Technical Note: Flagging inconsistencies in flux tower data

This is definitely a technical note, and as long as the editors of this journal are willing to consider it, I will give it due diligence as a referee. Otherwise, this document could be used as a white paper or grey literature to supplement processing on the ICOS or Fluxnet web sites.

I appreciate the utility of having a set of well defined and accepted flags for data by the community, as data has certain quality due to time and place. Sadly we are aware that many data users who are not involved in the details and rigors of the measurements, processing and interpretation often ignore these flags. But no harm in producing and providing them.

What is unique and distinct here is the production of a complementary set of consistency flags (C2F) for flux tower data, which rely on multiple indications of inconsistency among variables, along with a methodology to detect discontinuities in time series. I am a fan of multiple constraints, so I am willing to read the case before me.

As I read this work I have some disagreements with conditions which they may flag.

For example it is stated that most frequent flags were associated with photosynthesis and respiration during rain pulses under dry and hot conditions. I have spent a good part of my career studying such pulses. They are real and sustained following rain events. To remove them is faulty and will cause biases in sums. Yes, I concur during the rain event itself data such be flagged when sensors are wet. But following the rain, huge amounts of respiration can and will occur.

2 Materials and methods

As I read this paper back and forth, I wonder about the wisdom of its organization. In the Methods section there are 8 figures or so. This seems more like a Results and Discussion. I also suspect much of the excess material could be in an Appendix or Supplemental material. For a Technical Note, this paper is really long and excessive.

Section 2.2.1. It is important to benchmark and monitor the relation between PAR and Rg. It is our experience that quantum sensors tend to drift over time, if not frequently calibrated. PAR is used to upscale fluxes with remote sensing and if those relationships are built on faulty values of PAR, the derived products in time, space and trends will be in error. Hence, looking at these flags can have important implications. Too often this issue has been overlooked, so it would be important to know the consequences of improving these data.

The relation between Rn and Rg is important to examine, but realize it will change with season as albedo and surface temperature changes. So be careful and do not make your flags by using one annual dataset for the site.

Comparing GPP and Reco with Day vs Night partitioning methods may be interesting, but not sure which is right. We know there is down regulation in dark respiration during the day, it is hard to measure reliable CO2 fluxes at night under stable conditions and with tall vegetation and appreciable storage and or sloping terrain. These can be points of reference and maybe the daily sum is better than hour by hour measurements, as some errors cancel.

I have learned from Dario the value of plotting CO2 flux vs u* and developed a matlab subroutine to do so. The threshold can be uncertain as u* has some autocorrelation with the flux. Of course we don't

want to set high thresholds as they are based on a diminishing number of data points as high u* values are rare compared to low ones.

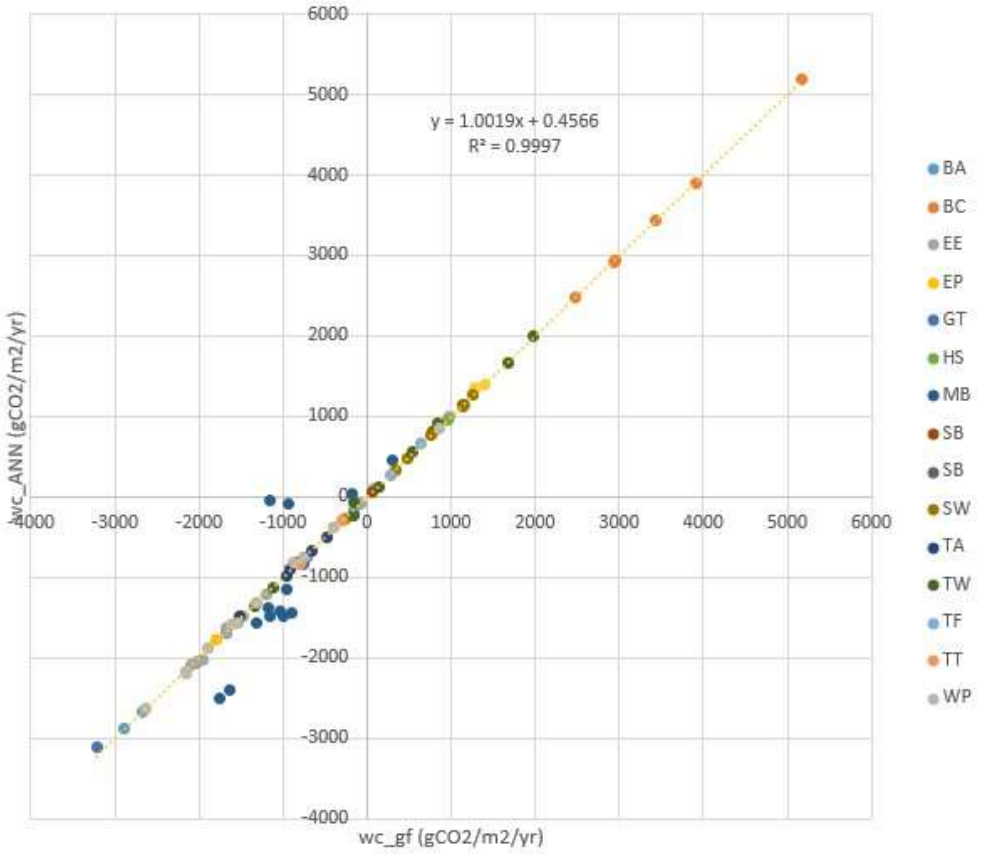
I see one of the constraints is LE +H vs Rn..What about G or storage in water column? I think these tests are only instructional. We know that there are many differences in sampling areas and representative of radiation and fluxes. It is dangerous to indict one or the other. And with wetlands it is really hard to measure water storage. We have a data set with nearly closed energy balance and then flooded the system and it all went to hell. Same sensors, same processing, ideal fetch and site. Just water moves heat in and out and it is hard as hell to sample well and well enough.

I must admit I am having a problem coming up with a salient point of this paper and how it will help me do better. I am at the point where an outlier score is proposed. It seems ok, but it is a lot like the college ratings, that depend upon an arbitrary set of metrics and scores.

I often advise use the set of sites that help you ask and answer the questions you are asking, relating to climate, function and structure. Just because these sites and data are in the fluxnet database does not mean we have to use them all. Maybe this should be the point of this paper.

Figure 1. It is a comparison between machine learning and flux data. Not sure what I am to learn and extract here. Which is right or wrong? Machine learning ultimately is a fancy least squared fit to a bunch of transfer and nodes.

Here is a set of data comparing annual carbon fluxes with machine learning methods from my sites. They are almost indistinguishable from the direct flux measurements they are derived from. In this case we know our site and develop the machine learning model with the most appropriate and representative biophysical forcings. In the figure given in this paper, I have no idea how appropriate the machine learning model may be for this situation, as the answer is based on independent variables they chose to use or omit.



Regarding the comparison of radiometers I know during some seasons our guy wires may shade the quantum sensor for certain angles of the sun. surely those data are not fit and I hope such a method may help detect these biases and errors.

Figure 2 seems to be a nice case study to show the attributes of your ideas. Maybe start with that one first. It is clear and more understandable, as we know PAR and Rg are closely related. So when there are differences it can help us think about why and which is more plausible and better.

Fig 3. Maybe I am just tired, or thick, but I don't follow the logic and rationale of the flag for light used for GPP. It would only give me pause on the accuracy of the machine learning calculations, but not the eddy fluxes.

Fig 5. I am trying to get my head around the issue of the comparison of the daytime vs nighttime methods. Again, I would argue one is better than the other. Personally, I like the idea of multiple constraints and see if the two methods are converging for confidence, more than anything. Not sure what you all are doing, but in early days working with Eva Falge, we estimated respiration during the day by the extrapolation of the CO2 flux vs light response curve. Now one of the limits is basing a regression and extrapolation on only a few points when the response function is linear, and the fact that during the sunrise sunset period steady state conditions don't hold. It is these reasons why I argue against one being better or worse, but if they both converge at least we may assume the fluxes may be good enough.

The reality is that pulses due to rain or insects passing through the path of the IRGA or sonic are problematic. Or those from electrical noise (a rarity today). We also see problems with CO2 fluxes over open water as there is a covariance with w and RSSI of the sensor that yields fluxes in the wrong direction and that are not physical. Those should be filtered. But I don't hear about that here.

Fig 6 seems to align with my suggestions that some sites may not be the best for some analyses and just toss them. Nothing lost as we oversample in many situations.

Fig 7. Curious as to why there is a systematic jump in LE. Eddy covariance should be immune from just a jump as we are doing mean removal. So even if sensors change and they are properly calibrated we should not expect such a marked difference. This is not like comparing two separate sensors, that can have offsets.

Fig 8. Illustration of the outlier score. This is needed to support the method described here. Has taken a long time to get to this point. Line 350!

Results

Fig9 demonstrates the point of this method. As expected met variable values tend to have few outliers.

Fig 10 provides a needed diagnostic as to when data may be rejected

Fig 11. Would think this would be a function of open vs closed path sensors

Fig 13. The jumps in NEE seem to be with site management. So Know Thy Site. Just don't blindly process long term data. This is why we have phenocams at our tower, to look at the vegetation when things are 'weird'.

Jumps in sensors can, will and do happen. This is why we make big efforts to write notes and log our sensor systems. Users have to remember Cavet Emptor and use the data wisely and when there are jumps look to reasons, and not mis interpret the data. Us data providers cant hand hold all users. They must do due diligence when using data too. Getting back to my point one should not use all the data. Use what is best and most fitting.

Fig14. Interesting

Discussion

Factors for potential false positive and false negative flagging

Glad to see something on this. But it leaves begging the point I make that respiration pulses are real.

Detection and interpretation of discontinuities in the time series

As I have mentioned, these are expected with long term sites as management can make changes..The site history needs to be considered too.

4.3.1 Flagged data points

I have already made my point about the danger of flagging rain pulses that are real. We have studied this with eddy fluxes, chambers, soil probes and they are consistent.

4.3.2 Flagged discontinuities in time series

It is reasonable to flag discontinuities, but aren't they flagged already?

Concluding points

I find this paper on the opaque side. It is a slog to read through, very engineering in spirit, style and narrative.

I must confess given the energy and time to write any paper, this is one I would not have spent writing.

I am missing the 'so what' message and being convinced I need to apply another set of flags to what I am already doing or what is being done in fluxnet, especially something that is automated and may not be applicable for the sites I may need in my synthesis.

The scoring method seems on the arbitrary side and reminds me of the scoring system for the 'best' world universities. Each scoring system yields a different ranking and group. I suspect this would apply to the application of this method, too.

I want to know how often this automated method suffers from type 2 errors, calling an error when there really isn't one.

I want to know how often this automated method suffers from type 2 errors, calling an error when there really isn't one. This concern also revolves around my complaints about flagging real respiration rain pulses. These pulses are real and sustained and should not be flagged (except for the period when the sensors are wet).

At this point I really feel it is up to the editor whether or not they are interested in publishing such a paper. My suspicion is that it may not be cited much, but again I may be wrong. As I look at the data from a different perspective being a data generator and knowing what to belief and accept as reasonable.