Response to Referee Comment (RC1) on

Quality control and error assessment of the Aeolus L2B wind results
from the Joint Aeolus Tropical Atlantic Campaign
(https://doi.org/10.5194/amt-2022-223)

We appreciate the referee's very insightful and helpful remarks on our manuscript. The responses to the individual comments and the corresponding changes that will be made to the manuscript are presented in the following.

General comment:

The paper has a bit of a gift wrap structure. The methods and goals are only rather vaguely touched in the beginning of the introduction, and become clear more and more only as one reads on. I recommend to give a more specific outline for what the reader is to expect.

Somewhat related, I find the naming / separation of sections 3 and 4 a bit confusing. Both sections are structured pretty much identically, Sec. 3 showing results for the model evaluation against ECMWF, Sec. 4 showing results for the AVATAR-T evaluation. The methods are indeed introduced in Sec. 3 as the name suggests, but it is somewhat of a results / methods hybrid.

I can see the benefit of introducing the methods on the examples with ECMWF, but I think a dedicated methods-only section that gives a summary (perhaps even a schematic figure?) of all methods that are going to be used could help the reader to have a clear picture on what to expect. As mentioned earlier, right now the exact methodology reveals itself only over long stretches of the manuscript.

Response to General Comment:

Thank you for this comment. We agree that the goal and structure of the paper should be outlined more clearly. Therefore, we have extended the introduction to better describe the objective of the paper and the methodical approach used to achieve this (new text is underlined):

> *[…] Therefore, a more detailed treatment of different QC schemes and how they affect the resulting statistics is necessary for comparable validation results and for a more objective assessment of the Aeolus wind data quality. Moreover, it is an important aspect with regards to the operational data assimilation in NWP centres and allows for a more rigorous error characterization of the Aeolus winds.*
>
> *This paper aims to raise the awareness to the influence of the chosen QC schemes on the validation results, particularly when using the L2B EE. It also demonstrates the usefulness of specific statistical tools for the purpose of outlier removal and the assessment of normality, which are necessary to retrieve*

> *the Aeolus wind errors in accordance with the MRD. The presented methods are applied in the context of the AVATAR-T validation campaign in 2021 by comparisons against the ECMWF model background winds and the 2-µm DWL wind data. The specifics of the campaign and available datasets are outlined in Sect. 2 together with a description of the L2B Rayleigh-clear and Mie-cloudy EE and their temporal evolution over the past three years. The model comparison in Sect. 3 serves as an example to introduce the reader to the detailed treatment of the Aeolus wind data in terms of QC and error assessment. In particular, the modified Z-score (Sect. 3.2) and normal quantile plots (Sect. 3.4) are discussed as powerful tools for removing gross errors and assessing the normality of the wind error distribution, respectively. In addition, the impact of the QC settings on the results from the model comparison is elaborated (Sect. 3.5). In Sect. 4, the statistical methods are then applied to the comparison of Aeolus wind observations against 2-µm DWL data. The paper concludes with a summary and outlook to future studies of the L2B wind error characteristics in Sect. 5.*

Regarding the structure and content of sections 3 and 4, we think that the statistical methods can be best illustrated by a concrete example rather than by an abstract explanation in a dedicated methods section. Most of the presented statistical parameters and tools refer to the Aeolus wind error, which is defined as the difference between the wind speed measured by Aeolus and that determined by the model or instrument employed for the validation. Consequently, introducing, for instance, the normal quantile plot in section 3.4 or the combined bar and line graph in section 3.5, necessitates a wind reference to demonstrate the relevance of these plots for assessing the wind error characteristics and the influence of the used QC scheme. For the purpose of keeping the text as concise as possible, despite the complexity of the topic, we decided to use the model comparison as an example to guide the reader through the steps of a careful statistical analysis of the Aeolus wind error. In section 4, these guidelines are then transferred to the validation against the 2-µm DWL. We think that with the extended introduction, as described above, this approach will become clear to the reader and facilitate the understanding of the statistical methods.

Also, we have renamed section 2 from “Datasets and methods” to “Datasets and the L2B estimated error”, as this header better describes the content. In this manner, it becomes clear that the methods are introduced in section 3 on the example of the model comparison.

Specific comment #1:

L25: I find the term "biased gross errors" a bit unusual and unclear.

Response to Specific comment #1:

The term was changed to “positively biased wind results”.

Specific comment #2:

L68: Is a Gaussian distribution actually appropriate for wind retrievals? Since wind speed is a bound variable, shouldn't the error distribution get more and more skewed as the retrievals get closer to the limit? I appreciate this is a common assumption for many variables that has also some practical reasons, but since such a strong focus is put on forcing the data into a normal distribution, some words on that might be helpful.

Response to Specific comment #2:

Our study follows the error definitions that are formulated in the Mission Requirements Document (MRD) (ESA, 2016), which assume Gaussian distributions for the Rayleigh and Mie wind error with respect to other wind observations or the model background. This is justified by the fact that the wind error is dominated by Poisson-distributed photon noise on the detectors, particularly for the Rayleigh channel. For the Mie channel, deficiencies in the signal analysis (Mie Core algorithm) give rise to gross errors which additionally contribute to the error distribution. The assumption in the MRD that the Mie gross errors are uniformly distributed did not prove correct, as pointed out in Sect. 3.3. Nevertheless, after sorting out the gross errors, the assumption of a Gaussian distribution is also valid for the Mie channel. When the wind retrieval gets closer to the limit, e.g., at low signal levels, the random error will increase with the noise accordingly, but without skewing the wind error distribution.

We have emphasized these considerations in the revised manuscript in the context of the error definitions (Sect. 3.1):

> *The definitions given here are in line with the those stated in the Aeolus MRD, which assumes that the wind error with respect to other wind observations or the model background can be described by a Gaussian distribution whose centre and width represent the accuracy and precision of the Aeolus winds. This is justified by the fact that the wind error is dominated by Poisson-distributed photon noise on the detectors, particularly for the Rayleigh channel. For the Mie channel, deficiencies in the signal analysis (Mie Core algorithm) give rise to gross errors which additionally contribute to the error distribution, as will be pointed out in Sect. 3.3. Nevertheless, after sorting out the gross errors, the assumption of a Gaussian distribution is also valid for the Mie channel.*

Specific comment #3:

L91-93: This statement is distracting and unnecessary here. Also, I have the impression that "Rayleigh-clear" and "Mie-cloudy" are used more often than not later on, so perhaps just delete this statement altogether (or stick with the simple notation consistently, which I'd actually prefer).

Response to Specific comment #3:

We removed the statement from the text.

Specific comment #4:

L103: Where does this strong signal decrease over time actually come from?

Response to Specific comment #4:

The root cause analysis of the decline in the atmospheric return signal is still ongoing. Laser-induced contamination, laser-induced damage and bulk darkening of the instrument's optics are the most probable causes, in addition to clipping losses at the instrument field stop. We have added a short comment on the potential root causes for the signal loss to Sect. 2:

> *The root cause analysis of the decline in the atmospheric return signal was still ongoing as of the writing of this paper. Laser-induced contamination, laser-induced damage, and bulk darkening of the instrument's optics are the most probable causes, in addition to clipping losses at the instrument field stop.*

Specific comment #5:

L121/L321: "Whereas" is a rather unusual conjunction to start a sentence with and sounds a bit awkward to me. Perhaps better use "while"?

Response to Specific comment #5:

The text was changed accordingly.

Specific comment #6:

L297: "the distribution is far from normal" sounds a bit funny to me. Perhaps better "far from Gaussian"?

Response to Specific comment #6:

The text was changed accordingly.

Specific comment #7:

L298: The value of 3.5 seems purely empirical. What was the decision criterion of Iglewicz and Hoaglin (1993), and is it likely to make it a good choice for your study as well?

Response to Specific comment #7:

The motivation of *Iglewicz and Hoaglin (1993)* for using a threshold value of 3.5 to detect outliers was based on a simulation study where the portion of outliers (as identified by the modified Z-score) in a random normal distribution was determined in dependence on the sample size and threshold value. The results were based on 10,000 replications for each sample size and showed that the portion of observations that are labelled as outliers does not vary much with the sample size if a threshold value of 3.5 is chosen. The so derived outlier-labelling rule as they call it, however, still contains a certain degree of arbitrariness. Nevertheless, we found in our studies that threshold values ranging from 3.0 to 3.5 ensure wind error distributions with a high degree of normality, i.e. small residuals in the normal quantile plots, as stated in l. 404ff.:

> *For the model comparison of Rayleigh-clear winds, a Z-score limit ranging from 3.0 to 3.5 was found to yield a high degree of normality, which is in accordance with the recommendation by Iglewicz and Hoaglin (1993).*

We then decided to use a threshold value of 3.5 to allow more data points to pass the QC, thereby providing more robust statistics. Also, we have checked the variability of the statistical parameters within the threshold range from 3.0 to 3.5. The outcome is discussed in l. 477ff.:

> *For the model comparison discussed above, the statistical parameters change by less than 7% if the Z-score limit is reduced from 3.5 to 3.0. The largest influence is found for the Rayleigh standard deviation which decreases to 7.3 m·s^{-1} (compared to 7.8 m·s^{-1} for a Z-score limit of 3.5), as the portion of outliers accounts for 4.5% (compared to 3.2%).*

Specific comment #8:

L335: remove the comma after "Gaussian distribution"

Response to Specific comment #8:

Done.

Specific comment #9:

L345: ECMWF are considered as absolute "truth" in the presented analyses. I'm wondering how likely it is that some of the supposed "gross errors" are actual rare extreme wind occurrences that were not modelled properly but captured correctly by the observations?

Response to Specific comment #9:

It is true that, in addition to representativity errors, model deficiencies in terms of parametrization and resolution can cause large discrepancies between the model background winds and the Aeolus wind observations. As stated in l. 601ff., such model errors are mainly located in highly-convective regions, e.g. in the tropics, and can exceed 10 m·s^{-1}:

> *Moreover, discrepancies between the 2-µm DWL and model background wind data can result from model deficiencies that are caused by imperfect parametrization or too low resolution. Errors of the model background, i.e. before the assimilation of Aeolus winds, are found to be especially large in convective areas in the tropics, exceeding even 10 m·s^{-1} on several occasions (Rennie et al., 2021).*

However, it is very unlikely that the modified Z-score filter will sort out these wind observations, as the threshold value is sufficiently relaxed. Taking the scaled MAD values from Table 2 before applying the Z-score filter (k = 5.1 m·s^{-1} for Mie, k = 7.2 m·s^{-1} for Rayleigh) and a Z-score threshold of 3.5, only wind results that deviate by more than ≈18 m·s^{-1} (Mie) or ≈25 m·s^{-1} (Rayleigh) from the model are considered as outliers (see also Fig. 2c,d). Hence, the risk of misinterpreting extreme wind occurrences as gross errors is very low.

The following paragraph was added to section 3.2 to clarify this point:

> *Note that the Z-score filter is sufficiently relaxed to avoid rejection of wind observations that might be due to extreme wind occurrences. Taking the scaled MAD values for the Mie and Rayleigh winds before applying the filter (5.1 and 7.2 m·s^{-1}, respectively) and considering that the median of the wind speed differences is close to zero, only wind results that deviate by more than ≈18 m·s^{-1} (Mie) or ≈25 m·s^{-1} (Rayleigh) from the model are considered as outliers. This is well above typical model errors that are caused by deficiencies in terms of parametrization and resolution, and can in extreme cases, like in highly-convection regions, exceed 10 m·s^{-1} (Rennie et al., 2021).*

<u>Specific comment #10:</u>

*L352: Perhaps change to "it was *assumed* that"?*

<u>Response to Specific comment #10:</u>

The text was changed accordingly.

<u>Specific comment #11:</u>

L386: Is 10 ms^{-1} a commonly used threshold? Perhaps add a rationale / reference?

<u>Response to Specific comment #11:</u>

As summarized in the introduction (l. 70ff.), the Rayleigh EE threshold typically ranges between 7 and 12.5 m·s^{-1}, with 8 m·s^{-1} being the most commonly used value in many validation studies. We chose 10 m·s^{-1} as an appropriate threshold given the fact that the EE of the Rayleigh winds had already significantly increased towards the AVATAR-T campaign (median of 4.5 to 5 m·s^{-1}, see Fig. 1) compared to the start of the FM-B period due to the signal loss and increased noise levels accordingly. The threshold also leads to portion of identified outliers (5.5%) that is comparable to that when the QC is solely based on the modified Z-score (3.2%), thereby providing better comparability between the two cases shown in Fig. 5.

The choice of the threshold value was specified in the revised manuscript as follows:

> *Here, a threshold of 10 m·s^{-1} was chosen, which is slightly larger than the most commonly used value of 8 m·s^{-1}, but regarded as an appropriate threshold given the increased noise levels during the campaign compared to the start of the FM-B and elevated EE of the Rayleigh winds accordingly (see Fig. 1).*

Specific comment #12:

L429/Fig 6.: The graph contains a whole lot of information which is hard to grasp as a whole. Perhaps add a more high-level description of what the purpose of the plot is before describing the axes and lines, etc. specifically?

Response to Specific comment #12:

Agreed. We have added the following sentence to section 3.5:

> *The diagram intends to give an overview of the most relevant statistical parameters (μ, σ, k, fraction of valid winds that is considered in the statistics) in dependence on the chosen EE threshold.*

Specific comment #13:

L499: Perhaps change to "1:1 line"?

Response to Specific comment #13:

The term was changed accordingly.

Specific comment #14:

L573: There are no orange bars. I assume this should be "black bars"?

Response to Specific comment #14:

The text was changed accordingly.

Specific comment #15:

L599: Replace "On the contrary" with "In contrast"?

Response to Specific comment #15:

Done.

Specific comment #16:

L639: The presented graph is a great way to summarize and visualize a lot of information, but I, personally, find it a bit exaggerated to call it "developing a new graph". I think it just draws the attention a bit away from the actually interesting part, which is the proposed systematic approach for selecting an EE threshold in a somewhat more objective manner.

Response to Specific comment #16:

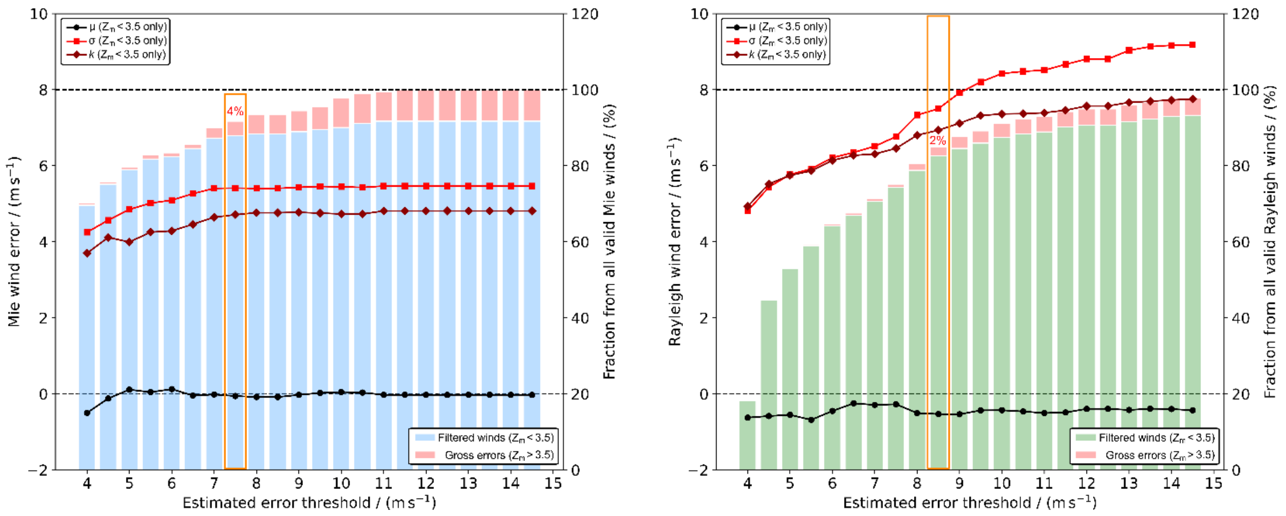
The text was changed to “a combined bar and line graph can be used to illustrate […]”.

Specific comment #17:

L654-657: I'm wondering how the statistics against ECMWF predictions would look like when using only the exact same locations as those in the AVATAR-T campaign? That is, how much of the mentioned deviations are just related to spatial representativeness?

Response to Specific comment #17:

Very good point indeed. For answering this question, we produced another combined line and bar graph depicting the statistics for the model comparison after restriction to those Aeolus wind results that were also validated by the 2-µm DWL (see below).



Same as Fig. 6, but after restriction to those Aeolus wind results that were also validated by the 2-µm DWL in Fig. 8.

In contrast to Fig. 6, this plot is thus based on only the first five underflights when the DWL was still operable, i.e. with a number of data points that is comparable to the DWL comparison. Also, for the sake of clarity, it does not include the statistical parameters (μ, σ, *k*) without applying the modified Z-score which are shown as dashed lines in Figs. 6 and 8. When comparing the new figure to Figs. 6 and 8, it becomes obvious that the statistical results are quite comparable to those of the model comparison of the entire dataset for the Mie winds. The mean bias is close to zero, while the standard deviation and scaled MAD converge toward 5.5 m·s^{-1} and 4.8 m·s^{-1}, respectively, as the EE threshold is relaxed beyond 10 m·s^{-1}. These values are slightly higher than those obtained for the entire model dataset (σ = 5.3 m·s^{-1}, k = 4.3 m·s^{-1}) which might be due to the different spatial representativeness of the model and the 2-µm DWL or simply because of the restriction of the dataset. As for the Rayleigh-clear winds, the discrepancies are larger with a more negative mean bias of -0.4 m·s^{-1} and significantly larger standard deviation and scaled MAD (σ = 9.1 m·s^{-1} compared to σ = 7.8 m·s^{-1} and k = 7.7 m·s^{-1} compared to k = 6.9 m·s^{-1}) if no QC based on the EE is applied. When using the two-step QC approach with an EE threshold of 8.5 m·s^{-1}, the values are reduced to σ = 7.5 m·s^{-1} and k = 6.9 m·s^{-1}, which is closer to the values obtained from the comparison against the 2-µm DWL (Fig. 8) (σ = 8.2 m·s^{-1}, k = 7.2 m·s^{-1}), but still lower. This result suggests that the spatial representativeness indeed contributes to the deviations in the validation results when comparing Figs. 6 and 8 in addition to the restricted overlap of the Aeolus data with the DWL winds. These aspects are covered in the discussion of the different overlap regions of the 2-µm DWL data coverage with the Mie and Rayleigh winds. Other factors that are mentioned in the text, e.g. the model errors, additionally contribute to the observed discrepancies.

Specific comment #18:

L670-- I find the concluding remarks straying a bit off-topic.

Response to Specific comment #18:

We have split the last section into two sections named “Discussion and summary” (sect. 5) and “Conclusion and outlook” (sect. 6) to better distinguish between the resume of the article and some remarks about the relevance of the results with a view to the Aeolus data assimilation and forthcoming studies on the Aeolus error characteristics. We think that the final remarks are important to highlight the impact of the paper, as it was also requested by the other two referees.