Comments to : Betsy.Weatherhead@Colorado.edu June 30, 2017

Response to Reviewer

I thank the authors for taking the suggestions of the first round of reviews into consideration. While manuscript still does not describe actual application of the techniques to atmospheric data, the introductory material on merging of ozone and temperature measurements does add value to the manuscript. And the structure and readability have improved compared to the first version.

I believe however that there are some errors and inconsistencies with the formulas, calculations and quantitative results of the study. I urge the authors to triple check their calculations, addressing the comments listed below.

1. The factor of *12 added to Eqs 2, 3 and 5 is wrong: to convert months into years one should divide by 12, not multiply. But I would not advocate this: since the quantities in each equation are based on monthly means, and the value of n used to determine the scaling factor for the conversion to a confidence interval (1.96 or other) is the number of months, I think it's best to write these formulas a "number of months…" *Change made. The reviewer is right, much to my chagrine. I won't bother with explanations. The reviewer is right. The corrections, however, were different for each of the equations. EQ2 doesn't need any factor of 12, Equations 3 and 5 just need to be represented as number of months—as the reviewer suggested. The errors were introduced in our attempt to respond to the reviewer's first set of comments.*

2. Spot checking Table A1 turns up a number of errors. For example, using Eq 2 and the numbers provided, I calculate time periods (months or solar rotation cycles) to identify an offset of 0.0008 watts m-2 nm-1 for SOLSTICE-SIM of 5.9 and 6.6, not 5 and 6 as listed in the table. Also, the value of 5.8 years to identify a drift of 0.0001 watts m-2 nm-1 year-1 in the SOLSTICE-SIM timeseries is inconsistent with Figure 5, which suggests around 2.2 years. *We have redone Table 1 to be more clear on when we are using the magnitude of variability and autocorrelation of the raw data and when we are using the magnitude of variability and autocorrelation of the de-trended data. We have also included more significant digits to allow better comparisons. Thank you.*

3. Apparently, the calculations behind Figure 5 use different values for sigma and phi than used for the prior calculations in section 3, and used in Table A1. This is not explained, and is therefore extremely confusing. *Yes, thank you. we discsussed several times as co-authors whether to address this directly in the text. Some thought it was obvious, others thought it was distracting. We have now included the text: ,,"σ and φ are the magnitude of variability and autocorrelation, respectively, of the differenced monthly data once any existing trend is removed." The reason behind the different sigmas and phis is that residuals of a statistical fit (and their associated characteristics summarized by sigma and phi) are always different when one uses a different statistical model. We have also added more text to make clear which drift, percentage drift, sigma and phi we are using in each section and updated the reference on the Figure.*

4. It doesn't make sense to me why the "drift" in Eq 3 (and 5) should be used in units of yr-1, when the sigma and phi are based on monthly timeseries. Perhaps this explains why different sigma and phi values are used in the construction of Fig 5, these could be estimates of the

standard deviation and autocorrelation of the annual mean timeseries, but could this be reliably done with just 3 years of data? It seems better to use a "drift" in units of month-1, and the original values of sigma and phi. When I do this, I get values pretty close to what is shown in Fig 5, but a little larger. For example, here is the calculation the way I think it has been done in Fig 5, for a "drift" of 1*10-4:

(1.96*8.58e-5/(1e-4)*sqrt((1+0.58)/(1-0.58)))^(2/3) = 2.1994,

and the way I suggest it probably should be done:

(1.96*1.7e-4/(1e-4/12)*sqrt((1+0.89)/(1-0.89)))^(2/3)/12 = 2.5144

We are trying to stay consistent with the literature on this subject. One way to think of the confusion in units is that drift, as expressed in units per year, is being compared with variability, which can be described by variability of monthly, quarterly, or annual data. However the data are collected (monthly, quarterly or annual, the magnitude of variability and autocorrelation will change appropriately. But, the number of years to detect is strongly based on this ratio (variability to drift) and length of time. We agree with the reviewer that he /she presents an acceptable alternative way to express things.We have added text to show the values we use in each section and believe the results are now easily reproducible.

5. The abstract states "For relative drift to be identified within 0.1% yr-1 uncertainty, the overlap for these two satellites would need to be 2.6 years", while Fig 5 suggests rather that ~2.6 years is needed to identify a drift of 10%! This is a big difference. If the 10% value is correct, it has a pretty substantial practical implication for the study—it seems unlikely that any reasonable overlap period (of a small number of years) will be able to do much to constrain drifts of any but the most egregious magnitude.

We think the results are now more easily reproducible by inclusion of the values used and why. Part of the problem with matching the abstract information with the Figure 5, is that they are both working off of different "relative" numbers. The text is now more clear in each case. We are balancing the fact that as scientists, we'd like to understand the relative drift in our data, for instance the SOLSTIC data; in this mindset we want the drift relative to the mean of SOLSTIC. But overlap periods only allow us to get insight into the relative drift and the amount of information is based on the differences between the two datasets. As such, the results are based on the characteristics of the differenced data and we present results in Figure 5 based on the relative percentage of the differenced data. We strengthened the text to make this more clear. We do want this paper to be useful and thank the reviewer sincerely for his/her careful checking of our results. We conclude that the errors are all due to the variety of calculations performed and think that the strengthening of the text will make the paper more useful.

Some specific comments

P1, l27: Perhaps pedantic, but it's the satellite missions that overlap, not the satellites themselves. *Change is made.*

P1, l29: this seems to be a result from another study, not this study, so probably shouldn't be in the abstract. *We'd like to keep this one in. This was work done for this study—in fact this was the exact point of the collaboration-- and is directly important to futher application of this work.*

P1, l32: actually 6 months (5.9), see major comments. *We have double checked this and have corrected the text.*

P2, l48: another pedantic point: the missions should overlap, not the launches *Thank you. Change is made.*

P5, Fig 2: these temperatures must be for a specific altitude range? *We now make it more clear that MSU Channel 4 and AMSU Channel 9 and refer to the lower stratosphere: "Both channels are designed to observe the lower stratosphere."*

P5, Fig 2: the relevance of Fig 2 to this study is questionable. It shows differences between 2 merged datasets with the offset and drift removed. But the offset and drift estimation is specifically the theme of this paper! *We've now added text to clarify: "These plots show the variability in overlap is highly dependent on latitude, as is often the case with Earth observations."*

P6, l8: "our" as in the authors', or more generally? *"Our" has now been removed.*

P6, l20: what "model" is being discussed here? *Clarification has been made: "model" has been replaced with "climate model."*

P6, l32: why can only "small" problems be identified? *Thank you. The sentence has been changed to "… and <u>even</u> small problems can be identified."*

P7, l18: "respect" is a unique word choice, and I don't know exactly what the authors mean by it. Thank you. *We have changed "respect" to "appropriately incorporate" so the sentence now reads: In order to appropriately analyze satellite observations, it is necessary to understand and appropriately incorporate the available information on the pre-flight calibration of instruments and in-flight expected behavior.*

P8, l8: So many atmospheric measurements show variability at time scales longer than the annual cycle due to modes of internal variability (e.g., ENSO, NAO) or responses to external forcings (like solar variations!). So I'd be careful about implying that one year of data will "cover the full range". *We fully agree. We have adjusted the sentence to clarify the point: While Earth observations often require a minimum of a one year overlap to cover the full range of expected observations, such arbitrary criteria ignore longer timescale phenomena including ENSO and NAO, and are impractical for covering a full 11-year solar cycle in a planned overlap period.*

P9, l6: These 3 points are contained in a paragraph describing Fig 3: it would be nice of point 1. connected the pre-flight calibration estimates with what is shown in Fig 3. *We have now tied the text more directly to the figure. We have also made clear that the corrections that were*

understood were already made in the construction of the datasets. In each case, evaluation of how best to characterize the drift takes place. For the Solstice-SIM data overlap, we noted that the differences between the two sets of data, showed lower variability than the ratio of the data, indicating an offset could be modeled as an additive adjustment."

P9, l9: It's perhaps impossible in this case to know whether there is a true drift between the instruments, or a "multiplicative bias". This might be worth pointing out at some point. *Thank you, we had checked this out, but hadn't included it in the paper. We've added the following text:*

1. *There are jumps in the time series related to spacecraft and instrument anomalies. Significant events are identified in instrument and spacecraft housekeeping telemetry and changes in behavior before and after these events can be characterized and corrected in the timese. Expamples of these phenomena are seen if Figurue 3 where SOLSTICE experienced a failure of the mechanism that changes the entrance slit from the solar to the stellar mode on 27 January 2006. The slit was moved back into position for continuous solar observations but did not return to the exact same position so the optical path through the instrument changed and therefore disrupted the degradation corrections and the wavelength scale. Similarly, a spacecraft safe-hold event on 14 May 2007 caused the instruments to become very cold and significantly changed the SIM wavelength scale and perhaps the transmission properties of the instrument. The change in the SIM wavelength grid is apparent in the uncorrected data, but in Figure 3 the data are interpolated onto a standard mission-length wavelength scale and does not appear as a jump in this figure. The 2007 safe-hold event had little effect on the performance of the SOLSTICE. The jump associated with the 2006 SOLSTICE slit anomaly has also been corrected and the change in character seen SOLSTICE data at this time represents the best compromise over the full wavelength range of the instrument.*

P9, l19: Does SIM also show a jump here? *We now address this point in the text: The change in the SIM wavelength grid is apparent in the uncorrected data, but in Figure 3 the data are inter"polated onto a standard mission-length wavelength scale and does not appear as a jump in this figure."*

P11, Fig 4: Is this the same data as in Fig 3, just shown now in monthly means? The wavelength of the measurements should be mentioned. *We have now made clear that Figure 3 shows the same data. Wavelength is now mentioned in abstract, Figure 3 caption and four other locations..*

P12, l7: Actually, depending on how sampling is dealt with, monthly means can exacerbate sampling differences between instruments compared to shorter-term averages. See, e.g., Toohey et al., 2013. *Thank you, I spoke with Dr. Toohey at some length on this general issue of monthly means and we jointly constructed the sentence, " Monthly averages can remove higher frequency noise and some sampling match-up problems, but they can also obscure important details and can often introduce their own biases, especially when sampling is irregular in time or space (Toohey et al., 2013). " I want to thank the reviewer for pointing us in the direction of both that paper and the equally valuable, Toohey and von Clarmann, 2013 paper.*

P12, l26: actually 6 months. *We get 5.2 months. I think the difference between 5.2 and 5.9 is*

due to the rounding errors of us reporting sigma and phi to only two significant digits. Partly for this reason, we have updated the table in the appendix to include more significant digits.

P13: l29: This sentence, which includes Eq 3, is not grammatically correct. *Fixed. Thank you.*

P14, l4: grammatical issue with sentence 2 of Figure caption. *Fixed. Thank you.*