Biogeosciences Discuss., 9, C8806–C8812, 2013
www.biogeosciences-discuss.net/9/C8806/2013/

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 the referee 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 the referee's suggestions. Below is a detailed account of how this was done. We welcome any further feedback that the editor and/or reviewer may have.

-Use of central limit theorem: Use of the central limit theorem really only works well at representing the distribution of a statistic if the nominal distribution is nearly normal -

which temperature statistics (max, min, variability) generally are. For time series which have statistics that exhibit bimodal or more complex distributions the resulting randomly sampled distributions will not accurately represent the limits of the true data. (An example can be seen at the web site (http://www.cs.uic.edu/ wilkinson/Applets/clt.html) if one selects the arcsine or binomial choices.) For many variables this will not be an issue, but for variable which can exhibit distinctly bimodal characteristics - for example precipitation and radiation and other closely associated factors - using the central limit theorem will result in distinctly un-representative parameter values.

This is an important point that the reviewer has made. We would like to further point out that the sampling distributions that have been suggested here are typically not applicable to the entire timeseries, but for a specific parameter statistic (such as daily maxima). In most of these cases, the distribution of these parameters (such as maxima) is, indeed, Gaussian. Of course, there are many measurements for which there is an insufficient amount of data to properly make these determinations, so this is indeed a potential risk. We have added a paragraph to the Discussion Section in which we explore the assumptions and limitations associated with the central limit theorem.

-Page 18184 lines 21-24: Cases such as the examples given are not 'indefinable'. It is more likely that their definition is trivial, such as a maximum wind direction of 360 degrees. A measured wind direction greater than 360 degrees would clearly indicate an error in either the instrumentation or algorithms used to determine wind direction, and as such would benefit from a plausibility maximum of 360 degrees.

Agreed, we have modified the wording to read: "It should also be noted that the application of the range test to some variables may not be statistically defined and/or quantitatively trivial (e.g., minimum wind speed or maximum wind direction) but can still have merit for detecting implausibility (i.e. a wind direction greater than 360 degrees is definitely implausible)."

-Page 18185 lines 11-13: The concept of multi-stage, or cross-referenced QC is

quite important. Statistical tests are not sufficient on their own and other comparative tests are required. For the reader, knowing what comparisons would be beneficial is paramount as it dictates what instrumentation needs to be deployed, and where it needs to be placed and how frequently it needs to be sampled in order to be useful for QC testing. It would be beneficial if this paper could present the physical concepts required for such cross-referenced QC and give a concrete example. Essentially, this means that sensors used for cross-reference QC must measure or be influenced by a common environmental property. For example a precipitation sensor could be used in QC of wetness sensors, humidity sensors, many radiation sensors, and soil moisture sensors to name a few. What readers need to know is how such QC relationships should be established.

We whole-heartedly agree with the reviewer and look forward to outlining all of our cross-instrument consistency tests in a rigorous fashion in a future publication. As stated in the introduction, this paper is the first in a series of papers that are anticipated to outline a number of quality assurance approaches that will be employed at NEON.

-Page 18186 lines 5-6: It is not clear how values are being flagged in the step test. Because the step test uses the differential of the time series a large step value will actually represent two values from the time series, one which is nominally good and an adjacent value which is nominal bad and separated from the good value by a relatively large change in value. Does the step test flag both the good and bad values from the original time series – which would mean that 2.5

The reviewer's comment appears to end rather abruptly, leaving the meaning of "2.5" somewhat ambiguous. We agree that some clarity should be added here. To address this, we have added the following: "However, for this test to be applied to paired datapoints in an automated fashion, it is simplest to flag both points, thereby resulting in more flags than the 2.5% would indicated. A subsequent processing of the flagged data (i.e. in the "data verification" stage of QC) could then help identify which of these flagged values is a distinct spike. However, if there is a sort of step-function change

in the mean of the timeseries, then additional verification will be required. It is for this reason that caution must be taken . . ."

-Page 18186 lines 10-24: The null/gap tests seem to be of a different nature than the preceding statistical tests. Although not clear from the text, it appears as if you are measuring the number of missing values in a statistical interval and assigning a single QC flag to that interval based on the result. In the preceding tests, I assume that the QC flags are applied directly to the measured time series values and not to any associated interval of derived statistics.

This is a direct vestige of NEON's approach to data products derived from time series. For most automated, sensor-based measurements, NEON will produce data products for 1 minute and 30 minute average descriptive statistics. The intent of this paper was not to have the tests be overly NEON-specific and, therefore, this degree of detail was left out (see the final comment by the second reviewer). To better address the reviewers comment, we have added the following: "In almost all cases, these two tests cannot be applied to a raw timeseries without defining a sampling period in which a known number of samples is expected"

-There also seems to be a logical fault to the null test. By enforcing a mu + 2 sigma upper limit on the number of missing values would imply that a time series which generally has very few missing values could result in statistical intervals consisting of primarily good data being flagged as bad because of a small number of missing values.

We appreciate the reviewers point here. The sensitivity of these tests is determined by the threshold of the parameter chosen. Similarly, the likelihood of false-positive errors will also be determined by this test. We think this should ultimately be left to the readers' discretion and have added some detail in the discussion section about this very point.

-Page 18188: The temporally and spatial weighted statistics defined by equations 4-7 would likely be relevant to derived statistics such as hourly means, daily maximum or

such. It is not clear how such weighting would be applied to time series collected at higher frequency (eg 10Hz or 0.1 Hz) as the weighting functions may only be relevant or much smaller spatial and temporal scales. A concrete set of examples are needed in order to show how such a weighted statistical test value could be employed with the measured NEON time series data.

The reviewer is arguing that the statistical approach outlined in equations 4-7 fails at very high frequency sampling rates. We agree that this is true. We also believe that this approach will fail over longer time scales, such as hundreds of years. We also will concede that there are other failure modes for these tests that are not described in this paper. In the discussion section, we have endeavored to explore the applicability of these tests to different observing systems and have included some of the limitations, but we do not think it is fruitful to provide drawn-out examples of failure modes.

-Page 18190 line23: Should '. . .5.2 x 106 data points.' Really be '. . . 5.2 x 106 data points.' ?

Yes, this should be an exponent as in scientific notation. The authors have made this correction in the text and will be sure to watch for this in the typesetting of the manuscript.

-Data driven approach: The authors promote a 'data driven approach' to quality control which minimizes the subjective nature of QC inherent in human driven approaches (page 18192 line21). This is a laudable goal but the example and caveats given in this paper suggest to this reviewer that we may still be some way from achieving such a goal. Specifically: the lower limit variance test was not applied, the Delta test was not applied, and an arbitrary gap test size was employed.

In preparing this example, we thought it would be illustrative to show that not all of the tests will always be applicable. Certain measurements require some tests, while others do not. 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, 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.

-Page 18193 line 3: Can the authors specify references for the 'other QC approaches' which employ 'numerous assumptions'

We have added a reference to Table 1, that highlight the different approaches employed at various other observing networks.

-Page 18194: The wording of this section is very in-definite. The use of the words: 'ideally', 'should', and 'potentially' give the impression that NEON doesn't really have a plan worked out for quality control of the data.

We appreciate the reviewer's point here. This is, again, a case of the authors erring on the side of caution in an effort to avoid the paper being overly NEON-centric. We have modified the wording to sound more certain.

-Page 18194 line 22: Will NEON make available non-quality controlled data to allow the community to test/apply their own quality measures?

Yes, every datum, quality-controlled and otherwise, will be freely available to the community. We think that the sentence: "As such, no data will ever be deleted and the raw data will be permanently maintained by NEON and freely available to interested data-users," clearly conveys this message.

-Page 18195 line 22-23: The sentence 'By adopting a uniform-basis for data QC, future network interoperability can be assured.' is somewhat of an overstatement. Simple uniformity to QC does not guarantee interoperability; certainly it is possible to have uniformly bad or inappropriate QC.

This is a very interesting perspective. We did not consider the failure mode of having "uniformly bad QC" during the drafting of this manuscript. We have attempted to clarify

this by rewording the sentence to read: “The future of network interoperability can only be enhanced when a well-planned, uniform approach to data QC is adopted.”

-Page 18202: Should the sigma test include tests for both “mu-2 sigma” and “mu+2 sigma” as opposed to the calculation specified in the table?

Yes, there are cases where either (or perhaps both) of these conditions would be of interest for plausibility testing. We have corrected this.

-Page 18209: The clipping of the large (but obviously good) PRT values in this figure emphasize that the blind application of ‘data driven’ QC tests has the potential be very inappropriate. It may be that the tests outlined in this paper to not scale well to the short time series used in this example. If so, then the authors need to use a longer time series example or better, to devise test methods that do scale to short or long time series, so that readers can benefit from this work.

We have expanded our discussion to address details surrounding these “false positive” test results in the aforementioned figure.

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