Biogeosciences Discuss., 9, C8813–C8817, 2013
www.biogeosciences-discuss.net/9/C8813/2013/

Biogeosciences
Discussions

Interactive comment on "Automated quality control methods for sensor data: a novel observatory approach" *by* J. R. Taylor and H. L. Loescher

J. R. Taylor and H. L. Loescher

jtaylor@neoninc.org

Received and published: 5 March 2013

The authors wish to thank Dario for his insightful comments and feedback on the manuscript entitled Automated Quality Control Methods for Sensor Data: A Novel Observatory Approach that is currently being considered for publication in Biogeosciences. To the best of our effort, we have endeavoured to address each of his suggestions. Below is a detailed account of how this was done. We welcome any further feedback that the editor and/or reviewer may have.

1) It would be better to define at the beginning (in section 2.1, or in table 2) that when you talk about mean and standard deviation you don't refer to the average and sigma

of the data but, for example for the sigma_test, the "standard deviation of the standard deviations" (page 18185, lines 5-6). This would improve the clearness.

Thank you, we have added the following sentence for clarification: "It, in essence, scrutinizes the 'standard deviation of the standard deviations.'"

2) I found the differences between null_test and gap_test not very clear; where is the threshold in number of gaps between the null and the gap tests? In addition it is not clear which action is suggested if a period is flagged for the null_test (that means that there are a number of single measurement points missing higher than expected): the whole period is flagged?

We appreciate that Dario would like more details about specific instances of how to apply some of these tests. The overarching goal of this paper is to identify general techniques that are useful for automated quality control of sensor-based measurements in large observatory networks. These techniques are defined in a general sense so that they can have broad applicability to many different measurements in multiple networks. We have intentionally avoided discussing numerous implementation details for specific sets of instruments as our hope was that the example provided would be sufficient for this. For this reason, we would like to not get into the details of whether a whole period should be flagged if a given number of null tests are failed, etc. It is our view that these decisions should be made on a case-by-case basis for the QC approach that is most suitable for the measurement of interest.

We have provided some clarity to the description of the null test and gap test that we hope will better distinguish them. The following sentence has been added: "The null test is intended to look for individual, missing data points within a given sampling period, while the gap test is meant to look for an extended period of missing data."

3) The previous point leads to a more general question: also the others test are based on analysis of "designed periods of time". This is quite arbitrary and could lead to different conclusions if period length is different also because the "flagging" is related

to the whole period. I suggest that more information and guidelines are given in the period definition; for example one option to explore could be that if a period doesn't pass one of the tests, it could be then analyzed using shorter periods in order to better identify the measurements points that create the problem.

Again, this is an implementation detail that we feel is best left to the reader's/user's discretion. For example, a different period of time and test parameters will need to be chosen for precipitation measurements than for those of temperature measurements. The overall quality assurance plan of various decision points and "what-if" statements will, necessarily, be specific to each network and set of observations. We respectfully disagree with Dario that this sort of detail should be included here.

4) The method proposed to calculate mean and standard deviation based on temporally and spatially adjacent observations is interesting (Eq. 4 and 5) but it presents two problems that should be better discussed: the arbitrary definition of the parameters z and t and the fact that in this way errors in the others sensors used are added on. I'm not sure that it is worthwhile to take this approach.

The variables "z" and "t" in the weighting factors of equations 4 and 5 are based on the formalism of Barnes Interpolation (as cited). Neither of these values is meant to be arbitrary, but rather they are defined by the respective temporal and spatial lag of the measurement in question. For example, a slow-changing variable, such as ambient temperature, can have considerable coherence across space and time, whereas variables related to turbulence change rapidly over very small regions of space. As with the details requested above, we feel that it is best left to the reader's/user's discretion to determine these values. We have added the following sentence to help clarify: "When considering the values of these parameters, it is necessary to assess the coherent structure of the measurement variable and assign appropriate spatio-temporal scales."

5) The Result section should be called something like "Test example"

We have modified this section to be entitled "Results and Test Examples"

6) One of the main point that I think should be changed in the paper is the example presented. With more than 10 years of micrometeorological measurements and a quite dense network, I would like to see the example based on a long term dataset (to test the sigma and delta tests, not applied now), with adjacent sites (to test the equations 3 and 4) and possibly with a second sensor that could be taken as reference "true" value to be used as validation. I don't think that it will be difficult to find such dataset (I can help if needed to find a site), for example one could do the analysis using sonic temperature and dataset to be quality checked and thermometer measurements as reference in a site cluster. The use of data coming from exclusively from a NEON tower is not relevant for the scope of the paper that wants to present a methodology.

We are struggling somewhat with this suggestion. We agree that it is important to not have the data or measurements used in this example be specific to NEON. This example was meant to show quality control for a very tangible and familiar measurement (temperature) that provided use-cases for most, but not all, of these automated plausibility tests. We considered the option of adding more examples with measurements that would require a different subset of plausibility tests (such as the micrometeorological measurements that Dario has suggested), but we thought that the current length of the discussion paper (35 pages) was already quite long and didn't think that the increased length would really benefit the reader. I think that conducting an exercise would be very useful, and we look forward to exploring this opportunity with Dario and others, but documenting it within this paper would only serve to make the manuscript less accessible. Should the editor feel otherwise, we would be happy to defer to his judgment in this matter.

7) I found not very clear the sentence at page 18190, lines6-8

In constructing a sampling distribution from a given population of data, it is necessary that the data be extracted in an independent, random fashion (in an abstract sense, we are "creating" a Gaussian Distributed sample of the data). This sentence is merely stating that if we were to choose specific values (such as daily maxima) then we would

be violating the requirement of random sampling, and the sampling distribution would not be representative. This is a subtle point and we can see how the reader may not interpret this upon first reading. We have clarified the sentence to read as follows: "It should be explicitly noted that daily extrema were not used in constructing these sample distributions as this would not allow the data be independent and randomly sampled, as required in the construction of sampling distributions (although, in practice, a sufficiently large volume of data would remove this restriction)."

8) My final comment is more a personal feeling I had reading the paper that however I want to share with the authors: in the paper you refer many times to NEON (cited 24 times in the text). It is important to describe what is NEON to give a background about why you are proposing these tools. However I think that these references and the sentences related to "what NEON will do" are too many, and I had to come back on the title of the manuscript while reading it to check if you were presenting a general method (as in the title) or NEON and his approach. In my opinion a paper less focused on what NEON planned to do (e.g. section 4.2) and more on the method proposed would be better and increase the clearness because more focused.

We very much appreciate this comment and were concerned about this point as we drafted the manuscript. Our intention is to present general methods that could be used at any network, but will be used at NEON. We have attempted to walk a fine line in how much detail is presented about NEON's approach versus how general these methods are. In contrast to Dario's view of how "NEON-centric" the paper appears, the first reviewer has asked for more details about how NEON will implement many of these tests. This places the authors in the difficult position of trying to balance contrasting expectations from the scientific community. We hope that, although NEON has been mentioned many times in the text, the methodology and examples have been presented in a general enough fashion so as to be applicable to a broad range of networks and measurements.

Interactive comment on Biogeosciences Discuss., 9, 18175, 2012.