We thank the reviewer for her/his helpful comments, which certainly improved the manuscript. The detailed replies on the reviewer's comments are structured as follows: The individual reviewer comments are given in bold letters, followed by our reply. Changes/additions made to the text are enclosed in quotation marks.

l. 114: "The non-imaging infrared thermometer has a larger and more sensitive detector..." this is not obvious as Table 1 show a larger NETD value

We thank the reviewer for pointing at this issue. We corrected it in a meaningful way.

"The non-imaging nadir pointing infrared thermometer has a smaller FOV and a lower sensitivity than the TIR imager (compare Tab. 1). However, the sensor calibration is more stable, does not need to be cooled and, therefore, is well suited to serve as a reference for the TIR imager. The spectral window of the infrared thermometer covers the spectral channel 3 of the TIR imager (compare Tab. 2), what allows for a measurement comparison between the two instruments and cross-calibration checks."

l. 243, Eq. (1): What does the "imager optics" mean? The external lens of the optic is certainly transparent enough in the infrared to have a low emissivity and to let pass an important fraction of the radiation emitted by "the bottom" of the optic (the detector?). Are the imager and its optics (lens, detector, etc.) assumed to be isothermal? How are the temperature T_opt and the emissivity epsilon_opt determined?

The imager optics refers to the lens of the imager, which also emits radiation. This emitted radiation is partly reflected back by the window and finally contributes to the detected signal. Although, the lens is designed to have a high transmissivity in the considered wavelength range, it is not perfectly transparent. Therefore, it absorbs and reemits radiation with its own temperature. Isotherm conditions in the camera, especially between lens and detector, are not given, because the detector is cooled to lower temperatures than the camera environment. The temperature of the lens is directly measured by the imager itself. Its emission coefficient is provided by the manufacturer. The different temperatures of the lens, the instruments body, and detector are all accounted for at some point in the calibration process (please see also our changes in Sect. 3.2.1). However, to avoid any confusion, we now call it "lens" and not "imager optics". We also renamed the parameter subscripts from "opt" to "lens". We made some further revisions to describe the sources of the different parameter values.

"…, originating from the imagers lens emitted with its own temperature T_{lens}."

"… The specific absorption/emission is given by ε, the reflection by R, and the transmission by T. The subscripts denote the two coefficients of the window (win) and the imager's lens (lens). The spectral ε, R, and T of the window as provided by the manufacturer and validated by cross-calibrations with a black body are shown in Fig. 3. Overall, the Germanium window has a high average transmissivity of 93.95 % in the wavelength range from 7.7 µm to 12 µm. The spectral behavior of the reflection coefficient is rather constant over the entire range with about 5 % on average, while the absorption/emission coefficient is almost negligible for the VELOX channels 2 and 3, but affects longer wavelengths (up to 10 % for Channel 5 and 6). The emission coefficient of the lens is 0.15. Although, this value seems to be quite large, it results in a rather low contribution to the composed signal (≈ 0.75 %), because it only corresponds to the radiation emitted by the lens. For the application of Eq. 1 the window parameters were integrated for the filter response function of the selected spectral channel.…"

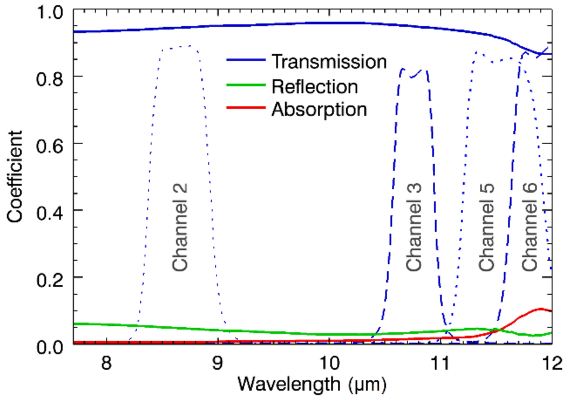


Figure 3: Spectral transmission, reflection, and absorption/emission coefficients of the Germanium window for the wavelength range covered by the TIR imager. Included are in addition the response functions (transmission coefficients, dashed/dotted lines) of the four narrow-band channels.

l. 297: how is the “no cloud-free” condition determined?

The highest brightness temperatures observed along the flight track are attributed to the cloud-free regions as we can assume that the highest temperatures are related to the warm ocean surface. Changes of the brightness temperature of the ocean surface are expected to be small in comparison to sudden temperature drops induced by clouds. Therefore, if the highest brightness temperature observed within a 60-second sequence is reduced by more than 3 % compared to the previous 60-second sequence, it is highly likely that clouds were present within this time frame. In this case, the calculated maximum brightness temperature envelope is set to the value of the previous cloud-free sequence. Using the 2D images, this method was visually validated for different cloud situations. We revised the part and added more information:

“… If a 60-second sequence is fully covered by clouds, the maximum values of the previous cloud-free sequence is used for the envelope. This is justified, because temperature changes of the ocean surface can be assumed to be spatially (temporally) weaker compared to the effect of clouds. A 60-second sequence is defined fully cloudy, if its maximum brightness temperature is reduced by more than 3 % compared to the previous sequence. ...”

l. 357: which of the two cloud fraction, "most likely cloudy" or "probably cloudy" is used for this comparison?

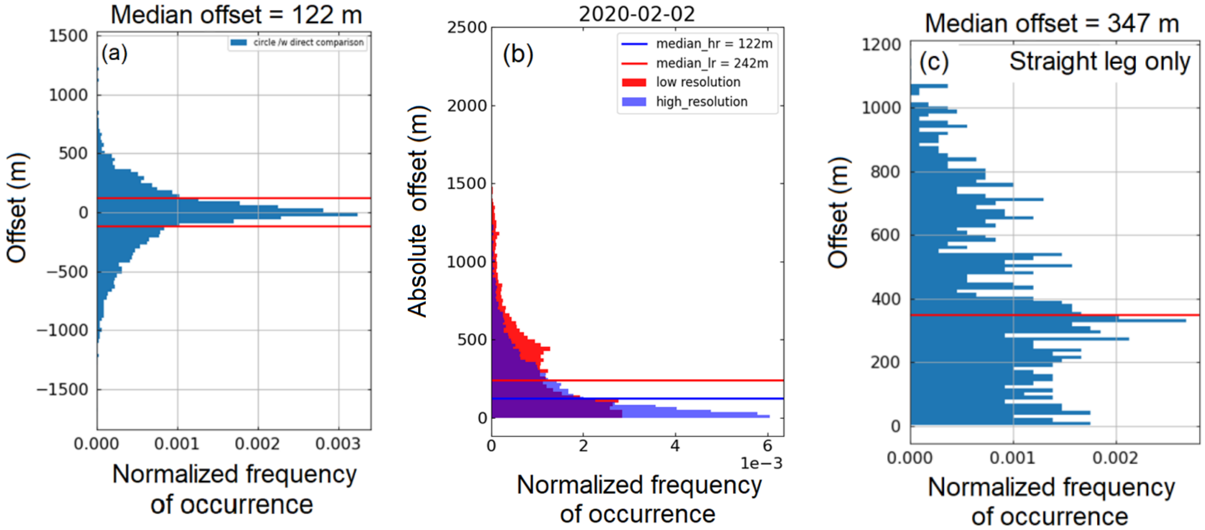
The “most likely cloudy” cloud fraction is used for this comparison. We’ve now added this information in the text.

“The comparison, based on the "most likely cloudy" threshold, highlighted the different sensitivities of the instruments to detect clouds, with VELOX always showing slightly larger cloud fractions compared to the other instruments (Konow et al., 2021).”

l. 376: I understand that the distance to the nearest dropsonde introduces errors, but why would these errors be systematic enough to generate a bias? I would rather expect a random error.

The reviewer is right. Our explanations of the dropsonde uncertainties were a bit too sparse. Together with other comments related to the cloud top altitude retrieval, we revised the entire section. We now use the VELOX cloud top altitude only as a first guess, which is then directly cross-calibrated by the data from WALES. To answer your question in more detail:

These offsets are randomly distributed (see. Fig R1a). The mean of the offset distribution is located close to 0 m for all flights. The 25 % percentile is about 120 m. For a flight leg with low dropsonde coverage (see Fig. R1c), 50 % of the VELOX cloud top altitudes had an offset smaller than 400 m (for low temporal resolution of 1 Hz) in pointwise comparisons with WALES. This offset could be further reduced by the use of a temporally higher resolved time series (full VELOX resolution, Fig. R1b).
This improvement indicates that the offsets are affected by the differences between the VELOX and WALES cloud masks, especially due to mismatches at cloud edges. A second source for larger offsets is attributed to the distance to the closest dropsonde, which introduces uncertainties in the radiative transfer simulations and the parametrization of the cloud-free atmospheric temperature offset. This is confirmed by the analysis of flight sections with a low number of dropsonde launches (2 February 2020). In these cases, the offsets increase as illustrated in Fig. R1c.



Figur R1: For EUREC^{4}A flight on 2 February 2020, **(a)** offsets beween cloud top altitude from VELOX and WALES, **(b)** Absolute offsets beween cloud top altitude from VELOX and WALES compared for 1 Hz VELOX data (red) and full VELOX temporal resolution (purple), and **(c)** offsets for flight sections with low dropsonde density. Red/purple lines indicate median values.

The revised section reads now like this:

"The cloud top temperature measured by VELOX is closely linked to the cloud top altitude. This relation is commonly used in cloud top altitude retrievals from satellite observations. Here, a similar approach is used for the images from VELOX and extended by a cross-calibration with nadir-pointing cloud top altitude measurements from WALES (Wirth et al., 2021). This method allows to extend the nadir measurements of WALES to 2D maps of cloud top altitudes, which resolve the horizontally structure of shallow cumulus. To apply the cross-calibration, a first guess of cloud top altitude from VELOX is needed. It is derived from the measured brightness temperature of the thermal imager's broadband channel 1. This first guess is necessary since there is no fixed direct relation between cloud top altitude derived from WALES and the VELOX brightness temperature along the flight path. It rather varies in time with the changing influence of the atmosphere. For the first guess, the brightness temperature is combined with atmospheric profiles from dropsondes (George et al., 2021) and radiative transfer simulations of the cloud-free atmosphere. The simulated brightness temperatures are parametrized as a function of the distance to the cloud top and used to invert the measurements.
In a second step this first guess of the VELOX cloud top altitude is cross-calibrated with the WALES cloud top altitude. The cross-calibration uses the cloud mask ("most-likely-cloudy" threshold) of VELOX (cloud mask based on the central 10 by 10 spatial pixels) and WALES. If both instruments detect a cloud, the cross-calibration is applied, which links the first guess of the VELOX cloud top altitude to the WALES cloud top altitude in a linear relationship. At this juncture, the correction of the first guess VELOX cloud top altitude ranges between 100 m and 300 m.

Two major reasons for these uncertainties were identified; (i) an increased distance to the next dropsonde leads to uncertainties in the cloud-free simulations, and (ii) missmatches in the cloud mask by VELOX and WALES. The latter can be reduced when using the full temporal resolution of VELOX.

Considering the NETD of VELOX, the full approach allows a retrieval of 2D maps of cloud top altitudes with a vertical resolution of 40 m. As an example, Fig. 11c shows the derived cloud top altitude for the cloud scene from 9 February 2020. Cloud top altitudes below 600 m might be nonphysical and are related to very thin clouds or cloud edges. These low cloud top altitudes probably results from a contamination of the signal by the emission of the ocean below."

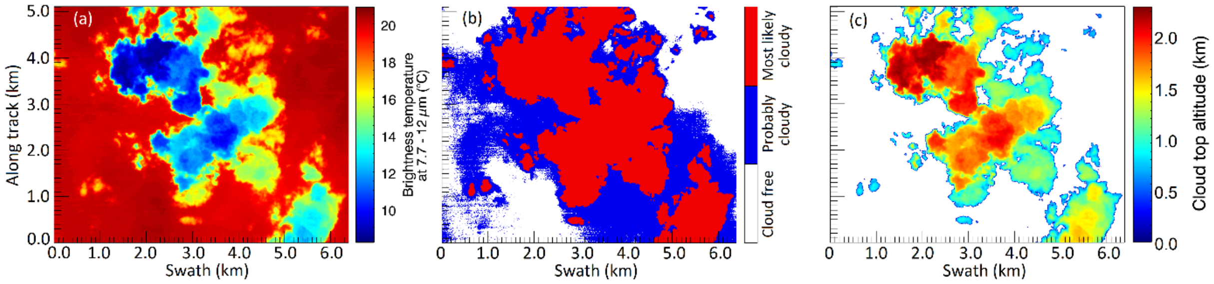


Figure 11. (a) Two-dimensional field of brightness temperature measured at a flight altitude of approximately 10 km with the VELOX broadband channel between 7.7 µm and 12 µm during the EUREC4A field campaign on 9 February 2020 at 15:05:21 UTC. For the same scene, panel (b) shows the combined cloud mask and panel (c) the retrieved cloud top altitude.

l.404: the bias in the retrieved altitude should be mentioned here

The reviewer is right. With the former method this would have been necessary information at this point. However, since we've revised the cloud top altitude section and now calibrate the VELOX data directly with WALES the offsets are not really treated as a bias anymore, they are rather a calibration factor. Therefore, we give the accuracy of 40 m here instead, which is related to the NETD of VELOX.