This work has been financed by the Italian Ministry of the Economic Development in the framework of the SolarCloud project

21ST INTERNATIONAL TOVS STUDY CONFERENCE, Darmstadt, Germany

Retrieval of Temperature and Water Vapor vertical profile from ATMS Measurements with Random Forests technique

F. Di Paola[1], A. Cersosimo[1,3], D. Cimini[1,2], D. Gallucci[1], S. Gentile[1,2], E. Geraldi[1], S. T. Nilo[1,3], E. Ricciardelli[1], F. Romano[1], M. Viggiano[1]

[1] Institute of Methodologies for Environmental Analysis, National Research Council, 85050 Tito Scalo (PZ), Italy
[2] CETEMPS, Department of Physics, University of L'Aquila, 67100 L'Aquila, Italy
[3] School of Engineering, University of Basilicata, 85100 Potenza, Italy

The ***Advanced Technology Microwave Sounder*** (ATMS) is a cross-track scanning microwave (MW) radiometer currently flying on the *Suomi National Polar-orbiting Partnership* (SNPP) satellite mission, that provides passive observations in the oxygen absorption band at 50÷58 GHz, in the water vapor absorption bands at 22 GHz and 183 GHz, as well as in some window frequencies, useful to retrieve **Temperature** (T) and **Water Vapor** (WV) **atmospheric vertical profiles**. Using spatial and temporal coincidences between ATMS observations and two different datasets of vertical T and WV, a global training dataset for machine learning purpose was built for the whole 2016. For each ATMS Brightness Temperatures (BT) acquisition, 32 levels of T (between 10 and 1000 hPa), and 23 levels of WV (between 200 and 1000 hPa), are used to train an algorithm based on **Random Forests** (RF) regression technique. **A single RF was trained for each level and atmospheric variable**, using the evaluation on the *Out of Bag* (OOB) error to optimize the number of random selection of the input variables at each node splitting step (hereinafter "*Ntry*"), the number of trees in each forest ("*Ntrees*") and the minimum leaf size parameter("*Nleaf*"), to avoid overfitting problem and obtain an accurate retrieval. Considering that the sounding below the precipitation level becomes unreliable, the precipitation-affected observations were removed from the training dataset by means of a pre-screening test based on BT.

Input variables

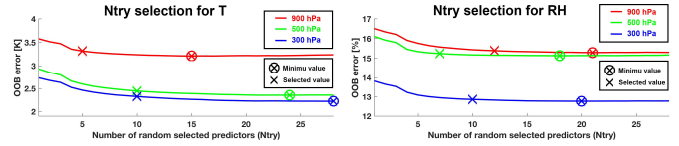
The decision tree built through the *Classification and Regression Trees* (CART) algorithm is able to discard less important variables since only the best predictor is selected at each split in the growing tree process, although the choice of the input for both T and WV at each pressure level would be more suitable from the physics point of view. For this reason, for each RF all the **22 BT** of ATMS are selected as input, jointly with ATMS **scan angle**, **land/sea flag**, **latitude**, **longitude**, **julian day** and **solar zenith angle**.

Tuning parameters

An optimal selection of *Ntry*, *Ntrees* and *Nleaf* parameters is necessary to design the appropriate RF for each pressure level and atmospheric variable. To this end, the OOB error evaluations - unbiased estimates of the true prediction error - are used. For each decision tree in the RF, a different bootstrap sample, generated by random sampling with replacement, was used. This means that, for each tree, about two-thirds of the dataset were used for the growing trees, while one-third remainds *out of the bootstrap aggregation* (Out of Bag). Using the OOB data over the trees in the forest, it was possible to evaluate the error - in terms of the Root Mean Square Error (RMSE) - of the grown trees. Starting from Ntrees=100 and Nleaf=1, the subsequent tree steps were iterated with the results of the previous iteration, until the Ntry, Ntrees and Nleaf values have become stable. Only two iterations were used, because the third iteration did not show any variation.

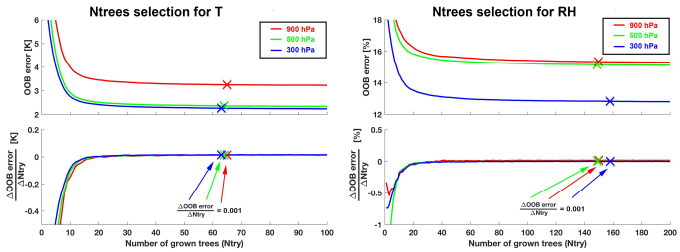
1) Ntry evaluation

Ntry, the number of random subset of predictors selected at each node, ranges from 1 to 28 (total amount of the input variables). Generally, high values of this parameter increase the chance of finding the best splitting predictor so to improve the result, but they also increase the correlation between trees and consequently the probability of overfitting the train dataset. To balance these over- and underfitting risks, the Ntry values, for each level and variable, were set by using the 0.1 arbitrary threshold above the minimum value of the 3-point moving average of the OOB error. This is a conservative choice to preserve the ability to generalize the results, paying at most the error of 0.1K for T and and 0.1% for Relative Humidity (RH).



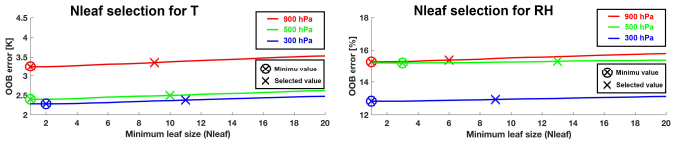
2) Ntrees evaluation

Ntrees is the number of trees in each forest. In general, the more the trees the better the results, without risk of overfitting the dataset. However, this improvement decreases as the number of trees increases, while the computational cost and the amount of stored trees rises. For this reason, Ntrees has been set using the arbitrary threshold of 0.001 for the 5-point moving average of the OOB error gradient, rounded to the higher tenth.



3) Nleaf evaluation

Nleaf, the minimum size of terminal node (*leaf*), sets the depth of the tree implicitly. Small values of this parameter can fit more complex functions, but also increase the risk of overfitting the training dataset. As done for Ntry, also for Nleaf the optimal value was selected considering the 0.1 threshold above the minimum value of the 3-point moving average of OOB error.



Training dataset

Two different datasets were chosen for the RF training:

- The radiosonde observation from the *Integrated Global Radiosonde Archive* (IGRA) that collects measurements from 275 stations distributed worldwide (variable vertical resolution, and 12-hour temporal resolution).
- The global atmospheric reanalysis from the *European Centre for Medium-Range Weather Forecasts* (ECMWF) *ERA-Interim* dataset (37 levels of vertical resolution, 0.75° horizontal resolution, and 6-hour temporal resolution).

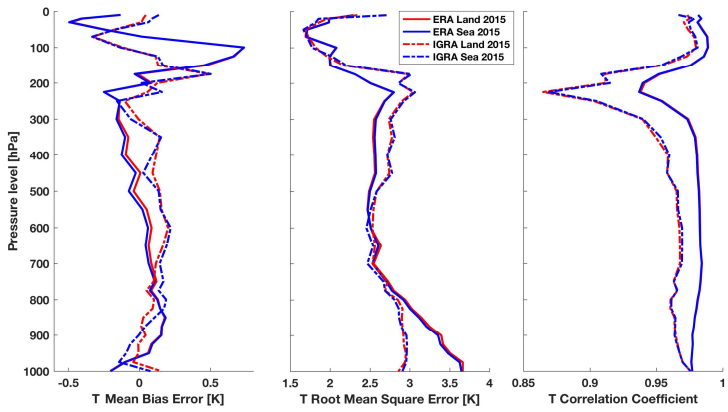
The first dataset offers the strength of a direct measurements, but with the weakness of a limited spatial distribution. On the other hand, the second one offers simulated results on a regular global grid. By using spatial and temporal coincidences between ATMS observations and two datasets for all the 2016, 6M and 12k of vertical profiles were selected for Era-Interim and IGRA, with a delay tolerance of 1 minute and 10 minutes respectively.

In order to build a homogeneous and unique training dataset the following steps were performed:

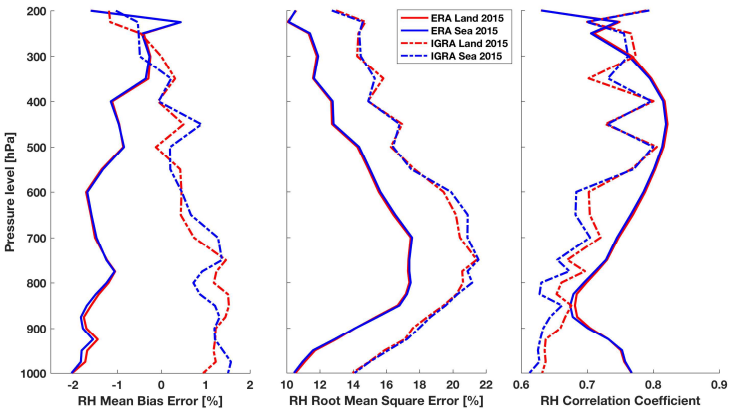
- the IGRA vertical profiles were interpolated linearly on ERA-Interim 32 vertical levels (between 10 and 1000 hPa) for T and 23 vertical levels (between 200 and 1000 hPa) for WV;
- ERA-Interim dataset was calibrated with IGRA measurements to find an additive bias for each pressure level and for 6 latitude intervals (from 90°S to 90°N with a step of 30°), separately for land and sea;
- ERA-Interim dataset was reduced to 12k profiles, using the sampling strategy suggested in Chevallier 2002, so to have the two datasets numerically comparable.

The final training dataset was built by joining the two datasets. It is formed of about 24k profiles, one half from ERA-Interim and the other half from IGRA soundings.

T validation results



RH validation results



As done for the training process, a validation dataset was generated for all the 2015, with about 6M and 12k of vertical profiles for ERA-Interim and IGRA respectively. The validation results for T profiles show that the MBE are generally within ±0.3K below 300hPa and degrade slightly to the highest levels, RMSE is less than 3.6K with the best performance at the highest levels, while CC is greater than 0.95 everywhere, except a small degradation around 200hPa. For RH profiles, MBE is within ±2%, RMSE is less than 22% with the worst performances around 850 hPa while the CC is greater than 0.6 with the best performances around 450 hPa. For both T and WV, there are no evident differences between land and sea. These results show an overall ability of the algorithm to retrieve T and WV vertical profiles in line with expectations.

Breiman L., Random forests, *Machine Learning*, **45** (**1**), 5-32, 2001
Breiman L., Friedman J. H., Olshen, R. A., Stone, C. J. Classification and Regression Trees. Wadsworth, Belmont, 1984
Durre I., Vose R.S., Wuertz D.B.: Overview of the Integrated Global Radiosonde Archive. *J. Climate*, **19**, 53–68, 2006
Chevallier F.: Sampled databases of 60-level atmospheric profiles from the ECMWF analyses, NWP SAF Report No. NWPSAF-EC-TR-004, 2002
Loh W.Y.: Regression Trees with Unbiased Variable Selection and Interaction Detection, *Statistica Sinica*, **12**, 361–386, 2002