AC#1

We thank the reviewer for his careful reading of the article. His constructive comments should undoubtedly contribute to improving the paper. As suggested, we have added a "discussion" section to include some details on the comparison between the turbulence indexes, and also to show the inhomogeneity of the turbulence detections according to the position of the balloons.

Follows a point-by-point response to the reviewer's remarks and comments.

Main comments

3.1 **Similarity of results from correlation method and Richardson method.**

The reviewer's comment is quite relevant. The fact that the diagnoses of the flow state, laminar or turbulent, are identical in 97 or 98% of the cases does not mean that the diagnoses of the turbulent cases alone are identical at that level. In fact, the detections are most consistent when the vertical stratification is high. In such cases, there is very little disagreement between the methods because the correlation levels, or Richardson numbers, are high.

However, the situation is quite different if we consider only the diagnoses of turbulent flows (between 3 and 6% of cases in the average). For these cases, the differences can reach a factor of two. Thus, for flight 7 (07_STR2), the percentages of detection of turbulent sequences vary from 3.3% (Ri_TSF) to 6.3% (r_P). We believe that these differences result mainly from the fact that the threshold values, zero correlation or Ri = 0.25, correspond to the tails of the distributions of these estimates (Figs. 9 and 13 of the paper). When the atmosphere is weakly stratified, threshold effects are likely to be important, leading to important differences in the diagnosis of the flow conditions. Also, the differences between the Ri and correlation methods can be partly due to the thresholds values of the hypothesis tests of a null correlation (i.e. choice of a confidence interval for the null correlation).

As suggested by the referee, we compared the detections by the 4 methods taking Ri_{TSF} as a reference indicator. We have considered the four possibilities for each of the three estimators: correct-turbulent and false-turbulent, correct-laminar and false-laminar.

Table 1: Percentages of true (identical) and false detection of the various methods compared to the Ri_LSF method.

Ri_{TSF}	Turbulent		Laminar	
	True	False	True	False
Ri_{LSF}	85.4	14.6	99.2	0.8
Pearson Corr	76	24	99.1	0.9
Spearman Corr	83.2	16.8	99.5	0.5

The table shows the percentages of correct (i.e. identical) and incorrect diagnoses by the Ri_{LSF} and correlation methods compared to the Ri_{TSF} detections. The bar chart below shows the same thing in graphic form. It can be seen that the diagnoses are identical in more than 99% of the cases if the flow is detected as laminar. On the other hand, the diagnosis are identical for about 80% of the cases if the flow is detected as turbulent. We attribute these poorer performances to the fact that the critical thresholds, Ri = 0.25 and correlation = 0, belong to the tails of the distributions of the statistics and that the edge effects are more important for these rare events.

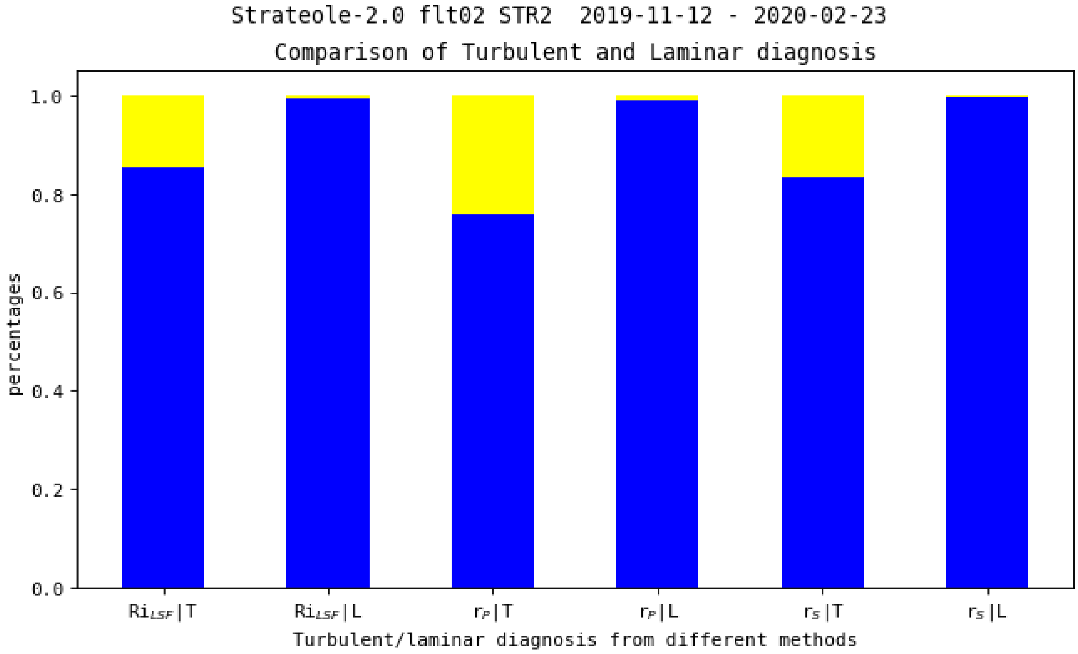

Figure 1: Percentages of true and false detections for the turbulent (T) and lamimar (L) episodes compared to the Ri_{TSF} *method. The three methods* Ri_{LSF}, r_P, rS *are compared.*

We have added a paragraph and a figure in the article to clarify this fact.

3.2 **High occurrence rate of negative Richardson number**

Ri and N^2 time series for flight 2 (02_STR2) are shown in figure 11 and 13 . The probability for Ri (N^2) to be negative is not zero since occurrences of negative values are visible in the time series. For the considered flight, the occurrence frequency of negative N^2 is 3.4% (from the Theil-Sen regression performed on T_C and Z_P). Such negative Ri (N^2) can result from both the dispersion of the temperature gradients estimates or the occurrences of episodes of unstable stratification. Note that we have corrected the histogram of N^2 in Figure 11. They were not plotted correctly in the original version of the paper since only positive classes were defined. Negative occurrences are now visible.

In the present study, negative estimates of Ri (or N^2) may be due to the precision of the estimates of the temperature gradients (scattered around a value close to -10°/km in case of quasi neutral stratification). Temperature gradients at the balloon flight level are estimated from the covariance of increments of temperature and displacements, these covariances being computed over one-hour time segments. Assuming neutral stratification, the covariances are expected to be scattered around 0, implying some negative estimates of N^2 .

However, unstable stratifications ($N^2<0$) seem to occur in the lower stratosphere since they have been reported in the literature. For instance, detection of turbulence by the Thorpe method from insitu measurements is based on observations of, $\partial\theta/\partial z<0$, i.e. $N^2<0$ (Thorpe, 1977). The probability of occurrence of such unstable layers likely depends on the vertical resolution of the profiles (see for instance Wilson et al., 2011, Geller et al., 2021) but it not zero. It is exact that for Kelvin-Helmholtz instabilities, turbulence is expected to be triggered for $0<Ri<1/4$, i.e. for $N^2>0$, but once it is developed, the stratification can become almost neutral ($N^2\approx 0$), or even unstable ($N^2<0$), as a result of stirring and mixing. Therefore, it is plausible that the occurrences of such unstable episodes may also contribute to negative values for Ri (N^2) based on covariances calculated on one-hour time segments.

Estimates of Ri, or N^2, from radiosondes, when applying the Thorpe's method, are made from the sorted potential temperature profiles – anywhere increasing with altitude - and therefore they cannot be negative. However, the measured profiles show decreasing potential temperature with altitude in some places (i.e. the stratification is unstable and $N^2<0$). This is at the base of the Thorpe detection method.

3.3 **Possible influences by warm downwash from the balloon**

The referee's remark is quite relevant. Indeed, due to the vertical oscillations of the balloon, the T-sensors are possibly in the wakes of the balloon or of the flight chain. Notice that we expect the balloon wake to be warm during daytime and cold during nighttime, the balloons being cooler than the ambient air during nighttime.

The diameter of the balloons is either 11 m (TTL) or 13 m (STR). The temperature sensors are located 27 m below the balloon base (except for TTL3 flight) and 15 m below the EUROS gondola. On all but TTL3 flights, the T sensors are located 7 m below the last gondola in the flight chain. Flight 03_TTL3, carrying the RACHuTS system, is an exception since the temperature sensors are located 30 cm away from the EUROS gondola.

The probability of the T-sensors being in the wake of the balloon or gondolas is clearly non-zero. If there is no horizontal wind shear, the T-sensors should enter the wake of the balloon as soon as they enter the area in which the balloon is oscillating (about 30 m wide). Taking into account the distance between the balloon and the T-sensors (27 m), the T-sensors can enter the balloon’s wake only if the amplitude of the balloon oscillations is larger than ~13.5 m (27 m peak-to-peak), that is for slightly more than 50% of the time (the median value for amplitudes is 15 m). Anyway, the T measurements could still be perturbed by the wake of the flight chain (gondola(s), parachute, wires). The only case where the T-sensors should not enter the wakes is when the wind shear is sufficiently large (about 5 m/s/km).

The issue of the possible impact of wakes was considered during this study. Indeed we calculated the statistics, vertical gradients and correlations, considering only the phases when the balloon descends. During these phases, the temperature sensors (which are located at the lower end of the flight chain) sample the "fresh" air if a minimum shear exists. The figure below shows the time series of Pearson/Spearman correlations for the flight presented in the article (Fig. 9) but considering only the phases when the balloon descends. The resulting time series are noisier since we only consider about half of the samples. However, both time series of correlations and temperature gradients have similar characteristics to those calculated when considering all samples, showing the same succession of stable and unstable periods. We therefore conclude that the impact of the wake during the ascending motions does not affect significantly the estimated correlations and covariances and because of the increase of noise we choose to consider all the samples. We did not mention it in the initial version of the article because we did not observe an important impact. There is now a paragraph in the “discussion” section about that point.

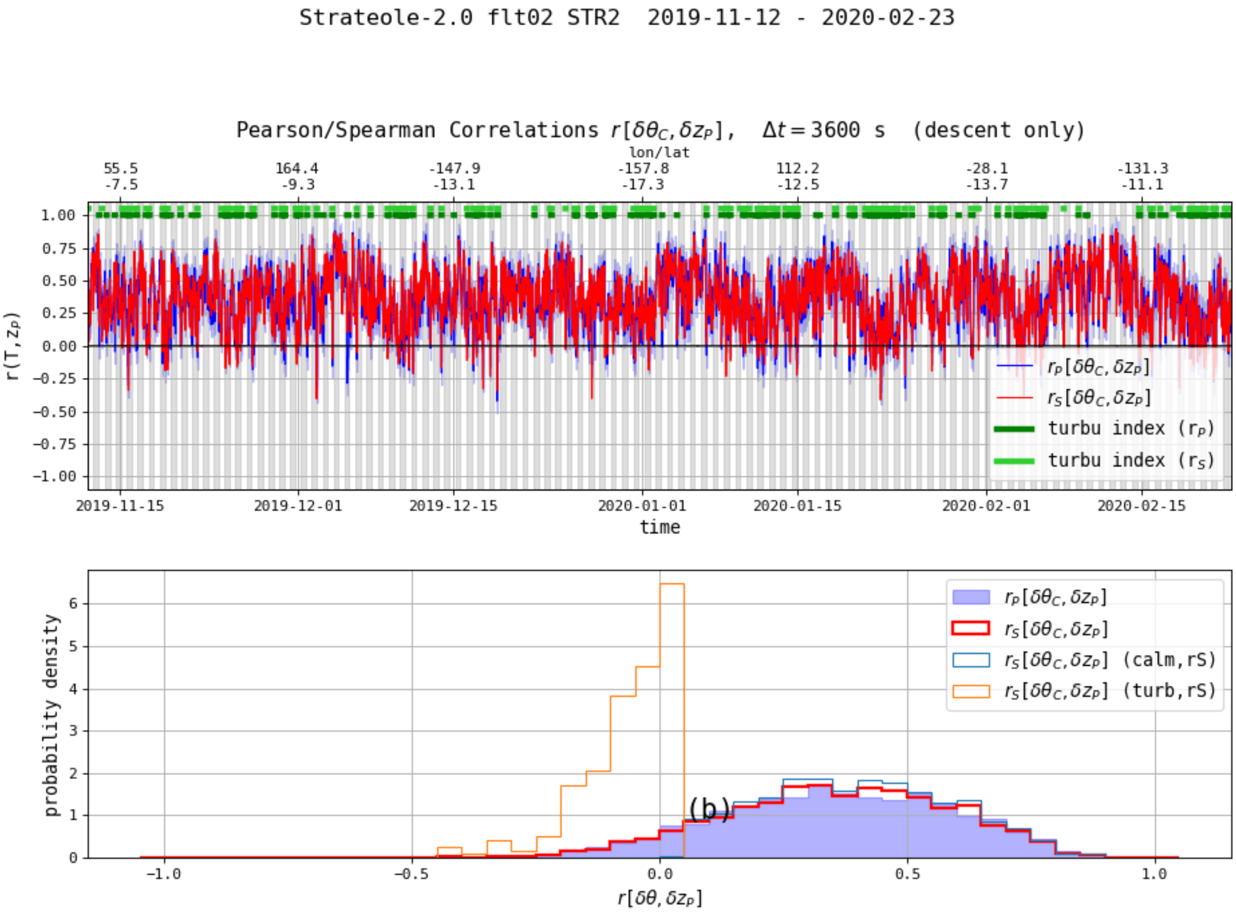

Figure 2: Time series and histograms of correlations but considering only the descending phases of the SPB.

3.4 **Nature of turbulence than can be detected by the correlation method**

As the reviewer recalls, numerical simulations indicate that large temperature gradients are expected at the edges of turbulent layers (Fritts et al., 2003; Werne and Fritts, 1999). These strong gradients are also commonly observed from radiosonde profiles when turbulence detection is performed by the Thorpe method. It is clear that the sampling by the balloons, drifting within an air mass and not cutting vertically through it as a radiosonde, will not allow to identify such temperature gradients at the edge of turbulent regions. Only the central part of the turbulent region, in which stratification is almost zero, can be detected. We added a few sentences in the document to clarify this fact.

Minor comments

We warmly thank the reviewer for his suggestions (which we all followed) and for pointing out the typos, which were corrected.

AC#2

We thank the referee for its constructive remarks which should undoubtedly contribute to improve the present paper.

Follows a point-by-point response to the referee's remarks.

Major remarks

1. The sampling period of measurements is 30 s, i.e. $f_{Nyquist}=1/60$ Hz. The oscillation periods of the balloon vary around 220 s, i.e. $f_{NBO}\approx f_{Nyquist}/3.5$. We found that the variability of measured quantities in this narrow frequency domain is dramatically affected by balloon oscillations (see the *w* spectrum of Fig.6). This conclusion is given in lines 200-204 and 401-404. In fact, our first attempt to detect turbulence was to look for an excess of variance or a spectral signature associated with fluctuations of the measured parameters in the $f_{NBO}-F_{Nyquist}$ frequency range. This method was unsuccessful. The main reason, we think, is that the high-frequency fluctuations are dominated by the balloon's natural oscillations. Therefore, we turned to statistical methods to detect the effects of turbulence from sensors that do not allow to directly measure turbulence.

2. As noted by the reviewer, the purpose of this paper is not to describe geophysical results. We indicated however that the detection is not uniform around the globe but seems to depend on the position. The following figure showing the occurrence frequency of the turbulence detection supports this statement. The variability according to the position appears to be very large, ranging from zero to about 25%. However, we think that the interpretation of these results is beyond the scope of the article and we do not think to publish it here. A PhD thesis by one of the co-authors (CP) is in progress on the geophysical exploitation of strateole-2 data.

We furthermore think that the turbulence we observe in the stratosphere is primarily due to wave breaking. As commonly accepted, deep convection is a major source of waves in the tropics, and those waves propagate vertically in the atmosphere. We therefore do not expect much difference between the occurrences of turbulence in the TTL and STR balloon flights: those are only 2 km apart. On the other hand, we expect a significant longitudinal modulation.

From Table 5 indicating the fraction of time during which the flow is turbulent does not show any clear contrast between the TTL and STR balloons, but it is not conclusive as only two TTL balloons are reliable (excluding flight 3, very noisy).

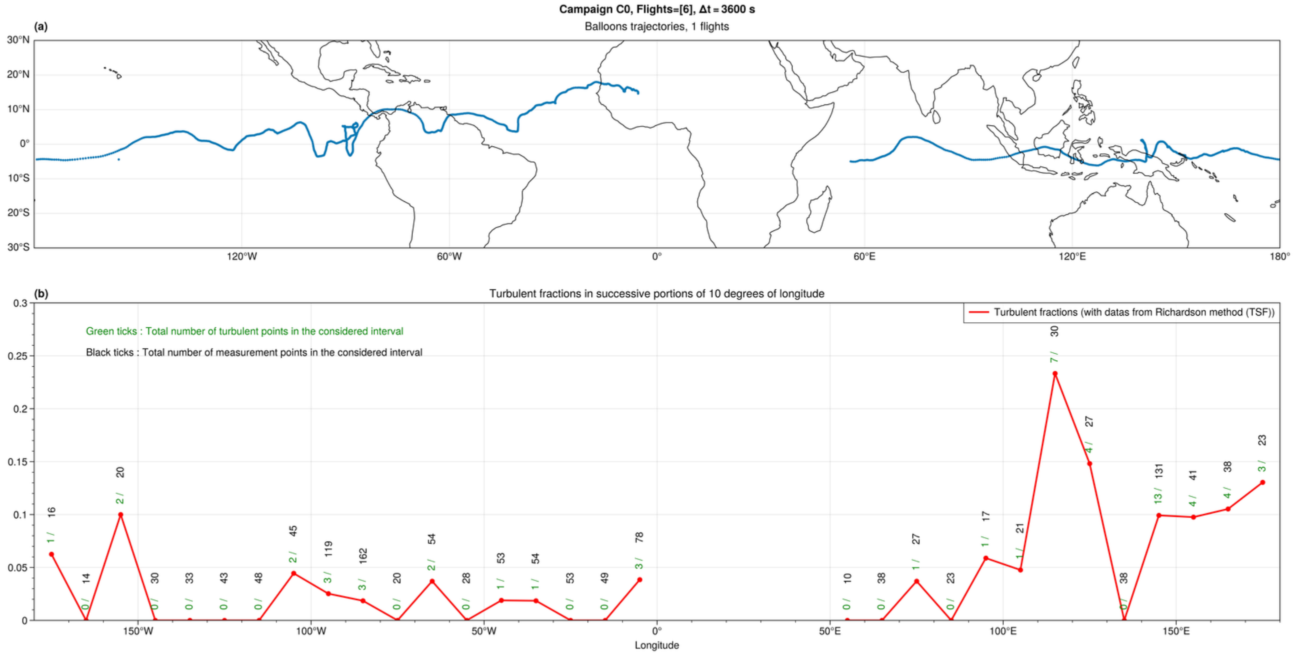

Figure a: Trajectory of flight 6 (STR1) of the C0 campaign (top) and percentage of turbulent detection by the Ri method (bottom).

3. We thank the reviewer for pointing out the (too) many typos for which we apologize. The mentioned typos – and others - have been corrected.

Minor remarks

- **line 10**: The mode of the distribution of the vertical displacements associated with the oscillatory motions of the SPBs is ~ 15 m (see figure 4).

- **line 30**: Gravity waves transport vertically energy and momentum. As long as waves do not saturate, there is no impact on the dynamics or chemistry of the atmosphere. Heat, momentum and minor-constituent fluxes occur when wave saturation occurs, i.e. when waves break into turbulence.

- **line 75**: We wrote "under high pressure balloons because the measurements are carried out on board of a gondola and temperature sensors located under the balloons.

- **line 115**: we downgrade the RACHuTS measurement to 30 m 1) to improve the measurement precision, and 2) in order to reach similar vertical resolution for TSEN and RACHuTS, TSEN estimates being evaluated in layers of about 30 m thick (+/- 15 m) -.

- **line 125**: The instrumental noise level is quantified by the standard deviation of the high frequency fluctuations. This standard deviation is estimated from the variance of the nth order increments as described in Appendix 1. Because both the balloon motions and the instrumental noise contribute to the high-frequency variability, we consider only the smallest 10% of estimates to minimize the influence of the balloon motions.

- **Eq 1**: $\pi \approx 3.14159\ldots$

- **line 195**: The referee's remark is quite right. Most large amplitude oscillations are associated with depressurization events. We have added a sentence to clarify this fact:"A few large amplitude oscillations (>100 m) are observed. We found that they are most often associated with depressurization events. ".

We did not handle balloon depressurization events because they do not appear to have impact on turbulence detection (we do not detect turbulent events during depressurization events).

- **line 220**: Since the horizontal wind is typically of the order of 5 m/s, the horizontal spatial scale is typically 18 km. But this horizontal scale does not correspond to the spatial extension of the observed episodes (turbulent or laminar) since the balloon is advected by the flow, i.e. the observations are Lagrangian. Lagrangian observations of turbulent or laminar events are rather related to the lifetime of the events.
- **Fig 5**: The red line at t_1 indicates the potential temperature level that the balloon will reach at t_1+30 s. It is drawn to indicate schematically that the whole layer is vertically displaced in the 30 s time interval. Therefore, the observed θ changes are not simply proportional to the vertical θ gradient.
- **Fig 6**: The two spectra in Fig. 6 are representative of the power spectra of *w* for all flights: almost flat up to the frequencies corresponding to the oscillatory motions of the balloons. Similar power spectra have already been published (Podglagen et al.).
- **line 270**: Thanks to the reviewer for pointing out the (too) many typos. We have corrected those mentioned, and others.

Fig 7: The colors of lines of the plot have been corrected

line 332: “the time series” → ‘parameters’

line 374: The probabilities are indicated as the y-axis of the cumulative distribution function (bottom right). The percentages are deduced.

line 387: The differences about the turbulence detections from the different methods are discussed in the paragraph “Discussion”, section 5.1 and are shown in the new figure 14.

The detection methods are based on statistics (two correlations and two regressions) applied on independent measurements. The percentages of turbulence differ by using different estimators or different parameters: the percentage differences can reach a factor of two (table 5, flight 07_STR2, 6.3% vs. 3.3%). These differences can be partly due to the thresholds values of the hypothesis tests of a null correlation (choice of a confidence interval). Likely more important is the fact that the thresholds values (null correlations or $Ri=0.25$) are associated with the tails of the distributions of these statistics (see the histograms of figures 9 and 13). The differences occur mainly when the atmosphere is close to a neutral stratification, i.e. when the estimates are close to the thresholds. When the atmosphere is somewhat stratified (about 95% of the time) the statistical estimators are very consistent, the agreement between the estimates being larger than 99% (Fig. 14 of the paper).

The following figure shows the turbulent periods detected by the four different methods for flight two. As stated in the paper, we privileged the combination of T_C and z_P as they appear to be less noisy than T_S and z_{GPS}. Overall, the agreement is good, even differences are visible.

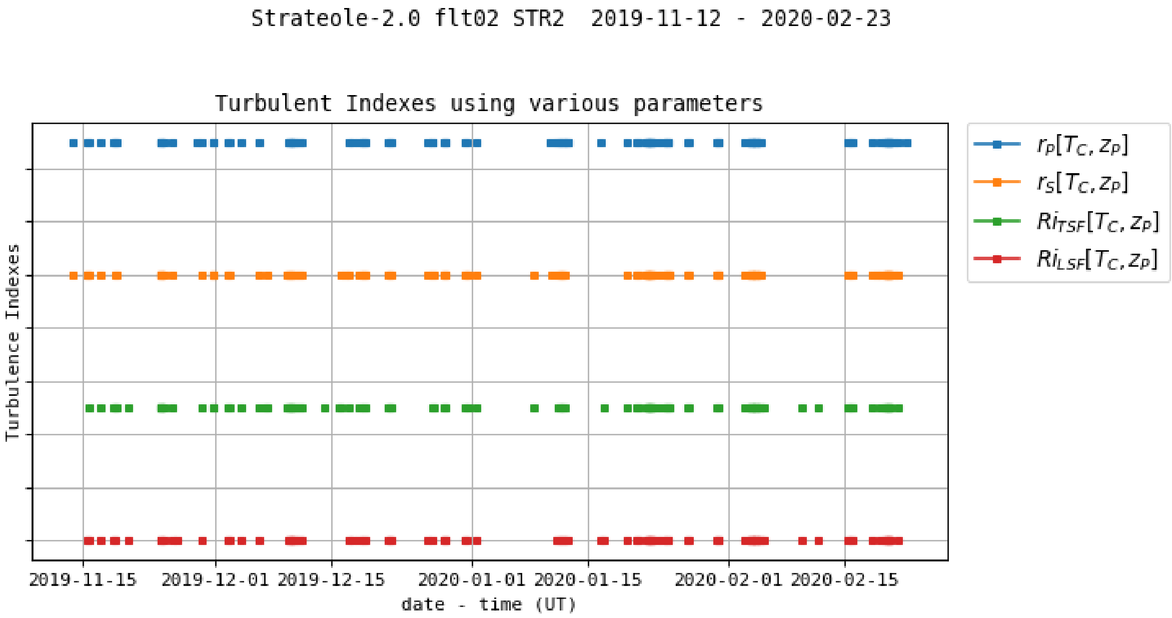

Figure b: Time series of the turbulence detection from four estimators.

In short, the used statistics (correlations and Ri) allow to describe the state of stratification of the atmosphere in which the balloons float. When the atmosphere is close to a neutral state, all estimates, correlation or *Ri*, reach values belonging to the tail of their distributions, close to threshold values. Some differences in detecting the turbulent episodes occur due to threshold effects. We added a subsection about the comparison of the different indexes (.

Appendix A: The two typos in the equations have been corrected.

AC#3

We thank the reviewer for his careful reading of the article. His constructive comments should undoubtedly contribute to improving the paper.

Follows a point-by-point response to the referee's remarks.

- **The referee states that the title of the paper is misleading because we do not present turbulence data.** It is correct that we do not present turbulence measurements as the oscillating motions of balloons does not allow direct observations of turbulence. However, the paper presents methods for detecting turbulence based on estimates of the local stratification. The underlying assumption of such an approach is similar to that of Thorpe's detection (Thorpe, 1977) or to detection methods based on the Ri criterion: in case of unstable flow, i.e. when $Ri < ¼$ or $\theta_z \leq 0$, the flow is assumed to be turbulent. It is true that these conditions may be the cause of turbulence - in the initial phase of an instability - before turbulence develops. However, the statistics on which stratification is estimated, correlations or covariances, are calculated over one-hour time segments. It appears very unlikely that an unstable flow could persist for such a time without generating turbulence.. Therefore, we believe that the detection of neutral or even unstable stratification is, or has been recently, associated with turbulent episodes. We therefore think that the current title is appropriate.

- **About the lack of other evidence of turbulence**: our first attempt to detect turbulence was to look for high-frequency excesses of variance or a high-frequency spectral signature of the measured fluctuations. This method was unsuccessful. We believe this is because the high frequency variance is dominated by the natural oscillations of the balloon. The sampling period of measurements is 30 s, i.e. $f_{Nyquist} = 1/60$ Hz. The oscillation periods of the balloon vary around 220 s, i.e. $f_{NBO} \approx f_{Ny}/3.5$. We found that the variability of measured quantities in this narrow frequency domain is dramatically affected by balloon oscillations (see the *w* spectrum of Fig.6). This point is discussed in the paper (lines 199-203 and 401-404). Therefore, we turned to statistical methods to detect the effects of turbulence from sensors that do not measure turbulence directly.

  Despite the fact that we cannot directly measure the turbulence intensity from measurements under SP balloons, we show that the vertical temperature gradients estimated by statistical methods from TSEN measurements are consistent with the RACHuTS measurements. Also, observing $\omega_{NBO} < \omega_B < N$, as modeled by Podglagen et al. (2016, see the supplementary information) gives confidence in the estimates of the vertical gradients since ω_{NBO} and *N* depend on them (Eqs. 1 and 2). Therefore, we are confident about the estimates of vertical gradients and correlations revealing periods with little or no stratification of the flow.

- **About the lack of relationship between the times and places where turbulence occurs and the synoptic conditions**: the objective of the paper is clearly methodological and we did not investigate for the causes of turbulence occurrences. We have only mentioned the fact that the frequency of occurrence of turbulence is about 5% in the average and is not uniform around the globe. It appears to be greater over convective areas such as the maritime continent. Such an assertion is justified by the following figure which shows the positions of the turbulence detections (Ri method) as green dots for the 8 flights of the C0 campaign. At the present stage, this result is qualitative and need to be deepened. Following the suggestion of the reviewer we have included this figure in the new version of the paper. The exploitation of the Strateole-2 data set with the presented methods is the subject the PhD thesis of one of the co-authors (C.P.).

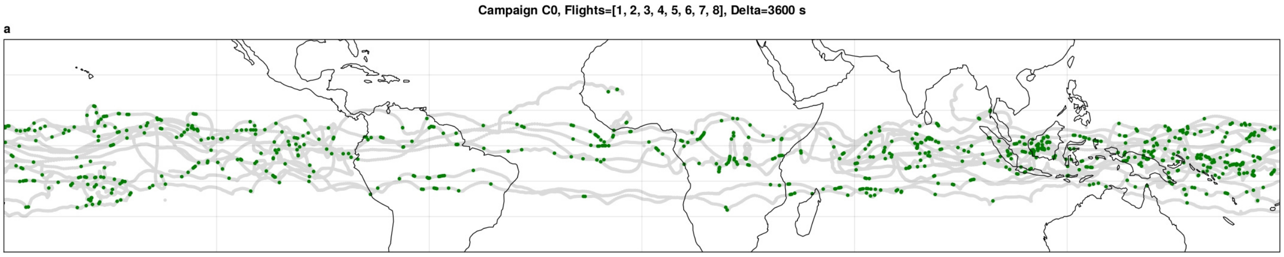

Fig. 3.1 Positions of the turbulence detection for the height flights of the C0 campaign.

- **About the assumption that layers with neutral stratification are turbulent**: There is no doubt that instability precedes the development of turbulence. We agree that this instability condition (Ri < 1/4) may be due to dynamic processes such as waves or shear, and not to turbulence. But such instability cannot persist and will necessarily generate turbulence. Therefore, if unstable, or neutral, stratification is observed, either turbulence is developed or it will develop. Since the statistics are based on one-hour segments, we favor the first hypothesis.

  We modified the sentence of line 205 to precise this point: “A consequence of turbulence is to restore stability from a preceding unstable state of the flow, which is achieved by locally mixing the fluid” and added a sentence: : “Note that neutral or even unstable stratification conditions may precede turbulence, but such conditions cannot persist and will cause turbulence.”

- **About the impact of turbulence on vertical transport**: the question of the impact of turbulence on the vertical transport is the major motivation for turbulence study. This does not imply that we can answer this question with this article, which is only methodological. An estimate of vertical transport will have to rely on a physical model constrained by observations. The fact that we detect turbulence in the average for about 5% of the time does not allow us to quantify the vertical transport. The question of the impact of turbulence on transport is the central question of the PhD thesis of C.P..

- **About the turbulent fraction of the atmosphere**: the turbulent fraction in the UTLS is very poorly known. The few estimates from radiosondes or instrumented aircraft suggest values between 0 (no detection at all) and about 5%. It is difficult to compare these values because they are based on instruments with different sensitivities.

  Beyond an average value for the tropical UTLS, we found that the spatial distribution of turbulent events around the globe is far from uniform. The differences are very large since the turbulent fraction can be higher than 20% in some regions and almost zero in others. Interestingly, these inhomogeneities provide information on the mechanisms behind these turbulent episodes.

  In addition, the Lagrangian observations enabled by SPBs provide information on the lifetime of turbulent events. This parameter is crucial in some models (Alisse & Haynes) to quantify the turbulent transport.

- **line 60**: Indeed, there are many studies of turbulence from aircraft measurements. We added a reference to a recent work (Dörnbrack et al. 2022) with numerous references. However, to our knowledge, these studies do not involve tropical UTLS (with the exception of Podglajen et al., 2017, who used data obtained above &the ceiling of most research aircraft). Note that the current lack of turbulence measurements in that region is a major motivation for the present research (lines 33-35).

- **Table 1**: All balloons carry TSEN sensors. This has been specified in the legend of Table 1.
- **line 115**: Following the remark of the reviewer, the sentence has been modified: "… which we degraded to about 30 m in order to (1) improve the raw 1-m accuracy of the altitude measurements and (2) achieve similar vertical resolution from the TSEN and RACHuTS measurements (since the amplitude of the balloon oscillations is typically +/- 15 m)."
- **line 152**: The following figure (Fig. 3.2) shows the relative variations of volume of a Strateole-2 balloon during a two and a half month flight.

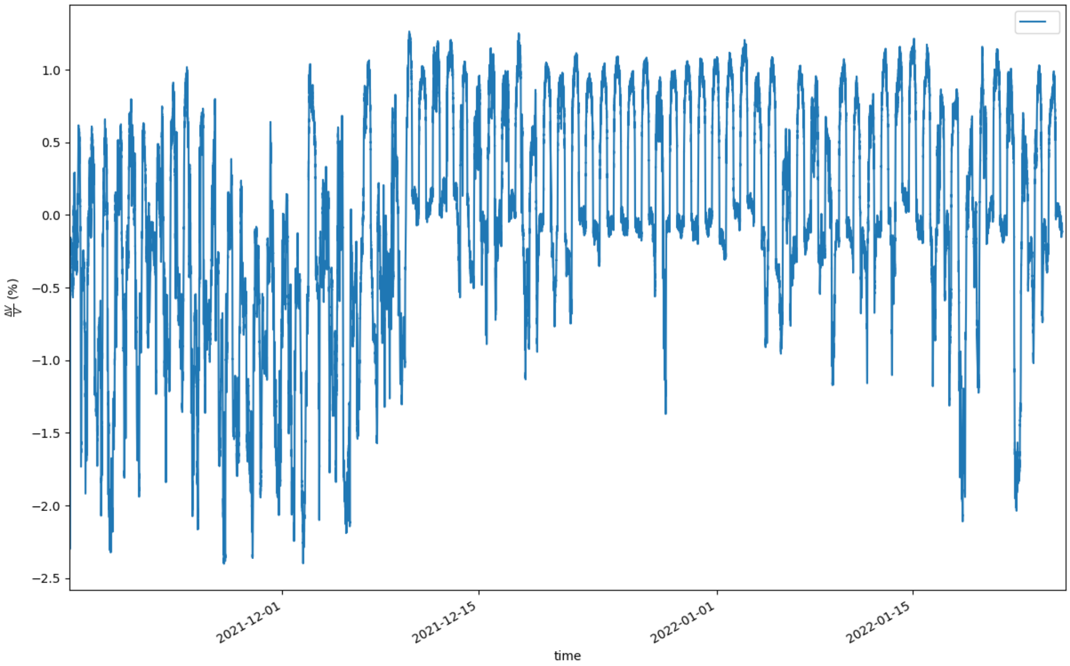

Fig. 3.2 Time series of the relative variations of the balloon volume during a flight.

The volume of the balloon varies by less than 3% peak to peak, the variations of diameter are thus lower than 1%. This precision has been added in the paper. The volume variations are mainly due to the diurnal cycle of the solar heating.

- **Fig. 8**: The large orange polygon extending to the left of -10 K/km represents the histogram of vertical temperature gradients for the only time intervals during which the flow is detected as turbulent. The green histogram (the color has been corrected) corresponds to the cases when the flow is stable.
- **line 395**: As suggested, we have removed the word "mainly" in the sentence.
- **line 404**: In the initial phase of this study, we looked for a high frequency signature as a proxy for turbulence, excess of variance or spectral signature. This method proved unsuccessful because the high frequency fluctuations ($N/2\pi < f < 1/60$ Hz) are affected by the natural oscillations of the balloon.

However, it is planned for the next Strateole-2 campaign to perform fast (1s) velocity measurements with a sonic anemometer. This instrument, called VATA, is currently under development. These measurements should allow to directly measure turbulence.

- **line 424**: As suggested, we have included in the paper the plot of the positions of the turbulence detections (Ri method) for the 8 flights of the C0 campaign (Figure 3.1). The study to determine the origins of instabilities is ongoing.