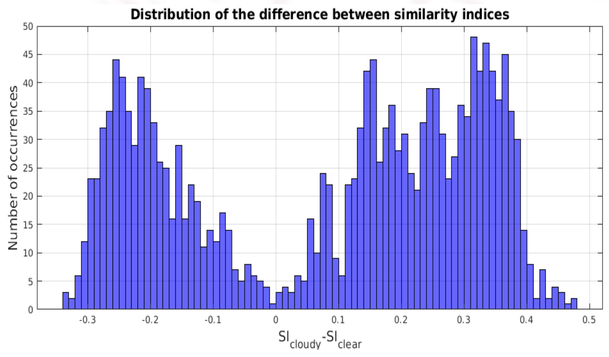

Figure 3. Representation of the bimodal SID distributions for the tropical region simulations

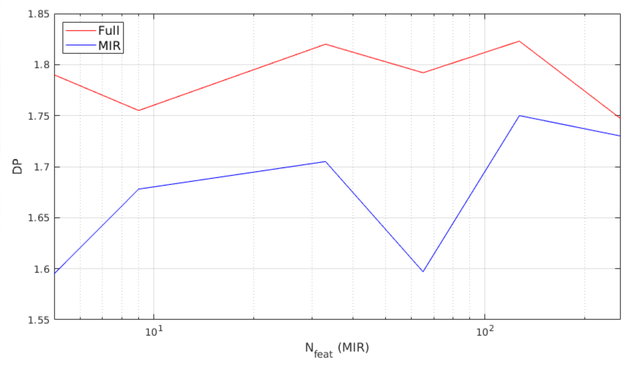

Figure 4. Detection performance (DP) as a function of the number of features (N_{feat}), using the MIR or the Full spectrum (MIR and FIR), for the tropical region simulations

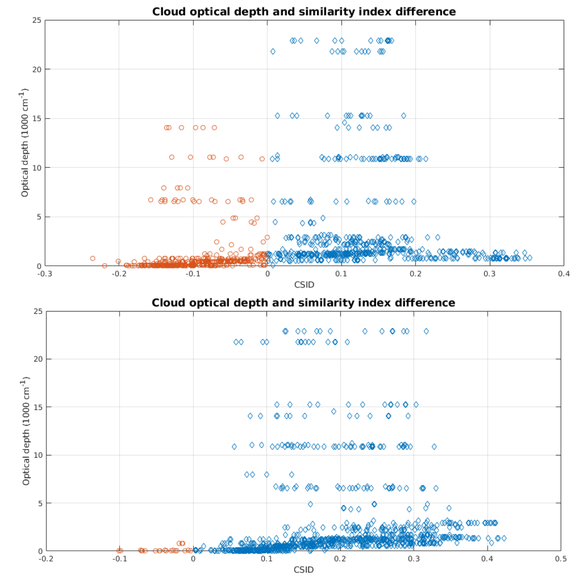

Figure 5. Corrected similarity index difference (CSID) and cloud optical depth, using just MIR (65 channels – upper panel) and Full spectrum (65 MIR + 63 FIR channels – lower panel). Misclassified spectra are plotted in orange circles.

Conclusions

- ICECAP allows clear/cloud identification and cloud classification (cloud phase and type) from IR high spectral resolution satellite data.
- ICECAP depends on a limited number of input parameters and can be applied to operational satellite analysis systems.
- ICECAP is used to evaluate the role of FIR in cloud identification and classification (FORUM Phase-A studies).
- It is shown that the use of FIR channels enhances the detection performances of cirrus clouds and the cloud classification.

Support:
FORUM studies are supported by projects of the Italian Space Agency, **ASI,** (SCIEF project) and of the European Space Agency (**ESA**).