This manuscript of Vardag *et al*. presents an analytical approach to evaluate the CO_2 source signature $\delta^{13}C_{S,}$ using continuous, high resolution time-series of CO_2 and $\delta^{13}C$, recorded with an FTIR. The analysis is based on the Keeling-plot method, where a time-window of 5 hours is continuously moved across the whole data set, resulting in a continuous source signature estimate over the observation period of four-years. The manuscript is generally well written, uses an outstanding data-record and validates the proposed method using pseudo data from the STILT model. However, the major findings, like the strong limitation of the Keeling-plot method for urban catchment areas with multiple and variable sources as well as the seasonal variation of the source signatures are known since many years and discussed in a vast number of publications, some of which are also referenced by the authors. Although, it is useful (but not novel) to see the difficulties of estimating the year-round CO_2 fossil fuel or biosphere share in urban atmosphere using the CO_2 and $\delta^{13}C$ data only, the reader is left with vague alternatives and a method, which is empirically tuned to a specific spatial and temporal setting, rejecting about 85% of the estimated values. This manuscript would strongly gain scientific value by including further tracers such as $^{14}CO_2$, $^{18}CO_2$, CO, and ^{222}Rn, discussing the advantages and pitfalls of such a combined approach, and deducing measurement strategies for future monitoring activities. As the authors have the above mentioned data (see e.g. Vogel *et al*. Tellus, 65, 2013) and a detailed model investigation (Vardag *et al*, ACP, 15, 2015), I strongly recommend using these in a concerted fashion to facilitate a better and clearer understanding of the limiting factors, requirements and identification of best practice for an efficient and unbiased monitoring of CO_2 source signatures. Without such major revision, the manuscript does not fulfill the high standards required for publication in Biogeosciences.

General comments:

A more appropriate title should be given. A "running Keeling approach" is awkward. First, the terminology broadly accepted by the community is the ''Keeling plot approach (or method)". Second, the mathematical operation applied in the described approach is a moving average or moving time window. In addition, the method does not differ (except the trace gas species and window size) from the method published by Röckmann *et al.*, so I strongly recommend to not increase the number of nomenclatures unnecessarily and stick with the name of "moving Keeling plot method" as proposed by Röckmann *et al*.

If the authors write four-years in the title then they should also give the signatures for all these years and not only limit to one particular year. Otherwise, give a reason why this year was selected as representative case and give estimates how the findings for 2012 can be extended to other years.

The abstract should also reflect the major drawbacks of the method: 85% of the data are rejected, because they do not fulfil the filtering criteria, mainly night-time periods are considered, and the selected criteria are empirical and specific to a particular urban area. Furthermore, an additional smoothing (100 h window) is applied to the estimated values.

The manuscript would greatly benefit from a more conventional structure, such as Introduction, Methods, Results and Discussion. Several sub-sub-sections are not necessary and hinder the text flow, e.g. by adding many cross-references. More specifically, I recommend merging the subsections 3.1 and 3.2 into section 3 as paragraphs. Similarly, sub-subsections 4.2.1 – 4.2.5 can be included in the main text using simple paragraph-spacing.

The averaging window was selected to be 5 hours, but the motivation is weak. In principle, the FTIR is able to produce 9 minute averaged values, so why not include the resulting 33 data points into the Keeling-plot intercept determination? The higher temporal resolution should lead to a more robust fit, and a better insight into the dynamics of source signature variations, which could eventually be used as a more objective filtering instead of the empirical criteria. Just consider Figure 1 with 10 fold better resolution. Arguing with the model resolution of 1 hour is not appropriate in this context. Similarly, the argument of being a period in which the source-mix does not change significantly is ambiguous because the large amount of rejected source signature estimates. For the reader it would be very useful to learn about the optimal temporal resolution but the respective limitation of the model and the instrument does, unfortunately, not allow to draw the corresponding conclusions.

How representative are the STILT model data for urban areas? A city with its complex network of buildings and street canyons generates turbulent flows at scales that are certainly beyond the resolution of STILT. Also, what is the model sensitivity at various sampling heights within an urban area?

The filter criteria used in the manuscript are mainly fulfilled for nighttime, so it would be good to know the uncertainty of the transport model for nocturnal data.

Advection and vertical mixing can significantly influence the urban CO_2 signal, leading to vertical gradients. Therefore, wind speed and direction data are most likely needed to adequately interpret the observed CO_2 values. Thus, a discussion about the representativeness and sensitivity of the sampling site to wind speed and direction as well as its location and height would be highly recommended.

The isotopic source signature of the biosphere is found to be more depleted than previously published value, but the analysis in the present work is mainly based on nighttime data, where photosynthesis is negligible and respiration dominates. Furthermore, distinguishing between respiration, coal burning and gasoline is difficult, because they have similar $\delta^{13}C$. The authors should discuss this potential bias on their δ_{bio} estimates. For such situations, the oxygen isotope ratio ($\delta^{18}O$) could be used to distinguish between biogenic and anthropogenic CO_2 as the evaporative enrichment of $H_2{}^{18}O$ in plants and soils imparts a unique signature. At the observed regional scale, it should be possible to provide the necessary model input.

In the same context, even the pseudo data shown in Fig 2a indicate a systematic bias for the summer period between the filtered and unfiltered cases. This discrepancy should be discussed in terms of influence in determining source signatures.

The source signature value (-32.5‰) found in this work is significantly different from the value (-25‰) published by the same authors for the same year (Vardag *et al*, 2015a). A discussion about this discrepancy is required.

Specific comments:

Abstract, L5: “without introducing biases” is a very strong statement and probably not applicable. “reducing biases” would be more appropriate.
Abstract, L6: state which model.

Abstract, L7: are these bias values for the model data? If so, state this explicitly.
Abstract, L13: This statement should be much more quantitative, which implies significant additional information and possibly research in the main section of the paper.

Pg2, L1: use plural for optical techniques, since there are various approaches available on the market.
Pg2, L2: thereby
Pg2, L21: "bias-free", see remark above
Pg2, L27: the "classical" is not necessary, because up to date there is only this method.
Pg3, L4: this sentence is awkward, I recommend reformulating it.
Pg3, Eq3: revise the formula, the CO_2bg has a positive sign.
Pg3, L14: "the Keeling plot" instead of "a Keeling plot".
Pg3, L28: why not to use measured data to test the different fit models? There should be no reason for synthetic data to deliver different results when applying different forms of the linear fitting routines. The situation can though be different when using real data.
Pg3, L29: for the very same criteria statement another reference is used (Sect 2.2 instead Sect. 2.3., see Pg3, L14)
Pg3, L30: specify, how the weights are determined?
Pg3, L31: revise the section name (see comment above regarding title)
Pg3, L33: "running" Keeling approach, again see above and delete this sentence.
Pg4, L19: The threshold criterion of 2‰ error has no objective motivation. Try to give its meaning in the context of some quantity like a confidence interval or in terms of source allocation error.
Pg4, L21: check wording "as a decrease of would be"
PG4, L28: how does this compare with a situation of 6 hour period and 4 or 6 ppm increase criteria? Is there a way to generalize these filter criteria?
Pg5. L7: give a reference for the STILT model.
Pg5. L24: what was the decision criterion for smoothing the source signatures with 100 hours window size? Evaluating the smoothing effect on pseudo data and assuming its validity on real data can be prone to errors.
Pg6, L11. Remove "Heidelberg" before "CO_2".
Pg6, L16. The explanation of outliers is weak and hard to understand. What do you mean by "statistical"? The filtering criteria were selected to be rather strict, so what else determines the uncertainty of the method?
Pg6, L18: are the values for inter-quartile ranges are for the smoothed data?
Pg7, L18 replace "we ask here, if we can" with "the question is whether it is possible to"
Pg9, L1. this section has nothing to do with accuracy evaluation, being more a qualitative description of various scenarios. Revision is recommended. See also suggestion above regarding text-flow.
Pg9, L15. This section is basically a repetition of what was already mentioned previously.
Pg10, L12: replace "real measured data set in Heidelberg" with "real data set measured in Heidelberg"
Fig.3 add the measured $\delta^{13}C_S$ to the figures.
Fig.5 it is somehow strange that if one considers the periods between January-April and October-December, where the measured $\delta^{13}C_S$ and assumed (or estimated) $\delta^{13}C_F$ show little deviation for both scenarios, the $\delta^{13}C_{bio}$ exhibits extreme fluctuations (Fig.5b). Furthermore, the fact that the agreement is good between model and observed $\delta^{13}C_S$ data would imply that

the summer period should look similar for the $\delta^{13}C_{bio}$ as well. In other words, what would the situation look like, when fixing both end members $\delta^{13}C_{bio}$ and $\delta^{13}C_F$, and estimating $\delta^{13}C_S$?

Appendix A, L7: *Röckmann et al.* found that fossil-fuel related emissions may be overestimated in EDGAR and using this inventory data leads to source signatures that are too enriched. Would this also apply to the CO_2 data presented in this work?.

Appendix A, L17-18: To what extent are the remote measurements made at Mace Head representative as background values for quantifying the regional atmospheric impact of urban CO_2 emissions in Heidelberg?