Tropospheric Relative Humidity Profile Statistical Retrievals and their Confidence Interval from Megha-Tropiques Measurements

Ramses Sivira, Hélène Brogniez, Cécile Mallet, Yacine Oussar

TROPOSPHERIC RELATIVE HUMIDITY PROFILE STATISTICAL RETRIEVALS AND THEIR CONFIDENCE INTERVAL FROM MEGHA-TROPIQUES MEASUREMENTS

Ramsés G. Sivira F.[1], **Hélène Brogniez**[1], **Cécile Mallet**[1], **and Yacine Oussar**[2]

[1]*Laboratoire Atmosphères, Milieux, Observations Spatiales, UVSQ/CNRS/IPSL, 11 Boulevard D'Alembert, 78280 Guyancourt, France. Email: ramses.sivira@latmos.ipsl.fr*

[2]*Laboratoire Physique et Etudes Matériaux, ESPCI-ParisTech, 10, rue Vauquelin, 75005 Paris, France. Email: Yacine.Oussar@espci.fr*

ABSTRACT

The combination of the two microwave radiometers, SAPHIR and MADRAS, on board the Megha-Tropiques platform is explored to define a retrieval method that estimates not only the relative humidity profile but also the associated confidence intervals.

A comparison of three retrievals models was performed, in equal conditions of input and output data sets, through their statistical values (error variance, correlation coefficient and error mean) obtaining a profile of seven layers of relative humidity. The three models show the same behavior with respect to layers, mid-tropospheric layers reaching the best statistical values suggesting a model-independent problem.

Finally, the study of the probability density function of the relative humidity at a given atmospheric pressure further gives insight of the confidence intervals.

Key words: Megha-Tropiques, Water Vapor profile restitution, Least Squares Support Vector Machines, Generalized Additive Models, Neural Networks.

1. INTRODUCTION

The amount of water vapor in the atmosphere is a key parameter of the climate system and the understanding of its evolution under a climate evolution relies on documentations of its horizontal and vertical distributions (e.g. [1]). Space borne radiometers are particularly adapted for this documentation thanks to their global coverage of earth system. However, they also bring the disadvantage of their indirect measurement and their reduced vertical resolution. Nowadays, restitution algorithms are more accurate and the vertical resolution are increasingly better.

The present study is motivated by a desire to explore the potential of recent statistical methods in this inverse problem. Three different kind of statistical models are implemented and tested to define their respective performance in this context.

2. PROBLEM DESCRIPTION: INPUTS AND OUTPUTS

• Predictors: satellite observations

Megha-Tropiques is an Indo-French satellite, launched in October 2011, that is dedicated to the observation of the energy budget and water cycle within the tropical belt ($\pm$ 30° in latitude). We focus on two of its instruments: MADRAS, a microwave imager for the observation of rain and clouds (Microwave Analysis and Detection of Rain and Atmospheric Structures) and SAPHIR, a microwave sounder of the tropospheric water vapor (Sondeur Atmosphérique du Profil d'Humidité Intertropicale par Radiométrie). SAPHIR is a cross-track sounder observing the Earth's atmosphere with 6 channels in the 183.31 GHz water vapor strong absorption line. MADRAS is a scanning imager with 9 channels ranging from 18.7 GHz to 157 GHz [2].

At the time of the study, no observation were yet available from the Megha-Tropiques platform, yielding to use synthetic data to overcome the problem. The transfer radiative model (RTTOV, [3]) is used to simulate the SAPHIR and MADRAS brightness temperatures (hereafter BTs) from the radiosoundings thermodynamic profiles. Whereas of the main goal of this work is to design a general retrieval algorithm for both land and oceanic surfaces, the present study focuses on the oceanic situations and puts aside the additional difficulties induced by the continental emissivities that contribute strongly to the microwave upwelling radiation [4]. Finally, with no information on the radiometric noise of the instruments were available, a null noise was applied.

The Fifteen BTs are normalized and a Principal Component Analysis (PCA) is performed on the input vector to obtain uncorrelated and linearly independent variables. For each model, performances obtained with normalized

BTs and PCA are compared to test their impact on the retrieval.

According to SAPHIR and MADRAS characteristics, their channels are centered in a specific frequencies of the spectrum with the aim to obtain different kinds of information. In the case of SAPHIR, channels take information from different altitudes (as far from 183.31 GHz, channels obtain information from higher altitudes of the atmosphere). In the case of MADRAS, channel 3 is centered at 23.8 GHz, which obtains information of total water vapor content. In consequence, the information registered from each channel has specific characteristics that would be important to specific atmosphere altitudes, in the other hand, the same channel would contributes with complementary information or even could decreases the information quality to another atmosphere altitude. Knowing this behavior, we have ranked for each layer the most relevant input variables using the Gram-Schmidt orthogonalization procedure. In the present study, we implement the GSO procedure method according to a wrapper approach that perform the selection iteratively [5]. The input vector of statistical retrieval algorithms are named $\mathbf{BT}$.

• The predictants: layered relative humidity profiles

The relative humidity profiles are provided by the operational radiosounding archive used in the ECMWF reanalyses assimilation process, which have been quality checked and reformatted by Laboratoire de Météorologie Dynamique in tropical oceans (30°S-30°N) over the 1990-2007 period. We also added a physical constraint on the relative humidity in order to remove the extremely dry profiles (RH $>$ 2 %) and the super-saturated layers encountered in the upper troposphere (RH $<$ 150 %, e.g. [6]).

Each profile consists in 22 levels of relative humidity ranging from surface (1000 hPa) to the stratosphere (86 hPa). For practical reasons we simplified these 22 levels profiles to seven layers reduced profiles. To accomplish this task, we used self-Organized maps to produce a low-dimensional and discretized representation of the output dataset with the aim to group the original levels in layers adapted to the model, taking into account their topology. In consequence, we build an output data set composed by 7 layers (86-106 hPa, 106-250 hPa, 250-380 hPa, 380-650 hPa, 650-850 hPa, 850-950 hPa and 1013 hPa) obtained from the original set.

In order to account for the well known exponential relation between brightness temperature and atmospheric optical thickness ([7]) the application of the exponential (EXP) allows to compare retrieved exp(RH) instead of RH [5].

3. REGRESSION METHODS

Three statistical models were tested, the first model is an additive model (generalized additive model, hereafter GAM, [8]). The second model is the Multi-Layer Perceptron (MLP) as defined by [9], that is the most widespread technique of non-linear regression [10]. The third technique implemented in our study is a Least-Square (LS) kernel method related to Support Vector Machine (SVM) called the LS-SVM method [11].

In an inverse problem, the quality (or the accuracy) of the results are conditioned on several parameters : the clarity of the input-output relationship, the relevance of the set of inputs, the adjustment of the parameters of the retrieval models for their optimization and the evaluation/validation method. Those aspects need to be optimized for each model in order to obtain the best possible configuration before confronting the results. For each model, the parameters optimization are:

• MLP: The MLP internal parameter, the matrix weights W of the connections must be learned from a training data set. These weights are determined in order to perform the optimal association that is to say to obtain the minimum of a cost function. We chose to minimize the mean quadratic error, $J(\mathbf{W})$, computed on the training data set.

To obtain the minimum of this multidimensional cost function we used the LevenbergMarquardt technique ([12]) as this technique is more powerful than the conventional gradient descent techniques. Theory shows that, if the architecture of the MLP is well-chosen, the minimization of J(W) is well achieved, and the observation set is consistent with the true field of variables, the MLP gives an accurate approximation of the conditional average of the relative humidity RH^i given a Brigthness Temperature vector $\mathbf{BT}$.

The learning phase may require long computations due to the minimization process. But during the operational phase the computation time is very fast because all the minimizations have been done during the learning phase and computations are only algebraic operations.

• LS-SVM: The LS-SVM optimization problem can be cast into a dual form with unknown parameters $\boldsymbol{\alpha}$ and b, $\boldsymbol{\alpha}$ being the vector of the Lagrange multipliers. The parameters can be computed by resolving the following system of linear equations:

$$\begin{bmatrix} \mathbf{K} + \frac{1}{2C}\mathbf{I_N} & \mathbf{1}_N \\ \mathbf{1}_N^T & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ b \end{bmatrix} = \begin{bmatrix} \mathbf{z} \\ 0 \end{bmatrix} \quad (1)$$

with $\mathbf{1}_N = [1, 1, ..., 1]^T$, $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, ..., \alpha_N]$ and $\mathbf{I_n}$ is the identity matrix. Finally, the LS-SVM model becomes:

$$f(\mathbf{BT}) = \widehat{RH^i} = \sum_{k=1}^{N} \alpha_k \, K(\mathbf{BT}, \mathbf{BT}_k) + b \quad (2)$$

where α and b are the solution to eq. (1).

• GAM: Generalized Additive Model can be described

by the following expression:

$$g(\mathbb{E}(RH^i|\mathbf{BT})) = \mu^i = \epsilon^i + f_1(BT_1) + f_2(BT_2) + ... + f_p(BT_p) \quad (3)$$

where g is a linearizing link function between the expectation of RH^i given $\mathbf{BT}$ and the additive predictors $f_j(BT_j)$, which are smooth and generally non-parametric functions of the covariates $BT_1, ..., BT_p$. Finally ϵ^i is the residual that follows a normal distribution. Here, penalized regression cubic splines are used as the smoothing functions and are estimated independently of the other covariates using the "back-fitting algorithm". Part of the model-fitting process is to choose the appropriate degree of smoothness, which is done through the generalized cross-validation criterion $n\,D/(n-dof)^2$, where dof is the effective degrees of freedom of the model, n is the number of data and D is the deviance of the model. Detailed information about this method can be found in [13].

4. STATISTICAL MODELS INTERCOMPARISON

The whole available data from a set of 1631 simulated examples ($BT_k, RH_k^1, ..., RH_k^7$). We randomly divided this dataset in two subset: The training and validating dataset, composed of 1140 examples. The remaining examples form the test set.

Figure 1 shows comparisons between the observed and the estimated relative humidity corresponding to layers 4^{th} and 7^{th} using the three studied models. Our first observation is the similar performances obtained from models in each layer, in fact, differences in correlation coefficient are lower than 5% at layers with highest accuracy (4^{th}) and lower than 20% at layers with lower accuracy (7^{th}). This characteristic suggest that the models' accuracy is layer-dependant, meaning that it is strongly constraint by the physical aspects of the inverse problem. This aspect is due to the distribution of SAPHIR channels, the overlapping of their weight functions at mid-tropospheric zones allows that the same atmospheric situation can be observed by all channels with a possible increase of the estimation accuracy; the contrary effect can be observed at extreme atmospheric zones, where these weight functions show their minimal values. This behavior could explain the accuracy differences between middle layers (3^{th} and 4^{th}), with a correlation coefficient bigger than 0.93 and 0.95 respectively, and extremes ones (1^{st} and 7^{th}), with a maximal correlation coefficient of 0.56 and 0.3 respectively. In order to inputs ranking suggested by the Gram-Schmidt process, we observed a bigger relevance for SAPHIR channels for high-altitude layers and this relevance decreases progressively to lower altitudes layers; inversely, MADRAS channels (specially the channel centered at 23.8 GHz) achieve their highest importance at lower layers, consistent with the characteristics of this channel, which is normally used for the restitution of the total water vapor content, knowing the great water vapor percentage inside this low-altitudes layers.

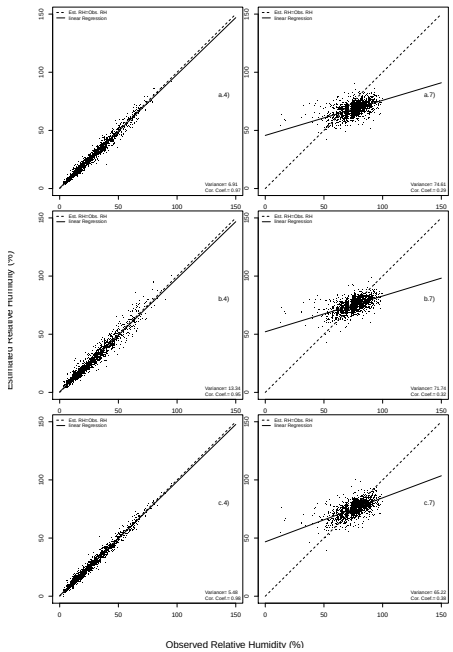

Figure 1: Comparisons between the relative humidity (in %) observed in the radiosounding and the relative humidity (in %) estimated by the statistical model for layers 4 (left) and 7 (right). Line (a) is for the MLP algorithm, line (b) is for GAM and line (c) is for LS-SVM. The $y = x$ line (dashed), the linear regression (plain) and the Pearson's correlation coefficient are also provided.

An analysis of the correlation coefficients and error variance reveals that the LS-SVM method gives the best result for 5 layers over 7 layers considered in this study. This technique also outperforms slightly the others on 2 layers. Theoretically, these 3 learning methods are equivalent. However, the conditions of their implementation are somewhat different. Since the LS-SVM are linear-in-their-parameters models, an exact validation method was implemented.

Finally, the non-biased characteristic for all models allows us to infer that the learning process was satisfactory, obtaining good generalization capabilities.

5. ESTIMATION METHODS OF CONDITIONAL PROBABILITY DISTRIBUTION AND CONFIDENCE INTERVALS

Once the relative humidity is estimated for each layer, the knowledge of corresponding uncertainties is required in order to know the accuracy of these estimations. To determine this characteristic we build a dependant confidence interval through the estimation of the error conditional

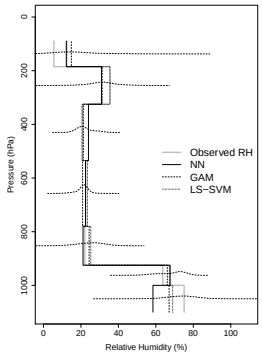

Figure 2: Example of Estimation of relative humidity profile. The observed profile is the thick gray line and the 3 estimations (plain(NN), dashed(GAM), dots(LS-SVM)) are in black. Conditional density distribution estimated from GAM model errors are showed for each layer (dashed) using Mixture Gaussian Model with m=2

probability density function ($p((RH - \widehat{RH})|BT)$). We build this probability function using the Mixture Gaussian Model [14] combining pondered gaussian functions as:

$$p(\epsilon_i|BT_i) = \sum_{j=1}^{m} \lambda_j * \phi(\epsilon_i|BT_i^T\beta_j, \sigma_j^2) \qquad (4)$$

where $\epsilon = (RH - \widehat{RH})$ and $\phi(\epsilon_i|BT_i^T\beta_j, \sigma_j^2)$ is the normal density with mean $\epsilon_i|BT_i^T\beta_j$ and variance σ_j^2. With this probability density function it is possible to estimate confidence intervals for each relative humidity value estimated.

Figure 2 shows a particular profile: the three estimated profiles corresponding to the three algorithms and the conditional probability density function obtained for the GAM model error. The error model, whose validation is in progress, has a Gaussian behavior for the middle layers. For extreme layers, the obtained density function is a mixture of two gaussian, with flatter behavior and sometimes clearly bimodal.

6. CONCLUSIONS

Through the combination of the two radiometers MADRAS and SAPHIR, the recent Megha-Tropiques mission gives the opportunity of important improvements in water vapor profile estimations. The present study explores the potential of three statistical methods for relative humidity profiles retrieval given a set of brightness temperature. Very similar results are obtained with the three considered approaches. However, significant differences in the accuracy of restitution are observed according to the altitude of considered layer, these differences are independent from the used method. Given a set of brightness temperature, a mixture gaussian model of the error probability density function is associated, so as to add a confidence level to the estimated humidity.

REFERENCES

[1] I. Held and B. Soden. Water vapour feedback and global warming. *Annu. Rev. Energy Environn.*, 25:441–475, 2000.

[2] Hélène Brogniez, Pierre-Emmanuel Kirstetter, and Laurence Eymard. Expected improvements in the atmospheric humidity profile retrieval using the Megha-Tropiques microwave payload. *Q. J. R. Meteorol. Soc.*, page doi:10.1002/qj.1869, 2011.

[3] M. Matricardi, F. Chevallier, G. Kelly, and JN. Thépaut. An improved general fast radiative transfer model for the assimilation of radiance observations. *American Meteorological Society*, 130:153–173, 2004.

[4] R. Bennartz and P. Bauer. Sensitivity of microwave radiances at 85-183 GHz to precipitating ice particles. *Radio Sci.*, 38:doi:10.1029/2002RS002626, 2003.

[5] R. Sivira, H. Brogniez, C. Mallet, and Y. Oussar. Comparison of non-linear methodes for the retrieval of relative humidity profiles in the context of the megha-tropiques measurements. *Journal of Atmospheric and Oceanic Technology*, 2012.

[6] K. Gierens, U. Schumann, M. Helten, H. Smit, and A. Marenco. A distribution law for relative humidity in the upper troposphere and lower stratosphere derived from three years of MOZAIC measurements. *Ann. Geophysicae*, 17:1218–1226, 1999.

[7] F. Ulaby, R. Moore, and A. Fung. *Microwave remote sensing: Active and passive.*, volume 1. Addison-Wesley, 1981.

[8] T. Hastie and R. Tibshirani. *Generalized Additive Models*. Chapman & Hall/CRC, 1990.

[9] D. Rumelhart, G. Hinton, and R. Williams. *Learning internal representation by error propagation. in: Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1. D. E. Rumelhart and J. L. McClelland, 1986.

[10] S. Haykin. *Neural Networks: A Comprehensive Foundation*. IEEE Press, New York, NY, USA, 1994.

[11] J. A. Suykens, T. Van Gestel, J. de Brabanter, B. de Moor, and J. Vandewalle. *Least Squares Support Vector Machines*. World Scientific, 2002.

[12] D. Marquardt. An algorithm for least squares estimation of non-linear parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11:431–441, 1963.

[13] S. Wood. *Generalized Additive Models, an Introduction with R*. Chapman & Hall/CRC, 2006.

[14] T. Benaglia, D. Chauveau, D. Hunter, and D. Young. mixtools: An r package for analyzing finite mixture models. *Journal of Statistical Software*, 32:1–29.