Response to Reviewers

"The MAESTRO turbulence dataset derived from the SAFIRE ATR42 aircraft"
by Jaffeux et al.

April 3, 2026

On behalf of all co-authors, we thank both reviewers for their thorough and constructive reading of the manuscript. Their comments have significantly improved the quality and clarity of the paper and dataset. Below we address each remark in turn. Reviewer comments are reproduced in italics; our responses follow in normal font. Changes to the manuscript are quoted in the responses.

Minor changes unrelated to the reviews:

The names of horizontal wind components in the three frames of reference used has been changed to be in line with common practises in atmospheric turbulence:

1. U_l and V_t for the wind-aligned frame of reference.
2. U_x and V_y for the trajectory-aligned (aircraft) frame of reference.
3. U_{geo} and V_{geo} for the earth-aligned (geographical) frame of reference.

This change has been consistently implemented in the text, figures, and dataset files.

As an introduction to the section presenting segment statistics, the strict use of wind-oriented components is explicitly stated and justified:

> "In the dataset, only horizontal components in the wind-aligned frame of reference (U_l and V_t) are used for the statistical description of individual segments. However, in the files, the suffixes are not repeated to maintain clarity. By projecting the wind onto the direction of the segment-averaged wind, the variance and covariance estimates represent fluctuations relative to the local flow, rather than with respect to fixed geographic or aircraft axes. This approach ensures that the statistics are physically interpretable and consistent across segments with differing sampling or wind directions."

Responses to Reviewer 2

Major Issue 1: Calibrations of the temperature and humidity instruments

1a)

It is a common practice to calibrate a potentially drifting fast-response instrument against a stable slow-response reference instrument. However, in my opinion the selection of the reference based on the highest correlation with the fast-response one is questionable. In principle, this choice should not involve the results from the instrument being calibrated but rather be directed by a priori known characteristics, reliability or laboratory tests of the possible reference sensors. The three reference hygrometers exhibit substantial systematic differences in the absolute values of water vapor mixing ratio. These offsets are carried on to the calibrated high-frequency record, depending on the choice of the reference sensor. Your correlation criterion does not take into account the absolute accuracy of the reference sensors. Yet, the accurate temperature and humidity are crucial if a user of the data wants to analyze relative humidity, saturation deficit or buoyancy effects.

We thank the reviewer for this important remark. We agree that the reference sensor should be selected on the basis of independently validated properties rather than by correlation with the fast instrument being calibrated.

In the revised manuscript, the justification for selecting the WVSS-2 as primary reference is now grounded in three independent lines of evidence, described in Section 3.1.2 and in the new Appendix B:

1. The 1011C chilled-mirror provides a direct condensation measurement and is in principle the most accurate instrument on board. A dedicated intercomparison between the WVSS-2 and the 1011C (Appendix B, Figs. B1–B3) shows that the WVSS-2 agrees with the 1011C within ±5% for 90.4% of all segments (509 out of 563), across the full range of conditions sampled during MAESTRO. This comparison was performed entirely independently of the fast sensors being calibrated. Only after this validation was the WVSS-2 designated as the primary reference.

2. The chilled-mirror is limited to a warming rate of $1\,\mathrm{K\,s^{-1}}$, which causes asymmetric ramping artefacts in heterogeneous conditions. This limitation specifically affects the cloud-base segments that are scientifically important in MAESTRO (Fig. B2). Applying post-processing inversion techniques to correct for thermal inertia would likely introduce additional uncertainty, particularly when combined with the downstream calibration of fast sensors. Using the 1011C as reference in heterogeneous conditions would therefore propagate a systematic slow-response artefact into the calibrated FAST-WAVE record.

3. The capacitive HUMAERO overestimates mixing ratio by more than 5% in 56.3% of all segments. This systematic bias is inconsistent across segment types and cannot be attributed to limitations of the reference instrument. It is therefore excluded a priori.

The $R^2 > 0.9$ criterion is therefore not used to choose between reference sensors: the WVSS-2 is designated as the reference on independent grounds. The R^2 serves solely as a per-segment quality filter, flagging cases where the low-frequency variability is insufficient for a reliable regression.

Regarding absolute accuracy: we acknowledge the reviewer's concern and have added a dedicated paragraph in Section 3.1.2. The dataset is optimized for turbulence analysis, where fluctuation amplitudes and spectral integrity are most important and mean offsets are of secondary importance. For users requiring accurate absolute values of water vapor mixing ratio (e.g. for relative humidity or saturation deficit), we still recommend using the calibrated FAST-WAVE / WVSS-2 values provided in the `SEGMENT_TIMESERIES` files. The absolute accuracy of the WVSS-2 relative to the 1011C is characterized in Appendix B.

1b)

Because the calibration is applied separately for each horizontal segment, the calibration of the instrument sensitivity might not work correctly in horizontally homogeneous thermodynamic conditions, i.e. where there are no significant variations in humidity/temperature along the segment. This can occur in weakly turbulent/laminar layers at higher altitudes and even in strongly turbulent flow without sources, sinks or mean gradients of these scalars. Then, the calibrated sensitivity is burdened with massive uncertainty and it might be better to recall the calibration from other segments. Please consider whether the problem is relevant for your measurements.

This is a valid and important concern. In particular, two reasons motivating per-segment rather than per-flight calibration are now stated in the manuscript:

> "The calibration is performed independently for each segment rather than once per flight or per campaign, for two reasons. First, fast sensors on airborne platforms can experience gradual or abrupt changes in their response characteristics during a flight, due to exposure to clouds, precipitation, salt, and wide ranges of temperature and humidity; a segment-scale calibration accounts for these slow drifts and environment-induced changes. And second, the MAESTRO flight legs sample distinct meteorological regimes at very different altitudes: temperature ranges from approximately +28 °C near the surface to −10 °C at 6000 m, and water vapor mixing ratio spans nearly an order of magnitude between the boundary layer and the mid-troposphere. Capacitive and spectroscopic sensors are known to exhibit environment-dependent sensitivities: the sensitivity of capacitive sensors to relative humidity is a nonlinear function of temperature, while laser absorption spectrometers can be affected by pressure and temperature variations. A segment-scale calibration ensures that the retrieved slope and offset are locally appropriate for the thermodynamic regime actually sampled, rather than imposing a single correction averaged over heterogeneous conditions."

We have added an explicit acknowledgement of the specific limitation being raised here in Section 3.1.2:

> "We acknowledge that this approach has a limitation in segments where the thermodynamic field is nearly horizontally homogeneous, for instance in weakly turbulent or laminar layers. In corresponding segments, the low-frequency variability available for the regression is small, the fitted slope is uncertain, and the calibration is less constrained. For these cases, the R^2 criterion ($R^2 > 0.9$) indicates where the low-frequency variability is potentially insufficient to produce a reliable regression. This is consistent with the general observation that weakly turbulent H-type segments, where horizontal homogeneity is expected, also show the lowest R^2 values in the calibration results."

Major Issue 2: Choice of the best sensors

Scientific approach is to test theories against experiments, not the other way round. I believe the agreement of the measurements with the Kolmogorov theory must not be used to assess their quality. Ideally, your measurements should be used to test whether the theory holds. In fact, there is much evidence that either the Kolmogorov assumptions (e.g. homogeneity, isotropy) are sometimes not met in atmospheric turbulence or the predictions of the theory (i.e. universal scaling) do not agree with observations.

We agree with the reviewer's point. In the revised manuscript, the spectral analysis has been removed. The spectral slopes are still extracted and included in the dataset. Controlling high frequency behavior of fast sensors is still required to be able to reliably document fine scale

dynamics. However, they are now presented as physically informative parameters. The primary basis for sensor selection is the independent intercomparison of reference sensors (Appendix B) and the calibration R^2 reflecting low-frequency amplitude agreement.

The text in Section 3.1.3 now explicitly states that deviations of the spectral slope from $-5/3$ carry physical information and may be indicative of buoyancy- or shear-dominated regimes, rather than indicating sensor malfunction. For the specific purpose of TKE dissipation rate computation, a slope close to $-5/3$ for the spectrum of the vertical wind is required for the underlying assumptions to hold.

The figures related to spectral comparison for individual sensors have been removed from the sensor-selection discussion. In the instrumentation Section 2.3, the high frequency flaws of the LI-7500 instrument are explicitly stated:

> "The LI-7500 exhibited a persistent narrowband spectral artifact peak (from 6 to 12.5 Hz) attributed to potential vibrations of the antenna, which strongly contaminates the high-frequency part of the humidity spectrum; therefore, this instrument is not retained in the final turbulence dataset."

For the sake of transparency, we provide an example spectrum here, that clearly shows this behavior.

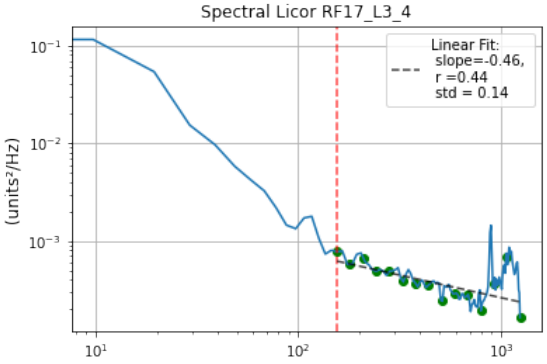

Figure 1: Typical power spectrum of the LI-7500 signal during the MAESTRO campaign

Major Issue 3: Segmentation algorithm

3a)

The algorithm to optimize segment placement should be thoroughly explained in the manuscript. In my opinion, the diagram in Fig. A1 is alone insufficient without proper description. In addition, even the diagram lacks the definition of some variables (e.g. "duramin") and implies some problems in the method. For instance, heterogeneity score seems to be computed over a data series of length "durati" whereas the segment stored afterwards has the length "seg_length".

The flowchart has been replaced by a dedicated subsection, describing of the segmentation algorithm in Section 3.1.1. All variables are now defined explicitly. The distinction between the search window (over which $\mathcal{H}$ is minimized) and the stored segment (always 270 s) is now unambiguous.

3b)

Please specify which temperature and humidity instruments are involved in the computation of heterogeneity score and hence in the segmentation procedure. It is also unclear, whether the

segmentation involved already calibrated wind, temperature and humidity because, on the other hand, the calibrations themselves were performed on the segments.

The processing chain has been updated to avoid any circular dependency. It follows the following structure:

1. **Wind correction** (campaign-wide, independent of segmentation): constant angle biases are determined from all stabilized legs and applied to the full dataset. This step is described in Appendix A and mentioned in the beginning of Section 3.
2. **Segmentation** uses only the corrected wind components in the geographical frame (U_{geo}, V_{geo}, W). Temperature and humidity are not involved in the heterogeneity score, avoiding any circularity.
3. **Humidity calibration** is performed independently for each segment, after segmentation.

For transparency, we provide the heterogeneity score distributiosns in Figure 2 calculated using only the wind components; to allow direct comparison with the previous five-variable formulation, the scores are scaled appropriately (by a 5/3 factor). Temperature and humidity introduced additional sources of heterogeneity (e.g., cloud microphysical variability), restricting the score to wind components makes the measure more closely reflect the homogeneity of the dynamic field alone. As a result, segments near cloud base now typically appear more homogeneous according to the wind-based heterogeneity metric. This shift is visible where the distribution of heterogeneity scores is systematically lower than in the original for B-type segments, five-variable version. Please note that $\mathcal{H}$ was redesigned according to your suggestions (see next comment), the one that is plotted here uses the old formulation to demonstrate clearly the impact of using 3 variables instead of 5. The core principle of the new formulation is the same, only the scaling has changed to allow it to be used and compared with different sample lengths and number of variables.

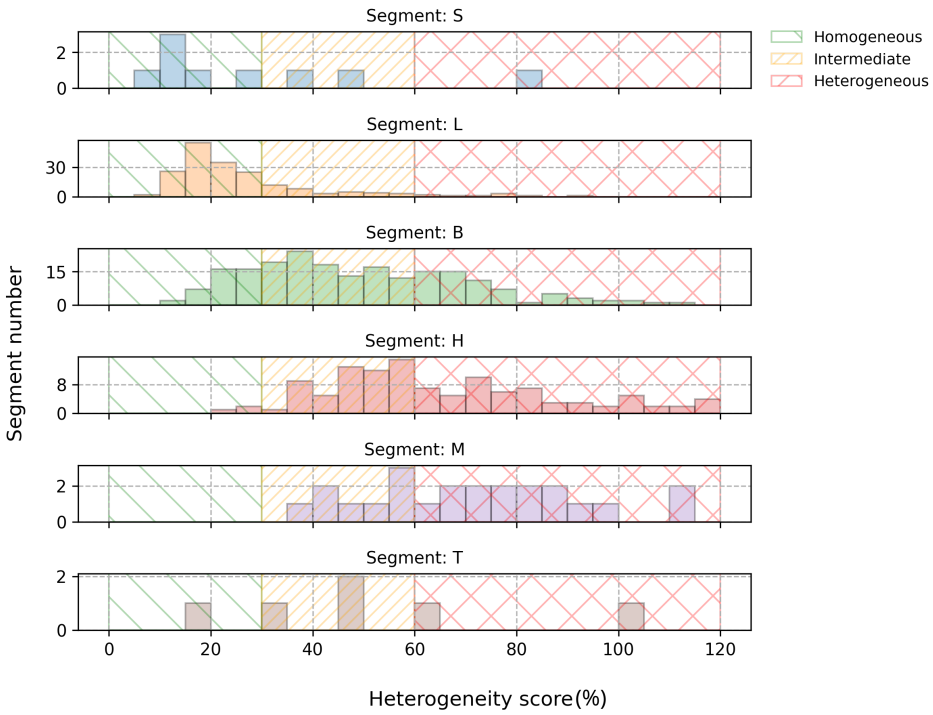

Figure 2: Distribution of the heterogeneity score $\mathcal{H}$ computed using only the three corrected wind components (U_{geo}, V_{geo}, W) for all MAESTRO segments (scaled by a factor 5/3 for direct comparison with the previous five-variable version). Scores are shown as histograms separated by segment type. The vertical dashed lines indicate the indicative thresholds for homogeneous ($\mathcal{H} < 30\%$, green), intermediate ($30\% \leq \mathcal{H} < 60\%$, orange), and heterogeneous ($\mathcal{H} \geq 60\%$, red) segments.

Major Issue 4: Heterogeneity score

I suggest revising the design and description of the heterogeneity score. If it has been introduced in literature before, please provide appropriate references.

4a)

The definition provided in Eq. (1) can be simplified to make your idea comprehensible: (i) Use either integration with respect to time (dt) or distance (dx), consistently. (ii) Do not use the same symbol T for time period and temperature. (iii) The factor 100 is actually irrelevant for segment definition and classification. It can be removed or placed in front of the whole expression.; (iv) Where appropriate, you could write an average over a segment denoted by overbar (as in Eq. (2)) instead of the integrals.

The equation has been completely rewritten in discrete form, which is also closer to its numerical implementation. The new formulation (Eqs. 2 and 3 in the revised manuscript) reads:

$$\mathcal{H} = \frac{1}{3 \cdot n} \sum_{p \in \{U_{\text{geo}}, V_{\text{geo}}, W\}} \sqrt{\sum_{i=1}^{n} \left(\frac{i}{n} - C_p(i) \right)^2} \tag{1}$$

where $C_p(i)$ is the normalized cumulative variance:

$$C_p(i) = \sum_{j=1}^{i} \frac{P_p(j)^2}{n\,\sigma_p^2} \tag{2}$$

This formulation addresses all four points of the remark: (i) consistent discrete indexing, no mixing of dt and dx; (ii) T no longer appears in this equation, instead $n = 6750$ the total number of data points in each segment is used; (iii) the factor of 100 is removed from the equation and $\mathcal{H}$ is described as expressed in percent in the text; (iv) the normalization by $n\sigma_p^2$ replaces the double integral and normalizes $\mathcal{H}$.

With the new formulation of the heterogeneity score presented here, we have revised the indicative heterogeneity thresholds downward. The new lower cutoffs were determined through visual inspection of representative time series, ensuring that the designated "homogeneous" range accurately excludes segments with obvious regime transitions or strong nonstationarities in any of the three wind components. Nevertheless, as before, these thresholds serve only as a guide for users and not as strict quality cutoffs; segment selection should still depend on the scientific context and the degree of homogeneity required for specific analyses.

4b)

As far as I understand the definition, the design of the heterogeneity score seems to be deficient in some aspects. (i) Due to employing the "cumulative variance", the score is not invariant to time reversal or shuffling the order of data points inside a series, which I would recognize as desirable properties for such a heterogeneity measure. A simple alternative, which can serve your purpose, might be the standard deviation of F^2 divided by the variance of F: std(F^2)/std(F)^2. (ii) Due to the last integration over time, the score depends on segment length, which should rather be avoided. I guess you want to normalize by ΔT as in Eq. (3). (iii) The second term inside the brackets representing steady linear accumulation of variance also needs to be normalized by ΔT. Otherwise it has different units than the first term in the brackets. (iv) Try to examine whether the range of possible heterogeneity values is bounded.

(i) the old and new discrete formulation are invariant to time reversal: reversing the time series mirrors $C_p(i)$ symmetrically, leaving the score unchanged. The revised manuscript states explicitly:

> "The score is invariant to time reversal but sensitive to the shuffling of time steps, which is the desired behavior for detecting regime transitions."

Regarding the suggested alternative std(F^2)/std$^2(F)$: this metric might capture intermittency well but is not sensitive to the spatial distribution of variance within a segment. Our score has been designed specifically to be sensitive to it, which is a necessary condition for the following optimization of segment boundaries. The objectives of such a definition are highlighted in the manuscript:

> "$\mathcal{H}$ has two purposes: (i) to be used in conjunction with the segmentation algorithm to optimize the partitioning of flight legs into segments representing single turbulent regimes, (ii) as a quantitative description of the wind field homogeneity for each sample."

(ii) The new formulation normalizes by n and uses the dimensionless index i/n. The score is therefore explicitly dimensionless and independent of segment length.

(iii) The new formulation compares two dimensionless quantities, $C_p(i) \in [0:1]$ and $i/n \in [0:1]$.

(iv) The minimum $\mathcal{H} = 0\%$ is achieved when variance accumulates exactly linearly. The maximum is achieved by a single Dirac-like step of variance, which gives the maximum possible value of 50 %. Indeed, the sum of i/n terms for i in $[1:n]$ is $(n+1)/2$, after dividing by n, the sum tends towards 0.5 for enough data points (here 6750). The score is therefore bounded in $[0\%, 50\%]$ and the revised text states:

"A single Dirac step of variance results in the maximum heterogeneity score $\mathcal{H} = 50\%$."

Major Issue 5: Manuscript composition and clarity

5a)

I suppose it would be natural if the composition of the manuscript reflects the order of the processing operations applied to the data, for instance: segmentation, calibrations, quality assessment, derivation of turbulence statistics and their errors. The dependencies among the processing steps should be explained clearly in the text. I got confused because they might have been convoluted, e.g. calibrated wind, temperature and humidity are needed to define the segments (through the heterogeneity score) but the calibrations are performed on the segments, implying the segments are defined previously.

Section 3 has been restructured to reflect the actual processing order:

1. Obtaining fluctuation time series, including segmentation algorithm and humidity calibrations
2. Statistical description of segments (moments, fluxes, integral length scales, spectral slopes in the inertial domain, TKE dissipation rate estimates)
3. Quality assessment (random and systematic errors)

As mentioned above, the circular logic was removed.

5b)

The volume of the presented material, in particular the number of plots, is large. For better readability I suggest focusing on the information which is most relevant for dataset users and removing or transferring to the appendix the secondary details. (i) As far as I know the methodology, the wind calibration accounting for the biases of the attack and sideslip angles is a standard practice for 5-hole probes. I expect the same or similar had to be executed in the past so that you get your initial wind signals. Thus, to my mind you basically updated such a calibration. If so, I see it rather as a detail of pre-processing. It can be briefly mentioned in the main part of the manuscript but described and evaluated in the appendix. The corrected wind can be directly used to derive all the dependent quantities in the dataset. (ii) Fig. 4 illustrating example time series for a random segment is rather superfluous because it does not convey any important information about the dataset. (iii) The analysis of the scaling of the inertial range (Figs. 9, 11, 13) is relevant only as far as it was important for the choice of the best sensors. Otherwise, only an example would be sufficient to illustrate the derived parameter included in the dataset, i.e. inertial range spectrum slope. (iv) Table 2 is not entirely consistent with the dataset content. Instead, the notations introduced there are rather confusing. Therefore, I suggest removing it.

We agree with the reviewer that there were many figures, but we still find relevant to keep information on the wind corrections. The following measures have been taken to reduce the volume of material presented in the core of the manuscript:

(i) Wind corrections are now described entirely in Appendix A, including the full wind retrieval equations and the reverse-heading validation. The main text refers to the appendix.

(ii) The example fluctuation time series figure has been removed from the main text.

(iii) The spectral comparison figures for individual sensors have been removed from the sensor-selection discussion.

(iv) Choice of the best temperature sensor, that mostly relied on high frequency arguments, is now dealt with in at the end of the instrumentation Section 2.3.

(v) The old Table 2 has been replaced by Table 3 of the revised manuscript. It provides a complete, readable, and accurate description of all three file types and their variables.

Major Issue 6: Instrument properties

The uncertainties of the directly measured quantities are not discussed. In sec. 2.2, I expected information about accuracy and response time for each sensor, not only the acquisition frequency. Please provide the complete model names and manufacturers of the instruments if they are commercially sold or relevant references documenting their properties if they have been developed exclusively for your research. It could be also helpful to state whether the sensors have been calibrated or tested in a lab, whether the slow-response ones (e.g. dew point hygrometer) have been corrected for their inertia or fast-response ones (e.g. spectroscopic hygrometers) for the impact of the flow around the fuselage.

A comprehensive instrumentation table (Table 2) has been added to Section 2.3. For each instrument, the table provides: model name, manufacturer, measurement principle, signal output frequency, internal response time, accuracy, and role in the final dataset.

Major Issue 7: Dataset composition

7a)

Instead of combining three different conventions (MAESTRO RFXX, ORCESTRA 2024-MM-DD, SAFIRE as2400XX), please select one of them.

The file naming convention is standardized among the MAESTRO and ORCESTRA communities, so we cannot change it as we would like. All turbulence dataset files distributed by national data centers use the consistent prefix `MAESTRO_ORCESTRA_ATR_TURBULENCE_` followed by the two-digit zero-padded MAESTRO flight number (`RFNN`), the SAFIRE identifier (`as2400ZZ`) as a secondary cross-reference, and the flight date (`YYYYMMDD`).

7b)

The keywords denoting three types of files - CALIBRATED, FLUCTUATIONS, MOMENTS for leg time series, segment time series and segment statistics, respectively - do not match their content. First, I understood all the files contain calibrated data, not only leg time series. Second, the segment time series contain not only the fluctuations but also the non-decomposed time series. Third, segment statistics involve not only turbulent moments but also integral length scales, inertial range slopes etc. You may consider LEG_ TIMESERIES, SEGMENT_ TIMESERIES, SEGMENT_ STATISTICS or similar.

The content of each file is now accurately described in Table 3.

7c)

Please use the same variable names and units for a given quantity consistently across the dataset, i.e. in the three types of files. In leg and segment time series, there are: ALTITUDE, LONGITUDE, STATIC_ PRESSURE, T (for potential temperature), U_ L, V_ T. In segment statistics there are: alt, lon, lat, PS, THETA, U, V. In segment statistics, most of the variables follow the pattern [statistics]_ [input_ variable], e.g. MEAN_ TS. However, there are a few which do not comply with it: UWE_ mean, VSN_ mean, WS_ mean, W_ mean, THETA_ mean, MR_ mean.

This brings unnecessary disorder. T in some places denotes temperature, in other potential temperature. On top of that, potential temperature is also sometimes denoted with THETA. I suggest a unified convention, e.g. T to denote potential temperature everywhere. In leg and segment time series, time is given in milliseconds; in segment statistics as a string.

Variable names have been harmonized across all three file types, as documented in Table 3. Potential temperature is uniformly denoted `T` with attribute `long_name = ''potential temperature''` and `units = ''K''` in all NetCDF files. Variables in `SEGMENT_STATISTICS` now consistently follow the `[STATISTIC]_[VARIABLE]` pattern. Time stamps in `SEGMENT_STATISTICS` are provided as strings.

7d)

Horizontal wind velocity components are considered in three coordinate systems related to: aircraft, mean wind and geographic directions. Wherever you mean a given component in a given coordinate system, please always use the same variable name and provide both the component name and coordinate system in the attributes. Noting only "streamwise", "longitudinal" or "transverse" without a coordinate system can be confusing because these can be understood to refer to the aircraft-oriented as well as mean-wind-oriented coordinate systems.

The revised Section 3.1 defines all three coordinate systems and their corresponding variable names: geographical frame (U_{geo}, V_{geo}; NetCDF: `U_geo`, `V_geo`), mean-wind frame (U_l, V_t; NetCDF: `U_l`, `V_t`), and aircraft frame (U_x, V_y; NetCDF: `U_x`, `V_y`). Attributes of horizontal wind variables in the NetCDF files now include `coordinate_system` specifying the coordinate system (either `geographical`, `mean-wind`, or `aircraft`).

Major Issue 8: Dataset content

The set of variables provided in the files do not agree with the description given in the manuscript (lines 430-441 and Table 2).

8a)

Segment statistics files include more variables than listed in Table 2, e.g. turbulent moments derived from detrended time series. To my mind, Table 2 is confusing and I consider it rather superfluous. The notations given there are not used neither in the manuscript nor in the dataset. In some cases, the notation goes against the convention adopted in variables names.

We agree that the old Table 2 was both incomplete and potentially misleading. It has been reworked entirely. In its place, the revised manuscript contains a single, comprehensive Table 3 that gives a complete and accurate inventory of all variables in all three file types (`SEGMENT_STATISTICS`, `SEGMENT_TIMESERIES`, and `LEG_TIMESERIES`). All variable names in the table match exactly those in the NetCDF files. Quantities derived from detrended series are explicitly listed with the `_DET` suffix. The table also documents quality metrics (Heterogeneity score, random and systematic errors, determination coefficient from humidity calibration) and additional derived quantities such as spectral slopes and TKE dissipation rate.

8b)

Segment time series files are supposed to contain 19 time series according to the text: 9 non-decomposed series (two horizontal wind components in three coordinate systems, vertical wind, water vapor mixing ratio and potential temperature), 5 fluctuation series from high-pass filtering (three wind components in one coordinate system, water vapor mixing ratio and potential temperature), 5analogous fluctuation series from detrending. I found only 14 and some of them lack

the attributes defining the coordinate system or derivation method (filtering or detrending). It looks like U_L_fluc, V_T_fluc are equal to U_c_fluc, V_c_fluc which makes me even more confused.

This was a genuine inconsistency between the manuscript description and the actual dataset content, and we thank the reviewer for identifying it precisely. The variables `U_L_fluc` and `V_T_fluc` were identical to `U_c_fluc` and `V_c_fluc`, respectively, due to a processing error. The aircraft-frame, now `U_x` and `V_y`, and mean-wind-frame, now $\texttt{U}_L and

8c)

Leg time series files are not precisely described in the manuscript. The text suggests they should include the same variables as for the segment time series but this is not the case.

We agree that the original description was ambiguous and could mislead users into expecting the same variable set as SEGMENT_TIMESERIES. The purpose and content of LEG_TIMESERIES files is now described explicitly in Section 4 and Table 3. These files cover entire flight elements, including legs with the MAESTRO leg-nomenclature, descents D, ascents A, and Velocity Azimuth Display maneuvers V, without any homogeneity constraint or segmentation. These additional flight elements (D, A, V) are now defined in the caption of Table 3 and in the text of Section 4. These files do not contain fluctuation time series. The revised Table 3 lists the LEG_TIMESERIES variables explicitly (15 in total): 6 coordinate/heading variables (time, ALTITUDE, LONGITUDE, LATITUDE, THDG, STATIC_PRESSURE) and 9 calibrated time series (U_l, V_t, U_geo, V_geo, U_x, V_y, W, T, MR).

8d)

For horizontal wind velocity, I recommend providing at least non-decomposed and high-pass filtered time series in all three coordinate systems (as the detrended can be easily recreated by a user). If there is a storage limitation involved, you can provide them in one selected coordinate system but then in the appendix please give the explicit equations to convert into other two coordinate systems.

The revised `SEGMENT_TIMESERIES` files include non-decomposed horizontal wind time series in all three coordinate systems (mean-wind, geographical, and aircraft frames), as well as high-pass filtered fluctuations in all three systems (6 non-decomposed and 6 fluctuation series for horizontal wind, plus the corresponding detrended equivalents). The computation of geographical wind coordinates is detailed in the new Appendix A. The conversion between coordinate systems is a rotation, and the definition of each frame is given explicitly in Section 3.1: the geographical frame (`U_geo`, `V_geo`) is fixed in space; the mean-wind frame (`U_l`, `V_t`) is obtained by projecting onto the direction of the segment-averaged wind; and the aircraft frame (`U_x`, `V_y`) is aligned with the aircraft longitudinal and transverse axes. No storage constraint prevents providing all three systems, and doing so avoids the need for users to perform coordinate rotations themselves.

Minor Issues (Manuscript)

Lines 28-39. Literature review extensively reports the results on shallow trade wind cumulus clouds. However, among the given targets of the MAESTRO campaign the equally important are deep convective clouds. If this is indeed the case, those could get a proportionate attention in the discussion of the past studies. In total, the review of previous works does not have to be long in such a paper describing a dataset.

The introduction has been improved to give more attention to deep convection. We added references and discussion of the campaign's sampling of mesoscale convective systems and the documented shift from trade-wind to deep-convection regimes in Section 2.1.

Lines 42–46. Please provide suitable references describing the listed campaigns.

References have been added as they became available for ORCESTRA (Stevens et al., 2025) PERCUSSION (Groß et al., 2026) and CLARINET (Bell, 2024). For campaigns without published references, they are described in the ORCESTRA overview article currently in preparation.

Sec. 2.3 (Flight sampling strategy). This subsection could get a simpler, logical structure if all the six types of horizontal transects are introduced together. Later you can discuss, which types were actually executed in which flights and why. It would be convenient to organize the subsequent figures so that the types appear in the order of their typical altitude, I suppose S, B, L, T, M, H. The division of the legs into homogeneous segments deserves a separate subsection.

Sec. 2.3 (Flight sampling strategy). How does the segmentation described here for turbulence measurements relate to the datasets from other instruments onboard ATR-42? Is the segmentation universal or specific for this one dataset, how can one match other observations with the turbulence data?

All six transect types (H, B, L, S, M, T) are now introduced together in a single bulleted list in Section 2.2. A sentence states:

> "This general leg nomenclature is consistent across all MAESTRO datasets, while the specific segmentation described in Section 3.1.1 applies to the turbulence dataset presented here."

T and M segments were flown at variable altitudes in the troposphere (from 1 to 3 km). For this reason, they are generally not presented in the Figures. All Figures describing variables for each leg/segment type, are presented in altitude increasing order, from top to bottom.

Lines 170–187. Are there any particular arguments behind the exact thresholds (30 and 60%) of the heterogeneity score?

With the new heterogeneity score definition, the thresholds have been revised to 10% and 20%, and the choice is justified by visual inspection of representative time series at each threshold level. The thresholds are explicitly presented as indicative guidance, not strict quality gates.

Fig. 5. The caption should describe the content of the panels, including what the dots denote and what "std" in the legend means.

The figure caption has been updated to describe the legend of panel (c) explicitly.

Line 226. Please justify why you decided to compute dissipation rate from the vertical velocity component.

The following justification has been added to Section 3.2.2:

> "Following standard practice for aircraft turbulence measurements (Lenschow, 1986), ε is computed from the vertical wind component W, which is less affected by mesoscale horizontal variability and more reliably captures isotropic fluctuations in the inertial subrange."

Another argument has been added which is to be able to readily compare with previous studies:

> "Similar methodological choices have been made in comparable aircraft studies (Brilouet et al., 2021; Shaw and Businger, 1985; Wacławczyk et al., 2020, 2017; Jen-La Plante

et al., 2016; Malinowski et al., 2013), enabling direct comparison with prior field campaigns."

Line 227. The inertial range can be resolved only up to the Nyquist frequency, hence 0.8 rad m^{-1}.

Corrected. The revised manuscript states: "the Nyquist wavenumber (12.5 Hz $\approx$ 0.8 rad m^{-1} at 100 m s^{-1})."

Line 230. σ_f is the square root of the definite integral of the vertical wind spectrum, not the "variance of the integrated vertical wind spectrum".

Corrected. The revised text now reads:

"$\sigma_f = \sqrt{\int_{k_{IL}}^{k_{\max}} S_W(k)\,\mathrm{d}k}$ is computed by integrating the W power spectral density over the inertial subrange."

Line 231. Please provide a source for the value of the Kolmogorov constant.

The Kolmogorov constant $\alpha = 0.52$ is now cited as obtained from Sreenivasan (1995) and Wyngaard and Coté (1971).

Line 270-273. You wrote earlier (line 246) that the average vertical wind was already zero in straight legs before applying the wind correction. How can you then obtain the biases of the attack and sideslip angles by minimizing the average vertical wind in stabilized legs even further? In general, the sideslip angle controls mostly the coupling between the horizontal velocity components and does not significantly influence the vertical wind retrieval in horizontal stabilized legs. How the minimization of the average vertical wind can provide the sideslip angle correction then? This aspect is actually indicated by your Fig. 12, taking into account the negligible bias obtained for the attack angle and the measurable bias for the sideslip angle.

We understand that the link between sideslip angle and vertical velocity might seem counter-intuitive at first. Two points clarify this: first, the averaged W value over all stabilized legs of the campaign had a systematic bias of $\sim$0.6 m s^{-1}. which is what the optimization minimizes. Second, the reviewer is correct that the sideslip angle primarily affects the direction and magnitude of the horizontal component in no roll conditions. However, for even small roll values, any sideslip bias will project horizontal components onto the vertical. In general, the magnitude of horizontal wind components is much larger than that of the vertical one, translating into correlation between roll and vertical velocity. This is then heavily reflected during turns where roll values are typically much higher. The simultaneous optimization of the incidence and attack angles on the stabilized legs allowed their combined contribution to be disentangled in all cases. The effectiveness of this approach is demonstrated by the reduction of the roll-W correlation from 25% to 1.4% during turns. All these elements are explained in Appendix A, which also include the full equations used to retrieve the wind components. The reader can now see that the roll angle ϕ appears in the expression of the vertical component.

Line 392. What exactly is meant by "only statistically significant population of segments"?

Replaced by:

"For the H category, which constitutes the only segment type with a statistically large sample size outside the boundary layer."

Sec. 3.3.7 (Systematic error). Using the term "systematic error" is misleading here because what you computed is not the systematic error as introduced by Lenschow et al. (1994). They considered the inherent systematic error of turbulent moment resulting from approximating the ensemble mean by the temporal average of a finite segment. Yours is a consequence of removing a range of scales by the particular choice of cutoff frequency in high-pass filtering. I would not call it "error" but rather "filtering effect" or similar.

This is somewhat true. However, we choose to keep this term, as it was employed in (Brilouet et al. 2021) for this exact error estimation. The revised text explicitly states:

> "It is not a systematic error in the classical sense (cf. Lenschow et al., 1994, who define systematic error as the inherent bias from approximating ensemble means with finite-sample temporal averages. Instead, it quantifies the variance suppressed by the choice of cutoff scale and method."

Line 473. The cited raw dataset seems to include only temperature from the Rosemount and humidity from the FAST-WAVE, not the entire campaign data involved in the preparation of the turbulence dataset (i.e. raw wind, aircraft state, temperature and humidity from other sensors). Please clarify that.

The Code and data availability section has been updated with two separate citations: Bony (2024a) for the fast temperature and FAST-WAVE humidity data, and Bony (2024b) for the full SAFIRE temperature and humidity 1Hz dataset.

Minor Issues (Dataset)

Please explain what it means when data records contain NAN values. Is it due to the failure of instrumentation, unacceptable quality of measurement or failure of derivation method?

NaN values indicate either instrument failure or missing data. This information has been added in the dataset summary (Section 4):

> "Nan values can be present in the statistics and legs time series files. In the moment files, inertial slopes may contain NaN values if the inertial range is too small to be able to derive a reliable slope estimate or if the regression coefficient is below 0.9. The TKE dissipation rate ε can also take NaN values if the inertial range of the vertical wind component spectrum is too small. In the leg time series, NaN values are associated with INS malfunctions affecting altitude, latitude, longitude, static pressure, true heading, and all wind components. NaN values can be found in 12 of the 269 leg time series files. In the segment time series, NaN values are avoided through the use of the segmentation algorithm."

(segment statistics) Together with the inertial range spectrum slopes, you can also provide the corresponding correlation coefficients in log-log space, which you analyze in the manuscript.

Following the revisions described in the response to Major Issue 2 (use compliance with Kolmogorov theory for sensor selection), the spectral correlation coefficients are no longer analyzed in the manuscript. We have therefore chosen not to include them in the dataset. The regression is performed on data points that are equally spaced in log-log space, which is a practically efficient choice but does not correspond to a statistically rigorous regression. In particular, the resulting determination coefficient R^2 value depends on the binning choice and does not have a straightforward probabilistic interpretation. We consider the spectral slope itself, combined with the other quality indicators already provided (heterogeneity score, random error, and variance), to be sufficient guidance for users wishing to assess the validity of the inertial-subrange assumption for a given segment.

(segment statistics) It seems the variables [VAR, M3, SKEW]_[U_L/U_L_DET/V_T/V_T_DET] are exactly equal to [VAR, M3, SKEW]_[U/U_DET/V/V_DET]. If so, one of these sets can be removed.

This was an issue in the processing (former `U_T` and `V_T` were equal to the former `U` and `V`, respectively) that is now fixed. It followed that derived fluctuations and statistics were therefore

the same. This has been fixed with wind components having correct values in each coordinate system.

(segment time series) The records are actually longer (5.5 min) then given in the text and indicated by the time bounds in the corresponding segment statistics (4.5 min). Please explain why.

A 30-second buffer on each side of the 4.5-minute core segment was used to mitigate edge effects when applying high pass filters. They should however not have been provided as such. Segments are now properly cropped and all have the standard 4.5-minute duration.

(leg time series) The files apparently involve additional leg types (D, A, V), which are not mentioned in the manuscript. Please correct the file names or define the extra types.

The leg types D (descent), A (ascent), and V (Velocity Azimuth Display) are now explicitly mentioned in Section 4 and in the caption of Table 3.

Technical Issues (Manuscript)

Line 4. Mesoscale cloud organization is already mentioned as a goal in the previous sentence.

The redundant phrase has been removed from the abstract.

Lines 8 and 81. The correct form in this meaning is probably "consists of".

Corrected in both locations.

Line 35. Please give details on the cited reference (Bony et al., in prep.).

The article has been published since then. Its reference is now cited properly.

Lines 172ff. The heterogeneity scores are given in % in the text but with no units in the plots.

Units (%) have been added to the x-axis labels of all heterogeneity score figures.

Line 176. Probably, you mean orange hashed area in Fig. 3.

Corrected.

Line 247. Correlation of what?

The whole wind correction part was reworked and moved to the dedicated appendix section A1. It now clearly states: "Initial analysis of the MAESTRO dataset revealed two significant indicators of wind measurement issues: (1) non-negligible averaged vertical velocity (W) values during stabilized legs (0.6 $\pm$0.1 m.s^1, whereas values close to 0 $\pm$0.1 m.s^1 are expected for sufficiently long averaging periods), and (2) substantial correlation (25%) between the roll angle ϕ and W outside of straight legs."

Line 251. Probably, you mean "from the air velocity with respect to the aircraft".

Corrected.

Line 323. Probably, you mean "wavelengths smaller than the integral length scale".

Clarified to: " The power spectral density is computed using a Fast Fourier Transform, and a linear regression in log-log space is performed over the inertial subrange, defined from the wavenumber corresponding to the integral length scale $k_{IL} = 2/L_x$ up to the Nyquist wavenumber."

Fig. 11. Probably, you mean temperature sensors in the caption.

Corresponding figure was removed.

Line 385. "Valid" might be better than "relevant" in this context.

This was rephrased.

Line 432. Probably, you mean segment instead of flight.

Corrected.

Technical Issues (Dataset)

We thank the reviewer for reviewing the content of the datasets and, in particular, for highlighting missing attributes in some variables.

The access via OpenDAP protocol through Thredds server does not work for leg time series and segment time series.

We thank the reviewer for reporting this. We have reported the issue to the AERIS team, but resolving it falls outside the scope of the present manuscript revision. The data remain fully accessible via direct HTTP download from the AERIS data portal at the DOI provided in the Data Availability section.

MEAN_ WD is probably given in radians instead of degrees.

Confirmed and corrected. `MEAN_WD` is now in degrees with `units = ‘‘degrees’’`.

THETA_ mean is given in K instead of Celsius.

The variable is now named `MEAN_T` with `units = ‘‘K’’` and `long_name = ‘‘mean potential temperature’’` consistently throughout.

MEAN_ TS has a wrong unit given in the attributes.

Corrected. `MEAN_TS` now has `units = ‘‘K’’`.

Time is given in different string pattern than specified in the attributes.

The time string format has been standardized to (`YYYY-MM-DDThh:mm:ss`) in all `SEGMENT_STATISTICS` files.

There are variables lacking attributes: date, VAR_ U_ L ... SKEW_ V_ T.

All variables now have complete CF-compliant attributes.

There are variables lacking attributes: U_ c, V_ c.

These variables have been renamed to `U_l` and `V_t` with full attributes.

For T, long_ name disagrees with standard_ name.

Corrected. `T` now has `long_name = ‘‘potential temperature’’` and `standard_name = ‘‘air_potential_tempe`

Responses to Reviewer 1

General remarks

We thank Reviewer 1 for the detailed reading. The general points raised have been addressed in the revised version.

The manuscript describes a data set of atmospheric observations measured above Cabo Verde using the ATR42 research aircraft as part of the MAESTRO campaign. The basic meteorological parameters are stored with a relatively high temporal resolution of 25 Hz. In total, data sets from 24 measurement flights between August 10 and September 10, 2024, are available. Most of the data was measured on horizontal flight paths, with flight altitudes ranging from low altitudes of up to 60 m above sea level to 6000 m, with the sections at higher altitudes being used primarily for downward-facing remote sensing methods.

The focus is on the high temporal resolution with associated spatial resolution of a few meters, which also allows turbulence investigations. While the wind vector and air temperature were measured using the aircraft's standard instrumentation, the campaign provided an opportunity to use new, high-resolution humidity sensors.

After an introduction that essentially presents the overarching ORCHESTRA initiative and the various aircraft campaigns associated with it, the actual motivation behind the MAESTRO campaign is briefly presented. Although the spatial distribution of clouds and their microphysical properties were also mentioned as motivation for the campaign, this dataset focuses more on the sub-cloud layer and the connection to the air masses below. The data set presented here does not contain any further information on cloud flights, which at first glance seems a little confusing given the motivation behind the campaign.

The second chapter presents the measurement campaign itself, the instrumentation, and the measurement strategy. However, I see major problems, particularly in the presentation of the instrumentation: in my opinion, the section lacks structure. Although reference is made to further literature, various details of the systems presented are selected and presented somewhat arbitrarily. Furthermore, it is very difficult to identify the sensors precisely, as they are often only mentioned using abbreviations without manufacturer details, etc. In my opinion, this chapter does not meet the standards for the introduction of instrumentation. I would have expected at least one sensor table with manufacturers, response time, and accuracy at this point. Particularly with regard to the typical acquisition frequencies of the individual sensors, there is often no clear distinction between the actual sampling frequencies and the native temporal resolution, i.e., the response time. Unfortunately, I cannot get an impression of the quality and accuracy of the sensors from this section. For example, two capacitive humidity sensors are mentioned that are supposedly calibrated, but there is no indication of how and against which standard this was done.

The description of the flight strategy and the associated classification into "types" is somewhat unusual, but it is certainly possible to do it this way. I am somewhat critical of the further subdivision of the legs into "homogeneous sections" with regard to the fluctuations of five selected parameters, and I am not sure what benefit this classification has for users of the data. I have the impression that this classification is only of interest to users who want to keep their own data analysis effort to a minimum. That is certainly legitimate, but it also limits the possibilities of data analysis somewhat. In my opinion, the explanation of the heterogeneity factor requires a few more clarifications (see detailed comments).

I see a general problem in the order of sections 2 and 3: Section 2 describes the division of observations according to flight patterns and homogeneous sections, but this is followed in section 3 by the more fundamental processing and calibration of the data. The order seems unfortunate to me, as it does not correspond to the actual order of post-processing. I suggest reconsidering

this order.

My last major point relates to the "calibration" of the sensors for rapid measurements, which is explained in great detail and with many illustrations. It is certainly standard practice to combine fast but rather inaccurate and fragile sensors with comparatively slow but more robust and easier to calibrate sensors. However, I am not yet entirely convinced by the method proposed here.

First, there is no mention of whether the slower sensors are corrected for their inertia before they are correlated with the faster sensors. Secondly, as already noted, there is no information about the accuracy of the sensors used as a reference. And finally, I do not understand why it is of interest to perform calibration for the different types of legs – is it assumed that the sensors behave differently at 60 m ASL than at 300 m ASL?

What I find completely missing in this data overview paper is a meteorological or synoptic classification of the 24 flights. Under what conditions were the flights carried out? This information would be very helpful for further use of the data, especially with regard to the cloud situation.

In summary, I believe that this manuscript definitely needs major revisions, although the data set itself is of great interest to external users.

A comprehensive instrumentation table (Table1) has been added, explicitly distinguishing the stored sampling frequency from the native instrument response time for each sensor.

The rationale for segment-scale calibration is now fully explained in Section 3.1.2 with explicit physical arguments. In particular, two reasons motivating per-segment rather than per-flight calibration are invoked:

> "The calibration is performed independently for each segment rather than once per flight or per campaign, for two reasons. First, fast sensors on airborne platforms can experience gradual or abrupt changes in their response characteristics during a flight, due to exposure to clouds, precipitation, salt, and wide ranges of temperature and humidity; a segment-scale calibration accounts for these slow drifts and environment-induced changes. And second, the MAESTRO flight legs sample distinct meteorological regimes at very different altitudes: temperature ranges from approximately +28 °C near the surface to −10 °C at 6000 m, and water vapor mixing ratio spans nearly an order of magnitude between the boundary layer and the mid-troposphere. Capacitive and spectroscopic sensors are known to exhibit environment-dependent sensitivities: the sensitivity of capacitive sensors to relative humidity is a nonlinear function of temperature, while laser absorption spectrometers can be affected by pressure and temperature variations. A segment-scale calibration ensures that the retrieved slope and offset are locally appropriate for the thermodynamic regime actually sampled, rather than imposing a single correction averaged over heterogeneous conditions."

Table 1 of the revised manuscript provides a synoptic and meteorological classification of all 24 flights, including cloud types and mesoscale features, cloud cover estimates, maximum cloud top, and adopted sampling strategy.

The processing section has been reordered so that the workflow follows the actual sequence of operations: segmentation, humidity calibration, and turbulence statistics. In addition, wind corrections have been moved to Appendix A.

Finally, a clarification has been added in the introduction to explain that, while the MAESTRO campaign targets both cloud properties and boundary-layer dynamics, the present turbulence dataset focuses on the sub-cloud layer and cloud-base environment; cloud microphysics and remote sensing data are published separately.

Abstract

I suggest to mention the true airspeed in addition to the frequency of the stored data.

In the abstract, we have added a mention of expected air speed:

> "25 Hz segmented time series [...] taken with an air speed of approximately 100 $\mathrm{m\,s^{-1}}$, relative to the aircraft."

Line 9ff: please explain what is meant by "turbulent moments"; you probably mean the statistical moments of the probability density function of the individual parameters. Please use precise terminology throughout the manuscript.

The term "turbulent moments" has been replaced throughout the manuscript by the more rigorous formulation "statistical moments".

Introduction

Line 20: Why only absorbing solar radiation? What about terrestrial irradiance and also reflection of solar irradiance?

Revised to:

> "These clouds interact with their environment by exchanging latent heat through evaporation and condensation, and by interacting with solar and terrestrial radiation."

Line 28/29: The statement is somewhat vague; above all, I would rather say that the warm ocean provides latent heat for convection, which is then also associated with turbulence.

Revised to:

> "the warm tropical ocean supplies the latent heat that drives convection and the associated turbulence."

Line 39 to 52: The transition from EUREC4A to the ORCHESTRA campaigns was quite complicated for me to read. In particular, the fact that ORCHESTRA is a combination of several measurement campaigns was mentioned rather late in the text, which I found somewhat confusing when reading it for the first time. Perhaps the section could be revised slightly to make it easier to understand.

The paragraph introducing ORCESTRA begins directly with its definition as a network of international field campaigns ("From August to October 2024, a network of international field campaigns, named ORCESTRA [...]"), unambiguously establishing its nature as a cooperation initiative of individual campaigns, that are listed right after. The connection from EUREC4A to ORCESTRA is also prepared by a transitional sentence highlighting that EUREC4A raised the scientific questions that motivated new field works. We believe the current structure is sufficiently clear.

Line 44/45: "This campaign follows ..." What exactly do you mean with "following"?

Revised to a more precise formulation:

> "This campaign pursues the scientific questions that arose from the earlier mentioned EUREC4A project."

Line 57/58: "in-situ turbulent scales ... " the scales are not turbulent, better: "in-situ observations of typical turbulence scales and fluxes in the sub-cloud layer.." ?

Revised to:

"in-situ observations of typical turbulence scales and fluxes in the sub-cloud layer."

Section 2.1: The MAESTRO field campaign: acquisition strategy

Line 1-3: However, none of the three objectives for the MAESTRO experiment primarily requires high-resolution turbulence data, correct?

The motivation to acquire turbulence data is now clearly expressed. MAESTRO objectives unrelated to turbulence are not mentioned outside the introduction anymore. The last paragraph of section 2.1 has been revised:

> "The MAESTRO project primarily aims to test hypotheses regarding the mesoscale organization of shallow and deep convection. Central to this investigation is the role of coherent structures in the boundary layer, such as cold pools and convective plumes, which are believed to significantly influence cloud development and thus contribute to mesoscale convective organization. These elements can affect the statistical properties of the flow (de Szoeke et al., 2017; Zilitinkevich et al., 2006), underscoring the importance of characterizing turbulence to identify their signatures.[...]"

Figure 1: The labels are quite small and bright - but might be okay. Are really all 24 flight patterns included in Fig 1?

We are aware of the difficulty to follow each flight on the map of Figure 1. All 24 research flights are indeed represented. Given the repeated sampling of the same geographical area, significant overlap between flight tracks is unavoidable. The primary purpose of this figure is to convey the overall spatial extent and geographical focus of the campaign rather than to allow the precise identification of each individual flights.

Section 2.2: Instrumentation

Section 2.2 on instrumentation should be improved in general: on the one hand, it states that the entire data set was published with a temporal resolution of 25 Hz; in Sec 2.2, the individual sampling frequency is specified again for each sensor, which is not really relevant; much more interesting is the actual temporal resolution/response time of the individual sensors.

For example, if I provide the dew point mirror data at 25 Hz even though I know that it cannot resolve this frequency, this is important information for the user of the data. If, on the other hand, I learn that the LiCor has a response time of 50 ms —i.e., 20 Hz— but the signal is sampled at 50 Hz and then made available at 25 Hz, I wonder what I am supposed to do with the individual pieces of information.

Finally, two capacitive humidity sensors are mentioned; one is sampled at 1 Hz, the other at 40 Hz, but what is the actual response time of the sensors? On the one hand, I am missing important information here, and on the other hand, rather unimportant information is provided - this should be sorted out a little and consideration given to what is really of interest to the user.

The caption of the new instrumentation table (Table 2) makes sure the reader is aware of this distinction:

> "The table distinguishes the sampling rate from the native instrument response."

Line 102: be consistent with the nomenclature of units in line 96 you write "m.s-1", here you write "°C/sec" which should be "K.s^{-1}", furthermore, it should be the "response" and not the "response time" which has unit of "s".

All units have been standardized throughout the manuscript. Response is now given in $\mathrm{K\,s^{-1}}$ for the 1011C warming rate, and notation "$\mathrm{m.s^{-1}}$" has been standardized to "$\mathrm{m\,s^{-1}}$" throughout. In

Table 2, the warming rate is still given in the "Internal Response Time" column, but with the "(warming rate)" mention.

Section 2.3: Flight sampling strategy

Why not describing all applied leg types together in the same way? In line 115 to 120 you describe three (major) types with bullets and the "S-type" later on in the text. And then in line 149/150 you define even more types. I suggest summarizing them and mentioning that some types were flown more frequently and others were more of an exception.

All six leg types (H, B, L, S, M, T) are now summarized together in a single bulleted list in Section 2.2. The text of this section clarifies which types were standard and which were occasional.

Line 154: "turbulent dataset" makes no sense, you probably mean a set of "turbulence data", same in the next line "5 turbulent fluctuations" makes no sense.

Corrected throughout: "turbulence" instead of "turbulent".

Equation 1: In my opinion, the explanation of the heterogeneity factor requires a few more clarifications, for example, what exactly the function F means. I also don't quite understand why the inner integrals in the fraction go over x and the outer integral then over time t? Did you come up with this parameter yourself or are there references for it?

The equation for the heterogeneity score has been completely rewritten in discrete form, which is also closer to its numerical implementation. The new formulation (Eqs. 2 and 3 in the revised manuscript) reads:

$$\mathcal{H} = \frac{1}{3 \cdot n} \sum_{p \in \{U_{\mathrm{geo}}, V_{\mathrm{geo}}, W\}} \sqrt{\sum_{i=1}^{n} \left(\frac{i}{n} - C_p(i) \right)^2} \tag{3}$$

where $C_p(i)$ is the normalized cumulative variance:

$$C_p(i) = \sum_{j=1}^{i} \frac{P_p(j)^2}{n\,\sigma_p^2} \tag{4}$$

The former function F is now clearly identified as the wind perturbation time series $P_p(j)$ for component p. The inconsistent mixing of dx and dt is eliminated by the discrete formulation. Although the score has been heavily modified, which is why it needs to be fully described in the text, the original idea on which it is based is now cited in the text:

> "The design of this score is based on work done by Bernard-Trottolo (2001). While the original definition only used the vertical wind component, we added both horizontal components to account for heterogeneity in these directions."

Line 206/207: "...projecting the geographical wind components onto the mean wind at the segment scale ...". I understand what you mean, but the sentence doesn't make sense. Please rephrase it.

Revised to the more rigorous formulation:

> "obtained by projecting the geographical wind components onto the direction of the segment-averaged wind."

Figure 5: panel c, label of y-axis: please delete the bracket; and about the title of panel c: "power density spectrum of water vapor mixing ratio" - not of an instrument.

Corrected. Panel (c) subtitle is now "Power density spectrum". This is more consistent with titles of panel (a) and (b) ("Autocorrelation function" and "Integral of autocorrelation function", respectively).

Equation 4: I think this equation requires some additional background information. For example, it would be important to know how large the error is if the spectrum (e.g., in Fig. 5c) does not have the theoretical slope of -5/3 that you assume here (but do not mention anywhere).

We thank the reviewer for this important remark. Deviations from the homogeneous isotropic turbulence are a common occurrence in the atmosphere. Understanding the physical reasons or mechanisms behind them is important and an active research topic (among others Nowak et al. 2025, Zilitinkevitch et al. 2021). This topic is of interest for the authors, who are currently investigating these deviations in the MAESTRO dataset. The streamwise and transverse horizontal wind components could also be used to compute alternative ε values. However, we do not wish to go beyond historically used methodologies for this estimation in the current article. We found, however, that evaluating the error induced by the spectral exponent assumption was a valid request.

A minimalist sensitivity analysis has been performed and is described together with its results in Section 3.2.2:

> "Let us consider the uncertainty of ε associated with the assumption $S_W(k) \propto k^{-5/3}$ throughout the integration range in Eq. (8). When the observed spectral slope deviates from $-5/3$, the integrated vertical wind variance σ_f^2 changes, introducing a bias in the retrieved ε. To quantify this sensitivity, we numerically evaluated Eq. (9) for a spectrum of the form $S_W(k) = \frac{4}{3}\alpha\varepsilon^{2/3}k^s$, where the slope $s = -5/3 + \delta$ is varied. ε is kept constant, and the retrieval of ε uses Eq. (9), still assuming the $-5/3$ slope and meaning the error associated with a slope deviation is not only sensitive on δ, but also slightly on the integration range as σ_f is still evaluated on the s-scaled spectrum.
>
> For parameters representative of L-type segments (integral length scale $L_W = 50$ m corresponding to $k_{IL} \approx 0.13$ rad m^{-1}, $k_{\max}$ equal to the Nyquist wavenumber ≈ 0.8 rad m^{-1}), a slope deviation of $\delta = +0.1$ (i.e., an observed slope of -1.57 instead of $-5/3$) leads to an underestimation of ε by approximately -18 %, while a deviation of $\delta = -0.1$ (slope of -1.77) causes an overestimation of approximately +23 %. For parameters representative of H-type segments (integral length scale $L_W = 200$ m corresponding to $k_{IL} \approx 0.03$ rad m^{-1}, $k_{\max}$ equal to the Nyquist wavenumber ≈ 0.8 rad m^{-1}), a slope deviation of $\delta = +0.1$ (i.e., an observed slope of -1.57 instead of $-5/3$) leads to an underestimation of ε by approximately -30 %, while a deviation of $\delta = -0.1$ (slope of -1.77) causes an overestimation of approximately +43 %. Segments where the parameter corresponding to the extracted spectral slope of the vertical wind (Slope_W) deviates substantially from $-5/3$ should be treated with caution when interpreting absolute values of ε."

For the sake of completeness, we add the full results of the sensitivity test to this review in following Figure 3.

Line 230: σ_f is a standard deviation, or maybe you missed the power 2?

Corrected, we do not use the term variance for σ_f anymore.

Line 231/232: I cannot follow this line of reasoning: where in Eq 4 can I see that the energy dissipation rate depends on the velocity gradients? And where does wind shear appear in the equation?

The original text was misleading. The statement about velocity gradients and shear has been

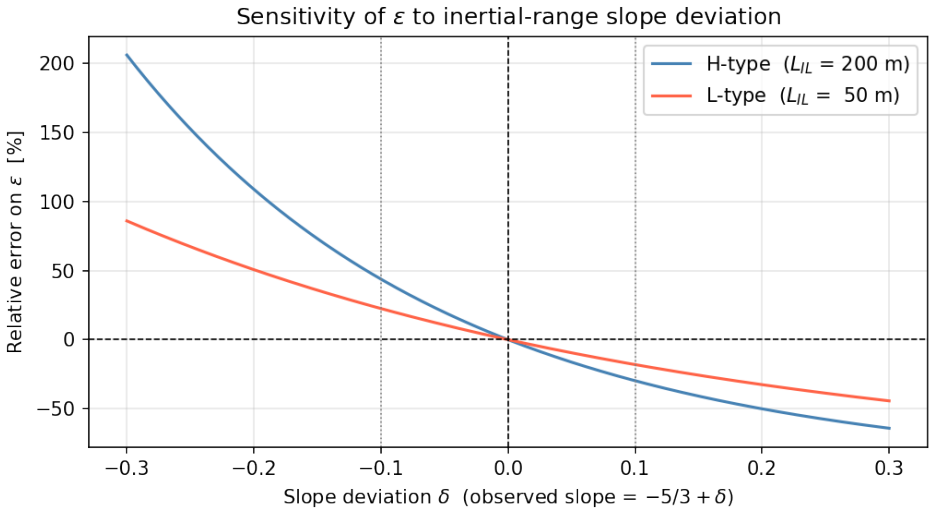

Figure 3: Sensitivity test for TKE dissipation rate ε (in percent), as a function of -5/3 deviation, for two typical integral length scale values. Positive/negative values of δ correspond to shallower/steeper slopes in the inertial subrange.

removed. The revised text presents Eq. (6) as a spectral method based on Kolmogorov's inertial-subrange scaling, without incorrectly invoking velocity gradients or shear.

Line 233/234: The distributions of energy dissipation rates shown in Fig. 17 show very low values, especially in the H segment, and I have serious doubts that a measurement system on an aircraft can accurately resolve dissipation rates down to 10-8 m2s-3. Can you please estimate the measurement accuracy or maximum resolution?

This concern is fully justified. We have added a detailed analysis of the spectral noise floor acknowledging the work of Muschinsky et al. 2001 (Section 3.2.2). Instrumental noise begins to influence the spectrum when spectral density values fall below approximately 2×10^{-6} m^1 s^{-2}. This affects exclusively some H-type segments; for all boundary-layer segment types (S, L, B), no noise contamination was detected. A dynamic k_{max} truncation based on a conservative noise threshold for spectral density (1×10^{-5} m^1 s^{-2}) prevents noise from biasing ε estimates. Our methodology allows the computation of very low ε values.

Section 3.2: Pre-Processing

The beginning of Section 3.2 is very confusing, as it appears to start with the introduction of a fast moisture sensor, but then leads indirectly to the issue of wind determination. This should be restructured somewhat, as the content does not really match the heading of Section 3.2.

The beginning of the pre-processing section has been restructured. The introduction of FAST-WAVE has been moved to Section 2.3 (Instrumentation). Section 3 now opens with a clear summary of the processing order.

Section 3.2.1: Wind corrections

Note: This section has now been moved to the appendix (Appendix A).

Am I correct in understanding that the wind vector is derived from the combination of the 5-hole arrangement and an additional Pitot-static tube? Does the central hole take on the function of a Pitot tube?

The five-hole radome measures differential pressures for angle of attack and sideslip. The central hole of the radome measures total pressure, while a separate Pitot-static system is used for static pressure. These details are provided in Table 2. The revised Appendix A clarifies the wind retrieval methodology and explicitly shows the equations used. True airspeed is derived from the total-to-static pressure difference. The full wind component equations and all variable definitions are now given in Appendix A.

Line 257: I don't quite see it that way: I also need the same, if not greater, accuracy in the measurements to determine the average wind vector. When calculating the fluctuations, average values are subtracted, and the difference in determining the horizontal wind during flight maneuvers, for example, will be significantly smaller.

We agree that accuracy in the mean wind is the same as for fluctuations, and we did not intend to imply otherwise.

Line 263: I think the initial guess is displayed in Fig 6a and not in 6b - correct?

Figures 6a and 6c (now A1a and c) displayed horizontal and vertical components of the original wind computations, respectively.

Line 268: Please check the panels which might disagree with the figure caption.

Caption has been verified and is correct.

Line 274: Why do you assume it's a "sensor defect"? There could be many different causes: installation errors with the IMU, calibration errors, and so on. With "sensor defect," I would really suspect a broken sensor.

The term "sensor defect" was indeed technically incorrect. It has been replaced by "calibration issue" and "systematic offset".

Section 3.2.2: Calibrations of temperature and humidity

Line 295ff: When describing the four steps, I don't quite understand the reference to the individual panels in Fig. 7. In the first step, for example, you mention low-pass filtering of all four moisture measurement time series and refer to panel 7a, which shows the unfiltered data — is that correct? The low-pass filtered data is shown in 7b. Does that match the description of the second step? And finally, in 7d, the "fast-wave" data is shown as corrected for offset and slope relative to the other three humidity sensors. However, the four time series in 7d still differ by at least one constant offset — what is the most likely solution?

Panel (a) shows the uncalibrated time series; panel (b) the low-pass filtered series; panel (c) the linear regression; panel (d) the calibrated FAST-WAVE alongside the three reference sensors. The residual offset visible in panel (d) reflects the known absolute differences between reference sensors, as characterized in Appendix B.

In terms of the absolute accuracy of the humidity sensors, the dew point mirror should have the best quality (although I am not very familiar with the other sensors), but it has the problem of high temporal inertia. Why is this property not consistently exploited and at least the offset determined via the mean value? I cannot quite follow this description of the calibration and therefore have serious doubts as to whether this is the right way to obtain high-quality humidity data.

This is addressed in Appendix B. The 1011C's thermal inertia causes asymmetric biases in highly heterogeneous conditions, relative to the adaptation speed of the sensor: it systematically underestimates the mean mixing ratio in segments with high horizontal humidity gradients, in particular when the airplane traverses clouds. Using the 1011C to determine even the mean

offset would introduce a systematic difference, depending on the presence of clouds. The WVSS-2, validated against the 1011C in 90.4% of segments, is the more reliable reference.

Line 315: The sentence doesn't make sense to me; you write something about fast calibration, but you mean the calibration of the data from the fast sensor — is that right?

Correct. Modified to:

> "The outcome of the calibrations of fast humidity sensors"

Line 318-320: In summary, I still have some doubts about the calibration of the fast sensors. Either I don't understand the procedure correctly, or the method is flawed. Using the correlation of two sensor signals to say something about the suitability of one sensor as a reference is somewhat risky, as the properties of a reference sensor are completely ignored. But as I said, maybe I have misunderstood something here and you can quickly convince me that this is the right way to do it.

In the revised manuscript, the justification for selecting the WVSS-2 as primary reference is now grounded in three independent lines of evidence, described in Section 3.1.2 and in the new Appendix B:

1. The 1011C chilled-mirror provides a direct condensation measurement and is in principle the most accurate instrument on board. A dedicated intercomparison between the WVSS-2 and the 1011C (Appendix B, Figs. B1–B3) shows that the WVSS-2 agrees with the 1011C within ±5% for 90.4% of all segments (509 out of 563), across the full range of conditions sampled during MAESTRO. This comparison was performed entirely independently of the fast sensors being calibrated. Only after this validation was the WVSS-2 designated as the primary reference.
2. The chilled-mirror is limited to a warming rate of $1\,\mathrm{K\,s^{-1}}$, which causes asymmetric ramping artefacts in heterogeneous conditions. This limitation specifically affects the cloud-base segments that are scientifically important in MAESTRO (Fig. B2). Applying post-processing inversion techniques to correct for thermal inertia would likely introduce additional uncertainty, particularly when combined with the downstream calibration of fast sensors. Using the 1011C as reference in heterogeneous conditions would therefore propagate a systematic slow-response artefact into the calibrated FAST-WAVE record.
3. The capacitive HUMAERO overestimates mixing ratio by more than 5% in 56.3% of all segments. This systematic bias is inconsistent across segment types and cannot be attributed to limitations of the reference instrument. It is therefore excluded a priori.

The $R^2 > 0.9$ criterion is therefore not used to choose between reference sensors: the WVSS-2 is designated as the reference on independent grounds. The R^2 serves solely as a per-segment quality filter, flagging cases where the low-frequency variability is insufficient for a reliable regression.

Line 323-326: I cannot understand the content of these sentences without further explanation. What do you mean by "resampled," for example? Would it help to have another illustration in which you explain the method using an example?

The paragraph has been rewritten and is placed in Section 3.2.2, when explaining spectral slope calculations. "Resampled" referred to equal-spacing of wavenumber-spectral density points on a logarithmic scale to perform the linear regression, which is now described with less details as follows:

> "a linear regression in log-log space is performed over the inertial subrange, defined from the wavenumber corresponding to the integral length scale up to the Nyquist wavenumber."

Section 3.3.3: Temperature

In principle, the same comments regarding the calibration of humidity sensors also apply here for temperature. However, there is another important point to consider: especially at the cloud base, droplet impaction could occur, or are you flying below the cloud base so that this can be ruled out?

Line 340: What evidence do you have to support the advantages of the 5-micron sensor? These are very vague statements.

Line 350/351: Since a properly performed calibration should increase the performance and reliability of a sensor, I cannot understand this statement. And why does the "de-iced" sensor not require calibration?

The temperature calibration section has been removed. The fourth item of the remarks at the end of the revised instrumentation section now justify the non-deiced Rosemount sensor choice for the dataset:

> "The non-deiced Rosemount E102AL was selected as the temperature sensor for the MAESTRO turbulence dataset. Fine-wire probes, despite their faster intrinsic response time, are highly sensitive to cloud droplet impaction: the less protective housing means that droplet collisions are more likely, rendering the measurement less reliable in cloudy conditions (Lenschow, 1986). With approximately 38% of all segments being B-type legs, where cloud droplet encounters are likely, this sensitivity represents a significant limitation for this dataset. The deiced Rosemount sensor avoids droplet impaction through resistive heating, but the heating element introduces a small systematic warm bias that degrades the accuracy of temperature fluctuation measurements in the clear-air conditions that constitute the majority of MAESTRO legs. The non-deiced Rosemount combines a sufficient acquisition frequency and response time for 25 Hz turbulence measurements with robustness across the full range of MAESTRO conditions, and was therefore retained as the reference temperature sensor."

Section 3.3.4: 3D Wind

In the chapter on determining the offsets for the slide and angle of attack, you wrote that once these have been determined, you assume constant values for the entire measurement campaign. Furthermore, these are probably installation errors or influences on the flow around the radome or similar. Why is the matter not settled with that? I don't understand the physical background as to why statistics are now being presented for uncorrected measurements in comparison to the corrected measurements — or have I misunderstood something? What exactly can I learn from Fig. 12?

In principle, a single calibration is applied to account for installation errors and influence of the flow around the sensors each time the airplane instrumental setup is modified. Former Figure 12 showed histograms of the variance ratio (corrected / uncorrected) for each wind component and segment type. Its purpose was twofold: (i) to document that the correction does indeed affect turbulence statistics in a non-negligible way (variance changes exceeding 20% were observed for some segments) and (ii) to demonstrate that the effect is physically consistent: the correction redistributes variance between the two horizontal components depending on the wind direction relative to the aircraft heading, while leaving the vertical wind variance and TKE nearly unchanged. This is expected from the geometry of the angle biases (sideslip primarily rotates the horizontal wind vector). We acknowledge, however, that this figure was primarily of diagnostic interest rather than directly actionable for dataset users. Following the Reviewer 2's comment and in the interest of conciseness, this figure has been removed from the main manuscript.

Fig. 11 caption: should be "temperature" instead of "humidity".

Figure 11 was removed.

Line 365/366: Which theoretical predictions of Kolmogorov's classical theory did you examine besides the spectral slope of $-5/3$?

Only the exponent of the spectrum in the inertial range was tested. In the revised manuscript, the discussion is limited to the spectral slope as a dataset parameters. In the revised manuscript, the spectral analysis has been removed. The spectral slopes are still extracted and included in the dataset. Controlling high frequency behavior of fast sensors is still required to be able to reliably document fine scale dynamics. This However, they are now presented as physically informative parameters. The primary basis for sensor selection is the independent intercomparison of reference sensors (Appendix B) and the calibration R^2 reflecting low-frequency amplitude agreement.

The text in Section 3.1.3 now explicitly states that deviations of the spectral slope from $-5/3$ carry physical information and may be indicative of buoyancy- or shear-dominated regimes, rather than indicating sensor malfunction. For the specific purpose of TKE dissipation rate computation, a slope close to $-5/3$ for the spectrum of the vertical wind is required for the underlying assumptions to hold.

The spectral comparison figures for individual sensors have been removed from the sensor-selection discussion. In the instrumentation Section 2.3, the high frequency flaws of the LI-7500 instrument is explicitly stated:

> "The LI-7500 exhibited a persistent narrowband spectral artifact peak (from 6 to 12.5 Hz) attributed to potential vibrations of the antenna, which strongly contaminates the high-frequency part of the humidity spectrum; therefore, this instrument is not retained in the final turbulence dataset."

For the sake of transparency, we provide an example spectrum here, that clearly shows this behavior.

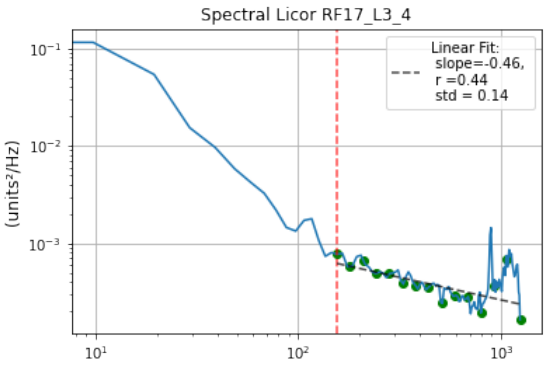

Figure 4: Typical power spectrum of the LI-7500 signal during the MAESTRO campaign

Section 3.3.7: Systematic error

Line 400/401: I don't understand this sentence; why is information lost during high-pass filtering? Of course, the choice of method for defining the fluctuations is somewhat arbitrary, but no information is lost in this context. It is also unclear to me why determining the fluctuations by subtracting the linear trends is used as a reference, so to speak, and why equation 7 can be used to determine or estimate a systematic error.

We agree that “information is lost” is imprecise. The revised text now reads:

> “The definition of the mean field directly influences the magnitude of the fluctuations, and the choice of a 5-km high-pass cutoff removes variance at scales larger than 5 km.”

The linear detrending method is used as a reference because it represents the minimal, assumption-free removal of the mean, retaining all variance at scales larger than the segment length, and is straightforward for any user to replicate. The quantity ϵ_f measures the sensitivity of the statistics to the choice of mean-removal method.

Section 4: Data Summary

The summary in Section 4, and especially Fig. 16, is of course highly simplified when averaging all 24 flights and showing the subdivision "only" for the leg types. This can be done, but in my opinion, the benefit is rather limited, as can be seen from the conclusions drawn from Fig. 16. It is not a new finding to emphasize that mechanical turbulence is strongest near the ground. Somewhat more surprising is the higher variance of the thermodynamic parameters at the cloud base; but why should the strongest gradients of T and q be found there? A cloud base is not necessarily the place where one would classically expect to find an inversion (cf. line 421).

In my experience, the arguments described in section lines 419–423 are more characteristic of cloud tops than cloud bases. For example, a temperature inversion at the cloud base would prevent convection and thus make cloud development rather impossible. Of course, there are also decoupled cloud layers, but if that is what is meant here, it should also be substantiated by observations. The term "entrainment" is also more associated with the cloud top than with the cloud base.

Former Figure 16 was not averaged over all flights but showing distributions corresponding to most values of the dataset (more than 500 samples). This presentation allows the transparent showcase of the range and variability of the most commonly used variables extracted for the dataset.

The original sentence referred to horizontal gradients of temperature and humidity along the flight track. The enhanced variance reflects the along-track heterogeneity of a cloud-base region. We acknowledge that the original phrasing could be misread as implying a vertical temperature inversion. The text has been revised to make clear that the enhanced T and q variance in B-type segments results from horizontal variability encountered as the aircraft samples a patchy cloud-base environment, including intermittent in-cloud and clear-sky conditions, which produce local fluctuations in temperature and humidity through condensation and evaporation. To avoid misinterpretation the sentences were changed to:

> "For temperature and water vapor mixing ratio, cloud-base segments (B-type) are associated with the highest variance values. B-type legs are horizontal transects flown at cloud-base level. Two possible contributing factors can be identified: (1) the aircraft may alternately encounter clear air and cloud interior during partially cloudy segments, generating along-track horizontal gradients in both T and q due to the contrasting thermodynamic properties of these two environments, and (2) even in segments that are fully within the clouds, the cloud interior itself is a region of strong thermodynamic heterogeneity, where active condensation, evaporation, and mixing processes produce large local fluctuations in temperature and humidity."

The energy dissipation rates shown in Fig. 17 are really very low, and I have serious doubts that values below 10^{-6} m^2 s^{-3} can be statistically significantly resolved. The spectral noise floor should be determined (see Muschinski et al., Boundary-Layer Meteorology 98:219–250, 2001).

If my assessment is wrong, that's not a problem, but then a thorough analysis of the data is

required. In any case, the spectral noise floor of the wind measurements should be determined in order to be able to estimate a meaningful resolution for the dissipation rate (see, for example, Muschinski et al. in Boundary-Layer Meteorology 98: 219–250, 2001).

We thank the reviewer for pointing to the Muschinski et al. (2001) reference. We considered the methodology employed in this study and addressed this issue accordingly as was explained earlier.

We hope that the revisions and the detailed responses above fully address the concerns raised by both reviewers.

Sincerely,
Louis Jaffeux, on behalf of all co-authors.