Atmos. Meas. Tech. Discuss., 7, 8983–9023, 2014
www.atmos-meas-tech-discuss.net/7/8983/2014/
doi:10.5194/amtd-7-8983-2014

Atmospheric Measurement Techniques Discussions

A relative humidity profile retrieval from Megha-Tropiques observations without explicit thermodynamical constraints

R. G. Sivira[1], H. Brogniez[1], C. Mallet[1], and Y. Oussar[2]

[1]Université Versailles St-Quentin, Sorbonne Universités, UPMC Univ. Paris 06, CNRS/INSU, LATMOS-IPSL, Guyancourt, France
[2]Laboratoire de Physique et d'Etudes des Matériaux (LPEM), UMR8213, ESPCI-ParisTech, Paris, France

Received: 1 August 2014 – Accepted: 15 August 2014 – Published: 3 September 2014

Correspondence to: H. Brogniez (helene.brogniez@latmos.ipsl.fr.)

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

Abstract

A statistical method trained and optimized to retrieve relative humidity (RH) profiles is presented and evaluated with measurements from radiosoundings. The method makes use of the microwave payload of the Megha-Tropiques plateform, namely the SAPHIR sounder and the MADRAS imager. The approach, based on a Generalized Additive Model (GAM), embeds both the physical and statistical characteritics of the inverse problem in the training phase and no explicit thermodynamical constraint, such as a temperature profile or an integrated water vapor content, is provided to the model at the stage of retrieval. The model is built for cloud-free conditions in order to avoid the cases of scattering of the microwave radiation in the 18.7–183.31 GHz range covered by the payload. Two instrumental configurations are tested: a SAPHIR-MADRAS scheme and a SAPHIR-only scheme, to deal with the stop of data acquisition of MADRAS in January 2013 for technical reasons. A comparison to retrievals based on the Multi-Layer Perceptron (MLP) technique and on the Least Square-Support Vector Machines (LS-SVM) shows equivalent performance over a large realistic set, promising low errors (bias $< 2.2\,\%$) and scatters (correlation > 0.8) throughout the troposphere (150–900 hPa). A comparison to radiosounding measurements performed during the international field experiment CINDY/DYNAMO/AMIE of winter 2011–2012 confirms these results for the mid-tropospheric layers (correlation between 0.6 and 0.92), with an expected degradation of the quality of the estimates at the surface and top layers. Finally a rapid insight of the large-scale RH field from Megha-Tropiques is discussed and compared to ERA-Interim.

1 Introduction

The atmospheric water vapor is a key parameter of the climate system and the understanding of its variation under a climate evolution relies on a thorough documentation of its horizontal and vertical distributions (Held and Soden, 2000; Roca et al.,

AMTD
7, 8983–9023, 2014

2010; Sherwood et al., 2010). It is a major greenhouse gas, part of a strong positive feedback that amplifies the warming caused by increases of greenhouse gases in the atmosphere (Spencer and Braswell, 1997; Hall and Manabe, 2000; Held and Soden, 2006), and, because of its short life cycle compared to other species, its distribution is mainly influenced by natural processes that occur at all scales, from the large scale cells of the atmospheric circulation to the scale of the hydrometeor (e.g. Houze and Betts, 1981; Pierrehumbert and Roca, 1998; Pierrehumbert et al., 2007).

While direct measurements by radiosondes are the most simple ways to look at the vertical structure of the relative humidity (RH) field, the network of stations (permanents or not) is inequally distributed between the two hemispheres and there is a clear gap of data over the oceans (Durre et al., 2006). The climate record built by aggregating the observations from the various operational sensors used worldwide (e.g. Vaïsala, MEISEI, IM-MK3, MODEM) requires regular intercomparison campaigns, such as those organized by the World Meteorological Organization (Nash et al., 2005) and the development of dedicated correction schemes in order to correct most of the observational errors (such as the drying effect of the radiative heating on the Vaïsala sensor or the insensitivity of the MEISEI system under dry conditions). Quite recently Wang and Zhang (2008) have summarized the systematic instrumental biases between several versions of the Vaïsala system that, if uncorrected, would affect analyses of the global moisture field. An alternative is the fleet of space-borne radiometers with channels located in spectral bands sensitive to the absorption by water vapor. Such instruments provide a more global sight of the distribution of the water vapor field since the late 70's, in the thermal infrared part of the spectrum (in the 6.3 µm band, e.g. on the METEOSAT satellites of the European Organization of the Exploitation of Meteorological Satellites agency -EUMETSAT-) and in the microwave range (near the 183.31 GHz absorption line, e.g. on the Advanced Microwave Sounding Unit-B -AMSU-B- radiometer of the National Oceanic and Atmospheric Administration -NOAA- satellites). These are however indirect estimations of the RH from the measured upwelling radiation and are thus strongly linked to the constraints of the underlying inverse problem (RH = f(radiation)).

AMTD
7, 8983–9023, 2014

The aim of this study is to explore the potential of the Megha-Tropiques mission, operating since October 2011, to retrieve atmospheric RH profiles without explicitly prescribing thermodynamical parameters. In fact, this work is also motivated by a desire to explore the potential of a purely statistical methods in the following problem: given a set of brightness temperatures (BTs) provided by a space-borne radiometer, what is the vertical distribution (i.e. a profile) of RH and what are the expected limits of such an approach? Many retrieval approaches exist but, to our knowledge, none of them estimate the RH profile from a simple input dataset restricted to the BTs. Indeed, most of the approaches are physically-based iterative techniques: a n dimensional variational algorithm that converges to the least biased profile using also other inputs as prior knowledge of the system under study (such as surface emissivity, temperature profile and sometimes a prior water vapor profile for BT simulations). These variational techniques are well established (Kuo et al., 1994; Cabrera-Mercadier and Staelin, 1995; Rieder and Kirchengast, 1999; Blankenship et al., 2000; Liu and Weng, 2005) and it would be unnecessary to reinvent a similar algorithm. The focus here is really on the microwave payload of the satellite, namely SAPHIR and MADRAS, both dedicated to improve the documentation of the atmospheric water cycle, and not on the choices of the relaxation scheme or the a priori constraints that could improve the retrieval. The approach is to learn the relationship between the inputs (i.e. the BTs) and the output (i.e. the RH at a specific atmospheric layer) directly from a training set that implicitly contains all the relevant information such as the statistical distribution of the atmospheric RH or the radiative transfer equation from the set of BTs.

In a previous paper, Brogniez et al. (2013) showed the expected improvements for the estimation of the RH profiles thanks to the combination of those two instruments, highlighting the gain of information for both ends of the troposphere when only a subset of the channels of MADRAS are used combined to SAPHIR measurements. As in Brogniez et al. (2013), the retrieval technique is based on the Generalized Additive Model (hereafter GAM, Hastie and Tibshirani, 1990) and its ability to model multi-variate and non-linear relationships. The choice of GAM over other retrieval techniques is relatively

AMTD
7, 8983–9023, 2014

subjective. So to ensure that the main patterns are independant from the choice of the statistical model, a comparison against two other models is done. We consider two machine learning regression algorithms as references for this purpose. The first one is the Multi-Layer Perceptron (MLP) as defined by Rumelhart et al. (1986), generally considered as reference because it is the most common approach to develop non-parametric and non-linear regression in various application domains. As shown in Thiria et al. (1993) MLPs are able to model complex inverse functions in processing noisy data. MLPs have been successfully applied in remote sensing application, with or without prior information (e.g. Mallet et al., 1993; Cabrera-Mercadier and Staelin, 1995; Aires and Prigent, 2001; Karbou et al., 2005; Aires et al., 2010). The second one is the Least Square-Support Vector Machines (LS-SVM) (Suykens et al., 2002), that belongs to the family of kernel methods. LS-SVMs are models linear in their parameters and their training process consists in a quadratic minimization under constraints. These properties confer to them the ability to build models with high generalization capabilities by avoiding overfitting and controlling model complexity. Numerous analysis involving real data in other areas (Balabin and Lomakina, 2011; Wun-Hua et al., 2006) have shown that SVM-based techniques are comparable in efficiency to MLPs.

The description of the data at hand and of the context of the work are made in Sect. 2. The three non-linear models, GAM, MLP and LS-SVM and their design for the study are detailled in Sect. 3, while Sect. 4 is dedicated to the evaluation of the estimations over a realistic dataset in order to have a large sample of evaluation. The application to Megha-Tropiques measurements is discussed in Sect. 5 with a comparison to radiosounding measurements in order to consider true data. Section 6 finally draws a conclusion on the study and discuss the ongoing work.

AMTD
7, 8983–9023, 2014

2 Data and context

2.1 Overview of the Megha-Tropiques mission

Megha-Tropiques is an Indo-French satellite that is dedicated to the observation of the energy budget and of the water cycle within the tropical belt (±30° in latitude). The platform carries four instruments: MADRAS, a microwave imager for the observation of rain and clouds (Microwave Analysis and Detection of Rain and Atmospheric Structures), SAPHIR, a microwave sounder of tropospheric RH (Sondeur Atmosphérique du Profil d'Humidité Intertropicale par Radiométrie), ScaRaB, a wide band instrument for the measurement of radiative fluxes at the top of the atmosphere (Scanner for Radiation Budget), and ROSA, a GPS receiver (Radio Occultation Sounder for the Atmosphere). In this study, we focus on the combined use of SAPHIR and MADRAS observations, whose characteristics are listed in Table 1 together with their in-flight radiometric sensitivities as estimated by the CNES space agency (Karouche et al., 2012). SAPHIR is the main instrument for RH profiling with 6 channels in the 183.31 GHz strong absorption line of water vapor. The first channel is close to the center of the line and is aimed at reaching the upper levels of the troposphere while the sixth channel is located on the wings of the absorption line and provides a deeper sounding of the atmosphere. In the context of the RH estimations, the measurements provided by MADRAS (dedicated to rainfall estimation) will obviously better constraint the problem since 23.8 GHz measurements are generally used for the determination of the total water vapor content (Schaerer and Wilheit, 1979) and the two 157 GHz channels can help removing the surface contribution to the upwelling radiation (English et al., 1994). However, due to a mechanical anomaly affecting the rotating mechanism of MADRAS, its measurements are considered invalid since 26 January 2013 and only SAPHIR observations are available to the scientific community after this date (joint CNES and ISRO communication done on 24 September 2013).

AMTD
7, 8983–9023, 2014

2.2 Data description

High quality RH soundings sampling the tropical troposphere, reasonably collocated in space and time with Megha-Tropiques observations are quite scarce, yiedling to use a synthetic training set to overcome the problem. This set is made of thermodynamical profiles representative of the 30° N–30° S atmosphere and of the associated BTs simulated using a radiative transfer model. This method embeds both the physical and statistical characteristics of the inverse problem in the training phase.

2.2.1 The radiosounding profiles

The RH profiles come from the operational radiosounding archive used in the ECMWF reanalyses assimilation process, which have been quality checked and reformatted by the Laboratoire de Météorologie Dynamique (namely the ARSA database, http://ara.abct.lmd.polytechnique.fr/index.php?page=arsa). The main aspects of the applied treatments are (i) the discarding of incomplete profiles both in temperature (threshold of 30 hPa) and humidity (threshold of 350 hPa), (ii) a vertical extrapolation of the profiles up to 2.10^{-3} hPa, considered as the top of the atmosphere, using climatologies and (iii) a projection on a 43-level fixed pressure grid with a surface level at 1013.25 hPa. Here the profiles are vertically restricted to the troposphere, from the surface up to 85 hPa, thus yielding to have 22-levels profiles, and an additional physical constraint on the RH has been applied in order to remove the extremely dry profiles (RH < 2 %) and the super-saturated layers encountered in the upper troposphere (RH > 150 %, e.g. Gierens et al., 1999; Read et al., 2007, 2001, the RH being defined with respect to ice or liquid water depending on pressure and temperature). In the following the term RH will refer to the relative humidity computed with respect to the liquid phase of water only.

Currently, only clear-sky conditions are considered. Indeed, as underlined by Brogniez et al. (2013), the representation of the cloudy conditions in a training database still presents a limit because reference profiles of cloudy situations with known uncertainties

AMTD
7, 8983–9023, 2014

are difficult to gather which introduces additional errors in the methodology. Moreover, while a similar effort (not shown) has been done over land, we restrict the present analysis to oceanic situations in order to put aside the consideration of the strongly variable continental emissivity (Karbou et al., 2005). This particular case and its treatment are mentioned in the last section. The base is finally made of 1631 thermodynamic profiles that cover the tropical oceans (30° S–30° N) over the 1990–2007 period.

2.2.2 Definition of the considered atmospheric layers

Given a set of BTs, the expected accuracy in the estimated RH will obvioulsy highly depend on the atmospheric area under consideration. Therefore for a specific pressure level, the relevant inputs will not be necessarily the same than for another level. One can indeed expect that the estimation of RH of mid-tropospheric levels will not significantly benefit from MADRAS measurements, while these should be an asset for a surface layer. This is why layer-dependent models are considered here. The RH profiles were analyzed to group the 22 original levels in relatively homogeneous layers. First, the analysis of the variance-covariance matrix determined groups of correlated successive levels. Then, self-organized maps (SOM) (Kohonen, 1982, 2001) are used to visualize the 22-D original profiles as 22 2-D images (not shown). This visual analysis allows to group the original levels with similar patterns taking into account linear and nonlinear relationships. This visualization is also used to analyse the patterns of the errors of estimation. The analysis of these SOM yields to combine the original pressure levels with a a semi-empirical iterative method in order to have layers with minimal variance of RH and minimal mean-median distance. From this reduction, the training RH dataset is composed of 7-layer profiles (grossly: 85–100, 130–250, 275–380, 425–650, 725–850, 900–955 and 1013 hPa).

Figure 1 shows the result of this vertical reduction using box-and-whiskers diagrams in order to present the main characteristics of the atmospheric layers (median, first and third quartiles, upper and lower limits of the distribution). The superimposed weighting functions of SAPHIR are drawn to recall that this radiometer is designed to focus on the

AMTD
7, 8983–9023, 2014

free troposphere (layers 3 to 6), with very few informations near the tropopause and in the boundary layer.

2.2.3 Synthetic Megha-Tropiques observations

The RTTOV fast radiative transfer model, version 9.3 (Radiative Transfer for Television and Infrared Observation Satellite Operational Vertical Sounder, Matricardi et al., 2004), is used to simulate SAPHIR and MADRAS BTs from the thermodynamic profiles described above. Because the surface emissivity contributes strongly to the upwelling radiation in the microwave domain (Ulaby et al., 1981; Bennartz and Bauer, 2003) its implementation is important for realistic radiative transfer simulations. Indeed, the surface emissivity affects the observed microwave upwelling radiation in the 2 lower channels of SAPHIR (183.31 ± 6.8 and ±11 GHz, with a difference of BT of up to 5K for some cases) and of all the 9 channels of MADRAS. In RTTOV-9.3, the oceanic surface emissivities are computed with the FASTEM-3 surface model (Fast Emissivity Model, Deblonde and English, 2001) using the 10 m wind speed. Here we use the wind extracted from a 18 year climatology from the ECMWF ReAnalysis (ERA) Interim model covering the same period as the thermodynamic profiles (1990–2008). Over continental surfaces, the emissivity atlas of Prigent et al. (2006), elaborated from 10 years of Special Sensor Microwave Imager (SSM/I) observations, is prefered over the internal module of RTTOV. Finally SAPHIR BTs are simulated only in the nadir geometry whereas the simulations of MADRAS BTs are performed at the radiometer's constant viewing angle of 53.5°.

The simulations also make use of the instrumental noise to have a realistic base of work. The radiometric sensitivity is often considered as the intrumental noise since it gives the minimum variation in the measured upwelling radiation that a specific channel can detect (noise-equivalent ΔT: NEΔT, in K). This noise may be considered as additive and modeled as realisations of a random variable following a normal distribution with a zero mean and a standard deviation equal to the NEΔT value for each channel.

AMTD
7, 8983–9023, 2014

The simulated datasets are re-built by aggregating 10 noisy samples for each original sample.

With the stop of MADRAS after almost 15 months of measurements, two configurations of a RH retrieval method have been considered: a SAPHIR-only scheme and a SAPHIR-MADRAS scheme, the latter being associated to a selection of the optimal channels. For the former configurations all SAPHIR channels are used. For the latter configuration, a selection of the BTs is performed because the BTs that will be significantly relevant in the RH retrieval at a given layer will not necessarily be the same set when considering another layer. For this purpose the optimal subset of channels is determined thanks to the Gram-Schmidt orthogonalization (GSO) procedure (see Chen et al., 1989). Here, since the size of the whole input set (the BTs) does not exceed 15 elements, the GSO procedure is implemented according to a wrapper approach. This is performed for each atmospheric layer.

3 Description of the non-linear models

3.1 General aspects

To ensure the consistency between the mathematical descriptions of the three statistical models, the notation will be as follows: the estimation of the RH^i at a specific layer i ($i \in [1;7]$), namely the output, is performed from a vector of BTs, the inputs, which is a p dimensional covariate noted **BT** ($p \in [1;15]$). Thus, for each layer i the training dataset is made of $(p+1)$-tuples $\left\{\mathbf{BT}_k, RH_k^i\right\}_{k=1}^{N}$, where the cardinality of the set N is 16310 (1631 profiles × 10 noisy reproductions).

The GAM, MLP and LS-SVM models learn the relationship between **BT** and RH^i of a layer i by building a model directly from the data at hand without other a priori information. For one particular channel, the nonlinearity between the measured BTs and the RH^i is more or less strong (Soden and Bretherton, 1993; Stephens et al.,

AMTD
7, 8983–9023, 2014

1996; Brogniez and Pierrehumbert, 2006; Brogniez et al., 2013). This is illustrated on Fig. 2 for the BT of the 183.31 ± 1.1 GHz channel of SAPHIR and the RH^3 and RH^4, extracted from the synthetic base. Hence the approach choosen is to adjust and optimize the **BT**-to-RH^i relationships separately for each of the 7 layers.

The dataset described in Sect. 2 is randomly divided into two subsets: a subset of 2/3 of the N samples (∼ 11 000 samples) is dedicated to the training and to the validation of the models while the remaining 1/3 forms the test set (∼ 5000 samples). Some parameters of the three modeling methods have to be adjusted and the selected models are those with the best generalization capabilities. These parameters are tuned to minimize the validation error which is an empirical estimation of the generalization error. Thus the selection of the models consists in the involvement of an efficient validation method. Various validation techniques exist in the literature (Hastie et al., 2009). The most popular techniques are probably the cross validation method and the leave-one-out (LOO) technique, which are implemented according to the modeling method. Note that since the three modeling methods will be compared, we focus on efficient validation techniques and pay less attention to the computational burden they involve.

The input vector **BT** is normalized (zero mean and unit variance). While such normalization does not affect the estimation provided by GAM (but only the relative weight of each predictor in the fit), the normalized input dataset is the same for all models in order to simplify the process. A Principal Component Analysis (PCA) is also implemented on the **BT** to feed each statistical model with uncorrelated and linearly independent data. Indeed, the weighting functions of the 6 channels of SAPHIR slightly overlap each other to cover the entire absorption line. As a result, while each channel receives mainly the radiation emitted by a given layer of the atmosphere, contributions from layers above and below are not negligible, yielding to some interdependencies between the channels. Finally, in order to account for the known exponential relationship between the BT in the 183.31 GHz line and the atmospheric RH (for instance at 183.31 ± 1.0 GHz, see Spencer and Braswell, 1997 and Buehler and John, 2005), the use of the exponential function is also considered, which has also the advantage to ensure the retrieval of

AMTD
7, 8983–9023, 2014

positive values. The effect of the PCA and of the exponential function have been evaluated for each statistical model for each layer i. The configuration with the smallest validation error was selected.

3.2 Generalized Additive Model

GAMs have recently started to be used in environmental studies as a surrogate to traditional MLP thanks to their ability to model nonlinear behaviors while providing a control of the physical content of the statistical relationships (Wood, 2006). Therefore, among the recent works, one can cite the use of GAM to perform a statistical downscaling of precipitations (e.g. Beckmann and Buishand, 2002; Vrac et al., 2007), to analyze time series (Davis et al., 1998; Mestre and Hallegatte, 2009; Underwood, 2009) and more recently to solve inverse problems (e.g. Brogniez et al., 2013). A reasonable number of papers provide in-depth descriptions of the GAM algorithm, and one can refer to Wood (2006) for a detailed presentation of the background and the implementation issues of such model. We provide here only briefly its main characteristics. A GAM infers the possible nonlinear effect of a set of p predictors $(\mathrm{BT}_1, \ldots, \mathrm{BT}_p)$ to the expectation of the predictant RH^i. It is expressed as followed:

$$g(\mathbf{E}(\widehat{\mathrm{RH}}^i|\mathbf{BT})) = \epsilon^i + f_1(\mathrm{BT}_1) + f_2(\mathrm{BT}_2) + \ldots + f_p(\mathrm{BT}_p), \quad (1)$$

where g is a linearizing link function between the expectation of $\widehat{\mathrm{RH}}^i$ given **BT** and the additive predictors $f_j(\mathrm{BT}_j)$, which are smooth and generally non-parametric functions of the covariates $\mathrm{BT}_1, \ldots, \mathrm{BT}_p$. 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" (Hastie and Tibshirani, 1990). Part of the model-fitting process is to determine the appropriate degree of smoothness, which is done through a penalty term in the model likelihood, controlled by a smoothing parameter λ. λ determines the trade off between

AMTD
7, 8983–9023, 2014

the goodness of fit ($\lambda \to 0$, gives a wiggly function) of the model and its smoothness ($\lambda \to \infty$).

Part of the GAM fitting process is to choose the appropriate degree of smoothness of the regression splines. The smoothing parameter λ is adjusted to minimize the generalized cross validation score (GCV). One can refer to Wood (2004, 2006) for more details on the training algorithm.

3.3 Multi-layer perceptron algorithm

The MLP (Rumelhart et al., 1986) belongs to the family of artificial neural networks (Haykin, 1994). MLPs are attractive candidates thanks to various well known properties. Indeed a MLP is a universal function approximator and thus can represent any arbitrary functions (Bishop, 1995) so they are widely used for the approximation of non-linear transfer functions. Moreover MLP have been shown to be able to deal with noisy data. In our case, defining an architecture consists of: (i) selecting the relevant input variables and (ii) setting the number of neurons in the hidden layer. A fixed architecture defines a function family $F(\cdot)$ in which we seek the best function allowing us to invert BTs. It is possible to express this MLP model in a mathematical way as:

$$\widehat{\mathrm{RH}}^{i} = F(\mathbf{W}, \mathbf{BT}), \tag{2}$$

where $F(\cdot)$ and $\mathbf{W}$ correspond respectively to the transfer function and the synaptic weights matrix of the model. The main critical point with the MLP method is the way to choose the optimal architecture and to adjust the corresponding MLP internal parameters (the weights). These parameters are determined so as to minimize the mean quadratic error computed on the training dataset. As our goal is to create a nonlinear model with good generalization capabilities, the problem of overfitting must be considered. To avoid overfitting, the LOO validation method is implemented to check the possible overfitting and to optimally select model parameters such as to minimize the validation error.

AMTD
7, 8983–9023, 2014

3.4 Least Square – Support Vector Machine

SVM are kernel methods (Scholkopf and Smola, 2002). They are attractive candidates for nonlinear modeling from data. Thanks to various desirable properties, they have the ability to build models with high generalization capabilities by avoiding overfitting and controlling model complexity. A Least Square formulation of SVM called LS-SVM were proposed to make the SVM approach for modeling more generally applicable, such as for dynamic modeling (Qu, 2009) or for implementing sophisticated validation techniques (Cawley and Talbot, 2007). The SVM technique and its derived formulations have found applications in atmospheric sciences, such as in statistical downscaling of precipitation (Tripathi et al., 2006; Anandhi et al., 2008), in regression problems (Sun et al., 2005) or in classification from remote sensing measurements (Lee et al., 2004).

The LS-SVM training procedure consists in estimating the set of adjustable parameters w and b by the minimization of the cost function:

$$J(w,e) = \frac{1}{2} w^T w + \frac{1}{2} C \sum_{k=1}^{N} e_k^2 \tag{3}$$

with e_k the prediction error for example k and N the size of the training set. C is an hyperparameter that controls the tradeoff between the prediction error and the regularization. This 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. Thus, the parameters can be computed by resolving a set of $(N+1)$ linear equations.

Since LS-SVM models are linear in their parameters models, the solution of the training phase is unique and can be computed straightforwardly using the set of $(N+1)$ linear equations as stated above. Here the validation error is estimated using the virtual LOO (or VLOO) method. This method, first proposed for linear models (Belsley et al., 1980) and later extended to nonlinear models (Laurent and Cook, 1993), allows to estimate the validation error by performing one training involving the whole available data. This estimation is exact when dealing with linear-in-their-parameters models, such as

AMTD
7, 8983–9023, 2014

LS-SVM models, while it remains an approximation for models which are nonlinear with respect to their parameters. More recently, a framework described by Cawley and Talbot (2007) implements the VLOO method for LS-SVM models. This method gives a fast and exact estimation of the validation error, which is a great benefit for reducing the computational burden involved by other validation techniques such as the cross validation method (Hastie et al., 2009).

4 Performance over the synthetic dataset

The retrievals of RH profiles provided by GAM, MLP and LS-SVM are compared for the two schemes. The following criteria are computed over the test set (∼ 5000 samples) for each atmospheric layer i: the mean error (refered to as the "bias"), the standard deviation of the error (SD) and the Pearson's correlation coefficient (R) between the estimated $\widehat{\mathrm{RH}}$ and the reference RH, using the variance-covariance matrix (cov):

$$\mathrm{SD}^i = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(\mathrm{RH}_k^i - \widehat{\mathrm{RH}}_k^i)^2} \qquad \mathrm{bias}^i = \sum_{k=1}^{N}(\mathrm{RH}_k^i - \widehat{\mathrm{RH}}_k^i)$$

$$R^i = \frac{\mathrm{cov}(\mathrm{RH}^i, \widehat{\mathrm{RH}}^i)}{\sqrt{\mathrm{SD}^2(\mathrm{RH}^i)\cdot \mathrm{SD}^2(\widehat{\mathrm{RH}}^i)}}. \tag{4}$$

Moreover the notation %RH will be used to make easier the discussion between relative units (in %) and RH units (in %RH). The size of the test set allows to consider that all the results are significant at the 99.9 % level of confidence.

4.1 On the optimization of the models

As mentioned in Sect. 2.2.3, it is important to underline that the optimized models are different for each atmospheric layer. Indeed, in the case of the SAPHIR-MADRAS

AMTD
7, 8983–9023, 2014

scheme a selection of the relevant channels is performed using the GSO procedure. The GSO procedure helps to reduce the complexity of the algorithms by reducing the number of inputs of the available set of data. It is implemented in the present case with a reasonable threshold of 10 % on the variation of the variance. This means that the inputs that enhance the error variance less than 10 % are considered as irrelevant. Of course the same inputs could be used for the different models with small deterioration. For example, for the layer 4 (425–650 hPa), a sensitivity analysis has shown that, when using GAM, the best set of inputs is {S3, S4, S5, S6, M3, M4} and if M9 is added, the SD decreases from 4 to 3.8 %RH. In fact, when the RH retrieval is based on the MLP approach, the SD increases from 2.8 to 3 %RH. In these two cases the difference is relatively small. In fact, an in-depth study of the relevancy of the channels reveals that the selected inputs are only weakly dependent on the retrieval model but are highly dependent on the atmospheric layer.

For the SAPHIR-only scheme, all channels are used. An impact study of the pre-processing of the data on the accuracy of RH retrieval shows that, whatever the atmospheric layer or algorithm considered, the improvement obtained with PCA is negligible (< 3 % of the error variance). The use of uncorrelated inputs is thus not necessarily required for the considered models. Finally, the linearization of the problem with the exponential function is beneficial only for the MLP: in this case it leads to a decrease of the error variance of about 50 %, while no significant improvement is observed for LS-SVM and GAM (< 3 % of the error variance).

4.2 Performance of GAM against the two other models

From here on, noise-free BTs are considered in order to only assess the statistical approaches. The radiometric noise of the two instruments are implemented for the evaluation of the retrieval of RH with RH values considered as reference values. Vertical profiles of mean biases, SD and R between the observed RH and the estimated RH are presented on Fig. 3. At first sight, the analysis of one layer at a time clearly shows that the overall quality of the retrieval is layer-dependent, meaning that it is strongly

AMTD
7, 8983–9023, 2014

constrained by the physical limits of the inverse problem. Thus, the layers covering the free troposphere (layers 2 to 6) are quite well modeled, with small SD reaching values between 2.6 and 7.8 %RH, and are characterized by a small scatter, with R lying in the 0.85–0.97 interval. The combined use of SAPHIR and MADRAS BTs is enough to explain more than 70 % of the variability of the RH at these layers. The retrieval of the RH of the extreme layers (layer 1 for the top of the atmosphere, layer 7 for the surface) seems more delicate and is clearly limited by the inputs at hand: as illustrated on Fig. 1, the 6 channels of SAPHIR observe the emitted radiation grossly between 150 and 850 hPa, and although MADRAS brings some additional relevant measurements, other information such as the surface emissivity might contribute significantly to better constraint the retrieval near the surface.

The LS-SVM technique provides overall the best results, with the highest correlation coefficients and the lowest variance for 5 layers over the 7 considered in this study. In fact, theoretically, these 3 learning methods are equivalent but the conditions of their implementation are somewhat different. First, since the LS-SVM are linear-in-their-parameters models, an exact validation method was implemented. The resulting procedure of selection of the relevant inputs is quite efficient. In addition, MLP models are nonlinear with respect to the adjusted parameters, and their training amounts to a nonlinear optimization. Several trainings with different initializations must be performed with no guarantee to achieve the best generalization capability given a network architecture. From this point of view, the LS-SVM approach is thus more successful. Finally, concerning the GAM approach, the smoothing splines used guarantee a nonlinear behavior, continuity and smoothness which are important characteristics in a learning algorithm. Another convenient characteristic for splines is that they are monotonic: the backpropagation algorithm can estimate parametric and non-parametric components of the model simultaneously.

The 3 methods perform equivalently: R and SD are very close to each other. The MLP approach provides slightly more biased estimations of the RH throughout the troposphere while the GAM and LS-SVM methods are centered. This distinction is more

AMTD
7, 8983–9023, 2014

pronounced for the surface layer with retrievals of RH characterized with a 6.9 %RH bias when using the MLP whereas the bias is 0.06–0.07 %RH with GAM and LS-SVM. A sample of profiles is presented on Fig. 4, with the observed relative humidity and the 3 estimations using the 3 approaches. As discussed above the top layer is the less well retrieved from the set of BTs, whatever the approach, while the mid-tropospheric layers (3 to 6, i.e. 350 down to 950 hPa) are pretty well estimated.

The errors obtained from the GAM estimation are projected on the 10×10 Kohonen maps that were obtained during the stage of clustering of the atmospheric layers (Sect. 2.2.2) and give a structural view of the errors. The projections are shown on Fig. 5. This allows to analyze the retrieval errors with respect to the clusters of RH revealed by the maps, and allow for a deeper analysis related to meteorological situations than the global biases and SD. A pattern of a large bias (~ 44 %RH) clearly stands out of the map of layer 1 (near tropopause), and this bias is associated to the neurons related to a moist structure at this top layer. This suggests that GAM has difficulties when dealing with a moist upper troposphere, that could be due to an under-representation in the training set. A similar statement can be made for the 6th layer, with the neurons associated to the largest bias in the upper left corner (negative in this case) being this associated to the more dry neurons of this layer. There is no clear pattern standing out of the remaining layers, even for the surface layer, meaning that the errors are uniformly distributed.

4.3 Performance for the two instrumental schemes

In the following, noisy BTs are used in order to discuss the results over the realistic instrumental configurations. Two GAMs are optimized for each atmospheric layer, one for each instrumental scheme: a SAPHIR-MADRAS scheme and a SAPHIR-only scheme. The evaluations over the validation set are summarized on Table 2, with biases, SD and R. An illustration of the scatter is given with Fig. 6 for two atmospheric layers: layer 4 ($\sim$ 425–650 hPa) and layer 6 ($\sim$ 900–955 hPa). These statistics allow for a discussion on the influence of MADRAS BTs on the quality of retrieval of the RH.

AMTD
7, 8983–9023, 2014

MADRAS channels are an asset for the estimation of the RH profile since their use reduce the scatter (improvement of R and reduction of SD). The pattern of scatter follows the distribution of the weighting functions of SAPHIR: the best estimations are obtained for the mid-tropospheric layers (R = 0.83 to 0.97, over layers 2 to 5) where the fonctions strongly overlap, and the quality of the estimations decrease towards the edges. One can also note that the retrieval model of the 7th layer use all 15 BTs of the microwave payload but this does not allow for a robust estimation of the RH (R = 0.54, corresponding to a R^2 value of 0.29). An estimation of the RH profile down to 955 hPa seems reasonable if no other constraint is added to the model.

When the SAPHIR-only scheme is used, such statement can be extended to the top layer (R = 0.57, R^2 = 0.32), thus limiting the estimation of RH from layer 2 to layer 6. For these atmospheric layers, the biases are small and range between 2.69 and −1.53 %RH. The impact of MADRAS BTs on the retrieval of RH is important to keep in mind when specific analysis of temporal and spatial variations of the RH field will be performed over the MT lifetime.

5 Application to Megha-Tropiques measurements

5.1 Some considerations on the Megha-tropiques observations

As other similar radiometer with varying viewing geometry, SAPHIR observations are subject to the so-called “limb effect”, described for instance in Goldberg et al. (2000). This means that, at SAPHIR frequencies, the pixels on the edge of the swath have BTs artificially lower than the pixels located in the center, the atmosphere of the former having a larger optical depth than the latter. For the same thermodynamical profile, this limb effect yields to shift upward the sounding altitude of the outermost pixels. Of course this needs to be taken into account in any retrieval processes (e.g. Karbou et al., 2005; Buehler et al., 2004). Possibilities are (i) to have one dedicated model per viewing angle (as done by Buehler and John, 2005), (ii) to include explicitely the viewing

AMTD
7, 8983–9023, 2014

angle in the retrieval method traditionnally done in iterative schemes (see Soden and Bretherton, 1993; Liu and Weng, 2005), or (iii) to apply a correction that brings all the viewing angles to an equivalent nadir position, before the retrieval itself (Brogniez and Pierrehumbert, 2006). Here, the GAMs have been optimized using the nominal viewing angle of MADRAS (53.5°) and limited to the nadir geometry of SAPHIR. In fact, the observed relationship between the BT and the viewing angle can be accurately approximated by a multi-variate linear function, as noticed by Goldberg et al. (2000) and Buehler et al. (2004). Knowing the means and variances of this relationship for each angle is enough to assimilate this function in the normalization method which is based on standard scores. These have been computed every 2° from nadir to 52° (the maximum viewing angle of SAPHIR is 50.7°) using the training database.

5.2 Comparison to radiosoundings measurements: the CINDY/DYNAMO/AMIE dataset

Observed RH profiles gathered from the CINDY/DYNAMO/AMIE international field experiment are used to evaluate the estimated RH profiles. With the 1st orbit of Megha-Tropiques executed on 13 October 2011, this large scale campaign is ideal to perform such exercise. It took place over the period October 2011–March 2012 in the Indian Ocean and was dedicated to better understand the processes involved in the initiation of the Madden-Julian Oscillation and to improve its simulation and prediction (Cooperative Indian Ocean Experiment on Intraseasonal Variability in the Year 2011/Dynamics of the Madden–Julian Oscillation/ARM Madden-Julian Oscillation Investigation Experiment, hereafter C/D/A). Measurements related to the atmospheric and oceanic states have been collected from radars, microphysics probes, a mooring network and an upper air sounding network. One can refer to Clain et al. (2014) for a discussion on the quality of the RH profiles and their use in the context of the evaluation of SAPHIR measurements. Here we focus on the oceanic sites and on the period October to December 2011 to evaluate the RH estimations: over that period, MADRAS performed optimally. Clain et al. (2014) found a systematic bias in the BT space that increases

AMTD
7, 8983–9023, 2014

with the distance of the observing channel from the central frequency. Such biases are eliminated by the normalization procedure of the retrieval scheme. Overall, among the 10 000 high-resolution soundings collected during the campaign (Ciesielski et al., 2014), only about 50 profiles match to our collocation criteria: a $\Delta t \leq \pm 45$ min and a $\Delta x \leq 50$ km.

The restriction of the training of the GAMs to clear-sky conditions requires a cloud mask. Therefore, cloud-free cases are detected from the radiosounding record itself (RH limited to 100 %RH) and are associated to the Hong et al. (2005) method to detect the precipitating scene (the convective overshootings, a threshold method based on the depression induced by the scattering of the microwave radiation by the precipitating particles) from the SAPHIR observations. One point of concern here is the availability of the Megha-Tropiques archive over this period which is not 100 %, with a lower availability for MADRAS. The completion of this archive until the date of launch is still a major point of concern for the two space agencies CNES and ISRO, in order to maximize the size of the MADRAS record.

For each of the 7 layers, the observed RH is defined by the mean of the measurements that fit into the pressure boundaries, assuming that this mean will be representative of the layer. This assumption is very simple, especially since the tropospheric RH is characterized by its strong vertical gradients induced by complex transport and thermodynamic processes (e.g. Pierrehumbert et al., 2007; Sherwood et al., 2010). However, a comparison (not shown) between such smooth mean and a discrete mean as defined from the training profiles show no systematic differences. Figure 7 shows the comparison between the observed and estimated RH using profiles of R and biases, for the two instrumental configurations. Figure 7 summarizes the results. Since the sample size is quite small ($N - 2 = 48$ degrees of freedom), a Student t test (Student, 1908) is performed to test the independance of the samples and assuming that they follow Gaussian distributions. The 99.9 % level of confidence is indicated on Fig. 7 and t values below this level are not given. Box-and-whiskers diagrams are used to represent the distributions of the differences and show the similarity/differences of the

AMTD
7, 8983–9023, 2014

estimations when using both SAPHIR and MADRAS or only SAPHIR. As expected from the synthetic data analysis, the mid-tropospheric layers 2 to 5 are very well retrieved, with quite good correlations (0.6–0.92) when SAPHIR and MADRAS are combined. Additional analysis show that the SD of the differences reach a maximum of 10 %RH (layer 4). The removal of MADRAS clearly affect the estimation of RH for most layers, while for layer 3 there is no significant effect. This is expected from the distribution of the weighting functions that present a large overlap around 300 hPa (see also Fig. 1).

5.3 Large scale structures

Figure 8 shows an example of RH estimation using the SAPHIR-only scheme, for the 4th atmospheric layer (425–650 hPa) observed on 14 November 2011 (observing time 17:00–18:55 TU). The RH of the ERA-Interim re-analysis dataset of the same date (at 18:00 TU). The large-scale patterns are clearly identical in the two maps, such as the large dry area over West Africa, the moist and thin filamentary structure North-West of it, or the moist area over Central America. The amplitude of the two fields present some discrepancies, but is is important to focus specifically on the cloud-free zones. The high and mid-level clouds in ERA-Interim are shaded in black while the low clouds are delimited by the grey contour. Over these areas the amplitudes are similar, with minima of RH around 10 %RH. No note that cloud-mask is yet available for the Megha-Tropiques observations and a current effort is on the use of the cloud mask and classification performed by the SAFNWC (Satellite Application Facility of EUMETSA) for the belt of geostationnary satellites, adjusted to Megha-Tropiques.

6 Conclusions

Microwave observations from the SAPHIR and MADRAS microwave radiometers of the Megha-Tropiques satellite are used to retrieve RH profiles. For this purpose, optimized GAMs were trained for each atmospheric layer over a realistic set of synthetic

AMTD
7, 8983–9023, 2014

observations. This set is composed of 18 years of radiosounding profiles covering the tropical belt (±30°) used in combination with a radiative transfer model (RTTOV9.3) to get the associated synthetic BTs. Our approach consists in using only the satellite measurements as inputs of the retrieval method. The training phase of the model considers implicitly the role of temperature, humidity and surface characteristics of the tropical atmosphere.

To assess the performance of GAM, MLP and LS-SVM were also trained and optimized using sophisticated validation methods. The novelty lies in the implementation of the LS-SVM modeling technique since, to our knowledge, it has never been applied for remote sensing retrievals whereas it solves the major problem of local minima, a common pitfall when using neural networks. While the 3 modeling methods come from different theoritical backgrounds, they achieve roughly the same performance. Moreover the LS-SVM approach provides roughly better results. We assume that these improvements come from their built-in regularization mechanisms, but they are associated to a heavy computational burden that compromises their implementation when considering large datasets (such as satellite measurements).

The intercomparison of the three models points towards the definition of the problem given the inputs at hand. The combination of SAPHIR and MADRAS or the use of SAPHIR-only makes it possible to perform a robust estimation of RH in the 150–950 hPa part of the troposphere with a small error (absolute maximum bias of 1.53 %) and scatter (min correlation of 0.49). Near the tropopause and at the surface, the retrieval capacity is clearly constrained by the information content brought by the inputs, whatever the configuration. Of course, the use of a retrieval technique (e.g. neural network or 1d-variational) using prior physical information should further improve the estimation: for instance, the surface layer should clearly benefit from prior knowledge of the surface temperature and total water vapor content. In fact, a comparison with existing works based on methods combining physical constraints with statistical tools (Kuo et al., 1994; Cabrera-Mercadier and Staelin, 1995; Rieder and Kirchengast, 1999; Liu and Weng, 2005; Aires et al., 2013) applied to on similar radiometers with less

AMTD
7, 8983–9023, 2014

channels in the 183.31 GHz line, such as AMSU-B or MHS, shows that the current approach gives similar performance (root mean square errors of about 10 % in the mid-troposphere). A 1D-var technique exploring SAPHIR data should further improve the estimation of RH.

Despite the focus of the present study on oceanic and cloud-free conditions, this approach has been extended to continental cases, where the influence of the surface emissivity on the measured brightness temperature at the top of the atmosphere needs to be taken into account. For this extension, the problem is constrained with the emissivity atlas of Prigent et al. (2006) as input of the radiative transfer model for the simulation of the BTs. While SAPHIR channels are less affected, in mean, by the surface emissivity, the consideration of realistic emissivity is an asset for the definition of a realistic training set. Comparisons (not shown) to in situ measurements from a continental site in Burkina Faso (a dedicated summer 2012 field campaign) reveals similar performance in the mid-tropospheric layers, the surface layers being slightly better estimated.

Following this work, our current efforts focus on the estimation of the conditional error associated to the retrieval itself. Indeed, because the widths and altitudes of the weighting functions of SAPHIR are strongly dependent on the thermodynamical state of the atmosphere (the drier the atmosphere, the wider the layer; the maximum of sensitivity shifting from the upper troposphere towards the mid-troposphere), it is clearly expected that the robustness of the RH estimation will be conditioned by the state of the atmosphere. The aim will be to provide the probability density function of the relative humidity on given BTs (a given state of the atmosphere) and thus address the issue of non-Gaussian distribution of the relative humidity at a given height. The knowledge of such information fits into the current work done within the Global Energy and Water Cycle Experiment (GEWEX) Water Vapor Assessment (G-VAP: http://www.gewex-vap.org) to better characterize the observational records, together with their uncertainties.

Acknowledgements. The authors thanks the LMD/ABC(t)/ARA group for producing and making available to the community their "ARSA" radiosounding database (available from http://ara.abct.

AMTD
7, 8983–9023, 2014

lmd.polytechnique.fr/index.php?page=arsa). Gaelle Clain is also deeply acknowledged for her work on the CINDY/DYNAMO/AMIE profiles and their collocation with Megha-Tropiques. The help of R. Johnson, P. Ciesielski (CSU) and J. Wang (NCAR) with the CINDY/DYNAMO/AMIE radiosounding dataset was greatly appreciated. The advices of P. Eriksson have been welcomed and helped to improve the manuscript. The different statistical models have been implemented and optimized using either Matlab (the somtoolbox package is freely available from the Laboratory of Computer and Information Science) or the R free software (the gam and AMORE packages are freely available from the Comprehensive R Archive Network – CRAN).

References

Aires, F. and Prigent, C.: A new neural network approach including first guess for retrieval of atmospheric water vapor, cloud liquid water path, surface temperature, and emissivities over land from satellite microwave observations, J. Geophys. Res., 106, 14887–14907, 2001. 8987

Aires, F., Bernardo, F., Brogniez, H., and Prigent, C.: An innovative calibration method for the inversion of satellite observations, J. Appl. Meteor. Climatol., 49, 2458–2473, 2010. 8987

Aires, F., Bernardo, F., and Prigent, C.: Atmospheric water-vapour profiling from passive microwave sounders over ocean and land. Part I: Methodology for the Megha-Tropiques mission, Q. J. Roy. Meteorol. Soc., 139, 852–864, 2013. 9005

Anandhi, A., Srinivas, V., Nanjundiah, R., and Kumar, D.: Downscaling precipitation to river basin in India for IPCC SRES scenarios using support vector machine, Int. J. Climatol., 28, 401–420, 2008. 8996

Balabin, R. and Lomakina, E.: Support vector machine regression (SVR/LS-SVM). An alternative to neural networks (ANN) for analytical chemistry? Comparison of nonlinear methods on near infrared (NIR) spectroscopy data, Analyst, 136, 1703–1712, 2011. 8987

Beckmann, B. and Buishand, T.: Statistical downscaling relationships for precipitations in The Netherlands and North Germany, Int. J. Climatol., 22, 15–32, 2002. 8994

Belsley, D., Kuh, E., and Welsch, R.: Regression Diagnostics, 1980. 8996

Bennartz, R. and Bauer, P.: Sensitivity of microwave radiances at 85-183 GHz to precipitating ice particles, Radio Sci., 38, 8075, doi:10.1029/2002RS002626, 2003. 8991

Bishop, C.: Neural networks for pattern recognition, vol. 1, Clarendon Press, 1995. 8995

AMTD

7, 8983–9023, 2014

Blankenship, C., Al-Khalaf, A., and Wilheit, T.: Retrieval of Water Vapor Profiles Using SSM/T-2 and SSM/I Data, J. Atmos. Sci., 57, 939–955, 2000. 8986
Brogniez, H. and Pierrehumbert, R.: Using microwave observations to assess large-scale control of free tropospheric water vapor in the mid-latitudes, Geophys. Res. Lett., 33, L14801, doi:10.1029/2006GL026240, 2006. 8993, 9002
Brogniez, H., Kirstetter, P.-E., and Eymard, L.: Expected improvements in the atmospheric humidity profile retrieval using the Megha-Tropiques microwave payload, Q. J. Roy. Meteorol. Soc., 139, 842–851, doi:10.1002/qj.1869, 2013. 8986, 8989, 8993, 8994
Buehler, S. and John, V. O.: A simple method to relate microwave radiances to Upper Tropospheric Humidity, J. Geophys. Res., 110, D02110, doi:10.1029/2004JD005111, 2005. 8993, 9001
Buehler, S. A., Kuvatov, M., and John, V. O.: Comparison of microwave satellite humidity data and radiosonde profiles: A case study, J. Geophys. Res., 109, D13103, doi:10.1029/2004JD004605, 2004. 9001, 9002
Cabrera-Mercadier, C. and Staelin, D.: Passive microwave relative humidity retrievals using feedforward neural networks, IEEE Trans. Geosci. Remote Sens., 33, 1324–1328, 1995. 8986, 8987, 9005
Cawley, G. and Talbot, N.: Preventing Over-Fitting during Model Selection via Bayesian Regularisation of the Hyper-Parameters, J. Machine Learn. Res., 8, 841–861, 2007. 8996, 8997
Chen, S., Billings, S., and Luo, W.: Orthogonal least squares methods and their application to non-linear system identification, Int. J. Control, 50, 1873–1896, 1989. 8992
Ciesielski, P., Yu, H., Johnson, R., Yoneyama, K., Katsumata, M., Long, C., Wang, J., Loehrer, S., Young, K., Williams, S., Brown, W., Braun, J., and Van Hove, T.: Quality-controlled upper-air sounding dayaset for DYNAMO/CINDY/AMIE: development and corrections, J. Atmos. Oceanic Technol., 31, 741–764, doi:10.1175/JTECH-D-13-00165.1, 2014. 9003
Clain, G., Brogniez, H., Payne, V. H., John, V. O., and Ming, L.: An assessment of SAPHIR calibration using quality tropical soundings, J. Atmos. Oceanic Technol., in press, 2014. 9002
Davis, J., Eder, B., Nychka, D., and Yang, Q.: Modeling the effects of eteorology on Ozone in Houston using cluster analysis and generalized additive models, Atmos. Environ., 32, 2505–2520, 1998. 8994
Deblonde, G. and English, S.: Evaluation of the FASTEM-2 fast microwave oceanic surface emissivity model, Tech. Proc., 2001. 8991

AMTD
7, 8983–9023, 2014

Durre, I., Vose, R., and Wuertz, D.: Overview of the Integrated Global Radiosonde Archive, J. Climate, 19, 53–68, 2006. 8985

English, S., Guillou, C., Prigent, C., and Jones, D.: Aircraft measurements of water vapour continuum absorption at millimetre wavelengths, Q. J. Roy. Meteorol. Soc., 120, 603–625, 1994. 8988

Gierens, K., Schumann, U., Helten, M., Smit, H., and Marenco, A.: A distribution law for relative humidity in the upper troposphere and lower stratosphere derived from three years of MOZAIC measurements, Ann. Geophys., 17, 1218–1226, doi:10.1007/s00585-999-1218-7, 1999. 8989

Goldberg, M., Crosby, D., and Zhou, L.: The Limb Adjustment of AMSU-A Observations: Methodology and Validation, J. Appl. Meteor., 40, 70–83, 2000. 9001, 9002

Hall, A. and Manabe, S.: Effect of water vapor feedback on internal and anthropogenic variations of the global hydrological cycleø, J. Geophys. Res., 105, 6935–6944, 2000. 8985

Hastie, T. and Tibshirani, R.: Generalized Additive Models, Chapman & Hall/CRC, 1990. 8986, 8994

Hastie, T., Tibshirani, R., and Friedman, J.: The elements of statistical learning, Springer, 2009. 8993, 8997

Haykin, S.: Neural Networks: A Comprehensive Foundation, IEEE Press, New York, NY, USA, 1994. 8995

Held, I. and Soden, B.: Water vapour feedback and global warming, Annu. Rev. Energy Environn., 25, 441–475, 2000. 8984

Held, I. and Soden, B.: Robust responses of the hydrological cycle to global warming, J. Climate, 19, 3354–3360, 2006. 8985

Hong, G., Heygster, G., Miao, J., and Kunzi, K.: Detection of tropical deep convective clouds from AMSU-B water vapor channels measurements, J. Geophys. Res., 110, D05205, doi:10.1029/2004JD004949, 2005. 9003

Houze, R. and Betts, A.: Convection in GATE, Rev. Geophys., 19, 541–576, 1981. 8985

Karbou, F., Aires, F., Prigent, C., and Eymard, L.: Potential of Advanced Microwaves Sounding Unit-A (AMSU-A) and AMSU-B measurements for atmospheric temperature and humidity profiling over land, J. Geophys. Res., 110, D07109, doi:10.1029/2004JD005318, 2005. 8987, 8990, 9001

AMTD
7, 8983–9023, 2014

Karouche, N., Goldstein, C., Rosak, A., Malassingne, C., and Raju, G.: Megha-Tropiques satellite mission: in flight performances results, Geoscience and Remote Sensing Symposium (IGARSS), doi:10.1109/IGARSS.2012.6350420, 2012. 8988, 9014

Kohonen, T.: Self-organizing formation of topologically correct feature maps, Biological Cybernetics, 46, 59–69, 1982. 8990

Kohonen, T.: Self-Organizing maps, Springer series in Informations Sciences, 3rd Edn., 2001. 8990

Kuo, C., Staelin, D., and Rosenkranz, P.: Statistical iterative scheme for estimating atmospheric relative humidity profiles, IEEE Trans. Geosci. Remote Sens., 32, 254–260, 1994. 8986, 9005

Laurent, R. T. ST., and Cook, R.: Leverage, local influence and curvature in nonlinear regression, Biometrika, 80, 99–106, 1993. 8996

Lee, Y., Wahba, G., and Ackerman, S.: Cloud classification of satellite radiance data by Multicategory Support Vectore Machines, J. Atmos. Oceanic Technol., 21, 159–169, 2004. 8996

Liu, Q. and Weng, F.: One-dimensional variational retrieval algorithm of temperature, water vapor and cloud water profiles from Advanced Microwave Sounding Unit (AMSU), IEEE Trans. Geosci. Remote Sens., 43, 1087–1095, 2005. 8986, 9002, 9005

Mallet, C., Moreau, E., Casagrande, L., and Klapisz, C.: Determination of integrated cloud liquid water path and total precipitable water from SSM/I data using a neural network algorithm, Int. J. Remote Sens., 23, 661–674, 1993. 8987

Matricardi, M., Chevallier, F., Kelly, G., and Thépaut, J.: An improved general fast radiative transfer model for the assimilation of radiance observations, Am. Meteorol. Soc., 130, 153–173, 2004. 8991

Mestre, O. and Hallegatte, S.: Predictors of tropical cyclone numbers and extreme hurricane intensities over the North Atlantic using Generalized Additive and Linear Models, J. Climate, 22, 633–648, 2009. 8994

Nash, J., Smout, R., Oakley, T., Pathack, B., and Kurnosenko, S.: WMO intercomparison of high quality radiosonde systems: Final report, WMO Report, 118 pp., 2005. 8985

Pierrehumbert, R. and Roca, R.: Evidence for control of Atlantic subtropical humidity by large-scale advection, Geophys. Res. Lett., 25, 4537–4540, 1998. 8985

Pierrehumbert, R., Brogniez, H., and Roca, R.: On the relative humidity of the Earth's atmosphere, in: The Global Circulation of the Atmosphere, Princeton University Press, 143–185, 2007. 8985, 9003

AMTD
7, 8983–9023, 2014

Prigent, C., Aires, F., and Rossow, W.: Land surface microwave emissivities over the globe for a decade, Bull. Am. Meteorol. Soc., 87, 1572–1584, 2006. 8991, 9006
Qu, Z.: Cooperative Control of Dynamical Systems, 2009. 8996
Read, W. G., Waters, J. W., Wu, D. L., Stone, E., Shippony, Z., Smedley, A. C., C.C.Smallcomb, Oltmans, S., Kley, D., Smit, H. G. J., Mergenthaler, J., and Karki, M.: UARS Microwave Limb Sounder upper tropospheric humidity measurement: Method and validation, J. Geophys. Res., 106, 32207–32258, 2001. 8989
Read, W. G., Lambert, A., Bacmeister, J., Cofield, R. E., Christensen, L. E., Cuddy, D. T., Daffer, W. H., Drouin, B. J., Fetzer, E., Froidevaux, L., Fuller, R., Herman, R., Jarnot, R. F., Jiang, J. H., Jiang, Y. B., Kelly, K., Knosp, B. W., Kovalenko, L. J., Livesey, N. J., Liu, H.-C., Manney, G. L., Pickett, H. M., Pumphrey, H. C., Rosenlof, K. H., Sabounchi, X., Santee, M. L., Schwartz, M. J., Snyder, W. V., Stek, P. C., Su, H., Takacs, L. L., Thurstans, R. P., Vömel, H., Wagner, P. A., Waters, J. W., Webster, C. R., Weinstock, E. M., and Wu, D. L.: Aura Microwave Limb Sounder upper tropospheric and lower stratospheric H2O and relative humidity with respect to ice validation, J. Geophys. Res., 112, D24S35, doi:10.1029/2007JD008752, 2007. 8989
Rieder, M. and Kirchengast, G.: Physical-statistical retrieval of water vapor profiles using SSM/T-2 sounder data, Geophys. Res. Lett., 26, 1397–1400, 1999. 8986, 9005
Roca, R., Bergès, J.-C., Brogniez, H., Capderou, M., Chambon, P., Chomette, O., Cloché, S., Fiolleau, T., Jobard, I., Lémond, J., Ly, M., Picon, L., Raberanto, P., Szantai, A., and Viollier, M.: On the water and energy cycles in the Tropics, C. R. Geoscience, 342, 390–402, 2010. 8984
Rumelhart, D., Hinton, G., and Williams, R.: Learning internal representation by error propagation. in: Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1, D. E. Rumelhart and J. L. McClelland, 1986. 8987, 8995
Schaerer, G. and Wilheit, T. T.: A passive microwave technique for profiling of atmospheric water vapor, Radio Sci., 14, 371–375, 1979. 8988
Scholkopf, B. and Smola, A.: Learning with Kernels, MIT Press, 2002. 8996
Sherwood, S., Roca, R., Weckwerth, T., and Andronova, N.: Tropospheric water vapor, convection and climate, Rev. Geophys., 48, RG2001, doi:10.1029/2009RG000301, 2010. 8985, 9003

AMTD
7, 8983–9023, 2014

Soden, B. and Bretherton, F.: Upper Tropospheric Relative Humidity from the GOES 6.7 Channel: Method and Climatology for July 1987, J. Geophys. Res., 98, 16669–16688, 1993. 8992, 9002
Spencer, R. and Braswell, W.: How dry is the tropical free troposphere? Implications for a global warming theory, Bull. Am. Meteor. Soc., 78, 1097–1106, 1997. 8985, 8993
Stephens, G., Jackson, D., and Wittmeyer, I.: Global observations of Upper-Tropospheric water vapor derived from TOVSQ radiance data, J. Climate, 9, 305–326, 1996. 8992
Student: The probable error of a mean, Biometrika, 6, 1–25, 1908. 9003
Sun, B.-Y., Huang, D.-S., and Fang, H.-T.: Lidar signal denoising using Least-Squares Support Vector Machine, IEEE Signal Processing Lett., 12, 101–104, 2005. 8996
Suykens, J. A., Gestel, T. V., de Brabanter, J., de Moor, B., and Vandewalle, J.: Least Squares Support Vector Machines, World Scientific, 2002. 8987
Thiria, S., Mejia, C., Badran, F., and Crepon, M.: A Neural Network Approach for Modelling Non Linear Transfer Functions: Application for Wind Retrieval from Spaceborne Scatterometer Data, J. Geophys. Res., 98, 22827–22841, 1993. 8987
Tripathi, S., Srinivas, V., and Nanjundiah, R.: Downscaling of precipitation for climate change scenarios: a support vectore machine approach, J. Hydrol., 330, 621–640, 2006. 8996
Ulaby, F., Moore, R., and Fung, A.: Microwave Remote Sensing Active and Passing Vol. 1: Microwave Remote Sensing Fundamentals and Radiometry, vol. 1, Addison-Wesley, 1981. 8991
Underwood, F.: Describing long-term trends in precipitation using generalized additive models, J. Hydrol., 364, 285–297, 2009. 8994
Vrac, M., Marbaix, P., Paillard, D., and Naveau, P.: Non-linear statistical downscaling of present and LGM precipitation and temperatures over Europe, Clim. Past, 3, 669–682, doi:10.5194/cp-3-669-2007, 2007. 8994
Wang, J. and Zhang, L.: Systematic Errors in Global Radiosonde Precipitable Water Data from Comparisons with Ground-Based GPS Measurements, J. Climate, 21, 2218–2238, 2008. 8985
Wood, S.: Stable and efficient multiple smoothing parameter estimation for generalized additive models, J. Am. Statist. Assoc., 99, 673–686, 2004. 8995
Wood, S.: Generalized Additive Models, an Introduction with R, Chapman & Hall/CRC, 2006. 8994, 8995

AMTD
7, 8983–9023, 2014

Wun-Hua, C., Jen-Ying, S., and Soushan, W.: Comparison of support-vector machines and back propagation neural networks in forecasting the six major Asian stock, Inderscience Enterprises Ltd, 1, 49–67, 2006. 8987

AMTD
7, 8983–9023, 2014

Table 1. Observational characteristics of SAPHIR and MADRAS. θ_{zen} is the viewing zenith angle. H and V correspond respectively to the horizontal and vertical polarizations of the observed electromagnetic field. f_0 corresponds to SAPHIR central frequency at 183.31 GHz. The instrumental noises obtained from in-orbit performance are also indicated (from Karouche et al., 2012).

instrument	central frequency		bandwidth	resolution (km)	In-Orbit Instrumental
	(name)	(GHz)	(MHz)	(along × across track)	Noise (NEΔT (K))
SAPHIR cross-track ($\theta_{zen} = \pm 50.7°$)	S1	$f_0 \pm 0.2$	±200	from (10 × 10) km^2 at nadir to (14.5 × 22.7) km^2 on the edge of the swath	1.44
	S2	$f_0 \pm 1.1$	±350		1.05
	S3	$f_0 \pm 2.8$	±500		0.91
	S4	$f_0 \pm 4.2$	±700		0.77
	S5	$f_0 \pm 6.6$	±1200		0.63
	S6	$f_0 \pm 11.0$	±2000		0.54
MADRAS conical scan ($\theta_{zen} = 53.5°$)	M1 & M2	18.7 (H & V)	±100	(67.25 × 40) km^2 (16.81 × 10) km^2 (10.1 × 6) km^2	0.48 & 0.56
	M3	23.8 (V)	±200		0.49
	M4 & M5	36.5 (H & V)	±500		0.40 & 0.40
	M6 & M7	89.0 (H & V)	±1350		0.55 & 0.53
	M8 & M9	157.0 (H & V)	±1350		1.59 & 1.49

AMTD
7, 8983–9023, 2014

Table 2. Mean bias (in %RH), standard deviation (SD, in %RH) and correlation coefficient (R) for the 7 layers, and defined between the observed RH and the estimated RH. The estimated RH is obtained using the GAM approach from the two configurations: SAPHIR-MADRAS joint measurements and SAPHIR-only measurements. For the SAPHIR-MADRAS configuration, the relevant channels selected using the GSO procedure are listed using the labels S_i and M_j indicated in Table 1.

Layer #	Scores	SAPHIR & MADRAS relevant channels		SAPHIR all channels
# 1 (85–100 hPa)	bias (%)	S1, S2, S3, S5	1.98	2.36
	SD (%)	M1, M2, M3, M4, M5,	8.92	9.91
	R	M6, M7, M8, M9	0.67	0.57
# 2 (130–250 hPa)	bias (%)	S1, S2, S3	−0.01	−0.09
	SD (%)	M1, M2, M3, M4	5.96	6.02
	R	M5, M6, M7	0.92	0.91
# 3 (275–380 hPa)	bias (%)	S1, S2, S3, S4, S5, S6	0.48	0.48
	SD (%)	M1, M2, M3, M5, M6	3.67	3.79
	R		0.95	0.94
# 4 (4250–650 hPa)	bias (%)	S1, S3, S4	0.45	0.08
	SD (%)	M1, M3, M5, M7	3.56	4.72
	R		0.97	0.95
# 5 (725–850 hPa)	bias (%)	S1, S3, S5, S6	0.95	2.69
	SD (%)	M3, M4, M7, M9	8.55	11.68
	R		0.91	0.83
# 6 (900–955 hPa)	bias (%)	S1, S3, S4, S5, S6	0.11	−1.53
	SD (%)	M1, M3, M4, M5, M6,	6.72	11.65
	R	M7, M8, M9	0.91	0.70
# 7 (1013 hPa)	bias (%)	S1, S2, S3, S4, S5, S6	0.36	-0.02
	SD (%)	M1, M2, M3, M4, M5,	8.69	9.67
	R	M6, M7, M8, M9	0.54	0.34

AMTD
7, 8983–9023, 2014

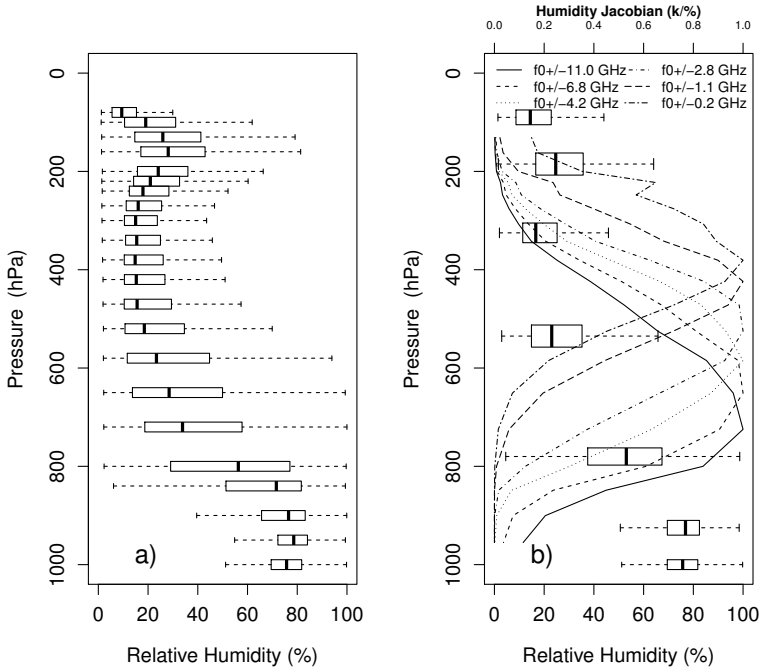

Figure 1. Relative humidity profiles from the database at the initial 22-level resolution **(a)** and at the reduced 7-layer resolution after clustering **(b)**. For each layer, the box-and-whiskers diagram indicates the median (the central vertical line), and the lower and upper quartiles (left and right edges of the box). The whiskers indicate the lower and upper limits of the distribution within 1.5 times the interquartile range from the lower and upper quartiles, respectively. The 6 weighting functions of SAPHIR, computed for a standard tropical profile observed at nadir viewing angle using the RTTOV radiative code, are also represented.

AMTD
7, 8983–9023, 2014

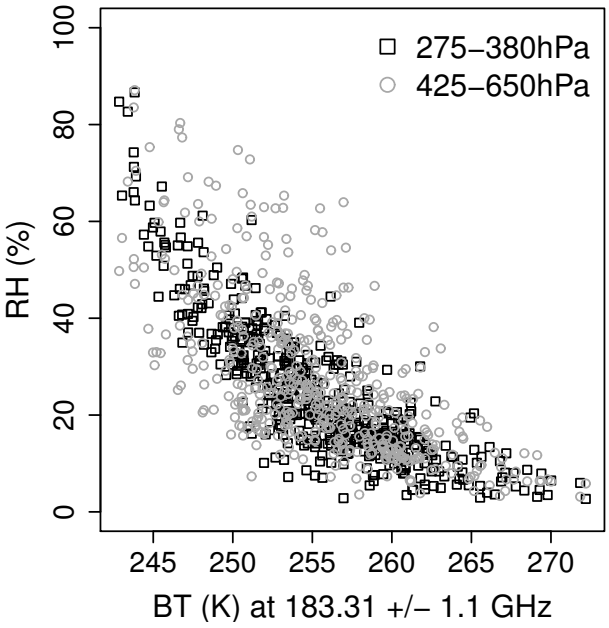

Figure 2. Relationship between the RH of two atmospheric layers (in %RH) and the associated BT (in K) at 183.31 ± 1.1 GHz (2nd channel of SAPHIR). The RH of the 275–380 hPa layer is represented with black squares, while the gray circles are for the RH of the 425–650 hPa layer.

AMTD
7, 8983–9023, 2014

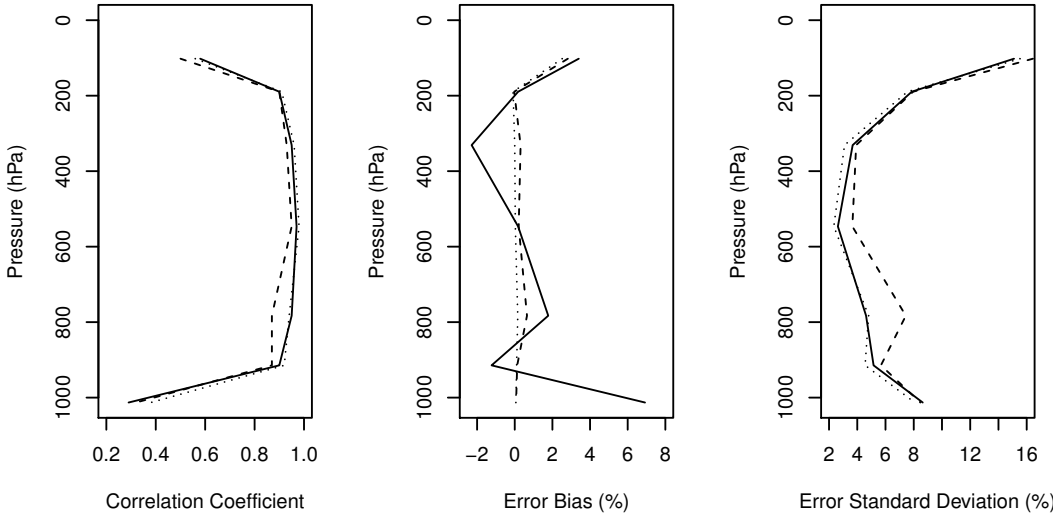

Figure 3. Vertical profiles of *R* (left), biases (center, in %RH) and SD (right, in %RH) for the MLP (solid line), the GAM (dashed line) and the LS-SVM (dotted line) models, trained on noise-free SAPHIR and MADRAS data.

AMTD
7, 8983–9023, 2014

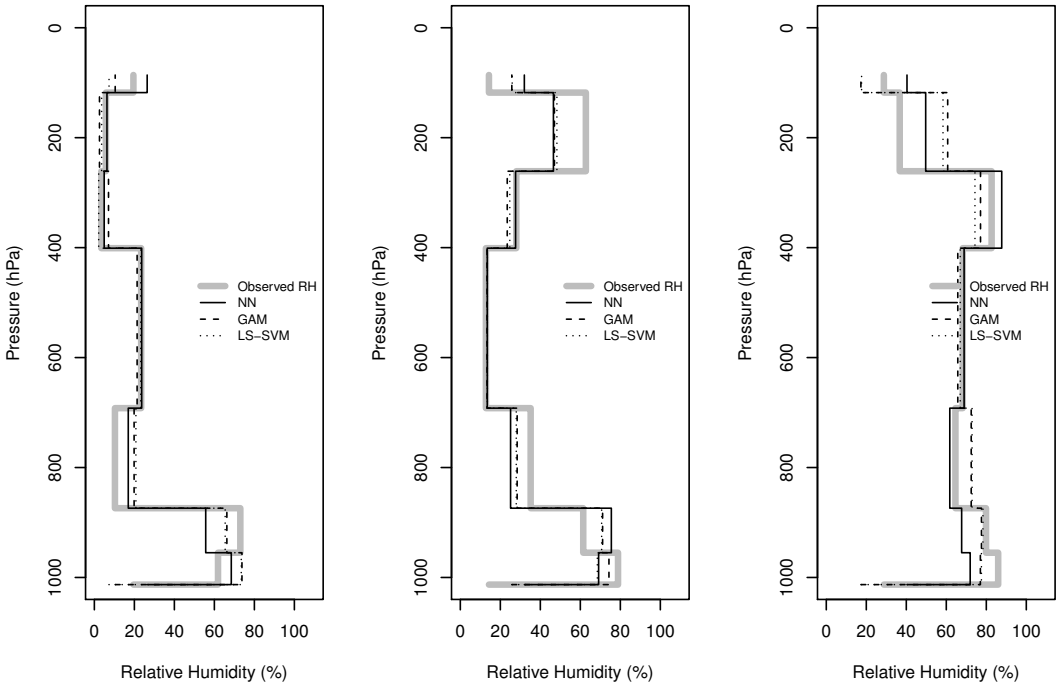

Figure 4. Examples of 3 estimations of RH profiles (in %RH) extracted from the database using the SAPHIR-MADRAS configuration. The observed profile is the thick gray line and the 3 estimations (plain, dashed, dots respectively for MLP, GAM and LS-SVM) are in black.

AMTD
7, 8983–9023, 2014

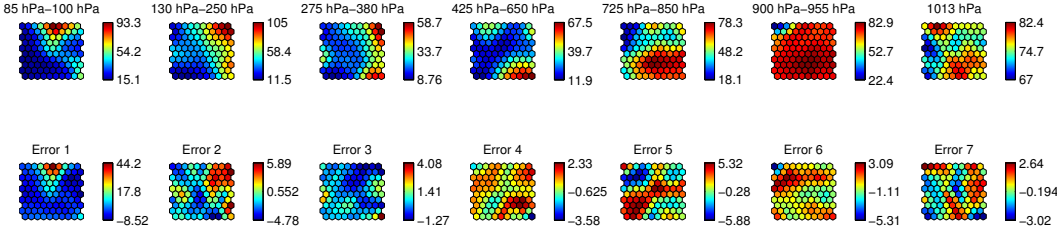

Figure 5. RH and the associated errors (both in %RH) projected on the 10 × 10 self-organizing maps obtained from the step of clustering of the original RH profiles (see Sect. 2.2.2): the upper row shows the mean RH for the 7 layers, and the lower row shows the errors of estimation using GAMs. Note that the color scales are adjusted to each map.

AMTD
7, 8983–9023, 2014

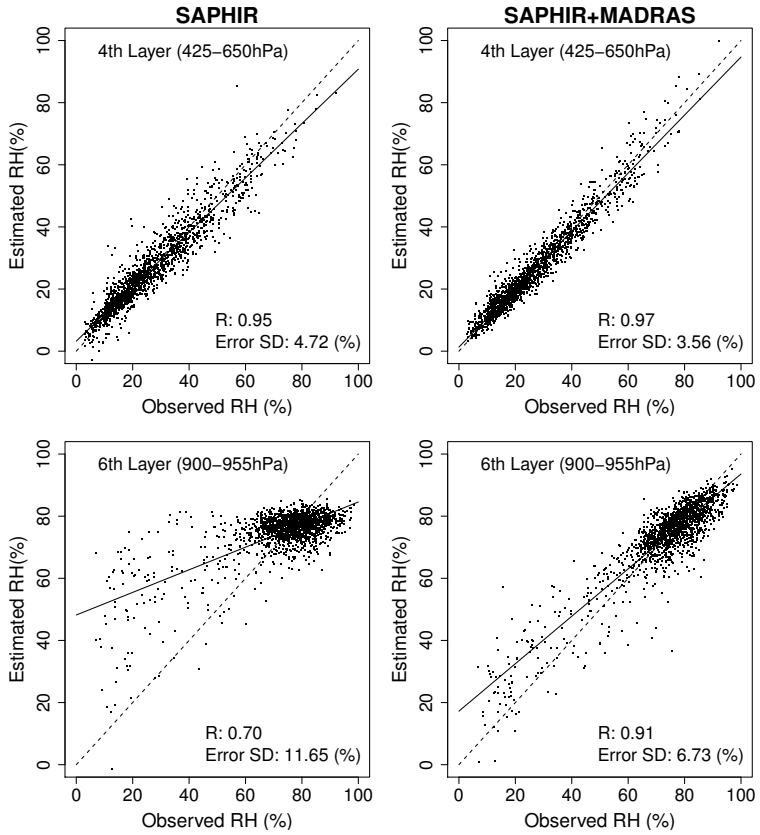

Figure 6. Scatter-plots of the observed RH versus the estimated RH (in %RH) for layer 4 (top row) and layer 6 (bottom row). The estimations are done using GAMs trained from SAPHIR-only BTs (left-hand side column) and from SAPHIR and MADRAS BTs (right-hand side column). The dashed line is $y = x$ line and the solid line represents the linear regression. The correlation coefficient (R) and the standard deviation of the error (SD) are provided within each panel.

AMTD
7, 8983–9023, 2014

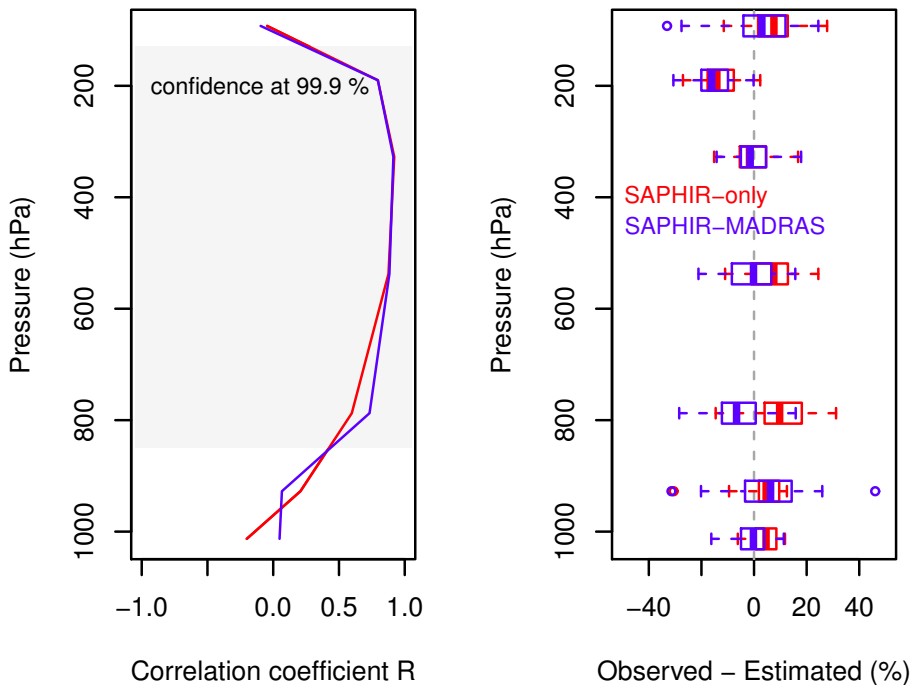

Figure 7. Vertical profiles of *R* (left) and differences (right, in %RH) for the SAPHIR-only (red) and the SAPHIR-MADRAS (blue) retrievals, computed over the subset of 50 RS from the CINDY/DYNAMO/AMIE campaign. For the profiles of the differences between the observed and estimated RH, the box and whiskers diagram indicates for each layer the median (the central vertical line) and the lower and upper quartiles (left and right edges of the box). The whiskers indicate the lower and upper limits of the distribution within 1.5 times the interquartile range from the lower and upper quartiles, respectively.

AMTD
7, 8983–9023, 2014

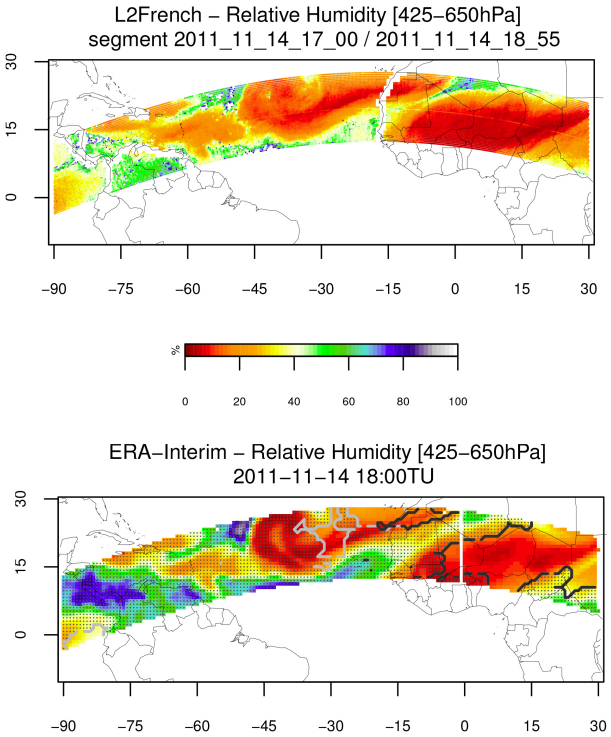

Figure 8. Relative humidity (in %) of the layer 400–600 hPa as estimated from Megha-Tropiques/SAPHIR measurements (top) and by the ERA-Interim reanalysis (bottom) for 14 November 2011. For the map of ERA-Interim RH, the black contour delineates the clear sky and the grey contour delineates the areas with low-level clouds, while the dotted areas are covered with high or mid-level clouds.

AMTD
7, 8983–9023, 2014