An empirical parameterization of the subgrid-scale distribution of water vapor in the UTLS for atmospheric general circulation models

Audran Borella[1], Étienne Vignon[2], Olivier Boucher[1], and Susanne Rohs[3]

[1]Institut Pierre-Simon Laplace, Sorbonne Université / CNRS, Paris, France
[2]Laboratoire de Météorologie Dynamique, IPSL, Sorbonne Université / École Polytechnique / ENS, CNRS, Paris, France
[3]Institute of Energy and Climate Research 8 - Troposphere, Forschungszentrum Jülich GmbH, Jülich, Germany

Key Points:

- The fine-scale distribution of water vapor in the upper troposphere is fully parameterized for use in atmospheric general circulation models
- The parameterization is empirical and based on 257 millions airborne observations made between 1995 and 2021
- This parameterization aims to improve the quality of the simulation of ice supersaturation, and the formation of cirrus and contrail clouds

Corresponding author: Audran Borella, audran.borella@ipsl.fr

Abstract

Temperature and water vapor are known to fluctuate on multiple scales. In this study 27 years of airborne measurements of temperature and relative humidity from IAGOS (In-service Aircraft for a Global Observing System) are used to parameterize the distribution of water vapor in the upper troposphere and lower stratosphere (UTLS). The parameterization is designed to simulate water vapor fluctuations within gridboxes of atmospheric general circulation models (AGCMs) with typical size of a few tens to a few hundreds kilometers. The distributions currently used in such models are often not supported by observations at high altitude. More sophisticated distributions are key to represent ice supersaturation, a physical phenomenon that plays a major role in the formation of natural cirrus and contrail cirrus. Here the observed distributions are fitted with a beta law whose parameters are adjusted from the gridbox mean variables. More specifically the standard deviation and skewness of the distributions are expressed as empirical functions of the average temperature and specific humidity, two typical prognostic variables of AGCMs. Thus, the distribution of water vapor is fully parameterized for a use in these models. The new parameterization simulates the observed distributions with a determination coefficient always greater than 0.917, with a mean value of 0.997. Moreover, the ice supersaturation fraction in a model gridbox is well simulated with a determination coefficient of 0.983. The parameterization is robust to a selection of various geographical subsets of data and to gridbox sizes varying between 25 to 300 km.

Plain Language Summary

Temperature and water vapor fluctuate in the atmosphere on different scales, from micrometers to thousands of kilometers. In this study we use airborne measurements of temperature and water vapor to study the spatial variability of humidity in the upper troposphere and lower stratosphere (UTLS). The observations are used to build a simple modelling of water vapor distribution on scales from tens of kilometers to hundreds of kilometers, which is designed to be used in atmospheric general circulation models (AGCMs), the atmospheric components of Earth system models. This new modelling of water vapor fluctuations aims to increase the physical representation of cirrus clouds and aviation-induced cloudiness in AGCMs. The observed water vapor distributions are modelled with a beta distribution, whose parameters are completely determined as empirical functions of two major variables of AGCMs, the average temperature in a gridbox, and the average water vapor in a gridbox. Overall, the modelled distributions fit very well those observed.

1 Introduction

Ice supersaturation is an ubiquitous phenomenon in the upper troposphere whereby the partial pressure of water vapor is higher than the saturation value with respect to the ice phase, thus being thermodynamically unstable (Gierens et al., 2012). Ice supersaturated regions (ISSRs) occur at temperatures lower than 273.15 K (0°C), with a lifetime that can be as long as 24 hours (Irvine et al., 2014) and spatial scales that vary from tens of kilometers to a thousand of kilometers (Spichtinger & Leschner, 2016). Formation, extent, and lifetime of ISSRs are affected by vertical motions associated with convective systems or extratropical cyclones as well as small-scale gravity waves and turbulence (Gierens et al., 2012; Kärcher et al., 2014). They are also strongly affected by the weather pattern, and thus highly vary in space and time (Lamquin et al., 2012). These regions are a prerequisite for the formation of persistent condensation trails created by aviation, which themselves have a significant impact on the climate (Schmidt, 1941; Appleman, 1953; Lee et al., 2021). Indeed, aircraft fly in the upper troposphere and lower stratosphere (UTLS), a region where ISSRs occur frequently, but with a high spatial and seasonal variability (Spichtinger et al., 2003).

Natural cirrus can also form *in situ* in ISSRs. Although they play a major role in the radiative balance of the Earth, climate feedbacks involving such clouds are still uncertain (Ceppi et al., 2017; Kärcher, 2017; Hill et al., 2023). This feedback can be estimated using atmospheric general circulation models (AGCMs), by simulating the changes in cloud radiative effect in a warming climate. One of the main and long-standing challenge of current AGCMs is the accurate representation of the formation and evolution of clouds. As the spatial scale of cloud processes is much smaller than the size of an AGCM gridbox, they must be parameterized. AGCMs generally consider a distribution of water inside each gridbox, which may operate on the total water (e.g., Smith, 1990; Bony & Emanuel, 2001; Tompkins, 2002) or only the water vapor in the clear-sky part of the gridbox (e.g., Tiedtke, 1993; Tompkins et al., 2007; Muench & Lohmann, 2020). An associated probability density function (PDF) can then be used to diagnose cloud properties within a gridbox, by calculating the quantity of water inside the newly formed clouds as well as the corresponding fraction of the gridbox occupied by clouds. Indeed, all the water that is distributed beyond a given threshold called the condensation threshold is converted into cloudy water, and some properties of the formed clouds can also be inferred from the distribution. A diagnostic scheme generally uses a total water distribution and diagnoses cloud amount and properties at each timestep. On the contrary, a prognostic scheme, which considers a balance of cloud formation and destruction terms at each timestep, often only requires a distribution of the water vapor in the clear-sky part of the gridbox.

At temperatures higher than 273.15 K (0°C), all the water that exceeds liquid saturation can be considered to be instantly condensed into clouds through a saturation adjustment process (Pruppacher & Klett, 2010). Therefore, the distribution of water vapor is easily separated into a clear and a condensed part, with the liquid saturation as a threshold. On the contrary, ice crystals can be formed within ISSRs from various processes and at various supersaturation levels. For example, in the so-called cirrus temperature regime below 235 K (−38°C), ice crystals can form through either homogeneous or heterogeneous nucleation (Kärcher, 2003). Which process is involved depends on the supersaturation level, the temperature, and the quantity and properties of the atmospheric aerosols. Homogeneous nucleation, which in this range of temperature refers to the homogeneous freezing of solution aerosol droplets, occurs in the absence of ice nucleating particles (INPs) and if supersaturation is high enough. The supersaturation value above which homogeneous nucleation occurs, often referred to as the Koop's threshold, depends on e.g., the ambient temperature, the aerosol particle size and activity, and their chemical composition (Koop et al., 2000; Gierens, 2003; Vignon et al., 2022; Baumgartner et al., 2022), but this threshold can be reduced to an approximate function of temperature only (Ren & Mackenzie, 2005). Heterogeneous nucleation, which here refers to heterogeneous freezing of solution aerosol droplets, can nevertheless occur at supersaturations comprised between the saturation and the homogeneous nucleation threshold, depending on the quantity and properties of INPs (Kärcher et al., 2022). There is therefore a continuous range of thresholds for condensation, instead of a single one as it is the case for temperatures higher than 273.15 K. A cloud formation parameterization thus requires the knowledge of the distribution of supersaturation to take into account a continuous range of potential condensation thresholds. The parameterization of the distribution of water vapor must therefore be as highly representative as possible of the true state of the atmosphere.

Previous studies have focused on the distribution of humidity in the UTLS, from observations or simulations (e.g., Gierens et al., 1999; Tompkins, 2002; Reutter et al., 2020; Petzold et al., 2020). However, they characterize the global distribution of humidity and its dependence on the season, altitude or geographical location. In contrast, a parameterization in an AGCM is built from state variables of the model, such as temperature or specific humidity within each gridbox at each timestep, rather than season, altitude, latitude or longitude. Therefore, these distributions from previous studies are

of great help to evaluate the outputs of an AGCM, but are not adequate to design a new physically-based parameterization.

This study thus aims to parameterize the fine-scale distribution of clear-sky water vapor from observations for AGCMs resolution scales, typically from 25 to 300 km. The only explanatory variables used in the parameterization are two of the prognostic variables of an AGCM, namely temperature and specific humidity. The objective is to build a sub-grid scale distribution which can be used for further parameterization development in such models. It is not straightforward that gridbox-averaged temperature and specific humidity are sufficient to faithfully and exhaustively represent the variability of water vapor within an AGCM gridbox. However, we refrain to use other variables than prognostic variables of an AGCM in order to build a cloud parameterization that does not depend on other physical parameterizations, which may introduce additional inter-dependencies, and, among prognostic variables, temperature and specific humidity are essential in explaining the origin of the variance of water vapor (Gierens et al., 2007).

Most types of observations at high altitude, such as measurements from airborne campaigns or radiosoundings, are usually sparse temporally and geographically (Krämer et al., 2020; Wolf et al., 2023), which prevents deriving significant statistics at climate temporal and spatial scales. We use here the IAGOS (In-service Aircraft for a Global Observing System) observational product, which is composed of airborne measurements made in a wide geographical zone and provides data since August 1994, thus increasing considerably the statistical representativity of our results. IAGOS is a European Research Infrastructure for global observations of atmospheric composition from commercial aircraft (Petzold et al., 2015, 2017). It is composed of a few commercial aircraft equipped with multiple sensors, in particular humidity and temperature sensors. The main advantage of IAGOS is the large number of *in situ* measurements within the UTLS in all seasons, which we use to construct the distributions. Recognizing the importance of these data, Gierens et al. (1997) and Gierens et al. (2007) also used the IAGOS dataset to investigate mesoscale distributions of humidity in the UTLS, but those distributions were not analyzed for AGCM parameterization purposes.

The paper is structured as follows. In Section 2 we present the IAGOS dataset and the methodology we use to analyse the data. In Section 3 we present the distributions as observed by IAGOS and analyse their main properties, as well as the dependence of their standard deviation and skewness to the larger-scale temperature and humidity. In Section 4 we derive the parameterization and conduct sensitivity studies. Results and assumptions are discussed in Section 5, and Section 6 closes the paper with a summary and conclusions.

2 Dataset and methods

2.1 IAGOS dataset

IAGOS measurements include concentrations of different gases (e.g., CO, O_3, CO_2, water vapor), aerosol, dust, cloud particles, and basic meteorological data. Two main datasets comprise IAGOS: MOZAIC (1994-2014) and IAGOS-CORE (from 2011). In our study, we use all available data in the 175–325 hPa pressure range from flights that took place between 1995 and 2021. Measurements are mainly performed in the mid-latitude Northern Hemisphere, especially in the North Atlantic corridor (Fig. 1), but we investigate the spatial dependency of our results in Sections 3.3 and 4.4.

The IAGOS measurements are not specifically made in clear-sky, and pre-2017 data lack differentiation between cloudy and clear-sky conditions. From 2017 onwards, a backscatter cloud probe was installed on some aircraft, allowing for such a differenciation. About 5 % of those measurements sampled cloudy air at most (Sanogo et al., 2023). Be-

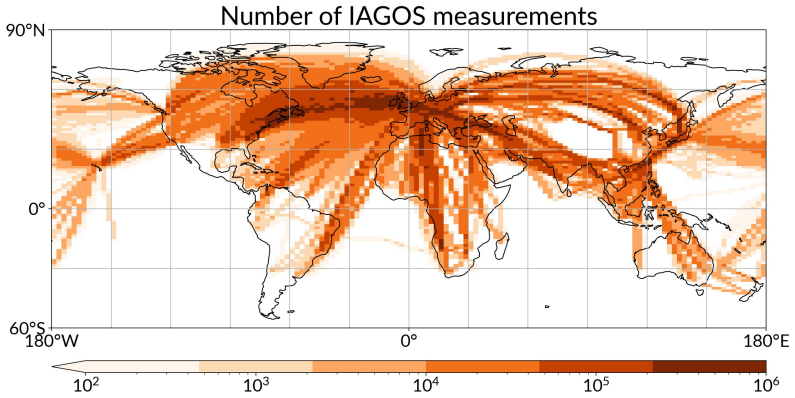

Figure 1. Number of measurements per box from the IAGOS infrastructure around the globe, between 1995 and 2021. The boxes measure 2.5° in longitude and 1.3° in latitude. Note the logarithmic scale.

cause this percentage is limited and that we can only differentiate clear- and cloudy-sky for a limited subset of the data, we assume that the measurements are all mostly representative of clear-sky conditions, and use all the available data to keep a high number of measurements.

2.2 Data processing

Relative humidity and temperature are measured with the ICH (IAGOS Capacitive Hygrometer) instrument, which consists of a capacitive relative humidity sensor and a platinum resistance sensor for the temperature measurement at the humidity sensing surface. Uncertainties stem from various sources, including instrumentation limitations and environmental conditions during data collection. The absolute uncertainty on temperature measurements is estimated as 0.5 K and the relative uncertainty on relative humidity w.r.t. liquid measurements as 5-6 % (Petzold et al., 2015; Rolf et al., 2023). The ICH samples relative humidity and temperature every 4 s, corresponding to about 1 km in flight. The temperature sensor has a time resolution of 4 s, however the humidity sensor has a time resolution which varies from 1 s at 300 K to 120 s at 210 K, the latter corresponding to about 30 km in flight (Neis et al., 2015).

As this study aims to investigate the variability of moisture on scales between 25 and 300 km, the low time resolution of the humidity sensor is a major concern. To study the subgrid-scale distribution of water on scales as low as 25 km, the measurements must be independent on scales much lower than 25 km. Previous studies performed a running mean on the data to dampen the lag effect (e.g., Gierens et al., 2007). However, such an averaging process reduces the water vapor variability and smoothes out peaks in the data that correspond to realistic small spatial scales fluctuations. To address this problem, we developed a reconstruction algorithm which partly solves the time resolution issue. The rationale of the algorithm is to reconstruct high-resolution data from the measured data, which is shown to differ by a temperature-dependent exponential moving average (Neis et al., 2015). Following Neis et al. (2015), the relationship between the time response of the sensor and the temperature is found using collocated measurements of reference high-resolution sensors that were conducted in specific dedicated campaigns for a few flights. The application of this reconstruction algorithm leads to a significant improvement of the data quality, as further detailed in Appendix A.

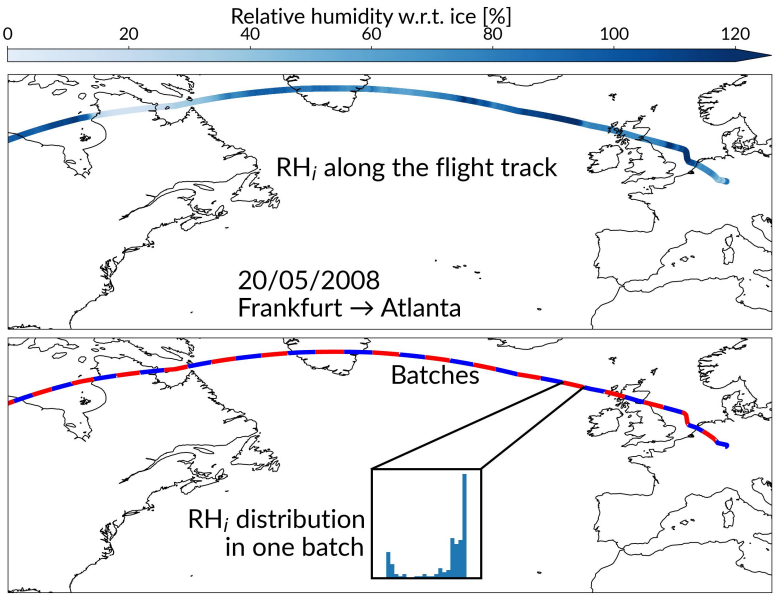

Figure 2. Illustration of the methodology used for grouping the data. Measurements from a given flight (a) are grouped into batches of length L_{batch} (b) and pooled into the corresponding $(\overline{T},\overline{\text{RH}_i})$ bin on the basis of their average values.

Data are then screened with different filters in order to remove unreliable values. The measurements are provided with a quality flag, and data for which temperature and relative humidity are not labelled as "good" or "limited" are discarded from the analysis. The reconstruction algorithm is then applied to the time series of water vapor, which neither adds nor deletes any data point, and which is not affected by data gaps. Next, the measurements are screened to the 175–325 hPa pressure layer. Two additionnal filters are then applied: (1) the first one removes data when the aircraft is climbing or descending too fast because we want statistics at a given pressure level, and (2) a second filter removes data for which there is less than 0.5 measurement per km available, to ensure that statistics are significant enough. The dataset after screening is composed of 60,304 flights, for a cumulative number of 257 million measurements. The time series of the measured relative humidity w.r.t. liquid is then converted into a time series of specific humidity q, knowing temperature and pressure for each measurement.

Each flight trajectory is then subdivided into batches of a fixed length L_{batch}, set at 200 km (Fig. 2). These batches represent the approximate size of a typical gridbox of an AGCM, and the sensitivity of our results to L_{batch} is investigated in Section 4.4. We compute the average temperature $\langle T\rangle$ and average specific humidity $\langle q\rangle$ in each batch. Hereinafter, we will consider the variable $\langle \text{RH}_i\rangle$, which we define such that:

$$\langle \text{RH}_i\rangle = \langle \text{RH}_i\rangle\left(\langle T\rangle, \langle q\rangle, \langle p\rangle\right) = 100 \cdot \frac{\langle p\rangle \cdot \langle q\rangle}{\varepsilon\ p_{\text{sat},i}\left(\langle T\rangle\right)}, \tag{1}$$

where ε is the ratio between the molar mass of water and that of dry air, $p_{\text{sat},i}$ is the saturated pressure w.r.t. ice (Sonntag, 1990), and $\langle p\rangle$ is the average pressure in the batch. We use $\langle \text{RH}_i\rangle$ because it gives direct information about cloud formation and saturation. Since $\langle \text{RH}_i\rangle$ is computed from $\langle T\rangle$, $\langle q\rangle$ and $\langle p\rangle$ directly, $\langle \text{RH}_i\rangle$ and $\langle q\rangle$ can be estimated from one another in an AGCM gridbox without further assumption.

$\langle T\rangle$ is associated to the corresponding temperature bin j centered around $\overline{T}^j$ with a width of 0.5 K $[\overline{T}^j - 0.25$ K, $\overline{T}^j + 0.25$ K$]$, where $\overline{T}^j$ varies from 200.25 to 254.75 K

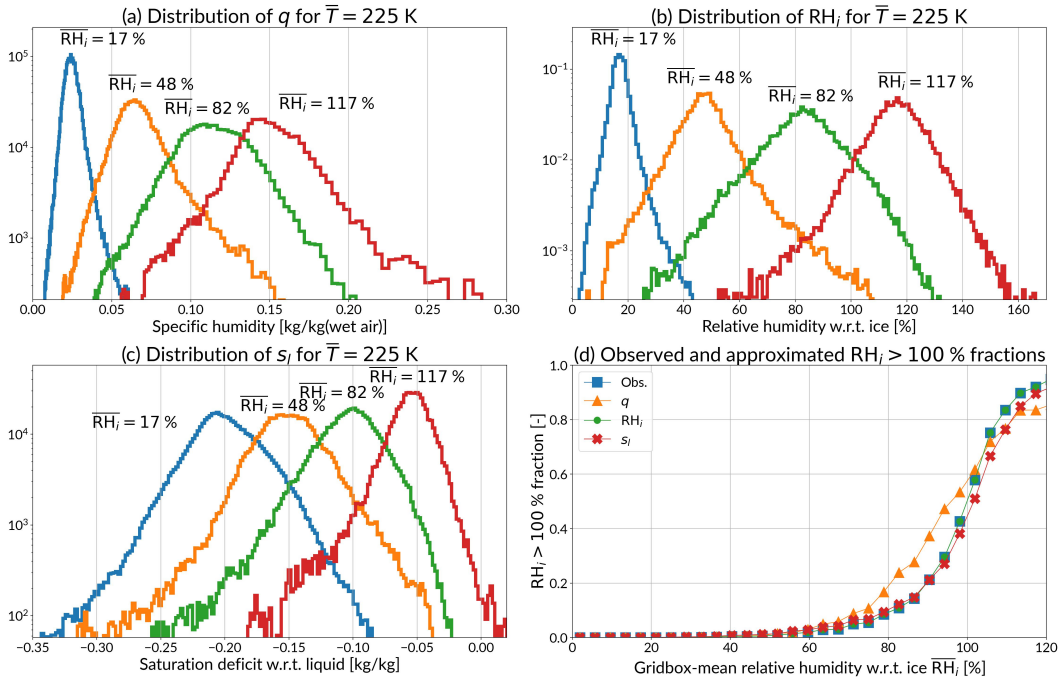

Figure 3. Distributions of (a) specific humidity, (b) relative humidity w.r.t. ice, (c) liquid saturation deficit for four bins of $\overline{\mathrm{RH}_i}$ (colored), and (d) observed and approximated ice supersaturation fractions when $\overline{T}$ is used to compute the saturation threshold for each humidity variable, with $\overline{T}$ fixed to the 225 K bin.

for a total of 110 different bins. Similarly, $\langle \mathrm{RH}_i \rangle$ is associated with a humidity bin, whose average value and width depend on $\overline{T}^j$. When converted to relative humidity w.r.t. liquid, this k bin, centered aroung $\overline{\mathrm{RH}_l}^k$ with a width of 2.4 %, is $[\overline{\mathrm{RH}_l}^k - 1.2\ \%, \overline{\mathrm{RH}_l}^k + 1.2\ \%]$, where $\overline{\mathrm{RH}_l}^k$ varies from 1.2 to 106.8 % for a total of 45 different bins. The corresponding binned average value is noted $\overline{\mathrm{RH}_i}^k$ when converted to relative humidity w.r.t. ice.

We retrieve the measurements of temperature and specific humidity for each batch and associate it with the binned average values $\overline{T}^j$ and $\overline{\mathrm{RH}_i}^k$, hereinafter simply referred to as $\overline{T}$ and $\overline{\mathrm{RH}_i}$. This is done for all batches in each flight, for all the flights in the dataset. The measurements are then used to calculate the distribution of humidity for each ($\overline{T}$, $\overline{\mathrm{RH}_i}$) bin.

3 Analysis of the distributions of water vapor in the UTLS

3.1 Example of individual distributions

We obtain distributions of specific humidity q, relative humidity w.r.t. ice RH_i, and liquid saturation deficit s_l, three humidity variables commonly used in cloud formation schemes of AGCMs for ($\overline{T}$, $\overline{\mathrm{RH}_i}$) bins. In this section, we illustrate the results for four values of average humidity covering a wide range, with $\overline{T}$ arbitrarily fixed to the 225 K bin. The distributions are analysed using their scale, which is quantified using standard deviation, and their shape, which is quantified by skewness.

The scale and shape of the distributions for all three humidity variables highly depend on $\overline{\mathrm{RH}_i}$ (Fig. 3a,b,c). The standard deviations of the distributions of q, RH_i, and

s_l, vary from 0.001 to 0.04 g.kg^{-1}, from 1 to 17 %, and from 0.02 to 0.03 g.kg^{-1}, respectively. This concurs with the result obtained by Gierens et al. (2007) that additionally showed that, for a fixed $\overline{\mathrm{RH}}_i$, standard deviation strongly depends on $\overline{T}$ (see their Fig. 7).

However, the dependence of scale and shape to $\overline{\mathrm{RH}}_i$ is different for the three humidity variables. The standard deviation of the q distributions increases with $\overline{\mathrm{RH}}_i$, and the skewness decreases but remains positive (Fig. 3a). The skewness of the RH_i distributions also decreases, but reaches 0 around $\mathrm{RH}_i = 80$ % and becomes negative at higher values of RH_i (Fig. 3b). The corresponding standard deviation increases, peaks and then decreases, with a maximum value also reached around $\mathrm{RH}_i = 80$ %. Contrarily to q, the distributions of s_l show a decreasing standard deviation with increasing $\overline{\mathrm{RH}}_i$, and a negative and decreasing skewness with increasing $\overline{\mathrm{RH}}_i$ (Fig. 3c).

Those differences stem from local fluctuations in temperature, as pressure is fixed. From a modelling perspective, this implies that the choice of the humidity quantity has an impact on the cloud formation scheme, because the saturation threshold is usually computed using $\overline{T}$. To further illustrate this, we calculate the approximated ice supersaturation fraction of each distribution of humidity using $\overline{T}$ instead of local T to compute the saturation threshold, as usually done in AGCMs (Fig. 3d). For low values of $\overline{\mathrm{RH}}_i$, the approximated fraction is almost zero for all three humidity variables. However, for $\overline{\mathrm{RH}}_i > 70$ %, the fraction computed using q distributions can be more than twice the one observed. This suggests that q is the less adapted humidity variable among the three to approximate ice supersaturation fraction in AGCMs. In the following sections we conduct the analysis using RH_i.

3.2 Statistics of the distributions

To generalize our findings, we systematically study the distributions as functions of $\overline{T}$ and $\overline{\mathrm{RH}}_i$. Fig. 4a shows how the 257 millions observations are distributed across the 4950 ($\overline{T}$, $\overline{\mathrm{RH}}_i$) bins. The highest numbers of measurements are found at low $\overline{\mathrm{RH}}_i$, and are usually made in the lower stratosphere. Observations at $\overline{\mathrm{RH}}_i$ higher than 40 % are usually made in the upper troposphere or at the tropopause. The presence of very few data points above the liquid saturation curve highlights the uncertainties of the IAGOS data, showing sometimes non-physical humidity values. In the following, we study the standard deviation and skewness of the distributions as a function of $\overline{T}$ and $\overline{\mathrm{RH}}_i$. These statistics are computed for bins with more than 2516 measurements available, such that all the bins above the homogeneous nucleation curve are discarded.

Standard deviation (σ) is a quadratic function of $\overline{\mathrm{RH}}_i$ for a fixed $\overline{T}$, peaking around $\overline{\mathrm{RH}}_i = 70$ % and falling to 0 % at $\overline{\mathrm{RH}}_i = 0$ %, and to about 10 % at $\overline{\mathrm{RH}}_i = 110$ % (Fig. 4b,d). At saturations higher than 110 %, the pattern of σ is noisier, yet increases with increasing $\overline{\mathrm{RH}}_i$. The shape of the described pattern is the same for all values of $\overline{T}$, but the maximum of σ highly depends on $\overline{T}$, ranging from 6 % at 205 K, to about 23 % at 245 K.

Skewness is linear between $\overline{\mathrm{RH}}_i = 0$ % and 110 % for a fixed $\overline{T}$, decreasing from about 5 to a value between 0 and -1 (Fig. 4c,e). Its zero value is reached at about the same $\overline{\mathrm{RH}}_i$ value than the peak value of σ. For $\overline{\mathrm{RH}}_i > 110$ %, the pattern of skewness becomes noisier, but reaches a plateau. The slope of the linear dependence of skewness to $\overline{\mathrm{RH}}_i$ does not depend on $\overline{T}$, but its intercept does, decreasing by about 1 unit when $\overline{T}$ increases from 210 K to 245 K.

3.3 Spatial sensitivity of the statistics

The distribution of RH_i highly depends on the pressure level relative to the tropopause and on the geographical zone considered (Reutter et al., 2020; Sanogo et al., 2023). As we aim to parameterize the distribution of RH_i for an AGCM as a function of $\overline{T}$ and $\overline{\mathrm{RH}}_i$

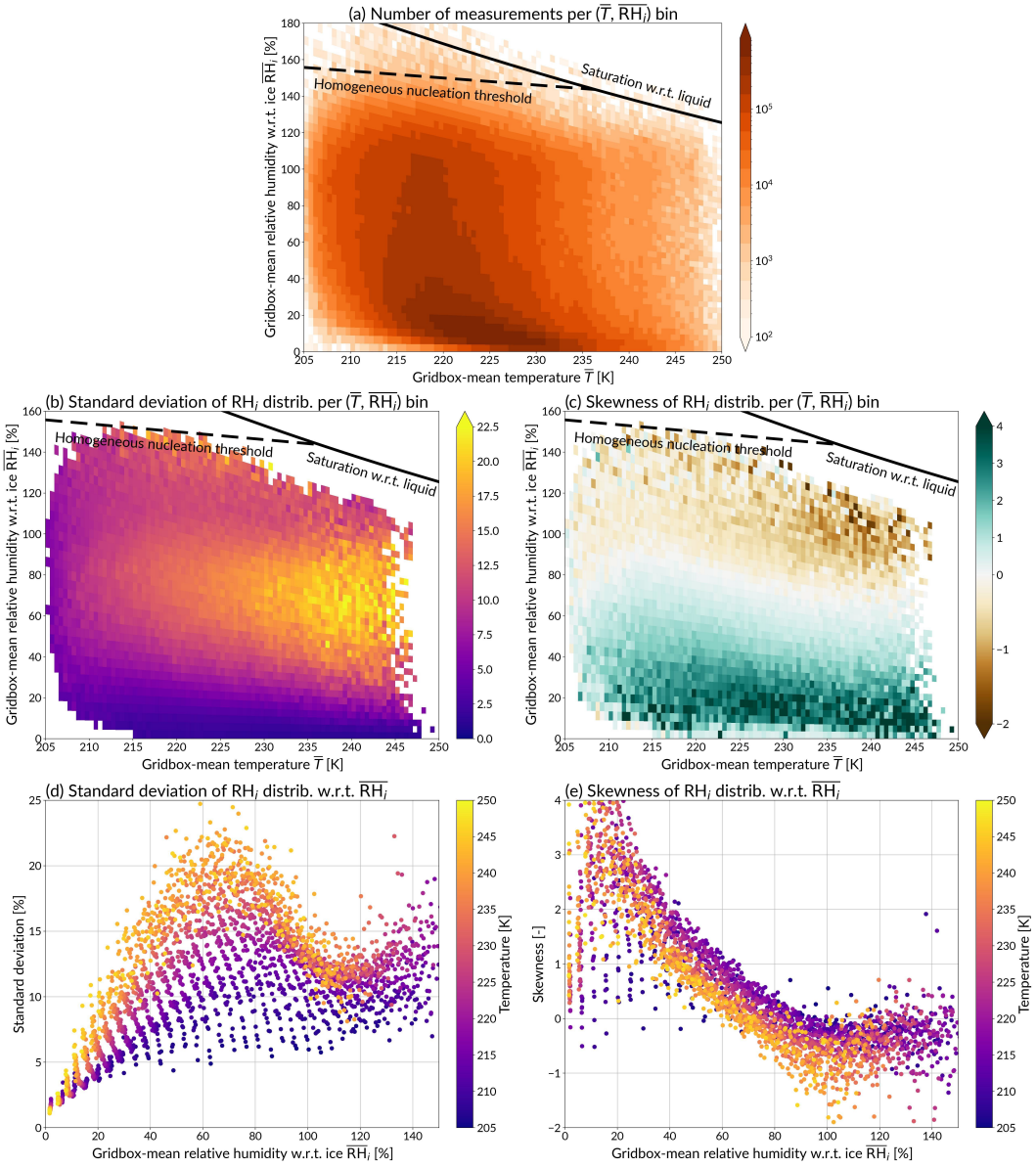

Figure 4. Number of measurements (a), standard deviation (b and d) and skewness (c and e) of the distributions as functions of $\overline{T}$ and $\overline{\mathrm{RH}}_i$. The first and second rows show the quantities in the ($\overline{T}$, $\overline{\mathrm{RH}}_i$) plane, while the third row shows them as a function of $\overline{\mathrm{RH}}_i$ only, with $\overline{T}$ indicated through the color scale. The saturation w.r.t. liquid (solid line) and homogeneous nucleation (dashed line) threshold are plotted. Homogeneous nucleation depends on other than temperature so data points above the curve are not unphysical. The homogeneous nucleation threshold plotted is the Ren and Mackenzie (2005) fit of the Koop et al. (2000) data.

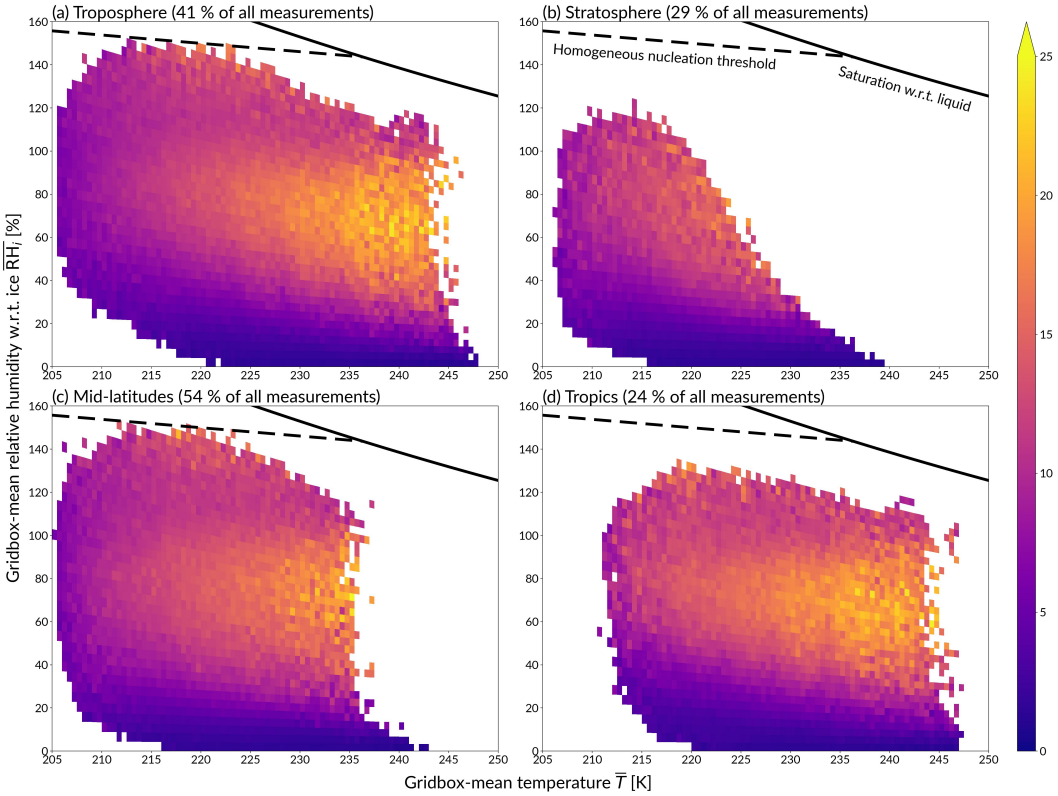

Figure 5. Standard deviation [%] of the RH_i distribution per ($\overline{T}$, $\overline{\mathrm{RH}_i}$) couple, for the upper troposphere (a), the lower stratosphere (b), the NH mid-latitudes (c), and the tropics (d). The lines are the same as in Fig. 4.

only, we conducted sensitivity runs to check the dependency of our findings to the pressure and geographical zone by screening the IAGOS data to four different regions: the troposphere, the stratosphere, the North Hemisphere (NH) mid-latitudes, and the tropics. We call the base case the case with all the data. The troposphere and stratosphere measurements are screened using the chemical definition of the tropopause: the troposphere is associated to ozone concentrations lower than 130 ppb, and the stratosphere to ozone concentrations higher than this value (Bethan et al., 1996; Gierens et al., 1999). We use the ozone measurements from IAGOS for the selection. The NH mid-latitudes are defined by the band of latitudes comprised between 40°N and 60°N, and the tropics by the band of latitudes comprised between 30°S and 30°N. We develop here the results of this sensitivity study for standard deviation, but similar conclusions can be drawn for the skewness (not shown).

For the troposphere, the NH mid-latitudes and the tropics, although the value of the peak has a lower amplitude than in the base case, the results are consistent with our previous findings (Fig. 5a,c,d) and the impact of such a screening is limited. Such a result is *a priori* not obvious since the process at the origin of the humidification of the UTLS at mid-latitudes and tropics are not the same (Gierens et al., 2012).

On the contrary, there is for the stratosphere an increase in standard deviation with increasing $\overline{\mathrm{RH}_i}$, with no peak (Fig. 5b). The results are difficult to generalize, because there is a clear lack of data for most of the ($\overline{T}$, $\overline{\mathrm{RH}_i}$) space, and the uncertainty of the humidity sensor is larger in stratospheric air. This is however not a major issue, as our ultimate goal is to accurately simulate the formation of tropospheric clouds in AGCMs.

Table 1. Compliance of usual probability distributions to our four criteria. Criterion 1: tails can be long. Criterion 2: skewness can be positive and negative. Criterion 3: the distribution can be left-bounded. Criterion 4: moments can be analytically inverted.

Distribution law	Reference example	Criterion 1	Criterion 2	Criterion 3	Criterion 4
Dirac	Lohmann and Kärcher (2002)	X		X	X
Triangular	Smith (1990)		X	X	X
Uniform	Tompkins et al. (2007)			X	X
Gaussian	Muench and Lohmann (2020)				X
Beta	Tompkins (2002)	X	X	X	X
Generalized lognormal	Bony and Emanuel (2001)	X		X	X
Lorentz	Gierens et al. (1997)	X			
Skew normal	-		X		X
Weibull	-	X	X	X	

4 Parameterization of the RH_i distribution

4.1 Fit of the distributions to a usual law

The first step to parameterize the distributions is to find a distribution law which fits well the observations, and which has intrinsic properties that would make it suitable for an implementation in an AGCM. We define four criteria which are based on the results found in the previous section and on other needs. The law needs to allow for a long tail (criterion 1). The law must allow both positive and negative values of skewness (criterion 2). The law needs to be bounded to the left to allow for a physical bound of RH_i (criterion 3). An additional but fundamental criterion is that the moments of the distribution can be analytically inverted, to allow for an implementation in a numerical AGCM (criterion 4). We test various distributions against our criteria in Table 1 and find that the most appropriate law is the beta distribution, which was already proposed by Tompkins (2002) to model total water distributions.

The beta distribution is bounded to the left by a, and to the right by c. Its location parameter is $\mathrm{loc} = a$, and its scale parameter is $\mathrm{scale} = c - a$. The beta distribution is further defined by two shape parameters p and q, so that its probability density function (PDF) is expressed as:

$$\begin{aligned} P_{p,q,a,c}(\mathrm{RH}_i) &= f_{p,q}\left(\frac{\mathrm{RH}_i - a}{c-a}\right)/(c-a) \\ \text{with} \quad f_{p,q}(x) &= \frac{x^{p-1}(1-x)^{q-1}}{\mathcal{B}(p,q)} \\ \text{and} \quad \mathrm{RH}_i &\in [a,c] \end{aligned} \tag{2}$$

where $\mathcal{B}$ is the beta function. The equations linking the mean, the standard deviation and the skewness of the beta distribution to its four parameters can be found in Appendix B.

Each distribution associated with a bin $(\overline{T}, \overline{\mathrm{RH}_i})$ is fitted to a beta law, for which the parameters are found by minimizing the negative log-likelihood function. Additionally, the parameters are constrained to ensure that some of the requirements previously stated are met. The location parameter a must be greater than 0 %, so that no negative value of RH_i is permitted. The shape parameters p and q are arbitratily constrained to be lower than 50, so that the numerical computation of the PDF of the beta distribution does not use high-exponent values. In the following, the quality of the fit is evaluated by computing each determination coefficient $R^2_{\mathrm{fit},k}$ between the cumulative density function (CDF) of the observed and fitted distributions, for the k^{th} bin of $(\overline{T}, \overline{\mathrm{RH}_i})$.

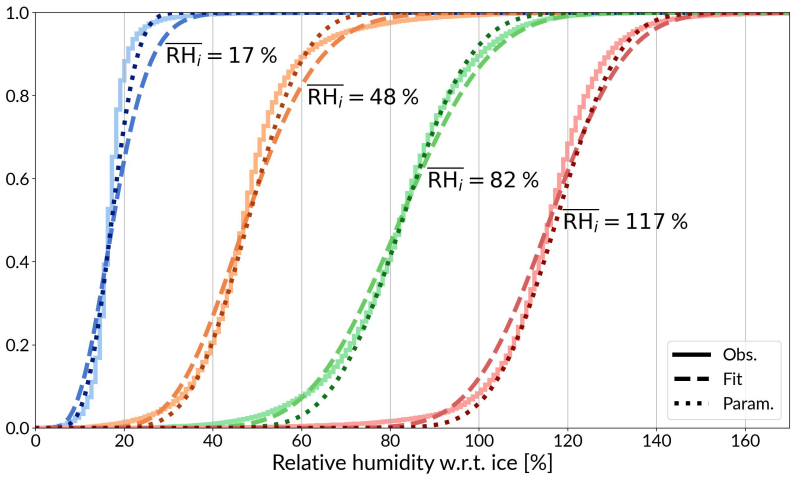

Figure 6. Cumulative density functions (CDF) of the observed distributions (solid lines and light colors), the fitted distributions (dashed lines and medium colors) and the parameterized distributions (dotted lines and dark colors), for four values of $\overline{\mathrm{RH}_i}$ (colored) and $\overline{T}$ fixed to 225 K. The observed distributions are the same as those shown in Fig. 3b.

It is more appropriate to fit the CDF than the PDF because in statistical cloud schemes, cloud formation, supersaturation, cloud water content, are estimated by the integration of the distribution of water vapor, which is the quantity provided by the CDF.

The fit of the observed distributions to the beta law shows good agreement (Fig. 6; compare the solid and dashed curves). The determination coefficients $R^2_{\mathrm{fit},k}$ between the fitted and the observed CDFs are equal to 0.989, 0.997, 0.999 and 0.998 for the four values of $\overline{\mathrm{RH}_i}$ presented, 17 %, 48 %, 82 % and 117 %, respectively. However, the skewness of the fitted distributions are underestimated for distributions with low standard deviations, as it is the case for the $\overline{\mathrm{RH}_i}$ = 17 % distribution. This is because of the constraint on the scale parameters p and q, which limits the capability of the distribution to adapt its shape to match as much as possible the CDF. This difference between the observed and fitted skewness explains the lower $R^2_{\mathrm{fit},k}$ for the $\overline{\mathrm{RH}_i} = 17$ % distribution than for the other three distributions. The determination coefficient $R^2_{\mathrm{fit},k}$ has a mean of 0.997, with a minimum value of 0.97, which depicts a very good agreement overall.

4.2 Empirical formulation of the standard deviation and skewness

In a next step, we seek to express the four parameters as functions of $\overline{T}$ and $\overline{\mathrm{RH}_i}$ in order to have a fully parameterized distribution. The location parameter a is fixed to $\mathrm{RH}_i = 0$ %, because its value was almost always 0 % when fitting the distributions (not shown). The four parameters can be formally expressed as functions of the average, standard deviation and skewness (see Appendix B), and we express these statistical measures as a function of $\overline{T}$ and $\overline{\mathrm{RH}_i}$, instead of the parameters. The average is set to $\overline{\mathrm{RH}_i}$ so as to conserve the water vapor mass. We then construct empirical formulations of the standard deviation σ and skewness γ as functions of $\overline{T}$ and $\overline{\mathrm{RH}_i}$, based on the analysis conducted in Section 3.2. The following paragraphs describe the empirical formulations, and Fig. 7 depicts them for four values of $\overline{T}$ 210, 225, 240 and 255 K.

Following the analysis in Section 3.2, we parameterize the standard deviation σ as a quadratic function of $\overline{\mathrm{RH}_i}$ which peaks at $\overline{\mathrm{RH}_{i,\max}}$, for $\overline{\mathrm{RH}_i}$ lower than a threshold value $\overline{\mathrm{RH}_{i,0}}$. For $\overline{\mathrm{RH}_i} > \overline{\mathrm{RH}_{i,0}}$, we assume a linear relationship between σ and $\overline{\mathrm{RH}_i}$. This lin-

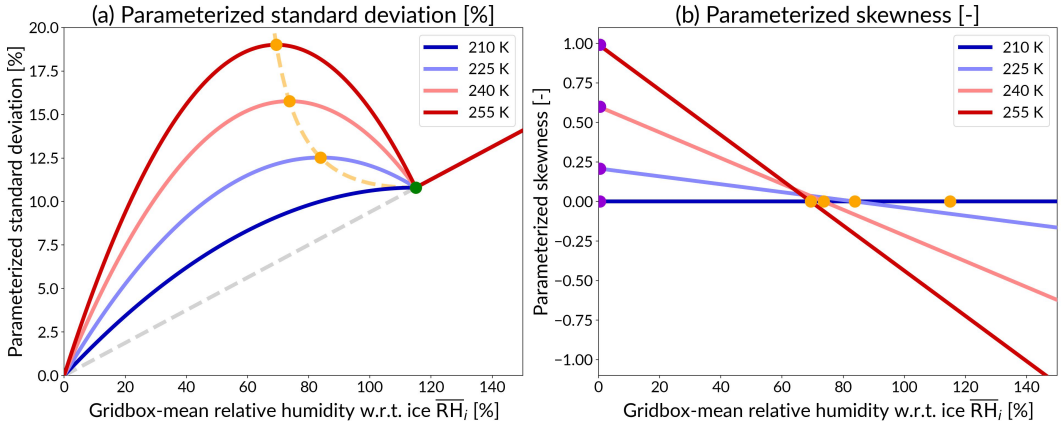

Figure 7. Empirical formulations of (a) the standard deviation and (b) the skewness as a function of $\overline{\mathrm{RH}}_i$ for four values of $\overline{T}$. In (a), the orange dots represent the points ($\overline{\mathrm{RH}}_{i,\max}$,$\sigma_{\max}$) and the orange dashed line its dependence with temperature, while the green dot represents the point ($\overline{\mathrm{RH}}_{i,0}$,$\sigma_0$). The dashed grey line illustrates the linear relationship of standard deviation to the right of the green dot. In (b), the orange dots represent the cancellation points of skewness ($\overline{\mathrm{RH}}_{i,\max}$,0), and the purple dots represent the points at the origin (0,γ_0).

ear relation is drawn in Fig. 7a to the right of the green dot, and it is constructed following the dashed grey line from the origin to the green dot. These different assumptions lead to the following expression for σ:

$$\sigma(\overline{T},\overline{\mathrm{RH}}_i) = \begin{cases} \alpha(\overline{T})\overline{\mathrm{RH}}_i(\overline{\mathrm{RH}}_i - \beta(\overline{T})) & \text{if } \overline{\mathrm{RH}}_i \le \overline{\mathrm{RH}}_{i,0}; \\ \sigma_0 \overline{\mathrm{RH}}_i / \overline{\mathrm{RH}}_{i,0} & \text{if } \overline{\mathrm{RH}}_i > \overline{\mathrm{RH}}_{i,0}. \end{cases} \tag{3}$$

$\overline{\mathrm{RH}}_{i,0}$ and σ_0 are constant values, which are represented by the green dot on Fig. 7a. α and β are two functions of $\overline{T}$ defining the quadratic part of σ. These two functions are determined by fixing $\sigma(\overline{T},\overline{\mathrm{RH}}_{i,0}) = \sigma_0$ for all $\overline{T}$, and by parameterizing the value of the peak of the curve $\sigma_{\max}$ as follows:

$$\sigma_{\max}(\overline{T}) = \kappa_v \max\left(0, \overline{T} - T_{\mathrm{thresh}}\right) + \sigma_0, \tag{4}$$

where T_{thresh} and κ_v are constant values. The values of $\overline{\mathrm{RH}}_{i,\max}$ and $\sigma_{\max}$ are represented by orange dots in Fig. 7a, and the orange dashed line. T_{thresh} is the temperature threshold below which σ shows no more peak, i.e., when the orange dot and the green dot overlay. κ_v is the rate of increase of the peak value of σ with temperature, i.e., the rate at which the orange dot goes higher. The functions α and β, as well as $\overline{\mathrm{RH}}_{i,\max}$, can then be derived as:

$$\beta(\overline{T}) = \frac{2\overline{\mathrm{RH}}_{i,0}}{1+\sqrt{1-\frac{\sigma_0}{\sigma_{\max}(\overline{T})}}} \tag{5}$$

$$\alpha(\overline{T}) = -\frac{4\sigma_{\max}(\overline{T})}{\beta(\overline{T})^2} \tag{6}$$

$$\overline{\mathrm{RH}}_{i,\max}(\overline{T}) = \frac{\beta(\overline{T})}{2} = \frac{\overline{\mathrm{RH}}_{i,0}}{1+\sqrt{1-\frac{\sigma_0}{\sigma_{\max}(\overline{T})}}}. \tag{7}$$

Again following the results of Section 3.2, we parameterize skewness so that it cancels at a $\overline{\mathrm{RH}}_i$ value that also corresponds to the maximum of σ, $\overline{\mathrm{RH}}_{i,\max}$, which is represented by orange dots on Fig. 7b. The linear dependence between skewness and $\overline{\mathrm{RH}}_i$

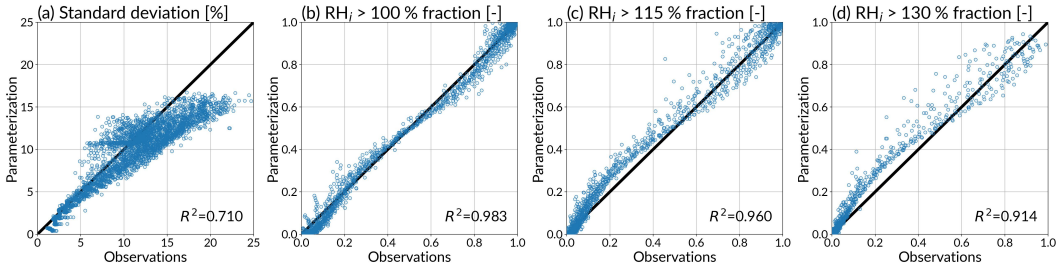

Figure 8. Scatter plots of (a) the standard deviation (in %), (b) ice supersaturated fraction, (c) fraction for which RH_i is higher than 115 % and (d) fraction for which it is higher than 130 % from the parameterized distributions (y-axes) versus the observed ones (x-axes).

is parameterized using T_{thresh}, because the lower the standard deviation, the lower the skewness (see Fig. 4). These assumptions lead to the following formulation for the skewness:

$$\gamma(\overline{T}, \overline{\mathrm{RH}_i}) = \gamma_0(\overline{T}) \left(1 - \frac{\overline{\mathrm{RH}_i}}{\overline{\mathrm{RH}_{i,\max}}(\overline{T})}\right) \quad (8)$$

$$\text{with} \quad \gamma_0(\overline{T}) = \kappa_s \max\left(0, \overline{T} - T_{\mathrm{thresh}}\right), \quad (9)$$

where κ_s is the rate of evolution of γ_0, the skewness at $\overline{\mathrm{RH}_i} = 0$ % with temperature. γ_0 is represented with purple dots on Fig. 7b, falling to 0 if $\overline{T}$ is lower than T_{thresh}.

4.3 Evaluation of the parameterization

With these empirical formulations of the average, standard deviation, and skewness, the distribution of RH_i is fully parameterized as a function of $\overline{T}$ and $\overline{\mathrm{RH}_i}$. The value of the parameters T_{thresh}, $\overline{\mathrm{RH}_{i,0}}$, σ_0, κ_v and κ_s are obtained by fitting again each observed distribution to a beta law, but the four parameters p, q, a and c are now prescribed using the parameterization, with a being fixed. This means that instead of fitting each distribution with the four parameters of the beta law, all the distributions are now simultaneously fitted with the five free parameters of the parameterization T_{thresh}, $\overline{\mathrm{RH}_{i,0}}$, σ_0, κ_v and κ_s. The resulting distributions from this new fit are called the parameterized distributions, and the fit minimizes $\sqrt{\frac{1}{N}\sum_i^N (1 - R^2_{\mathrm{par},k})^2}$, where $R^2_{\mathrm{par},k}$ is the determination coefficient between the observed and parameterized CDF of the distribution for the k^{th} bin of ($\overline{T}$, $\overline{\mathrm{RH}_i}$).

The fits of the observed distributions to the parameterized beta law show good agreement (Fig. 6; compare the solid and dashed lines). $R^2_{\mathrm{par},k}$ has a mean of 0.997, with a minimum value of 0.917. This is lower than the minimum value of $R^2_{\mathrm{fit},k}$ which is 0.97, because some distributions, especially those corresponding to the edge of the ($\overline{T}$, $\overline{\mathrm{RH}_i}$) domain, are less well represented with the proposed parameterization. The values of the five parameters can be found in Table 2 (line "Base case").

The quality of the parameterization is also assessed by comparing the value of the observed and parameterized standard deviations, as well as the value of the observed and parameterized CDFs for three arbitrary values of relative humidity: $\mathrm{RH}_i = 100$ %, which represents the ice supersaturation fraction, 115 % and 130 %. The standard deviation σ is generally slightly underestimated, especially for high values of observed σ (Fig. 8a), which may have an impact on supersaturation cloud formation schemes for AGCMs, in particular for the representation of water content. On the contrary, ice supersaturation fraction and the fractions of RH_i larger than 115 % and 130 % are much better captured (Fig. 8b, c, d). The determination coefficients R^2 are 0.98, 0.96 and 0.91, respectively.

Table 2. Number of observations, values of the five free parameters of the parameterization as well as mean and minimum value of $R^2_{\mathrm{par},k}$, for the base and different sensitivity cases.

	# obs.	T_{thresh} [K]	$\overline{\mathrm{RH}}_{i,0}$ [%]	σ_0 [%]	κ_v [%.K^{-1}]	κ_s [K^{-1}]	Mean $R^2_{\mathrm{par},k}$	Min. $R^2_{\mathrm{par},k}$
Base case	257 M	216	115	10.8	0.192	0.0177	0.997	0.917
25 km-batch	259 M	205	85	3.8	0.038	0	0.998	0.986
50 km-batch	258 M	209	107	5.4	0.070	0	0.998	0.986
75 km-batch	258 M	210	110	6.4	0.104	0.0090	0.998	0.983
100 km-batch	258 M	210	110	7.2	0.124	0.0161	0.998	0.986
150 km-batch	257 M	216	117	9.0	0.201	0.0307	0.998	0.979
300 km-batch	256 M	210	110	11.9	0.189	0.0133	0.996	0.863
NH mid-latitudes	139 M	210	110	10.0	0.142	0.0048	0.997	0.951
Tropics	61 M	222	107	12.2	0.155	0.0018	0.995	0.841
North Atlantic	53 M	210	110	8.2	0.202	0.0354	0.998	0.923
Troposphere	105 M	220	116	10.3	0.275	0.0406	0.997	0.948
Stratosphere	74 M	186	200	1.7	0.355	0.0157	0.996	0.928
No clouds	40 M	210	110	11.0	0.145	0.0169	0.995	0.840

For ice supersaturation, the fraction is estimated with a maximum error of 0.1 in most of the cases. Fractions lower than 0.5 are slightly overestimated, and fractions higher than 0.5 are slightly underestimated.

4.4 Sensitivity of the parameterization

We now assess the sensitivity of our parameterization to the length of the batches L_{batch} and to further screening of the data by region or airmass property. We change L_{batch} to simulate a change in the AGCM gridbox size, from 200 to 25, 50, 75, 100, 150 and 300 km. The regions used to screen the measurements are the tropics (30°S to 30°N), the Northern Hemisphere (NH) mid-latitudes (40°N to 60°N), the North Atlantic (40°N to 60°N and 65°E to 5°E), the troposphere and the stratosphere. We also screen the clear-sky conditions measurements, using a threshold of $N_i = 10^{-3}$ cm^{-3} for the minimum concentration of crystals defining a cloud, following Sanogo et al. (2023).

When L_{batch} varies from 25 to 300 km, the mean determination coefficient $R^2_{\mathrm{par},k}$ is always greater than 0.996 (Table 2). However, the quality of the fit strictly decreases with L_{batch}, with a minimum $R^2_{\mathrm{par},k}$ going from 0.986 for a batch length of 25 km, to 0.863 for 300 km. This indicates that the lower L_{batch}, the better our parameterized distributions simulate water vapor variability. κ_v and σ_0 globally increase with L_{batch}, meaning that the variability of water vapor is increasing with L_{batch}. There is however no clear trend for T_{thresh} and κ_s, suggesting that large-scale dynamical processes are not the source or sink of skewness.

Screening the data to specific regions does not significantly reduce the mean value of $R^2_{\mathrm{par},k}$ but it remains always very high, showing that the parameterized subgrid-scale distributions can simulate all the situations considered here. However, the parameters differ slightly when the measurements are screened for clear-sky condition than in the base case, suggesting that our assumption that all measurements are made in clear-sky may not be completely accurate.

5 Discussion

5.1 Are IAGOS observations reliable and sufficiently robust to build a parameterization?

We now discuss the potential limitations of our study, focusing on the capability of IAGOS observations to build a parameterization of the subgrid scale distribution of water vapor for an AGCM. This study has been conducted with the underlying assumption that the IAGOS dataset contains sufficient information to achieve our goal. However, some limitations of IAGOS question this assumption. Gierens et al. (2007) argued that pooled data such as those used in this study cannot be directly employed in a statistical cloud scheme, because each individual distribution (i.e., the distribution in one single batch) is mixed up in the resulting statistical distribution. They further argued that such data can only be used for validation. However, we think that if the initial objective is well set, in particular if the variables we want our parameterization to depend upon are well defined, there is no such limitations to use observational data. What is represented and what is hidden in the pooling process, and therefore in the parameterization, is not a result of the algorithm, but a decision as to which explanatory variables we want to use.

The parameterization we have built could have also been derived from the output of high resolution models e.g., Cloud Resolving Models. Such models could help to parameterize such a distribution, because they are resolved enough to study the mesoscale variability of water vapor, and low-resolved enough to consume a reasonable amount of resources per simulation. They have however their own limitations: for example, they rely on parameterizations of turbulence and cloud microphysics, which have their own deficiencies and uncertainties, particulary in extremely cold environements. Moreover, generalizing results for a wide range of temperatures and humidities, as done with the IAGOS dataset in this study, would still need a large number of simulations, large enough to become too computationnally expensive.

AGCMs that use the distribution of water vapor to simulate the formation of new clouds assume that this distribution is representative of an atmospheric state that has not yet "experienced" condensation of the water vapor. However, if we see condensation as a fast process, real-world observations have, by nature, already experienced condensation and the distribution that we derive is more representative of an atmospheric state after than before condensation takes place. This means that distribution that we have derived needs to be slightly modified before it can be used in an AGCM. We address this issue by leaving free parameters in the parameterization which can be modified. We anticipate that up to five parameters can be left for tuning within some *a priori* range when the parameterization will be implemented in an AGCM.

5.2 Technical limitations of IAGOS

The IAGOS dataset consists of measurements made by *in situ* sensors, mounted on commercial aircraft. Operating commercial aircraft aims at maintaining the highest security level while minimizing the operating costs. This implies that aircraft avoid as much as possible convective regions, adopt trajectories that decrease the flight time, and perform other maneuvers that may bias the sampling of data. Most of all, flight paths are not well distributed around the globe (see Fig. 1). Therefore, measurements are biased toward specific meteorological conditions and geographical regions, and our results cannot be representative of all the situations. The shape of the parameterized distribution is relatively insensitive to the geographical region (see Section 4.4), but we cannot exclude the potential effect from a meteorological sampling bias.

Additionally, *in situ* measurement are collected along 1-D lines in space. Nonetheless, the parameterized distributions are meant to be used in 3D gridboxes. We there-

fore implicitly assume that the 3D variability is somewhat isotropic and fully characterized with 1D transects.

The limitations presented in this Section are inherent to the IAGOS dataset, and their associated biases can hardly be estimated. They can all modify the shape and scale of the humidity distributions studied. We assume that those biases have a low impact on the overall shape of the distributions and on the general evolution of skewness and standard deviation, as the sensitivity study shows almost no dependence to the geographical and altitude region. However, having such biases means that the law used to fit the distributions, the empirical formulations found for the standard deviation and skewness, and finally the five parameters of the parameterization, are associated with uncertainties. We can account for a part of these uncertainties by providing ranges of plausible values for the parameters of the parameterization. With a modelling perspective, these ranges can be used to tune the AGCM using a framework such as proposed by Mignot et al. (2021).

6 Summary and conclusion

In this study, we have parameterized the distribution of water vapor at the mesoscale, typically 200 km, as a function of the average temperature and specific humidity. This parameterization is meant to predict supersaturation and cloud formation in an AGCM. The distributions of water vapor are built from the IAGOS observational product, which is composed of 27 years of airborne measurements of atmospheric properties, such as temperature and relative humidity, corresponding to a total of 257 millions measurements. We applied a new reconstruction algorithm to increase the quality and reliability of the data.

The observed distributions of water vapor are expressed using relative humidity w.r.t. ice, and their standard deviation and skewness are investigated as a function of the average temperature and specific humidity. Clear patterns ermerge for how standard deviation and skewness evolve, with a noticeable increase in magnitude with increasing temperature. For a fixed temperature, the standard deviation shows a quadratic behavior, between 0 % to a relative humidity higher than 100 %. Beyond this value, it increases again but with a less clear pattern. Skewness is correlated to the standard deviation: when the latter has a quadratic behavior, skewness has a linear one, decreasing from a positive value to a negative value with increasing average relative humidity. The zero value is reached at about the same average relative humidity as where standard deviation is maximum.

The distributions are then fitted to a beta law, with a very high determination coefficient. The parameters of the fitted distributions are parameterized with empirical functions of average temperature and humidity with five parameters, for potential direct application in AGCMs. The distributions are again fitted with this parameterization, and the determination coefficient is high, always greater than 0.917. In particular, the parameterization predicts the observed ice supersaturation fraction with a very good accuracy. The sensitivity of the parameterization to different geographical regions is investigated, and indicates that for a same set of parameters, the parameterization successfully captures different situations around the Earth, which is a major requirement for an implementation in an AGCM.

This parameterization is designed to be implemented in AGCMs to better represent the formation and evolution of high clouds and condensation trails. Future work will focus on testing its implementation and tuning in an AGCM.

Appendix A Reconstruction algorithm

The idea of the algorithm is to reconstruct a reference time series of relative humidity w.r.t. liquid RH_l, from the measured time series $\widetilde{\mathrm{RH}_l}$. RH_l is a high-temporal resolution time series, and $\widetilde{\mathrm{RH}_l}$ is the time series measured by the capacitive hygrometer from IAGOS. Neis et al. (2015) (hereinafter N15) showed that $\widetilde{\mathrm{RH}_l}$ can be modelled as an exponential moving average (EMA) smoothing of RH_l. The EMA is defined as a recursive linear transformation of this quantity $\mathrm{RH}_l(t)$, to the smoothed quantity $\widetilde{\mathrm{RH}_l}(t)$, with a dependence on the sensor temperature T_S:

$$\widetilde{\mathrm{RH}_l}(t) = \widetilde{\mathrm{RH}_l}(t-\Delta t) + \alpha(T_S, \Delta t) \cdot (\mathrm{RH}_l(t) - \widetilde{\mathrm{RH}_l}(t-\Delta t)) \tag{A1}$$

where Δt is the time between two measurements and α is a function of T_S and Δt.

The methodology of the reconstruction algorithm we use is based on N15, which itself relies on high-resolution colocated measurements of RH_l from the CIRRUSIII and AIRTOSS-ICE campaigns (Krämer et al., 2016, 2020). The preprocessing and grouping of the data follow the same procedure as in N15 and are not detailed here. The major difference between N15 and this work is that N15 provided an algorithm to construct $\widetilde{\mathrm{RH}_l}$ from RH_l, but in this study we provide an algorithm to reconstruct RH_l from $\widetilde{\mathrm{RH}_l}$. A reconstruction of the time series in similar conditions has already been done by Ehrlich and Wendisch (2015), but the α term of Eq. A1 is constant in their work, while here we make it depend upon temperature.

Following Ehrlich and Wendisch (2015), we first smooth the raw data to reduce the noise using a Blackman window, defined by:

$$w(t) = 0.42 - 0.5\cos(2\pi\ t/t_{B,1}(T_S)) + 0.08\cos(4\pi\ t/t_{B,1}(T_S)), \tag{A2}$$

where $t_{B,1}$ is the size of the window in seconds. As α, this value depends on the sensor temperature T_S, and will need to be evaluated. We then reverse the EMA, using Eq. A1:

$$\mathrm{RH}_l(t) = \widetilde{\mathrm{RH}_l}(t-\Delta t) + \alpha(T_S, \Delta t)^{-1} \cdot (\widetilde{\mathrm{RH}_l}(t) - \widetilde{\mathrm{RH}_l}(t-\Delta t)) \tag{A3}$$

$$\text{with} \quad \alpha(T_S, \Delta t) = 1 - \exp\left(-\frac{\Delta t}{\tau(T_S)}\right) \tag{A4}$$

Finally, to remove the additional noise created by this operation, and following once again Ehrlich and Wendisch (2015), we smooth the result with a new Blackman window following Eq. A2, with another window size $t_{B,2}$.

When we apply the methodology of N15 to this new reconstruction algorithm, we find the following formulations for the three temperature-dependent calibration functions τ, $t_{B,1}$ and $t_{B,2}$:

$$\tau(T_S) = \exp\left(-80.5 + 0.765\ T_S - 0.00171\ T_S^2\right) \tag{A5}$$

$$t_{B,1}(T_S) = \exp\left(-26.3 + 0.343\ T_S - 0.000886\ T_S^2\right) \tag{A6}$$

$$t_{B,2}(T_S) = \max\left(1, 68.5 - 0.25\ T_S\right) \tag{A7}$$

We compute the determination coefficient between the RH_l time series measured by the high-resolution instrument and those (1) measured by the IAGOS instrument, (2) measured by the IAGOS instrument to which we applied the reconstruction algorithm, and (3) measured by the IAGOS instrument to which we applied a moving average of $\Delta t = 1$ min, as done in Gierens et al. (2007). This is done for the 12 AIRTOSS flights (AIR1 to AIR12), and to 5 of the CIRRUSIII flights (CIR1 to CIR5) for which the IAGOS instrument was installed. Our reconstruction algorithm is overall increasing the quality of the IAGOS measurements (Table A1). However, an important assumption of this algorithm is that it can be applied with the same parameters to all the IAGOS flights,

Table A1. Determination coefficient R^2 computed between the time series of the high-resolution measurements of RH_l and the IAGOS measurements (raw), the IAGOS measurements to which the reconstruction algorithm is applied (reconstruction), and the IAGOS measurements to which an averaging procedure of $\Delta t = 1$ min is applied (average), for 17 flights.

Flight	R^2 raw	R^2 reconstruction	R^2 average
AIR1	0.891	0.965	0.885
AIR2	0.829	0.857	0.827
AIR3	0.938	0.960	0.935
AIR4	0.826	0.843	0.817
AIR5	0.863	0.891	0.860
AIR6	0.775	0.737	0.753
AIR7	0.983	0.981	0.937
AIR8	0.576	0.566	0.554
AIR9	0.968	0.973	0.951
AIR10	0.947	0.960	0.926
AIR11	0.964	0.985	0.959
AIR12	0.968	0.989	0.960
CIR1	0.484	0.691	0.486
CIR2	0.758	0.701	0.758
CIR3	0.597	0.732	0.602
CIR4	0.716	0.710	0.722
CIR5	0.668	0.735	0.672
Mean	**0.809**	**0.840**	**0.800**

and that the quality of the data will overall increase. This assumption has only been validated for the 17 used flights which flown in similar meteorological conditions. Additional co-located measurements using high-resolution sensors along with the IAGOS sensors in different meteorological conditions would strenghten this assumption.

Appendix B Properties of the beta law

The average μ, the standard deviation σ and the skewness γ of a beta law are expressed as a function of the location parameter a, the scale parameter $c-a$ and the two shape parameters p and q as:

$$\begin{cases} \mu = \frac{\bar{\mu}-a}{c-a} \\ \sigma = \frac{\bar{\sigma}}{c-a} \\ \gamma = \frac{2(q-p)\sqrt{p+q+1}}{(p+q+2)\sqrt{pq}} \end{cases} \quad \text{with} \quad \begin{cases} \bar{\mu} = \frac{p}{p+q} \\ \bar{\sigma} = \sqrt{\frac{pq}{(p+q)^2(p+q+1)}} \end{cases} \tag{B1}$$

Therefore, the three parameters c, p and q can be determined as a function of the average, the standard deviation, the skewness and the location parameter such that:

$$\begin{cases} p = \frac{\nu}{\xi^2(\nu+1)+1} \\ q = \nu - p \\ c = a + \frac{p+q}{p}(\mu - a) \end{cases} \quad \text{with} \quad \nu = 2\frac{\xi^2 - \gamma\xi - 1}{\gamma\xi - 2\xi^2} \quad \text{and} \quad \xi = \frac{\sigma}{\mu - a} \tag{B2}$$

Open Research Section

The processing code will be made freely available on Zenodo if the paper is accepted. In the meantime, it is accessible on a gitlab: `https://gitlab.in2p3.fr/audran.borella/iagos-water-vapor-distributions-parameterization`. The IAGOS data can be downloaded from the IAGOS data portal at https://doi.org/10.25326/20 (Boulanger et al., 2018). This study used IAGOS data on their 01/01/2024 version.

Acknowledgments

This research has been supported by the French Ministère de la Transition écologique (N° DGAC 382 N2021-39), with support from France's Plan National de Relance et de Résilience (PNRR) and the European Union's NextGenerationEU. To process the IAGOS data, this study benefited from the IPSL mesocenter ESPRI facility which is supported by CNRS, Sorbonne Université, Labex L-IPSL, CNES and École Polytechnique. MOZAIC/CARIBIC/IAGOS data were created with support from the European Commission, national agencies in Germany (BMBF), France (MESR), and the UK (NERC), and the IAGOS member institutions (`http://www.iagos.org/partners`). The participating airlines (Lufthansa, Air France, Austrian, China Airlines, Hawaiian Airlines, Air Canada, Iberia, Eurowings Discover, Cathay Pacific, Air Namibia, Sabena) supported IAGOS by carrying the measurement equipment free of charge since 1994. The data are available at `http://www.iagos.fr` thanks to additional support from AERIS. The CIRRUS-III project was mainly financed by FZ Jülich and the German ScienceFoundation (SFB 641 – The Tropospheric Ice Phase) and supported by DLR, the Max PlanckSociety, ETH Zurich, and Droplet Measurement Technology Inc. (Boulder, CO, USA). Main funding for the AIRTOSS-ICE project was provided by the German Science Foundation (DFG) through the "Spatially Inhomogeneous Cirrus: Influence on Atmospheric Radiation" project (BO1829/7-1). We thank Sidiki Sanogo for the insightful discussions about the IAGOS data, and the study in its entirety.

References

Appleman, H. (1953). The formation of exhaust condensation trails by jet aircraft. *Bulletin of the American Meteorological Society*, *34*(1), 14–20. doi: 10.1175/1520-0477-34.1.14

Baumgartner, M., Rolf, C., Grooß, J.-U., Schneider, J., Schorr, T., Möhler, O., ... Krämer, M. (2022). New investigations on homogeneous ice nucleation: the effects of water activity and water saturation formulations. *Atmospheric Chemistry and Physics*, *22*(1), 65–91. doi: 10.5194/acp-22-65-2022

Bethan, S., Vaughan, G., & Reid, S. J. (1996). A comparison of ozone and thermal tropopause heights and the impact of tropopause definition on quantifying the ozone content of the troposphere. *Quarterly Journal of the Royal Meteorological Society*, *122*(532), 929–944. doi: 10.1002/qj.49712253207

Bony, S., & Emanuel, K. A. (2001). A parameterization of the cloudiness associated with cumulus convection; Evaluation using TOGA COARE data. *Journal of the Atmospheric Sciences*, *58*(21), 3158–3183. doi: 10.1175/1520-0469(2001)058⟨3158:APOTCA⟩2.0.CO;2

Boulanger, D., Blot, R., Bundke, U., Gerbig, C., Hermann, M., Nédélec, P., ... Ziereis, H. (2018). *IAGOS final quality controlled Observational Data L2 – Time series, Aeris* [dataset]. IAGOS. (accessed 16 November 2023, `https://doi.org/10.25326/06`)

Ceppi, P., Brient, F., Zelinka, M. D., & Hartmann, D. L. (2017). Cloud feedback mechanisms and their representation in global climate models. *WIREs Climate Change*, *8*(4), e465. doi: 10.1002/wcc.465

Ehrlich, A., & Wendisch, M. (2015). Reconstruction of high-resolution time series from slow-response broadband terrestrial irradiance measurements by

deconvolution. *Atmospheric Measurement Techniques*, *8*(9), 3671–3684. doi: 10.5194/amt-8-3671-2015

Gierens, K. (2003). On the transition between heterogeneous and homogeneous freezing. *Atmospheric Chemistry and Physics*, *3*(2), 437–446. doi: 10.5194/acp-3-437-2003

Gierens, K., Kohlhepp, R., Dotzek, N., & Smit, H. G. (2007). Instantaneous fluctuations of temperature and moisture in the upper troposphere and tropopause region. Part 1: Probability densities and their variability. *Meteorologische Zeitschrift*, *16*(2), 221–231. doi: 10.1127/0941-2948/2007/0197

Gierens, K., Schumann, U., Helten, M., Smit, H., & Marenco, A. (1999). A distribution law for relative humidity in the upper troposphere and lower stratosphere derived from three years of MOZAIC measurements. *Annales Geophysicæ*, *17*(9), 1218–1226.

Gierens, K., Schumann, U., Smit, H. G. J., Helten, M., & Zängl, G. (1997). Determination of humidity and temperature fluctuations based on MOZAIC data and parametrisation of persistent contrail coverage for general circulation models. *Annales Geophysicæ*, *15*(8), 1057–1066. doi: 10.1007/s00585-997-1057-3

Gierens, K., Spichtinger, P., & Schumann, U. (2012). Ice Supersaturation. In U. Schumann (Ed.), *Atmospheric Physics: Background – Methods – Trends* (pp. 135–150). doi: 10.1007/978-3-642-30183-4_9

Hill, P. G., Holloway, C. E., Byrne, M. P., Lambert, F. H., & Webb, M. J. (2023). Climate models underestimate dynamic cloud feedbacks in the tropics. *Geophysical Research Letters*, *50*(15), e2023GL104573. doi: 10.1029/2023GL104573

Irvine, E. A., Hoskins, B. J., & Shine, K. P. (2014). A Lagrangian analysis of ice-supersaturated air over the North Atlantic. *Journal of Geophysical Research: Atmospheres*, *119*(1), 90–100. doi: 10.1002/2013JD020251

Koop, T., Luo, B., Tsias, A., & Peter, T. (2000). Water activity as the determinant for homogeneous ice nucleation in aqueous solutions. *Nature*, *406*(6796), 611–614. doi: 10.1038/35020537

Krämer, M., Rolf, C., Luebke, A., Afchine, A., Spelten, N., Costa, A., ... Avallone, L. (2016). A microphysics guide to cirrus clouds – Part 1: Cirrus types. *Atmospheric Chemistry and Physics*, *16*(5), 3463–3483. doi: 10.5194/acp-16-3463-2016

Krämer, M., Rolf, C., Spelten, N., Afchine, A., Fahey, D., Jensen, E., ... Sourdeval, O. (2020). A microphysics guide to cirrus – Part 2: Climatologies of clouds and humidity from observations. *Atmospheric Chemistry and Physics*, *20*(21), 12569–12608. doi: 10.5194/acp-20-12569-2020

Kärcher, B. (2003). A parameterization of cirrus cloud formation: Heterogeneous freezing. *Journal of Geophysical Research*, *108*(D14), 4402. doi: 10.1029/2002JD003220

Kärcher, B. (2017). Cirrus clouds and their response to anthropogenic activities. *Current Climate Change Reports*, *3*(1), 45–57. doi: 10.1007/s40641-017-0060-3

Kärcher, B., DeMott, P. J., Jensen, E. J., & Harrington, J. Y. (2022). Studies on the competition between homogeneous and heterogeneous ice nucleation in cirrus formation. *Journal of Geophysical Research: Atmospheres*, *127*(3), e2021JD035805. doi: 10.1029/2021JD035805

Kärcher, B., Dörnbrack, A., & Sölch, I. (2014). Supersaturation variability and cirrus ice crystal size distributions. *Journal of the Atmospheric Sciences*, *71*(8), 2905–2926. doi: 10.1175/JAS-D-13-0404.1

Lamquin, N., Stubenrauch, C. J., Gierens, K., Burkhardt, U., & Smit, H. (2012). A global climatology of upper-tropospheric ice supersaturation occurrence inferred from the Atmospheric Infrared Sounder calibrated by MOZAIC. *Atmospheric Chemistry and Physics*, *12*(1), 381–405. doi: 10.5194/acp-12-381-2012

Lee, D., Fahey, D., Skowron, A., Allen, M., Burkhardt, U., Chen, Q., ... Wilcox, L. (2021). The contribution of global aviation to anthropogenic climate forcing for 2000 to 2018. *Atmospheric Environment*, *244*, 117834. doi: 10.1016/j.atmosenv.2020.117834

Lohmann, U., & Kärcher, B. (2002). First interactive simulations of cirrus clouds formed by homogeneous freezing in the ECHAM general circulation model. *Journal of Geophysical Research: Atmospheres*, *107*(D10), AAC 8–1–AAC 8–13. doi: 10.1029/2001JD000767

Mignot, J., Hourdin, F., Deshayes, J., Boucher, O., Gastineau, G., Musat, I., ... Silvy, Y. (2021). The tuning strategy of IPSL-CM6A-LR. *Journal of Advances in Modeling Earth Systems*, *13*(5). doi: 10.1029/2020MS002340

Muench, S., & Lohmann, U. (2020). Developing a cloud scheme with prognostic cloud fraction and two moment microphysics for ECHAM-HAM. *Journal of Advances in Modeling Earth Systems*, *12*(8), e2019MS001824. doi: 10.1029/2019MS001824

Neis, P., Smit, H. G. J., Rohs, S., Bundke, U., Krämer, M., Spelten, N., ... Petzold, A. (2015). Quality assessment of MOZAIC and IAGOS capacitive hygrometers: insights from airborne field studies. *Tellus*, *67B*(1), 28320. doi: 10.3402/tellusb.v67.28320

Petzold, A., Krämer, M., Neis, P., Rolf, C., Rohs, S., Berkes, F., ... Wahner, A. (2017). Upper tropospheric water vapour and its interaction with cirrus clouds as seen from IAGOS long-term routine in situ observations. *Faraday Discussions*, *200*, 229–249. doi: 10.1039/C7FD00006E

Petzold, A., Neis, P., Rütimann, M., Rohs, S., Berkes, F., Smit, H. G. J., ... Wahner, A. (2020). Ice-supersaturated air masses in the northern mid-latitudes from regular in situ observations by passenger aircraft: vertical distribution, seasonality and tropospheric fingerprint. *Atmospheric Chemistry and Physics*, *20*(13), 8157–8179. doi: 10.5194/acp-20-8157-2020

Petzold, A., Thouret, V., Gerbig, C., Zahn, A., Brenninkmeijer, C. A. M., Gallagher, M., ... Volz-Thomas, A. (2015). Global-scale atmosphere monitoring by in-service aircraft – current achievements and future prospects of the European Research Infrastructure IAGOS. *Tellus*, *67B*(1), 28452. doi: 10.3402/tellusb.v67.28452

Pruppacher, H., & Klett, J. (2010). *Microphysics of Clouds and Precipitation* (Vol. 18). doi: 10.1007/978-0-306-48100-0

Ren, C., & Mackenzie, A. R. (2005). Cirrus parametrization and the role of ice nuclei. *Quarterly Journal of the Royal Meteorological Society*, *131*(608), 1585–1605. doi: 10.1256/qj.04.126

Reutter, P., Neis, P., Rohs, S., & Sauvage, B. (2020). Ice supersaturated regions: properties and validation of ERA-Interim reanalysis with IAGOS in situ water vapour measurements. *Atmospheric Chemistry and Physics*, *20*(2), 787–804. doi: 10.5194/acp-20-787-2020

Rolf, C., Rohs, S., Smit, H. G. J., Krämer, M., Bozóki, Z., Hofmann, S., ... Petzold, A. (2023). Evaluation of compact hygrometers for continuous airborne measurements. *Meteorologische Zeitschrift*. doi: 10.1127/metz/2023/1187

Sanogo, S., Boucher, O., Bellouin, N., Borella, A., Wolf, K., & Rohs, S. (2023). Variability of the properties of the distribution of the relative humidity with respect to ice: Implications for contrail formation [preprint]. *Atmospheric Chemistry and Physics*. doi: 10.5194/egusphere-2023-2601

Schmidt, E. (1941). Die Entstehung von Eisnebel aus den Auspuffgasen von Flugmotoren. In *Schriften der deutschen akademie der luftfahrtforschung.*

Smith, R. N. B. (1990). A scheme for predicting layer clouds and their water content in a general circulation model. *Quarterly Journal of the Royal Meteorological Society*, *116*(492), 435–460. doi: 10.1002/qj.49711649210

Sonntag, D. (1990). Important new values of the physical constants of 1986,

vapour pressure formulations based on the ITS-90, and psychrometer formulae. *Zeitschrift fuer Meteorologie*, *40*(5), 340–344.

Spichtinger, P., Gierens, K., & Read, W. (2003). The global distribution of ice-supersaturated regions as seen by the Microwave Limb Sounder. *Quarterly Journal of the Royal Meteorological Society*, *129*(595), 3391–3410. doi: 10.1256/qj.02.141

Spichtinger, P., & Leschner, M. (2016). Horizontal scales of ice-supersaturated regions. *Tellus*, *68B*(1), 29020. doi: 10.3402/tellusb.v68.29020

Tiedtke, M. (1993). Representation of clouds in large-scale models. *Monthly Weather Review*, *121*(11), 3040–3061. doi: 10.1175/1520-0493(1993)121⟨3040:ROCILS⟩2.0.CO;2

Tompkins, A. M. (2002). A prognostic parameterization for the subgrid-scale variability of water vapor and clouds in large-scale models and its use to diagnose cloud cover. *Journal of the Atmospheric Sciences*, *59*(12), 1917–1942. doi: 10.1175/1520-0469(2002)059⟨1917:APPFTS⟩2.0.CO;2

Tompkins, A. M., Gierens, K., & Rädel, G. (2007). Ice supersaturation in the ECMWF Integrated Forecast System. *Quarterly Journal of the Royal Meteorological Society*, *133*(622), 53–63. doi: 10.1002/qj.14

Vignon, E., Raillard, L., Genthon, C., Del Guasta, M., Heymsfield, A. J., Madeleine, J.-B., & Berne, A. (2022). Ice fog observed at cirrus temperatures at Dome C, Antarctic Plateau. *Atmospheric Chemistry and Physics*, *22*(19), 12857–12872. doi: 10.5194/acp-22-12857-2022

Wolf, K., Bellouin, N., & Boucher, O. (2023). Long-term upper-troposphere climatology of potential contrail occurrence over the Paris area derived from radiosonde observations. *Atmospheric Chemistry and Physics*, *23*(1), 287–309. doi: 10.5194/acp-23-287-2023