Reply to RC1

Thank you very much for your reviewing our manuscript and providing us with valuable comments and suggestions. We reprocessed the IMS-100-GDP because we found some minor issues in the processing data used in the previous manuscript. Furthermore, we applied the updated screening of data following review comments. The number of samples has increased from 57 to 59. However, the main message of the manuscript has unchanged.
Hereafter, Cx represents the referee's comments and Rx represents the reply to Cx.

Main comments
C1) Overall, I find a logical/expositional gap between method section (§4) and result section (§5). In particular, §4 refers to layer averages and their standard deviation in the ensamble of M=57 dual flights (see below for more details). Instead in §5, following Immler's approach which the reader expects a consistency check at the single measurement level. So, these two levels should be better connected. It is worth noting that with known uncertainties called type B evaluation. For type A evaluation, Immler et al. (2010) suggests using Students't
R1) We have taken two approaches in comparing the two data products. The first method, the layer mean in the ensemble of dual flights, is to obtain a profile of differences between the products, which is necessary to create a homogenized data set for the climatological discussion at a site where there were instrument changes. The second method, the consistency verification using the uncertainty per single measurement level following Immler's approach (Type B evaluation), is necessary to know if the results obtained in the first verification can be regarded as being significantly different.
Also Immler et al. (2010) concluded that Type A evaluation of uncertainty is not expected to play an important role within GRUAN. Thus we did not discuss it in this study.
The explanation for concepts of the two methods is added to the beginning of Sect. 4 as follows:
"Temporally simultaneous measurements were compared using the two statistical approaches adopted by Kobayashi et al. (2019) to evaluate differences in the data products. The first method, the layer mean in the ensemble of dual flights (described in Section 4.4), is to obtain a profile of differences between products, which is necessary to create a homogenized data set for the climatological discussion at a site where there were instrument changes. The second method, the consistency verification using the uncertainty per single measurement level following type B evaluation in Immler et al. (2010) (described in Section 4.5), is necessary to know if the results obtained in the first method can be regarded as being significantly different."

Furthurmore, a brief note about Type A/B evaluation is added to Section 4.5 as follows:

“Immler et al. (2010) proposed an expression for the degree of consistency as shown in Table 8. This approach is a Type B evaluation of uncertainty. For Type A evaluation, Immler et al. (2010) concluded that it is not expected to play an important role within GRUAN, so it is not considered in this study.”

C2) I would like to see some more details about the Immler’s consistency check mentioned in §4.5 and in Figure 15 (and 19).

R2) The description of the consistency check is added as:

“Immler et al. (2010) proposed terminology for comparing pairs of independent measurements of the same quantity for consistency using estimated uncertainties as described in the following: Consider two independent measurements, m1 and m2, of the same measurand with standard uncertainties, u1 and u2, respectively. Assuming the hypothesis that m1 = m2 is true and uncertainty is normally distributed, the probability that occurs only by chance, is roughly 4.5% for k = 2 and 0.27% for k =3. If Eq. 32 is true for k = 2, it is very likely that the two measurements did in fact not measure the same thing, probably due to an unrecognized or unaccounted-for systematic effect in one or both measurements. Immler et al. (2010) proposed an expression for the degree of consistency as shown in Table 8. “

“For statistical consistency check, the total consistency ranks shown in Table 8 (1: consistent, 2: in agreement, 3: significantly different, or 4: inconsistent) between RS92 and iMS-100 within a specific pressure layer for a particular parameter are estimated as the 95 % percentile value of consistency ranking numbers within the layer.”

C3) What about missing values? (only large gaps mentioned in the introduction). Recently GRUAN community is considering the importance of interpolation and its uncertainty (see e.g. Fassò et al. (2020, *https://doi.org/10.5194/amt-13-6445-2020*) for T, and Colombo et al. (2022, https://iopscience.iop.org/article/10.1088/1361-6501/ac5bff/pdf) for RH ). These considerations should be present in the state of the art literature of the present manuscript.

R3) Thank you for the references (we were aware of the work by Fassò et al. (2020), These are very good works; we are afraid, however, that the actual implementation to the GDPs will take more some time. In our current paper, we would like to limit the statement to a reference for future product development. The citation is added to the summary section as:

“The interpolation and the estimation of uncertainty for data missing periods are discussed in some articles. For example, Fassò et al. (2020) proposed a method for temperature data using the Gaussian process, and Colombo and Fassò (2022) attempted to apply it to RH data. These studies will be considered for future improved versions of the IMS-100-GDP. “

C4) The title focus on GDP comparison, an essential part of the GDP is the measurement uncertainty assessment. How do the uncertainties of the two products compare? I expected to see something about this

R4) The uncertainty components and their estimation methods (some components depend on correction methods) vary by product and are not unified, so they are not simply comparable. Therefore, we have not discussed the comparison of uncertainties themselves.

Typos and minor comments

C5) L.18: heare is here

R5) Corrected.

C6) L.37-38: a verb is missing?

R6) Corrected.

C7) L.40: RS-11G GDP is RS-11G GDP.1?

R7) Corrected.

C8) L.38-43: are these results from Kobayashi et al.? be more explicit

R8) The reference is shown again in the revised manuscript.

C9) L.49: Sect.5 and Sect.6 add a comma

R9) Corrected.

C10) L.50: “See Appendix A for a summary ...” the style is inconsistent with the rest of the paragraph, use passive form.

R10) Corrected.

C11) L.55: ant is and

R11) Corrected.

C12) L.245: “M in” is “M is”

R12) Corrected.

C13) L.245 Is M=57?, please be more explicit

R13) Yes, but M is 59 due to reconsideration of the screening process. Revised as “M is the number of data sets, here 59.”

C14) L.252-253: To define a standard deviation, in Eq. (18) summands must be squared. After this, eq. (18) defines the standard deviation of the (mean) differences given by eq.(16). The ensamble mean difference is given by (17) and its standard deviation (taking the ensamble as a random sample) is not given by eq. 18, as stated in L.252 ("The standard deviation of the ensemble mean difference for individual pressure layers") but is given by sigma(A_k)/sqrt(M). Although the phrasing in L.252 is present in literature, it is imprecise and should read "the ensamble standard deviation of the mean differences"
R14) Eq. 18 is corrected, and the phrase "the ensemble standard deviation of the mean difference" is used.

C15) L.255: "... error the ..." is "... error of the ..."
R15) Corrected.

C16) §4.5: this sect. is a 3-line section. For self-contentedness, I suggest to briefly report the Immler check, or to avoid heading this sentence as section.
R16) The description of the consistency check is detailed in the revised manuscript.