Author reply to the comments by Anonymous Referee #3 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 #3 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.

This study present a radically new way to process eddy-covariance data. It combines R-coded EC software that are wrapped in a portable Docker image that can be used on various platforms. It is meant to be scalable and to make use of parallel processing of large quantities of data.

Author reply: Many thanks for this succinct summary.

Changes in the manuscript: No changes performed.

Major comments

In line with the other reviewers, I think that the paper currently lacks a clear scientific question. I could image that for GMD a clear description of a software environment would suffice, but this paper seems to describe "work in progress".

Author reply: As stated by the Referee, the aim of manuscript is to introduce the novel eddy4R-Docker software development model 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 three tests of geoscientific applications are provided in favor of a single in-depth scientific survey. One core component of GMD model description papers is "…evaluation against standard benchmarks…" which is addressed in Sect. 3.3. To demonstrate completion we created an executable example workflow accompanying the revised manuscript.

In addition, as detailed in the replies to Referee #1, we clarify 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. In doing so, we clarify that although the specific software implemented by this developmental model is a work in progress, the developmental model itself is complete and robust, as shown by the test applications.

Changes in the manuscript: Please see the detailed responses below as well as the responses to Referee #1 for changes made in the manuscript.

I am a big fan of Docker and directly downloaded the Docker image. I was disappointed in the fact that the image did not contain clear examples (e.g. the three examples outlined in the paper). I could see that the eddy4R.base and eddy4R.qaqc packages were part of the Docker image. I think it is a missed opportunity not to provide examples of (simple) data processing and plotting. Now the advantage of Docker images remains untraceable to the readers and remains rather theoretical.

Author reply: We could not agree more with the Referee in that an application example would add much value for the reader and potential user. For this reason, we created an executable example workflow accompanying the revised manuscript. It utilizes the functionality of both R-packages presented here, eddy4R.base and eddy4R.qaqc, and contains a user-extensible data read-in, processing and plotting workflow.

Changes in the manuscript: Sect. 2.6 now introduces the executable example workflow:

> "To demonstrate some basic capabilities and provide a template for potential eddy4R-Docker users, an executable example workflow and data are included in the eddy4R-Docker image. Once the eddy4R container is started, the example workflow, input data (NEON dp0p HDF5 file) and output data (NEON dp01 HDF5 file) are available from the Docker-internal directory `/home/eddy/`. The example workflow is located at `/home/eddy/flowExmp/flow.turb.tow.neon.exmp.dp01.R`, and provides a selection of the processing steps that yield the EC dp01 data on the NEON data portal (https://w3id.org/smetzger/Metzger-et-al_2017_eddy4R-Docker/portal/0.2.0). The example workflow is fully documented to guide readers through the various processing steps. These include data and metadata import from the input HDF5 file, data assignment to file-backed objects, processing of 1 minute and 30 minute data statistics and data quality, and writing the output HDF5 file. In addition, outputs from the quality flag and quality metric model are visualized."

For instance, the HDF5 section (2.4) is clear but a rather standard description that is available on internet (meta-data, directory structure, self-documenting). Again, this is a missed opportunity to guide users through an example (download raw data, process the data, and HDF5 output and visualization of results). You want to convince the "traditional ASCII" community.

Author reply: Agreed. The executable example workflow includes HDF5 read-in, write-out examples with attributed metadata to demonstrate the utility of having metadata attached to the data.

Changes in the manuscript: Please see our replies to the Referee comment above. Among others we demonstrate the utility of the HDF5 file format in the executable example workflow, and example HDF5 input and output files are already pre-compiled into the Docker image.

Section 2.5 presents the way NEON wants to deploy Docker images. Again, this remains rather high level, while the stated goal is to "empower the Science community at large by putting the key to the scientific algorithms into the hand of scientists". Again, a clear running example in a Docker container would convince these scientists more than a NEON brochure.

Author reply: We believe that this concern is addressed through the executable example workflow, which is described in Sect. 2.6.

Changes in the manuscript: No changes to Sect. 2.5.

Section 2.6 would be an ideal starting point for further "Docker-assisted" data analysis, but unfortunately stops at a reference to the eddy4R wiki pages.

Author reply: In response to the Referee comment, we introduce the executable example workflow incl. "Docker-assisted" data analysis in Sect. 2.6.

Changes in the manuscript: Please see our replies to the Referee comments above.

In section 5 there is a reference to the raw data, but again unfortunately no examples are given in which a Docker image automatically reads, processes, and presents results. In the remainder of the paper, three examples are given, which is basically fine, but without a traceable and "hands-on" exercise does not add much. It is (and should be) part of the standard software testing.

Author reply: We address this concern through providing the executable example workflow.

Changes in the manuscript: Please see our replies to the Referee comments above.

In summary, I very much like the concept presented in this paper. However, without more in depth possibilities for potential users of the software, the papers seems more suitable for internal documentation than convincing readers that this is a promising way for the community to process eddy covariance data.

Author reply: We thank the Referee for sharing the positive impression of the paper's concept. We agree that more in-depth possibilities for potential users of the software will help demonstrate the utility of the software development approach.

Changes in the manuscript: We have created an executable example workflow accompanying the revised manuscript.

Minor comments

Page 1: line 34: mention where the NEON site is and also where the aircraft data were collected.

Author reply: This information is provided as part of the test applications in Sects. 3.1 and 3.2.

Changes in the manuscript: Added "USA" for the aircraft test application in Sect. 3.2.

Page 1, line 38: "streaming generation of science-grade EC fluxes": please explain better what this means.

Author reply: Adjusted.

Changes in the manuscript: Changed to "…automated generation of science-grade EC fluxes…"

Page 6, line 185: current recent

Author reply: Adjusted.

Changes in the manuscript: Changed to "…most recent eddy4R source code…"

Page 6, Figure 3, introduced at line 191. This hardly adds anything. A link would do here. Also figure 4 and figure 7 seem illustrations that do not add much.

Author reply: We agree that Figure 3 can be removed without losing much information. The GMD instructions for "model description papers" require a "user manual"-like component: Figure 4 introduces the HDF5 format and structure used in this study and for NEON data portal downloads of EC data (https://w3id.org/smetzger/Metzger-et-al_2017_eddy4R-Docker/portal/0.2.0). Figure 7 presents the development environment user interface. As both of these are fairly new to the EC community we are under the impression that retaining Figure 4 and Figure 7 provides clarity for some readers.

Changes in the manuscript: Removed Figure 3.

Page 7, line 231: CI?

Author reply: Cyberinfrastructure, as introduced in Sect. 2.

Changes in the manuscript: No changes.