Atmos. Meas. Tech. Discuss., 7, C2541–C2547, 2014
www.atmos-meas-tech-discuss.net/7/C2541/2014/

Atmospheric Measurement Techniques Discussions

Interactive comment on "Technical note: Detecting outliers in satellite-based atmospheric measurements" *by* P. E. Sheese et al.

Anonymous Referee #1

Received and published: 9 September 2014

General comments:

This paper concerns the detection of unphysical outliers in satellite-based observations of chemical composition. While technical in nature, this topic is an important one for users of atmospheric composition data, and is little discussed in the literature. Therefore I find the subject matter appropriate for AMT.

It is clearly difficult to devise a consistent method which can be used for all atmospheric species at all latitudes/heights as the authors have attempted to do here. Also, the success of any method is quite difficult to quantify, as it is difficult to objectively determine which outliers are "unphysical", and the result of defective observations or retrievals, or just "natural" outliers, due to rare dynamical or chemical conditions.

I do however have serious concerns about the methods used by the authors, and the justifications for these methods. I doubt that the methods introduced here would be used by other satellite data users. I also fear that the authors' conclusion that the median absolute deviation (MAD) should not be used in outlier detection would actually hinder future attempts at outlier detection.

I strongly encourage the authors to revise the outlier detection method used, perhaps using some of the suggestions in the comments below, in order to try and find a simpler method. I also strongly encourage the authors to include more information and discussion about the potential sources of outliers in their data set, both natural and unphysical, in order to frame the difficulties and consequences of outlier detection.

Specific comments:

I am not convinced of the quality of the outlier detection procedure described in this work. The distribution fitting procedure, step 1, is likely to run into problems when trace gas distributions are non-Gaussian, especially for instance at low concentrations when distributions are often log-normal. Secondly, as it is applied to the full data set (within latitude bands), it fails to detect outliers for species with strong seasonal variability, when those outliers lie within the variability of the full year. Lastly, performing distribution fits for each species at each altitude seems a lot of computational expense for what actually amounts to little more than defining an "edge" of the distribution, especially when this edge is then arbitrary shifted by some amount based on the SD of the data. It is also not clear if the first step provides any added value compared to the second step, i.e., does it detect any outliers that wouldn't be detected by the second step?

The second step, based on the standard deviation (SD) of data within a 15 day window, is problematic in its use of the SD rather than the MAD. Outlier detection based on the MAD is generally regarded as superior to methods based on the SD (Leys et al., 2013). The authors claim that using the SD and MAD leads to different thresholds,

and that “using the standard deviation when there are extreme outliers can allow for the acceptance of data that should clearly be rejected” and that “using the MAD on asymmetrically distributed data can lead to the rejection of ‘good’ data”. However, these conclusions are based on quite arbitrary thresholds, which are clearly different for the SD and MAD-based techniques since the distribution of the data they are being applied to is clearly non-Gaussian. If the distribution was different, different conclusions might be found. The authors go on to state that “Since the distributions of ACE-FTS data across regions (altitude/latitude/season/local time) tend to be asymmetric, the screening process does not make use of the median absolute deviation.” Asymmetry of the distribution is clearly an obstacle, but not an unsurmountable one. For example, asymmetry can be dealt with by considering separately the median deviation on either side of the median (Rosenmai, 2013), using a so-called “double” or “two-sided” MAD. Furthermore, the authors later introduce a method based on the SD of the monthly data, which should have the very same problem if the data distribution is asymmetric, in addition to being highly sensitive to the outliers (necessitating the use of an iterative method). I have the feeling that the use of such a “double MAD” could quite improve the outlier detection method used in this paper, and remove the need for the iterative SD-based method.

The current title is very general, although the work involves a very specific instance, involving a single data set (ACE-FTS) and one (multistep) method. The title should be made more specific.

The introduction requires an explanation of what “anomalous data” is. This is a subtle issue which requires some clear description – in fact, the use of the word “outliers” needs to be introduced, since “outliers” refers generally to points which lay far from the center of the distribution, but such “outliers” need not necessarily be “wrong”, or faulty, just rare. As this paper seeks to remove outliers resulting from errors in the observations or retrievals, and not (I assume) outliers due to anomalous chemical or dynamical conditions in the atmosphere, some careful consideration and language is

needed to differentiate natural vs unnatural outliers.

The introduction is sparse in terms of referencing prior work concerning the use of probability distribution functions in regards to atmospheric measurements. I would suggest at least including citations to (Sparling, 2000), and (Lary and Lait, 2006).

Line 12: has been

Line 34: The technique doesn’t assume that the distribution is symmetric, but application of the MAD as an outlier detector may run into problems if the distribution is asymmetric.

Line 40: not clear why sampling bias is important. Sampling bias is an error in a mean due to insufficient sampling. The next sentence refers to the distribution may not be normal, but this is not clearly connected to sampling biases. More explanation is needed.

Line 41: There is no reason the atmosphere need be normally distributed, and it’s not clear why one would need to identify regions where it is.

Line 42: Very unclear what is meant here. Obviously, the atmosphere has some degree of predictability. But this has little to nothing to do with normality of distributions of atmospheric quantities.

Line 44: The use of “normal” is confusing, as it’s unclear whether it should mean “typical” or something connected to the “normal” or Gaussian distribution.

Line 70: this sentence concerns “flagging” vs. the option of removing them from the data set I assume? This is not perfectly clear.

Line 71: Again, some clearer description of what “good” and “bad” means is needed.

Line 78: This is repeating the discussion of the introduction, which, again, should not be a major hurdle to outlier detection.

Line 80: This is a strong statement... of course if someone is interested in polar processes, they will focus on measurements there, but the fact that different processes are important in different locations does not mean one should partition the data.

Line 81: maybe point out that this is for all latitudes.

Line 84: the use of "assuming a normal distribution" is somewhat misleading... 1.428*MAD will equal the SD in the case where the distribution is Gaussian, but you don't need to assume a Gaussian distribution to make use of 1.428*MAD for some purpose... I know you are not saying otherwise, but some care is needed with the language in order not to confuse non-experts.

Line 86: How does one know this data should be rejected?

Line 89: Clearly the temporal variability of H2O (and other species) in the Antarctic vortex leads to an asymmetric VMR distribution. Later you use a 15-day running mean, but in theory one could use a MAD on each month of data rather than the full data set? In this case, the median and "rejection" thresholds shown in Fig 1 would have an annual cycle.

Line 110: Defining a threshold in terms of the "expectation distribution" makes much less sense than doing so in terms of the relative pdf, since your threshold then depends on the number of measurements.

Line 112: The meaning of "1" is hard to understand, and seems totally arbitrary. If the threshold was in terms of the relative pdf, at least you could place the threshold at a 99% confidence interval (or 99.99%, as stated), for example.

Line 116: these statements must assume a typical value of N, unless I have totally misinterpreted the preceding description.

Line 121: What is the justification for separating sunrise and sunset measurements?

Line 125: What is the justification for using 3 Gaussians? In the cases shown, the

method produces a good fit, but it doesn't need to always be so good. As there is no physical justification for the distributions to be a mix of Gaussians, this method seems arbitrary, and the quality of fits should, in principle, get better as more Gaussians are included. Plus, the quality of the fits will depend on the presence/absence of outliers, unless the fitting procedure is robust.

Line 130: The pre-screening of the data deserves some more attention, before the description of the outlier detection procedure would be best.

Line 137-138: Which fit are these metrics for? The CH4 fit looks much worse than the NO2 fits.

Line 139: why are extreme events undersampled in ACE-FTS data?

Line 146: this is a disturbingly complicated method. It is not clear how this is different than simply adjusting the threshold value used. I would assume that use of a double MAD would be superior to this method.

Line 154: This is not really the fault of the expectation distribution, but actually of the method of applying the method to the full year of data at once. If one lumps all the data together, any method you use will only remove the outliers on the edge of the distribution, and if the species has strong cyclical variability, it will only find outliers at the extremes of the cycle.

Line 160: This iterative technique is exactly what the use of MAD can let one avoid. And the same problem MAD has with an asymmetric distribution is also affecting a SD-based outlier detection procedure, so it is not clear why the MAD is not used here. In fact, I would think that a temporally varying double-sided MAD filter could be all that is needed, making the fitting of Gaussians unnecessary.

Line 165: Extreme outliers sure, but even moderate outliers can affect the standard deviation.

Line 178: "Robust" has a specific meaning in statistics, meaning insensitive to outliers.

Thus, the use of “robust” in this sentence is potentially confusing.

Line 179: “keep the data as inliers”: the outliers are data also. Some more careful wording required.

Line 187: In addition to the identified instances, it appears as if the method is identifying a number of outliers for high altitude NO and CO at low concentrations which appear to be within the typical ranges of VMRs. The distribution of these species is almost certainly poorly approximated as Gaussian, perhaps log-normal would be a much better fit to the data distribution. It could be that the distribution fitting procedure is running into problems here.

Line 202: Again, “outliers” should be further qualified – “unphysical outliers” or such.

Line 223: More description of “screening the data based on the percent error” is needed, as this has little or nothing to do with the material presented in the manuscript.

References

Lary, D. J. and Lait, L.: Using probability distribution functions for satellite validation, IEEE Trans. Geosci. Remote Sens., 44(5), 1359–1366, doi:10.1109/TGRS.2005.860662, 2006.

Leys, C., Ley, C., Klein, O., Bernard, P. and Licata, L.: Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median, J. Exp. Soc. Psychol., 49(4), 764–766, doi:10.1016/j.jesp.2013.03.013, 2013.

Rosenmai, P.: Using the Median Absolute Deviation to Find Outliers, [online] Available from: http://eurekastatistics.com/using-the-median-absolute-deviation-to-find-outliers (Accessed 3 September 2014), 2013.

Sparling, L. C.: Statistical perspectives on stratospheric transport, Rev. Geophys., 38(3), 417, doi:10.1029/1999RG000070, 2000.

Interactive comment on Atmos. Meas. Tech. Discuss., 7, 8393, 2014.