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, which 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 database by means of a pre-screening test based on BT.

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.

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). 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) and 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.

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 ±0.3 K below 300 hPa and degrade slightly to the highest levels, RMSE is less than 3.6 K with the best performance at the highest levels, while CC is greater than 0.95 everywhere, except a small degradation around 200 hPa. 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.