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.

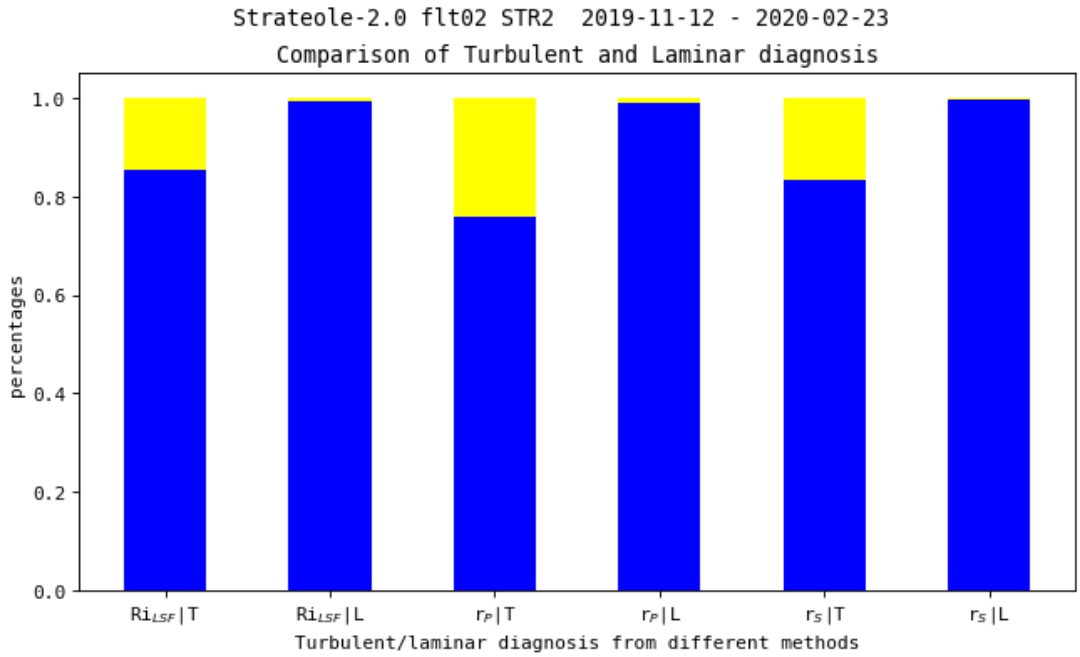

Figure 1: Percentages of true and false detections for the turbulent (T) and lamimar (L) episodes compared to the Ri_{TSF} method. The three methods Ri_{LSF}, r_P, rS are compared.

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 in-situ 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.

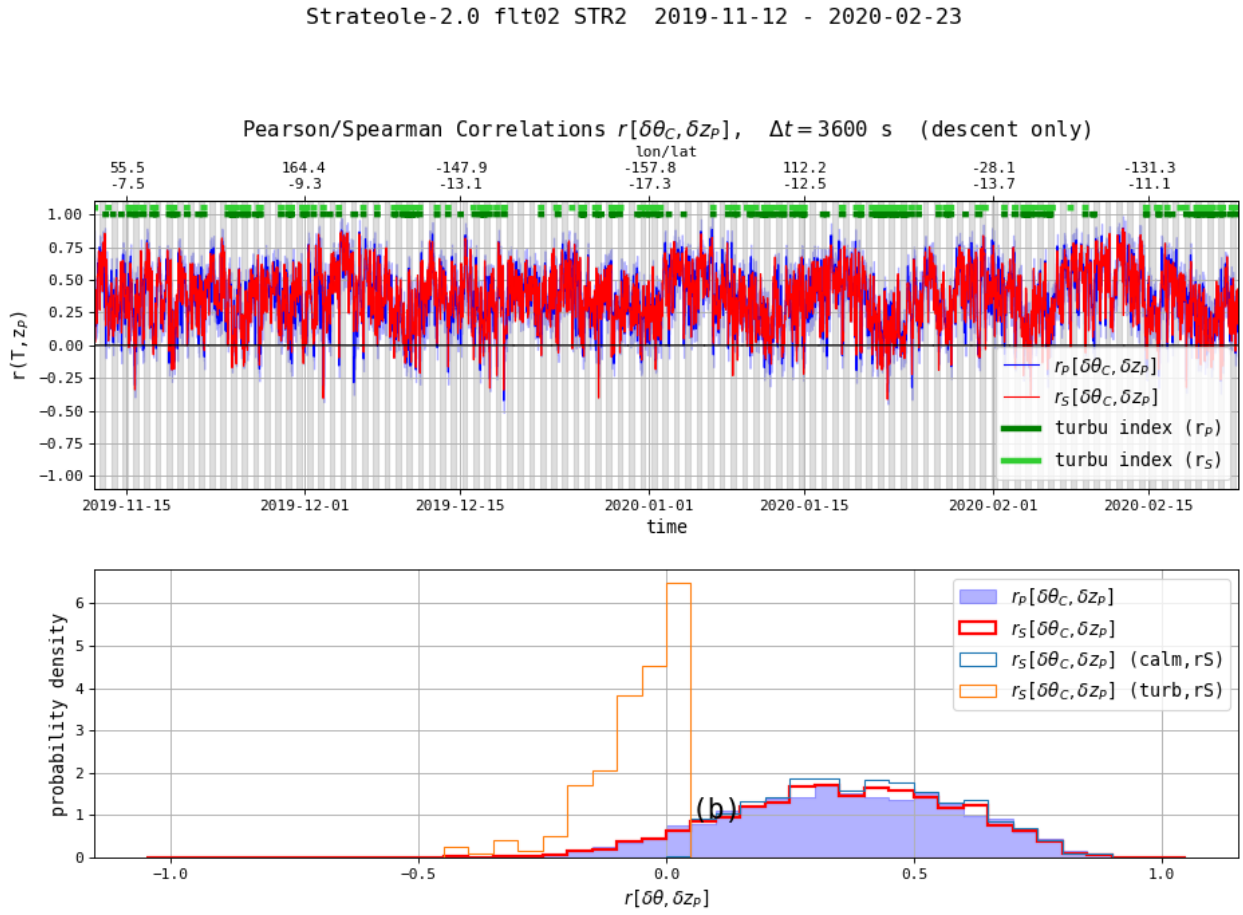

Figure 2: Time series and histograms of correlations but considering only the descending phases of the SPB.

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.