Integrated water vapor and liquid water path retrieval using a single-channel radiometer

Anne-Claire Billault-Roux[1] and Alexis Berne[1]

[1]Environmental Remote Sensing Laboratory, Swiss Federal Institute of Technology, Lausanne

Correspondence: Alexis Berne (alexis.berne@epfl.ch)

Abstract. Microwave radiometers are widely used for the retrieval of Liquid Water Path (LWP) and Integrated Water Vapor (IWV) in the context of cloud and precipitation studies. This paper presents a new site-independent retrieval algorithm for LWP and IWV, relying on a single-frequency 89-GHz ground-based radiometer. A statistical approach is used, based on a neural network, which is trained and tested on a synthetic data set constructed from radiosonde profiles worldwide. In addition to 89-GHz brightness temperature, the input features include surface measurements of temperature, pressure and humidity, as well as geographical information and, when available, estimates of IWV and LWP from reanalysis data. An analysis of the algorithm is presented to assess its accuracy, the impact of the various input features, as well as its sensitivity to radiometer calibration and its stability across geographical locations. While 89-GHz brightness temperature is crucial to LWP retrieval, only moderately does it contribute to IWV estimation, because of the lower sensitivity of this channel to water vapor emission. The algorithm is shown to be quite robust although its accuracy is inevitably lower than that obtained with state-of-the-art multi-channel radiometers, with a relative error of 7.2 % for LWP (on cloudy cases with LWP > 30 $\mathrm{g\,m^{-2}}$) and 5.2 % on IWV. The highest accuracy is obtained in mid-latitude environments with a moderately moist climate, which are more represented in the training dataset. The new method is then implemented and evaluated on real data that were collected during a field deployment in Switzerland and during the ICE-POP 2018 campaign in South Korea. ~~The new algorithm is shown to be quite robust, especially in mid-latitude environments with a moderately moist climate, although its accuracy is inevitably lower than that obtained with state-of-the-art multi-channel radiometers.~~

1 Introduction

Clouds play a key, though complex, role in the atmosphere's radiative balance and global circulation (Hartmann and Short, 1980; Slingo, 1990; Hartmann et al., 1992; Wang and Rossow, 1998; Stephens, 2005; Mace et al., 2006; McFarlane et al., 2008), and cloud studies have thus been propelled to the forefront of climate research. One of the core challenges is the monitoring, quantification and modelling of cloud liquid water, which has a significant contribution to radiative processes on a global scale. In this perspective, highly accurate methods were developed to retrieve liquid water path (LWP) as well as integrated water vapor (IWV) from microwave radiometer measurements, relying on the fact that water in its liquid and vapor phases is the main atmospheric contributor to brightness temperatures in millimeter wavelengths, outside of the oxygen window. On a different note, quantifying cloud liquid water content is also relevant to the field of snowfall studies. Identifying

the presence of supercooled liquid water during a snowfall event is of paramount importance to the understanding of snowfall microphysics, for it drives riming of snow particles, which in turn ~~affect~~ affects the efficiency and the spatial distribution of precipitation (Saleeby et al., 2011), as well as wet deposition of aerosols (Poulida et al., 1998). Improving the monitoring of cloud liquid water processes is thus valuable to climatological, meteorological and hydrological applications.

The quantitative retrieval of LWP from ground-based or satellite measurements of brightness temperature (T_B) at a single millimeter wavelength is an underdetermined problem. This brightness temperature results from the radiative contribution of gases ~~, aerosols~~ and hydrometeors across the atmospheric column, and depends on the vertical profile of temperature. To lift this underdetermination, state-of-the-art retrievals of LWP and IWV rely on multi-frequency radiometers, which provide T_B measurements in several microwave channels. This allows to separate the contributions of water vapor and liquid water (e.g. Westwater et al. (2001)) and, to some extent, retrieve the full profile of liquid water content and humidity in the atmospheric column (Löhnert et al., 2004).

Such multi-frequency instruments are however not always available. It was shown (Küchler et al., 2017) that a radiometer channel at 89 GHz could be added to a W-band cloud radar operating at 94 GHz, thus allowing to have collocated measurements of radar variables and brightness temperature. This paves the way for an improved understanding of cloud and precipitation physics. Küchler et al. (2017) proposed a method to derive LWP estimates from the single-frequency measurements of brightness temperature, and the present study builds on those findings.

Two approaches are commonly considered for the retrieval of LWP and IWV, as described in Turner et al. (2007) and Cadeddu et al. (2013). The first method relies on the reconstruction of atmospheric profiles, with a physical model that is iterated until modeled T_Bs match the measured ones. Although this method is formally the most accurate (Turner et al., 2007), it ~~is computationally heavy and cannot be implemented accurately when only one frequency is available~~requires more than one radiometer frequency to lift the problem's fundamental underdetermination, and is thus not applicable for this study. The other way to tackle the problem is to derive statistical relationships between T_Bs and LWP and/or IWV based on synthetic datasets. This approach has been widely used, both for ground-based and satellite applications, with varying degrees of complexity in the algorithms (linear, quadratic, log fitting or using neural network architectures) (Karstens et al., 1994; Löhnert and Crewell, 2003; Mallet et al., 2002; Cadeddu et al., 2009). The retrieval coefficients that are computed with this method are usually site-specific, since they incorporate during the learning or regression stage the climatological features at the location of the dataset. The geographical range within which a site-specific algorithm could be reliable is difficult to estimate, especially if the orography of the region is complex, as highlighted by Massaro et al. (2015). In general, implementing a site-specific algorithm in a location with a different climatology is likely to yield erroneous retrievals (Gaussiat et al., 2007). In order to implement such an algorithm at another site, a new parameterization should be performed using a suitable dataset; but there might not always be enough reliable data available for this purpose. In order to avoid this lengthy process, and in the case of instruments that are intended to be deployed in various locations, a site-independent algorithm is more adequate (Liljegren et al., 2001).

The purpose of this study is to present a new site-independent method for the retrieval of both LWP and IWV, that relies on a single radiometer frequency. The regression is performed through a neural network, whose input consists of brightness temperature at 89 GHz, as well as surface measurements and geographical information. Those additional input features are

especially key to the retrieval of IWV, to which T_B at 89 GHz is less sensitive than to LWP. Although it implies a loss of ~~performance~~ precision in comparison with state-of-the-art multi-frequency retrievals, the advantage of this method is to be applicable in any location with a constrained uncertainty.

The following section describes the data used in the different steps of this study, from the design steps to the validation of the new method. Section 3 outlines the forward model that is used to build the synthetic ~~database~~ dataset on which the LWP and IWV retrieval algorithms are trained. In Sect. 4, the design of the algorithms is detailed, and the results on the synthetic dataset are reviewed and analyzed in Sect. 5. An independent validation of the method is presented in Sect. 6 using two contrasted datasets that were collected during field deployments in Payerne (Switzerland) and in the Taebaek mountains (South Korea). A summary and conclusions are provided in Section 7.

2 Data

The present work is based on two types of data: a multiyear collection of radiosonde observations across the world (for training and testing of the retrieval algorithms) as well as sets of measurements from ~~a~~ an 89-GHz radiometer deployed in various regions during field campaigns limited in time. Those two types of data are described below.

2.1 Radiosonde dataset

The design of a statistical algorithm requires a large dataset on which to perform statistical learning. Here, this ~~database~~ dataset was built using radiosonde profiles collected in over 180 stations throughout the world, available through the University of Wyoming portal (Oolman, 2020). In total, $\sim 10^6$ radiosonde profiles are used, from 20 years of data (2000-2019). It was ensured that the ~~database~~ data included radiosonde stations from all climatic regions and covering a wide range of altitudes (0 to 4000 m). However, lack of available data in some areas inevitably results in an unbalanced dataset, where polar and tropical areas are under-represented compared to mid-latitudes, especially Europe. The possible impact on the performance of the algorithm is further discussed in Sect. 5.

~~The~~ A quality check was performed on each of the relevant variables (pressure, temperature, relative humidity), through the following steps: first, the minimum and maximum P (resp. T, RH) in a given range of altitudes were extracted from each radiosonde. When examining the distributions that are obtained, outliers were visible, which were then removed with a 10^{-4} quantile (upper and lower quantile). The atmospheric column was split into 9 ranges of altitudes, and this routine was performed for each. In total, 6395 profiles were flagged out and removed. It was ensured that this did not result in the systematic removal of some geographical locations. Following this step, the vertical profiles of pressure, temperature and relative humidity are ~~extracted from the radiosonde data and then~~ used as input to the forward model, ~~as~~ described in Sect. ~~3.~~ 3. The vertical resolution is relatively low (0.37 km on average).

2.2 Field deployments

In the validation ~~steps~~ stage of this work, the new method was implemented ~~on~~ using real 89-GHz radiometer data, that were collected during campaigns described below.

2.2.1 Instrument

The main instrument that was used ~~in this study~~ for the implementation of the algorithm is the one described in Küchler et al. (2017), which is here referred to as WProf. This radar-radiometer system, conceived and built by RPG, consists of a 94-GHz Frequency-Modulated Continuous Wave (FMCW) cloud radar with an 89-GHz radiometer channel, which allows for joint active and passive retrievals of cloud and precipitation. In the data presented here, WProf was deployed together with a weather station that provided surface measurements of temperature, pressure and relative humidity.

2.2.2 Payerne 2017

The first data set on which the new algorithm was evaluated was collected during a field deployment in Payerne (Switzerland), at 450 m of altitude, in late spring 2017 (May 15th – June 15th). As a means of comparison, data from the Swiss meteorological institute (MeteoSwiss) was used. MeteoSwiss's facilities in Payerne comprise a multi-frequency radiometer with tipping-curve calibration, HATPRO (Rose et al., 2005; Löhnert and Maier, 2012). This state-of-the-art instrument retrieves LWP and IWV with a nominal accuracy of respectively 20 $\mathrm{g\,m^{-2}}$ and 0.2 $\mathrm{kg\,m^{-2}}$ (RPG Radiometer Physics GmbH, 2014). During this deployment, both WProf and HATPRO measured brightness temperatures with a high temporal resolution of the order of a few seconds. The instruments were located approximately 65 m apart; this distance is small enough that it should in general not affect the comparison of the retrieved values from the two instruments. However, in some rare cases, it is possible that a cloud would overpass one of the radiometers, but not the other, leading to a discrepancy in the measured brightness temperatures.

In addition, radiosondes are launched twice daily in Payerne by MeteoSwiss, allowing for the direct computation of IWV values, which are used as a further source of validation for the IWV retrieval algorithm.

2.2.3 ICE-POP 2018

The second dataset on which the new algorithm was tested was gathered during the ICE-POP 2018 campaign, which took place in South Korea during the 2017-2018 winter, in the context of the 2018 Olympic and Paralympic winter games in Pyeong Chang~~(Gehring et al., 2020)~~. A description of the data is presented in Gehring et al. (2020). During this campaign, the weather was generally cold and dry; nine precipitation events were recorded, and occasional fog was present (about 25 occurrences during the campaign timeframe). WProf was deployed from November 2017 to April 2018 in Mayhills, 50 km south-east of Pyeong Chang, at 789 m of altitude. This allows for an implementation of the algorithm in a different context than Payerne: i.e. in winter conditions and ~~at~~ in a fully different geographical ~~location,~~ setting, located at a lower latitude and closer to the sea.

In this case, unlike in Payerne, no independent measurements of LWP are available; however, radiosondes were launched every 3 hours, thus providing a means of comparison for IWV retrievals, although only with a lower temporal resolution.

3 Forward model

In order to develop a statistical algorithm, a large amount of data is required to reliably perform the statistical learning phase. For this purpose, a synthetic ~~database~~ data set was built, using as a starting point the radiosonde profiles ~~worldwide~~ described in the previous section. A two-step forward model was implemented, first to identify clouds in each profile and derive the corresponding liquid water content, then to compute the resulting 89-GHz brightness temperature. The different steps of this forward model are illustrated in the flowchart in Fig. 1.

3.1 Cloud liquid model

To derive profiles of liquid water content (LWC) from radiosonde profiles of atmospheric variables, the cloud model from Salonen and Uppala (1991) was used. Cloud boundaries are identified using a threshold U_c on relative humidity, this threshold being pressure- and temperature-dependent, according to Eq. 1.

$$U_c = 1 - \alpha\sigma(1-\sigma)[1+\beta(\sigma-0.5)] \quad (1)$$

Here, $\sigma = \frac{P}{P_0}$ with P and P_0 denoting respectively atmospheric pressure at the current level and at the ground. Corrections from Mattioli et al. (2009) are used for the coefficients α and β of the Salonen model. ~~More~~ Within the cloud layers, the liquid water profile is then calculated as a function of temperature and height above cloud base, following Eq. 2.

$$LWC = w_0\left(\frac{h-h_b}{h_r}\right)^a f(T) \quad (2)$$

where $f(T) = 1 + cT$ for $T \geq 0$ and $f(T) = exp(cT)$ for $T < 0$, with T in °C, $a = 1.4$, $c = 0.04°C^{-1}$, $w_0 = 0.17\ g\ m^{-3}$, $h_r = 1.5$ km, h and h_b denoting height and height of cloud base. There are some limitations to assuming a single universal cloud model, since it may fail to capture specific cloud properties in certain environments: more sophisticated and accurate models ~~can~~ could be defined on a local geographical scale to counter this (e.g. Pierdicca et al. (2006)), ~~but~~ . However, given the stated objective of this study to design a non-site specific algorithm, it was considered preferable to assume a single universal liquid cloud model, in spite of its potential ~~limitations~~drawbacks.

A further limitation of the cloud model is related to the relatively low resolution of the atmospheric profiles extracted from the radiosonde data (c.f. Sect. 2.1) that are used as an input. This might result in a misrepresentation of the cloud layers in their detection and their size. In order to ensure that this forward model generated the least possible bias, its results were compared against LWP values from ERA5 reanalysis data (Copernicus Climate Change Service, 2020). Even though the model might fail, on a given occurrence, to reproduce the actual liquid water profile in the atmospheric column, it should not produce a significant bias on average. This condition guarantees that the synthetic dataset that is used for training contains realistic – if

not real – profiles, and this should therefore not degrade the quality of the retrieval algorithm. This cloud model was chosen over other commonly used ones (Decker model, Salonen model without correction, c.f. Mattioli et al. (2009)) for it was found to produce the least bias when compared to ERA5 LWP values (mean bias of 14 $\mathrm{g\ m^{-2}}$ vs. 26 $\mathrm{g\ m^{-2}}$ (resp. -24 $\mathrm{g\ m^{-2}}$) for the unadjusted Salonen model (resp. the Decker model with 95 % threshold).

3.2 Radiative transfer model

Ground-level brightness temperatures ($\mathrm{T_B}$) at 89 GHz are simulated for each profile using the Passive and Active Microwave TRansfer Model (PAMTRA (Maahn, 2015; Mech et al., 2020)) available at https://github.com/igmk/pamtra (last access: ~~May 27th~~Nov 18th, 2020). As input to the radiative transfer calculations, vertical profiles of temperature, pressure, hydrometeor mixing ratio and water vapor mixing ratio are used. Gaseous absorption is calculated using the default parameters in PAMTRA, i.e. with the model proposed by Rosenkranz (1998) and modifications from Liljegren et al. (2005) and Turner et al. (2009). It should be kept in mind that some irreducible uncertainty remains tied to the choice of these parameters in the radiative transfer model.

The cloud droplet size distribution (DSD) is chosen as a ~~gamma distribution following Karstens et al. (1994)~~monodisperse distribution with radius $r_c = 20\ \mu\mathrm{m}$ following Cadeddu et al. (2017), and scattering calculations are performed with Mie equations, assuming spherical particles. Let us note here that the exact choice of the DSD has little impact on $\mathrm{T_B}$ modeling as long as the droplets are in the Rayleigh regime for the given frequency, since the emission cross-section in this regime is quasi-linearly related to the particle's volume. When the droplet size deviates from this regime, for instance as droplets grow larger near the onset of precipitation, then the Rayleigh assumption falls short and higher-order terms in the Mie equations become non-negligible, which alters the modeling of $\mathrm{T_B}$ (e.g., Zhang et al., 1999). This implies that the algorithm will output biased results when applied to raining cases, and should not be trusted in those ~~cases~~circumstances. This shall be considered as an intrinsic limitation to the algorithm.

In order to have a more rigorous grasp on when and how this drawback might affect the synthetic data set that is constructed – and hence the retrieval –, criteria from Karstens et al. (1994) were used. In their study, the authors distinguished three types of liquid water clouds based on the value of LWC at a given altitude; for each category of cloud, a different characteristic radius is chosen for the DSD. Assuming that Mie effects start to become an issue in the second category of clouds (*cumulus congestus*), identified for LWC > 0.2 $\mathrm{g\ m^{-2}}$, this corresponds on average to LWP $\geq$ 830 $\mathrm{g\ m^{-2}}$, and around 2% of the entire dataset fall into this category (i.e. the LWC threshold is exceeded in at least one altitude range). Taking the third category (*cumulonimbus*) with LWC > 0.4 $\mathrm{g\ m^{-2}}$, this applies to 1% of the entire dataset and the threshold increases to LWP > 1400 $\mathrm{g\ m^{-2}}$. Those values can serve as a benchmark to identify LWP values where Mie effects can typically contaminate the retrieval. However, edge cases can also exist where the total LWP is quite low, but a small layer of nearly-precipitating or drizzling cloud still contaminates the retrieval, without featuring extremely high total LWP.

Finally, the forward model that is presented here does not include the contribution of ice clouds and snowfall. While radiative emission from ~~solid hydrometeors~~ ice and snow particles has a minor influence on brightness temperature when compared to emission from liquid droplets and water vapor, ~~they~~ and is in general negligible, solid hydrometeors do contribute to microwave

brightness temperature through the backscattering of surface radiation. Scattering from snowfall particles is difficult to model accurately, but Kneifel et al. (2010) suggest that this effect could be notable during snowfall, in a way that is highly dependent on the microphysical properties of snowfall particles, and that would increase with their size. The present study does not take into account this process and could therefore yield biased results during intense snowfall events.

4 Design of the IWV and LWP retrieval algorithms

4.1 Input features

When a single frequency is available for the measurement of T_B, the problem's underdetermination can be partially relieved by including other available information in the retrieval's measurement vector. Adding further information also allows to retrieve IWV, for which 89-GHz T_B alone would be insufficient. In this study, several categories of variables were included in the input features. The first category consists of T_B and higher order polynomials (up to fourth degree) and is expected to have the greatest importance in the retrieval of LWP. The effect of higher order terms will be discussed further on. In order to simulate realistic measurements, a random Gaussian noise was added to the modeled brightness temperatures, with a mean and standard deviation of resp. 0 K and 0.5 K ; those values were identified by Küchler et al. (2017) as the characteristics of the measurement noise of the 89-GHz radiometer. Secondly, surface measurements are included (temperature, sea-level pressure and relative humidity); in the case of the radar-radiometer set-up that is used here, a weather station is collocated, meaning those measurements are available at the location of the instruments. The third category of input features comprises geographical descriptors: latitude, longitude, altitude; the day of year is also included in this group of features, as a means to account for seasonal variability in atmospheric and meteorological conditions. When available, a fourth category is added to the input features with reanalysis data (precipitable water and liquid water) from ERA5 (Copernicus Climate Change Service, 2020). The spatial and temporal resolution of this reanalysis data is too low for it to be held as ground truth, but it can serve as a reasonable rough estimate ~~in~~ and thus bring some improvements to the statistical learning process – although it could not be included as such in a physical model. Those four groups of features are used both for the retrieval of IWV and that of LWP. In the case of LWP, an additional input feature can be added, which is the output of the IWV retrieval algorithm. The impact of each of those feature groups on the retrieval will be discussed in Sect.5.

4.2 Dataset preprocessing

~~Strong rain~~ Rain events should be excluded from the training set, since they are out of the scope of the algorithm's range of validity, as explained in Sect. 3. Profiles with LWP > 1000 $\mathrm{g\ m^{-2}}$ are therefore removed ~~(i.e. in the range of heavy rain according to Cadeddu~~ (i.e. in the range of heavy rain according to Cadeddu et al., 2017, and in view of the discussion conducted in Sect. 3.2). The resulting dataset contains $\sim 10^6$ profiles and is used for the design of the IWV retrieval algorithm.

4.2.1 Further preprocessing for LWP dataset

In the case of LWP retrieval, additional pre-processing is needed, since the forward model produced a large majority of clear-sky cases. If left as such, the training phase ~~will~~ would result in a strong bias of the retrieval toward low LWP values (a bias of $\sim 100\ \mathrm{g\ m^{-2}}$ for LWP $> 400\ \mathrm{g\ m^{-2}}$ was noted in the development stages of the algorithm): this is a common artefact in statistical learning algorithms, as an effect of unbalanced training set. In order to avoid this, the dataset was subsampled so that clear-sky and cloudy cases (up to $600\ \mathrm{g\ m^{-2}}$) would be equally represented; the value chosen for this threshold results from a trade-off between bias reduction and preservation of overall accuracy. The resulting histogram is shown in Fig. 2, and the LWP dataset thus contains $\sim 10^5$ profiles. In the case of IWV, the distribution is also not uniform, but it suffers from a much smaller ~~bias~~ assymetry than the initial LWP data set. After some trials, it was considered preferable to use the full IWV data set rather than go through subsampling steps, which did not seem to bring significant improvements in this case. It should also be noted here that the additional pre-processing that was necessary for the LWP retrieval algorithm led us to design two separate algorithms, rather than a single one that would retrieve IWV and LWP at once. Indeed, while LWP retrieval is mostly relevant in cloudy cases, IWV can show some significant variability in clear-sky cases, which should therefore not be excluded from the training stage.

4.3 Statistical retrieval using a neural network

~~Both~~ After preprocessing, LWP and IWV datasets were randomly split into training, validation and testing set (70 %-15 %-15 %), and normalized using mean and standard deviation of each input feature in the training set. The validation set is used for tuning the hyperparameters of the neural network, while the final evaluation metrics are computed on the testing set. A densely-connected neural network architecture was chosen over linear regression and decision-tree-based retrieval techniques for it was found to produce more reliable results, with higher accuracy than the former and less prone to overfitting than the latter. The algorithm was designed using the Keras library in Python (Chollet et al., 2015). The neural ~~networks' hyperparameters were tuned on the validation set~~network was trained through mini-batch gradient descent, using RMSprop optimizer which allows for learning rate adaptation and is often used for statistical regression problems (Chollet, 2017). As comes across from the training curve of the LWP retrieval on Fig. 3, the training dataset is large enough to ensure that the algorithm is not prone to overfitting. Indeed, the error on the validation set quickly drops when the size of the training set, then plateaus with a slight decrease. In other words, the accuracy of the algorithm is not limited by the amount of data used in the training stage. Figure 4 and Table 1 summarize the resulting architecture and relevant parameters of the algorithm. These include the description of the neural network's structure (number of neurons and hidden layers) as well as training parameters such as the batch size and number of epochs, i.e. the number of iterations through the entire dataset in the learning phase. Different versions of the algorithm were trained, using various sets of input features, to assess the importance of each category (discussed below).

5 Results on synthetic dataset

In this section, the algorithm is evaluated on the synthetic dataset (testing set), through different criteria. Overall, results are encouraging and the retrieval appears to be robust. Some limitations can be identified, which will be discussed here. Additionally, an analysis on the impact of the various input features on the retrieval of IWV and LWP is conducted.

5.1 Error curves

~~Figure 5 illustrates~~ Fig. 5 presents the distribution of the error ~~over the validation set~~on the testing set, for the best version of the algorithm, which is the one that uses the full set of input features. In panels c) and d), the target variable, respectively IWV and LWP, is binned to intervals on which the root mean square error (RMSE) is calculated. This illustrates the behavior of the algorithm across the entire range of values, rather than summarizing the performance with a single metric such as total RMSE, which can conceal specific behaviors related to the distribution of the target variable in the dataset. Along the same line, we emphasize that comparing those total RMSE values to those from other studies should be done carefully because they strongly depend on the dataset from which they are calculated. In a similar way, panel e) (resp. f) illustrates the distribution of the mean bias across the range of IWV (resp. LWP) values.

Figure 6 shows how this total error, represented by RMSE (left panels) and by the square correlation coefficient (R^2) (right panels), is affected by the addition or removal of input features. For each set of input features, a full tuning of the algorithm was performed, and the results that are presented correspond to those from the tuned – i.e. best – version on the ~~validation~~ testing set.

5.1.1 IWV algorithm

~~The~~ Overall, the IWV retrieval algorithm yields a RMSE of 1.6 $\mathrm{kg\ m^{-2}}$ on the testing set, which corresponds to a relative error of 5.2 %~~, and is reasonably well distributed across the dataset. There is however a bias for large IWV values (> 60), which~~ . For comparison, the ERA5 data alone has a higher RMSE (3.4 $\mathrm{kg\ m^{-2}}$) on the same data set. Looking at Fig. 5 a), c) and e), it comes across that the retrieval performs quite well over the full range of IWV values, and the error distribution is relatively homogenenous. For high IWV values, however, a significant negative bias is present (as large as -6 $\mathrm{kg\ m^{-2}}$). Because such high values are underrepresented in the dataset, ~~and which the algorithm thus tends to underestimate (see Fig. 5 a)~~they are not well captured during the statistical learning stage, which leads to a systematic underestimation. However, these are by definition "border" cases, for which a decrease in accuracy is inevitable.

From Fig. ~~a)~~ 6 a) and b) it comes across that the IWV retrieval is significantly improved by the addition of ~~secondary~~ multiple input features. ~~Including~~ The highest accuracy is obtained with the full set of input features, and corresponds to a RMSE of 1.53 $\mathrm{kg\ m^{-2}}$. On the other hand, including solely $\mathrm{T_B}$ measurements in the input deteriorates the RMSE to nearly 6 $\mathrm{kg\ m^{-2}}$. ~~It appears that the most important secondary feature~~ If only one input feature were available, all the versions would predict worse results than those given by reanalysis data. Including $\mathrm{T_B}$ in the retrieval does not lead to the same leap in accuracy than for LWP (discussed in the following subsection), which was expected because the microwave frequency that is ~~ERA5~~

~~estimates, but that IWV is also significantly correlated to climatological features such as surface atmospheric variables as well as geographical and temporal information .~~ used here (89 GHz) is much less sensitive to water vapor. However, excluding T_B from the input features degrades the RMSE to 2.56 kg m^{-2}, i.e. + 67 % error, which clearly shows that some information is extracted from the brightness temperature in the retrieval.

~~Overall, the IWV retrieval is reliable and accurate, in particular when using all available secondary information.~~ An analysis was conducted to identify the importance of higher order polynomials in the algorithm, a summary of which can be found in Appendix. It was found that the most accurate retrieval is obtained by including T_B and T_B^2. If higher order terms are added, this slightly reduces the accuracy of the retrieval, and also degrades its robustness to T_B miscalibration. On the other hand, including only T_B, while it makes the algorithm slightly more stable, does not appear as the best solution for it has lower accuracy. Hence, the results which are presented here and in the following sections are those obtained using T_B and T_B^2.

5.1.2 LWP algorithm

The LWP retrieval algorithm has a RMSE of 86 g m^{-2} at best on the testing set (~~see Fig. c), i.e. 24.6 % in relative error on the total~~ training set: 84 g m^{-2} and validation set: 86 g m^{-2}). This corresponds to a relative error of 19 % on the testing set. ~~If~~ Let us underline that the subsampling which is performed on the dataset for the retrieval of LWP is applied to training, validation and testing sets: the results that are presented here are therefore computed on the testing set with a truncated distribution – i.e. after subsampling. Additionally, if clear-sky cases are removed using 30 g m^{-2} as a threshold value, following Löhnert and Crewell (2003), the relative error is ~~17.5 %.~~ 7.2 %. As already mentioned, the total RMSE values given here should be taken with care since they depend on the data set's distribution. For comparison, when the retrieval is implemented on the full dataset, i.e. without the subsampling step, the total RMSE drops to 40 g m^{-2}. The RMSE is here again rather homogeneous across the range of LWP values (Fig. 5.d), with however a ~~bias~~ small bias of around 20 g m^{-2} for low LWP values (visible in Fig. 5 f), which are slightly overestimated, and ~~for~~ while there is an underestimation of large LWP (LWP > 800 g m^{-2})~~which are underestimated (see Fig. 5 b and d). The latter point is due to~~ , with a negative bias down to -100 g m^{-2}. Both biases result from an effect of regression towards the mean, which is a intrinsic artefact of statistical algorithms. The significant negative bias for large LWP values is enhanced by the lack of data in this range~~, and~~ . It is likely acceptable, for it would correspond ~~to raining events~~ mostly to raining cases (light to moderate) which the retrieval does not aim to capture; yet this highlights once again that those cases are out of the algorithm's scope ~~. As already mentioned, the total RMSE values given here~~ and that retrievals with high LWP should be taken with care~~since they depend on the data set's distribution. For comparison, when the retrieval is implemented on the full dataset, i. e. without the subsampling described in Sect. 4, the total RMSE drops to 40~~ .

~~Figure~~ The analysis of higher order terms' importance in the case of LWP retrieval shows that the best results are obtained by using T_B polynomials up to the fourth order (see Appendix), while this does not affect significantly the stability of the retrieval to errors in T_B. Let us highlight that in the case of a linear regression, one would expect the error to diverge when high-order polynomials are included. This is not the case here, because of the non-linear behavior of the neural network. Therefore, in the results which are shown here and further, "T_B" implies that T_B polynomials up to the fourth order are used.

Figures 6 c) ~~shows that the LWP retrievalis less affected than IWV by secondary input features : most categories other than brightness-temperature-derived features have a relatively small impact on the error. This is reasonable since liquid water typically has a greater contribution to T_B than water vapor (Löhnert and Crewell, 2003). Furthermore,~~ and d) show that for LWP retrieval, input features other than T_B only bring second-order improvements, while they were shown to be crucial in the IWV retrieval. For instance, the addition of reanalysis data significantly improves the IWV retrieval, but only in a relatively minor way does it increase the accuracy of LWP retrieval. On the contrary, excluding T_B from the input features leads to RMSE near 200 $\mathrm{g\,m^{-2}}$ and $R^2 < 0.5$, i.e. to values with make the retrieval not relevant. This highlights once again that brightness temperature at 89 GHz is much more sensitive to liquid water than to water vapor. An additional reason for this high dependence on T_B is that LWP at a given location can have a high temporal variability due to cloud dynamics in the atmospheric column, which might not always be captured in the time series of surface atmospheric variables~~. Similar reasons can help explain why the addition of reanalysis data significantly improves the IWV retrieval, but only in a minor way does it increase the LWP retrieval's accuracy. Liquid water content can vary on a shorter~~ , nor by ERA5 models which have a comparatively low spatial and temporal ~~scale than that captured by ERA5 models~~resolution.

Still, the accuracy of the algorithm drops severely when no other features are considered than brightness temperature (RMSE of 140 $\mathrm{g\,m^{-2}}$). This means that, albeit second-order when taken individually, and somehow redundant when all used together, the secondary input features are efficient in incorporating statistical trends and climatological information to the retrieval during the training phase.

Adding IWV prediction as an input feature to the LWP retrieval has a very minor impact. For clarity, it was only included in Fig. 6 c) in the best-case scenario and not for every other combination of input features. This is not surprising, since it is itself the output of an algorithm that relies on essentially the same input features. However, the slight improvement that is seen can be understood by recalling that the IWV retrieval algorithm was trained on a much larger dataset, which includes in particular a larger number of clear-sky cases (cf. Sect. 3).

5.2 Sensitivity to instrument calibration

In order to assess the stability of the algorithm with respect to potential miscalibration or calibration drift of the radiometer, T_B offsets were virtually added to the testing dataset before implementing the retrieval. Figure 7 illustrates the behavior of the algorithm when such a miscalibration, with a constant offset is present (varying from 0 to 5K). Panel a) shows that a 5 K offset in T_B results in a 30 % increase in RMSE for the IWV estimations, which is non-negligible. Ensuring proper radiometer calibration thus seems crucial in constraining the error of this retrieval. For comparison, the 89-GHz radiometer presented in Küchler et al. (2017) has a nominal accuracy of 0.5 K, after calibration. If the calibration cannot be ensured, and if there is no means to correct for miscalibration (of > 3K), it is preferable for IWV retrieval to use the algorithm that does not rely on T_B, shown with the black dashed line.

In terms of relative impact, the LWP algorithm is less affected (Fig. 7 b)) with an increase of the RMSE of less than 10 % for an offset of 5 K in T_B, which makes it reasonably stable to inaccuracy of T_B measurements. It also appears that the different versions are affected in a similar way by offset T_B values. ~~When examined closely, the general trend follows the intuition that the retrieval is slightly more stable when more input features are included, which counterbalances the effect of the miscalibration~~However, the algorithm which includes the prediction of IWV in the input features diverges faster than the others. This is understandable, for the error on T_B propagates through the IWVpred input feature, in addition to the T_B features themselves. Therefore, in the case of uncertain calibration, more robust results would be obtained without including this feature.

5.3 Geographical distribution of the error

One of the motivations of this study was to design an algorithm that could be used across the globe with a constrained uncertainty. Figure 8 illustrates the geographical distribution of the error for both LWP and IWV retrievals, using the synthetic radiosonde-based dataset. Two approaches were used to assess this error: first, RMSE values were calculated on the entire set of data available for each location, excluding LWP greater than 1000 $\mathrm{g\,m^{-2}}$. Second, the RMSE was normalized by the mean value of LWP (resp. IWV) for each site, excluding low values (LWP less than 20 $\mathrm{g\,m^{-2}}$, i.e. using a conservative threshold to exclude clear-sky cases). Note that this normalized error is not equal to the relative error; rather, it gives an idea of how large the RMSE of the retrieval is, compared to the mean values that are observed at a given location.

From the non-normalized error (left panels of Fig. 8), it can be seen that most high- and mid-latitude locations have a constrained RMSE around 20-60 $\mathrm{g\,m^{-2}}$, while tropical sites are not as well captured, with RMSE exceeding 120 $\mathrm{g\,m^{-2}}$ in some locations. The temperature and humidity conditions, as well as the strong precipitation events that typically occur in those regions, are probably responsible for this discrepancy. Cases with high LWP are more common under such climatic conditions, and it was observed in Sect. 5.1.2 that the accuracy of the algorithm decreases in that range. Tropical climates are underrepresented in the dataset, for little data is available from this region in comparison with mid-latitude areas: their specificity might therefore not be fully captured during the learning stage of the algorithm. This accounts at least partly for the enhanced error over the Indian peninsula and South-Eastern Asian islands.

The normalized error (right panels of Fig. 8) shows that the error is overall of the same order of magnitude across the globe. However, a few regions stand out from this analysis, which typically feature arid climates: the stations of Dalanzadgad (Mongolia), Salalah (Oman), Minfeng (China, north of Tibet), Jeddah (Saudi Arabia) all have a normalized error on LWP higher than 0.7, and are in the desert. In a similar way, it appears that the IWV retrieval algorithm performs poorly – in terms of normalized error – in cold environments where absolute humidity is low, as in Sermersooq (Greenland). In such regions, the new algorithm is not sensitive enough to capture accurately the fine variations of atmospheric vapor and liquid water content: if detailed studies of those areas were to be conducted, more than one radiometer frequency would likely be necessary, along with specific training sets on which to perform the statistical learning, as was done in the Arctic by Cadeddu et al. (2009).

6 Evaluation in two contrasted datasets

As a further step in the validation process, the algorithm was applied to data from two campaigns involving WProf, first in Payerne, Switzerland, then near Pyeong-Chang, South Korea (see Sect. 2 for the full description of the datasets). In both cases, the output of the retrieval is compared with values retrieved through other methods, either a multi-channel radiometer or – in the case of IWV – radiosonde data.

6.1 Payerne 2017

6.1.1 IWV retrieval

The results of the new IWV retrieval algorithm are compared to those from MeteoSwiss' operational radiometer HATPRO, and to the radiosonde-derived values. From Fig. 9 a) and c) it appears that the IWV retrieval has relatively limited spread but has a constant bias (-1.8 $\mathrm{kg\,m^{-}2}$), which is visible both in the comparison against HATPRO (a) and radiosonde-derived measurements (c). This might be due to a bias in ERA5 data during this timeframe over the region (with a value of -4.1 $\mathrm{kg\,m^{-}2}$), which is visible in ERA5 records during the entire campaign (not shown here) and for which there is no clear explanation at this stage. This bias points to one of the limitations of the IWV retrieval algorithm, which is sensitive not only to radiometer miscalibration but also to possible biases in other input variables, which can be difficult to monitor and assess – as in the case of ERA5 values in Payerne. In spite of this, the top panels of Fig. 10 ~~and 11 shows~~ (which illustrate the error vs. HATPRO measurements) and Fig. 11 (error vs. value derived from radiosounding) show that overall, the implementation of the different versions of the algorithm on the Payerne dataset matches the conclusions from the testing set results: ~~the best retrievalis obtained with the full set~~ more features lead to an enhanced precision of the retrieval. The accuracy drops when only one or two groups of input features ~~, and~~ are included, but no single group of features seem to increase the accuracy alone. There is however a difference between Fig. 10 b) and Fig. 11 b): in the latter, higher R^2 (and similar RMSE) is actually obtained from the algorithm that does not use $\mathrm{T_B}$ in input, than with the ~~greatest improvements result from the addition of ERA5 products, with surface measurements as the next most important category~~ full set of input features. This is at first surprising, but it was explained by taking a closer look at the results: the algorithm without $\mathrm{T_B}$ leads to IWV values which are more smooth and less sensitive to short-time variations. These are not reflected in the comparison against radiosonde data, for which a 30-min averaging was implemented.

~~However, the bias that is seen here points to the limitations of the IWV retrieval algorithm , which is sensitive not only to radiometer miscalibration but also to possible biases in other input variables, which can be difficult to monitor and assess – as in the case of ERA5 values in Payerne~~When variations over a small timeframe are considered, the inclusion of $\mathrm{T_B}$ improves the retrieval, as comes across from the comparison against HATPRO's measurements in Fig. 10.

6.1.2 LWP retrieval

Figure 9 b) shows that LWP values retrieved with the new algorithm are in general agreement with those obtained thanks to HATPRO, although a larger spread is observed than in the IWV retrieval. A saturation effect can be seen near precipitation onset, when LWP values from HATPRO reach 600 $\mathrm{g\,m^{-2}}$. Additionally, outliers are visible as vertical and horizontal bars close to the axes, for which two hypotheses are considered. One is that the distance between the two instruments was big enough, that in some cases a liquid water cloud would overpass one of the two instruments but not the other. Hence, HATPRO would measure a non-zero LWP while WProf would indicate a clear sky, or vice-versa. Also, measurement artefacts cannot be excluded, e.g. due the persistence of a liquid water film on the radome of either radiometer, after precipitation or due to condensation.

For comparison, the method described in Küchler et al. (2017) was implemented (further on referred to as K17), by performing a quadratic regression on a dataset consisting solely of radiosonde profiles collected in Payerne. As proposed by the authors, a first version (K17A) relies on a measurement vector consisting of $\mathrm{T_B}$, $\mathrm{T_B^2}$, as well as the IWV estimate from reanalysis data $\mathrm{IWV_{ERA5}}$ and $\mathrm{IWV_{ERA5}^2}$. Another version (K17B) includes only $\mathrm{T_B}$ and $\mathrm{T_B^2}$. Theoretical RMSEs derived for those quadratic regressions on the synthetic dataset (19 720 profiles) are 21 $\mathrm{g\,m^{-2}}$ and 43 $\mathrm{g\,m^{-2}}$, respectively, which is similar to the values obtained by the authors on radiosonde data from De Bilt (the Netherlands), i.e. 15 $\mathrm{g\,m^{-2}}$ and 44 $\mathrm{g\,m^{-2}}$.

K17A and K17B were applied to Payerne campaign dataset, and their results are compared to those from the new algorithm in Fig. 10. The error metrics are calculated using HATPRO's values as a reference. The algorithms perform in a similar way, with slightly better results for the new algorithm when at least one of the secondary input features is included. We remind that K17A and K17B were specifically tuned on Payerne data, while the new algorithm was tuned globally, on a dataset that did not ~~include~~ comprise radiosonde profiles from Payerne.

6.2 ICE-POP 2018

As detailed in Sect. 2, the South Korean deployment of WProf in 2017-2018 also offers an opportunity to compare results from the IWV retrieval to IWV from radiosonde measurements.

The analysis of the $\mathrm{T_B}$ timeseries showed that a miscalibration of the radiometer led to unrealistic – negative – values for which a correction had to be implemented, through the addition of a constant offset to $\mathrm{T_B}$ measurements. The value of this offset (~~18~~ 20 K) was determined by computing theoretical brightness temperatures from clear-sky radiosonde profiles and comparing them to measured ~~TBs~~$\mathrm{T_B}$s, following the approach of Ebell et al. (2017). This is however only a first-order correction whose output should be taken with care, especially after the analysis of Sect. 5.2 which underlined the importance of $\mathrm{T_B}$ accuracy for IWV retrieval.

After this correction, the IWV retrieval gives coherent results (see Fig. 12), with a total RMSE that is slightly lower than that obtained on the testing data set (1.2 $\mathrm{kg\,m^{-2}}$~~at best). Here again, the~~). The best results are found when several input features are included and drop severely when no secondary input features are used~~. From Fig. 12 it also comes through~~ , which corresponds to the results on the synthetic data set presented in Sect. 5. Because of the relatively low sensitivity of $\mathrm{T_B}$ at 89 GHz to water

vapor, the algorithm largely relies on non-radiometric features; this is even more the case in cold and dry environments like that of ICE-POP, where IWV is low. In fact, slightly better results are obtained with all input features except brightness temperature. The miscalibration of the radiometer, which may not have been perfectly corrected by the addition of a constant offset, might emphasize this error. This also corresponds to what was noted in Payerne: when the results are averaged over 30 minutes, brightness temperature brings little, if any, improvement to the results. T_B is relevant when a higher temporal resolution is considered (cf. Sect. 6.1.1) – for which no comparison was available during ICE-POP – or when ERA5 data is significantly off. In this case however, it comes across from Fig. 12 that the algorithm is consistently outperformed by ERA5 products~~, which~~ : they have both a lower RMSE and a higher R^2, which makes the algorithm less relevant for the study of this specific campaign. The high accuracy of ERA5 data during ICE-POP also explains the high correlation coefficient of the retrieval that uses ERA5 and Geographical input features: since the geographical parameters are constant, the temporal variability is that of the reananalysis data and therefore the correlation coefficient of the retrieval is close to that of ERA5 data alone. Let us highlight that although reanalysis data outperforms the retrieval for ICE-POP, this was not the case in Payerne nor in the ~~radiosonde database~~full radiosonde data set, where the algorithm has a ~~greater~~ higher accuracy than ERA5 values. Possibly, the dry and cold weather that was observed during the ICE-POP campaign ~~(Gehring et al., 2020)~~ featured little short-term variability and ~~were~~ was associated with stable atmospheric conditions that were particularly well captured in ERA5 reanalyses. Snowfall events during the campaign, as well as occasional fog, can also bias the retrieval by enhancing brightness temperature.

~~Bringing the~~ The analysis of the ICE-POP data ~~set~~ was taken a step further ~~, it~~ to explore the latter point. It appears that the IWV retrieval is most reliable in non-precipitating or cold conditions, i.e. when little liquid water is expected in the column. ~~Periods~~ To visualize this, periods with no precipitation or fog are identified using WProf's radar measurements as time steps with low radar equivalent reflectivity (Ze < -10 ~~dB~~dBZ) in the lower gates (first kilometer above the radar), and temperature time series are provided by the weather station coupled to WProf. ~~It appears that the algorithm~~ Fig. 13 shows the scatter plot of the error – for the algorithm that includes all input features – color-coded in a way to differentiate dry from precipitating or fog conditions: black triangles correspond to dry timesteps, and circles to timesteps with Ze > -10 dBZ, with their color indicating surface temperature. The algorithm yields a larger bias in rain ~~, –~~ as was expected in the design steps of the algorithm (Sect. ~~3) ,~~ 3) – but also during snow events with relatively warm temperatures, close to or slightly above 0 °C (Fig. 13). Changes in the dielectric properties of snowflakes during the melting process can explain this increased error; additionally, the process described by Kneifel et al. (2010) and which was recalled in Sect. 3 suggests that snowfall events with large snow particles (typically present with relatively mild temperatures) could have a non-negligible contribution to brightness temperature, which might explain the enhanced error in those cases.

7 Summary and conclusions

A new site-independent method was designed for the retrieval of LWP and IWV from a single-channel ground-based radiometer. In addition to 89-GHz brightness temperature, additional input features were used for the retrieval, such as surface

atmospheric variables (temperature, pressure and humidity) and information on the geographical location and season. A neural network architecture was chosen for the statistical learning.

Training and testing were performed on a synthetic data base that was built using radiosonde profiles worldwide. The geographical distribution of the error shows that the algorithm performs better in mid-latitudes and regions with a moderate climate than in areas with a extreme climates – either arid or very moist – which include both tropical and polar regions, that are not well represented in the training dataset due to lack of available data. Also, the forward model that was used should most likely be revised in order to capture finely the atmospheric conditions in such specific environments. In addition, the training dataset lacks data from locations with complex orography, and more in-depth investigations should be conducted regarding the reliability of the retrieval in such terrain (Massaro et al., 2015).

The algorithm was then applied to two contrasted data sets, one reflecting summertime weather conditions in Switzerland, and the other in winter conditions in South Korea. For this application, measurements from RPG's cloud radar-radiometer system were used.

In Payerne, the new LWP retrieval was found to perform slightly better than the algorithm proposed by Küchler et al. (2017) for the same instrument, which was specifically trained using radiosonde data from Payerne. When compared to radiosonde measurements of IWV, the IWV retrieval was found to be less accurate than that of a state-of-the-art multi-channel radiometer (HATPRO), although both instruments yield errors within the same order of magnitude. In the South-Korean winter dataset, the IWV retrieval proved ~~in general quite robust, but~~ relatively robust, in spite of a slight bias ~~was present~~ during some snowfall events, that could be related to the scattering properties of snow particles, which were not taken into account in the forward model. In the case of ICE-POP, reanalysis data was actuallly more accurate than the IWV retrieval when compared with radiosonde measurements, but its temporal resolution remains low, which makes the use of the algorithm still relevant for retrievals with high temporal resolution.

Further steps in the improvement of the current algorithm would include coupling information from the radar and the radiometer channel (Ebell et al., 2010; Cadeddu et al., 2020). The detection of clear-sky cases with radar data (Mätzler and Morland, 2009) could help monitor the calibration of the radiometer, and introduce T_B offsets for correction when necessary (Ebell et al., 2017). Radar moments could be used to distinguish cloudy from drizzling or rainy cases, similar to the approach conducted by Cadeddu et al. (2020), and to use appropriate DSDs for each case to account for non-Rayleigh scattering by precipitating droplets. However, forward modeling of radar data requires further assumptions on microphysical properties and atmospheric conditions, for which generalization to a global geographical scale is a real challenge. Additionally, the retrieval that is presented here uses reanalysis data as an optional feature, which shown to be valuable. In the case where near real-time retrievals were necessary, the user could choose to use a version of the algorithm that does not rely on ERA5, but this would be detrimental, especially for the IWV retrieval. Another option would be to implement the algorithm with the output of forecast models (IWV and LWP) instead of reanalysis data. This approach was however not tested as this stage.

Overall, the LWP and IWV retrieval methods that were designed within this study were shown to ~~perform reasonably well~~be robust, both when applied to synthetic and to real datasets, although their performance is inevitably lower than that of multi-

channel radiometers specifically designed for LWP and IWV retrieval. While retrieving IWV based on T_B at 89 GHz alone does not lead to accurate results – because of the lower sensitivity of this channel to water vapor emission, compared to liquid water – this study showed that reliable retrievals could be achieved by including surface and geographical information, as well as reanalysis data if available, among the input features. The new algorithms should be seen as a valuable tool for atmospheric liquid water and vapor monitoring in the context of radar-radiometer studies. They are non-site specific, and thus do not require further tuning before use on a new site, which makes them easy to implement, while their accuracy is well characterized.

Code availability. The code developed in this study can be made publicly available upon request to the authors.

Author contributions. ACB and AB designed and conducted the study. ACB processed the data and run the analyses. ACB prepared the manuscript with contributions and supervision from AB.

Competing interests. The authors declare that they have no competing interests.

Acknowledgements. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824310. Additionally, the authors are greatly appreciative to the participants of the World Weather Research Programme Research and Development and Forecast Demonstration Project International Collaborative Experiments for Pyeongchang 2018 Olympic and Paralympic Winter Games (ICE-POP 2018), hosted by the Korea Meteorological Administration. In particular, we are thankful to the High Impact Weather Research Center of the Korea Meteorological Administration for providing us with radiosonde data from Daegwallyeong, South Korea. We are grateful to Kwonil Kim and GyuWon Lee from Kyungpook National University for their help in the installation of the W-band cloud radar on site. We would further like to thank Maxime Hervo and Giovanni Martucci from MeteoSwiss for providing us with radiometer and radiosonde data from Payerne, Switzerland. ~~Finally, special~~ Special thanks go to Josué Gehring, Alfonso Ferrone and Christophe Praz for the deployment of the instruments on the field. Finally, we are thankful to the three anonymous reviewers whose comments helped improve this study.

References

Cadeddu, M., Liljegren, J., and Turner, D.: The Atmospheric radiation measurement (ARM) program network of microwave radiometers: instrumentation, data, and retrievals, Atmospheric Measurement Techniques, 6, 2359–2372, https://doi.org/10.5194/amt-6-2359-2013, 2013.

Cadeddu, M. P., Turner, D. D., and Liljegren, J. C.: A Neural Network for Real-Time Retrievals of PWV and LWP From Arctic Millimeter-Wave Ground-Based Observations, IEEE Transactions on Geoscience and Remote Sensing, 47, 1887–1900, 2009.

Cadeddu, M. P., Marchand, R., Orlandi, E., Turner, D. D., and Mech, M.: Microwave Passive Ground-Based Retrievals of Cloud and Rain Liquid Water Path in Drizzling Clouds: Challenges and Possibilities, IEEE Transactions on Geoscience and Remote Sensing, 55, 6468–6481, https://doi.org/10.1109/TGRS.2017.2728699, 2017.

Cadeddu, M. P., Ghate, V. P., and Mech, M.: Ground-based observations of cloud and drizzle liquid water path in stratocumulus clouds., Atmospheric Measurement Techniques, 13, https://doi.org/10.5194/amt-13-1485-2020, 2020.

Chollet, F.: Deep learning with Python, 2017.

Chollet, F. et al.: Keras, https://keras.io, 2015.

Copernicus Climate Change Service: ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate, accessed on 2019-05-15, https://cds.climate.copernicus.eu, 2020.

Ebell, K., Löhnert, U., Crewell, S., and Turner, D. D.: On characterizing the error in a remotely sensed liquid water content profile, Atmospheric Research, 98, 57 – 68, https://doi.org/10.1016/j.atmosres.2010.06.002, 2010.

Ebell, K., Löhnert, U., Päschke, E., Orlandi, E., Schween, J. H., and Crewell, S.: A 1-D variational retrieval of temperature, humidity, and liquid cloud properties: Performance under idealized and real conditions, Journal of Geophysical Research: Atmospheres, 122, 1746–1766, https://doi.org/10.1002/2016JD025945, 2017.

Gaussiat, N., Hogan, R. J., and Illingworth, A. J.: Accurate Liquid Water Path Retrieval from Low-Cost Microwave Radiometers Using Additional Information from a Lidar Ceilometer and Operational Forecast Models, Journal of Atmospheric and Oceanic Technology, 24, 1562–1575, https://doi.org/10.1175/JTECH2053.1, 2007.

Gehring, J., Ferrone, A., Billault-Roux, A.-C., Besic, N., Ahn, K. D., Lee, G., and Berne, A.: Radar and ground-level measurements of precipitation collected by EPFL during the ICE-POP 2018 campaign in South-Korea, Earth System Science Data Discussions, 2020, 1–29, https://doi.org/10.5194/essd-2020-134, 2020.

Hartmann, D. L. and Short, D. A.: On the Use of Earth Radiation Budget Statistics for Studies of Clouds and Climate, Journal of the Atmospheric Sciences, 37, 1233–1250, https://doi.org/10.1175/1520-0469(1980)037<1233:OTUOER>2.0.CO;2, 1980.

Hartmann, D. L., Ockert-Bell, M. E., and Michelsen, M. L.: The Effect of Cloud Type on Earth's Energy Balance: Global Analysis, Journal of Climate, 5, 1281–1304, https://doi.org/10.1175/1520-0442(1992)005<1281:TEOCTO>2.0.CO;2, 1992.

Karstens, U., Simmer, C., and Ruprecht, E.: Remote sensing of cloud liquid water, Meteorology and Atmospheric Physics, 54, 157–171, https://doi.org/10.1007/BF01030057, 1994.

Kneifel, S., Löhnert, U., Battaglia, A., Crewell, S., and Siebler, D.: Snow scattering signals in ground-based passive microwave radiometer measurements, Journal of Geophysical Research: Atmospheres, 115, https://doi.org/10.1029/2010JD013856, 2010.

Küchler, N., Kneifel, S., Löhnert, U., Kollias, P., Czekala, H., and Rose, T.: A W-Band Radar–Radiometer System for Accurate and Continuous Monitoring of Clouds and Precipitation, Journal of Atmospheric and Oceanic Technology, 34, 2375–2392, https://doi.org/10.1175/JTECH-D-17-0019.1, 2017.

Liljegren, J. C., Clothiaux, E. E., Mace, G. G., Kato, S., and Dong, X.: A new retrieval for cloud liquid water path using a ground-based microwave radiometer and measurements of cloud temperature, Journal of Geophysical Research: Atmospheres, 106, 14 485–14 500, https://doi.org/10.1029/2000JD900817, 2001.

Liljegren, J. C., Boukabara, S. ., Cady-Pereira, K., and Clough, S. A.: The effect of the half-width of the 22-GHz water vapor line on retrievals of temperature and water vapor profiles with a 12-channel microwave radiometer, IEEE Transactions on Geoscience and Remote Sensing, 43, 1102–1108, https://doi.org/10.1109/TGRS.2004.839593, 2005.

Löhnert, U. and Crewell, S.: Accuracy of cloud liquid water path from ground-based microwave radiometry 1. Dependency on cloud model statistics, Radio Science, 38, https://doi.org/10.1029/2002RS002654, 2003.

Löhnert, U. and Maier, O.: Operational profiling of temperature using ground-based microwave radiometry at Payerne: prospects and challenges, Atmospheric Measurement Techniques, 5, 1121–1134, https://doi.org/10.5194/amt-5-1121-2012, 2012.

Löhnert, U., Crewell, S., and Simmer, C.: An Integrated Approach toward Retrieving Physically Consistent Profiles of Temperature, Humidity, and Cloud Liquid Water, Journal of Applied Meteorology, 43, 1295–1307, https://doi.org/10.1175/1520-0450(2004)043<1295:AIATRP>2.0.CO;2, 2004.

Maahn, M.: Exploiting vertically pointing Doppler radar for advancing snow and ice cloud observations, Ph.D. thesis, Universität zu Köln, 2015.

Mace, G. G., Benson, S., Sonntag, K. L., Kato, S., Min, Q., Minnis, P., Twohy, C. H., Poellot, M., Dong, X., Long, C., Zhang, Q., and Doelling, D. R.: Cloud radiative forcing at the Atmospheric Radiation Measurement Program Climate Research Facility: 1. Technique, validation, and comparison to satellite-derived diagnostic quantities, Journal of Geophysical Research: Atmospheres, 111, https://doi.org/10.1029/2005JD005921, 2006.

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, International Journal of Remote Sensing, 23, 661–674, https://doi.org/10.1080/01431160110045959, 2002.

Massaro, G., Stiperski, I., Pospichal, B., and Rotach, M.: Accuracy of retrieving temperature and humidity profiles by ground-based microwave radiometry in truly complex terrain, Atmospheric Measurement Techniques, 8, 3355, https://doi.org/10.5194/amt-8-3355-2015, 2015.

Mattioli, V., Basili, P., Bonafoni, S., Ciotti, P., and Westwater, E. R.: Analysis and improvements of cloud models for propagation studies, Radio Science, 44, 1–13, 2009.

Mätzler, C. and Morland, J.: Refined Physical Retrieval of Integrated Water Vapor and Cloud Liquid for Microwave Radiometer Data, IEEE Transactions on Geoscience and Remote Sensing, 47, 1585–1594, https://doi.org/10.1109/TGRS.2008.2006984, 2009.

McFarlane, S. A., Mather, J. H., Ackerman, T. P., and Liu, Z.: Effect of clouds on the calculated vertical distribution of shortwave absorption in the tropics, Journal of Geophysical Research: Atmospheres, 113, https://doi.org/10.1029/2008JD009791, 2008.

Mech, M., Maahn, M., Kneifel, S., Ori, D., Orlandi, E., Kollias, P., Schemann, V., and Crewell, S.: PAMTRA 1.0: A Passive and Active Microwave radiative TRAnsfer tool for simulating radiometer and radar measurements of the cloudy atmosphere, Geoscientific Model Development Discussions, 2020, 1–34, https://doi.org/10.5194/gmd-2019-356, 2020.

Oolman, L.: Atmospheric Soundings, http://weather.uwyo.edu/upperair/sounding.html, 2020.

Pierdicca, N., Pulvirenti, L., and Marzano, F. S.: A model to predict cloud density from midlatitude atmospheric soundings for microwave radiative transfer applications, Radio Science, 41, 1–12, 2006.

Poulida, O., Schwikowski, M., Baltensperger, U., Staehelin, J., and Gaeggeler, H.: Scavenging of atmospheric constituents in mixed phase clouds at the high-alpine site jungfraujoch—part II. Influence of riming on the scavenging of particulate and gaseous chemical species, Atmospheric Environment, 32, 3985 – 4000, https://doi.org/10.1016/S1352-2310(98)00131-9, 1998.

Rose, T., Crewell, S., Löhnert, U., and Simmer, C.: A network suitable microwave radiometer for operational monitoring of the cloudy atmosphere, Atmospheric Research, 75, 183 – 200, https://doi.org/10.1016/j.atmosres.2004.12.005, 2005.

Rosenkranz, P. W.: Water vapor microwave continuum absorption: A comparison of measurements and models, Radio Science, 33, 919–928, https://doi.org/https://doi.org/10.1029/98RS01182, 1998.

RPG Radiometer Physics GmbH: Humidity And Temperature PROfilers, https://www.radiometer-physics.de/products/microwave-remote-sensing-instruments/radiometers/humidity-and-temperature-profilers/#tabs-container-0, 2014.

Saleeby, S. M., Cotton, W. R., and Fuller, J. D.: The Cumulative Impact of Cloud Droplet Nucleating Aerosols on Orographic Snowfall in Colorado, Journal of Applied Meteorology and Climatology, 50, 604–625, https://doi.org/10.1175/2010JAMC2594.1, 2011.

Salonen, E. and Uppala, S.: New prediction method of cloud attenuation, Electronics Letters, 27, 1106–1108, 1991.

Slingo, A.: Sensitivity of the Earth's radiation budget to changes in low clouds, Nature, 343, 49–51, https://doi.org/10.1038/343049a0, 1990.

Stephens, G. L.: Cloud Feedbacks in the Climate System: A Critical Review, Journal of Climate, 18, 237–273, https://doi.org/10.1175/JCLI-3243.1, 2005.

Turner, D. D., Clough, S. A., Liljegren, J. C., Clothiaux, E. E., Cady-Pereira, K. E., and Gaustad, K. L.: Retrieving Liquid Wat0er Path and Precipitable Water Vapor From the Atmospheric Radiation Measurement (ARM) Microwave Radiometers, IEEE Transactions on Geoscience and Remote Sensing, 45, 3680–3690, https://doi.org/10.1109/TGRS.2007.903703, 2007.

Turner, D. D., Cadeddu, M. P., Lohnert, U., Crewell, S., and Vogelmann, A. M.: Modifications to the Water Vapor Continuum in the Microwave Suggested by Ground-Based 150-GHz Observations, IEEE Transactions on Geoscience and Remote Sensing, 47, 3326–3337, https://doi.org/10.1109/TGRS.2009.2022262, 2009.

Wang, J. and Rossow, W. B.: Effects of Cloud Vertical Structure on Atmospheric Circulation in the GISS GCM, Journal of Climate, 11, 3010–3029, https://doi.org/10.1175/1520-0442(1998)011<3010:EOCVSO>2.0.CO;2, 1998.

Westwater, E. R., Han, Y., Shupe, M. D., and Matrosov, S. Y.: Analysis of integrated cloud liquid and precipitable water vapor retrievals from microwave radiometers during the Surface Heat Budget of the Arctic Ocean project, Journal of Geophysical Research: Atmospheres, 106, 32 019–32 030, https://doi.org/10.1029/2000JD000055, 2001.

Zhang, G., Vivekanandan, J., and Politovich, M.: SCATTERING EFFECTS ON MICROWAVE PASSIVE REMOTE SENSING OF CLOUD PARAMETERS, in: The 8th Conference on Aviation, Range, and Aerospace Meteorology, pp. 497–501, 1999.

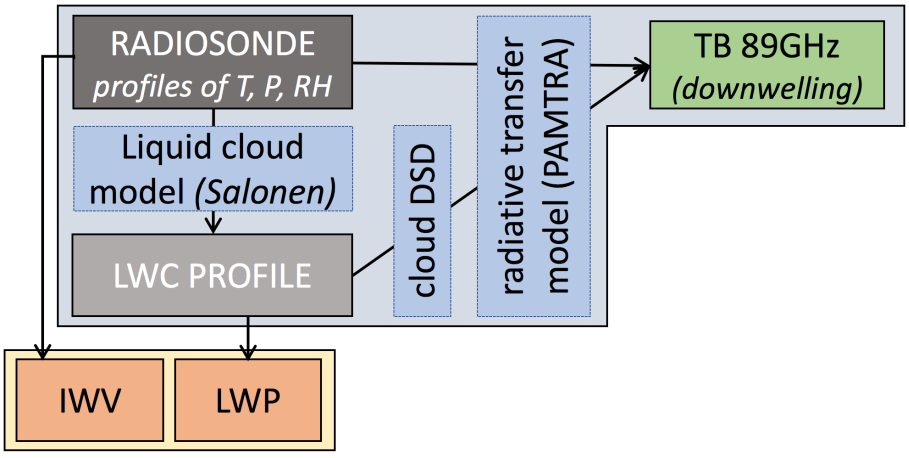

Figure 1. Illustration of the different steps of the forward model.

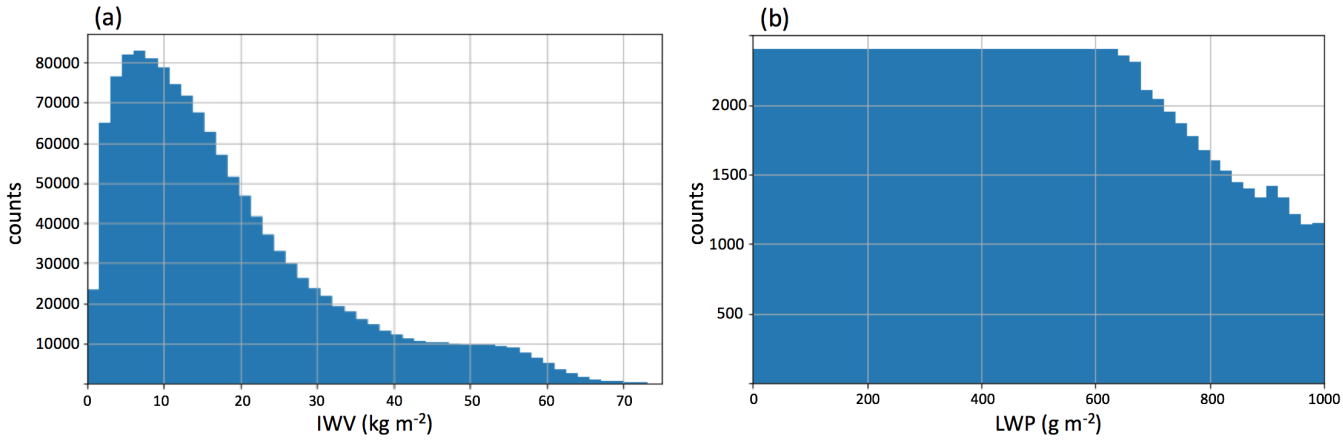

Figure 2. Distribution of the target variables (IWV and LWP, resp. in panels (a) and (b)) in the synthetic dataset, after preprocessing.

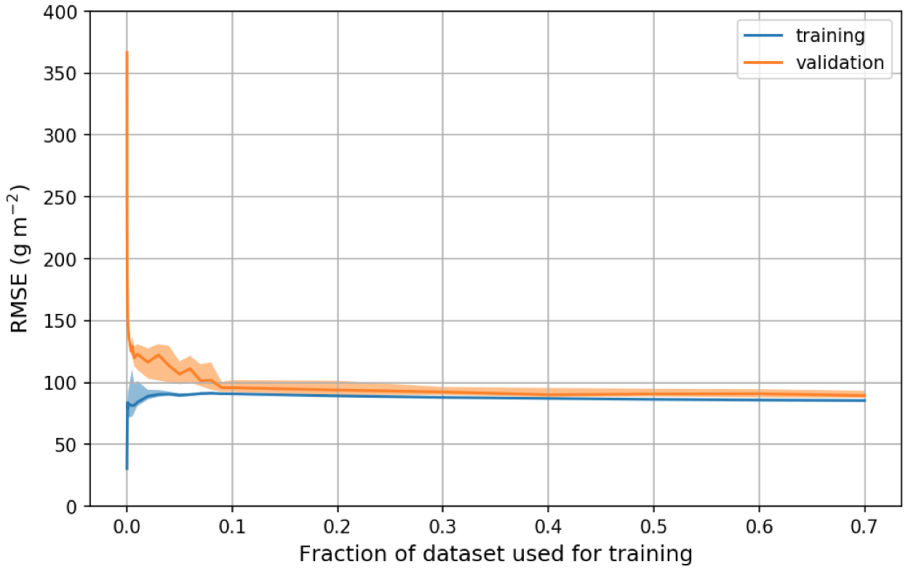

Figure 3. Learning curves for the LWP retrieval, showing the RMSE on training and validation set with varying training set size. Shaded areas correspond to the interquartile range calculated over 50 realizations of random splitting of the dataset into training and validation sets, bold lines are the median.

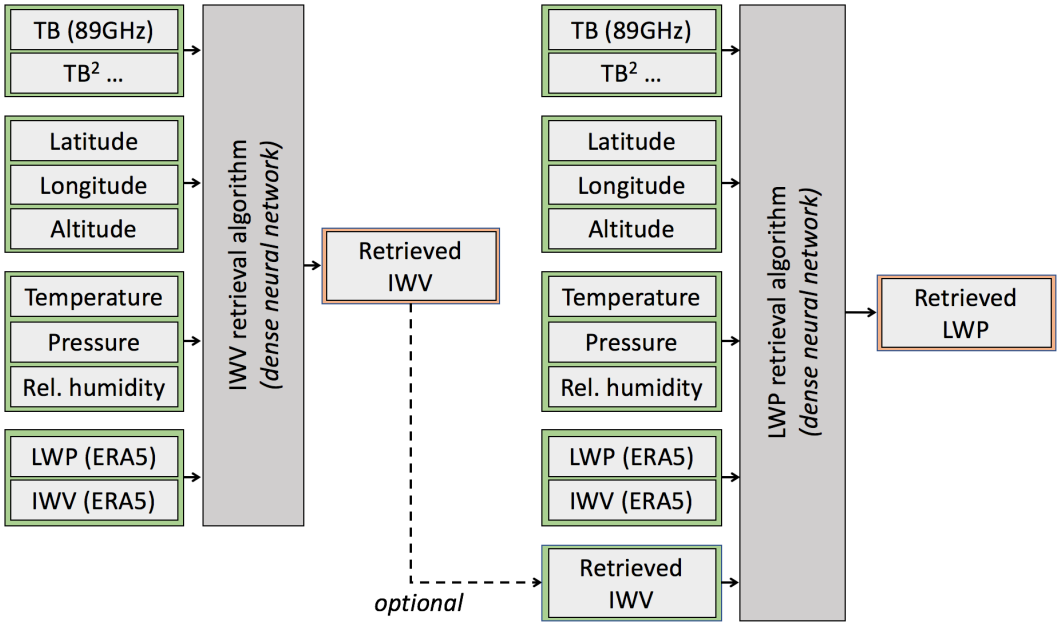

Figure 4. Structure of the retrieval algorithms. Some versions of the LWP retrieval include, among the input features, the output of the IWV retrieval. Note that the IWV and LWP algorithms are trained on different datasets.

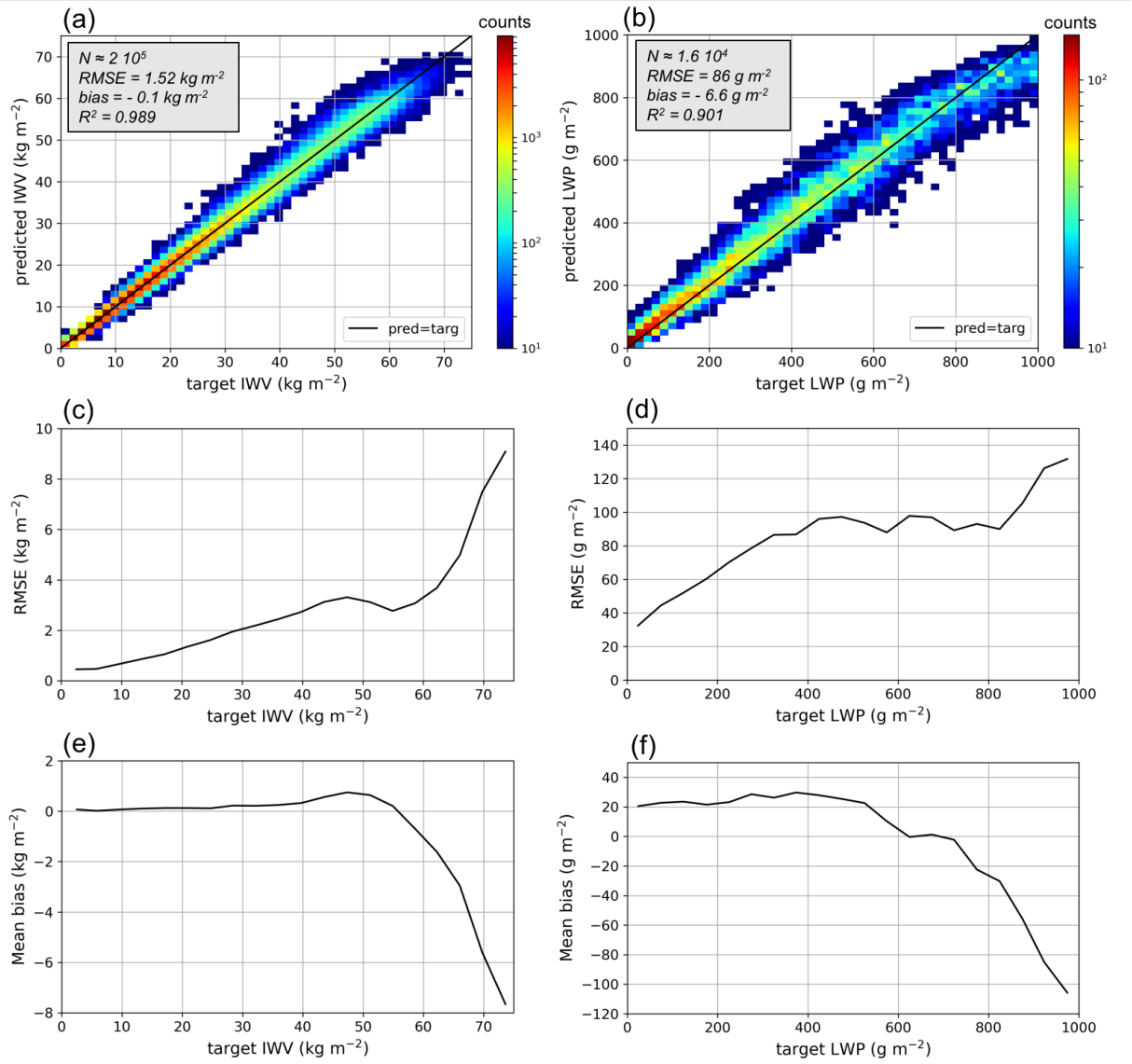

Figure 5. Results of the retrieval algorithms on the synthetic testing dataset. The best versions of the algorithms are presented, i.e. the ones which use the full set of input features. Panels (a) and (b) show the distribution of predicted vs. target values of resp. IWV and LWP. The size of the testing set is indicated (N) as well as relevant error metrics (RMSE, mean bias, R^2). Panels (c) and (d) illustrate the distribution of the RMSE across the range of IWV and LWP values, binned into intervals of resp. 4 kg m^{-2} and 50 g m^{-2}. Similarly, panels (e) and (f) show the distribution of mean bias across the range of IWV and LWP values.

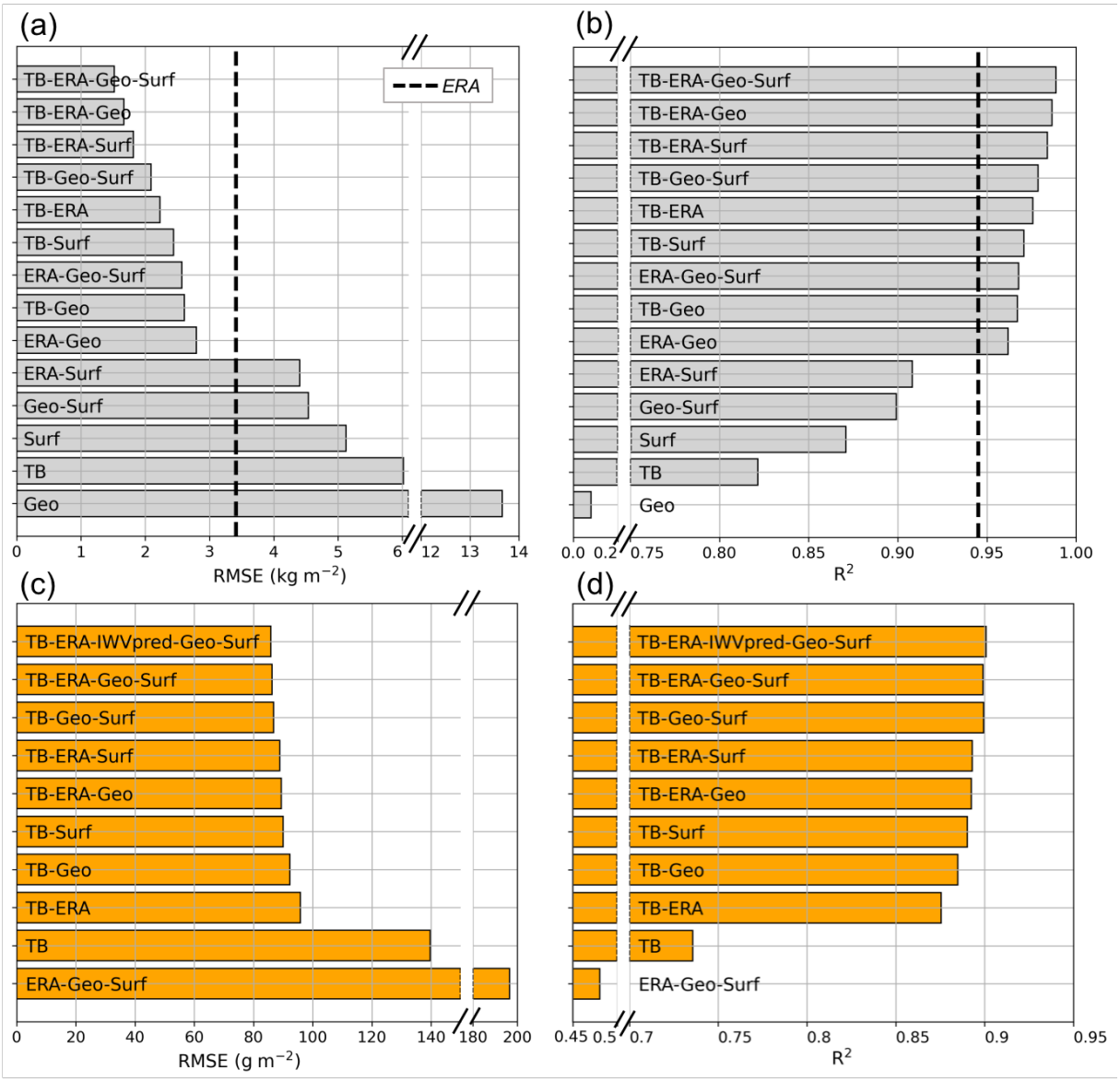

Figure 6. Global error metrics (RMSE on the left panels and correlation coefficient ~~R2~~ R^2 on the right) computed on the testing ~~(grey) and training (orange) sets~~ set for different versions of the (a,b) IWV and (c,d) LWP retrievals. Each bar shows the result of a version whose input features are specified in the label. For example, ~~"~~"ERA-IWVpred-Geo-Surf~~"~~" corresponds to the version of the LWP retrieval algorithm that uses the following categories of input features: ERA5 variables, IWV obtained from the IWV retrieval, geographical information and surface measurements. The bars are sorted with increasing RMSE. For the IWV retrieval, the accuracy of the algorithm is compared to that of reanalysis data alone (dashed lines).

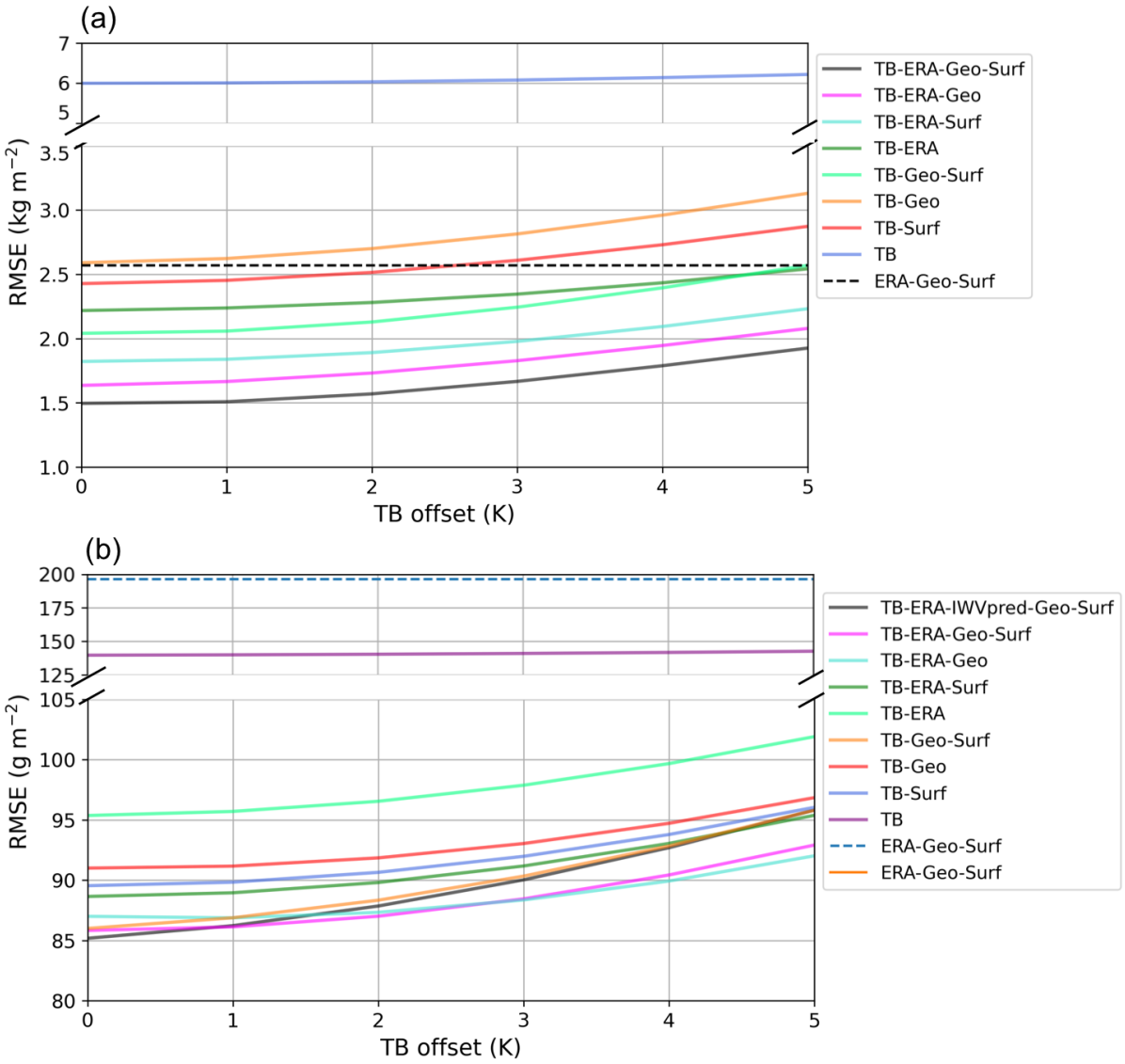

Figure 7. RMSE on testing set of the different versions of the (a) IWV and (b) LWP retrieval, after addition of a constant T_B offset in the input. Dashed lines ~~include~~ show the ~~prediction of IWV~~ retrievals without T_B in the input features.

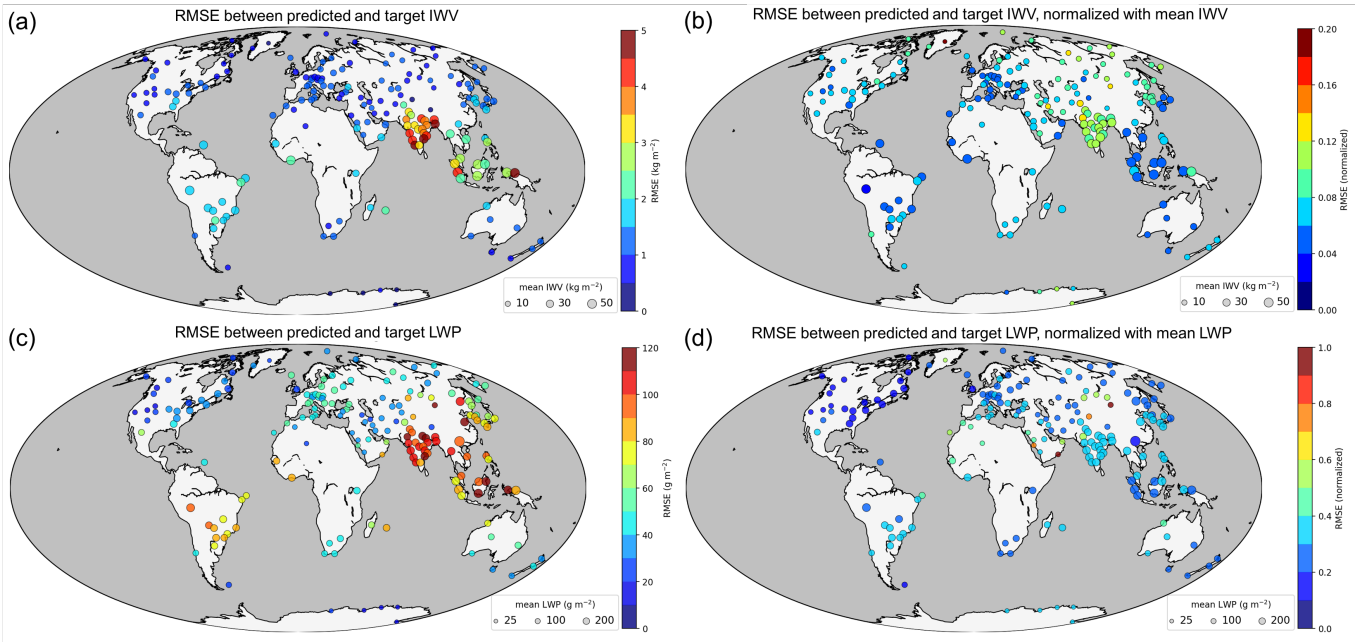

Figure 8. Geographical distribution of the error on the synthetic dataset. Left panels (a) and (c) illustrate the total RMSE on resp. IWV and LWP. In panels (b) and (d) is shown the normalized error, i.e. the RMSE normalized by the mean value of IWV (resp. LWP) at each location. For the evaluation of LWP, clear-sky as well as strong rainy cases are removed (resp. LWP < 20 g m^{-2} and LWP > 1000 g m^{-2}). The size of the disks represents the mean value of IWV or LWP at each site, while the color codes for the error of the retrieval.

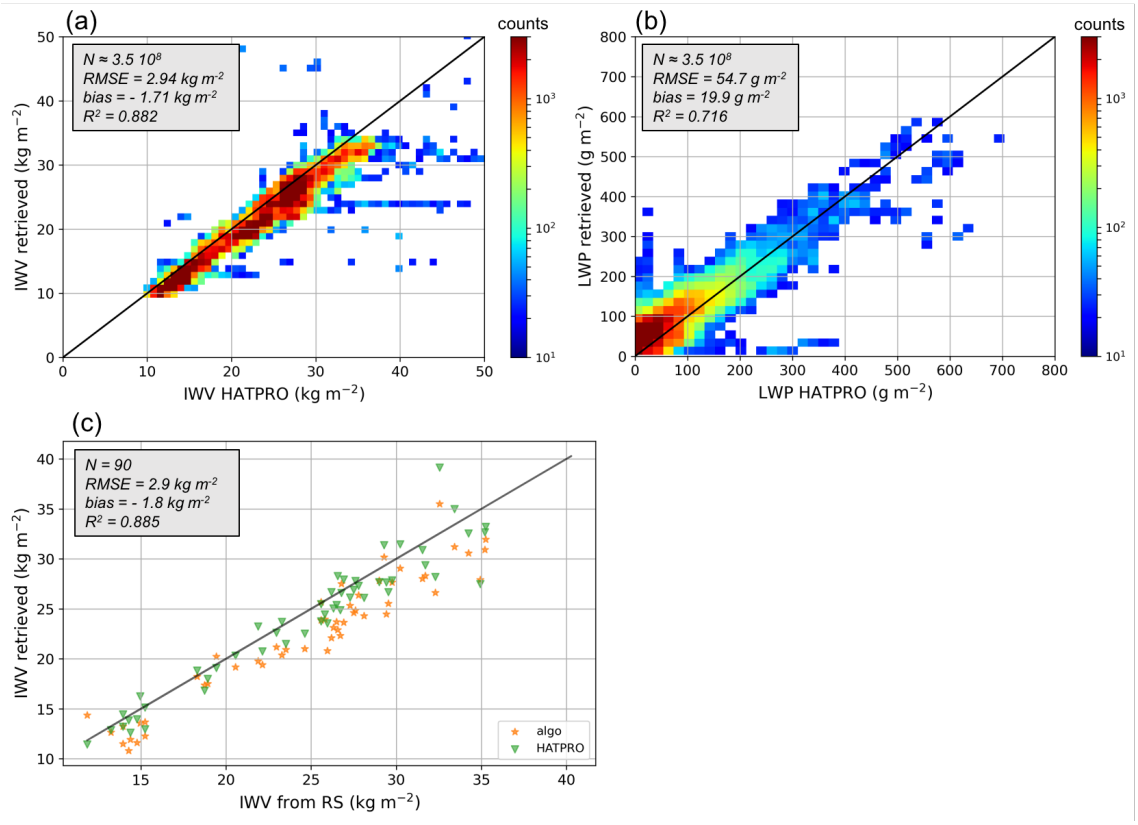

Figure 9. Comparison of (a) IWV and (b) LWP retrieved over Payerne with the new algorithm, using the full set of input features, against the retrieval from MeteoSwiss' radiometer HATPRO. Panel (c) shows IWV retrieved from the new algorithm and from HATPRO against the from radiosonde measurements; a 30 minute time averaging is used for radiometer measurements. The size of the data set is indicated (N) as well as relevant error metrics (RMSE, mean bias, R^2).

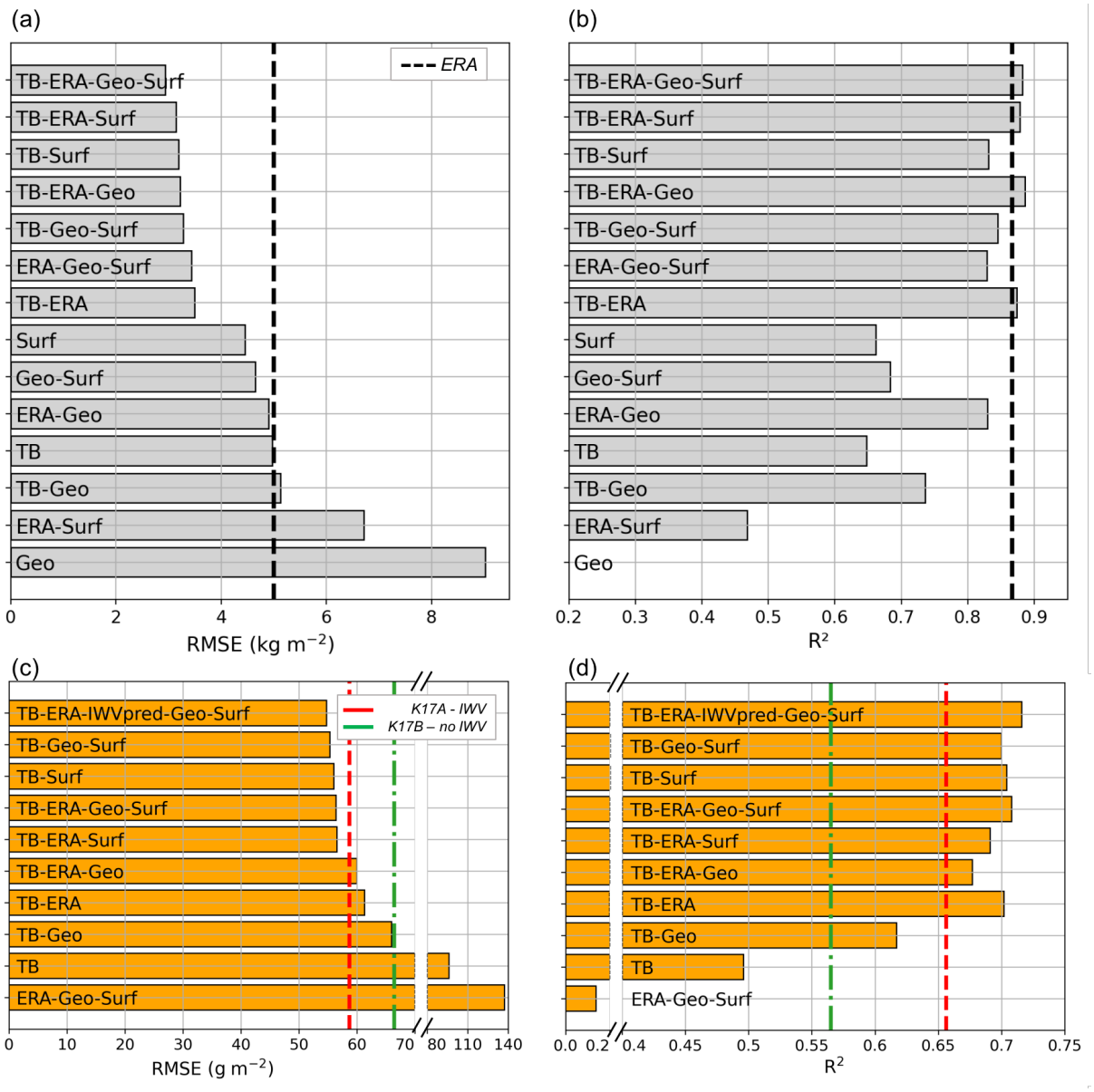

Figure 10. Error of the new retrieval algorithms over Payerne compared to HATPRO retrievals. In panels (a, c) the RMSE of resp. IWV and LWP is calculated ~~in comparison with HATPRO retrievals~~for different versions of the algorithm. Similarly, R^2 is shown in panels (b, d). Each bar shows the result of a version whose input features are specified in the label. In panels (a, b), the black dashed line shows the error of IWV from ERA5 reanalysis data. In (c, d), the dashed lines present the results from K17A and K17B, as defined in the body text.

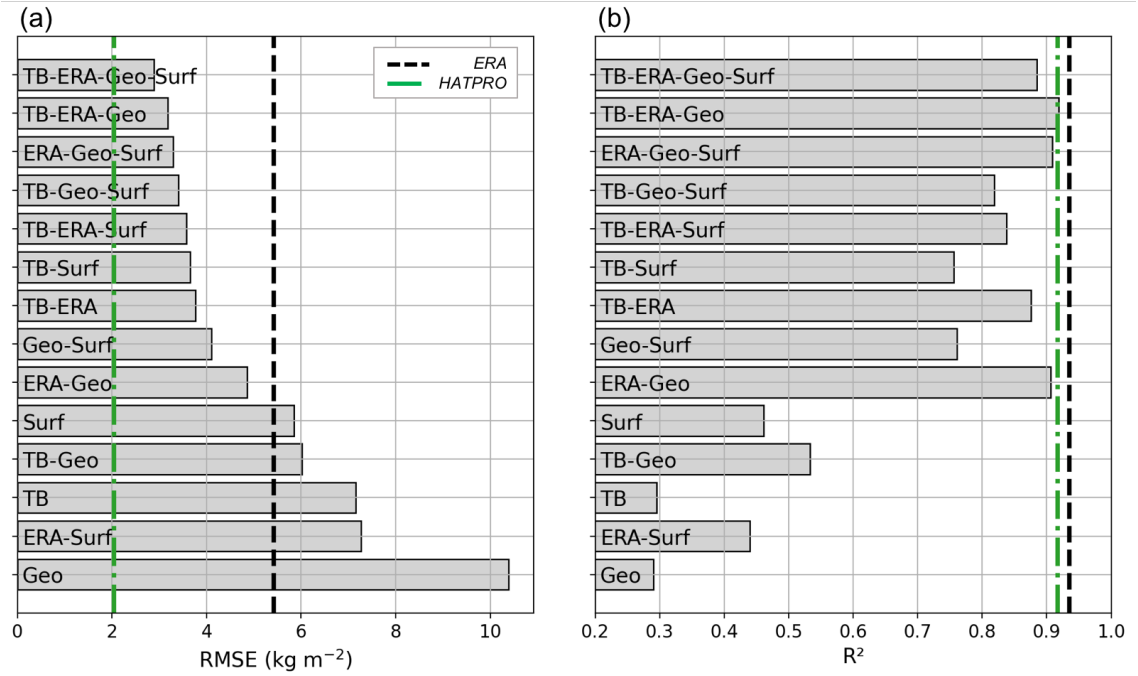

Figure 11. Results of the IWV retrieval in Payerne compared to radiosonde measurements (a: RMSE and b: R^2)~~compared to radiosonde measurements~~. The radiometer measurements are averaged over 30 minutes. For comparison, the dashed lines illustrate the error of HATPRO (green) and ERA5 (black) vs. radiosonde-derived IWV.

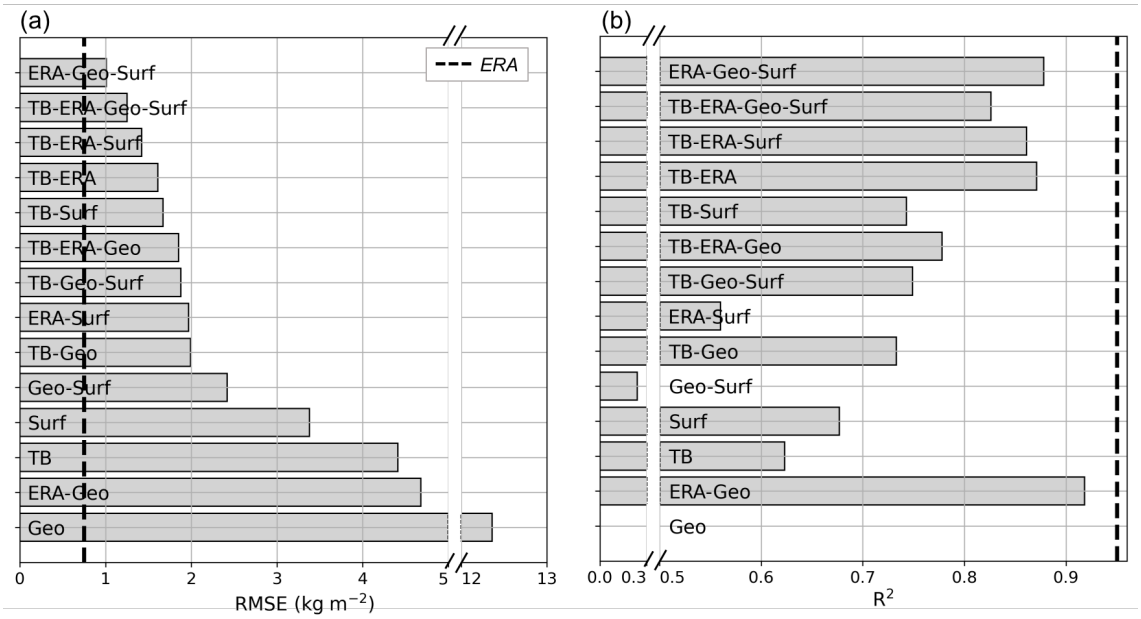

Figure 12. Error of the IWV retrieval during the ICE-POP campaign with different versions of the algorithm. The RMSE is computed against IWV from radiosonde profiles, after 30 minutes of temporal averaging in the radiometer data. The dashed line shows the error of IWV from ERA5 reanalysis data.

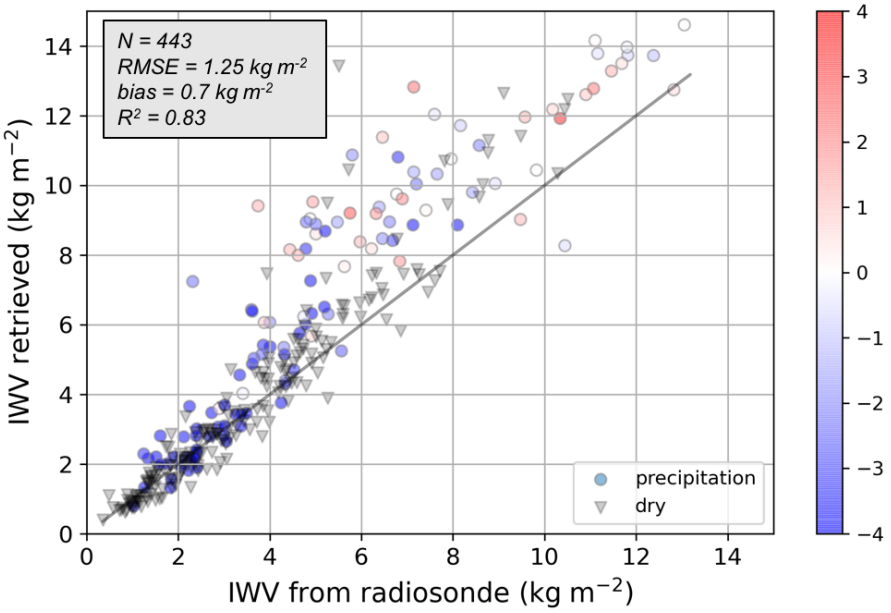

Figure 13. Scatter plot of retrieved IWV vs. IWV computed from radiosonde profiles. The algorithm used for the retrieval is the one with the full set of input features. The color indicates the surface temperature (in degrees Celsius). Dry conditions are identified with the equivalent radar reflectivity in the first kilometer above the radar (with a -10 dB threshold), and are coded as black triangles; precipitating conditions are denoted with circles. The size of the data set is indicated (N) as well as relevant error metrics (RMSE, mean bias, R^2).

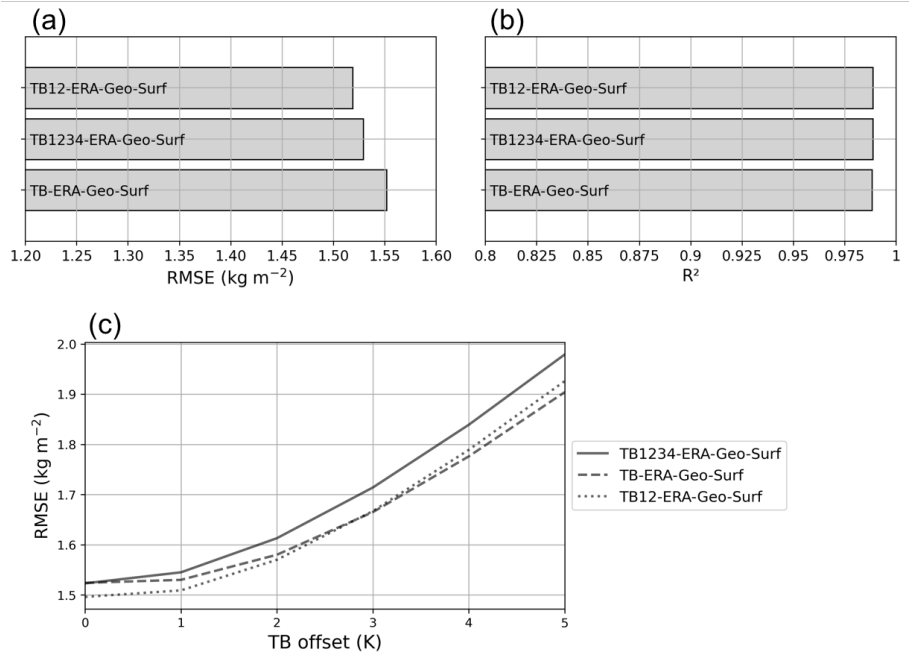

Figure A1. Effect of higher-order T_B polynomials on IWV retrieval. Panels a) and b) show the RMSE and R^2 on the testing set. It comes across that the best results are obtained with T_B and T_B^2. Adding T_B^3 and T_B^4 leads to similar results, with slightly higher RMSE. Panel c) illustrates how the RMSE changes when a constant T_B offset is added to the testing input, simulating a miscalibration of the radiometer. In terms of relative increase, the retrieval with T_B only is slightly less affected, but it does not bring a large enough improvement to be considered preferable in comparison with the retrieval with T_B and T_B^2.

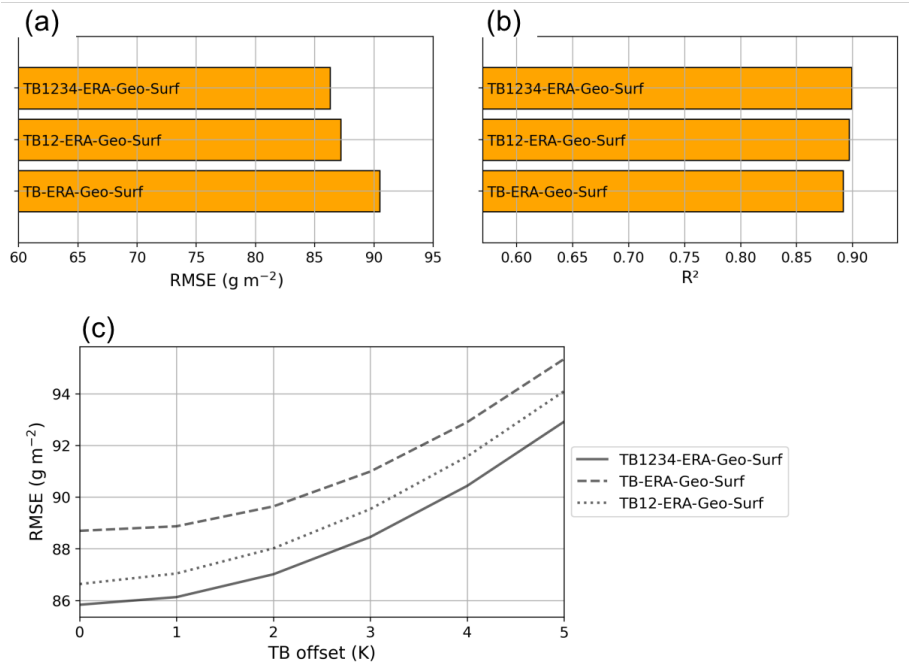

Figure A2. Effect of higher-order T_B polynomials on LWP retrieval. Panels a) and b) show the RMSE and R^2 on the testing set. It comes across that the best results are obtained with the full set of T_B polynomials up to the fourth order. Using only T_B and T_B^2 leads to similar results, with slightly higher RMSE. Panel c) illustrates how the RMSE changes when a constant T_B offset is added to the testing input, simulating a miscalibration of the radiometer. In terms of relative increase, the retrieval with T_B only is slightly less affected, but the RMSE remains higher than that of the other retrievals.

Table 1. Main parameters of the neural networks and training process.

Target	Neurons	Layers	Cost function	Optimizer	Activation	Epochs	Batch size
IWV	120	7	Mean square error	RMS ~~propagation~~ prop	ReLU	70	512
LWP	150	6	Mean square error	RMS ~~propagation~~ prop	ReLU	90	512

Integrated water vapor and liquid water path retrieval using a single-channel radiometer

amt-2020-211

Responses to reviewers

Anne-Claire Billault-Roux and Alexis Berne

November 2020

First, we would like to thank the reviewers for their constructive comments that helped us improve our retrieval method and perform a more thorough analysis of its quality. In the present document, we provide our responses to the comments of the three referees. Each section corresponds to the comments of a referee. The comments of the reviewers are reported in italic, our responses in normal font and the corresponding modifications in the manuscript in blue.

In addition to the modifications related to referee comments, other minor changes were made to the manuscript, which are listed below:

Section 3.2 "*The cloud droplet size distribution (DSD) is chosen as a gamma distribution following Karstens et al. (1994)*"
The cloud droplet size distribution (DSD) is chosen as a monodisperse distribution with radius rc = 20 μm following Cadeddu et al. (2017)

It was considered preferable to use a unique cloud droplet size distribution, rather than one that introduces empirical thresholds. As explained in Sect. 3.2, the exact choice of the DSD does not affect the radiative transfer model as long as droplets are within the Rayleigh range for the considered frequency. This corresponds to the range of validity of the algorithm, so this modification does not substantially affect the results.

Section 6.2 "*The value of this offset (18 K) was determined by computing theoretical brightness temperatures from clear-sky radiosonde profiles*"
The value of this offset (20 K) was determined by computing theoretical brightness temperatures from clear-sky radiosonde profiles

Calculations to evaluate the miscalibration were performed once again. The exclusion of precipitating cases requires to define some thresholds, which were adjusted during this new analysis. This led us to modify the value of the offset in TB. However, this value should be taken with care, since a fully satisfactory correction of such a high miscalibration is a quite challenging task.

1 Anonymous referee #1

The authors developed an algorithm to retrieve liquid water path (LWP) and integrated water vapor (IWV) from ground-based passive microwave observations at 89 GHz. The algorithm is based on a neural network approach and uses synthetic cloud data derived from global radiosonde observations of atmospheric profiles. The algorithm is tested at two locations – Switzerland and South Korea, and the testing results are reasonably good. Because location and time information are used in the input of the algorithm, the algorithm can be used anywhere over the globe without further tuning required. The algorithm is targeted for the WProf radar-radiometer instrument, which was deployed during the ICE-POP 2017 field experiment. I found that the work is quite useful, and the writing is generally clear. However, I do like to see some revisions before recommending publication. In particular, I'd like the authors to clearly state the fundamental limitation of the approach: the 89 GHz observation is not optimized for IWV retrieval, most of the IWV info gained from the algorithm is from secondary input parameters, such as surface temperature/humidity obs, ERA reanalysis, or even the correlation between LWP and IWV. I recommend minor revision, but would like to see the authors clearly address the above comment.

We thank the reviewer for the positive comment and the constructive feedback. This raises a valid point that we did not emphasize in the first version of the manuscript: the frequency of the microwave radiometer which is used here is mostly sensitive to liquid water, and to a smaller extent to water vapor. Initially, the algorithm was only designed for LWP retrieval, but in the course of investigations it was found that it could also provide reasonable estimates of IWV, provided that other input features are taken into account. We tried to make this more clear in the revised version, for example through the following items.

[Abstract] While 89-GHz brightness temperature is crucial to LWP retrieval, only moderately does it contribute to IWV estimation, because of the lower sensitivity of this channel to water vapor emission.

[Section 1] Those additional input features are especially key to the retrieval of IWV, to which TB at 89 GHz is less sensitive than to LWP.

[Section 5.1.1] If only one input feature were available, all the versions would predict worse results than those given by reanalysis data. Including TB in the retrieval does not lead to the same leap in accuracy than for LWP (discussed in the following subsection), which was expected because the microwave frequency that is used here is not highly sensitive to water vapor. However, excluding TB from the input features degrades the RMSE to 2.56 kg m^{-2}, i.e. + 67 % error compared to the best version, which clearly shows that some information is extracted from the brightness temperature in the retrieval.

[Section 5.1.2] This also highlights once again that brightness temperature at 89 GHz is much more sensitive to liquid water than to water vapor: for LWP retrieval, input features other than TB only bring second-order improvements, while they were shown to be crucial in the IWV retrieval.

[Section 6.2] Because of the relatively low sensitivity of TB at 89 GHz to water vapor, the algorithm largely relies on non-radiometric features; this is even more the case in cold and dry environments like that of ICEPOP, where IWV is low.

[Section 7] While retrieving IWV based on T_B at 89 GHz alone does not lead to accurate results – because of the lower sensitivity of this channel to water vapor emission, compared to

liquid water – this study showed that reliable retrievals could be achieved by including surface and geographical information, as well as reanalysis data if available, among the input features.

---

1. *Lines 9-11. The description of algorithm performance is too qualitative (don't carry much information). Please use some quantitative measures, such as R2, Bias, RMSE. Specify the performance separately for IWV and LWP retrievals. Also, indicate which input parameter(s) has (have) the most impact on the algorithm's performance.*

   We added some detail on the algorithm's performance in the abstract. As such, we believe it might be misleading to cite in the abstract which of the secondary input features has the greatest effect, since we demonstrate that they each contribute to the retrieval in a relatively similar way. The take-home message is that overall, including diverse features (rather than using TB alone) improves significantly the algorithm.

   The new algorithm is shown to be quite robust although its accuracy is inevitably lower than that obtained with state-of-the-art multi-channel radiometers, with a relative error of 7.2 % for LWP (on cloudy cases with LWP $>$ 30 g m^{-2}) and 5.2 % on IWV. The highest accuracy is obtained in mid-latitude environments with a moderately moist climate. Furthermore, the additional input features are found to improve significantly the accuracy of the retrieval.

---

2. *Line 27. Aerosols contribute to millimeter wave radiation?*

   We thank the reviewer for noting this mistake, which we removed.

---

3. *Line 40. I don't understand why physical model is "computationally heavy and cannot implemented accurately when only one frequency is available". I think the fundamental problem is that you have only one piece of information (TB), and yet you want to obtain two unknowns (IWV and LWP). You need secondary info and/or climatological relation between the unknowns. It is a highly statistical problem anyway; there is no need (not helpful) to go through the physical route.*

   Physical methods are used in some state-of-the-art multi-frequency radiometer retrievals, as described for example in Cadeddu et al. 2013 and Turner et al. 2007. In those cases, such methods are more accurate than statistical retrievals, but require more initial information on the atmospheric column; in addition, they have a higher computational expense. As highlighted by the referee in this comment, it is in our case not possible to go through such a physical method because of the lack of microwave measurements. Thus, it was misleading to mention the high computational expense - which was in this case not the problem.

   Although this method is formally the most accurate (Turner et al. 2007), it requires more than one radiometer frequency to lift the problem's fundamental underdetermination, and is thus not applicable for this study

4. *Section 3.1: Please give more detailed info on how "clouds" are created based on atmospheric profiles, in particular, on how LWP depends on temperature and humidity. Since the observation is TB from one channel only, it cannot separate between liquid and vapor information. Therefore, if there is a correlation between humidity (or temperature) and cloud liquid, the retrieved LWP will be always correlated with IWV because the relation is built in the a priori training dataset. This correlation may not be valid in natural clouds, at least for some types of clouds. So, the authors should let readers know about this issue, so that readers can take precaution when interpreting retrieved results. I suggest that the authors rewrite this section, provide full information on the liquid cloud model.*

More detail was included on the liquid cloud model in Sect. 3.1. It is indeed true that the model might not be accurate in certain cases, and can fail to output the correct LWP for a given atmospheric profile. However, the most important here is that the model is bias-free and outputs a *realistic* – if not *true* – profile, which is then used to create the dataset and to train the retrieval algorithm. The Salonen model was compared to a few commonly used models (described for example in Mattioli 2009, Loehnert and Crewell 2003), and selected because it provided the least mean bias when compared to ERA5 data. This ensured that the model is realistic enough to produce a dataset on which to perform the statistical learning.

To derive profiles of liquid water content (LWC) from radiosonde profiles of atmospheric variables, the cloud model from Salonen 1991 was used. Cloud boundaries are identified using a threshold U_c on relative humidity, this threshold being pressure- and temperature-dependent, according to Eq. 1.

$$U_c = 1 - \alpha\sigma(1-\sigma)[1+\beta(\sigma-0.5)] \tag{1}$$

Here, $\sigma = \frac{P}{P_0}$ with P and P_0 denoting respectively atmospheric pressure at the current level and at the ground. Corrections from Mattioli et al. 2009 are used for the coefficients α and β of the Salonen model. Within the cloud layers, the liquid water profile is then calculated as a function of temperature and height above cloud base, following Eq. 2.

$$LWC = w_0(\frac{h-h_b}{h_r})^a f(T) \tag{2}$$

where $f(T) = 1 + cT$ for $T \geq 0$ and $f(T) = exp(cT)$ for $T < 0$, with T in °C, $c = 0.04°C^{-1}$, $w_0 = 0.17$ g m^{-3}, $h_r = 1.5$ km, h and h_b denoting height and height of cloud base. There are some limitations when assuming a single universal cloud model, since it may fail to capture specific cloud properties in certain environments: more sophisticated and accurate models can be defined on a local geographical scale (e.g. Pierdicca et al. 2006). However, given the stated objective of this study to design a non-site specific algorithm, it was considered preferable to assume a single universal liquid cloud model, in spite of its potential drawbacks.
A further limitation of this cloud model is related to the relatively low resolution of the atmospheric profiles extracted from the radiosonde data (c.f. Sect. 2.1) that are used as an input. This might result in a misrepresentation of the cloud layers in their detection and their size.
In order to ensure that this forward model resulted in the least possible bias, its results were compared against LWP values from reanalysis data. Even though the model might fail, on a given occurrence, to reproduce the actual liquid water profile in the atmospheric column, it should not produce a significant bias on average. This condition guarantees that the synthetic dataset that is used for training contains realistic – if not real – profiles, and this should therefore not degrade the quality of the retrieval algorithm. This cloud model was chosen over other commonly used ones (Decker model, Salonen model without correction, c.f. Mattioli 2009) for

it was found to produce the least bias when compared to ERA5 LWP values (mean bias of 14 g m^{-2} vs. 26 g m^{-2} (resp. -24 g m^{-2}) for the unadjusted Salonen model (resp. the Decker model with 95 % threshold).

---

5. *Lines 140-141. Higher order terms of TB are used as input of the algorithm. Why? Do the higher order terms actually have "the greatest importance"? I didn't see the importance has been shown in the sensitivity test section. Please clarify.*

   Indeed, an analysis of the importance of higher order terms was missing from the first version of the manuscript and was included in Sect. 5. Figures to support this analysis are presented in Appendix.

   [Section 5.1.1] An analysis was conducted to identify the importance of higher order polynomials in the algorithm, a summary of which can be found in Appendix. It was found that the most accurate retrieval is obtained by including TB and TB^2. If higher order terms are added, this slightly reduces the accuracy of the retrieval, and also degrades its stability to TB miscalibration. On the other hand, including only TB, while it makes the algorithm slightly more stable, does not appear as the best solution for it has lower accuracy. Hence, the results which are presented here and in the further sections are those obtained using TB and TB^2.

   [Section 5.1.2] The analysis of higher order terms' importance in the case of LWP retrieval shows that the best results are obtained by using TB polynomials up to the fourth order (cf. Appendix), while this does not affect significantly the stability of the retrieval to errors in TB. Let us highlight that in the case of a linear regression, one would expect the error to diverge when high-order polynomials are included. This is not the case here, because of the non-linear behavior of the neural network. Therefore, in the results which are shown here and further, "TB" implies that TB polynomials up to the fourth order are used.

---

6. *Line 145-146: I am concerned about the algorithm replying on reanalysis data, in particular directly replying on IWV and LWP in the reanalysis. This makes the 'observation" completely mixed with "model", something worries me if I try to use "observation" to validate models.*

   The logic behind this point is that reanalysis data can contain information on the meteorological and geographical setting of a given measurement. The spatial and temporal resolution is not sufficient for the reanalysis to be considered as ground truth, and is therefore not used for the validation of the algorithm. As such, ERA5 data is not incorporated in a physical model, but merely serves to improve the statistical retrieval.

   The spatial and temporal resolution of this reanalysis data is too low for it to be held as ground truth, but it can serve as a reasonable rough estimate and thus bring some improvements to the statistical learning process – although it could not be included as such in a physical model.

---

7. *Line 152: What are considered to be "strong rain events"? 0.1 mm/hr, 1 mm/hr, or 10 mm/hr? If I am a data user, I will stop using the retrieval when rainrate exceeds 0.1 mm/hr because all assumptions used in the algorithm development become invalid once rain/drizzle starts. So, I wouldn't call "strong rain" here.*

   We changed "strong rain" to "rain" and included a more detailed discussion of this topic in Sect. 3.2, in relation with one comment from Referee #3.

---

8. *Line 157. I am not sure I understand why if clear cases are left in the training set, "the training phase will result in a strong bias of the retrieval toward low LWP values." I thought the algorithm should be able to retrieve zero LWP (clear) cases. If all/most cases in the training set are cloudy cases, will the retrievals be most likely to have LWP >0 ? That is not what really happens in nature. For most locations, clear-sky has more frequency of occurrence than cloudy-sky. The fact that the case counts in Fig.2b is the same values for LWP from 0 to 600 g m^{-2} also puzzles me. Maybe I misunderstand the concept here. Please explain.*

   If the training set is kept as such – without subsampling –, the algorithm has a huge bias toward low values: because the training set contains almost only values of LWP=0, the algorithm has very good performance by predicting LWP=0 all the time. Which is not what is desired... This is an example of regression toward the mean, a common behavior for regression methods. To correct for this, we truncated the distribution. This does not mean excluding clear-sky cases (they are still present in the training set, although in a proportion more comparable to cloudy cases), and the algorithm is still able to identify LWP=0. The choice of the 600 g/m2 threshold resulted from a trial-error method. We modified the explanation to clarify this point.

   If left as such, the training phase would result in a strong bias of the retrieval toward low LWP values (a bias of $\sim$ 100g m^{-2} for LWP $>$ 400g m^{-2} was noted): this is a common artefact that is observed in statistical learning algorithms, as an effect of unbalanced training set. In order to avoid this, the dataset was subsampled so that clear-sky and cloudy cases (up to 600 g m^{-2}) would be equally represented; the value chosen for this threshold results from a trade-off between bias reduction and preservation of overall accuracy.

---

9. *Line 168. What is the difference between validation set and testing set?*

   The hyperparameters of the neural network (number of neurons, layers...) are tuned on the validation set while the testing set is for the final evaluation of the algorithm.

   The validation set is used for tuning the hyperparameters of the neural network, while the final evaluation metrics are computed on the testing set.

---

10. *Fig.6: Explanation to Fig.6 is severely lacking. It is really hard to understand what input are critical, although I guess the purpose of having this figure is exactly tended to show the importance of each input variable. Please think a better way to explain. But one thing seems to be clear: no extra inputs but TB alone will result in a horrible retrieval. It also seems to me that ERA and the combination of Geo+Surf are redundant, i.e., you only need either, but doesn't have to have both, although the authors didn't conclude in such a way. If my interpretation is correct, I would get rid of ERA in the algorithm, and wouldn't say "the most important secondary feature is ERAs estimates." (lines 192-193).*

We tried to make the figure more readable by changing the labels and only writing the input variables which are used for each version of the algorithm. We also expanded the explanation and analysis of this figure in Section 5.1 which was largely rewritten.

[Section 5.1] Figure 6 shows how this total error, represented by RMSE (left panels) and by the square correlation coefficient (R^2) (right panels), is affected by the addition or removal of input features. For each set of input features, a full tuning of the algorithm was performed, and the results that are presented correspond to those from the tuned – i.e. best – version on the testing set.

[Section 5.1.1] From Fig. 6 a) and b) it comes across that the IWV retrieval is significantly improved by the addition of multiple input features. The highest accuracy is obtained with the full set of input features, and corresponds to a RMSE of 1.53 $kg\ m^{-2}$. On the other hand, including solely T_B measurements in the input deteriorates the RMSE to nearly 6 $kg\ m^{-2}$. If only one input feature were available, all the versions would predict worse results than those given by reanalysis data. Including T_B in the retrieval does not lead to the same leap in accuracy than for LWP (discussed in the following subsection), which was expected because the microwave frequency that is used here (89 GHz) is much less sensitive to water vapor. However, excluding T_B from the input features degrades the RMSE to 2.56 $kg\ m^{-2}$,i.e. + 67 % error compared to the best version, which clearly shows that some information is extracted from the brightness temperature in the retrieval.

[Section 5.1.2] Figures 6 c) and d) show that for LWP retrieval, input features other than T_B only bring second-order improvements, while they were shown to be crucial in the IWV retrieval. For instance, the addition of reanalysis data significantly improves the IWV retrieval, but only in a relatively minor way does it increase the accuracy of LWP retrieval. On the contrary, excluding T_B from the input features leads to RMSE near 200 $g\ m^{-2}$ and $R^2 < 0.5$, i.e. to values with make the retrieval not relevant. This highlights once again that brightness temperature at 89 GHz is much more sensitive to liquid water than to water vapor. An additional reason for this high dependence on T_B is that LWP at a given location can have a high temporal variability due to cloud dynamics in the atmospheric column, which might not always be captured in the time series of surface atmospheric variables, nor by ERA5 models which have a comparatively low spatial and temporal resolution.

---

11. *Section 5.1. I would like to see a test in which you only use secondary info (without TB observation), and see how IWV and LWP "retrievals" will look like. I'd guess IWV retrievals could be reasonably good. If that is true, it means TB observation does not add much info to the retrievals.*

This is an important point that we did not discuss in the first version of the manuscript. In the

revised version, a discussion of this point was included in Sect. 5.

[Section 5.1.1] Including T_B in the retrieval does not lead to the same leap in accuracy than for LWP (discussed in the following subsection), which was expected because the microwave frequency that is used here (89 GHz) is much less sensitive to water vapor. However, excluding T_B from the input features degrades the RMSE to 2.56 $kg\ m^{-2}$, i.e. + 67 % error compared to the best version, which clearly shows that some information is extracted from the brightness temperature in the retrieval.

[Section 5.1.2] On the contrary, excluding T_B from the input features leads to RMSE near 200 $g\ m^{-2}$ and $R^2 < 0.5$, i.e. to values with make the retrieval not relevant.

---

12. *Lines 245-247. Any ideas why error in these areas are large? Underrepresentation of radiosonde profiles in the training set, or cloud model not suitable for these regions?*

It is difficult to identify the most important factor that causes a high error in these areas. The underrepresentation of this range of latitudes and corresponding climates is certainly a critical point: unfortunately, there is to our knowledge little data available from these areas.
We would expect the cloud model to also be less accurate in such areas than in the midlatitude regions for which it was initially designed. However, this should not transpose directly to such a high error in the results of the algorithm on the synthetic dataset (which does not consist of "true" LWP but of LWP derived from this cloud model; so the bias from the cloud model should not be visible).

Tropical climates are underrepresented in the dataset, for little data is available from this region in comparison with mid-latitude areas: their specificity might therefore not be fully captured during the learning stage of the algorithm. This accounts at least partly for the enhanced error over the Indian peninsula and South-Eastern Asian islands.

---

13. *Lines 261-262 "This ... stage". This sentence doesn't make sense to me. ERA5 just happened to have IWV bias in this region for this time period, or ERA5 generally has IWV bias in general?*

It is a bias that was noted by MeteoSwiss at Payerne, at least during this timeframe. This is not a general statement on ERA5 values. We tried to make this sentence clearer.

This might be due to a bias in ERA5 data during this timeframe over the region, which is visible in ERA5 records during the entire campaign (with a value of $-4.1kg\ m^-2$) and for which there is no clear explanation at this stage.

---

14. *Fig.10. I have difficulties to interpret the results of the lowest two rolls in Figs.10c and. Adding ERA data makes LWP results worse? I am confused here.*

There was unfortunately a major mistake in the labels of this figure. We corrected for this, please refer to the corresponding Fig. 10 in the revised version.

---

15. *Lines 296-297. I always believed that the 89 GHz is not a good frequency for IWV. It works well for LWP. When IWV and LWP are not well correlated, IWV retrievals will suffer.*

Indeed, TB at 89 GHz is much more sensitive to liquid water than to water vapor. However, in general, removing TB from the input features still leads to a decrease in the accuracy of the algorithm, although in a much smaller way than for LWP, as discussed in Sect. 5.

The best results are found when several input features are included and drop severely when no secondary input features are used, which corresponds to the results on the synthetic data set presented in Sect. 5. Because of the relatively low sensitivity of TB at 89 GHz to water vapor, the algorithm largely relies on non-radiometric features; this is even more the case in cold and dry environments like that of ICEPOP, where IWV is low.

---

16. *Lines 303-304. I am not sure that the results in this paper indicated this statement. Please explain.*

This is an observation deduced from Fig. 13. The sentence was rephrased to avoid confusion.

Snowfall events during the campaign, as well as occasional fog, can also bias the retrieval by enhancing brightness temperature.
The analysis of the ICE-POP data was taken a step further to explore the latter point. It appears that the IWV retrieval is most reliable in non-precipitating or cold conditions, i.e. when little liquid water is expected in the column. To visualize this, periods with no precipitation are identified using WProf's radar measurements as time steps with low radar equivalent reflectivity ($Ze < -10dBZ$) in the lower gates (first kilometer above the radar), and temperature time series are provided by the weather station coupled to WProf. Fig. 13 shows the scatter plot of the error, color-coded in a way to differentiate dry conditions from precipitating cases.

---

17. *Section 7. The summary appears mostly focusing on IWV retrieval. Please also mention some results on LWP. To repeat my main concern: I think this instrument is not good for IWV, but rather good for LWP. I hope that the authors can state something to reflect this point.*

We nuanced the results on IWV and highlighted that the retrieval of IWV is a corollary to that of LWP, which we did not make clear in the original manuscript. For a reasonably accurate IWV retrieval, it is critical to have several ground measurements available; TB has a role but certainly not as crucial as in the LWP retrieval.

While retrieving IWV based on T_B at 89 GHz alone does not lead to accurate results – because of the lower sensitivity of this channel to water vapor emission, compared to liquid water – this study showed that reliable retrievals could be achieved by including surface and geographical information, as well as reanalysis data if available, among the input features.

---

18. *Fig.2. The x-axis label in Fig.2b should be "LWP (g mˆ-2). Fig.3 and Table 1 are not explained well in the text. I am not sure what the curves and symbols mean. Please explain more as not everyone is familiar to the Keras library in Python.*

The label of Fig. 2 was corrected.
Some more detail was included on the design and training of the algorithm in section 4.3, to explain Fig. 3 and make more clear the notations of Table 1.

The neural network was trained through mini-batch gradient descent, using RMSprop optimizer which allows for learning rate adaptation and is often used for statistical regression problems (Chollet, 2017). As comes across from the training curve of the LWP retrieval on Fig. 3, the training dataset is large enough to ensure that the algorithm is not prone to overfitting. Indeed, the error on the validation set quickly drops when the size of the training set, then plateaus with a slight decrease. In other words, the accuracy of the algorithm is not limited by the amount of data used in the training stage. Figure 4 and Table 1 summarize the resulting architecture and relevant parameters of the algorithm. These include the description of the neural network's structure (number of neurons and hidden layers) as well as training parameters such as the batch size and number of epochs, i.e. the number of iterations through the entire dataset in the learning phase.

---

2 Anonymous referee #2

The manuscript presents a single channel retrieval for IWV and LWP using the 89 GHz frequency. The retrieval uses a neural network trained with a synthetic dataset from a collection of radiosonde data worldwide. The coefficients are applied to the test dataset and to real data collected during two field campaigns. The retrievals provide robust, albeit not excellent results. The quality of the retrievals is however good and robust enough to be acceptable when more sophisticated instruments are not available. I found the paper interesting and the results useful considering the difficulty of deploying full radiometric suites in many locations. Overall the exposition is clear and well organized.

We thank the reviewer for their positive comment.

---

1. *Line 140: "The first category consists of TB and higher order polynomials (up to fourth degree)" -Do higher order polynomials actually add information to the network? I would imagine that non-linearity is accounted for in the network structure.*

   While neural networks are able to fit non-linear behaviors, they are not always able to resolve polynomial features – or require more parameters to do so, with the risk of producing overfitting. Polynomial input features are thus sometimes used in the design of NNs. The concern of how much those higher order powers benefit to the retrieval is fully valid, and we included a more in-depth discussion of this topic in Sect. 5. Figures to support this analysis are presented in Appendix.

   [Section 5.1.1] An analysis was conducted to identify the importance of higher order polynomials in the algorithm, a summary of which can be found in Appendix. It was found that the most accurate retrieval is obtained by including TB and TB^2. If higher order terms are added, this slightly reduces the accuracy of the retrieval, and also degrades its stability to TB miscalibration. On the other hand, including only TB, while it makes the algorithm slightly more stable, does not appear as the best solution for it has lower accuracy. Hence, the results which are presented here and in the further sections are those obtained using TB and TB^2.

   [Section 5.1.2] The analysis of higher order terms' importance in the case of LWP retrieval shows that the best results are obtained by using TB polynomials up to the fourth order (cf. Appendix), while this does not affect significantly the stability of the retrieval to errors in TB. Let us highlight that in the case of a linear regression, one would expect the error to diverge when high-order polynomials are included. This is not the case here, because of the non-linear behavior of the neural network. Therefore, in the results which are shown here and further, "TB" implies that TB polynomials up to the fourth order are used.

---

2. *Line 157: "In order to avoid this, the dataset was subsampled so that clear-sky and cloudy cases (up to* $600g\ m^{-}2$*) would be equally represented" -I am not sure I entirely agree with this approach. The point here is that neural networks perform better when the training dataset reproduces statistically the true occurrence of the events. Statistically clear sky cases occur more often than cloudy cases so I am afraid that modifying that distribution may actually cause the*

network to bias the LWP. Note that this is different from what you did earlier to avoid the uneven geographical sampling. In that case the problem was due to an uneven distribution of the monitoring network and the resampling was legitimate.

If the training set is kept as such – without subsampling –, the algorithm has a huge bias toward low values: because the training set contains almost only values of LWP=0, the algorithm has very good performance by predicting LWP=0 all the time. Which is not what is desired... This is an example of regression toward the mean, a common behavior for regression methods. To correct for this, we truncated the distribution. This does not mean excluding clear-sky cases (they are still present in the training set, although in a smaller proportion), and the algorithm is still able to identify LWP=0. The choice of the 600 g/m2 threshold resulted from a trial-error method. We modified the explanation to clarify this point.

If left as such, the training phase would result in a strong bias of the retrieval toward low LWP values (a bias of $\sim$ 100g m^{-2} for LWP $>$ 400g m^{-2} was noted): this is a common artefact that is observed in statistical learning algorithms, as an effect of unbalanced training set. In order to avoid this, the dataset was subsampled so that clear-sky and cloudy cases (up to 600 g m^{-2}) would be equally represented; the value chosen for this threshold results from a trade-off between bias reduction and preservation of overall accuracy.

---

3. *Section 5.1.1, line 193: Does the retrieval contribute anything to the ERA estimates of IWV? i.e. if you compute the RMSE of the ERA water vapor on the same dataset would you get the same RMSE as the retrievals (1.6 kg/m2) or worse? I see you show this information later on, but it would be useful to also comment on it here.*

The information was added in this section.

For comparison, the ERA5 data alone has a higher RMSE (3.4 kg m^{-2}) on the same data set.

---

4. *Section 5.1.2, line 210-2-13: "Similar reasons can help explain why the addition of reanalysis data significantly improves the IWV retrieval, but only in a minor way does it increase the LWP retrieval's accuracy. Liquid water content can vary on a shorter spatial and temporal scale than that captured by ERA5 models." -Another reason is that the 89 GHz is not a water vapor resonance therefore the information content for vapor is less than for liquid water path.*

This is an important aspect which was also pointed out by the other referees, and which we did not emphasize enough in the first version of the manuscript. We mention it here and also recall it in the summary section.

This also highlights once again that brightness temperature at 89 GHz is much more sensitive to liquid water than to water vapor: for LWP retrieval, input features other than TB only bring second-order improvements, while they were shown to be crucial in the IWV retrieval.

---

5. *Figs 6, 10, 11, and 12 are a little bit difficult to interpret in my opinion. I am not sure I understand them entirely. The general conclusion that I would draw from them is that the RMSE is very similar for almost all combinations of input parameters except for two or three combinations. However, looking at the various combinations, is hard to understand the rationale for the different performances. As an example if I look at Fig. 10a I see that the combination noERA-Geo-noSurf has a vastly different RMSE than noERAnoGeo-Surf. Does this mean that the effect of having surface parameters is equivalent to having ERA data? Similarly, for example in fig. 10b I see that the combination ERAnoPWVpred- Geo-Surf has much higher RMSE than ERA-noPWVpred-noGeo-noSurf. I am not sure how to interpret that.*

We tried to make these error plots more readable by changing the labels and only writing the input variables which are used for each version of the algorithm. We also expanded the explanation and analysis of figure 6 in Section 5.1 which was largely rewritten. There was unfortunately a major mistake in the labels of Fig. 10, which was corrected for.

[Section 5.1] Figure 6 shows how this total error, represented by RMSE (left panels) and by the square correlation coefficient (R^2) (right panels), is affected by the addition or removal of input features. For each set of input features, a full tuning of the algorithm was performed, and the results that are presented correspond to those from the tuned – i.e. best – version on the testing set.

[Section 5.1.1] From Fig. 6 a) and b) it comes across that the IWV retrieval is significantly improved by the addition of multiple input features. The highest accuracy is obtained with the full set of input features, and corresponds to a RMSE of 1.53 $kg\ m^{-2}$. On the other hand, including solely T_B measurements in the input deteriorates the RMSE to nearly 6 $kg\ m^{-2}$. If only one input feature were available, all the versions would predict worse results than those given by reanalysis data. Including T_B in the retrieval does not lead to the same leap in accuracy than for LWP (discussed in the following subsection), which was expected because the microwave frequency that is used here (89 GHz) is much less sensitive to water vapor. However, excluding T_B from the input features degrades the RMSE to 2.56 $kg\ m^{-2}$, i.e. + 67 % error compared to the best version, which clearly shows that some information is extracted from the brightness temperature in the retrieval.

[Section 5.1.2] Figures 6 c) and d) show that for LWP retrieval, input features other than T_B only bring second-order improvements, while they were shown to be crucial in the IWV retrieval. For instance, the addition of reanalysis data significantly improves the IWV retrieval, but only in a relatively minor way does it increase the accuracy of LWP retrieval. On the contrary, excluding T_B from the input features leads to RMSE near 200 $g\ m^{-2}$ and $R^2 < 0.5$, i.e. to values with make the retrieval not relevant. This highlights once again that brightness temperature at 89 GHz is much more sensitive to liquid water than to water vapor. An additional reason for this high dependence on T_B is that LWP at a given location can have a high temporal variability due to cloud dynamics in the atmospheric column, which might not always be captured in the time series of surface atmospheric variables, nor by ERA5 models which have a comparatively low spatial and temporal resolution.

3 Anonymous referee #3

1.) It's very good and important to see, that an approach has been made to derive a globally valid retrieval algorithm for IWV and LWP observations from single channel microwave observations. This can be beneficial for different science applications, e.g. weather & climate but also in astronomy and radio propagation.
2.) The paper is well written and makes clear points. It nicely addresses the fact, that the combination of microwave radiometry in synergy with re-analysis output and standard environmental conditions can be advantageous.
3.) Although I strongly favor short and concise papers, the results presented here (especially in Sections 5 and 6) are – to an extent – kept rather minimal. The paper would benefit from a more detailed and quantitative discussion. See also my specific comments below.

We thank the reviewer for their feedback. Following their comment, we have included a more detailed discussion on the algorithm's results in Sect. 5 and 6, which were largely rewritten.

---

1. *Lines 25-27: If the authors are referring to TBs in the microwave spectrum, please omit "radiative contribution of...aerosols"*

   We thank the referee for noting this mistake, which we removed.

---

2. *Lines 68 onward: I assume you performed a quality control for the radiosonde profiles used for retrieval development, if not please consider doing so. Depict checks concerning range (e.g. min/max) of atmospheric parameters, maximum ascent height, consistency checks concerning pressure and/or temperature gradient, etc...*

   Initially, no specific check was performed on the profiles, but unrealistic values were excluded at various stages (especially during the computation of LWP, IWV and TB with PAMTRA). However, this did not provide a full insight on the quality of the data, therefore a more thorough approach was implemented.

   A quality check was performed on each of the relevant variables (pressure, temperature, relative humidity), through the following steps: first, the minimum and maximum P (resp. T, RH) in a given range of altitudes were extracted from each radiosonde. When examining the distributions that are obtained, outliers were visible, which were then removed with a 10^{-4} quantile (upper and lower quantile). The atmospheric column was split into 9 ranges of altitudes, and this routine was performed for each. In total, 6395 profiles were flagged out and removed. It was ensured that this did not result in the systematic removal of some geographical locations. Following this step, the vertical profiles of pressure, temperature and relative humidity are used as input to the forward model, described in Sect. 3.

---

3. *Line 76: Please discuss what the relatively low vertical resolution could imply for the retrieval development. E.g., the coarse resolution of the relative humidity profile will influence where and*

how many liquid layers are detected. What happens if this is significantly different for different radiosonde sites? This discussion could also be part of Section 3.1.

This is a real concern that we encountered when designing the forward model. The crucial point was to ensure that the dataset that was used to train the algorithm was a realistic one, i.e. that it did not contain a statistical bias. We added a discussion in Sect. 3.1.

A further limitation of this cloud model is related to the relatively low resolution of the atmospheric profiles extracted from the radiosonde data (c.f. Sect. 2.1) that are used as an input. This might result in a misrepresentation of the cloud layers in their detection and their size.
In order to ensure that this forward model resulted in the least possible bias, its results were compared against LWP values from ERA5 reanalysis data (Copernicus Climate Change Service, 2020). Even though the model might fail, on a given occurrence, to reproduce the actual liquid water profile in the atmospheric column, it should not produce a significant bias on average. This condition guarantees that the synthetic dataset that is used for training contains realistic – if not real – profiles, and this should therefore not degrade the quality of the retrieval algorithm. This cloud model was chosen over other commonly used ones (Decker model, Salonen model without correction, c.f. Mattioli 2009) for it was found to produce the least bias when compared to ERA5 LWP values (mean bias of 14 g m^{-2} vs. 26 g m^{-2} (resp. -24 g m^{-2}) for the unadjusted Salonen model (resp. the Decker model with 95 % threshold).

---

4. *Lines 92/93: Describe how far WProf and HATPRO were apart (in meters) and if you expect any corresponding uncertainties during retrieval application.*

This is a relevant information which was added to the text. We do not expect that it affects greatly the comparison, but there could be some edge cases where a cloud would be in the field of one radiometer and not the other.

The instruments were located approximately 65 m apart; this distance is small enough that it should in general not affect the comparison of the retrieved values from the two instruments. However, in some rare cases, it is possible that a cloud would overpass one of the radiometers, but not the other, leading to a discrepancy in the measured brightness temperatures.

---

5. *Section 3.2: I am missing a specification of the absorption models used for water vapor, oxygen and cloud liquid water. As previous studies have shown (e.g. Cimini et al. 2018, https://doi.org/10.5194/acp-18-15231-2018), this can be decisive for the absolute accuracy of your retrieval results. This aspect is nowhere discussed in the paper, should be, however.*

The description of the absorption model used was added to this section (Rosenkranz model for gases, which corrections from Turner 2009 and Liljegren 2005). It is the default one implemented in PAMTRA, and was used in some recent similar studies (e.g. Cadeddu 2017). As pointed out by the reviewer in this comment, there remains an irreducible uncertainty associated to those models, which the reader ought to have in mind.

Gaseous absorption is calculated using the default parameters in PAMTRA, i.e. with the model proposed by Rosenkranz 1998 and modifications from Turner 2009 and Liljegren 2005. It should be kept in mind that some irreducible uncertainty remains tied to the choice of these parameters in the radiative transfer model.

---

6. *Line 127: Please specify at which LWC (together with the assumed parameters of your gamma distribution, specify these as well), Mie effects become non-negligible and in how many cases in your data set this threshold is exceeded.*

Regarding the DSD, it was decided for the final version of the algorithm that a monodisperse distribution (as in Cadeddu 17) was more robust than the Gamma proposed by Karstens 1994, which introduced several empirical thresholds that could lead to artefacts and unrealistic discontinuities. This is also justified by the fact that, as long as the Rayleigh regime holds – i.e. in the range of validity of the algorithm – the choice of the DSD does not impact TB calculations. For the identification of potentially problematic profiles in terms of droplet size, criteria from Karstens 1994 were used and a discussion was included on this topic.

The cloud droplet size distribution (DSD) is chosen as a monodisperse distribution with radius $r_c = 20\ \mu m$ following Cadeddu 2017, and scattering calculations are performed with Mie equations, assuming spherical particles.

In order to have a more rigorous grasp on when and how this drawback might affect the synthetic data set that is constructed – and hence the retrieval, criteria from Karstens 1994 were used. In their study, the authors distinguished three types of liquid water clouds based on the value of LWC at a given altitude; for each category of cloud, a different characteristic radius is chosen for the DSD. Assuming that Mie effects start to become an issue in the second category of clouds (*cumulus congestus*), identified for $LWC > 0.2g\ m^{-2}$, this corresponds on average to $LWP \geq 830\ g\ m^{-2}$ and around 2% of the entire dataset fall into this category (i.e. the LWC threshold is exceeded in at least one altitude range). Taking the third category (*cumulonimbus*) with $LWC > 0.4\ g\ m^{-2}$, this applies to 1% of the entire dataset and the threshold increases to $LWP > 1400\ g\ m^{-2}$. Those values can serve as a benchmark to identify LWP values where Mie effects can typically contaminate the retrieval. However, edge cases can also exist where the total LWP is quite low, but a small layer of nearly-precipitating or drizzling cloud still contaminates the retrieval, without featuring extremely high total LWP.

---

7. *Section 4.1: I don't find any indication in the manuscript of how you are dealing with random uncertainty of your measurement variables, e.g. radiometric noise and T/q/p sensor uncertainty. If you want to simulate a realistic retrieval behavior, you need to put noise on your measurements (training, validation and test data set), otherwise you are assuming a "perfect relationship" between measurement and LWP/IWV, which you will never have in reality.*

This is a valid concern which we had neglected in the first place. In the latest version of the retrieval – which is presented in this revised version – a random Gaussian noise was added with a standard deviation of 0.5K as noted by Kuechler et al, 2017. This explanation was added to

section 4.1. Regarding the other surface parameters (temperature, relative humidity, pressure), we consider that it is not necessary to add noise, for the values used are already real measurements (i.e. from real sensors with noise).

In order to simulate realistic measurements, a random Gaussian noise was added to the modeled brightness temperatures, with a mean and standard deviation of resp. 0 K and 0.5 K; those values were identified by Kuechler et al. 2017 as the characteristics of the measurement noise of the 89-GHz radiometer.

---

8. *Line 140: Since the relationship between LWP/IWV and the TBs is, well I'd say, linear to moderately non-linear, I ask myself why you need to use TB^4? Can you quantify the retrieval improvement when using only TB and TB^2 in comparison to higher order terms?*

   Following this comment and similar ones from the other referees, a more rigorous analysis was performed to determine the quantitative contribution of the higher-order terms. It was included in Sect. 5, and supporting figures are added in Appendix.

   [Section 5.1.1] An analysis was conducted to identify the importance of higher order polynomials in the algorithm, a summary of which can be found in Appendix. It was found that the most accurate retrieval is obtained by including TB and TB^2. If higher order terms are added, this slightly reduces the accuracy of the retrieval, and also degrades its stability to TB miscalibration. On the other hand, including only TB, while it makes the algorithm slightly more stable, does not appear as the best solution for it has lower accuracy. Hence, the results which are presented here and in the further sections are those obtained using TB and TB^2.

   [Section 5.1.2] The analysis of higher order terms' importance in the case of LWP retrieval shows that the best results are obtained by using TB polynomials up to the fourth order (cf. Appendix), while this does not affect significantly the stability of the retrieval to errors in TB. Let us highlight that in the case of a linear regression, one would expect the error to diverge when high-order polynomials are included. This is not the case here, because of the non-linear behavior of the neural network. Therefore, in the results which are shown here and further, "TB" implies that TB polynomials up to the fourth order are used.

---

9. *Begin with introducing your results in a general positive sense, make the reader feel like you are now going to present some great, interesting and relevant plots (which you mostly do!). I wouldn't begin Section 5.1 with two sentences that actually belong in the figure caption.*

   An introductory sentence was added in the beginning of this section to open the discussion section in a more appealing way.

   In this section, the algorithm is evaluated on the synthetic dataset (testing set), using different criteria; overall, results are encouraging and the retrieval appears to be robust. Some limitations can be identified, which will be discussed here. Additionally, an analysis on the impact of the

various input features on the retrieval of IWV and LWP is conducted.

---

10. *Section 5.1: I assume Figure 5 illustrates the retrieval including all input parameters? Please state this clearly in the text and figure caption.*

Indeed, Fig.5 illustrates the retrievals with all input parameters. This was clarified in the text and caption.

[Text] The distribution of the error on the testing set is shown in Fig. 5, for the best version of the algorithm, which uses the full set of input features.

[Figure caption] Results of the retrieval algorithms on the synthetic testing dataset. The best versions of the algorithms are presented, i.e. the ones which use the full set of input features.

---

11. *Lines 188 – 190: I think Figure 5 would benefit from two additional sub figures showing the bias as a function of binned IWV and LWP. Then you could quantify the statement you make in these two lines and elaborate a bit more on the bias behavior for smaller IWV.*

As suggested by the reviewer, two subplots were added that illustrate the bias as a function of binned LWP and IWV. Please refer to the new figure (figure 5). The following sentence was added to the caption:

Similarly, panels (e) and (f) show the distribution of mean bias across the range of IWV and LWP values.

---

12. *Lines 191 – 194: I'm a bit puzzled by Figure 6a. If you only use the 89 GHz TB then ERA5 performs significantly better. So, is there any sense of using this TB at all? I think you need to perform a retrieval without TB, just with all other parameters and add this one to your plots. This would help in putting the value of the 89 GHz TB in context. Then you need to discuss your results in more detail.*

The IWV retrieval algorithm without TB is indeed an relevant point that was missing from the first version of the manuscript. We added it to the revised version, and it gives some valuable insight into the role of TB. A discussion on this point was included.

From Fig. 6 a) and b) it comes across that the IWV retrieval is significantly improved by the addition of multiple input features. The highest accuracy is obtained with the full set of input features, and corresponds to a RMSE of 1.53 $kg\ m^{-2}$. On the other hand, including solely T_B measurements in the input deteriorates the RMSE to nearly 6 $kg\ m^{-2}$. If only one input feature were available, all the versions would predict worse results than those given by reanalysis data. Including T_B in the retrieval does not lead to the same leap in accuracy than for LWP (discussed

in the following subsection), which was expected because the microwave frequency that is used here (89 GHz) is much less sensitive to water vapor. However, excluding T_B from the input features degrades the RMSE to 2.56 $kg\ m^{-2}$, i.e. + 67 % error compared to the best version, which clearly shows that some information is extracted from the brightness temperature in the retrieval.

---

13. *Lines 196 and following: When you mention the RMSE of* $86gm^{-2}$*, I assume you are applying the retrieval derived from the equally LWP-distributed training data set to the equally distributed test data set? I'm not sure.. please make clear.*

Indeed, the RMSE is calculated on the testing set which was preprocessed in the same way as the training set. This was not clear enough in the initial version of the text so the following modifications were added:

[Beginning of section 4.3] After preprocessing, LWP and IWV datasets were randomly split into training, validation and testing set (70 %-15 %-15 %) (...)

[Section 5.1.2] Let us underline that the subsampling which is performed on the dataset for the retrieval of LWP is applied to training, validation and testing sets: the results that are presented here are therefore computed on the testing set with a truncated distribution – i.e. after subsampling.

---

14. *Lines 199 – 201: You write: "with however a bias for low LWP values, which are slightly overestimated, and for large LWP (>* $800gm^{-2}$*) which are underestimated". Please apply my comment to Lines 188-190.*

As suggested by the reviewer, two subplots were added that illustrate the bias as a function of binned LWP and IWV. Please refer to the new figure (figure 5). The following sentence was added to the caption:

Similarly, panels (e) and (f) show the distribution of mean bias across the range of IWV and LWP values.

---

15. *Line 211: You write: "but only in a minor way does it increase the LWP retrieval's accuracy". But it does!? Going from noERA-noIWVpred-noGeo-noSurf to ERA-noIWVprednoGeo- noSurf reduces the RMSE from roughly 140 to 90* gm^{-2}*. Or am I misinterpreting something wrong here?*

Fig. 6 was plotted again, with the most recent version of the algorithm including the modifications that were implemented following the reviewers' comments. The labels were made clearer. In the new figure, it is clearly visible that ERA5 input to the LWP retrieval does not bring a major improvement if it is used together with other input features. However, if considering only

TB+ERA as input features, this performs significantly better than TB alone.

Still, the accuracy of the algorithm drops severely when no other features are considered than brightness temperature (RMSE of 140 g m^{-2}). This means that, albeit second-order when taken individually, and somehow redundant when all used together, the secondary input features are efficient in incorporating statistical trends and climatological information to the retrieval during the training phase.

---

16. *Section 5.2: I like the idea of showing the sensitivity to instrument calibration offset, but only when I look at Figure 7, do I only see that you have looked at the effects for all different retrieval configurations and for continuously rising TB offset. Here again: please describe and discuss your results with more detail. To make your discussion complete, please add the "only-TB" retrieval to Figure 7a and 7b.*

Fig. 7 was updated and a little alleviated to make it more readable. The description and discussion of the figure was extended.

In order to assess the stability of the algorithm with respect to potential miscalibration or calibration drift of the radiometer, T_B offsets were virtually added to the testing dataset before implementing the retrieval. Figure 7 illustrates the behavior of the algorithm when such a miscalibration, when a constant TB offset is present (varying from 0 to 5K). Panel a) shows that a 5 K offset in T_B results in a 30 % increase in RMSE for the IWV estimations, which is non-negligible. Ensuring proper radiometer calibration thus seems crucial in constraining the error of this retrieval. For comparison, the 89-GHz radiometer presented in Kuechler et al, 2017 has a nominal accuracy of 0.5 K, after calibration. If the calibration cannot be ensured, and if there is no means to correct for miscalibration ($>$ 3K), it is preferable to use the algorithm that does not rely on TB, shown with the black dashed line.

---

17. *Section 5.3: Do you have any interpretation as of why the results over the Indian Peninsula are so much worse than elsewhere, even compared with sites at similar latitude? Is it possible that this is associated to the quality of the radiosondes or is there any other reason you can think of?*

18. *Line 242: Please describe how you think "humidity and temperature conditions" can lead to the discrepancy.*

We address those two comments jointly.
It is difficult to identify the most important factor that causes a high error in these areas. The underrepresentation of this range of latitudes and corresponding climates is most likely a critical point: unfortunately, there is to our knowledge little data available from these areas. The quality check that we performed on the radiosonde data did not suggest that there is a signficant and systematic difference in the quality of the data, although this cannot be fully excluded.

The temperature and humidity conditions, as well as the strong precipitation events that typically occur in those regions, are probably responsible for this discrepancy. Cases with high LWP

are more common under such climatic conditions, and it was observed in Sect. 4.1.2 that the accuracy of the algorithm decreases in that range. Tropical climates are underrepresented in the dataset, for little data is available from this region in comparison with mid-latitude areas: their specificity might therefore not be fully captured during the learning stage of the algorithm. This accounts at least partly for the enhanced error over the Indian peninsula and South-Eastern Asian islands.

---

19. *Figure 9: Can you explain the outliers (especially the vertical and horizontal "bar structures") in Figures 9a and 9b?*

We do not have clear-cut answer for this, but a few hypotheses that were added in the discussion (see below). Let us point out that the color scale is logarithmic and that these outliers are far less numerous than the bulk of the data, especially in the case of IWV.

Additionally, outliers are visible as vertical and horizontal bars close to the axes, for which two hypotheses are considered. One is that the distance between the two instruments was big enough, that in some cases a liquid water cloud would overpass one of the two instruments but not the other. Hence, HATPRO would measure a non-zero LWP while WProf would indicate a clear sky, or vice-versa. Also, measurement artefacts cannot be excluded, e.g. due the persistence of a liquid water film on the radome of either radiometer, after precipitation or due to condensation.

---

20. *I'm missing a discussion of Figure 11. One point would be, e.g. as also seen in Figure 6a, that the reanalysis performs better than the TB-only retrieval. Here, again it would help if you added a retrieval derived without any TB to discuss the overall TB value. Comparing such a retrieval against your ERA-Geo-Surf would tell you something about the impact of 89 GHz TB and if it's sensible to use it at all if you have the other parameters available.*

The retrieval without TB was included. Although brightness temperature does not have a similar impact as for LWP where it is the primary feature, it does improve the accuracy. It seems to have a similar weight as the other groups of features. In particular, it seems to increase the sensitivity to short temporal changes (visible in the comparison against HATPRO measurements). When averaging over 30 minutes, including TB does not result in significant improvement.

The top panels of Fig. 10 (which illustrates the error vs. HATPRO measurements) and Fig. 11 (error vs. value derived from radiosounding) shows that overall, the implementation of the different versions of the algorithm on the Payerne dataset matches the conclusions from the testing set results: more features lead to an enhanced precision of the retrieval. The accuracy drops when only one or two groups of input features are included, but no single group of features seem to increase the accuracy alone. There is however a difference between Fig. 10 b) and Fig. 11 b): in the latter, higher R^2 (and similar RMSE) is actually obtained from the algorithm that does not use T_B in input, than with the full set of input features. This is at first surprising, but it was explained by taking a closer look at the results: the algorithm without T_B leads to IWV values which are more smooth and less sensitive to short-time variations. These are not reflected in the comparison against radiosonde data, for which a 30-min averaging was implemented. When variations over a small timeframe are considered, the inclusion of T_B improves the retrieval, as comes across from the comparison against HATPRO's measurements in Fig. 10.

21. *Line 261: Please quantify the "constant bias"*

The value of this bias (negative bias of $-1.8kg\ m^{-2}$) was added, as well as a more quantitative description of the ERA5 bias over Payerne.

From Fig. 9 a) and c) it appears that the IWV retrieval has relatively limited spread but has a constant bias (-1.8 kg m^{-2}), which is visible both in the comparison against HATPRO (a) and radiosonde-derived measurements (c). This might be due to a bias in ERA5 data during this timeframe over the region (with a value of -4.1 kg m^{-2}), which is visible in ERA5 records during the entire campaign (not shown here) and for which there is no clear explanation at this stage.

---

22. *Lines 296 and following: In Fig 12, I'd also include results from a retrieval without TB. Another possibility for ERA5 outperforming the retrievals could maybe be fog? Could you include a discussion of the weather during ICE-POP?*

Figure 12 was also updated with the no-TB retrieval, which in the case of ICEPOP and the comparison against radiosonde measurement outperforms the other algorithms. A discussion on this is added to section 6.2.
Regarding weather conditions, a brief description is added to Sect. 2.2.3 and the topic is discussed further in Sect. 6.2.

[Section 2.2.3] A description of the data is presented in Gehring et al, 2020. During this campaign, the weather was generally cold and dry; nine precipitation events were recorded, and occasional fog was present (about 25 occurrences during the campaign timeframe).

[Section 6.2] Possibly, the dry and cold weather that was observed during the ICE-POP campaign featured little short-term variability and was associated with stable atmospheric conditions that were particularly well captured in ERA5 reanalyses. Snowfall events during the campaign, as well as occasional fog, can also bias the retrieval by enhancing brightness temperature.
The analysis of the ICE-POP data was taken a step further to explore the latter point. It appears that the IWV retrieval is most reliable in non-precipitating or cold conditions (...)

---

23. *Section 7: Can you elaborate a little on what the alternative would be to using the re-analysis? If, e.g. you would need quasi real-time retrieval results.*

A possibility would be to use, instead of reanalysis data, the output of forecast models (global models like IFS or GFS for instance, or regional models if available). They are likely to be reasonably accurate especially within a few hours of the model run. However, their performance might not be excellent in remote locations were few ground measurements and/or radisoundings are available. This drawback is also true for reanalysis data, although perhaps to a smaller extent.

Additionally, the retrieval that is presented here uses reanalysis data as an optional feature, that was shown to be very valuable. In the case where near real-time retrievals were necessary, the user could choose to use a version of the algorithm that does not rely on ERA5, but this would be detrimental, especially for the IWV retrieval. Another option would be to implement the algorithm with the output from forecast models (IWV and LWP) instead of reanalysis data. This approach was however not implemented as this stage.

---

24. *Figure 2: X-axis labelling of Figure 2b needs to read "LWP", not "IWV"*

We thank the referee for noting this mistake, which was corrected for.

---

25. *Figure 6: Do you need to make the bars orange with diagonal lines – just orange would probably make the text easier to read?*

The visual aspect of Fig. 6 was modified to make it more readable.

---

26. *Figure 9: It would help if you included in-plot statistics such as number of cases, bias, RMSE and R^2. Best do consistently with Figures 5 and 13.*

Following this suggestion, an annotation was added to the density plots and scatter plots (figures 5, 9 and 13) with statistical information: the number of plots, the root mean square error, as well as bias and correlation coefficient. The following sentence was added to the captions:

The size of the data set is indicated (N) as well as relevant error metrics (RMSE, mean bias, R^2).

---

27. *Figure 10: I think the ordering of the text in the bars is swapped, otherwise the plots make no sense to me.*

There was indeed a big mistake in the labels of the figure that was uploaded, the legend of the bars was swapped. This was corrected for.