Interactive comment on "A robust data cleaning procedure for eddy covariance flux measurements" *by* Domenico Vitale et al.

Domenico Vitale et al.

domvit@unitus.it

Received and published: 8 November 2019

We thank Andrew Kowalski (AK hereafter) for constructive comments and suggestions that helped us revise and clarify key concepts and the way we approached them in this study.

In the following, our replies to the referee comments.

AK - *The authors present a procedure for the "cleaning" of raw data from eddy covariance flux towers. If effective, such a procedure will be of great value. The procedure described is founded partially in previous studies of quality control for flux tower data, and partially based on statistical analyses that can be considered novel in this context. I am no expert in statistical analyses, but the approach put forth seems to be valid and defensible. The paper is well written, although I offer some minor suggestions intended for its improvement. My main problem with the paper is that the evaluation of the proposed scheme is based on synthetic time series, and I am left wondering whether such "data" are representative in terms of the types of problems that can be encountered in real, field data. If there were a means of assessing this key question, then I would suggest this as a "major" revision issue. In any event, however, I believe that the manuscript merits publication.*

Authors' reply - We understand and share the concerns raised regarding the use of synthetic data. However, the use of Monte Carlo simulations is attractive and constitutes a powerful and objective tool for the evaluation of methods.

Simulations allow, in fact, to get a full control of (1) the time series dynamics (since the simulated autoregressive processes are stationary with a pre-fixed, reference correlation), (2) the presence of a specific source of systematic error, and (3) the uncertainty due to random error component. As a consequence, a proper evaluation of the bias effect caused by systematic errors and a more objective performance evaluation of the tests involved for their detection, becomes feasible. Such evaluations are difficult to achieve with real, EC field data because the reference "true" value is unknown, therefore it is not possible to properly quantificate the bias effect, and replicates are not available, making it difficult the evaluation of the uncertainty associated with the estimates (either covariances and test statistics) due to the random error component.

We are aware that the model generating the simulated data differs from the actual data generating process. On the other hand, it is not an easy task to develop a stochastic

simulation model so complex as to include all the sources of error, either systematic or random, that are present in real-word EC time-series data. Building on the principle that "all models are wrong but some models are useful", we opted for a "useful" simulation model, having a simple dynamics that can reproduce, at least in part, the behavior of real EC data. If a QC test exhibits high false-positive and/or false-negative rates in such simple scenarios, it is most unlikely it will work properly with real data. In this case, the test would likely be methodologically robust but unusable in practice, affected by over-inflated type-I and type-II errors.

Thus, we are also aware that the robustness of our proposal needs to be investigated further. This paper is a starting point aiming at presenting a new strategy and a new framework that could hopefully stimulate innovative research on more complex EC data simulation designs, as well as on new and more robust QC tests for EC data to be used in the context of our data cleaning procedure.

In the revised version of the paper we will add a more clear note of caution on the fact that the evaluation of QC tests was performed on simulated time series affected only by a limited number of sources of systematic and random error.

AK - Specific comments (by line number)

- 38 "An integral part of the EC?"

- 83 "classified into partially versus completely data-driven"

- 87 "Such a call"

- 93 sacrifice

- 98 "because it is extremely"

- 100 "false-positive rates"

- 105 "selected, the results of these tests"

- 135 "proposed set, and iv) results in"

- 142 "a set of tests"

- 200 "with some instruments, especially older models or often"

- 374 "With respect to"

- 404 LOESS 414 loess

- 485 "was run"

- 519 "n0 indicates the percentage"

- 536 "was run"

- 549 the proposed tests were more selective

- 553 identify the same issue, which is why the ModEr statements

- 573 change "Exemplary" to "Representative".

- 643 "Previous research (refs) has highlighted"

- 648 Figure 14: Maybe change "DoY" to "Day of Year". To what year does it refer? Also, change "Density" to "Probability density"

Authors' reply - We thank AK for the above suggestions and corrections that will be implemented in the new version of the paper.

AK - 82 *This statement strikes is as excessively bold and over-simplified. For example, a random error that were to consistently reduce the covariance at a frequency of 0.01Hz (by introducing random noise at that frequency) would cause flux underestimation for the corresponding eddies. It could well be that these frequencies matter greatly for (convective) eddy transport during daytime, and far less so at night. The result would be an overestimation of NEE over the long term.*

Authors' reply - We agree with the reviewer statement that the presence of a specific source of error as the one reported in the example, could attenuate covariances and, consequently, introduce long-term biases.

We shall clarify that the distinction between random and systematic error affecting half-hourly flux estimates is based on the effects of the source of error on the quantity of interest and is thus not strictly linked to its features. In other words, the question to answer is: is the source of error responsible to introduce bias or to increase the uncertainty of the quantity of interest (e.g. covariance)?

With this definition, if the presence of some source of error is responsible for attenuating flux covariance estimates, then the source of error is systematic, even if the error component is a noise term having similar characteristic to random data. If, instead, the presence of some source of error is responsible for increasing the uncertainty associated with flux covariance estimates (i.e. standard deviation) then the source of error is classified as random. We are aware that, in practice, it is difficult to distinguish between random and systematic errors because some source of error can have both a random and a systematic component, there are no reference values to evaluate the presence of bias, and there are not replicates to consistently quantificate the random uncertainty.

However, we consider the above-mentioned classification more suitable to characterize the multitude of errors affecting eddy-covariance time series. We will better clarify these concepts in the introductory section of the revised version of the manuscript.

AK - 200 *Table 1: if the terms "fraction of missing records (FMR)" and "longest gap duration (LGD)" are defined with the acronyms in parentheses as in this sentence, then it makes it easier for the reader to find them when they encounter them for the first time in the text.*

Authors' reply. Thanks for the suggestion, we agree and so we will introduce the definition and acronym in the text.

AK - 248 *Again, I find greatly exaggerated the claim the random errors due not affect covariances. This depends on the magnitudes of the signal/noise variances, and also on their durations during the averaging period.*

Authors' reply - As explained before, the "unavoidable random measurement error" we are referring is the one that, by definition, is not responsible for introducing bias in flux estimates. To avoid confusion, we call them simply random error and add a reference to the definition of random/systematic error described in the introductory section.

AK - 372 *This seems particularly risky for high towers in situations with strong convection. Maybe this could be stated explicitly rather than left vague as "further considerations".*

Authors' reply - We rephrase the sentence as suggested, however what is reported is the suggestion by Mahrt in his paper proposing the test.

AK - 385 *The post-dawn transition from a stable to a growing/convective boundary layer seems to be a moment that is particularly relevant here. Perhaps you could elaborate on this or at least prepare the user for its importance.*

Authors' reply - It is true, although the change of conditions happening in the post-dawn period doesn't lead automatically to the presence of outliers (as per our definition). Nonetheless, we will add this as an example.

AK - 465 *A key point: can you either justify this or evaluate its importance?*

Authors' reply - We understand the importance as also discussed in the first part of the replies. We will add a justification that is based on the fact that a proper validation can only be done by comparing the results with true reference values. This is not possible with real, EC field data because we don't know with certainty whether data are error-free or not. For this reason the performance evaluation is carried out via Monte Carlo studies where we have a full control of time series dynamics, error contribution and its bias effects on the quantity of interest. The use of real data gives an idea of the impact the proposed schemes has in cleaning EC data and can help to understand which is the main source of systematic error affecting EC flux variables.

AK - 470 *You are careful to describe the methods for the eddy covariance calculations. However, whereas the determination of the storage terms is not trivial, you provide no methods.*

Authors' reply - We will add more information about the storage calculation (the best available since not all the selected sites had profile measurements available). However it is important to note that the effect of the storage doesn't influence the QC procedure

that is based on the turbulent data time series. The storage term does have an effect on the computed NEE that is used to identify outliers, however the effect is minimal. We will add however an explicit sentence on the fact that the storage term must be correctly measured and calculated, including proper references.

AK - 484 *The meaning of the subscript (i=1,...,6) is not at all clear to me. What are the differences between the six cases?*

Authors' reply - We recognize that the six cases are not clearly defined (the definition is in different parts of the text). We will add a detailed description of the 6 scenarios in the new version of the paper.

AK - 497 *Figure 2 has no units on either axis. Also, whereas this figure has labels for every panel, many of the following figures do not (but should).*

Authors' reply - We will add labels on all the axes of all the figures. About the letters, we used them only when needed to refer to specific subplots in the text, however it is not a problem to add them in all the figures.

AK - 516 *This is hardly a stand-alone figure (i.e., a reader would need to search broadly through the text in order to understand it). One might easily suppose that LSR means "least-squares regression", and not a "low signal resolution" test. Again, the subscripts are not clear. What is the difference between i=1 and i=4 (both of which give phi values of 0.9). Do "A-R" and "Amp Res" refer to the same thing? Generally speaking, I think that the average reader would initially obtain no information from this figure, and would become bewildered after studying it for ten minutes.*

Authors' reply - We agree with the reviewer on the complexity of the figure. However, the plots illustrated herein are crucial as they summarize important information. In order to improve readability and clearness of the figure, we will use labels, uniform the acronyms in the figure and throughout the text and improve the caption. The figure will then be self explaining and easy to get in the main message.

AK - 529 *Does size 6000 mean 10 minutes of data?*

Authors' reply - Yes, in case of time series sampled at 10Hz scanned frequency. We will clarify this in the text.

AK - 531 *Which 15% were replaced? I don't see a change in variance in Figure 5 (whose panels are not labelled, but I refer to Figure 5f.*

Authors' reply - It is 15% of simulated data points randomly selected, and the change in variance is evident when compared to the uncontaminated time series, which is however not shown. It will be added as a subplot of Figure 5 in the revised version of the manuscript. We will also add labels to each panel and provide more details in the figure caption.

AK - 572 *What are the units of the values 0.0004 and 0.004?*

Authors' reply - These values, as slope coefficients of deterministic time trend contaminating the simulated stationary autoregressive processes are dimensionless. They quantify the change in the mean level, per unit of time.

In case of time series of length $n = 18000$, as in EC raw data sampled at 10Hz in an

average period of 30 minutes, a slope coefficient equal to 0.0004 (0.004) increases the mean level of 7.2 (72) units of the response variable, respectively (e.g. 7.2 ms^{-1} for vertical wind speed, 72 μmol/mol for CO_2 concentrations, 72 K for sonic temperature).

We will better explain their meaning in the revised version of the manuscript.

AK - 573 *What are the units of the values 0.025 and 0.25?*

Authors' reply - The units of the pre-fixed standard deviation values used to generate stochastic trends contaminating the simulated stationary autoregressive processes are dimensionless.

Stochastic trends were generated as cumulative sum of n independent random variables normally distributed with mean 0 and standard deviation (s) equal to 0.025 or 0.25. Consequently, the sum of such variables is Normal with mean 0 and standard deviation $\sigma = \sqrt{n} \cdot s$.

In case of time series of length $n = 18000$, as in EC raw data sampled at 10Hz in an average period of 30 minutes, s equal to 0.025 (0.25) generates an ensemble of stochastic trajectories, responsible for non-stationary conditions, whose standard deviation at $n = 18000$ is $\sigma = 3.35(33.5)$ units of the response variable, respectively (e.g. ms^{-1} for vertical wind speed, μmol/mol for CO_2 concentrations, K for sonic temperature).

We will better explain their meaning in the revised version of the manuscript.

AK - 603 *Figure 9 has no units. Also, understanding the meaning of FW96 and M98 would require the reader to have access to older papers that may not be readily available. When a measure is used to compare against a threshold, that variable should be explicitly defined (so as to make the figure stand alone).*

Authors' reply - We agree that figures should be stand alone and so we will review all labels and captions in order to give all the elements to the reader to understand the concept and meaning. Specifically on Figure 9, we will report in the caption that test statistics (x-axis) and correlation (y-axis) between simulated time series are dimensionless. As the meaning of the stationary tests and the choice of threshold values were introduced in Section 2.2.3, we will provide more details and report the reference to the Section 2.2.3 in the figure caption.

AK - 634 *The appearance of references (to the works of Cleveland, Lomb, and Scargle) in the Results section indicates poor structure. All of these tests should have been previously defined in the methods section.*

Authors' reply - The references mentioned by the reviewer are present in the introduction and material and methods sections, and were reported also in the Result section aiming at improving its clearness. However, we agree that this may reduce the readability of this section. We will remove them from the Results section in the revised version of the paper.

AK - 639 *The y-axis labels of Figure 13 are particularly weak. What is "c" to which NEE has been added prior to computing a logarithm? What is "D"? The term "Irregular" is nowhere precisely defined (although it appears to derive from the work of Lomb/Scargle). What are its units? Also, the choice of max/min for the y-axes of the LRT and SRT variables (likewise not clear from the figure) are poor and give no idea regarding the ranges of these variables.*

Authors' reply - Recognizing the length and complexity of the paper we removed some

text and explanations, which ultimately made the Figure unclear for the readers. We will review labels and caption of Figure 13. In particular:

- c is a constant added to flux variable to ensure the argument of log is positive;

- D, LRT, SRT and Irregular indicate the components of the STL decomposition, as described in detail in Section 2.3.1 (we will report the details also in the figure caption);

- units: log-transformed variables are dimensionless. In Figure 13, only original NEE time series and the Irregular component (after inverse log-transformation) are in μmol CO_2 m^{-2} s^{-1}.

- y-axis range: we will adjust it.

AK - 648 *Figure 14: Maybe change "DoY" to "Day of Year". To what year does it refer? Also, change "Density" to "Probability density".*

Authors' reply - Thank you for this comment. We will change the labels accordingly.