Manuscript accepted for publication in J.Geophys.Res. doi:10.1002/jgra.50461(July 2013)

Simpson's paradox in trend analysis: An example from El Leoncito airglow data

Jürgen Scheer and Esteban R. Reisin,
Instituto de Astronomía y Física del Espacio
CONICET-UBA, Buenos Aires, Argentina

Corresponding author: J. Scheer, Instituto de Astronomía y Física del Espacio,
Ciudad Universitaria, C.C. 67 Suc. 28, 1428 Buenos Aires, Argentina (jurgen@iafe.uba.ar)

abstract

We use our mesopause region temperature data from El Leoncito (31.8°S, 69.3°W) to illustrate how the increased length of the dataset alone does not simplify trend analysis. This is because the adequate interpretation of trend results does not only depend on the statistical characteristics of the data time-series. A longer dataset may make unexpected features stand out, which require an explanation before definite conclusions on long-term trends can be drawn. While the rotational temperatures derived at El Leoncito from the OH(6-2) airglow band appear rather homogeneous at first sight, the O_2 temperatures measured with the same instrument and optical filter exhibit features strongly reminiscent of Simpson's classical statistical paradox, in that straight-forward trends derived from parts show signs opposite to those of the complete data set. The resolution of this paradox requires more efforts to diagnose and remove the impact of instrumental artifacts besides taking any other geophysical variation that does not directly contribute to long-term change into account. Intercomparison with other instruments is certainly useful, but may warrant the elimination of new uncertainties discovered in the act.

1. Introduction

It is a common notion that atmospheric trend analysis requires long datasets in order to average over the quasi-periodic fluctuations of shorter duration. To determine the minimum data length required, it is popular to cite the formulas by *Tiao et al.* [1990] and *Weatherhead et al.* [1998, 2002]. These formulas were originally designed for time series of monthly mean ozone concentrations and predict the number of years necessary to detect a given linear trend, if the standard deviation of the month-to-month variations and the autocorrelation, i.e. the quasi-deterministic relationship between the monthly data (expressed as the lag term in an autoregressive model of order one), are known. As usual with any statistical results, the conditions for their strict validity are idealized, but, while generally plausible, they are hard to ascertain precisely in any practical situation.

Especially, the formulas mentioned provide no way to distinguish atmospheric trends from instrumental drift. The absence of such a drift seems to be one of the "reasonable expectations" alluded to by *Weatherhead et al.* [2002] with regard to their formula 3. It may be possible to reduce instrumental drift effects if many similar and well intercalibrated instruments are used, so that the absence of drift becomes a resonable expectation. However, in general, the situation is likely to be more complex. It is generally assumed that longer datasets automatically lead to better trend results. However, this assumption is questionable on two grounds. From the theoretical point of view, it can be argued that the climatologically quiet time scale where the basic statistical assumptions become true is possibly beyond the expected period range [*Lovejoy*, 2013]. On the other hand, unexpected instrumental artifacts, in addition to other geophysical variations, may show up in longer datasets, which need explanations and eventually corrections before better trend results can be derived from the data. It is this second case which we will here focus on.

So, the purpose of this paper is not to present new trend results, but to discuss methodological difficulties inherent in trend analysis with examples from a real data set. We discuss the mesopause region temperature data from the El Leoncito site in Argentina and some technical issues at somewhat more detail than usually done in the literature, to illustrate our point.

2. Data

The dataset that we will use here as an example is not thought to be particularly fraught with difficulties for trend analysis, and parts of it have in fact been used for this purpose successfully, as we believe, in the past [*Reisin and Scheer*, 2002]. It consists of rotational temperatures from the OH(6-2) Meinel and the O_2b(0-1) Atmospheric airglow bands, corresponding to nominal altitudes of 87 km and 95 km, respectively. An advantage of the retrieval of rotational temperatures is that it only depends on the measurement of intensity ratios, but not on absolute intensities. The data are acquired with the tilting filter spectrometer designed specifically for long-term monitoring of atmospheric dynamics in the mesopause region [*Scheer*, 1987; *Scheer and Reisin*, 1990, 2001]. The small number of optical components (one interference filter, one lens) and moving parts (one filter mount tiltable between 0° and 30°) makes this conceptually simple instrument an instructive example for the present purpose. Although parts of the instrumentation have been improved and modernized several times in the past, the optics and hardware configuration as well as the control logic are still essentially the same. We here only consider the temperatures (but not the airglow intensities) measured from the astronomical observatory "Complejo Astronómico El Leoncito" (31.8°S, 69.3°W) since the beginning of automatic operation of the airglow instrument in 1998, until the end of 2011. For brevity, we will refer below to these temperatures as the "LEO" dataset.

Note that both airglow emissions and even the spectral background are observed with, and therefore depend on, only one filter (the inclination of which defines wavelength). The same filter has been used since 1996 (but had been acquired seven years earlier, so that potential aging effects should already have diminished). After 2002, there were practically no observations until 2006, when the instrument was again deployed after refurbishment, with a new photomultiplier, and after a new spectral calibration. The new calibration was mainly needed to reestablish the relation between motor steps and filter tilt angle (and therefore, peak wavelength) after repairs in the tilt mechanism, but also to take any other eventual changes in spectral characteristics into account. The unavoidable systematic errors in all the instrument parameters determined in the calibration lead to a final uncertainty in the derived rotational temperatures. This uncertainty has been calculated based on plausible estimates of the different contributing factors to be ±1.3 K for OH, and ±1.5 K for O_2 temperatures [*Reisin*, 1987]. These systematic errors are only approximate, but nominally still applicable today. While they are nearly negligible for most practical purposes, their combination is expected to cause an unknown discontinuity between the data before and after the data gap, which may easily amount to several kelvins. More details about the data acquisition and temperature retrieval will be given in section 5.

From 1998 to 2002, data were acquired during approximately 200 nights per year, with data gaps distributed more or less randomly. Since 2006, annual coverage rose from about 300 to more than 350 nights per year. Therefore, for both time spans, annual means can be expected to be approximately unaffected by the seasonal variation (which, at any rate, is small: only about 15 K peak to peak [*Scheer et al.*, 2005]). Since each annual mean is based on between 64,000 and 114,000 individual, statistically independent temperature samples (from each of the two airglow emissions), the noise content (standard error) of each annual mean is definitely not greater than a fraction of 0.1 K, and therefore completely negligible compared to interannual variability. This means that we can simplify arguments by dealing directly with annual mean temperatures.

3. Some Questionable Results

3.1 OH Temperature Trend

The annual means of OH temperatures suggest a negative trend, at first sight (see Figure 1). Indeed, a regression line with slope -2.1 (±0.6) K/decade is consistent with the data. According to formula 3 by *Weatherhead et al.* [2002], 7.8 years of data should be enough to define this trend "at a 90% significance level" (although this means that the uncertainty of the slope would be considerable), and one might think that with our data span of 13 years, we are on the safe side.

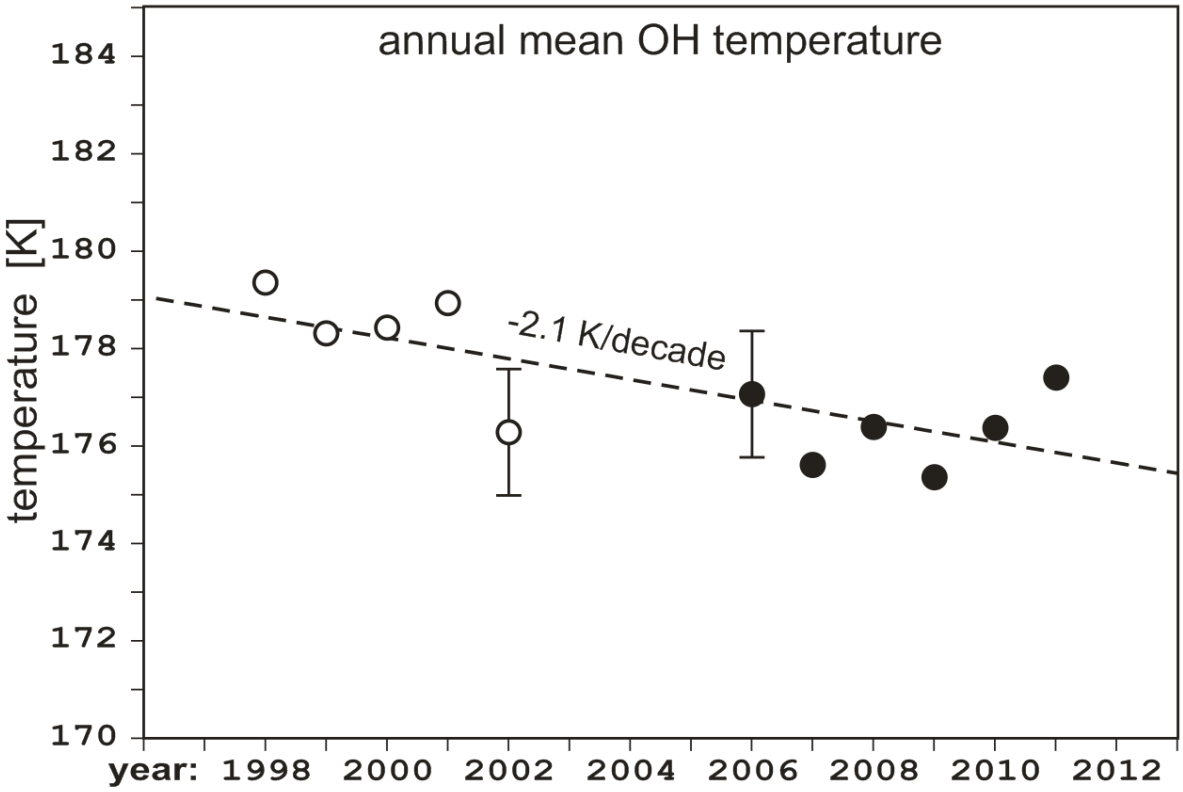

Figure 1. Annual means of ground-based OH rotational temperatures at El Leoncito (open circles for first group, filled circles for second group) and fitted trend (dashed line). The error bars represent the systematic instrument errors, independently for each data group (see text).

We could tentatively regard this as a valid result, under the following simplifying assumptions:
1. that there is no solar cycle effect. This is plausible, because of our previous findings for OH temperature using the double peak structure of solar cycle 23 [*Scheer et al.*, 2005].
2. that the deviations from the regression fit are random-like fluctuations. Deviations for each year are smaller than ±1.5 K, which is attributable to unspecified interannual variations.
3. that the systematic offset between the two data groups can be ignored. However, this is a point for which we have no evidence.

The size of the trend is in the range of literature values (and much smaller than our previous results based on earlier data, from 1986 to 2001, see *Reisin and Scheer* [2002]), which might add to a temptation to regard this as a plausible new result. However, the major weakness of this "trend" lies in its dependence on the unknown offset, i.e. the impossibility to ascertain assumption three. In addition, on closer inspection, we note that the mean slope of the data group from 1998 to 2002 is different from that of the group from 2006 to 2011. While the first group suggests a trend of -5.5 (±2.9) K/decade, the second group has a slightly positive mean slope of +0.8 (±2.1) K/decade.

3.2 O_2 Temperature Trend

The annual mean O_2 temperatures might be expected to be equally suitable for such a "simplified" trend analysis as the OH temperatures, since they depend on the same optical components of the same instrument and have been acquired virtually simultaneously. However, the corresponding figure suggests a problem (Figure 2). The negative trend is nearly twice that for OH, but the big scatter of the data make this hardly significant. On the other hand, the two data groups each signal an even much stronger positive trend, namely +9.0 K/decade, from 1998 to 2002, and +12.5 K/decade, from 2006 to 2011. Such a situation, when two groups of data show a statistical behaviour with one sign, but the complete data set shows the opposite behaviour has occasionally been encountered in statistical surveys [*Simpson*, 1951], and is known as "Simpson's paradox". We shall return to this topic, below.

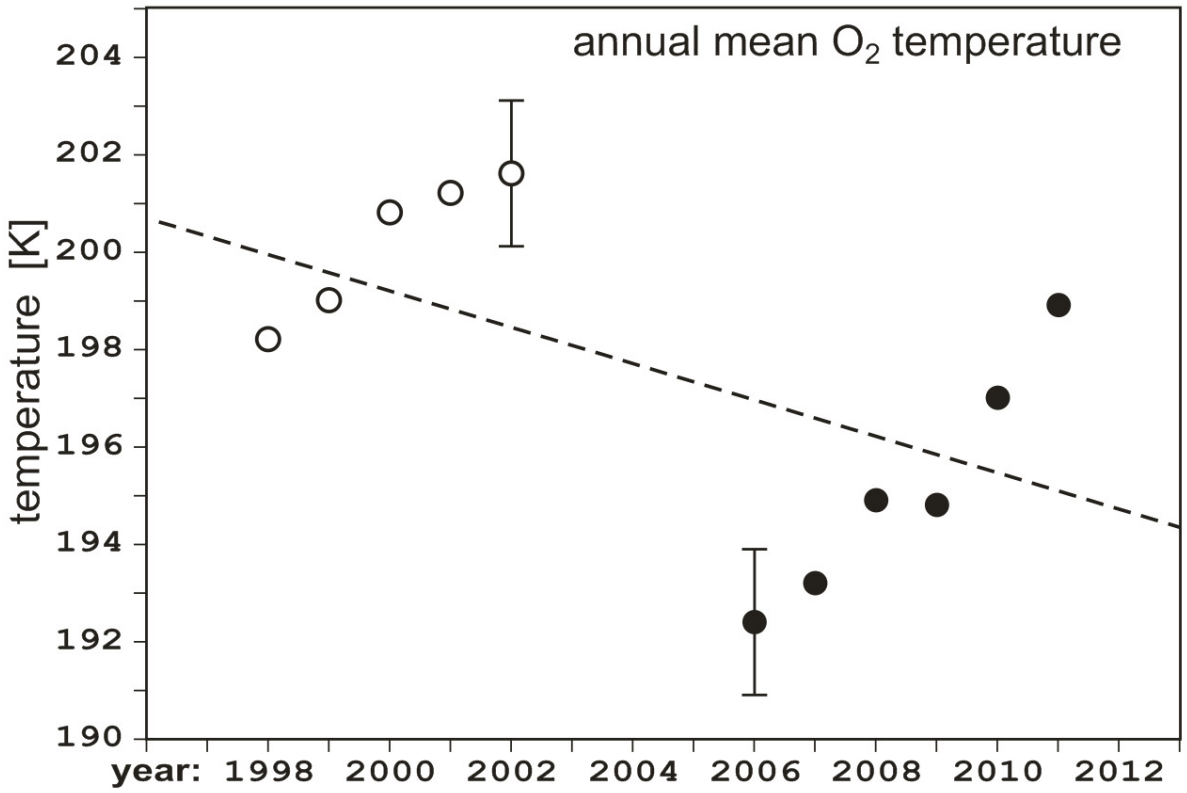

Figure 2. Annual means of ground-based O_2 rotational temperatures at El Leoncito (circles) and fitted trend (dashed line). Symbols and error bars as in Figure 1.

We must also resist the temptation to use multi-regression analysis to fit not only a linear trend, but also the supposed discontinuity created by the instrument modification. Such an exercise would make us believe that either, the more recent data group should be raised by 12.5 (±0.4) K, or the older data group lowered by 15.7 (±0.3) K, based on the assumption that either the strong positive trend of 1998 to 2002, or the one of 2006 to 2011, respectively, be true. Of course, both alternatives cannot hold simultaneously, and there is no hint at which one we should prefer. At any rate, without solid additional evidence, none may appear convincing.

3.3 Solar Cycle Effect on O_2 Temperature

The possibilities of numerical exercises are not yet exhausted (if they ever could be), because previous results for O_2 temperature [*Scheer et al.*, 2005; see also *Beig et al.*, 2008; *Beig*, 2011] suggest there might be a solar cycle effect, even without a trend. And indeed, we can fit a sinusoid to the data, which turns out to have a period of 12.2 years and an amplitude of 4.15 K (see Figure 3, panel a). The period looks "attractive", because it is not too far from the 11 years of a typical solar activity cycle. The amplitude is not outside the range of literature values, although about 70% greater than our previous result for O_2 temperature [*Scheer et al.*, 2005]. The maximum near 2001

is also not at odds with the solar activity of cycle 23 with its two peaks in mid-2000 and in early 2002. However, the activity minimum was not in mid-2007, as the fit suggests, but in late 2008 and early 2009 (Figure 3, panel b), and the present solar maximum is weak and its exact timing still uncertain. So, for sinusoidal fits in search of solar cycle effects, this seems to be a particularly unfortunate epoch.

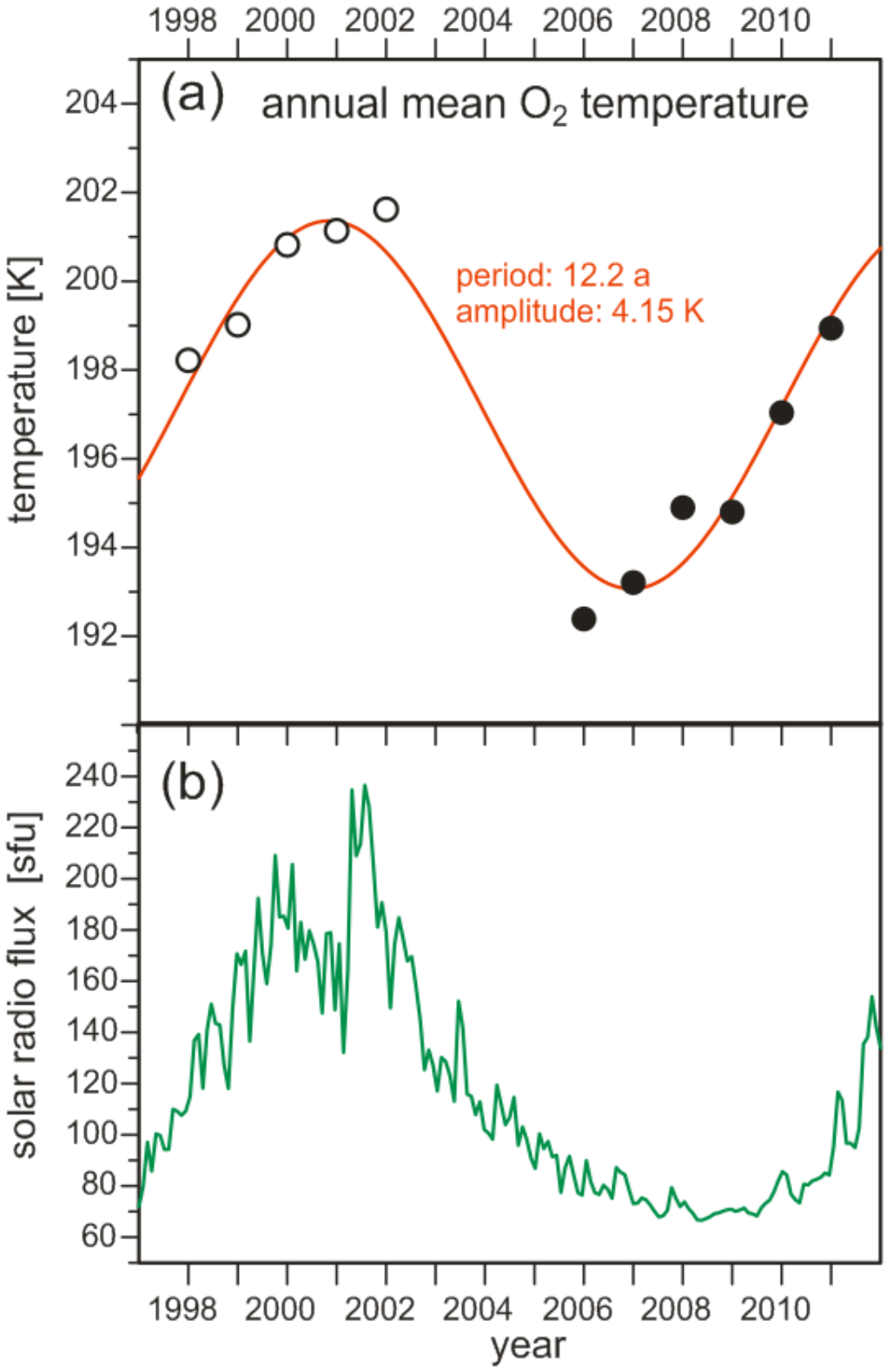

Figure 3. El Leoncito O_2 temperatures (circles) and least-squares fit of sinusoid (a); monthly means of Penticton observed F10.7 cm solar flux (b).

In situations like this, the scatter plot technique which we have used in the past [*Scheer et al.*, 2005] should be expected to be more appropriate. However, the results of such an approach do not lead to

a definite answer (Figure 4). The overall regression line with a slope of 6.2 (±0.9) K/100 sfu is faithful to the annual means of the first data group, where all residuals are smaller than 0.8 K. On the other hand, the distribution of points in the second group is clearly unsuitable for a linear fit (with most residuals between 1.4 K and 2.7 K). It is obvious from Figure 4 that this situation cannot be improved by any temperature offset between both groups. Nor does the inclusion of temporal trends help, as numerical experiments have shown. The introduction of time lags (i.e. a delayed atmospheric response) would even increase the discrepancies. So, although both approaches suggest the presence of a solar cycle effect, its contribution cannot presently be quantified with resonable confidence.

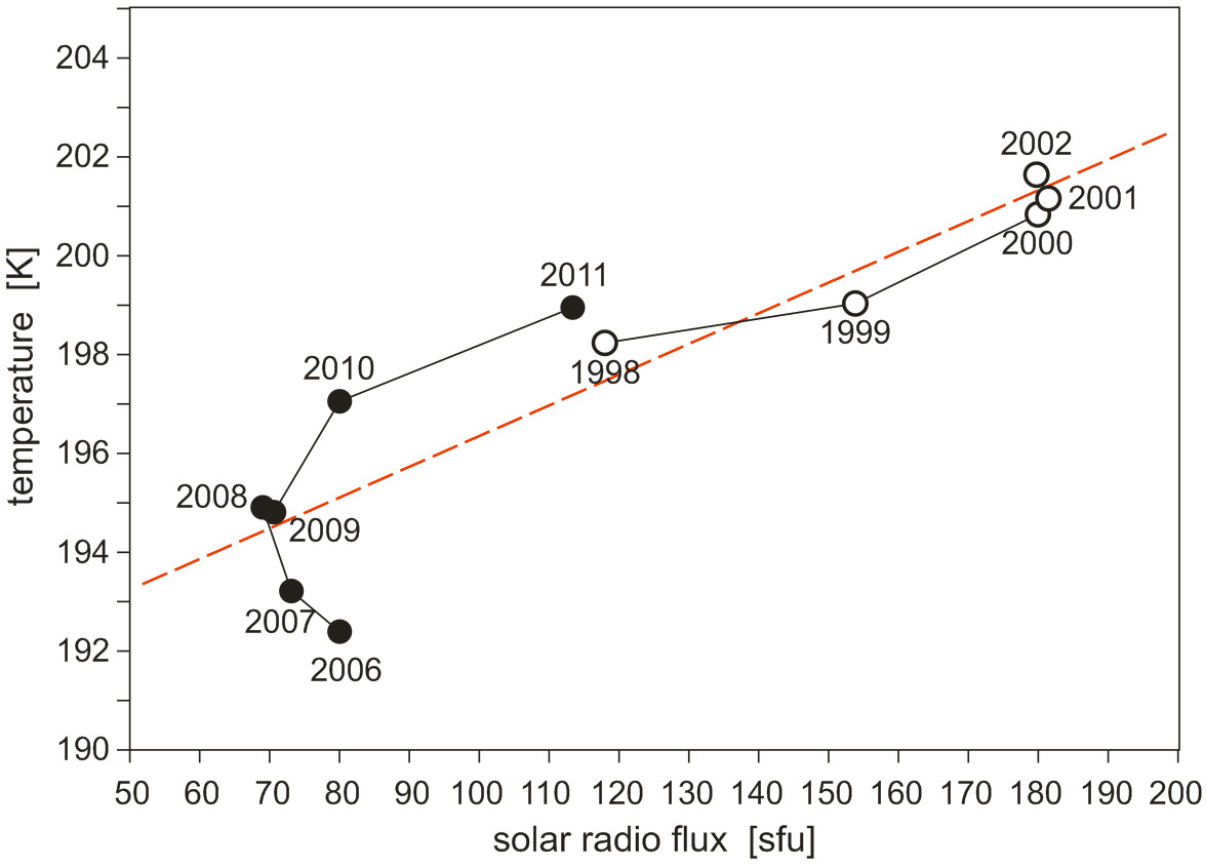

Figure 4. El Leoncito O_2 temperatures (circles) versus annual means of Penticton observed F10.7 cm solar flux, and the regression line through both data groups.

4. Model Uncertainty and Simpson's Paradox

But what is wrong with these putative results? Formally, there is nothing wrong, since the least-squares fit provides the best possible approximation to the data. And it cannot be denied that the characteristics thus determined might be defended on the ground of existing knowledge, as mentioned. But this is where a logical weakness becomes apparent, namely a certain circularity of argument, since the previous results also had to be defended on similar grounds. The decision to fit a sinusoid, or a regression line and nothing else, or both, is then the critical step. However, selecting the most appropriate mathematical description (the "model", in statistical terminology) is not an optimization problem in the same sense as a least-squares fit is. This is because the decision about what is appropriate must depend on context and purpose. There is a relation to the "model uncertainty" problem which statisticians found difficult to bring under control. This is clearly shown by the *Chatfield* [1995] paper that includes a debate among more than 30 professional statisticians which reflects the range of opinions existing in the statistical community on this topic, but also the consensus about the failure to address, in theory and teaching, the problem of choosing the "best model", i.e. the most appropriate formal description, to which the data are then fitted (while they agreed on the seriousness of the consequences that arise from wrong decisions).

In the context of trends in mesopause region temperature time series, the "model" consists of a combination of solar cycle, linear trend, but also instrumental artifacts. Some of these factors may be assigned a zero coeffient, a priori, or it may be left for the fitting algorithm to decide (at an additional cost in terms of required data length [*Weatherhead et al.*, 2002]).

As mentioned above, the situation with LEO O_2 temperature looks like a strong case of Simpson's paradox, when conclusions drawn from part of the data change sign, if the complete data set is analyzed. *Pearson et al.* [1899] and Pearson's disciple *Yule* [1903] first discussed weaker situations, where the conclusions from parts of the data did change appreciably, but the differences did not involve a change of sign. The behavior of our OH temperatures may qualify as a case in this category of weaker differences (at the level of the combined 1-σ errors given at the end of section 3.1). Half a century later, *Simpson* [1951] published an example of the more dramatic case, where it appeared necessary to draw the opposite conclusion from the complete dataset, than from any of the two parts of it, similarly to what we see in O_2 temperature.

The discussion of the reasons for these discrepancies, and more importantly, the question of which formal procedure should be recommended to avoid the paradox in all possible cases, continues until present times. The most serious practical consequences of the paradox come from medical statistics, but its logical and philosophical basis is still a topic of active investigation. Also in other areas of science, the consequences of wrong conclusions based on bad decisions are serious enough to warrant all efforts to avoid them.

According to *Pearl* [2009; mainly chapter 6], Simpson's paradox is due to a purely statistical instead of causal analysis, and the resolution of the paradox requires the introduction of a formally correct treatment of causality (J. Pearl is one of the main contributors to the development of such a formalism). This view has been contested by *Bandyopadhyay* et al. [2011], who, while agreeing that Simpson's paradox is not a logical paradox at all, but just a consequence of false human expectations, believe that a mathematically correct treatment can always arrive at a consistent description of what happens in parts of the data set, and in the combined version. In our view, Simpson's paradox appears to be just a special case of the general problem of (statistical) model uncertainty.

In the geophysical context as ours, causality of course plays the key role, in that physical mechanisms must account for all the quantitative description of reality that our data presumably refer to, and we cannot limit ourselves to an orthodox statistical analysis with disregard to causality. In this sense, there is nothing paradoxical in "Simpson's paradox". As essential as it is to distinguish between geophysical trends and instrumental drifts, as important must it also be for us to be able to finally distinguish between natural and anthropogenic change. This distinction is not feasible by statistical analysis, but requires geophysical models that correctly quantify all the relevant physical mechanisms ("models" is unfortunately the same term as used in statistical analysis, but refers to an altogether different concept).

5. Search for Instrumental Artifacts

"Instrumental artifacts" is a label for the consequences of all the physical processes within the measuring system which are not correctly taken into account in the data retrieval process. Access to quicklook and other subsidiary data and their inspection in search for clues about potential sources of instrumental artifacts like drift and other anomalies may be necessary to deal with instrumental effects. This, and remedial action in revised retrieval schemes are normally only accessible to the original investigators, and unavailable to mere data users. We shall here scrutinize some candidate mechanisms for potential sources of artifacts in our tilting filter spectrometer.

The instrument samples the airglow spectrum at seven spectral positions corresponding to predetermined wavelengths [*Scheer and Reisin*, 2001], to collect information on relative spectral intensity of three positions of the P-branch of the OH band (at 846.72 nm, 850.67 nm, 855.05 nm), three positions on the short wavelength flank of the O_2 band (at 861.4 nm, 862.3nm, 863.2 nm), and

one spectral background position (at 857.4 nm). There is an approximately quadratic relationship between wavelength and the number of motor steps. This relation also depends slightly on filter temperature. The airglow samples are covered by 2656 motor steps (for the more recent instrument configuration, and a filter temperature of 13°C).

As the filter temperature follows changes in ambient temperature, the filter tilt angle is adjusted to compensate for these changes, so that each sampling position (in terms of the number of motor steps corresponding to a given tilt angle) remains at its nominal wavelength. Ambient temperature changes are monitored continuously, but spectral peak shifts are also measured periodically by sampling at nine narrowly spaced positions close to a reference peak of the OH spectrum (the one at 850.7 nm). This is done repeatedly in both scan directions to arrive at statistically solid results free from backlash and short-term drift, and then a parabola is fitted through these nine samples to precisely determine peak position. This, and also the empirical relation between peak position and filter temperature are used to correct the seven airglow sampling positions. This is part of the normal data acquisition routine. The corresponding house keeping data are also saved as log files and used to derive further corrections during the final rotational temperature retrieval.

These archived house keeping data can be used to look for signs of possible long-term changes in instrument performance. This can be done by computing the mean positioning error for a data batch (i.e., over one or a few weeks), by averaging the individual (often more than a dozen) nightly positioning corrections.

The results over a time span of nearly 2400 nights, from 2006 to mid-2012, are shown in Figure 5. This plot represents the history of the positioning corrections that have actually been applied. There is clear evidence of a linear trend, and the slope of the corresponding regression line is 1.98×10^{-3} ($\pm8\times10^{-5}$) steps/day. This is a very small effect, but over the whole time span of Figure 5, a systematic positioning error of 4.5 steps would have accumulated. Although this is only a very small fractional displacement at any given sample position, if it had not been corrected for, then its effect on O_2 temperature would be quite noticeable. For the older 1998-2002 data block (not shown), the slope was 1.17×10^{-2} ($\pm7\times10^{-4}$) steps/day, that is, about a factor of six greater than in the later years (although corrected, as well). The mean positioning errors seldom deviate more than one motor step (corresponding to a linear displacement of the tilt mechanism by only 3 μm) from the linear trend (the standard deviation with respect to this line is 0.78 steps). The scatter about the trend line has not changed appreciably over the years, so there is no evidence for any other effect than the very slow apparently mechanical drift, which the positioning correction was meant to have taken into account.

If the drift were caused by an aging effect of the microswitch which defines the maximum tilt angle, then the presently used correction strategy should perfectly compensate for this. If wear in other parts of the tilting mechanism were the reason, the resulting higher-order deviations in the wavelength calibration for a given filter temperature might affect rotational temperatures, although more detailed knowledge of the process causing the drift would be needed to quantify the effect. Also a drift in the thermistor used for monitoring the filter temperature would not be automatically corrected for, as would any differential change in filter characteristics. However, it is hard to see how these hypothetical drifts could have caused the factor of six difference in the slope of the regression lines to the positioning data, before and after the instrument modification, while, on the other hand, resulting in the same order of O_2 temperature drift (if the 9.0 to 12.5 K/decade could at all be due to such an effect).

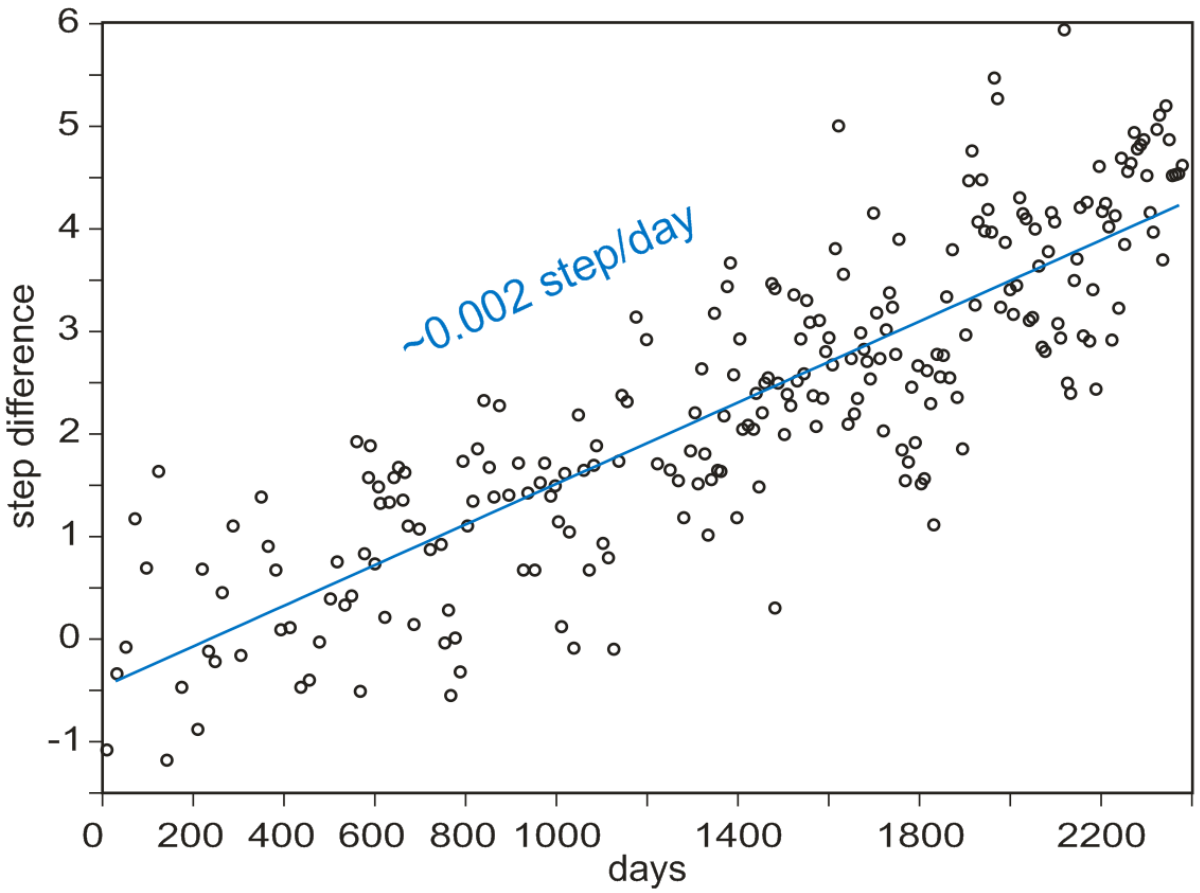

Figure 5. Mean sample position corrections for El Leoncito airglow spectrometer from early 2006 until mid-2011, and linear least-squares fit (see text for details).

The total spectral response of the instrument can be judged by two spectra of a neon lamp recorded almost 13 years apart (in 1998 and in 2011), that is, bridging most of the time span of the present dataset (see Figure 6). The shapes of the spectral peaks give a good idea of the instrument function (spectral width). There is no evidence of any remarkable change such as filter aging effects like differences in the width of the instrument function, which could affect instrument performance. The small differences in the neon spectra near 866 nm are too far from the three O_2 sampling positions, which lie between 861.4 and 863.2 nm, to affect O_2 temperature (the airglow sampling positions are marked by the arrows in Figure 6). This means that apart from eventual subtle higher-order effects, there is no obvious hint as to where the instrumental problem, if indeed there is one, might be hidden.

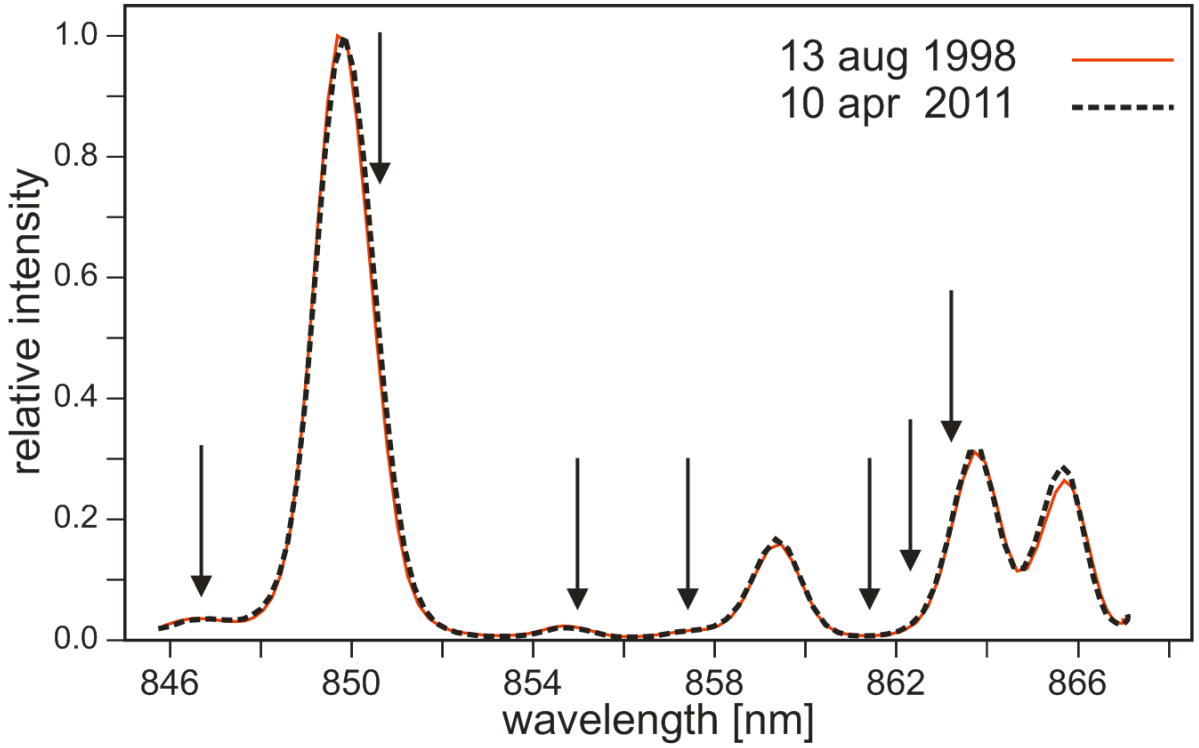

Figure 6. Neon lamp spectra measured with El Leoncito airglow spectrometer at the two dates shown. Arrows indicate the sampling positions used for airglow observations.

6. Intercomparisons

Intercomparison with satellite data used as transfer standard is in principle an ideal way to resolve incompatibility problems within a single dataset, as in the present example. However, as in a direct comparison of colocated ground-based instruments, the other instrument adds its own "degrees of freedom" in terms of potential instrumental drift effects to the number of problems to be resolved. The more complex the instrument hardware (and data retrieval), the more difficult the situation may become.

The data from the SABER instrument on the TIMED satellite should be suitable for such an intercomparison. Among many other atmospheric parameters, SABER also supplies temperatures in a wide altitude range that includes both airglow layers. The technique depends on the absolute photometry of incompletely thermalized CO_2 emissions, and involves a complex retrieval scheme [*García-Comas et al.*, 2008].

In an early (although probably failed) attempt to determine the systematic uncertainty between both LEO data groups, we used SABER version 1.07 temperatures during 4002 overpasses at El Leoncito within 1000 km miss distance, since the beginning of SABER data acquisition in early 2002 until the end of 2010. Each SABER temperature profile was averaged with a gaussian weight function approximating the nominal peak altitude and typical shape of either the OH, or the O_2 airglow layer. This leads to "airglow-equivalent" SABER temperatures corresponding to the altitudes of 87 and 95 km, respectively. The resulting annual means of these temperatures are shown in Figure 7, not only as average (arithmetic means), but also as median values. We show both variants, because we observed an unexpected pronounced asymmetry of the temperature histogram for the OH layer (see Figure 8, panel a). This asymmetry is practically absent for O_2 (Figure 8, panel b). As a consequence of the asymmetry, the average OH-equivalent temperatures are about one kelvin greater than the correponding medians. However, there is no systematic difference between averages and medians, for O_2 (i.e., at 95 km).

The main message of Figure 7 is that the SABER temperatures over El Leoncito exhibit strong negative trends at both altitudes, namely about -5 K/decade at 87 km, and even nearly -9 K/decade at 95 km. Without discussing the geophysical plausibility of this behaviour (which might be related

to the decline of solar activity towards the end of solar cycle 23, but also temperature trends, and eventually, instrumental drift, or retrieval problems), we will just take it here for granted in order to estimate the systematic uncertainty between both LEO data groups.

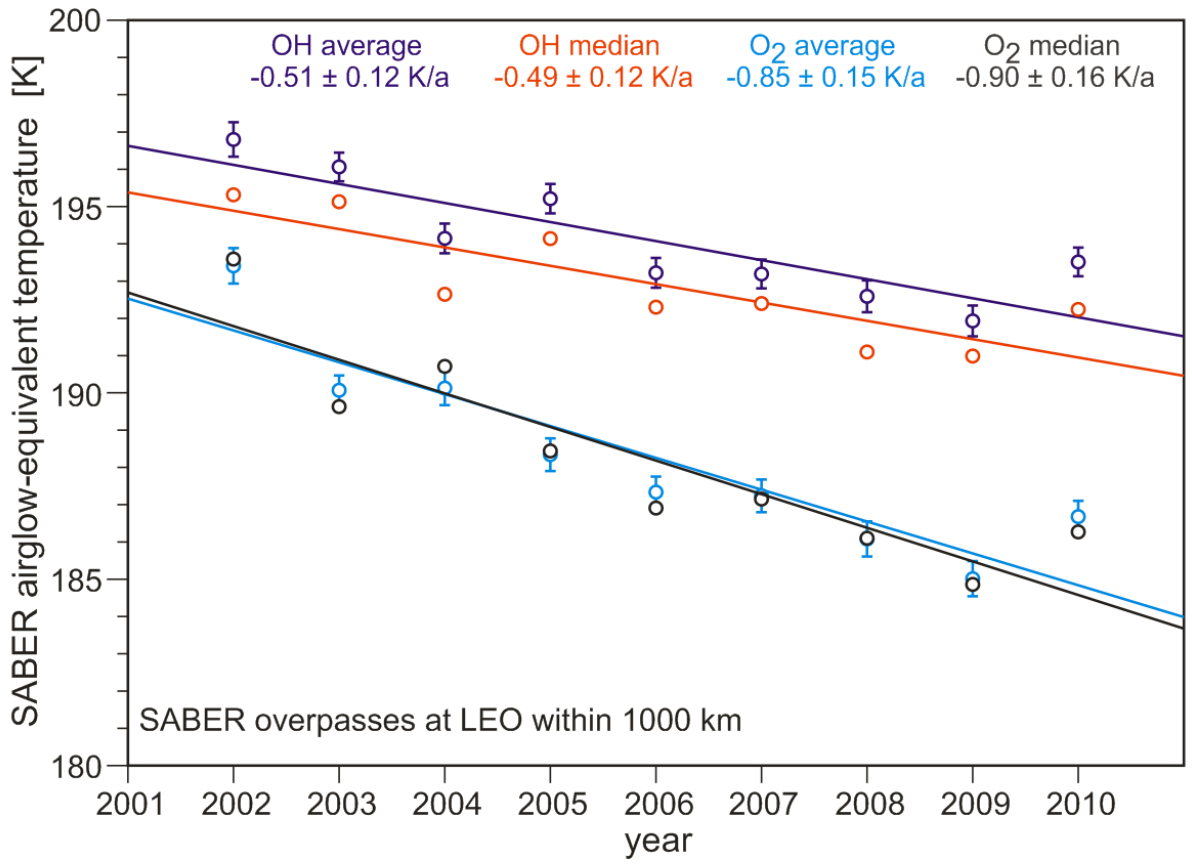

Figure 7. Annual means and medians of OH and O_2 airglow-equivalent SABER temperatures during overpasses at El Leoncito. Straight lines represent fits to the corresponding points (points, lines and labels in top of figure are color-coded).

Only the SABER data of 2002 are available for comparison within the earlier LEO data group. For this year, the annual mean O_2 temperature was 8.2 (±0.5) K greater than the average of the O_2-equivalent SABER temperature. The comparison for the second LEO data group was done for the years 2006 to 2010, and resulted in LEO O_2 temperatures 8.0 (±1.1) K greater than the corresponding SABER result, on average. So, these two offsets are equal within error bars, and we are led to deduce that the systematic uncertainty between both LEO data groups be zero. However, the quality of this putative result is poor, because of the opposite trends in the behaviour of the LEO and SABER temperatures, which cause the offset to grow from 5.1 K in 2006 to 10.4 K in 2010. This is another example where more data do not lead to better statistics but do reveal more complexity.

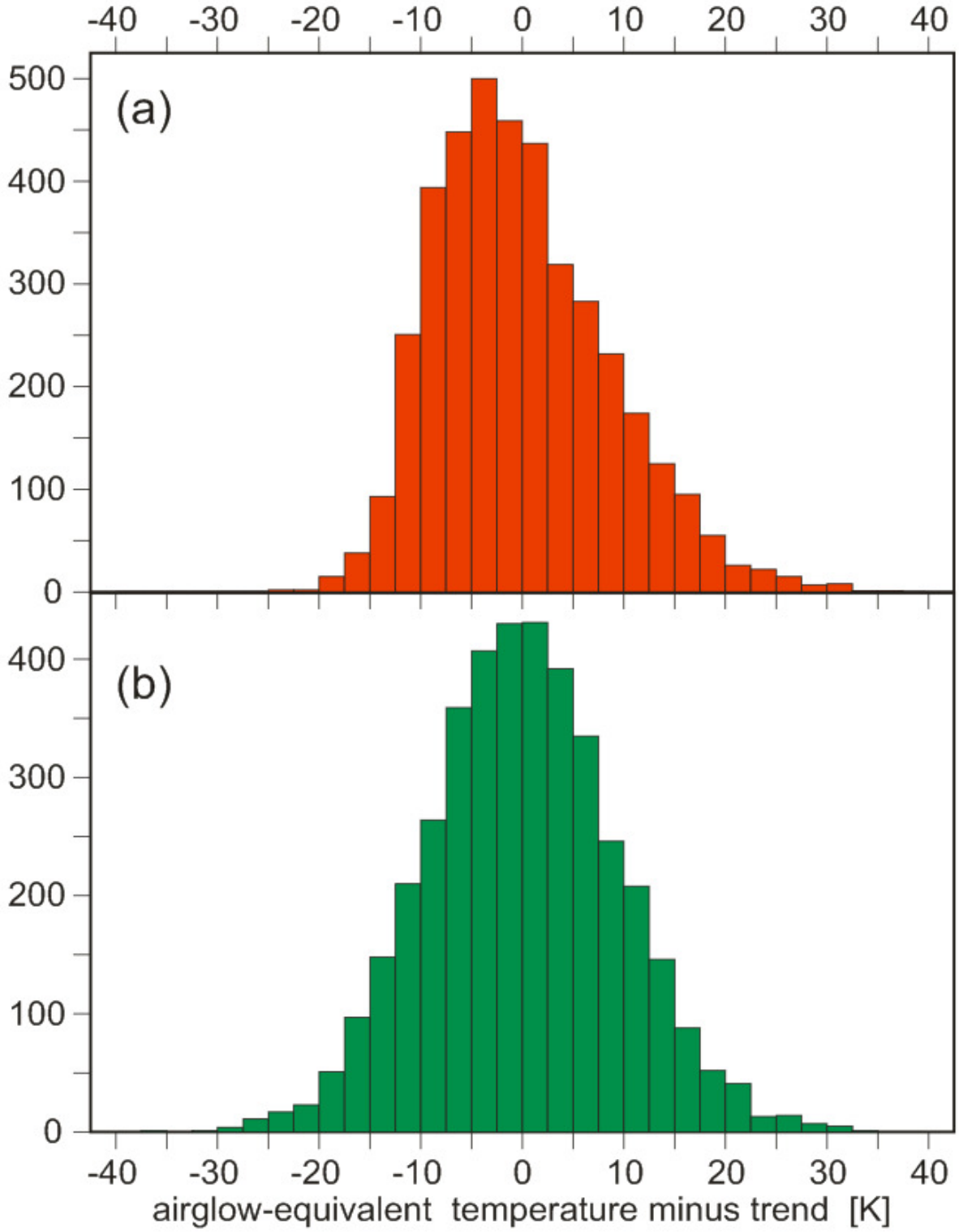

Figure 8. Histograms of the scatter of SABER airglow-equivalent temperatures about the linear trend fitted to the annual means. (a) for OH-equivalent temperature (at 87 km), (b) for O_2-equivalent temperature (at 95 km). Ordinates show numbers of cases within a 2.5 K bin, from the total of 4002 overpasses.

In an independent intercomparison for the Antarctic station Davis (68°S, 78°E), *French and Mulligan* [2010] found a strong positive trend in the bias of OH-equivalent SABER V1.07 temperatures with respect to ground-based OH rotational temperatures. This bias trend was 6.6 (±1.6) K/decade. On the other hand, those authors found a small negative trend in the bias of a different satellite instrument (Microwave Limb Sounder on the Aura satellite), again with respect to their ground-based observations. This contradiction was tentatively blamed on possible SABER instrument drifts or retrieval problems of data version 1.07. The use of modeled atomic oxygen mixing ratios in the temperature retrieval is a possible contributor to the uncertainty, as discussed

by *Smith et al.* [2010]. The forthcoming SABER temperature version 2.0 is expected to improve in various aspects (as signalled by *Stevens et al.* [2012], who used a pre-release V2.0), and it remains to be seen whether, and how, the long-term variations will be affected.

7. Conclusions

The airglow rotational temperatures obtained at the nominal altitudes of 87 and 95 km, which fall into two subsets separated by a three-year data gap, after which some instrument characteristics may have changed, serve to illustrate the limitations of an acausal, purely statistical, analysis.

Annual mean temperatures at 87 km, from both subsets (1998 to 2002, and 2006 to 2011) together show a rather smooth negative trend of 2.1 (±0.6) K/decade. Although this value does not include an adjustment for the unknown offset between both data subsets, the quality of this fit does not look so bad as to suggest the need to look for alternative statistical models to better fit the data.

For the corresponding data at 95 km, the situation is different, by offering several alternative views. While there is a strongly negative trend, for the complete dataset, this has a poor statistical quality because of the strongly positive trends for each subset. The situation looks like Simpson's paradox, to which the literature still does not provide universally acceptable, and implementable, solutions. For our data, the discrepancy cannot be resolved by an ad-hoc adjustment between both data subsets.

Another possibility, to fit a sinusoid to all the data, leads to a period of 12.2 years, with one maximum close to the peak of solar cycle 23, and an amplitude of about 4 K. On the other hand, a regression line fit to temperature versus solar radio flux, which should better represent the solar cycle effect, surprisingly gives much poorer results than the sinusoidal approach. At this stage, it is not clear how the unknown offset (different from the one at 87 km) between both data subsets, linear trends, and eventual instrumental artifacts might lead to a more satisfactory solar-geophysical interpretation of the data.

We have searched for potential sources of instrument drift as a contribution to the anomalous "Simpsonian" behaviour, and discussed some instrumental characteristics in detail, but have not found a convincing cause.

Intercomparison with satellite data are also inconclusive. SABER overpass data at El Leoncito show a surprisingly strong negative trend of about 9 K/decade at 95 km (unprecedented in the literature), and also a smaller negative trend at 87 km. On the other hand, other evidence from the literature casts doubt on the stability of SABER V1.07 temperature data, which hopefully will be resolved in the next V2.0 data version.

We think that although the examples discussed are specific to one individual instrument, in principle data from any source are likely to be subject to these, or to other problems. Intercomparison and thorough analysis of the behaviour of each measuring system is definitely the route to follow, to resolve uncertainties in trend analysis. However, we hope to have made clear that this inevitably demands the elimination of additional unknowns.

Acknowledgments. The authors are grateful to the SABER team for permitting access to the SABER data via the http://saber.gats-inc.com/ internet site, and to the CASLEO staff for technical support during the many years of data acquisition at El Leoncito. We also acknowledge funding by CONICET grant PIP 112-200801-00287, and thank an anonymous reviewer for his helpful comments.

References

Bandyopadhyay, P.S., D. Nelson, M. Greenwood, G. Brittan, and J. Berwald (2011), The logic of Simpson's paradox, *Synthese*, 181(2), 185-208, doi: 10.1007/s11229-010-9797-0.

Beig, G., J. Scheer, M.G. Mlynczak, and P. Keckhut (2008), Overview of the temperature response in the mesosphere and lower thermosphere to solar activity, *Rev. Geophys.*, 46, RG3002.

Beig, G. (2011), Long-term trends in the temperature of the mesosphere/lower thermosphere region: 2. Solar response, *J. Geophys. Res.*, 116, A00H12.

Chatfield, C. (1995), Model uncertainty, data mining and statistical inference, *J. Roy. Statist. Soc.*, A 158, 419-466.

French, W.J.R., and F.J. Mulligan (2010), Stability of temperatures from TIMED/SABER v1.07 (2002-2009) and Aura/MLS v2.2 (2004-2009) compared with OH(6-2) temperatures observed at Davis Station, Antarctica, *Atmos. Chem. Phys.*, 10, 11439-11446.

García-Comas, M., M. López-Puertas, B.T. Marshall, P.P. Wintersteiner, B. Funke, D. Bermejo-Pantaleón, C.J. Mertens, E.E. Remsberg, L.L. Gordley, M.G. Mlynczak, and J.M. Russell III (2008), Errors in Sounding of the Atmosphere using Broadband Emission Radiometry (SABER) kinetic temperature caused by non-local-thermodynamic-equilibrium model parameters, *J. Geophys. Res.*, 113(D24), D24106,doi:10.1029/2008JD010105.

Lovejoy, S. (2013), What is climate?, *EOS*, 94(1), 1-2.

Pearl, J. (2009), Causality: Models, Reasoning and Inference, Cambridge University Press, Cambridge UK.

Pearson, K., A. Lee, and L. Bramley-Moore (1899), Genetic (reproductive) selection: Inheritance of fertility in man, and of fecundity in thoroughbred race horses, *Phil. Trans. Roy. Stat. Soc. London*, Ser. A, 192, 257-330.

Reisin, E.R. (1987), Medición espectroscópica de temperaturas atmosféricas en la zona de la mesopausa, Tesis de Licenciatura de Física, Universidad de Buenos Aires, Buenos Aires, Argentina. [Available at http://www.iafe.uba.ar/aeronomia/semin.pdf].

Reisin, E.R., and J. Scheer (2002), Searching for trends in mesopause region airglow intensities and temperatures at El Leoncito, *Phys. Chem. Earth*, 27, 563-569.

Scheer, J. (1987), Programmable tilting filter spectrometer for studying gravity waves in the upper atmosphere, *Appl. Optics*, 26, 3077-3082.

Scheer, J. and E.R. Reisin (1990), Rotational temperatures for OH and 02 airglow bands measured simultaneously from E1 Leoncito (31°48'S), *J. Atmos. Terr. Phys.*, 52, 47-57.

Scheer, J. and E.R. Reisin (2001), Refinements of a classical technique of airglow spectroscopy, *Adv. Space Res.*, 27(6-7), 1153-1158.

Scheer, J., E.R. Reisin, and C.H. Mandrini (2005), Solar activity signatures in mesopause region temperatures and atomic oxygen related airglow brightness at El Leoncito, Argentina, *J. Atmos. Sol. Terr. Phys.*, 67(1-2), 145-154.

Simpson, E.H. (1951), The interpretation of interaction in contingency tables, *J. Roy. Stat. Soc.*, Ser. B, 13, 238-241.

Smith, A.K., D.R. Marsh, M.G. Mlynczak, and J.C. Mast (2010), Temporal variations of atomic oxygen in the upper mesosphere from SABER, *J. Geophys. Res.*, 115(D18), D18309, doi:10.1029/2009JD013434.

Stevens, M.H., L.E. Deaver, M.E. Hervig, J.M. Russell III, D.E. Siskind, P.E. Sheese, E.J. Llewellyn, R.L. Gattinger, J. Höffner, and B.T. Marshall (2012), Validation of upper mesospheric and lower thermospheric temperatures measured by the Solar Occultation for Ice Experiment, *J. Geophys. Res.*, 117, D16304.

Tiao, G.C., G.C. Reinsel, D. Xu, J.H. Pedrick, X. Zhu, A.J. Miller, J.J. DeLuisi, C.L. Mateer, and D.J. Wuebbles (1990), Effects of autocorrelation and temporal sampling schemes on estimates of trend and spatial correlation, *J. Geophys. Res.*, 95(D12), 20507-20517.

Weatherhead, E.C., G.C. Reinsel, G.C. Tiao, X.-L. Meng, D. Choi, W.-K. Cheang, T. Keller, J. DeLuisi, D.J. Wuebbles, J.B. Kerr, A.J. Miller, S.J. Oltmans, and J.E. Frederick (1998), Factors affecting the detection of trends: Statistical considerations and applications to environmental data, *J. Geophys. Res.*, 103(D14), 17149-17161.

Weatherhead, E.C., A.J. Stevermer, B.E. Schwartz (2002), Detecting environmental changes and trends, *Phys. Chem. Earth*, 27, 399-403.

Yule, G.U. (1903), Notes on the theory of association of attributes in statistics, *Biometrika*, 2, 121-134.