Second Review of manuscript bg-2020-445 by Vogl et al.

Reply to comments of Reviewer # 1:

We want to thank the anonymous referee again for their helpful comments. We indeed have missed to address the minor comments from their first review, which was due to human error, which we want to apologize for. We are addressing the raised comments (and the minor comments from round 1) in a point-by-point way below:

The authors addressed the major comments of my first review in a more or less satisfying way (see detailed comments below). However, they overlooked or disregarded all my "Minor Comments", which is not acceptable. In addition, the author responses (and revisions) led to some additional issues that need to be addressed before the paper can be published.

Important: The page and line numbers I use in the following refer to the revised manuscript version with markups (ATC1).

Again, we apologize for missing out on addressing the minor comments. We are addressing them in this review round.

COMMENTS

1) Please address the list of "Minor comments" in my first review. These comments have not been considered in this revised version (Some of the comments may have become obsolete due to other changes)

We are addressing the minor comments from round 1 in a point-by-point way below:

MINOR COMMENTS

P1, L15-16: It is not clear, which β approach this sentence is related to.

This comment has become obsolete. We agree that this sentence was unclear, and it was removed from the abstract.

P2, L2-5: Both sentences are formulated in a misleading way.

First sentence: the detection limit of the instrument does not limit the REA fluxes directly but the quality/uncertainty of the REA fluxes. Change e.g. to "...when the uncertainty of the REA flux quantification is not limited by ...".

Second sentence: change to: "For REA sampling differences close to the instruments detection limit …"

Thank you for pointing us to these unclear formulations. We appreciate your suggesting a better alternative. We have changed the sentences accordingly.

P6, Fig. 1: Please check if the position of the grey points in the right panel is correct. According to Eq. 4 and a β w value of about 0.6, the normalized vertical distance of the two grey points (=DELTAw/ SIGMA w) should be about 1.6, but in the figure this distance is much less than 1. Maybe the x- and y-axis need to be exchanged...?

Fig. 1 was updated during the first review round, and the grey points were removed. Please note also that the right panel in Fig. 1 refers to hyperbolic deadbands, and (former) Eq. 4 refers to the β w method (model 3).

P8, Table 1: The units for the roughness length are probably [cm], not [m]. Only indicate two significant digits in the roughness length values, because their accuracy is not so high.

Please also include the average canopy heights and the EC measurement heights in the table (better than scattered in the text). This would be advantageous for the reader.

Thank you for spotting the error and for the suggestion to include canopy and measurement heights in Table 1. The roughness length values were updated during the first review round (comment R2S15). We have added the estimated canopy and measurement heights in the table as suggested.

P9, L17: The formulation "...resulting in a total measurement height ..." is not logical (what results in what?). Please rephrase.

We have rephrased the sentence as follows: „The EC flux instrumentation (2 m high) was installed on top of a 31 m high scaffolding tower reaching above the highest tree tops, resulting in a total measurement height of 33 m above ground."

P9, last line: Correct to: "A diel course is still observed, but the flux is constantly directed ..."

Thanks for spotting this. We have changed the sentence as suggested.

P10, L18: What do you mean with "perturbation time scale"

*„perturbation time scale" is a term related to Reynolds averaging. It describes the **time scale** over which the average is computed; this average is used to derive the **perturbations** by subtracting it from each individual observation.*

P10, L25: Explain the "additional hard thresholding".

This issue was addressed in the answer to R2S20 as follows:

We applied physical plausibility thresholds to filter the data for unphysical outliers. These thresholds were different for each scalar and each data set, due to different biochemical and meteorological conditions, and different measurement systems used. More specifically, the thresholds were defined as follows in our code (for the Dry Valleys/ gravel site "DRYVEXA", the forest site "WS2016", and the meadow site "ExpMM2015", respectively):

```
if (d=="DRYVEXA"){phys=c(0,400,0,300,-0.020,0.010)} # sensible heat, latent heat, co2

if (d=="WS2016"){phys=c(-100, 550,-100,400,-0.010,0.010)}

if (d=="ExpMM2015"){phys=c(-100, 200,-80,250,-0.020,0.010)}
```

To improve the readability of our text, we added an explanatory sentence to Section 3.2, which combines this issue with another comment raised by Referee #1 (**R1C4**):

“In the final step, the same thresholds for physical plausibility which were applied to the computed EC fluxes were also used to remove unplausible REA flux estimates from the data sets. These thresholds were chosen individually for each scalar and each data set due to the wide range of meteorological and biochemical conditions covered in this study.”

P11, Line 1-2: I do not understand what "force it through zero" means here.

This question is related to the problems arising from the negligible CO_2 flux measured at the Antarctic gravel site. We added a constant correction of 0.00035 mol L^{-1} to each observation to obtain a mean CO_2 flux of 0. This value was determined empirically. Using this constant offset correction does not impact our other evaluations and was only done for physically correct visual representation, as no CO_2 flux is expected at that site due to absence of biological activity.

P11, Line 5: I do not understand why the slope m had to be computed in the present study. It is not necessary for the β w calculation according to Eq. 4. Moreover, the w'-c' statistics are not available in a real REA application (see also comment 3 above).

You are correct that m does not need to be computed, and that w‘-c‘ statistics are not available for real REA applications. We included these considerations for the sake of completeness, and for clarity of the theoretical derivation of REA / β.

P20 Fig. 11 middle and right panel: The symbol colors hardly distinguishable. Removing the black frame of the symbols may be helpful.

Thank you for the suggestion. We decreased the line width of the markers used in this plot.

P21, L7-10 ("The tested REA models .... along with the main results of this study.") This part should be omitted from the Conclusions because it is pure repetition.

We want to thank the reviewer for this comment. However, we think that this summary of the used methods and models at the beginning of the Conclusions is beneficial for the ‚quick reader‘ focusing on abstract and conclusions, and hence would rather like to keep this part. We leave the decision to the editor.

Figures 5-8: Indicate the units of the RMSE in the right panels.

Thank you for spotting the missing unit in the plots. They were added to the figures.

Table 2: In the second lowest row, "rxx" presumably should be replaced by "rxy"

Thank you for spotting this. The term „r_{xy}“ was changed to „r_{sp}“ (scalar – proxy) during the first review round, including in the table (where the phrase was wrong indeed)

2)

The reviewer's second comment refers to the discussion following R1C1 during the first review round, which is why we are first adding the complete discussion (R1C1 and our answer to this comment) below:

round 1 R1C1) Only the performance of the REA approaches for the H_2O flux is tested in the present study. This is done after an initial deadband optimization (using the reference EC dataset) for the same test scalar. This leads to a certain lack of independence in the method validation. Although the CO_2 flux and its correlation with the other scalar fluxes is introduced in Sections 3 and 4.1, the REA evaluations for the CO_2 flux are unfortunately not presented. Alternatively CO_2 could have served as second proxy scalar option beside the temperature T (at least for some sites) as indicated in Section 2.3.

The authors should more prominently (in abstract and objectives) declare that they are evaluating the REA approaches only for H_2O fluxes. In addition they need to discuss better, whether and why they assume that the results also apply to other scalars, despite a sometimes low scalar correlation as exhibited in Fig. 3.

round 1 authors' answer to R1C1) The reason why only the results for the H_2O flux are presented was to limit the analysis to a reasonable scope. Additionally, we decided to not present the CO_2 flux results because, for the gravel site (Antarctica), there is basically no measurable CO_2 flux due to lack of biological activity, which makes the interpretation difficult. However, we agree that, for method validation, considering another flux than the one for which the deadband size was optimized is required. Following the referee's suggestion, we propose adding an appendix (Appendix A), in which we present the hourly binned RMSE evaluation, which was done for H_2O in Fig. 9, but for the CO_2 flux. Alternatively, the below figure and interpretation could be included and discussed in the main manuscript. We would like to leave this decision to the editor. Regarding the second part of the comment, we state that the changes will be reflected in abstract and introduction.

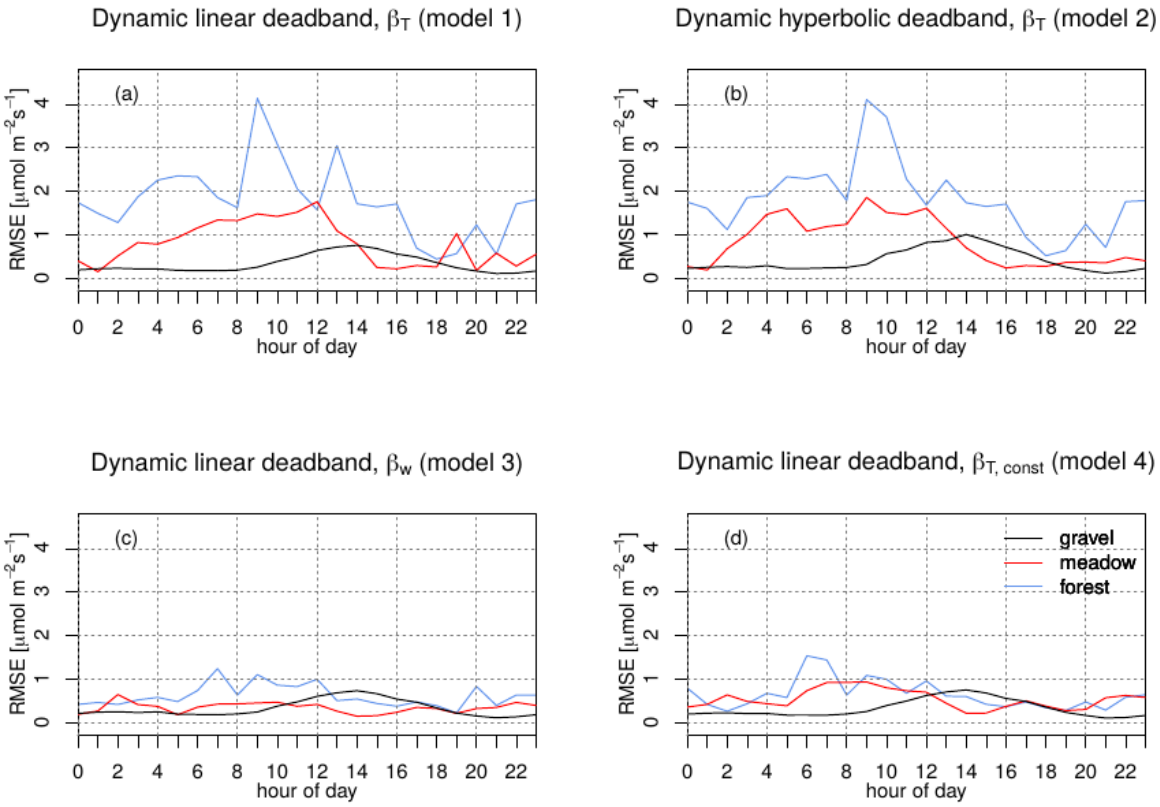

Fig. 11: Same as Fig. 9 but for the CO_2 flux. The gravel site results (solid black lines) should be regarded with caution as the magnitude of the CO_2 flux at this site is close to zero (compare to Fig. 2).

Interpretation: *The same findings that were drawn from the H_2O flux analysis are also apparent in the above figure: Both proxy approaches (panels (a) and (b)) result in higher values of the RMSE than the β_w (panel (c)) and the constant β (panel (d)) methods. The RMSE for both proxy approaches at the meadow site peaks during 13-14 UTC, the time when scalar-scalar correlation of sensible heat and CO_2 is lowest. At the forest site, the RMSE for the β_T approaches is highest when the magnitude of the CO_2 is largest. The RMSE for the gravel site is included in this figure even though the magnitude of the CO_2 flux is close to 0 throughout the daily course and thus no conclusions should be drawn from its RMSE.*

R1C1 response: I still miss the important discussion (in the Discussion section) about the assumption of scalar similarity, i.e. whether and why the results for H_2O investigated here (with minor results also for CO_2) can be applied to all other scalars, especially the ones for which REA is usually applied.

This is especially critical because the authors state on P24, Line 8: "Choosing the optimum proxy scalar is critical for the methods success". This sentence implies that a general similarity between all scalars is not expected.

Answer to the answer to R1C1:

We want to thank the reviewer for this comment. First of all, we would like to clarify that two different „instances" of scalar similarity are of importance in the presented study:

(i) scalar-proxy similarity assumed by models 1 and 2, which employ a half-hourly adjusted proxy-derived β p value according to Eq. (2):

$$\beta_p = \frac{\overline{w'p'}}{\sigma_w \cdot \Delta \overline{p}}$$

We can investigate the validity of this scalar-proxy similarity assumption e.g. using the scalar-proxy correlation coefficients $r_{s,p}$

(ii) scalar similarity with respect to the general validity of our presented results for scalars different from H_2O. E.g. whether the optimized deadbands presented in this study can be used also for REA flux measurements of other atmospheric compounds such as ammonia or aerosol particles. This is more difficult to answer, regarding the available data we have.

You are encouraging more in-depth discussion about (ii), while the sentence you cite from the Conclusions refers to (i). We are therefore suggesting to rephrase the sentence as follows:

„Concerning models 1, and 2, choosing the optimal proxy scalar is critical for the methods‘ success.“

It is however true that we did not further discuss (in the Discussion section) whether conclusions can be drawn about the flux estimation of other scalars than H_2O. We acknowledge that adding a more in-depth discussion about this issue would definitely improve the manuscript. Given the data we have at hand (fast-response observations of CO_2, H_2O and temperature), we can however only answer the question whether our results found for the H_2O flux are also valid for the CO_2 flux, if temperatures is used as the proxy .

We have recreated Figs. 5-8 but for the CO_2 flux:

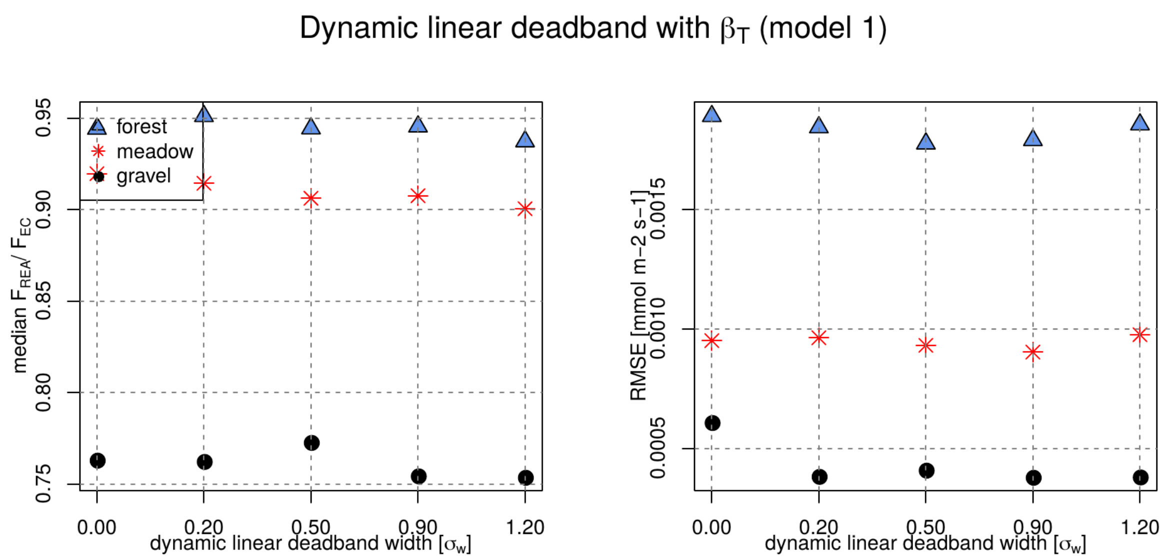

Figure 1: Errors as a function of dynamic linear deadband width. The x axis is the scaling factor a multiplied with the vertical wind standard deviation in Eq. 9 to define the deadband threshold. Left panel: Median F_{REA}/F_{EC} (CO_2 flux simulated with sensible heat as a proxy) ratio for each of the simulated dynamic deadband widths; right panel: RMSE for each of the simulated dynamic deadband widths

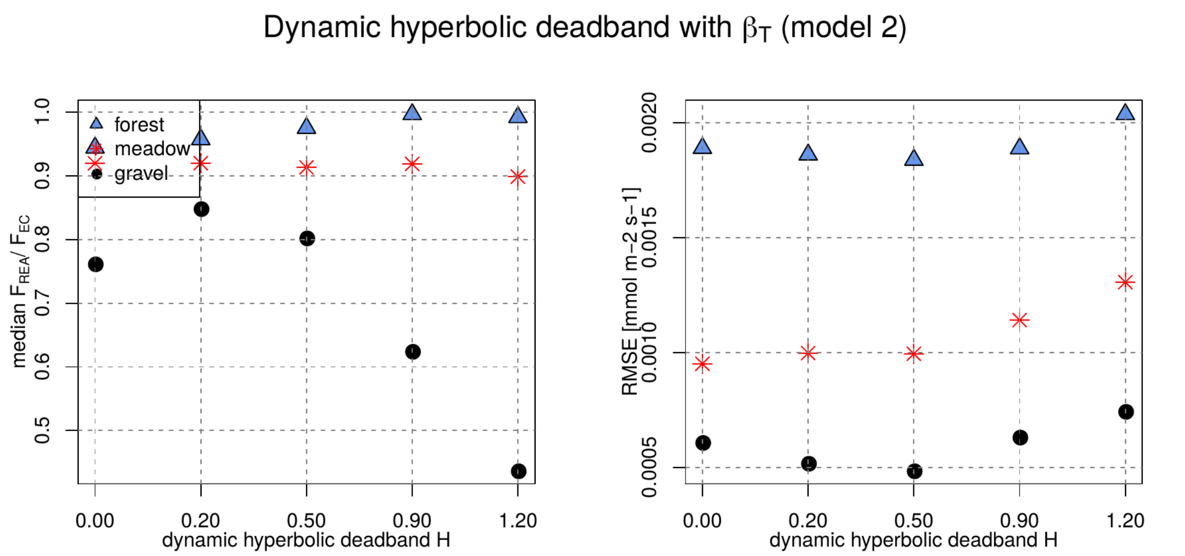

Figure 2: Errors as a function of dynamic hyperbolic deadband size. The x axis is the H parameter in Eq. 10, which defines the deadband size. Left panel: Median F_{REA}/F_{EC} (CO_2 flux simulated with sensible heat as a proxy) ratio for each of the simulated dynamic deadband sizes; right panel: RMSE for each of the simulated dynamic deadband sizes

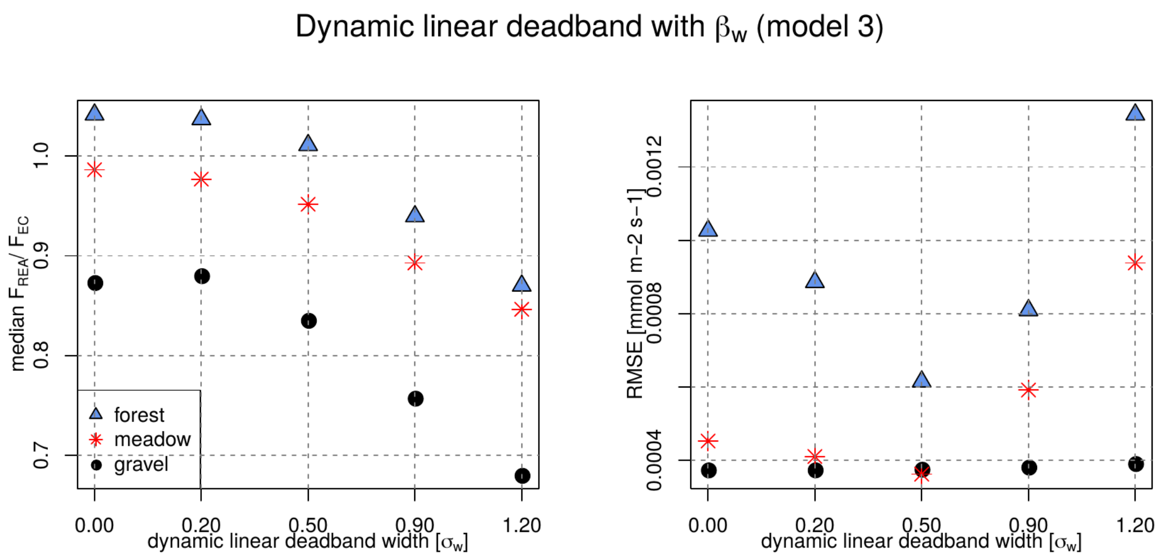

Figure 3: Errors as a function of dynamic linear deadband width. The x axis is the scaling factor a which is multiplied with the vertical wind standard deviation in Eq. 9 to define the deadband threshold. Left panel: Median F_{REA}/F_{EC} (CO_2 flux simulated using the REA approach described in Baker (2000)) for each of the simulated dynamic deadband widths; right panel: RMSE for each of the simulated dynamic deadband widths

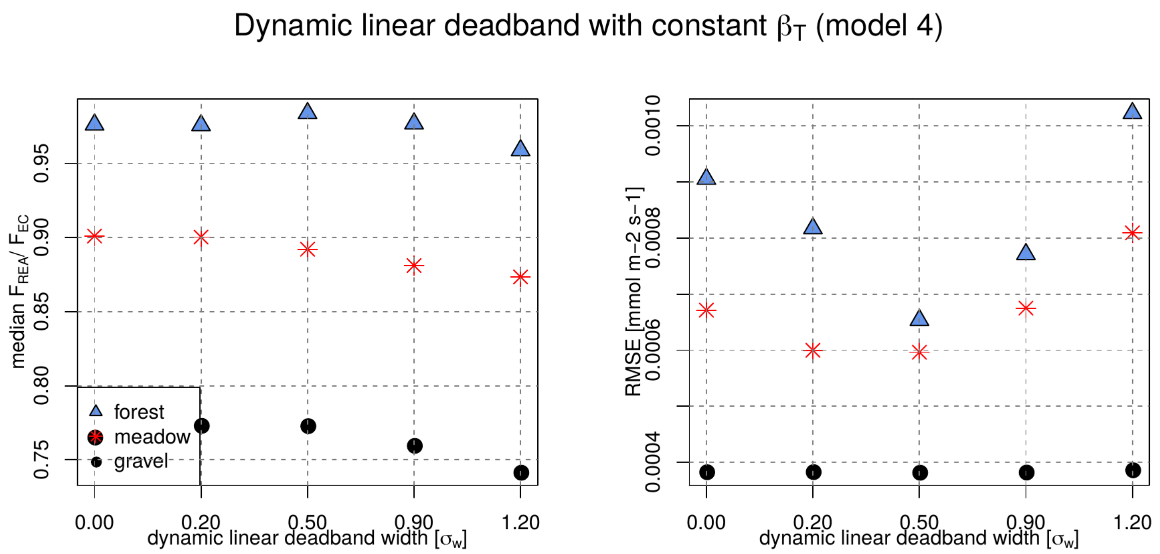

Figure 4: Errors as a function of dynamic linear deadband width. The x axis is the scaling factor a which is multiplied with the vertical wind standard deviation in Eq. 9 to define the deadband threshold. Left panel: Median F_{REA}/F_{EC} (CO_2 flux simulated using constant β T and dynamic linear vertical wind deadband) for each of the simulated dynamic deadband widths; right panel: RMSE for each of the simulated dynamic deadband widths

The optimal deadband sizes summarized in Table 2 prove to be also valid for the CO_2 flux: While it is hard to decide for an optimum deadband size for models 1 and 2, the deadband sizes found for H_2O (a=0.5; H=0.5) are among the best-performing options. For models 3 and 4, the deadband size found for H_2O (a=0.5 for both models) also exhibit an optimum (minimum) RMSE for the CO_2 flux.

During the first review round, we showed that applying these deadband sizes to CO_2 results in a similar pattern in the diurnal RMSE as was observed for the H_2O flux: Smaller RMSE for models 3 and 4 than for models 1 and 2, and the forest site exhibiting larger RMSE than the meadow site.

We propose adjusting the paragraph about the CO_2 flux at the end of section 4.3.1 as follows, adding a discussion about the applicability to other scalars:

„So far, only one proxy-scalar combination was investigated in this study. However, showing that the presented results are also valid for other scalars is critical for their applicability. The data sets allow for including CO_2 for additional validation. The CO_2 flux was simulated with the optimized models 1-4 (using the deadband sizes summarized in Table 2), with sensible heat as the proxy for models 1, 2 and 4. Comparison of F_{REA}/F_{EC} ratio and RMSE indicated that the optimum deadband sizes found for H_2O (Table 2) are also valid for CO_2. The hourly RMSEs are included in Appendix A in Fig. A1. A similar pattern in the diurnal RMSE as observed for the H_2O flux also emerges for CO_2: Models 1 and 2 both yield higher RMSEs than models 3 and 4, the forest site exhibiting larger RMSEs than the meadow site. These findings suggest that the results presented here for the H_2O flux are also valid for the CO_2 flux, and possibly other atmospheric compounds. However, we cannot arrive at a final conclusion for other (including reactive) scalars, for which REA is often applied, since fast-response analyzers are missing. Answering this question is beyond the scope of this study and should be considered in future research."

3) P1, line9: In the context of major comment 2 above, this statement is over-ambitious, especially concerning the part "...formulating universally applicable recommendations...", and I suggest to downgrade it to some extent.

We want to thank the reviewer for this comment. We propose to change the sentence as follows:

„This study evaluates a variety of different REA approaches with the goal of formulating recommendations applicable over a wide range of surfaces and meteorological conditions for an optimal choice of the β factor in combination with a suitable deadband."

4)

The fourth comment refers to the discussion following R1C4 during the first review round, which is why we are first adding the discussion regarding R1C4:

R1C4) How can it be that the zero deadband calculations result in RMSE of about 20 mmol m-2 s-1 for the forest site in Figs. 5 and 6, when the fluxes themselves are only between 0 and 4 mmol m-2 s-1 (Fig. 2) and the flux ratios in the left panels are close to 1? This seems very unplausible and needs a detailed explanation.

round 1 authors' answer to R1C4) Thanks for spotting this. The large RMSE compared to the median F_{REA}/F_{EC} ratio close to 1 was actually due to one single outlier. We decided to take the physical plausibility thresholds, which were applied to the EC fluxes, and also apply them to all simulated REA fluxes. This removes the outlier in question, and reduces the RMSE values for the forest site in Figs 5 and 6. However, the thresholding does not alter any of the other presented results significantly. The main finding presented in this section, i.e. that the proxy-based approaches result in a larger error compared to the β_w and $\beta_{T,const}$ approaches, remains still valid.

We propose to include the following explanation in Section 3.2, stating that the physical plausibility thresholds were applied to the simulated REA fluxes as well:

"In the final step, the same thresholds for physical plausibility which were applied to the computed EC fluxes were also used to remove unplausible REA flux estimates from the data sets. These thresholds were chosen individually for each scalar and each data set due to the wide range of meteorological and biochemical conditions covered in this study."

R1C4 response: Such a strong effect of one single outlier makes the suitability of the used evaluation method questionable.

We agree that the RMSE is prone to outliers in the data, which makes strict qualtity criteria necessary. Due to concerns regarding RMSE, we decided to not only make our choice of optimal deadband using the RMSE but to also take median F_{REA}/F_{EC} into account (which did not change after removing the outlier, but in turn are problematic when the magnitude of observed fluxes is small). Combining these two measures makes us confident with regards to the choice of optimal deadband size. The sensitivity of the RMSE to individual large outliers is inherent to its statistical definition, which is a well accepted metric when comparing methods.

5)

The fifth comment refers to the discussion following R1C6 during the first review round, which is why we are first adding the discussion regarding R1C6:

R1C6) For Figure 10 and 11 it is not indicated, which data are displayed. Are these all (valid) data for all three sites or only data from one site? This needs to be clearly stated in the Figure caption.

round 1 authors' answer to R1C6) Thanks for bringing up this issue. In Figs 10 and 11, all valid data from all three sites are combined. The observations from all three ecosystems fall along the same lines, which suggests that e.g. the findings of β_w vs. kurtosis as a function of deadband size presented in the left panel of Figure 11 are universally applicable.

For clarification, we are adding the following sentence to the caption of Fig. 10:

"This figure combines valid data points from all three sites."

and we are adding

"Valid data points from all three sites are combined in this panel.",

to the caption of Fig. 11.

R1C6 response: The pooling of data from all three sites in the evaluation can be

problematic. Does this imply that β_w and the w-statistics as well as β_T values are fully independent of the site conditions (canopy height, roughness, correlation between proxy scalar and scalar of interest, etc.) ? This issue needs some statements/discussion in the text.

We want to thank the reviewer for this comment. This is a very interesting point indeed. The findings of Ammann & Meixner (2002) pointed towards a systematic dependence of the β_p factor from the stability parameter z/L which we tried to reproduce in our data. However, we can only find the pattern described in their work for the forest site, and only if no or small deadbands are used. For the other sites and larger deadbands, the relationship described by Ammann & Meixner (2002) vanishes. Below are the plots of the results shown in Fig. 10, but for each site individually:

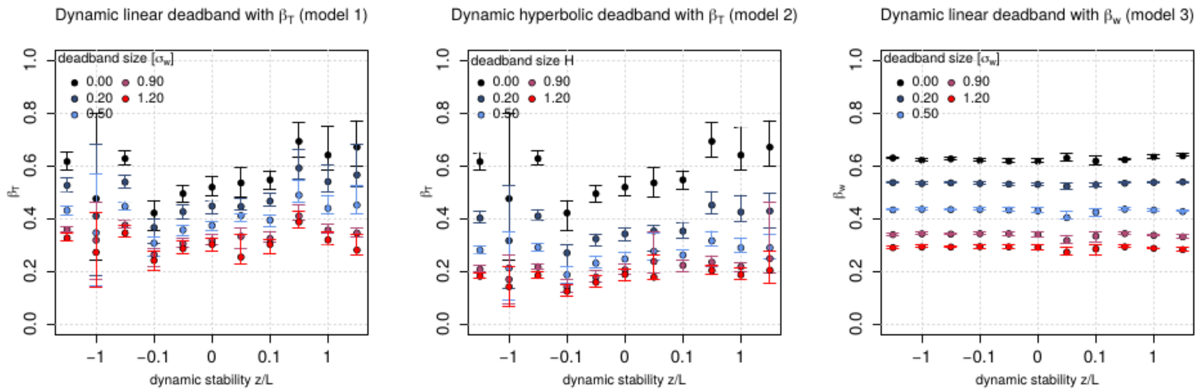

Figure 5: Same as Figure 10, but for the forest site only

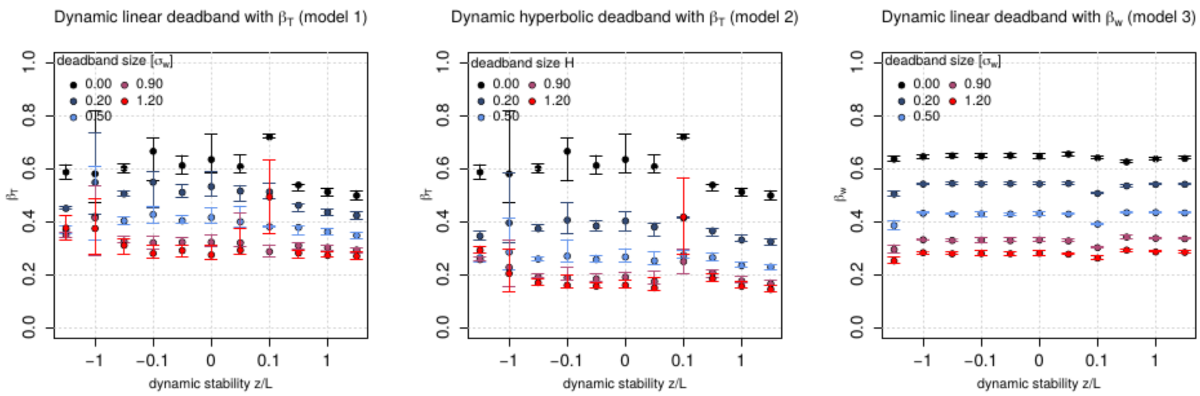

Figure 6: Same as Figure 10, but for the meadow site only

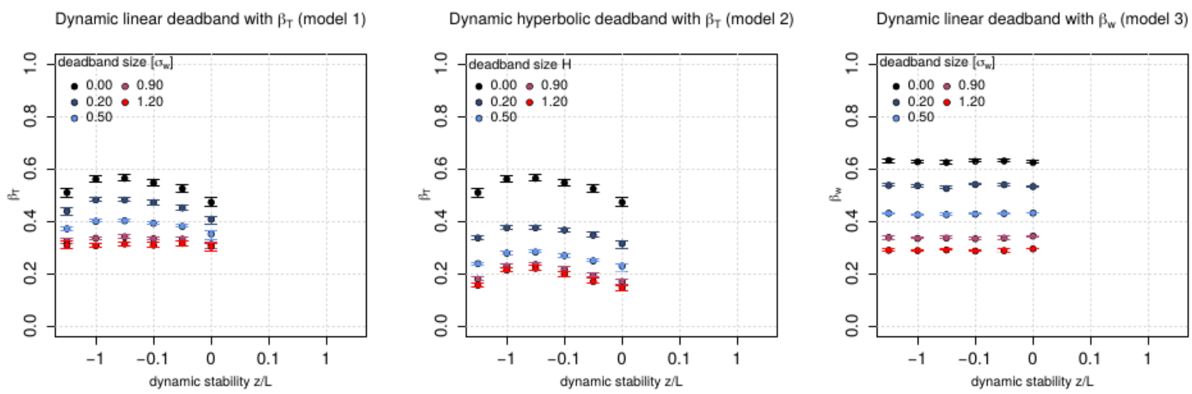

Figure 7: Same as Figure 10, but for the gravel site only

However, concerning the β_w factor, the situation is different. The right-hand panels in the above figures all show a strikingly similar pattern; also, if we are plotting the data in the left panel in Fig. 11 (which was the core of this comment) individually for the three sites, the results look very similar:

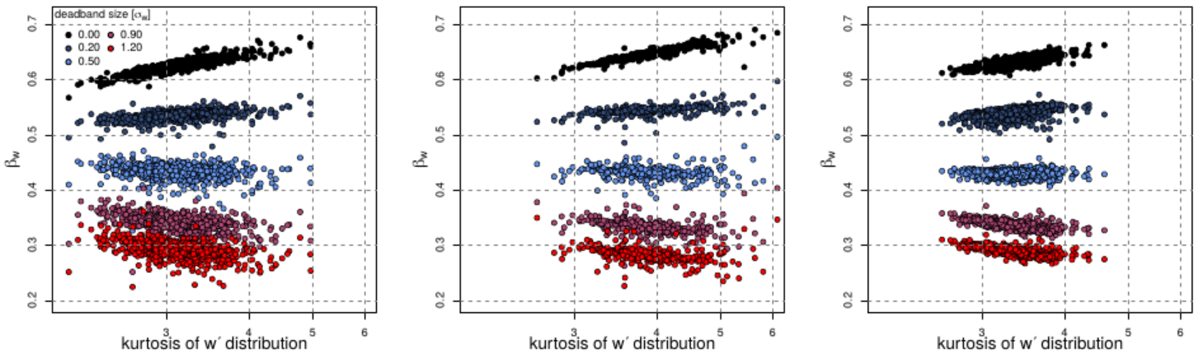

Figure 8: The left panel of Fig. 11 but for each site individually: Forest (left), meadow (center) and gravel (right).

To answer the reviewer's question: Yes, this implies that β_w and the w-statistics are fully independent of the site conditions. β_T on the other hand behaves differently for each site and cannot be described by z/L reliably.

We propose rephrasing part of the discussion about Fig. 10 as follows:

„For the two dynamic proxy models (models 1 and 2; left and center panel in Fig. 10), β_T without deadband approximately follows the relationship found by Ammann and Meixner (2002), i.e. a constant β T for unstable conditions, and an increase from neutral and stable conditions of z/L>= 0.06. However, this increase is associated with large statistical uncertainty and only due to the data from the forest site (please note that Fig. 10 combines the observations from all three sites). We therefore recommend exercising caution when using stability-dependent parameterizations of β_T. Variability of β_T generally decreases with increasing deadband size. Model 3 (right panel in Fig. 10) shows a very different behavior: β_W is apparently unrelated to dynamic stability, and displays a generally lower variabiliy than β_T."

Also, we propose to add the following to the discussion belonging to Fig. 11:

„This finding suggests that the turbulence statistics, including the β_w factor, are site-independent despite the significant differences in climate and surface characteristics across the three ecosystems (canopy height, roughness, etc.)."

6) R1C5 response: The newly added Table 2 is very important for the present study. There are two important questions arising from it that deserve some thoughts/discussion: i) can the average β w be considered as a site independent constant?; ii) Would the use of an average constant β w yield similarly good results like the use of half-hourly β w-values?

The (successful) use of an overall constant β w value would strongly simplify the REA measurements.

Thank you for pointing this out. These are indeed very interesting questions. Concerning (i):

β_w indeed was found to be independent from the site (see the answer to your comment above); and regarding (ii): It is shown e.g. in Fig. 11 that β_w does not vary a lot over time.

We hence propose to add the following to the discussion of Fig. 11:

„This is confirmed by the nearly identical average β_w values found for the three sites in Table 2 of 0.43-0.44. In connection with the small spread of β_w values in Fig. 11, and the strikingly similar RMSE for models 3 and 4 in Fig. 9, our results suggest that β_w can be considered a both site- and time-independent constant.“

7) In various parts of the text and the abstract the terms "proxy-based approaches" or "proxy-dependent approaches" are used synonymously for the models 1 and 2. However, this is misleading and not correct, because the model 4 approach is proxy-based as well (and also model 3 uses w as proxy, though it is not a scalar). I suggest to use a more suitable and specific expression for models 1 and 2 like e.g. "dynamic scalar proxy approaches" or just "model 1 and 2 approaches"

We thank the reviewer for this remark. However, we think that „dynamic scalar proxy approaches“ would impact the readability of our manuscript in a negative way. We have added the word „dynamic“ to the occurrences of „proxy-based“ and „proxy-dependent“, to distinguish these two models from model 4, which uses a constant proxy-based β factor.

8) Figure 10: correct the y-axis titles in the left and central panel to β_T.

Thanks for spotting this. We have made the suggested changes.

9) Figure A1: For H_2O in Fig. 9 an individually adjusted deadband width was used. Was the same deadband width also applied for CO_2, or was it indivdually adjusted for CO_2. In either case, does an individual adjustment yield the same optimum deadband widths for H_2O and CO_2?

We want to thank the reviewer for this good comment. We have already answered this question in our reply to (1).