Geosci. Instrum. Method. Data Syst. Discuss., 3, C9–C12, 2013
www.geosci-instrum-method-data-syst-discuss.net/3/C9/2013/

Geoscientific Instrumentation Methods and Data Systems
Discussions

Interactive comment on "The AmeriFlux data activity and data system: an evolving collection of data management techniques, tools, products and services" *by* T. A. Boden et al.

M. Syrjäsuo (Referee)

mikko.syrjasuo@aalto.fi

Received and published: 25 March 2013

General comments:

The manuscript describes an implementation of a data management system for observational data from a number of Carbon flux measurement sites. The systematic approach by the authors is highly recommended to all scientists working with data and the treatment is of interest to the readers of GI. Especially, I found Table 2 a magnificent example of things that need to be considered and evaluated given raw data from stations.

Specific comments:

I have two overall questions and a small number of more detailed questions. A minor revision is sufficient to polish the text for a wider audience.

1) In my opinion, the text could be shortened and made more concise. Also, the introduction should include a clearer overview of the data flow. Perhaps, you could simply move the data life cycle (Fig 7) to the introduction. A table of inputs and output formats would be a useful summary, I leave it to the authors to decide whether such a table works better in the introduction or conclusions.

2) I think the Conclusions could be shortened and a separate short chapter dedicated for "personal experience" would benefit the science community. There are things that are of significant practical importance in the era of international networks and virtual observatories. The following topics are examples of questions to which there probably is no right or wrong answer. However, the authors' collected knowledge in "making it work" is highly valueable to the readers of GI, who are developing their own data systems.

I would very much like to know the number of person-years needed to keep the system operational, what infrastructure beyond the obvious do you need etc. More often than not, one finds a problem in the database architecture after a year or two of actual use, which then results in a complete rewrite to re-organise the tables. Did you encounter something like this and what was the cause of it?

From your point of view, what is the best method to input data from the sites: email, FTP, wget? How did you solve the data policy issues if you had any? How do you satisfy site investigators who are, of course, providing the data but would also like their own research efforts cited?

Which parts of the data flow absolutely require manual analysis and which parts might or should be automated in future? If you could start this from scratch, what would you

do differently and what practices would you recommend?

——

3) Page 4, line 14: The sentence beginning with "AmeriFlux data have improved understanding..." What would not be understood if there were no AmeriFlux?

4) Page 7, consistency of raw data: is it the responsibility of the site investigators take care of consistency of their data due to, e.g., change of instrumentation? So, the raw data is processed to remove systematic errors due to ageing etc.? Are the measurement data calibrated by the site teams – and thus traceable to NIST – and is this recorded somehow?

5) Page 11, line 13: quality flags - what kind of flags do you use? Do you have a single flag or are you using many different flags?

6) Page 11, line 19: Does Level 4 also include quality flags? Do you also indicate which values were "gap-filled" and, perhaps, provide error estimates for the interpolated values?

7) Page 14, lines 19-20: is there an API to the database that would allow others to access the MySQL, too? Access to data via FTP works for most but some people may want to extend the database operations and link their own database systems to the one at CDIAC. You mention that downloading data directly from the MySQL database is possible, but are database operations possible?

8) Table 2. Having a numeric value to represent missing values is a common but outdated and not-recommended practice. Are you working around a specific "feature" (of MySQL?) in the software? It would be a much better solution to use a non-numeric value, which can then be appropriately dealt with. Commonly used representation are, for example, NA in R, "." in SAS etc. which allows the statistical software handle the missing values as appropriate in given context.

9) Table 2. Do you obtain "reasonable" values from the site investigator teams or do

you simply use older data to determine the range?

10) Table 3. In the "YES"-boxes, what happens if the procedures fail? For example, what if a 15-day window is not sufficiently long? In other words, is it possible to have gaps in the "gap-filled" Level 2 or 4 data?

——

Technical corrections:

11) Page 6, line 22: please spell out what PAR stands for (photosynthetically active radiation?)

12) Page, 6: typo, this should probably read "based on new files' names"

Interactive comment on Geosci. Instrum. Method. Data Syst. Discuss., 3, 59, 2013.