Earth, Planets and Space

FULL PAPER Open Access

Comparison of different machine learning approaches for tropospheric profiling based on COSMIC-2 data

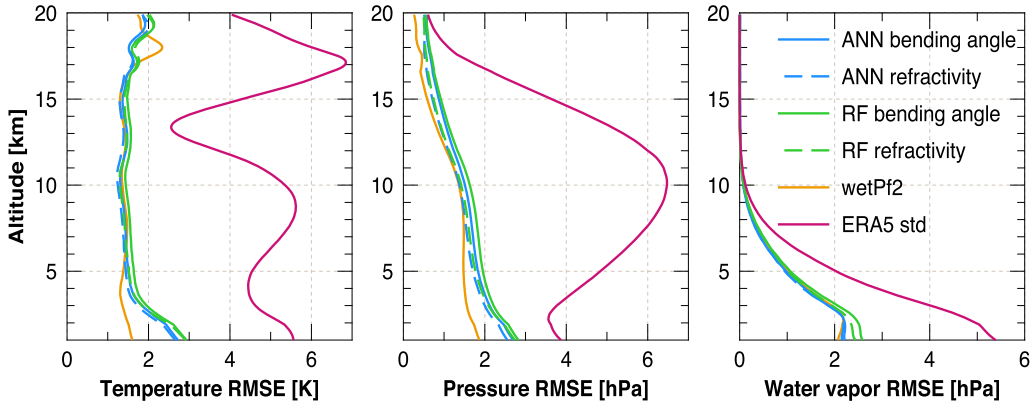
Elżbieta Lasota*

Abstract: Precise and reliable information on the tropospheric temperature and water vapor profiles play a key role in weather and climate studies. Among the sensors supporting the observations of the troposphere, one can distinguish the Global Navigation Satellite System Radio Occultation (RO) technique, which provides accurate and high-quality meteorological profiles. However, external knowledge about temperature is essential to estimate other physical atmospheric parameters. To overcome this constraint, I trained and evaluated four different machine learning models comprising Artificial Neural Network (ANN) and Random Forest regression algorithms, where no auxiliary meteorological data is needed. To develop the models, I employed 150,000 globally distributed (45°S–45°N) RO profiles between October 2019 and December 2020. Input vectors consisted of bending angle or refractivity profiles from the Formosa Satellite-7/Constellation Observing System for Meteorology, Ionosphere, and Climate-2 mission together with the month, hour, and latitude of the RO event. While temperature, pressure, and water vapor profiles derived from the modern ERA5 reanalysis and interpolated to the RO location served as the models' targets. Evaluation on the testing data set revealed a good agreement between all model outputs and ERA5 targets, where slightly better statistics were noted for ANN and refractivity inputs. Vertically averaged root mean square error (RMSE) did not exceed 1.7 K for the temperature and reached around 1.4 hPa and 0.45 hPa for the total and water vapor pressures. Additional validation with 477 co-located radiosonde observations and the operational one-dimensional variational product showed slightly larger discrepancies with the mean RMSE of around 1.9 K, 1.9 hPa, and 0.5 hPa for the temperature, pressure, and water vapor, respectively.

Keywords: Radio occultation, Tropospheric profiling, Machine learning, Bending angle, Refractivity, Random Forest

*Correspondence: elzbieta.lasota@upwr.edu.pl
Wrocław University of Environmental and Life Sciences, Institute of Geodesy and Geoinformatics, Grunwaldzka 53, 50356 Wrocław, Poland

Graphical Abstract



Introduction

The earth's troposphere is a complex, inhomogeneous, and highly variable environment, which is indispensable to life and the main place of human activity. In recent years, studying the earth's troposphere became a major research topic due to climate change and a growing number of severe weather events, such as intense storms and tropical cyclones. Monitoring and prediction of the aforementioned phenomena strongly depend on the understanding of multiple processes including shortwave and longwave radiative transfer, the regional and global water and energy cycles, land surface-atmosphere feedback, and mesoscale circulations, which in turn are affected by the distribution and temporal evolution of temperature and water vapor in the troposphere. Hence, knowledge on the temperature and water vapor profiles plays a crucial role in understanding these earth's system processes and is essential in climate and weather researches but also affects soil and hydrological studies (Wulfmeyer et al. 2015).

Reliable, precise, and accurate monitoring and prediction of the state of the troposphere in terms of temperature and water vapor profiles is a challenging task involving many sensors and techniques, which can be separated into two key groups: in situ measurements from weather stations, meteorological buoys and radiosondes; and remote sensing observations including radar, lidar, airborne and satellite soundings (Wulfmeyer et al. 2011). Among satellite observations dedicated to tropospheric profiling, one can distinguish the Global Navigational Satellite System (GNSS) Radio Occultation (RO) technique, which was first applied to probe earth's troposphere in the GPS/MET experiment (Ware et al. 1996) and nowadays RO serves as a standard source of information about the weather and climate (Kursinski et al. 1997; Anthes 2011). RO technique offers accurate and precise tropospheric profiles with high vertical resolution and global coverage in any weather conditions, which are successfully assimilated into numerical weather prediction (NWP) models greatly improving forecast quality (Huang et al. 2010; Rennie 2010).

RO is an active limb viewing technique, which employs the phase and amplitude of two L-band electromagnetic signals transmitted from GNSS satellites and received on low earth orbit satellites (Kursinski et al. 1997). GNSS signal propagating through the earth's atmosphere is primarily affected by the change of the air density, free electrons in the ionosphere and water vapor, which results in the signal's delay and bending. The latter can be assessed as a function of impact parameter from Doppler shifts frequency based on accurate and precise clocks, satellite orbits and velocity measurements. In the next step, ionosphere-corrected bending angle profiles are transformed to refractivity using Abel transform, where the assumption of local spherical symmetry is applied. Finally, refractivity profiles can be straightforwardly transformed to dry temperature and dry pressure profiles using only the dry term of refractivity equation in the regions where water vapor is negligible (above around 8–14 km altitude) and ideal gas and equilibrium assumptions can be applied (Scherllin-Pirscher et al. 2011). However, special attention must be paid to the lower troposphere, where the dry assumptions are no longer valid due to the presence of abundant water vapor. Hence, ancillary information about temperature, pressure or water vapor pressure is required to calculate the physical tropospheric parameters (Healy and Eyre 2000). Hitherto applied solutions encompass: (1) a direct approach exploiting external pressure and temperature profiles from radiosondes observations or weather models (Ware et al. 1996); (2) iterative methods using RO refractivity and independent temperature profiles (Gorbunov and Sokolovskiy

1993) or surface observations (O'Sullivan et al. 2000); (3) a commonly used one-dimensional variational (1DVar) retrieval method, where RO measurements are combined with background information from a weather model to resolve the one-dimensional tropospheric state in a statistically optimal way (Healy and Eyre 2000; Poli et al. 2002); and eventually, (4) recently proposed simplified linearized 1DVar algorithm, which integrates the direct method with optimal estimation (Li et al. 2019).

The main limitation of the currently used approaches to derive meteorological profiles from RO observations is a demand for meteorological information from external data sources. To overcome this problem, several authors have attempted to employ machine learning (ML) algorithms, the artificial neural network (ANN) in particular, where no independent information is needed. Bonafoni et al. (2009), as one of the first, trained different multilayer perceptron ANN to retrieve the meteorological profiles in the lower troposphere. They used 445 occultations over the land surface covered by vegetation and deserts within the tropics during the summer. GNSS RO refractivity profiles from Formosa Satellite 3/Constellation Observing System for Meteorology, Ionosphere, and Climate (COSMIC) mission served as an input whereas dry and wet refractivity profiles together with the dry pressure from European Centre for Medium-Range Weather Forecast (ECMWF) analysis formed a target during the learning process. They revealed the good performance of proposed algorithms with RMSE of slightly below 2 K for the temperature, 2 hPa and 0.7 hPa for the dry and wet pressures, respectively in the vegetation zones. The promising initial results encouraged Pelliccia et al. (2010) to test the developed methodology on COSMIC observations over tropical oceans. They confirmed the ANN feasibility to obtain high-quality meteorological profiles in the lower troposphere based on RO data. The currently exploited solution was modified in the next work of Pelliccia et al. (2011), who revised the ANN topology but also reduced the influence of the ECMWF model on the training process using RO refractivity weighted by fractional contributions of wet and dry components to the total refractivity in ECMWF model as the targets. Using more than one thousand profiles over the Arctic region in the winter season, they evaluated vertically averaged RMSE as almost 1 K for the temperature, 1.6 hPa for the pressure and 0.04 hPa for the wet pressure. The most recent evidence Shyam et al. (2016) employed almost 5000 samples comprised the month, latitude, bending angle, or refractivity from COSMIC mission to derive water vapor partial pressure profiles using a fully connected three-layer ANN. The results revealed better performance for refractivity as an input to the ANN than for bending angle with water vapor RMSE below 1.5 hPa for the testing data set. However, their analysis was restricted only to the monsoon season over India and the adjoining region.

The main objective of this study is to test and evaluate alternative methods to retrieve tropospheric profiles of temperature, pressure, and water vapor from globally distributed (45°S–45°N) RO measurements based on ANN and random forest (RF) regression, requiring no external meteorological data. Present works have shown the potential and benefits of ANN application in RO-based tropospheric profiling, although no one, to the best of my knowledge, has used RF regression for this purpose. It is widely considered that RF has a few advantages over ANN, which can make it an attractive tool for tropospheric profiling. RF belongs to the group of ensemble ML algorithms, which are characterized by stability, biasedness and less data required to accurately train the model. Furthermore, RF is easier to train due to no need for input preparation, such as scaling or normalization and a lower number of hyperparameters to tune. Very often RF models trained with default configuration bring very good results (Siroky 2009). Furthermore, the number of training examples engaged in developing ANN algorithms was limited to less than 5000 observations, and proposed solutions should be validated by a larger sample size, which may be done by exploiting RO measurements from the newest missions such as the new ones COSMIC-2 launched in 2019.

COSMIC-2 is a follow-on mission to the successful COSMIC mission and has considerable advantages over its precursor (Schreiner et al. 2020). First, the COSMIC-2 mission is designed to also track the signal from the Russian GLONASS, which contributes more observations with over 4000 profiles per day in the tropics and sub-tropics in near real time. Second, each COSMIC-2 satellite is equipped with a high-gain beam-forming RO antenna, which exhibits a higher signal-to-noise ratio (SNR). Higher SNR allows deeper penetration into the lower troposphere (50% of observations reach 200 m above the earth's surface) and reduces the thermal noise impact on bending angle errors. And finally, the received RO profiles present higher precision and accuracy. Derived refractivity profiles are in line with the ECMWF short-term forecasts retrievals and radiosonde observations (RAOBs) below 20 km altitude with minor negative biases reaching up to around 3% and 4%, respectively close to the surface (Schreiner et al. 2020; Ho et al. 2020). A comprehensive comparison between COSMIC-2 and multiple data sets such as RAOBs, dropsondes, and ERA5 reanalysis revealed mean absolute errors and standard

deviations in the troposphere of usually less than 0.5 °C and 1.5 °C for temperature and 1 hPa and 2.5 hPa for water vapor pressure (Chen et al. 2021).

Therefore, this study will be the first attempt in ML-based tropospheric profiling exploiting observations from the new COSMIC-2 mission. The growing number of high-quality observations will be a useful aid for the ML training data set. Furthermore, previous studies have been restricted to the particular seasons and areas such as tropics, Arctic region, or India and the global approach was beyond the scope of those studies.

Within the framework of the above-described criteria, I developed and tested RF and ANN models using 150,000 RO observations between October 2019 and December 2020, covering the land and wet areas. The input of the models consisted of both refractivity and bending angle vertical profiles as well as the month, hour, and latitude of the RO events. The training target comprised of the temperature, pressure, and water vapor partial pressures derived from the state-of-the-art ERA5 atmospheric reanalysis (Hersbach et al. 2020). Then the models' performance was evaluated during the testing phase based on the squared mean differences between ML outputs and ERA5 meteorological profiles, which are assumed to be the truth. However, ERA5, as any reanalysis, is only the best fit of the current state of the atmosphere, and it is affected by errors coming from the different sources such as observations or used assimilation system and using the ERA5 in models training and testing may produce inaccurate results. Therefore, the additional verification with the independent operational 1DVar COSMIC products and nearby RAOBs is performed; however, the number of the available RAOBs is limited to the land area only. Eventually, obtained in near real time with high accuracy and vertical resolution meteorological profiles afterward can serve as an important source of information in the weather and climate studies.

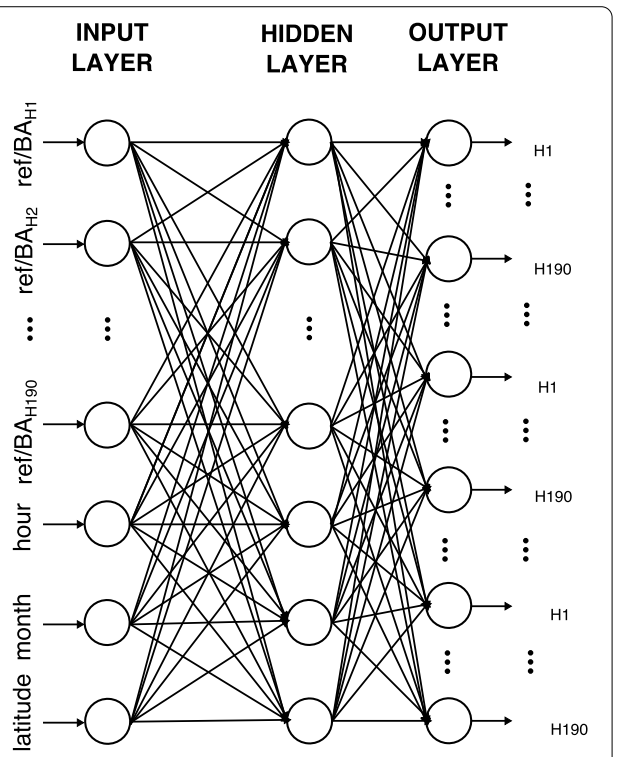


Fig. 1 Example of the MLP structure used in this study. The input consists of refractivity (ref) or bending angle (BA) profiles for 190 altitude levels (H_1–H_{190}) together with latitude, hour and month of RO event. Temperature (*T*), pressure (*P*) and water vapor pressure (*Vp*) profiles for the corresponding altitude levels form the output

Data and methods

In this study, different ML approaches are implemented for tropospheric profiling where no external information about temperature is needed. ML refers to the algorithms, which are capable of improving their performance based on past experience (Michie et al. 1994). For this purpose, I tested ANN and RF models with two separated inputs consisting of bending angle and refractivity, which eventually resulted in trained 4 different models. The whole process involved the following steps: (1) possession of the data, (2) preprocessing, (3) hyperparameters tuning, (4) training and testing the models, and finally, (5) validation with RAOBs.

Artificial Neural Network

ANN is a neurologically inspired ML algorithm that reflects the behavior of the neurons network present in the human brain (Hassoun 1995). ANNs are able to deal with highly non-linear problems and learn directly based on any kind of data. Due to its flexibility and good performance, ANN found application in a wide variety of fields, such as image classification, speech recognition, risk management or weather forecasting. One of the most commonly used ANN topologies is a feed-forward multilayer perceptron (MLP), which was employed in the present study. MLP is a supervised network, which transforms the input data into output based on experience gained during the training on the data set (Gardner and Dorling 1998). The model typically consists of three types of layers: an input, one or more hidden layers, and an output interconnected by multiple fundamental processing units called neurons or nodes. The number of hidden layers and the number of nodes in these layers are arbitrary and depend on the complexity and the amount of available data. An example of MLP model architecture used in this study is presented in Fig. 1. In a fully connected MLP network, each neuron in a certain layer is connected to every neuron in the adjacent layer, whereas the strength

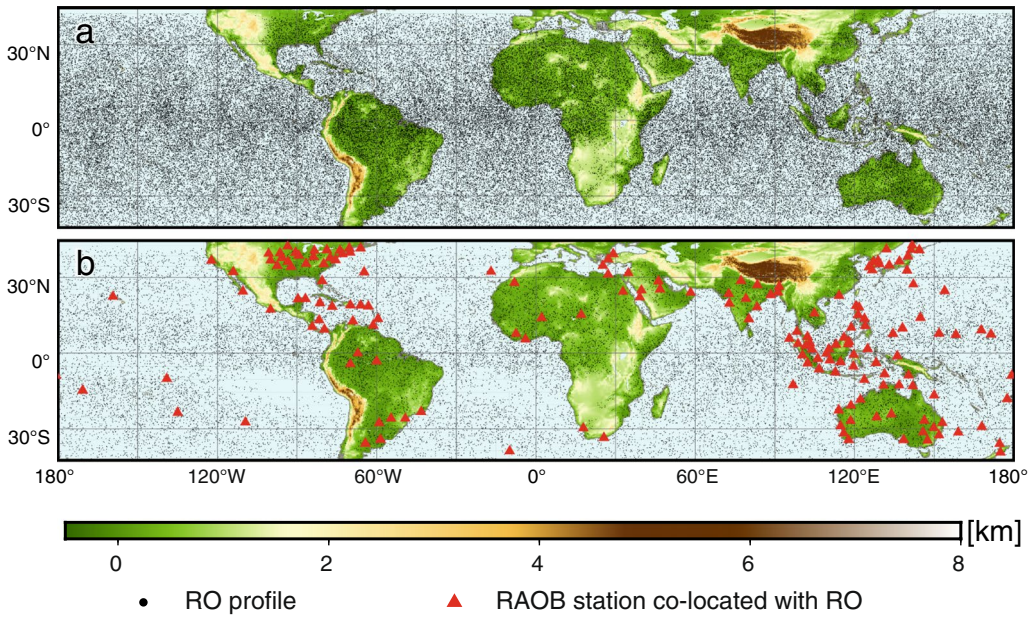


Fig. 2 Distribution of RO profiles employed in training (**a**) and testing (**b**) of machine learning models. Red triangles in (**b**) emphasizes radiosonde stations co-located with RO profiles. Background color shows terrain elevation. Note, the location and number of available RO soundings is restricted by the topography

of the particular connection is expressed by a numerical weight determined during training. Usually learning process is performed iterative using the backpropagation algorithm. The input data are repeatedly fed into the neural network, multiplied by connection weights, summed up and passed to the next layer. Eventually, in the last layer, the model's error is estimated based on the differences between predicted and real outputs. In the next step, the calculated error is fed back and used to adjust the connections weight, which minimizes the model's error and produces the outputs closer to the targets.

Random Forest regression

RF is a statistical nonparametric learning model based on a large ensemble of decision trees. Demonstrated for the first time by Breiman (2001), nowadays RF is one of the most widely and successfully used machine learning algorithms for both classification and regression tasks. However, until now, RF has not been fully explored in GNSS meteorology, especially in RO tropospheric profiling (Łoś et al. 2020). RF belongs to the group of the Bootstrap and Aggregation algorithms, commonly called bagging algorithms. Bagging refers to the random subsampling with the replacement of the original training data set and features to generate multiple base learning models. In RF, decision trees serve as the base models; each tree is constructed independently for the combination of the selected variables and there is no interaction between single trees. The final RF result is calculated as the mean of the outputs of all individual trees. It is proved that the application of RF as the bagging technique contributes to the lower variance and stability and, contrary to the simple decision tree, which is sensitive to the used data set, prevents overfitting (Breiman 1996; Ali et al. 2012).

Step 1: data acquisition

Input: RO bending angle and refractivity profiles

Near-real-time RO profiles from the COSMIC-2 constellation gathered in the latitude band 45°S–45°N for a period between October 1, 2019, and December 31, 2020, were the main products used in this study (Fig. 2). The COSMIC-2 is a follow-on mission of the greatly beneficial COSMIC program led by the Taiwanese National Space Organization, and U.S. National Oceanic and Atmospheric Administration, and other agencies (Schreiner et al. 2020). COSMIC-2 consists of a set of 6 satellites, which were successfully launched into low-inclination orbits on June 25, 2019. The satellites produce more than 4 000 atmospheric soundings a day within $\pm 50°$ of the north and south latitude band providing better insight into weather and climate. RO data at various processing levels is published in near real time by 0200 UTC the following day and freely available at the COSMIC

Data Analysis and Archive Center (CDAAC) website (UCAR COSMIC Program 2019). I focused on atmPrf and wetPf2 level 2 products. The first one contains vertical profiles of bending angle and refractivity, which constitute the inputs for the ML models. The wetPf2 files include meteorological profiles of temperature, pressure and water vapor with 100 m vertical resolution derived through the 1DVar approach with the ECMWF analysis as a background. In the 1DVar approach, a cost function J is minimized to estimate with the maximum likelihood the optimal tropospheric state profile x:

$$J(x) = \left(h(x) - y^0\right)^T (O+F)\left(h(x) - y^0\right) + \left(x - x^b\right)^T B^{-1}\left(x - x^b\right) \quad (1)$$

where y^0 is the observation vector, h is the observation operator, x^b is the background meteorological profile. The matrices B, O and F express the background, observation and observation operator error matrices, respectively (Poli et al. 2002). Then, the so-called wet profiles together with RAOBs were used to assess the accuracy of the ANN/RF retrievals. To balance the required number of observations to train the models and the need to profile the troposphere close to the surface as much as possible, only RO profiles, which reached 1 km altitude and below, were considered in this study.

Target: ERA5 meteorological profiles

The target variable consists of the meteorological profiles of temperature, pressure, and water vapor partial pressure. Specifically, the meteorological data from the ERA5 reanalysis was employed in the training and testing processes. The ERA5 is the most recent global atmospheric reanalysis produced by the ECMWF and replaced the previously used and very popular ERA-Interim reanalysis (Hersbach et al. 2020). The major developments encompass higher 1 h time resolution and 0.25° (31 km) spatial resolution, improvements in Data Assimilation (DA) system and rapid 5 day preliminary availability. The DA system implemented in ERA5 is based on a hybrid incremental 4DVar with 12 h windows and assimilates more than 200 types of conventional meteorological data and observations provided by satellites. Meteorological data are available for 37 pressure levels with a resolution of 25 hPa between 1000 and 750 hPa, 50 hPa below 250 hPa layer, and 16 irregularly spaced levels above with the top-level at 1 hPa. However, the ERA5 does not provide water vapor partial pressure V_p, which constitutes part of the ML output. Instead, it must be calculated from pressure P and specific humidity q (Wallace and Hobbs 2006):

$$V_p = \frac{q \cdot P}{\frac{M_w}{M_d} + \left(1 - \frac{M_w}{M_d}\right) \cdot q} \quad (2)$$

where $M_w = 18.0152$ g mol^{-1} and $M_d = 28.9644$ g mol^{-1} are molar masses of moist and dry air, respectively.

Radiosonde observations

In this study, RAOBs served as an additional validation data source. Meteorological profiles from the radiosonde stations located up to 70 km from the mean RO tangent point were downloaded from the National Oceanic and Atmospheric Administration Earth System Research Laboratory (NOAA/ESRL) radiosonde database (Govett 2020). The database provides temperature, pressure, and dew point depression measurements for at least 21 mandatory levels of constant pressure. The dew point depression T_{dd} can be transformed to the water vapor partial pressure based on the Clausius–Clapeyron equation (Perry 1950):

$$V_p = e_{s0} \cdot e^{\frac{l_v}{R_v}\left(\frac{1}{T_0} - \frac{1}{T_d}\right)} \quad (3)$$

where $e_{s0} = 6.11$ hPa is the reference saturation vapor pressure, $l_v = 2.5 \cdot 10^6$ J kg^{-1} denotes the latent heat of vaporization of water, $R_v = 461.525$ J K kg^{-1} is the gas constant for water vapor, $T_0 = 273.15$ K is the reference temperature and $T_d = T - T_{dd}$ stands for the dew point temperature in K.

Step 2: preprocessing

After the acquisition of the needed data and before the training, it was necessary to perform the preprocessing, which included vertical and 3D interpolation of the RO and ERA5 profiles, splitting the data into training and testing subsets, and eventually, co-location of RO and radiosonde observations and their vertical interpolation.

The models' inputs comprised of latitude, month, and hour of RO event as well as the vertical profiles of bending angle or refractivity linearly interpolated to 100 m resolution between 1 and 20 km resulting in 190 fixed levels. The upper boundary of 20 km was chosen as a compromise between computational speed, model complexity, and the altitude, above which water vapor becomes negligible.

To derive the target profiles of the temperature, pressure, and water vapor at the RO location, 3D interpolation (horizontal and vertical) was applied to the ERA5 meteorological data. Since the ERA5 stands out with the high 1 h resolution, temporal interpolation to the

time of the RO observations was omitted. Therefore, first, the vertical spacing of the ERA5 data was adjusted to the RO altitude (100 m resolution within 1–20 km altitude). I applied different interpolation strategies depending on the meteorological parameters. For the temperature, a simple linear interpolation was used, the water vapor partial pressure was interpolated exponentially, while the pressure at the particular height h was calculated based on the pressure P_i of the adjacent layer i at height h_i (Boehm and Schuh 2004; Wallace and Hobbs 2006):

$$P = P_i \cdot e^{-\frac{(h-h_i)\cdot g_m}{R_d \cdot T_v}} \tag{4}$$

where R_d is the gas constant for dry air, g_m denotes acceleration due to gravity, which can be calculated as a function of latitude and height (Kraus 2007), and T_v stands for the virtual temperature, which expresses the dry temperature with the same density as the moist air with the constant pressure:

$$T_v = \frac{T \cdot P}{P - \left(1 - \frac{M_w}{M_d}\right) \cdot V_p} \tag{5}$$

Afterward, the vertically uniform ERA5 profiles were interpolated horizontally to the mean RO tangent point position using bilinear interpolation.

In the next step, the pairs of input RO and output ERA5 profiles were subdivided into training and testing data sets. Since the total number of available RO observations exceeded 1.75 million, to reduce computational cost and satisfy memory constraints in the estimation of model parameters, the subsampling was performed. 10 000 random samples for each month between October 2019 and September 2020, giving 120,000 profiles in total, entered into a learning set and the 30,000 random observations between October and December 2020 were used for model testing (Fig. 2). It should be noted, that the Earth's topography was one of the factors, which constrained the RO technique to sound the atmosphere down below 1 km, which was set as a threshold in this study. Therefore, most (100,150, 83.5%) of RO profiles used for training were located over the wet areas and the rest 19,850 (16.5%) observations occurred above land. Similarly, the test data set consisted of 24,270 (80.9%) and 5730 (19.1%) samples, which took place over the oceans and land, respectively. The RO events included in the testing data set were co-located with nearby RAOBs using a 2 h time window and 70 km as a maximum spatial distance between the mean tangent point of RO profile and location of radiosonde station. This resulted in 477 pairs of co-located RAOB-RO cases, which distribution is presented in Fig. 2b.

Originally, RAOBs are provided at constant pressure levels, which correspond to different geometric heights depending on the current weather conditions. Hence, it was necessary to first determine the radiosonde data at common altitudes to be able to evaluate the ML model performance. Since the vertical resolution of radiosonde measurements is relatively sparse, I interpolated meteorological data at the chosen 10 rigid height levels of 1.5, 3.1, 5.8, 7.5, 9.6, 10.9, 12.4, 14.2, 16.6, and 18.7 km, which approximately equal to the mandatory pressure levels of 850, 700, 500, 400, 300, 250, 200, 150, 100, and 70 hPa.

Although the RF does not require any additional preprocessing steps before the training, for the ANN, it is recommended to perform data normalization, which leads to speeding up the learning, a more stable algorithm, and faster convergence. In this study, for convenience, min–max normalization was applied to input features (bending angle/refractivity, latitude, month, and hour) of both algorithms (ANN and RF), transforming them into the 0–1 range. Furthermore, it has to be noted that the vertical input profiles were scaled separately at each altitude level.

Step 3: hyperparameters tuning

Hyperparameters are configuration parameters, which control the learning process and, contrary to the model parameters, must be set in advance before the training. Hyperparameters can significantly influence the model performance and the selection set of the most optimal variables is one of the biggest challenges in model building. To mitigate this problem, a few hyperparameter optimization techniques have been developed, such as Random Search, Bayesian optimization, or Gaussian Process (Bergstra et al. 2011; Bergstra and Bengio 2012). In this study, I exploited the Random Search, separately for the bending angle/refractivity and ANN/RF models. The Random Search approach allows scanning a large domain of hyperparameters by selection and evaluation n random combinations. The number of n was set to 200 in this research. The evaluation of each set of hyperparameters was performed using the popular K-fold cross-validation. In the K-fold cross-validation, the training data set is randomly split into K equal-sized subsets, where one of the partitions is used in testing and the rest K-1 of data serves in the model learning. The process is repeated K times and the final output is estimated as the average of the K fitting results.

ANN and RF are characterized by different sets of available hyperparameters. For RF, the following hyperparameters were considered, where the numbers in the parentheses specify the ranges of values to try:

1. N estimators—the number of trees in a forest (50, 60, …, 600).
2. Max depth—the maximum depth of the tree (4, 6, …,26)
3. Min samples split—the minimum number of samples to consider for each split (30, 40, …, 90, 100, 150, 200,…, 500).
4. Min samples leaf—the minimum number of samples required to be at a leaf node (10, 15, 20, …, 45, 50, 60, …, 100).
5. Max features—the number of features consider when looking for the best split (10, 15, 20, …, 190, 194).
6. Bootstrap—defines whether some samples will be used multiple times in a single tree (True, False).

Note, the pre-pruning technique was used to mitigate the chance of overfitting. Tested ranges of hyperparameters, such as max depth, min samples split, and min samples leaf were constraint to prevent full growth of single trees.

For the ANN, I adjusted parameters related to the network structure (1–3) and training algorithm (4–6):

1. Layers—number of hidden units (1, 2, 3).
2. Neuron—number of units at each hidden layer. (195, 200, 205,… 500).
3. Dropout rate—a fraction of the neurons, which are randomly ignored during the training (0, 0.1, … 0.5).
4. Activation function (linear, rectified linear unit).
5. Epochs (200, 250, 300,…1600).
6. Batch—describes how many samples are processed before the update of internal ANN parameters (50, 100, …, 500, 1000, 1500, …, 9500, 10,000).

The weights and learning rates of the ANN were determined during the training using a popular and effective Adam optimization algorithm, which was presented for the first time by Kingma and Ba (2014).

It must be pointed out that many various sets of hyperparameters may result in a similar performance. Hence, to simplify the processing and save computational time, the RF models with the lowest number of trees or ANN with a minimum number of hidden layers and neurons were chosen. Eventually, the following ML models were revealed as the best during hyperparameters optimization:

- bending angle based ANN: 1 hidden layer, 455 neurons, 0 dropout rate, linear activation function, 850 epochs and batch size of 50 samples,
- refractivity based ANN: 1 hidden layer, 405 neurons, 0 dropout rate, linear activation function, 1150 epochs and batch size of 50 samples,
- bending angle based RF: 300 trees, max depth of 20, min samples split of 30, min samples leaf of 10, 20 max features, bootstrap option set False.
- refractivity based RF: 190 trees, max depth of 18, min samples split of 30, min samples leaf of 10, 15 max features, bootstrap option set True.

Results

Step 4: training and testing

Training accuracy of the different machine learning approaches was evaluated using the root mean square error (RMSE) between the outputs of different ML models and ERA5 target profiles on the testing data set. The corresponding RMSE were computed also on the training data set to check if algorithms are prone to either overfitting or underfitting. The vertically averaged RMSE for temperature, pressure and water vapor partial pressure together with the corresponding ERA5 standard deviations are comprised in Table 1. The same table also shows RMSE between 1DVar results provided by CDAAC in wetPf2 products and ERA5 targets. The attached ERA5 standard deviations were calculated based on the entire available data set (150,000 profiles) and reflect the climatological variability of the particular meteorological parameters. As to temperature, the RMSE varies between 1.51 K for the ANN based on refractivity up to 1.68 K for the RF with bending angle input. At the

Table 1 Fitting results obtained on the training and test data sets

		Artificial Neural Network				Random Forest				CDAAC wetPf2	ERA5	
		Bending angle		Refractivity		Bending angle		Refractivity				
		L	*T*	*L*	*T*	*L*	*T*	*L*	*T*			
Temperature (*K*)	RMSE	1.49	1.56	1.46	1.51	1.44	1.68	1.47	1.60	1.50	STD	4.82
Pressure (hPa)		1.33	1.34	1.20	1.22	1.10	1.42	1.12	1.26	1.05		4.23
Water vapor pressure (hPa)		0.44	0.44	0.43	0.43	0.43	0.48	0.43	0.46	0.45		0.93

Vertically averaged root mean square errors (RMSE) for the temperature, pressure, and water vapor partial pressure between ERA5 and obtained using different machine learning models on the training and test data sets indicated by columns 'L' and 'T', respectively or 1DVar approach stored in operational CDAAC wetPf2 product. The right column presents vertically averaged standard deviations for the corresponding meteorological parameters calculated from ERA5 reanalysis

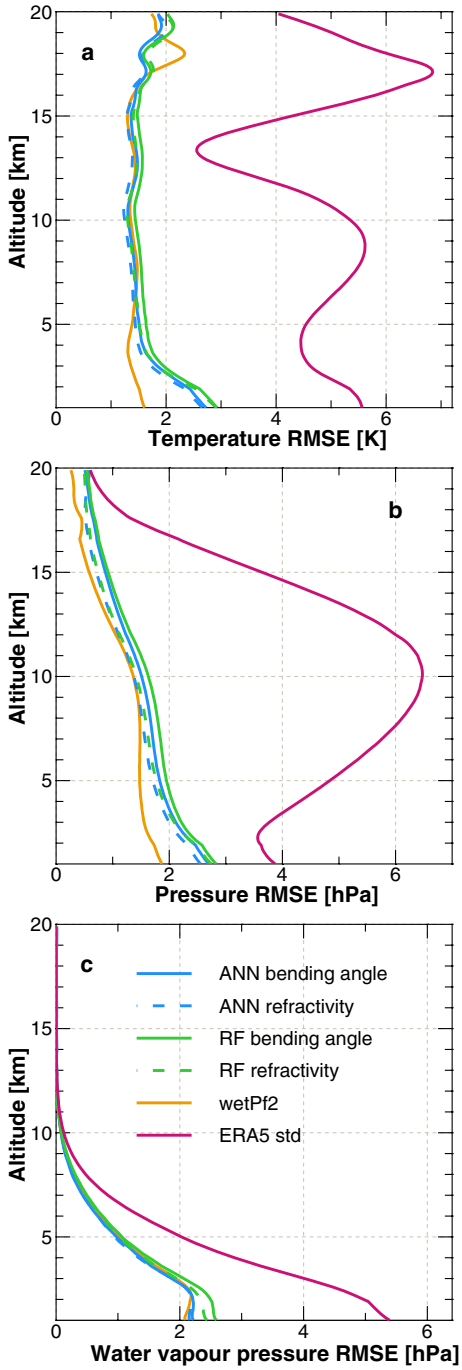


Fig. 3 Temperature (**a**), pressure (**b**) and water vapor partial pressure (**c**) RMSE vertical profiles. The errors are obtained on the testing data set from the operational 1DVar CDAAC wetPf2 product (orange lines), ANN (blue lines) and RF (green lines) approaches with bending angle (solid lines) or refractivity (dashed lines) inputs, with respect to the ERA5 targets. The pink lines present the ERA5 standard deviations of particular meteorological parameters

same time, the mean standard deviation of the ERA5 profile reaches slightly more than 4.8 K, which is significantly above the ML model errors. The comparison between ERA5 and wetPf2 temperatures revealed similar performance with RMSE of 1.50 K what confirms the good accuracy of temperature profiles retrieved using ML models.

Unfortunately, I obtained slightly worse fitting results for the pressure compared to the wetPf2 product. ANN and RF with refractivity input show the best performance with the vertically averaged RMSE of around 1.25 hPa, while the worst performance (1.42 hPa) is noted for the RF with bending angle input. The ERA5 mean standard deviation was 4.20 hPa, which is a factor almost 4 higher. Pressure RMSE for 1DVar solution was significantly smaller by around 0.2–0.3 hPa than for ML models. Almost the same performance is observed for 1DVar CDAAC product, both ANN and RF models with vertically averaged RMSE of around 0.45 hPa for water vapor pressure. Although the mean standard deviation of ERA5 water vapor reached 0.93 hPa and exceeds the fitting errors for all the models, it is only a factor 2 higher, which is small compared to the factors 3 for temperature and almost 4 for pressure. This result may reflect the complex and variable nature of water vapor, which is more difficult to predict. Differences between RMSE obtained on the training and test data sets were negligible for both ANN models and more pronounced for RF models, in particular for bending angle input. The temperature and pressure RMSE differences of around 0.2 K and 0.3 hPa indicate slight overfitting, although the pre-pruning was applied before hyperparameters optimization. The reasons for this result are not yet completely understood but may be attributed to distinct periods of training (October–September) and test (October–December only) data sets but also the difficulty of the problem to solve arising from the high number of input and output variables and complex nature of bending angle. This is an important issue for future research, which should cover feature engineering and validation on a larger data sample.

The models' performance was also evaluated in terms of training speed based on the mean of five model runs. The analysis was conducted on a machine with 32 Gb RAM and 16-cores Intel Xeon Silver 4216 CPU (2.10 GHz). Although RF models achieved a slightly worse accuracy, they outperform the training time of ANN with the mean execution times of 9.6 and 36.8 min for refractivity and bending angle inputs. Corresponding ANN models were computationally more demanding, which resulted in training times of 135.6 and 141.0 min, which are factors of around 14 and 4 higher.

The vertical RMSE profiles for temperature, pressure, and water vapor partial pressure with regard to

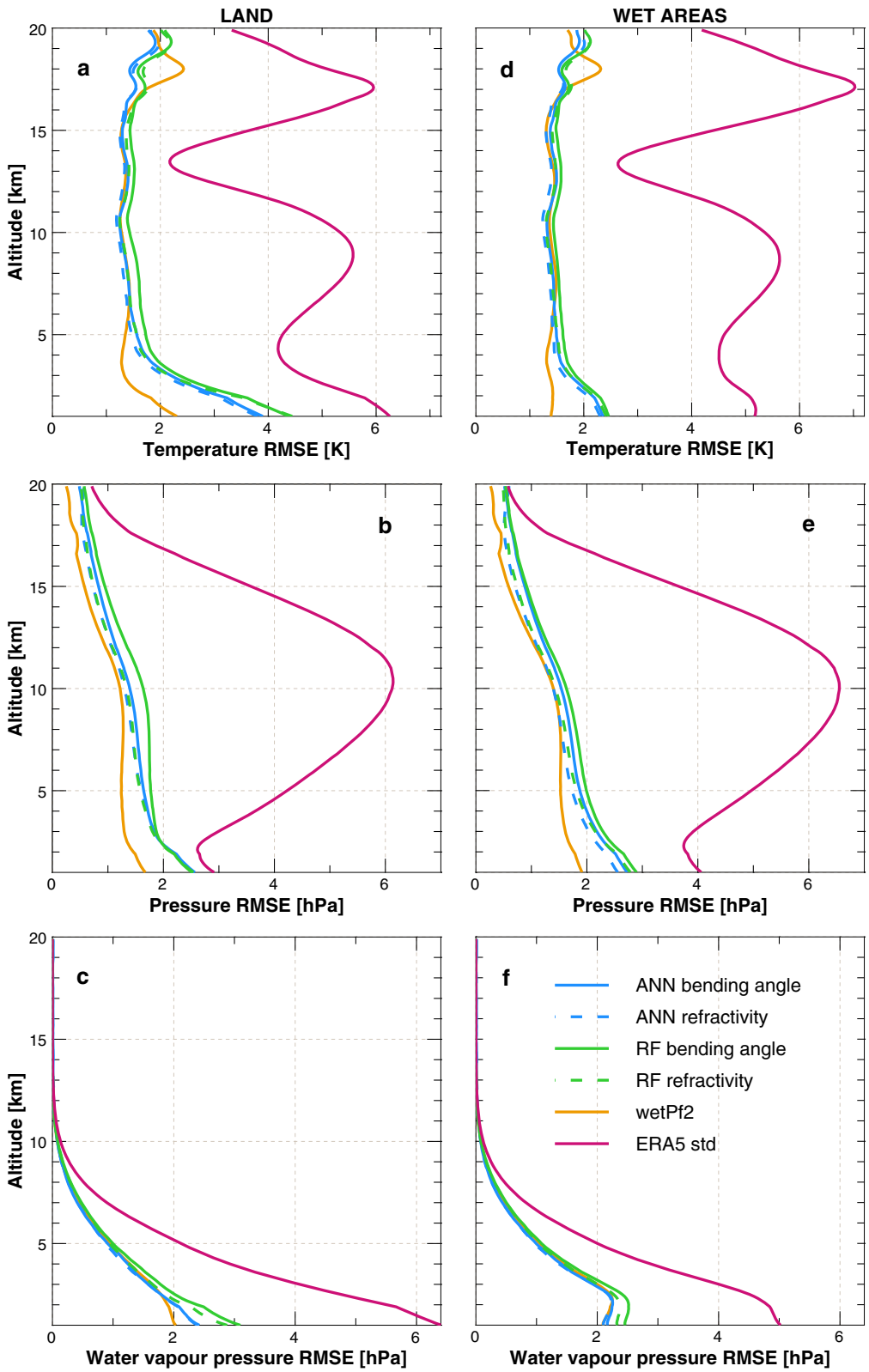


Fig. 4 Similar like Fig. 3 but for RO profiles located over land (**a**–**c**) and wet areas (**d**–**f**)

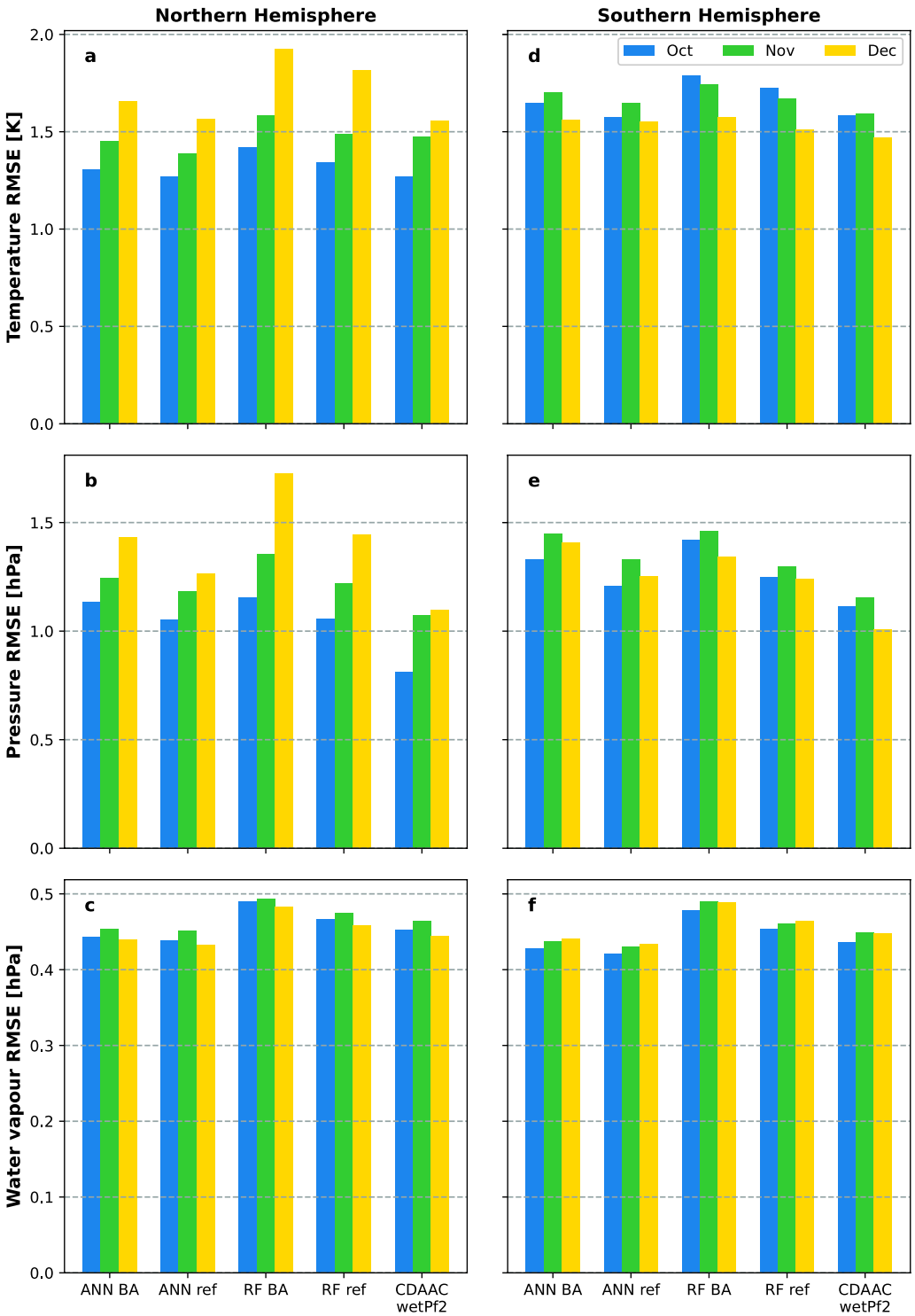


Fig. 5 Fitting results for different months on the Northern (left panels) and Southern (right panels) hemispheres. The bars show vertically averaged RMSE for temperature (**a**, **d**), pressure (**b**, **e**), and water vapor partial pressure (**c**, **f**) computed between ERA5 and different ML models or CDAAC wetPf2 product on test data sets. Blue, green, and yellow distinguish events in October, November and December

different ML models and operational 1DVar CDAAC product are presented in Fig. 3. The retrieval errors for the temperature (Fig. 3a) for all ML models and wetPf2 product are quite stable between 5 and 17 km altitude and equal to approximately 1.5 K. However, the RMSE for the RF model with bending angle input stands out slightly by around 0.2 K, what is in line with the highest vertically averaged RMSE. Unfortunately, below 3 km the temperature RMSE increases up to almost 3 K with negligible differences between the models, while the results for wetPf2 are significantly smaller by around 1 K. Between 17 and 19 km the opposite pattern is observed, ML RMSE are around 1.5 K, while 1DVar RMSE rises up to 2.5 K. At the same time, the ERA5 standard deviation greatly exceeds 4 K with an exception between 12 and 15 km with a drop below 3 K, which may be related to the tropopause height inducing lower temperature variability.

The pressure RMSE (Fig. 3b) for all ML models are in quite good agreement and drop from 2.7 to 1.0 hPa between 1.0 and 20.0 km. However, the RF model with bending angle input stands out with errors larger by around 0.2 hPa. A similar pattern but with slightly smaller RMSE is visible for the CDAAC product, where the maximum and minimum RMSE reached approximately 2 hPa and 0.5 hPa. While, the largest ERA5 standard deviation of more than 6 hPa is visible around 10.5 km altitude.

RMSE for the water vapor (Fig. 3c) shows a gradual decrease from 2.5 hPa to almost 0 between 2.5 to 12.5 km for all ML models and 1DVar solution and together with ERA5 standard deviation become negligible above the upper altitude. The results for all ML models and wetPf2 product are in good agreement; only in the lowest part below 2.5 km, RF models are characterized by larger errors by around 0.5 hPa. At the same time, ERA5 standard deviation reaches approximately 5.5 hPa at 1.0 km altitude, which is a factor of 2 larger. As mentioned before, most of the RO profiles took place over the oceans, which can have different characteristics and accuracy compared to the observations above the ground. Therefore, I computed RMSE separately for each group, which vertical profiles are depicted in Fig. 4. For temperature, the slightly worse performance of ML models is noted for land, in particular below 3 km, where RMSE gradually grows up to over 4 K for RF models with decreasing altitude. In comparison, RMSE for wet areas reaches less than 2.5 K, which may be related to the smaller ERA5 standard deviation. The different tendency is seen in pressure errors (Fig. 4b and d). Although ERA5 standard deviation below 6 km is larger by 1 hPa for wet areas, the accuracy of ML models is similar. The patterns of pressure RMSE for all ML models and CDAAC product are steady but shifted by around 0.1 hPa for wet areas compared to land. Similar to pressure, water vapor pressure errors are consistent for land and wet areas and differ only in the lowest part of the troposphere. A larger RMSE by around 0.5 hPa was noted for land, which may be a consequence of the larger water vapor variability reflected in the ERA5 standard deviation. The above findings suggest that the location of RO event or further, land cover, may have a significant impact on models training should be employed in a feature space.

In addition, fitting accuracy on the test data set was examined for different months and hemispheres. In the Northern Hemisphere (left panels in Fig. 5), temperature and pressure RMSE for all ML models and CDAAC wetPf2 gradually increase every month. The largest errors are noted in December, especially for the RF model with bending angle input with maxima of around 1.9 K and 1.7 hPa, respectively. Interestingly, temperature errors for all ML models and CDAAC product are similar in October and equal to around 1.3 K. Different pattern is visible in the Southern Hemisphere (right panels in Fig. 5), since it experienced reversed seasons and is covered mostly by water. The smallest temperature RMSE of around 1.5 K were in December, while bigger errors by around 0.1 K were in October and November. For pressure, the largest discrepancies are noted in November with values varying between 1.2 to just below 1.5 hPa. I obtained consistent results for water vapor partial pressure for both hemispheres and all tested months with the best accuracy of 0.42 hPa for the ANN model with refractivity input in October in the Southern Hemisphere.

Table 2 Results of validation with RAOBs

	Artificial Neural Network		Random Forest		CDAAC wetPf2
	Bending angle	Refractivity	Bending angle	Refractivity	
Temperature [K]	1.89	1.81	2.16	2.01	1.29
Pressure [hPa]	1.93	1.83	2.13	1.89	1.35
Water vapor pressure [hPa]	0.47	0.46	0.54	0.52	0.41

Vertically averaged RMSE between the temperature, pressure, and water vapor partial pressure obtained using different inputs and machine learning models, operational CDAAC wetPf2 and 477 co-located RAOBs with co-location criteria of 70 km and 2 h time window

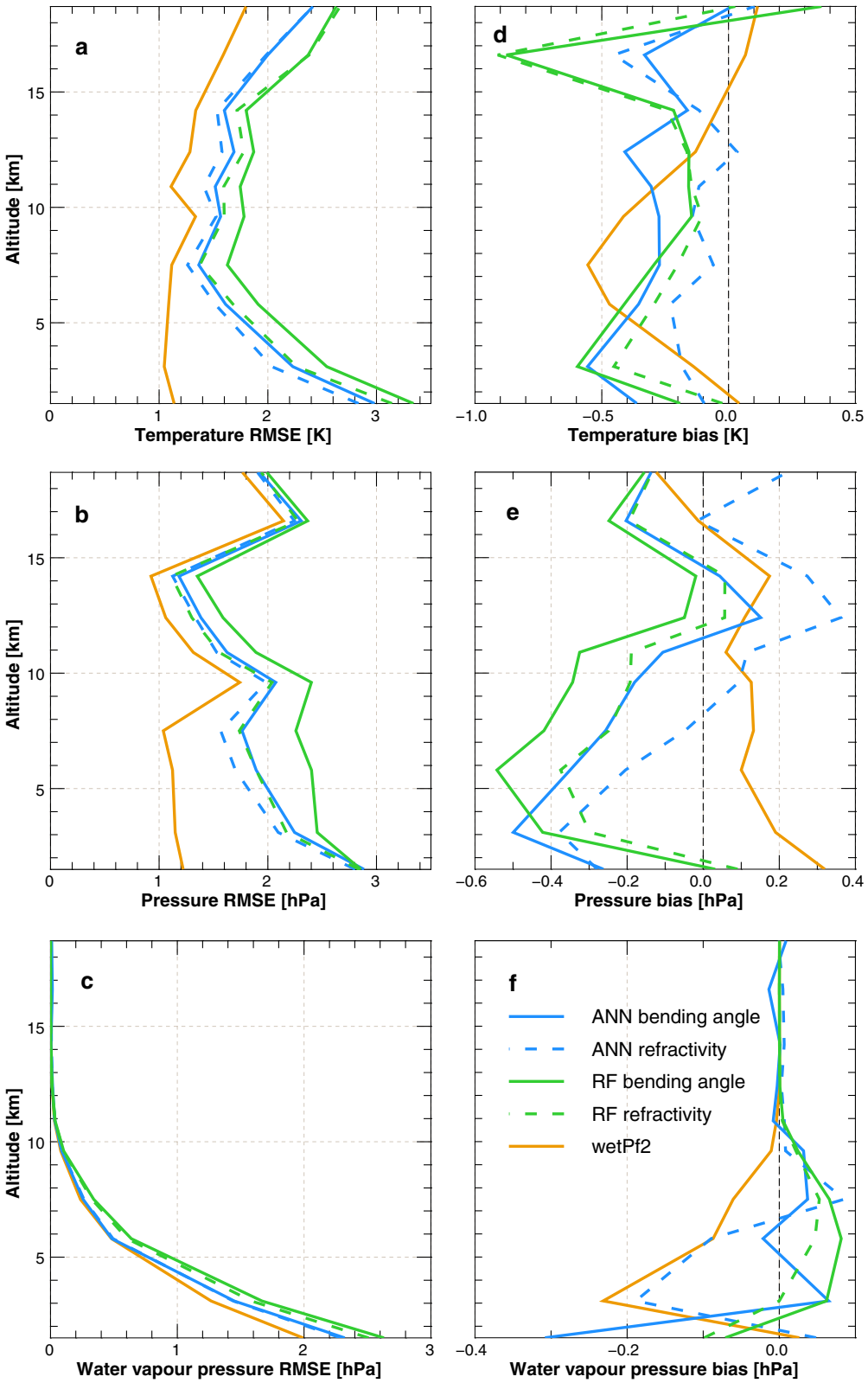


Fig. 6 Overall validation results for temperature (top), pressure (middle) and water vapor partial pressure (bottom). Vertical RMSE (**a**–**c**) and mean differences (**d**–**f**) profiles are obtained using ANN (blue lines) and RF (green lines) models with bending angle (solid lines) and refractivity (dashed lines) inputs, official CDAAC wetPf2 product (orange lines) and 477 co-located RAOBS with co-location criteria of 70 km and 2 h time window

Step 5: validation with radiosondes

In this study, RAOBs were used as an additional validation data source for the ML retrievals. Table 2 compares the vertically averaged RMSE for temperature, pressure and water vapor partial pressure between trained ML models, operational CDAAC 1DVar retrievals stored in wetPf2 products and 477 co-located RAOBs. For all meteorological parameters, the best performance among ML models is noted for the refractivity-based ANN, which is in line with the testing results presented in the previous subsection. The mean RMSE between the predicted and RAOBs temperature, pressure are equal to 1.87 K and 1.83 hPa, respectively. While the mean water vapor RMSE for all ML models are barely distinguishable with the smallest RMSE of 0.46 hPa. The worst results were obtained for RF with bending angle with the mean RMSE of 2.16 K, 2.13 hPa, and 0.54 hPa for the temperature, pressure, and water vapor. Furthermore, it is worth noticing that using the bending angle as an input vector in both, ANN and RF, models results in larger discrepancies than corresponding models with refractivity inputs. Unfortunately, the validation results for ML models are worse than calculated for the CDAAC product with the mean RMSE of 1.29 K, 1.35 hPa, and 0.41 hPa for the temperature, pressure and water vapor. However, only profiles over land were taken into account in the validation, which comes from the limitation of the location of radiosonde stations. As presented in Fig. 4, the ML learning accuracy decreases over land areas, especially in the lower troposphere. Thus, the above results further support the idea to include the information about wet/land location in the input features.

Comparison between the vertical profiles of RMSE and mean differences for temperature, pressure, and water vapor between ML, 1DVar CDAAC retrievals and co-located RAOBs is presented in Fig. 6. As mentioned above, the lowest temperature RMSE is observed for the wetPf2 product where the mean error is approximately 1.2 K up to 12.5 km and gradually reaches almost 2.0 K at 20 km altitude (Fig. 6a). A similar but shifted by over 0.5 K pattern is visible for the ML outputs. However, the mean RMSE at 1.5 km ranges from 2.8 to 3.3 K with the lowest and highest values for refractivity-based ANN and RF with bending angle input, respectively. Similar to the 1DVar results, there is a significant increase in RMSE above 12.5 km, which reaches more than 2.5 K at the highest analyzed level. Mean temperature biases for all ML models and wetPf2 product (Fig. 6d) are mostly negative with absolute values below 0.5 K. However, contrary to RMSE, the best results were obtained for ANN with refractivity input. The largest temperature discrepancy of around -0.9 K is present for both RF models at 16.6 km, which may be related to tropopause altitude. Interestingly, the largest wetPf2 bias of over -0.5 K is visible at 7.5 km.

The vertical profiles of pressure RMSE for ML models follow each other above 9.6 km, while below, there are significant discrepancies (Fig. 6b). The maximum difference of around 0.7 hPa is noted between RF with bending angle input and ANN with refractivity input, which presents the best performance. The largest RMSE of 3.0 hPa is seen at 1.5 km for bending angle ANN, while the lowest errors of around 1.2 hPa below 7.0 km are noted for the wetPf2 product. It is clear from Fig. 6e that ML models tend to overestimate pressure, especially below 10 km with maximum biases of over -0.5 hPa for bending angle based ANN and RF. Opposite Similar to the temperature and pressure, the best agreement is observed between RAOBs and CDAAC water vapor pressure, with RMSE slowly decreasing from just above 2.0 hPa to almost 0 above 12.4 km. However, the big peak of over -0.2 hPa is seen at 3.1 km in the mean bias profile (Fig. 6f) Among the ML models, the smallest RMSE is noted for ANN with bending angle input with a maximum of 2.2 hPa at 1.5 km altitude. Surprisingly, RF with refractivity input shows the best performance in terms of mean differences with absolute values staying within ± 0.1 hPa.

Figure 7 presents the differences of the temperature (a–c), pressure (d–f) and water vapor pressure (g–i) between ML, CDAAC outputs and RAOBs for 3 out of 477 co-located RO events. The events were selected based on the spatial and time differences below 15 km and 30 min for mid ($>35°$), subtropical (23.5°–35.0°) and tropical (0–23.5°) latitudes. The upper panel shows the RO case result, which occurred at 12:15 UTC on December 7, 2020, and RAOB from the station located in Topeka (95.62°W 39.07° N) in the USA with the distance of 8 km to the mean RO tangent point. For the temperature (Fig. 7a), differences for CDAAC and ML retrievals are mostly within ± 2 K below 16 km, while the largest errors of -4 K and 5 K are noted for RF with bending angle input at the most upper layer. There is a satisfactory agreement between various ML outputs for the pressure, which are mostly positive and less than 2 hPa. Surprisingly, the differences between the official CDAAC product and RAOB are almost always positive with the exception of a negative peak of over -2.5 hPa at 3.1 km. The water vapor differences (Fig. 7g) are often close to zero with an exception for RF models with large spikes of -1.5 to -3.0 hPa.

The second analyzed event (middle panels) took place at 23:12 UTC on December 15, 2020, where the closest available RAOB was located 14 km away in Delhi, Safdarjung (77.20° E 28.58° N) in India. Contrary to the previous event, all of the differences (Fig. 7b, e, h) are mostly negative. As to the temperature, the largest errors of over

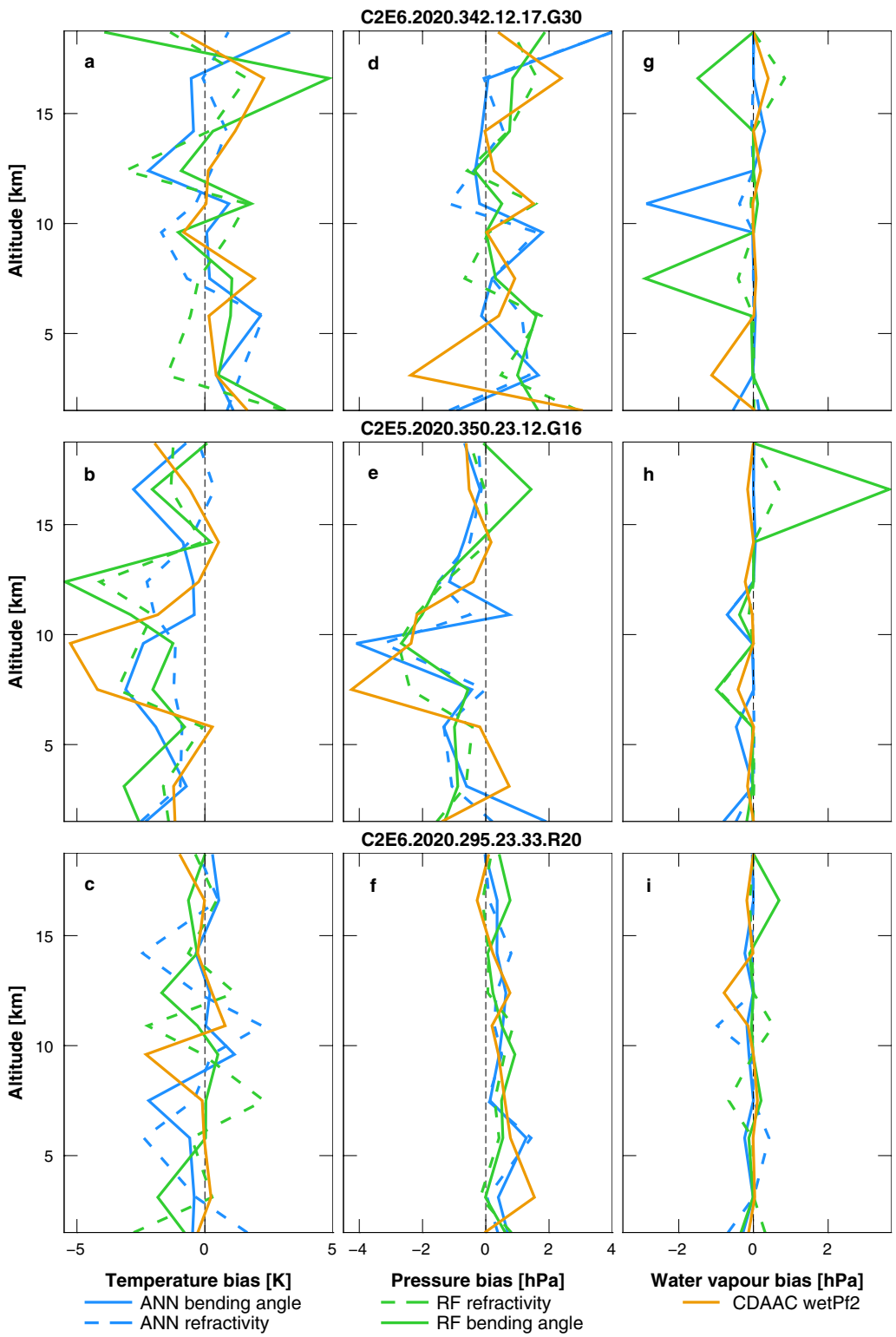


Fig. 7 Validation results for temperature (**a**–**c**), pressure (**d**–**f**) and water vapor partial pressure (**g**–**i**). Vertical profiles show differences between co-located RAOBs and RO retrievals obtained using different machine learning models or 1DVar approach contained in the official wetPf2 CDAAC product for radiosonde stations in Topeka (top panels), Delhi, Safdarjung (middle panels) and Legazpi, Luzon (bottom panels). Bold titles indicate the CDAAC profile ID and vertical dashed lines represent zero difference

−5 K are reported at 9.6 km for wetPf2 and at 12.4 km for RF with bending angle input. A similar negative peak for wetPf2 of slightly exceeding 4 hPa is seen at 7.5 km in the pressure profiles. The water vapor errors fluctuate around zero for all ML models and CDAAC product; however, the RF model with bending angle input stands out with an unexpected error of over 3 hPa at 16.6 km. Interestingly, the best overall performance was recorded for the ANN with bending angle input.

The bottom panel in Fig. 7 shows the results for the RO event, which happened at 23:31 UTC on October 21, 2020, with relation to the RAOB taken 6 km away in Legazpi, Luzon (123.73° E 13.15° N) in the Philippines. The zigzagging temperature differences for all ML models and 1DVar solution hardly exceeds 2.5 K, while the pressure differences (Fig. 7f) are quite stable and below 1.1 hPa and tightly follow each other for all altitudes. The best performance was recorded for water vapor pressure, where all the differences stay within 0.7 hPa, which is quite surprising, since the radiosonde station is located in the tropics characterized by abundant water vapor.

Discussion

RO technique provides meteorological profiles with high vertical resolution and accuracy. However, commonly applied solutions to derive tropospheric profiles depend on the a priori information about temperature or pressure. To overcome this problem, recent studies focused on ANN approaches, where no auxiliary data is needed. In one of the first studies, Bonafoni et al. (2009) reported RMSE between ECMWF analysis and ANN outputs reaching more than 3 K for the temperature and 2.5 hPa for the water vapor pressure for vegetation zone at around 1.5 km altitude. Slightly better results were observed for the tropics in the work of Pelliccia et al. (2010) where the maximum RMSE for the temperature was just above 3 K and for the water vapor pressure did not exceed 2.5 hPa. While the best performance was achieved by Pelliccia et al. (2011) with RMSE of less than 2 K, 4 hPa, and 0.2 hPa for the temperature, pressure, and water vapor, respectively. My findings agree well with these results in terms of temperature and pressure RMSE, which were mostly around 1.5 K and below 2 hPa. As expected, only for the water vapor, I computed larger RMSE of around 2.0–2.5 hPa at the lowest altitudes. The prime reason for the discrepancy is the restricted use of RO profiles only to the winter season in the north of the Arctic Polar Circle, where water vapor content is small and less variable. It is confirmed by the small standard deviation of water vapor from ECMWF analysis, which did not exceed 1.2 hPa in their work, whereas in the current study, I calculated the ERA5 standard deviation of more than 5 hPa. Furthermore, their comparison of 2 RO events co-located with RAOBs revealed huge pressure differences of more than −4 hPa, while in the current study, the differences between ML models and RAOBs were mostly within −3 to 3 hPa (Fig. 7) and rarely exceeded this range. The lower pressure differences obtained in the present study may also be a consequence of the output used to train the models. The former study used only ECMWF analysis, whereas the current work exploited state-of-the-art ERA5 reanalysis, which provides meteorological profiles with higher accuracy and higher spatial and temporal resolutions.

Shyam et al. (2016) tested two different inputs of bending angle or refractivity to derive water vapor profiles using ANN. Their study revealed better performance for ANN with refractivity input. The maximum RMSE between ANN with refractivity and bending angle inputs and CDAAC product was below 1.5 and 5 hPa, respectively. My results confirm the better performance of ML models with refractivity input. Nevertheless, the differences between water vapor RMSE for ML models with refractivity and bending angle inputs are smaller and do not exceed 0.2 hPa compared to the 3.5 hPa demonstrated by Shyam et al. (2016). These negligible differences can be explained by improved SNR in the COSMIC-2 mission, which contributes to the smaller bending angle errors.

Conclusion

This study compares different methods to determine tropospheric profiles of temperature, pressure, and water vapor pressure based on the RO observations from the COSMIC-2 mission. For this purpose, I trained and tested 4 ML models embracing ANN and RF algorithms with refractivity and bending angle profile as the inputs. ML is a powerful tool to estimate meteorological profiles using only RO data without external knowledge on temperature or pressure from weather models or other measurements. As an input, I exploited globally distributed 150,000 RO profiles between October 2019 and December 2020, whereas the training target consisted of ERA5 meteorological profiles of temperature, pressure, and water vapor interpolated to the position of the RO event.

I obtained acceptable and consistent results for all trained ML models. The vertically averaged RMSE between predicted and target ERA5 profiles were around 1.5 K for temperature, 1.3 hPa for pressure, and 0.5 hPa for water vapor. The largest errors of almost 3 K, 2.7 hPa, and a little more than 2.5 hPa are noted below 2.5 km altitude. Disappointingly, further comparison carried out with RAOBs revealed bigger discrepancies. Vertically averaged RMSE between ML outputs and observations varied from 1.8 to 2.2 K for temperature and 1.8–2.1 hPa for pressure. Only the water vapor errors

of around 0.5 hPa are in good agreement with the previous findings. Similar examination with respect to the operational 1DVar CDAAC product showed smaller errors of 1.3 K for temperature and 1.4 hPa for pressure, whilst RMSE of just above 0.4 hPa for water vapor pressure compares well with the ML models accuracy. The substantial disagreement between ML models and CDAAC product was noted below 5 km altitude. As well known, the lower troposphere is highly variable and difficult to predict environment, which consequently results in the decline of ML models performance. The obtained errors are higher than I expected and there is certainly room for improvement. Hence, to address this issue, I am planning to train separate ML models for the upper and lower troposphere, which are represented by different characteristics and processes. Taken together, apart from the slight disagreement for temperature and pressure, I believe the presented results compare quite well with CDAAC retrievals and encourages further research in this field.

To the best of my knowledge, it was the first attempt to apply the RF algorithm to profile the troposphere based on RO measurements. Current solutions focused only on the ANN application, whilst the use of RF was disregarded. In general, my results demonstrated a quite good agreement between ANN and RF retrievals with the RMSE differences between the models mostly of around 0.2 K for the temperature and 0.1 hPa for the pressure, and negligible for water vapor partial pressure. These differences can be explained by usually better prediction skills of good-tuned ANN models as the problem complexity and size of the training data set increase (Kayri et al. 2017) compared to RF, which is successfully applied in tasks, where limited data sample is available. However, one of the major ANN drawbacks is the long development time and a good way to mitigate this problem is using RF instead. Although RF models showed slightly worse accuracy, they were a few times faster to train and outperformed ANN with processing speeds of around 10 and 40 min.

It is plausible that a number of limitations could have influenced the results obtained. First, I developed and tested the models using only subsamples of all the available observations. Second, I did not include the contribution of hydrometeors in the output vector. Hydrometeors are products of the condensation or deposition of atmospheric water vapor, such as rain, snow, fog, or clouds. It is widely assumed that hydrometeors have a negligible impact on GNSS signal. However, the more recent evidence (Yang and Zou 2012; Zou et al. 2012; Lasota et al. 2018) emphasizes their importance in GNSS retrievals, where ignoring their contribution may lead to the significant errors of derived water vapor profiles. An additional possible source of the error may arise from the used features. Selection and extraction of relevant and independent features play a key role in ML models' training and can improve their speed and accuracy. Here, the features set consisted of the refractivity/bending angle profile and latitude, month, and hour of the RO event. Bonafoni et al. (2009) trained two separate ANN for vegetation and desert areas, which resulted in different model accuracy. As also presented in the comparisons between retrievals over land and oceans, different hemispheres and months, and indirectly in the validation with the RAOBs, extension of the feature set by the information about surface/land cover and elimination of less important characteristics should be undertaken in the next study. Furthermore, future work will concentrate on the detection of severe weather events, such as extratropical storms, heavy precipitation, or tropical cyclones and training appropriate ML models. Eventually, to further my research, I plan to exploit convolutional neural networks, which take into account spatial relationships, which may have a significant contribution to tropospheric profiling.

Abbreviations

ANN: Artificial Neural Network; CDAAC: COSMIC Data Analysis and Archive Center; COSMIC: Constellation Observing System for Meteorology, Ionosphere, and Climate; DVar: Dimensional variational; ECMWF: European Centre for Medium-Range Weather Forecast; GLONASS: Globalnaja Nawigacionnaja Sputnikovaya Sistema; GNSS: Global Navigation Satellite Systems; ML: Machine Learning; MLP: Multilayer Perceptron; NOAA/ESRL: National Oceanic and Atmospheric Administration Earth System Research Laboratory; NWP: Numerical weather prediction; RAOB: Radiosonde Observations; RF: Random Forest; RMSE: Root mean square error; RO: Radio Occultation; SNR: Signal-to-noise ratio.

Acknowledgements

The author would like to thank UCAR CDAAC for processing and sharing COSMIC-2 RO data and ECMWF for providing ERA5 reanalysis data sets.

Authors' contributions

EL convinced, planned and performed the experiment, analyzed and visualized the results. The manuscript was written and revised by EL. The author read and approved the final manuscript.

Funding

This research was conducted under the Leading Research Groups support project from the subsidy increased for the period 2020–2025 in the amount of 2% of the subsidy referred to Art. 387 (3) of the Law of 20 July 2018 on Higher Education and Science. This study was funded by Wrocław University of Environmental and Life Sciences.

Availability of data and materials

The COSMICS-2 profiles used to train and testing the models in the current study are available in the UCAR COSMIC Program repository, https://doi.org/10.5065/t353-c093. The ERA5 reanalysis used as a target in this study is available in the Copernicus Climate Data Store, https://cds.climate.copernicus.eu/. The radiosonde observations are available in NOAA/ESRL radiosonde database, https://ruc.noaa.gov/raobs/.

Declarations

Competing interests

The author declare that she has no competing interests.

Received: 19 August 2021 Accepted: 15 November 2021
Published online: 11 December 2021

References

Ali J, Khan R, Ahmad N, Maqsood I (2012) Random forests and decision trees. Int J Comput Sci Issues 9(5):272

Anthes RA (2011) Exploring Earth's atmosphere with radio occultation: contributions to weather, climate and space weather. Atmos Meas Tech 4:1077–1103. https://doi.org/10.5194/amt-4-1077-2011

Bergstra J, Bengio Y (2012) Random search for hyper-parameter optimization. J Mach Learn Res 13(1):281–305

Bergstra JS, Bardenet R, Bengio Y, Kégl B (2011) Algorithms for hyper-parameter optimization. In: Advances in neural information processing systems. pp 2546–2554

Boehm J, Schuh H (2004) Vienna mapping functions in VLBI analyses. Geophys Res Lett. https://doi.org/10.1029/2003GL018984

Bonafoni S, Pelliccia F, Anniballe R (2009) Comparison of different neural network approaches for the tropospheric profiling over the inter-tropical lands using GPS radio occultation data. Algorithms 2(1):31–45. https://doi.org/10.3390/a2010031

Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140

Breiman L (2001) Random forests. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324

Chen S-Y, Liu C-Y, Huang C-Y, Hsu S-C, Li H-W, Lin P-H, Cheng J-P, Huang C-Y (2021) An analysis atudy of FORMOSAT-7/COSMIC-2 radio occultation data in the troposphere. Remote Sens 13(4):717. https://doi.org/10.3390/rs13040717

Gardner MW, Dorling S (1998) Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmos Environ 32(14–15):2627–2636

Gorbunov ME, Sokolovskiy SV (1993) Remote sensing of refractivity from space for global observations of atmospheric parameters. Max-Planck-Institut für Meteorologie

Govett M (2020) NOAA/ESRL Radiosonde Database. In: URL https://ruc.noaa.gov/raobs/. Accessed 1 Jul 2020

Hassoun MH (1995) Fundamentals of artificial neural networks. MIT press, Cambridge

Healy S, Eyre J (2000) Retrieving temperature, water vapour and surface pressure information from refractive-index profiles derived by radio occultation: a simulation study. Q J R Meteorol Soc 126(566):1661–1683

Hersbach H, Bell B, Berrisford P, Hirahara S, Horányi A, Muñoz-Sabater J, Nicolas J, Peubey C, Radu R, Schepers D et al (2020) The ERA5 global reanalysis. Q J Royal Meteorol Soc. 146(730):1999–2049

Ho S-P, Zhou X, Shao X, Zhang B, Adhikari L, Kireev S, He Y, Yoe JG, Xia-Serafino W, Lynch E (2020) Initial assessment of the COSMIC-2/FORMOSAT-7 neutral atmosphere data quality in NESDIS/STAR using in situ and satellite data. Remote Sens 12(24):4099. https://doi.org/10.3390/rs12244099

Huang C-Y, Kuo Y-H, Chen S-Y, Terng C-T, Chien F-C, Lin P-L, Kueh M-T, Chen S-H, Yang M-J, Wang C-J et al (2010) Impact of GPS radio occultation data assimilation on regional weather predictions. GPS Solut 14(1):35

Kayri M, Kayri I, Gencoglu MT (2017) The performance comparison of Multiple Linear Regression, Random Forest and Artificial Neural Network by using photovoltaic and atmospheric data. In: 2017 14th International Conference on Engineering of Modern Electric Systems (EMES). pp 1–4

Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint, arXiv:14126980

Kraus H (2007) Die Atmosphäre der Erde: Eine Einführung in die Meteorologie. Springer-Verlag

Kursinski ER, Hajj GA, Schofield JT, Linfield RP, Hardy KR (1997) Observing Earth's atmosphere with radio occultation measurements using the Global Positioning System. J Geophys Res 102(D19):23429–23465. https://doi.org/10.1029/97JD01569

Lasota E, Rohm W, Liu C-Y, Hordyniec P (2018) Cloud detection from radio occultation measurements in tropical cyclones. Atmosphere 9(11):418. https://doi.org/10.3390/atmos9110418

Li Y, Kirchengast G, Scherllin-Pirscher B, Schwaerz M, Nielsen JK, Ho S, Yuan Y (2019) A new algorithm for the retrieval of atmospheric profiles from GNSS radio occultation data in moist air and comparison to 1DVar retrievals. Remote Sens 11(23):2729. https://doi.org/10.3390/rs11232729

Łoś M, Smolak K, Guerova G, Rohm W (2020) GNSS-based machine learning storm nowcasting. Remote Sens 12(16):2536. https://doi.org/10.3390/rs12162536

Michie D, Spiegelhalter DJ, Taylor CC (eds) (1994) Machine Learning, Neural and Statistical Classification. Ellis Horwood, Upper Saddle River, NJ, USA

O'Sullivan D, Herman B, Feng D, Flittner D, Ward DM (2000) Retrieval of water vapor profiles from GPS/MET radio occultations. Bull Am Meteor Soc 81(5):1031–1040

Pelliccia F, Pacifici F, Bonafoni S, Basili P, Pierdicca N, Ciotti P, Emery WJ (2011) Neural networks for arctic atmosphere sounding from radio occultation data. IEEE Trans Geosci Remote Sens 49(12):4846–4855. https://doi.org/10.1109/TGRS.2011.2153859

Pelliccia F, Bonafoni S, Basili P, Ciotti P, Pierdicca N (2010) Atmospheric profiling in the inter-tropical ocean area based on neural network approach using GPS radio occultations. Open Atmos Sci J. 4(1)

Perry JH (1950) Chemical engineers' handbook. ACS Publications

Poli P, Joiner J, Kursinski ER (2002) 1DVAR analysis of temperature and humidity using GPS radio occultation refractivity data. J Geophys Res 107(D20):ACL-14

Rennie MP (2010) The impact of GPS radio occultation assimilation at the Met Office. Q J R Meteorol Soc 136(646):116–131. https://doi.org/10.1002/qj.521

Scherllin-Pirscher B, Steiner AK, Kirchengast G, Kuo Y-H, Foelsche U (2011) Empirical analysis and modeling of errors of atmospheric profiles from GPS radio occultation. Atmos Meas Tech 4(9):1875–1890. https://doi.org/10.5194/amt-4-1875-2011

Schreiner WS, Weiss J, Anthes RA, Braun J, Chu V, Fong J, Hunt D, Kuo Y-H, Meehan T, Serafino W et al (2020) COSMIC-2 radio occultation constellation: first results. Geophys Res Lett 47(4):e2019GL086841

Shyam A, Gohil BS, Basu S (2016) Retrieval of water vapour profiles from radio occultation refractivity using artificial neural network. Indian J Radio Space Phys 42(6):411

Siroky DS (2009) Navigating Random Forests and related advances in algorithmic modeling. Stat Surv 3:147–163. https://doi.org/10.1214/07-SS033

UCAR COSMIC Program (2019) COSMIC-2 Data Products. https://www.cosmic.ucar.edu/what-we-do/cosmic-2/data/. Accessed 1 Mar 2021

Wallace JM, Hobbs PV (2006) Atmospheric science: an introductory survey. Elsevier, Amsterdam

Ware R, Exner M, Feng D, Gorbunov M, Hardy K, Herman B, Kuo Y, Meehan T, Melbourne W, Rocken C et al (1996) GPS sounding of the atmosphere from low earth orbit: preliminary results. Bull Am Meteor Soc 77(1):19–40

Wulfmeyer V, Behrendt A, Kottmeier C, Corsmeier U, Barthlott C, Craig GC, Hagen M, Althausen D, Aoshima F, Arpagaus M, Bauer H-S, Bennett L, Blyth A, Brandau C, Champollion C, Crewell S, Dick G, Girolamo PD, Dorninger M, Dufournet Y, Eigenmann R, Engelmann R, Flamant C, Foken T, Gorgas T, Grzeschik M, Handwerker J, Hauck C, Höller H, Junkermann W, Kalthoff N, Kiemle C, Klink S, König M, Krauss L, Long CN, Madonna F, Mobbs S, Neininger B, Pal S, Peters G, Pigeon G, Richard E, Rotach MW, Russchenberg H, Schwitalla T, Smith V, Steinacker R, Trentmann J, Turner DD, van Baelen J, Vogt S, Volkert H, Weckwerth T, Wernli H, Wieser A, Wirth M (2011) The Convective and Orographically-induced Precipitation Study (COPS): the scientific strategy, the field phase, and research highlights. Q J R Meteorol Soc 137(S1):3–30. https://doi.org/10.1002/qj.752

Wulfmeyer V, Hardesty RM, Turner DD, Behrendt A, Cadeddu MP, Girolamo PD, Schlüssel P, Baelen JV, Zus F (2015) A review of the remote sensing of lower tropospheric thermodynamic profiles and its indispensable role for the understanding and the simulation of water and energy cycles. Rev Geophys 53(3):819–895. https://doi.org/10.1002/2014RG000476

Yang S, Zou X (2012) Assessments of cloud liquid water contributions to GPS radio occultation refractivity using measurements from COSMIC and CloudSat. J Geophys Res. https://doi.org/10.1029/2011JD016452

Zou X, Yang S, Ray PS (2012) Impacts of ice clouds on gps radio occultation measurements. J Atmos Sci 69(12):3670–3682. https://doi.org/10.1175/JAS-D-11-0199.1

Publisher's Note