**Author reply to the comments by Anonymous Referee #2 of the manuscript gmd-2016-318**

# "eddy4R: A community-extensible processing, analysis and modeling framework for eddy-covariance data based on R, Git, Docker and HDF5"

**by S. Metzger et al.**

We thank Anonymous Referee #2 for the valuable feedback, which helped to improve the manuscript. Please find below the Referee comments recited in *blue, italics font*, followed by our point-by-point replies and corresponding changes in the manuscript in black, upright font.

*Metzger et al. describe a data processing framework which is illustrated on ad- hoc examples from NEON's eddy covariance tower and airborne measurement datasets. Overall this technical concept seems potentially valuable for streamlining automation of specific data processing steps from different measurement stations but it is extremely difficult to recognize the broader scientific values in the current version of the paper as written. I must admit that I was rather disappointed to find the description of the tools to be fragmented and poorly supported by the scientific results and conclusions. The whole analysis is very descriptive and in many cases misleading as to what is possible. There is little effort to synthesize what an be actually learnt from using the tool other than what its potential applications might be in the future. Most importantly, the paper does not specify scientific goals and does not even address the scope of modeling which is what the main interest of the journal's audience is. The manuscript seems to need much work to make the results and discussion useful for the scientific community but could be worthwhile to reconsider after major clarifications. The other reviewer has already provided a useful detailed guidance how this could be achieved and I agree with her/him. I also have other concerns which hopefully can be addressed in the revision.*

> Author reply: Many thanks for your summary. Please find specific replies below. As detailed in the responses to Referee #1, we have better clarified the problem statement of the paper and why the DevOps approach and specific tools used to implement are the answer. The problem statement is: "How do we collaboratively create portable, reproducible, open-source, scalable, and extensible software that improves reliability and comparability of eddy covariance data products?" We believe this clarification addresses the main concern of this Referee – that there is little of use to the scientific community presented by the description of software tools – by better communicating that the focus of the paper is not on the specific software implemented, but a model of

how the EC community can go about creating portable, reproducible, and extensible software.

As such the manuscript addresses a methodological rather than scientific question. For this reason, the GMD journal was chosen, and three tests of geoscientific applications are provided in favor of a single in-depth scientific survey. These applications serve not as scientific results, but as tests that the software is portable, reproducible, and extensible. One core component of GMD model description papers is the "…evaluation against standard benchmarks…" which is addressed in Sect. 3.3.

Changes in the manuscript: Please see responses below and also to Referee #1.

*General issues:*

*1) One fundamental issue in this paper intended for GMDD is that the work is not even connected to any model or modeling framework. The journal scope does not overlap with what paper is about or at least the connection is not made clear. Because there is no model, there is no model version – a requirement of the journal. There are only two words "model" in the whole paper, one of which is included in the last sentence of conclusions but probably in a different meaning: "We hope this framework can serve as a *model* for implementing community-sourced, distributed-development scientific code while combatting the deficiencies of current computational frameworks that limit accessibility, reproducibility, and extensibility."*

Author reply: The authors considered several journals before deciding where to submit our manuscript, and we came to this decision through taking into account the manuscript types requested on the Geoscientific Model Development (GMD) webpage. Specifically, we felt that our paper provides "…utility tools … such as coupling frameworks … with a geoscientific application".

In addition, as detailed in the replies to Referee #1, we clarified the problem statement of the paper: "The question we ask in this paper is: How do we collaboratively create portable, reproducible, open-source, scalable, and extensible software that improves reliability and comparability of eddy covariance data products?" We then introduce the DevOps approach in more detail and how it, along with the specific tools implemented in the eddy4R-Docker development model, solves this problem. The framework provides modular processing for surface-atmosphere exchange data with quality assurance and quality control as foundation for modelling exercises such as the test application in Sect. 3.2. This includes footprint modeling (GMD: Kljun et al., 2015), evaluation of large eddy simulations (GMD: Maronga et al., 2015), machine learning etc. The result is an end-to-end framework for model building, parameterization and assessment considering the large amounts of theoretical assumptions in eddy-covariance technique that require corrections to the data. The combination of these tools to address the concern of reproducibility was a major consideration when submitting to GMD.

Per suggestion of Referee #2 as well as the executive editor, in addition to Sect. 5 Code and data availability we now include the eddy4R-Docker development model version (now: 0.2.0) also in the manuscript title.

We further clarify in the revised manuscript that eddy-covariance data processing consists of employing a sequence of model algorithms. These often originate from scientific sub-fields with corresponding publications, and eddy4R-Docker provides an integrative, yet modular and extensible framework for their concerted application and continued development. In its current form eddy4R-Docker 0.2.0 encompasses the following models: plausibility tests (Taylor and Loescher, 2013), de-spiking (Brock, 1986), lag correction, data aggregation, and QA/QC budgeting (Smith et al., 2014).

Additional models are in preparation for future extension of the eddy4R-Docker framework presented here: coordinate rotation (Wilczak et al., 2001), spectral correction (Nordbo and Katul, 2012), turbulent mixing and stationarity (Foken and Wichura, 1996), detection limit (Billesbach, 2011), turbulent sampling error (Lenschow et al., 1994), footprint analysis (Kljun et al., 2015), storage flux term, and uncertainty budgeting.

Please note that e.g. Kljun et al. (2015) is itself published in GMD.

<u>Changes in the manuscript:</u> We clarify the objective of the paper in the introduction: ”The question we ask in this paper is: How do we collaboratively create portable, reproducible, open-source, scalable, and extensible software that improves reliability and comparability of eddy-covariance data products?”

We then added the following text later in Sect. 2: ”Thus, the DevOps model serves as the framework within which the scientific community can efficiently and robustly collaborate to produce, manage, and iterate community software. Through choosing appropriate tools to implement the DevOps workflow steps, the reproducibility, scalability and extensibility needs of software development communities (including EC) can be met.”

Lastly, we link the manuscript to the GMD literature realm and further clarify in Sect. 2.1: “Eddy-covariance data processing consists of employing a sequence of models. These often originate from scientific sub-fields with corresponding publications, and eddy4R provides an integrative, yet modular and extensible framework for their concerted application and continued development: eddy4R.base provides natural constants and basic functions for usability, regularization, transformation, lag-correction, aggregation and unit conversion ensuring consistency of internal units at any point in the workflow. Next, eddy4R.qaqc provides the general quality assurance and quality control (QA/QC) tests of Taylor and Loescher (2013), along the Smith et al. (2014) model for tracking quality information in large datasets, and functions for de-spiking (Brock, 1986; Fratini and Mauder, 2014; Mauder et al., 2013; Mauder and Foken, 2015; Metzger et al., 2012; Vickers and Mahrt, 1997). eddy4R.turb provides standard, Reynolds-decomposed turbulent flux calculation (Foken, 2017), accompanied by models for planar fit transformation (Wilczak et al., 2001) and spectral correction (Nordbo and Katul, 2012). Additional functionalities include Fourier transform, the determination of detection limit (Billesbach, 2011), integral length scales and statistical sampling errors (Lenschow et al., 1994), and flux-specific QA/QC models (Foken and Wichura, 1996; Vickers and Mahrt, 1997). Also, basic scaling variables, atmospheric stability and roughness length (Stull, 1988), as well as the flux footprint (Kljun et al., 2015; Kormann and Meixner, 2001; Metzger et al., 2012) can be determined. Lastly, edd4R.erf provides time-frequency de-composed flux processing and artificially intelligent functionality to determine an environmental response function model and project the flux fields underlying the EC observations (Metzger et al., 2013; Xu et al., 2017)."

References:

Billesbach, D. P.: Estimating uncertainties in individual eddy covariance flux measurements: A comparison of methods and a proposed new method, Agric. For. Meteorol., 151, 394-405, doi:10.1016/j.agrformet.2010.12.001, 2011.

Brock, F. V.: A nonlinear filter to remove impulse noise from meteorological data, J. Atmos. Oceanic Technol., 3, 51-58, doi:10.1175/1520-0426(1986)003<0051:anftri>2.0.co;2, 1986.

Foken, T., and Wichura, B.: Tools for quality assessment of surface-based flux measurements, Agric. For. Meteorol., 78, 83-105, doi:10.1016/0168-1923(95)02248-1, 1996.

Kljun, N., Calanca, P., Rotach, M. W., and Schmid, H. P.: A simple two-dimensional parameterisation for Flux Footprint Prediction (FFP), Geosci. Model Dev., 8, 3695-3713, doi:10.5194/gmd-8-3695-2015, 2015.

Lenschow, D. H., Mann, J., and Kristensen, L.: How long is long enough when measuring fluxes and other turbulence statistics?, J. Atmos. Oceanic Technol., 11, 661-673, doi:10.1175/1520-0426(1994)011<0661:HLILEW>2.0.CO;2, 1994.

Maronga, B., Gryschka, M., Heinze, R., Hoffmann, F., Kanani-Sühring, F., Keck, M., Ketelsen, K., Letzel, M. O., Sühring, M., and Raasch, S.: The Parallelized Large-Eddy Simulation Model (PALM) version 4.0 for atmospheric and oceanic flows: model formulation, recent developments, and future perspectives, Geosci. Model Dev., 8, 2515-2551, doi:10.5194/gmd-8-2515-2015, 2015.

Nordbo, A., and Katul, G.: A wavelet-based correction method for eddy-covariance high-frequency losses in scalar concentration measurements, Boundary Layer Meteorol., 146, 81-102, doi:10.1007/s10546-012-9759-9, 2012.

Smith, D. E., Metzger, S., and Taylor, J. R.: A transparent and transferable framework for tracking quality information in large datasets, PLoS One, 9, e112249, doi:10.1371/journal.pone.0112249, 2014.

Taylor, J. R., and Loescher, H. L.: Automated quality control methods for sensor data: A novel observatory approach, Biogeosciences, 10, 4957-4971, doi:10.5194/bg-10-4957-2013, 2013.

Wilczak, J. M., Oncley, S. P., and Stage, S. A.: Sonic anemometer tilt correction algorithms, Boundary Layer Meteorol., 99, 127-150, doi:10.1023/A:1018966204465, 2001.

*2) It is not apparent how exactly this technical set of workflows adopted by NEON can be useful for a broader scientist/modeler community and what scientific problems it can solve as the idea wraps around different open-source products dedicated essentially to crunching of eddy covariance measurement data. In the abstract, it is promised that the framework is applicable beyond EC but it is completely unclear how. Maybe one way to overcome this issue would be to make a strong connection to a modeling framework where measurement and model outputs are evaluated together or elucidate aspects where this data processing framework would add to novelty and usefulness for the broader GMD community*

Author reply: This suite of packages and Docker image is meant to provide a modularly extensible flux processing platform as foundation for modeling exercises (see reply above). The presented framework is motivated by a lack of collaborative coding and processing code development in the eddy-covariance community.

To address the Referee comment and demonstrate the ease of applied modelling, we prepared an executable example workflow that accompanies the revised manuscript and executes the QA/QC model by Smith et al. (2014).

The underlying functions are already included in the eddy4R.qaqc package. In addition, we highlight the extensibility that can be achieved with the modular packaging of the eddy4R-Docker framework: the eddy4R family of packages already includes the Environmental Response Function (ERF) model for flux upscaling to the landscape (manuscript Sect. 3.2.2), scheduled for future release.

Changes in the manuscript: We now introduce the executable example workflow in Sect. 2.6. Please see our reply to Referee #1, comment 17 for details.

References:

Smith, D. E., Metzger, S., and Taylor, J. R.: A transparent and transferable framework for tracking quality information in large datasets, PLoS One, 9, e112249, doi:10.1371/journal.pone.0112249, 2014.

*3) There are no clear scientific objectives of the paper and the title does not help either "eddy4R: A community-extensible processing, analysis and modeling framework for eddy-covariance data based on R, Git, Docker and HDF5". The use of "modeling framework" is misleading (see also comment 1) because the paper fails to present any modeling or prediction which could be achieved from this framework.*

Author reply: The aim of manuscript is to introduce the novel eddy4R-Docker software framework to address a methodological rather than scientific question: the portable, reproducible and extensible processing of eddy-covariance data. For this reason, the GMD journal was chosen, and test applications to three geoscientific use cases are provided in favor of a single in-depth scientific survey.

Based on the GMD manuscript types specifications, as well as existing papers from our community (e.g., Kljun et al, 2015; Maronga et al., 2015), we are under the impression that scientific hypothesis testing is not a typical component of a GMD model / framework description paper. On the other hand a core component of GMD model description papers is "…evaluation against standard benchmarks…" which is addressed in Sect. 3.3.

Changes in the manuscript: Please see reply to Referee comment 1) (above).

References:

Kljun, N., Calanca, P., Rotach, M. W., and Schmid, H. P.: A simple two-dimensional parameterisation for Flux Footprint Prediction (FFP), Geosci. Model Dev., 8, 3695-3713, doi:10.5194/gmd-8-3695-2015, 2015.

Maronga, B., Gryschka, M., Heinze, R., Hoffmann, F., Kanani-Sühring, F., Keck, M., Ketelsen, K., Letzel, M. O., Sühring, M., and Raasch, S.: The Parallelized Large-Eddy Simulation Model (PALM) version 4.0 for atmospheric and oceanic flows: model formulation, recent developments, and future perspectives, Geosci. Model Dev., 8, 2515-2551, doi:10.5194/gmd-8-2515-2015, 2015.

*4) The story basically presents a rather ambitious idea of automating data processing including quality control. The latter is not shown yet that it already works well so the product is not yet ready to be fully useful for the community. Once QC is implemented it could be interesting to see how it is done and how flexible the options are for the user. For instance, on page 14 L32 it is concluded "Once scientific QA/QC and uncertainty budget is implemented, the computational expense will likely increase by a factor of two to three. This suggests that eddy4R performs comparably to other flux processors." As presented, the value from another EC flux processor tool is unclear in where it would really help but what is interesting is that the development is directed to a modeling audience who might also be able to use this tool if it was better explained. However, without clearly stated goals and sufficient supporting material to assess its quality and usefulness, it is difficult to evaluate the code framework for all its ambitious features. The paper is incoherent in its presentation (e.g. different components, datasets are presented separately without a clear thread creating multiple fragmented methods and results) and in many places the quality is diverging from the standards of a scientific paper.*

Author reply: We agree that the implementation of the QA/QC framework substantially adds to the novelty of the methods included in this initial release of the eddy4R software. The QA/QC framework to deal with plausibility tests on the data is now fully implemented. Additional flux QA/QC tests are still be refined to accompany the full suite of eddy4R packages that are being released with the completion of NEON Construction. We hope to address the main concern by providing an example workflow accompanying the revised manuscript, which include the Taylor and Loescher (2013) high-frequency plausibility test model alongside the Smith et al. (2014) model for consolidating the results to a final quality flag. This highlights some of the capabilities of the eddy4R.base and eddy4R.qaqc packages, and provides a user-accessible and modifiable workflow template. Please also see our detailed QA/QC replies to Referee comment 6).

The different test applications are central to proving the flexibility of the eddy4R-Docker framework to process both tower and aircraft flux data. They demonstrate that the DevOps approach can be used in scientific software development.

Changes in the manuscript: We now introduce the executable example workflow in Sect. 2.6. Please see our reply to Referee #1, comment 17 for details.

References:

Smith, D. E., Metzger, S., and Taylor, J. R.: A transparent and transferable framework for tracking quality information in large datasets, PLoS One, 9, e112249, doi:10.1371/journal.pone.0112249, 2014.

Taylor, J. R., and Loescher, H. L.: Automated quality control methods for sensor data: A novel observatory approach, Biogeosciences, 10, 4957-4971, doi:10.5194/bg-10-4957-2013, 2013.

*Specific issues*

*5) The number of figures seems rather large and not all of them seem necessary. A heavy detail from different settings and configurations (e.g. Sect. 3.1.1, 3.2.1) could be nicely summarized in a table. The examples in Figures 9-13 require specific understanding of eddy covariance and do not help a modeler to adjust the framework for their needs. Even for the eddy covariance community, it might seem surprising that airborne and tower data can be automatically compared, because there is a comprehensive quality control that needs to be performed on these data and it is not easy to automate such EC comparison at different scales (e.g. Mahrt, 1998), at least without multiple user interactions. For example, in Sect. 3.3 "Validation and verification" it is stated: "eddy4R includes a verification script which automatically processes subsets of the tower and aircraft data introduced in Sect. 3.1 and Sect. 3.2, and verifies the results against a reference, e.g. generated with a different software." Where is this validation shown? Do you actually mean that you duplicate the processing (e.g. also with Eddy Pro) or just check selected files for consistency? The agreement in Figure 13 definitely seems surprising. It almost looks like the same dataset was plotted against the same dataset? The significant figures inconsistently range from 1 to 5. R2=1 is surprisingly good but not too meaningful (did you mean 1.0000, 0.9999 or 0.99)? I am also confused why the measured variables (e.g. w, q, CO2 mixing ratio) are compared with each other as they should have been the same unless the software interferes with the measurement data.*

Author reply:

We agree that Figure 3 can be removed without losing much information.

The intent of this paper has been clarified to demonstrate the applicability of the DevOps model to EC science code development (please see our reply to Referee comment 3 for details). One key attribute of the eddy4R-Docker methodology is its user-extensibility per requirements of the desired application. As such, no default workflow or settings exist that could be easily tabulated across applications. As mentioned by the Referee, specifics differ substantially e.g. among the tower and aircraft use cases. To demonstrate the complementarity of the eddy4R-provided functions and user-supplied workflow files, the corresponding workflows and settings are thus documented individually for each test case in Sects. 3.1, 3.2 and 3.3. The test applications in Figures 9 – 13 are central evidence to the claim of adjustability and expansibility.

We agree with the Referee that tower and aircraft data require careful QA/QC and interpretation, and in no part of these test applications airborne and tower data are automatically compared.

Rather, a comparison takes place as part of the DevOps **Verify** step: reference datasets generated with EddyPro have been stored and automated tests are performed prior to new code being incorporated. The results are shown in Sect. 3.3: calculations were performed independently at LI-COR (EddyPro) and U Wisconsin, with identical settings and based on the same input dataset as specified in the manuscript. Four significant digits were specified in the plotting routine for Figure 13, and the output of any uninformative zeroes is consequently suppressed. As discussed e.g. in Mauder et. al. (2008), discrepancies exist among software implementations also for the calculation of averages and variances, which we thus show alongside their corresponding fluxes.

Changes in the manuscript: Removed Figure 3.

References:

Mauder, M., Foken, T., Clement, R., Elbers, J. A., Eugster, W., Grunwald, T., Heusinkveld, B., and Kolle, O.: Quality control of CarboEurope flux data - Part 2: Inter-comparison of eddy-covariance software, Biogeosciences, 5, 451-462, 2008.

*6) The data quality control does not seem careful. For example, in Figure 9 the periods of latent heat and CO2 flux were not rejected when the friction velocities were at their minima which look definitely below 0.1 m/s on days 114, 115, 116, 119, and other. It is also unclear what the gaps correspond to (rejected data, power interruption). I would be surprised if it was not possible to choose an uninterrupted dataset in the NEON's large EC measurement network. It also seems weird in the same figure that the general temperature trend is anticorrelated with sensible heat flux (doy 115-118). Does the data output use normalized flux units for CO2?*

Author reply:

We absolutely agree that thorough QA/QC is paramount for evaluating scientific findings. This is especially true for measurement techniques involving substantial theoretical models and assumptions, such as eddy-covariance. However, the objective of this paper is to demonstrate the applicability of the DevOps model to scientific code development, and not the publication of scientific findings. Please refer to our responses to previous Referee comments for this clarification as part of the manuscript revision.

NEON sites are just beginning to collect eddy-covariance data, which are considered engineering-grade until Construction and Commissioning of the Observatory is complete. At this time, first, provisional NEON eddy-covariance data products produced by eddy4R-Docker are available for download on the NEON data portal (https://w3id.org/smetzger/Metzger-et-al_2017_eddy4R-Docker/portal/0.2.0). Analogous to the manuscript presentation, these data are accompanied by both, provisional quality and uncertainty summaries following the models of Salesky et al. (2012), Smith et al. (2014) and Taylor and Loescher (2013).

The results in this manuscript are shown identical to directly displaying the data downloadable from the NEON data portal, without sub-setting for quality or uncertainty. This highlights the need for the data user to determine and select the acceptable level of quality and uncertainty based on the particular use case (analogous to e.g. MODIS quality flags). During the remaining software development steps throughout NEON Construction, dedicated flux QA/QC metrics are being added to the already implemented plausibility tests. These are currently residing in the eddy4R.turb package, which is not released, and hence not applied here. Please also see our reply to Referee comment 4).

To address your last question about the temperature trend and sensible heat flux, we see the DOY 117 night was a very turbulent night with a lot of mixing resulting in less radiative cooling. This may explain some of the decorrelation.

Changes in the manuscript: In Sect. 3.1.2 we have clarified that the full flux QA/QC and uncertainty budget needs to be applied: "The spiky results preceding and following periods with >10% invalid data highlight the need for applying the full flux QA/QC and uncertainty budget to provide science-grade fluxes."

Many thanks for catching the mal-formatted unit of the $CO_2$ flux. We have corrected the figure axis label to µmol $m^{-2}\,s^{-1}$.

Additionally, please also see replies to above Referee comments for changes to the manuscript.

References:

Salesky, S., Chamecki, M., and Dias, N.: Estimating the random error in eddy-covariance based fluxes and other turbulence statistics: The filtering method, Boundary Layer Meteorol., 144, 113-135, doi:10.1007/s10546-012-9710-0, 2012.

Smith, D. E., Metzger, S., and Taylor, J. R.: A transparent and transferable framework for tracking quality information in large datasets, PLoS One, 9, e112249, doi:10.1371/journal.pone.0112249, 2014.

Taylor, J. R., and Loescher, H. L.: Automated quality control methods for sensor data: A novel observatory approach, Biogeosciences, 10, 4957-4971, doi:10.5194/bg-10-4957-2013, 2013.

*7) The results and discussion also do not focus on the science but rather on what the software can do before the QC/QA are implemented. The QC/QA are the most important component of* *any data processing, so I am a little bit shocked that this has not been done before the submission and only raw data are reported. It is all about QC/QA so if it does not work well, the whole infrastructure could be in vain. Was it not possible to wait until the QC steps are implemented? When will it happen?*

Author reply: Please see our replies to Referee comments 4) and 6).

Changes in the manuscript: Please see our replies to Referee comments 4) and 6).

*8) Sect. 3.2.1 "Algorithm setting and profiling". Can you define algorithm setting? There is no model algorithm here. By algorithm you probably mean the data processing routine which deals with technical issues of EC data handling such as "despiking of unphysical data". This and other similar sections can be confusing for the journal readers. It is also unclear what you mean by profiling in this context as it can also have different meanings (I suppose you meant vertical profiling of EC fluxes rather than algorithm profiling). The authors should be careful not to use ambiguous terms and define clearly what they mean by model, algorithm, and other terms where the meaning is not unambiguous in the modeling context.*

Author reply: An algorithm setting encompasses the selection of a specific model (e.g., Salesky et al., 2012) for a general application (e.g., random error calculation), alongside the corresponding parameter settings (e.g., number of low-pass filters). The eddy4R-Docker framework encompasses a multitude of these models, which are summarized in Sect. 2.1. The Sect. 3.2.1 in question describes these settings in detail for the aircraft test application, incl. literature references.

Changes in the manuscript: We now provide additional clarification in Sect. 2.1: "Eddy-covariance data processing consists of employing a sequence of model algorithms. These often originate from scientific sub-fields with corresponding publications, and eddy4R provides an integrative, yet modular and extensible framework for their concerted application and continued development:"

Given this clarification and that vertical soundings are nowhere mentioned in the manuscript, no changes were performed with regard to "algorithm profiling". The section heading in question reads "Algorithm settings and profiling": "algorithm" is the noun-modifier adjective to "profiling". That is, profiling is performed with regard to a piece of code, and profiling results are provided in the same section: "The analysis took 56 minutes with 8-fold parallelization and consumed <3 GB RAM thanks to the use of fast access file-backed objects."

References:

Salesky, S., Chamecki, M., and Dias, N.: Estimating the random error in eddy-covariance based fluxes and other turbulence statistics: The filtering method, Boundary Layer Meteorol., 144, 113-135, doi:10.1007/s10546-012-9710-0, 2012.

*9) The choice of example figures 10 and 11 is not optimal because they are not well explained or sufficiently informative for the story. The figures are described only superficially what they represent but are not interpreted scientifically. The blue areas represent high deposition of methane? I am not convinced it is fair to show these data without the discussion of uncertainties which are definitely different in airborne and ground fluxes. The results section 3.3.2 are less than a paragraph so it cannot be informative. This should be made more general or explained much better for general audience of GMD.*

Author reply: The intent of this paper has been clarified to demonstrate the applicability of the DevOps model to science code development. For this reason, test applications to three geoscientific use cases are provided in favor of a single in-depth scientific survey. Please see our responses to Referee comments 1) and 3) for additional detail. The figures in question represent the second test application: the analysis of airborne EC data which has been emphasized by Referee #1 as a highlight of the eddy4R-Docker development model. For this test application, uncertainties are discussed just below Figure 11: "Corresponding systematic and random statistical errors are calculated following Lenschow and Stankov (1986) and Lenschow et al. (1994), and the flux detection limit is calculated after Billesbach (2011)." The uncertainty results are then provided in Figure 12.

Changes in the manuscript: We have added explanation to the caption of Figure 10: "For each combination of aircraft position and eddy size, blue and red areas indicate transport toward and away from the surface, respectively."

*10) There are other issues which are uncommon to see in a peer-review paper. For example, on Page 7, L225-239 the information is shown as bullet points more like a web-based manual or technical report which almost feels like from a magazine advertising IT Systems. It is interesting, that not even one paper is cited from the GMD community, and majority of the references are authors' own papers published in specialist eddy covariance journals. I think it would make more sense to send the paper to one of those journals or make a better and balanced connection to the GMD literature realm. The EC data handling does seem promising but the approaches vary in various details among the groups so I found the author's EC method review particularly unbalanced.*

Author reply: The GMD instructions for "model description papers" require a "user manual"-like component. As such, list-style presentations are not untypical in the peer-reviewed literature, see e.g. Maronga et al. (2015) in GMD. Regarding the literature review and connection to GMD literature realm, please see our replies to Referee comment 1): out of 10 EC papers introduced there, only one is co-authored, while two others are GMD papers. Lastly, the intent of this paper has been clarified to demonstrate the applicability of the DevOps model to science code development. Manuscript Sect. 2 has been substantially expanded specifically to address this model framework aspect of GMD (please see replies to Referee #1 for details).

Changes in the manuscript: Please see our replies to Referee comment 1).

References:

Maronga, B., Gryschka, M., Heinze, R., Hoffmann, F., Kanani-Sühring, F., Keck, M., Ketelsen, K., Letzel, M. O., Sühring, M., and Raasch, S.: The Parallelized Large-Eddy Simulation Model (PALM) version 4.0 for atmospheric and oceanic flows: model formulation, recent developments, and future perspectives, Geosci. Model Dev., 8, 2515-2551, doi:10.5194/gmd-8-2515-2015, 2015.

*Overall, I like the general concept of the real-time data/model processing framework, but I expected the paper would be much more than just a teaser of an EC data- processing framework in progress. The revised paper should be guided by clearly defined science question(s) through a coherent story thread throughout the paper. If the intention is to publish the science in GMD, I would strongly recommend the authors to refocus the story on a solid connection between measurement data and modeling. One example could be a model-measurement testbed to validate models on observation data which could be very novel and useful for a larger audience including GMD.*

Author reply: Thank you for your thorough review. We hope that the proposed improvements by focusing on the DevOps approach for collaborative coding and including the QA/QC framework in the executable example workflow, in conjunction with our responses would sufficiently address your concerns.

Changes in the manuscript: Please see responses above for changes made to the manuscript.

*References:*

*Mahrt L. Flux sampling errors for aircraft and towers. Journal of Atmospheric and Oceanic technology. 1998 Apr;15(2):416-2*