Response to Anonymous Referee #1

We thank the Referee for their insightful comments. We have implemented a number of changes as outlined below.

General Comment: The paper by Wolfe et al. describes airborne eddy covariance measurements on a C-23B Sherpa aircraft. It summarizes results from flights in the eastern US. The general topic is suitable for AMT, but there are a couple of issues that need to be addressed before publication. In particular, the discussion on errors needs revision.

Specific Comments:

It is not clear whether a Webb correction was necessary for CO2 and H2O fluxes, and how this was incorporated in the flux analysis code. The 10Hz humidity correction for the LGR instrument mentioned on page 5 (line 31) seems tricky – since there was a redundancy of humidity measurements, a better experimental setup would have been to use a Nafion dryer for the EC system and just focus on CO2 and CH4 to avoid this problem all together. As the data are treated it is not clear to what extent the water vapor flux influences CH4 and CO2 fluxes, or how the correction procedure would degrade the precision of the flux calculation, given the large random errors of 10 Hz concentration datasets.

We do not perform a Webb correction, as the 10 Hz observations of CO2, CH4, and H2O (both LGR and DLH) are corrected to dry mixing ratios prior to calculation of fluxes. The interpolation of LGR H2O to the CO2 time base is not especially tricky, as 1) the gas sampling systems and native sampling rates are identical, and 2) the CO2-H2O correlation is sufficiently strong to provide good lag-correlation. We have added the following statement in the first paragraph of section 3.1:

Raw gas concentrations are provided as dry mixing ratios, eliminating the need for density corrections (Webb et al., 1980) to derived fluxes.

In our standard procedure, we calculate latent heat fluxes from DLH by first converting from mole fraction (moles per mole moist air, the native DLH measurement) to mixing ratio (moles per mole dry air), which theoretically negates the need for a Webb correction. As a test, we recalculated LE using mole fraction and applied a density correction following Eq. (23) of Webb et al. (1980). The full range of difference between the two methods is ±1%, and the normalized mean bias is 0.12%. The LGR instrument water corrections (Fig. S3) include both dilution and spectroscopic effects, thus it is somewhat more difficult to separate out the dilution component in our analysis code, but we expect a similar result. We further note that other researchers also avoid density corrections through similar approaches (Desjardins et al., 2018).

For the initial CARAFE deployments, we decided that not drying the GHG sample gas was preferable as the comparison of LGR H2O mixing ratios and fluxes against DLH provides a valuable performance cross-check. Also, redundancy is insurance against instrument failure. We will consider drying the sample for future deployments.

According to eq. 12 the turbulent random error should always be smaller than the combined error which includes instrument noise. Inspecting figure 7 actually shows the opposite for most tracers; the relative turbulent error is larger than REFS01 for T, H2O and CO2; this contradicts the theory. An explanation is needed – could there be a calculation error in the analysis code?

The turbulence random error defined in Eq. (11) represents an upper limit (note the ≤ sign). Thus, we expect that the empirical total random errors, RE_{FS01} and RE_{wave}, should be generally smaller than the root-sum-square of RE_{turb} and RE_{noise}. Fig. S9 illustrates this point for RE_{wave}. We have added a plot to Fig. S9 to show a similar correlation for RE_{FS01} and modified Sect. 3.4.2 as follows:

The maximum lag for the summation is set to 10 seconds based on comparison with the root-sum-square of RE_{noise} and RE_{turb}, <u>the latter representing a theoretical upper limit for total random error (Fig. S9a).</u>

Eq. 12 is cast in the time domain. For aircraft measurements the time domain is not really meaningful. The discussion of errors should be handled in the spatial domain. For example, a cut off frequency of 0.02 Hz corresponds to a distance of 3.75 km at the aircraft speed of the C-23B Sherpa. The same criterion would correspond to a 12 km distance on a G5-aircraft. The issue of spatial vs. temporal scale should be treated consistently throughout the manuscript. While the error discussion is treated in the time domain, some figures show a spatial, others a temporal scale. Figures 5 and 9 should be modified to show a spatial scale as well.

We agree that it is more appropriate to cast discussion in the spatial domain when referring to turbulence scales or wavelet-derived fluxes. For certain aspects, however, the time domain is meaningful – specifically, with regards to instrumentation. For example, the characteristic response times described in Sect. 3.4.1 are inherent to each instrument and independent of platform speed. A similar argument holds for the influence of instrument noise on spectra shown in Fig. 9(b). With specific regard to Eqs. (12) and (13), we would obtain the same results (for this particular set of measurements) regardless of whether we use the temporal or spatial domain, as the Sherpa cruise speed is fairly constant for most flux legs (81 ± 9 m/s). The use of temporal domain is mostly a matter of convenience.

We have added the following statement to the beginning of Sect. 3:

The following discussion references both the time and spatial domains as appropriate, the two coordinates being linked by leg-average aircraft speed.

In addition, we have added/modified text throughout Sections 3 and 4 to better address spatial vs temporal scales (see esp. Sect. 3.3), and we have added spatial scales to Figs. 4, 5, and 9.

Total error: Systematic errors inherent to unresolved scales always lead to an underestimation of fluxes and should be used to correct the data rather than adding these to a total error. Adding

systematic errors to the total error is generally only admissible, if they are not separable from other errors or if their sign cannot be defined. Neither is true for SErt and SEturb.

Based on current literature, there seems to be no consensus in the flux community on how to handle systematic errors. Some groups lump systematic errors into total error as we have done (Misztal et al., 2014; Vaughan et al., 2016). Gioli et al. (2004) applies only a high-frequency correction, while Mauder et al. (2013) derives long-wavelength errors based on energy closure but advocates against using these to correct fluxes. Still other groups seem to ignore uncertainties entirely in their analysis (Desjardins et al., 2018; Sayres et al., 2017). Furthermore, SE_{rt} is sometimes unreasonably large when spectra are noisy (P. 11, Line 29), and SE_{turb} represents an upper limit. For these reasons, we believe it best to report SE as separate data columns and allow data users to decide how to treat these errors.

We have added the following discussion to the end of Sect. 3.4.1:

Systematic errors can be applied as a correction factor to fluxes (if of known sign) or be included as part of the total uncertainty. Both practices are common among the airborne flux community (Gioli et al., 2004; Misztal et al., 2014). For the errors discussed above, SE_{acc} is of unknown sign, while SE_{turb} and SE_{RT} should both increase the flux. We are, however, reluctant to employ the latter two as correction factors. SE_{turb} represents an upper limit and thus may slightly "over-correct" the fluxes, while SE_{RT} can become unrealistically large when fluxes are small due to the amplification of high-frequency noise by Eq. (9). Thus, we elect to include all systematic errors in the total flux error and assume all error components are symmetric for simplicity. Total systematic error (SE_{tot}), given as a fraction of the flux over any interval, is then the root-sum-square of SE_{turb}, SE_{rt}, and SE_{acc}. Total systematic error is reported as a separate variable in flux archive files and may be used as part of the total error or as a correction factor (after removing the accuracy contribution) at the discretion of data end-users.

Additional systematic errors for surface fluxes arising from flux divergence are discussed separately but should probably be part of section 3.4.

We have moved this discussion to Sect. 3.4 and added sections on total error and error averaging.

Repeatability: it is mathematically not sound to simply average second moments as presented in Figure 11 (see for example: https://www.eol.ucar.edu/content/combining-short-term-moments-longer-time-periods). Within the uncertainty of the presented data it might not make a large difference for Figure 11, but it would be worth double checking using the correct averaging formula.

From the referenced link, the relevant formula here is

$$\overline{x'y'}^N = \frac{1}{N}\sum_{j=1}^{m} N_j(\overline{x'y'}^j + \overline{X}^j\overline{Y}^j) - \frac{1}{N}\sum_{j=1}^{m} N_j\overline{X}^j \frac{1}{N}\sum_{j=1}^{m} N_j\overline{Y}^j$$

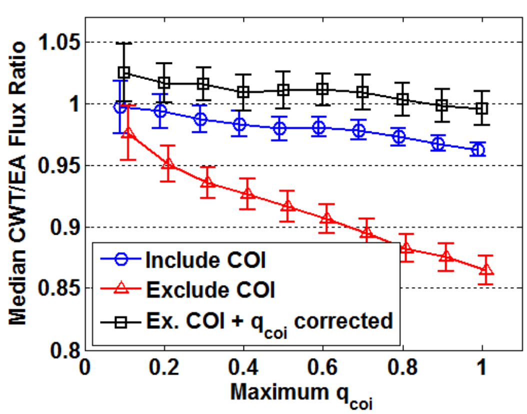
Here, x and y correspond to scalar and vertical wind measurements and there are m sub-intervals of length N_j included in the average over a total interval of length N. For any flux-relevant sub-interval, however, the mean vertical wind should be sufficiently close to 0 that the terms on the right are negligibly small. Also, all N_j are roughly equal in our averaging routines. In this case, the formula simplifies to an unweighted average, as used in this and many other studies.

We did check anyway, as the reviewer suggested. We find no appreciable difference in average fluxes using either averaging method.

Figure 6: the plotted differences are likely caused by a dramatic increase of systematic errors (eq. 7) towards the edges of the CWT – could the calculated flux ratios improve when accounting for these SE ? (e.g. by introducing a weighted SE along the CWT). To be more specific, the COI cuts off a substantial part of the frequency domain towards the edge of the CWT which should result in a systematic flux underestimation according to eq 7.

We assume that the reviewer is referring to the difference between the "exclude COI" and "include COI" cases. Exclusion of the COI from scale-averaging necessarily leads to a systematic underestimate of the true flux for the reasons the reviewer describes, and this is discussed in Sect. 3.3.1. Figure 6 provides an ensemble estimate for the resulting systematic error; in the case where we do not filter with the q_{coi} flag (rightmost points), the error (taken as the difference between the "include" and "exclude" cases) is ~10% of the flux. We have added some text to this section to clarify this point.

It is difficult to develop a robust (time-dependent along the CWT) estimate of the systematic error resulting from the COI, as this area by definition represents a region where the wavelet transform suffers from limited information. In theory the q_{coi} flag provides a rough means of doing this calculation: by first scale-averaging the CWT while excluding the COI, and then dividing by ($1-q_{coi}$) to correct for the fraction of cospectral power that was lost in the COI. The below plot shows the results of this calculation (black squares). The correction does indeed mitigate the systematic errors caused be excluding the COI for the ensemble of all fluxes. We hesitate to recommend such a correction for time-resolved CWT fluxes, however, as it inherently assumes that the globally-averaged ogive/cospectrum is representative of the local ogives/cospectra.



The purpose of the q_{coi} flag is to allow filtering of fluxes that may suffer from large COI-related systematic errors. We believe this conservative strategy is preferable to attempting to recover cospectral power within the COI via a correction factor.

Minor Comments: Figure 8: How high was zi?

1070 m. We have added this info to the figure caption.

Figure 9a: A label for the CO2 and CH4 instrument should be added (e.g. LGR)

The "LGR" is meant to distinguish between the two water measurements. Since there is only one CO2 and CH4 flux measurement, we feel this change is unneeded. Actually the label should be GHG as this is the name of the system; we have modified Figs. 7, 9 and 10 accordingly.

References

Desjardins, R. L., Worth, D. E., Pattey, E., VanderZaag, A., Srinivasan, R., Mauder, M., Worthy, D., Sweeney, C., and Metzger, S.: The challenge of reconciling bottom-up agricultural methane emissions inventories with top-down measurements, Agr. Forest Met., 248, 48-59, 2018.

Gioli, B., Miglietta, F., De Martino, B., Hutjes, R. W. A., Dolman, H. A. J., Lindroth, A., Schumacher, M., Sanz, M. J., Manca, G., Peressotti, A., and Dumas, E. J.: Comparison between tower and aircraft-based eddy covariance fluxes in five European regions, Agr. Forest Met., 127, 1-16, 2004.

Mauder, M., Cuntz, M., Drue, C., Graf, A., Rebmann, C., Schmid, H., Schmidt, M., and Steinbrecher, R.: A strategy for quality and uncertainty assessment of long-term eddy-covariance measurements, Agr. Forest Met., 169, 122-135, 2013.

Misztal, P. K., Karl, T., Weber, R., Jonsson, H. H., Guenther, A. B., and Goldstein, A. H.: Airborne flux measurements of biogenic volatile organic compounds over California, Atmos. Chem. Phys., 14, 10631-10647, 2014.

Sayres, D. S., Dobosy, R., Healy, C., Dumas, E., Kochendorfer, J., Munster, J., Wilkerson, J., Baker, B., and Anderson, J. G.: Arctic regional methane fluxes by ecotope as derived using eddy covariance from a low-flying aircraft, Atmos. Chem. Phys., 17, 8619-8633, 2017.

Vaughan, A. R., Lee, J. D., Misztal, P. K., Metzger, S., Shaw, M. D., Lewis, A. C., Purvis, R. M., Carslaw, D. C., Goldstein, A. H., Hewitt, C. N., Davison, B., Beevers, S. D., and Karl, T. G.: Spatially resolved flux measurements of NOx from London suggest significantly higher emissions than predicted by inventories, Faraday Disc., 189, 455-472, 2016.

Webb, E. K., Pearman, G. I., and Leuning, R.: Correction of flux measurements for density effects due to heat and water vapour transfer, Quarterly Journal of the Royal Meteorological Society, 106, 85-100, 1980.