Journal: ACP
Title: Assessing large-scale weekly cycles in meteorological variables: a review
Author(s): A. Sanchez-Lorenzo et al.
MS No.: acp-2011-944
MS Type: Review Article

Response to Thomas Bell

General remarks

Because many research papers have been published on whether or not weather statistics change with the day of the week, and because the research results reported are so disparate and confusing, it is high time for a review of the literature to try and bring some order to all this. A review of the literature advances science by both cataloguing and evaluating past research, helping to clarify what research survives careful scrutiny and what research may have reached conclusions that aren't supported by the paper's contents. A bibliographic list by itself is not enough to justify publishing a review of this topic in a research journal.

Although I can understand the authors' reluctance to make judgments about various papers in an organized and assertive way, I feel that one thing that is missing from this manuscript is something like a "grade" for each paper that would suggest how reliable the methods used in the paper are, so that, for instance, modelers looking for validation of their models know which results can be trusted and which can't. Sanchez-Lorenzo et al. do this in a kind of informal way in their discussion of each paper, and in their discussion of statistical techniques, but they don't summarize their opinions in a way that a reader can determine by quickly scanning the contents. And the opinions of the authors of this manuscript are frequently not so easy to determine. One can sometimes only discern their opinions by "reading between the lines." I describe below some of my own opinions about the problems that afflict the papers being reviewed. Ideally, in my opinion, Table 1, containing a list of the papers reviewed, should have an extra column or two containing some measure(s) of the reliability of each paper's results.

Many of the papers being reviewed in this manuscript do not control for the false discovery rate (FDR) (discussed recently by Wilks, 2006). [References given here may be found in the Reference section of the manuscript or at the end of this review.] The FDR can lead to results that make no physical sense, because some thing is identified as significant when in fact it is just noise. Or the FDR can lead to results that seem to vary inconsistently from one dataset to another. There is no shame in not having corrected for the FDR, because I don't think methods for dealing with it were very well developed until recently, even though climate researchers were aware of the problem at a "gut level". Methods of dealing with the FDR aren't as yet, perhaps, common knowledge. Nevertheless, papers that used methods susceptible to the FDR effect need to be flagged where possible.

Another problem I see in too many papers is a lack of details about the methods used. If one has any doubt about the results and wants to see if perhaps the methods used were not as strong as the authors credit them with, there is sometimes no way to evaluate for oneself the adequacy of

the method. Papers published in GRL, Science, and Nature are particularly prone to this problem because of the page limitations imposed by the journals. I believe that if a paper reports results that seem odd and one cannot determine by reading the paper whether the methodology was adequate, the paper should be flagged as such.

The purpose of this aspect of the authors' review would not be to "point fingers" at particular papers and authors, but to indicate that some topics need to be revisited with better data or better methods. This would be enormously helpful to this research area with all its confusing reports. (I should say that it is perhaps too glib to suggest that a single "grade" could adequately describe a paper. Most papers contain the results of multiple investigations. Some of the results reported are excellent and informative, and others are flawed. A single table might not be capable of handling this without an annoying number of footnotes. Moreover, the fact that some aspect of a paper is flawed does not mean the whole paper is "wrong"; all of the papers reviewed have valuable observations and insights to offer. It is only their estimates of statistical significance that may need revisiting.) But if the authors can devise a presentation that gives the reader more guidance for what papers contain results that are trustworthy, given what we now know, this review paper would be a great contribution. As I mentioned above, this might be done by adding a column or two to Table 1.

I believe the manuscript as it stands is an excellent beginning at reviewing this subject matter that will serve the community well, but I really believe that the one feature it lacks that would lift it to a far higher level of usefulness is some sort of assessment of each paper with a "grade" or code that helps the reader to decide what results can be relied on and what results need further confirmation. This is admittedly difficult and somewhat awkward, but given the fact that this journal permits an open discussion of whatever assessments are made in this paper, and an opportunity for authors of the papers being reviewed to provide their own viewpoints, I think not to do this would be to miss a real opportunity to bring some clarity to this research area. Otherwise this review paper will be treated as a convenient bibliographic list. It will probably be referred to a lot, since it saves future authors from having to enumerate all the previous studies in this area.

In my comments below, I try to point out some of the papers I myself have concerns about, but I don't discuss them all – that is, after all, the job of Sanchez-Lorenzo et al.

We really appreciate the reviewer's comments and his support for the publication of our paper in ACP. The authors of this manuscript would like to emphasize his tremendous work to review our manuscript, as we have never seen before such an effort by a reviewer.

The manuscript has been revised after considering most of his comments and suggestions. Regarding the "assessment" of the statistical analyses performed in the literature listed in Section 1, we have added a last column in Table 1 with a qualitative indication of the robustness of the analysis. In the answer to the item 6 the reviewer will find more details about this issue and our opinion regarding such assessment.

Major and minor comments are addressed below.

Major comments

1) P. 3, l. 11: It may not be the weekly cycle in commuter traffic but the cycle in commercial transportation (especially involving large diesel engines) that is at issue. A number of studies of road traffic (admittedly located in the Western hemisphere) find that commuter traffic volumes don't change so much with the day of the week – only the hours of the day when the traveling occurs changes. It's a different story for commercial traffic, which is observed to diminish on weekends.

Thank you for this information. "Commuter automobile traffic" has been replaced by "commercial transportation".

2) P. 3, l. 20 top. 4, l.3: The authors make a good point about the distinction between searches for a weekly cycle at a local (generally urban) level (often carried out as single-site studies) and weekly cycles over large scales. They should mention that there is a weekly cycle in pollution in rural areas as well as in urban areas, and measured as a percent change they may not be so different. Rosenfeld et al. (2008) point out that the effect of aerosols on storm development is not linear: aerosol effects on storm invigoration seem to peak at AOD (aerosol optical depths) of ~ 0.25–0.3. Above that they may moderate and even change sign. It is therefore well within the realm of physical possibility that weekly cycles can be quite different in different areas (even though the weekly cycle in aerosols on a percent basis is the same). The weekly cycle may disappear or even change sign. The same can be said for the weekly cycle in different decades. The fact that some studies in some areas (and time periods) see different weekly cycles in atmospheric response does not necessarily imply that some of the studies are faulty. It's also possible that the weekly cycle on a local level is too small to detect with a small amount of data, but weekly cycles in large-area averages may be detectable because the natural variability (the noise) is beaten down by the averaging, allowing the small signal to show through. (The weekly cycle may even be smaller near cities than away from them. See, for example, Bell et al., 2009a.)

We have added that weekly cycles in atmospheric pollution in rural areas have also been observed, which is also explained in Section 4.2.1 (*Evidences of WCs in pollution and anthropogenic aerosols*). Equally, we appreciate the information regarding the non-linear effect of the aerosols, which has been also pointed out in Section 4.2.2 (*Brief summary of direct and indirect aerosol effects*): "…In a convective cloud, if precipitation is reduced for liquid-water clouds, more liquid water may reach the freezing level, and freezing is delayed to higher altitudes, the so-called thermodynamic effect or convective invigoration effect (Koren et al., 2005). Such an effect would lead to deeper clouds, and more intense precipitation, although Rosenfeld et al. (2008) pointed out that effects of aerosols might be not linear on storm invigoration."

3) P. 4, line 1: "and therefore modify the radiation budget." This summary of possible explanations for large-scale weekly cycles doesn't include the mechanism proposed by Rosenfeld, reviewed in Rosenfeld et al. (2008), though the authors discuss this later in the manuscript. The "Rosenfeld" mechanism doesn't involve radiative effects at all, but involves

instead the thermodynamic consequences of the reduction in droplet size for cloud and storm development.

We have shortened this paragraph and this sentence does not appear in the current version that now reads as: "A different mechanism is required in order to explain the "large-scale" WCs, being the most common approach to link these changes, i.e. through atmospheric interactions at the mesoscale, with direct and indirect aerosol effects (see Section 4.2.2)."

On the other hand, the direct and indirect effect of aerosols are summarized and discussed in Section 4.2.2, as has been pointed out by the reviewer.

4) P. 4, ll. 14 –15: Instead of "a sequence of seven days can be described using a sinusoidal function with a periodicity of seven days," I would prefer to say "the variation of day-of -the-week averages can be described approximately by a sinusoidal function with a period of exactly seven days."

Done.

5) P. 4, ll. 16 –17: I don't understand the statement, "The latter approach does not necessarily mean that there is not a real seven-day cycle in the data." To me, a weekly cycle exists if the probability distribution $p(r,t)$ of an atmospheric variable r depends on the time t in a special way: it is periodic in time with a period of exactly 7 days. (In reality this can't possibly be true for the real atmosphere with its seasonal and interannual variation and decadal trends, but it is, one hopes, a useful approximation to reality.) The example given in the manuscript here, that the mean for one day of the week is different from the means for the other days of the week, seems to me a sign that there is a weekly cycle in the statistical sense described above.

In order to clarify the text we have deleted the sentence "The latter approach does not necessarily mean that there is not a real seven-day cycle in the data" in the revised manuscript.

6) Pp. 4 –17, i.e., Section 2: This section consists largely of a catalog of papers on the weekly cycle that the authors have chosen to discuss. The reader must wait till later sections to get any hint about the reviewers' conclusions about how credible each paper's results might be. In my opinion, it would be better for the authors to first review the statistical methods generally employed in the papers reviewed, and to point out the pitfalls in some of the methods. Later, when the papers are enumerated and discussed, the authors should indicate those papers whose results may be affected by statistical problems of which the original authors were unaware, or point out any deficiencies in the data used in the paper under discussion.

With this possible reorganization of the paper in mind, I will mention some of the papers discussed in Section 2 that might be flagged as having potential statistical problems. The open-discussion format of ACP will allow for the original authors to offer rebuttals to opinions expressed in the review.

The reviewer is probably right, but it is also true that the main objective of the current manuscript is to list publications dealing with the topic of the large-scales weekly cycles, with a

brief summary of their main conclusions. Secondly, the review tries to point out some of the main weaknesses in the statistical analysis observed in the literature as a whole (Section 3), as well as the possible causes that, if real, can explain the weekly cycles (Section 4).

Nevertheless, the reviewer will probably agree with us that, unfortunately, currently there is no consensus on the best way to proceed in order to evaluate the statistical and physical reliability of weekly cycles. The reviewer points out his opinion about this matter, as well as provides an excellent assessment of the main weaknesses of the statistical analysis used until the present. In fact, some of his opinion and suggestions will be added into our paper, and will substantially contribute to its improvement. On the other hand, we consider that a more detailed discussion is desirable in the future in order to compare different statistical approaches (e.g. using a benchmark dataset) and to find a consensus in the community dealing with the topic of the large-scale weekly cycles. Afterwards, it may be easier to make judgments about the methods applied in some of the papers, as well as provide a guideline to follow in future studies. This assessment is, however, beyond the scope of the current manuscript and we consider very risky to flag as "suspicious" some papers when there is still no consensus about this issue.

Equally, the open-discussion format of ACPD only allows writing comments during the public discussion, which is closed after some weeks. Thus, the original authors will not have a chance to offer rebuttals to our opinions. We consider that a future workshop/meeting should be desirable for this issue, as well as to create an international mailing list with the scientists interested to discuss this issue.

Nevertheless, following the reviewer's suggestion regarding the "assessment" of the analyses performed in the literature listed in Section 2, we have added a last column in Table 1 with a simple, qualitative indication of the complexity and robustness of the statistical analysis used and described in the different manuscripts. This last column is described and explained in Section 3 after summarizing the main weaknesses observed in the statistical analysis applied to the large-scale weekly cycles: "Taking into account all the recommendations summarized in this section, we have added a last column in Table 1 with a qualitative indication of the complexity and robustness of the statistical analysis used and described in the different manuscripts. An assessment of these methods is, however, beyond the scope of the current manuscript, as currently there is no consensus on the best way to proceed in order to evaluate the statistical reliability of WCs."

7) P. 6, ll. 3 –4: More precision might be helpful here: "… datasets over the Atlantic coast of the U.S. and neighboring oceanic areas".

Done.

8) P. 7, ll. 16 –20: It appears that the authors of the paper Sanchez-Lorenzo et al. (2009) and the authors of the Comments in the same journal by Hendricks Franssen et al. (2009) might have disagreed on the trustworthiness of the results in Sanchez -Lorenzo et al. (2009). Since some of the same authors are collaborating on this review, can the review authors summarize for the reader what conclusions they have reached about Sanchez -Lorenzo et al. and what research, if any, needs to be done to improve their results?

In fact, Hendricks Franssen et al. (2009) published a comment of a paper published by Sanchez-Lorenzo et al. in 2008 and not in 2009. In their reply published in 2009, Sanchez-Lorenzo et al. added new analyses regarding the diurnal temperature range and pressure weekly cycles in Spain, as well as suggesting a possible weekly cycle in some of the main circulation patterns in a window covering the 50% of the Northern Hemisphere centered over the Atlantic Ocean. Thus, the work summarized in the former p. 7, l. 16-20 is not criticized in Hendricks Franssen et al. (2009).

Overall, the reviewer' suggestion is already included in Section 2.5 with a brief discussion of the original paper published by Sanchez-Lorenzo et al. (2008), the subsequent comment by Hendricks Franssen et al. (2009) and the reply by Sanchez-Lorenzo et al. (2009) that read as: "Concerning Southern Europe, Sanchez-Lorenzo et al. (2008) claimed to find significant winter WCs over Spain for different meteorological variables (temperatures, rainfall, cloud cover, sunshine duration, and sea level pressure) using data for the 1961–2004 period. The results showed a tendency towards positive (negative) anomalies of rainfall, cloud cover (DTR, sunshine duration, and sea level pressure) during the central days of the weekdays, and the opposite anomalies for weekends. As they used series distributed over different geographical areas with different levels of urban influence, they argued that these WCs can hardly be related only to local effects, suggesting a possible link with periodicities in atmospheric circulation over Western Europe. Hendricks-Franssen et al. (2009) commented the paper, pointing out some deficiencies in the statistical analysis (mainly linked to neglecting the strong spatial auto-correlation in the statistical analysis). They claimed, on the basis of non-parametric testing, Monte Carlo bootstrap methods and a periodogram analysis, that the WCs of air pressure are not significant over Spain (for more details see Sect. 3). In their reply, Sanchez-Lorenzo et al. (2009) agreed with some of the deficiencies pointed out in the comment, although suggesting that the analysis should be applied for all the meteorological variables (not only air pressure), specially for the ones with low spatial auto-correlation (e.g. rainfall). Anyway, they also showed new evidences of winter WCs in DTR over Spain after the application of more robust statistical analysis by using PCA techniques and the non-parametric Kruskall-Wallis test."

9) P. 8, ll. 4–11: The paper by Kim et al. (2010) uses a sophisticated statistical technique, cyclostationary Empirical-Orthogonal-Function (CSEOF) analysis, and makes some extraordinary claims based on it. In particular, they claim that their analysis indicates that there is a "naturally occurring" weekly cycle and that weekly cycles in statistics need not be anthropogenic. In my opinion their conclusion is not justified. The CSEOF technique finds a temporal filter such that only data with temporal oscillations with periods near 7-days are passed. The wording of the paper is not perfectly clear to me, but I believe the authors may have actually imposed an assumption that the statistics of the data had exact 7-day periodicity. (In Kim and Roh, 2010, I believe they imposed only 365-day periodicity.) If they did not impose 7-day periodicity, the filter is not infinitely narrow, but passes oscillations with any period in the neighborhood of 7 days. An oscillation "in the neighborhood" of 7-days does not preserve its phase linked to the days of the week year after year and does not qualify as a weekly cycle. If they did impose 7-day periodicity, then the 7-day periodicity they "discover" is built into the method from the start. Kim et al. (2010) claim that the variability captured by the CSEOF method is Rossby-wave-like, but this is asserted rather than proved. The paper is interesting, but I don't believe its contents justify the sweeping conclusions it draws.

We have considered the reviewer opinion in Section 4.1, where the concerns about the reliability of the methods are introduced. Equally, in former p. 8, l. 4-11 the paragraph has been slightly changed in order to clarify that Kim et al. (2010) findings are not conclusive: "They proposed a method based on cyclostationary Empirical Orthogonal Function (EOF) analysis to remove a possible "natural" weekly cycles in the time series data (see section 4.1 for more details). After their attempt to isolate the naturally occurring weekly cycles, they claimed that there was a remaining signal, which is speculated to be anthropogenic, with positive anomalies on weekends in mid-western U.S. (mainly in Colorado and New Mexico) and a strong signal over the North-east, with the main centre around 40°N and 85°W. Nevertheless, they speculated that the possible natural component in the WC is of the same magnitude, if not stronger, than the anthropogenic component, although the proposed method need to be further scrutinized to ensure its robustness."

10) P. 9, l. 21: Rosenfeld and Bell (2011) looked at the statistics of the entire eastern half of the U.S. (east of 100W), not just the SE U.S., and found a statistically significant weekly cycle over this area. The earlier papers by Bell et al. confined themselves to analysis of latitudes south of 40N because the TRMM satellite data does not exist north of this latitude. The weekly cycle in rainfall was first detected in the TRMM data, and the later studies were partly motivated by the desire to buttress the TRMM results.

We have slightly changed the last sentence of this paragraph in order to clarify the studied area in Rosenfeld and Bell (2011) in comparison with earlier papers by Bell et al.: "In later studies, the same team showed more evidences of WCs over the land in south-eastern USA and nearby ocean during the summertime, using different datasets of lightning (Bell et al., 2009a) and storm heights (Bell et al., 2009b), as well as tornado and hailstorm activity in the entire eastern half of the U.S. (east of 100W) (Rosenfeld and Bell, 2011)."

11) P. 9, ll. 22 –25: The paper by Tuttle and Carbone (2011) used a technique suggested by Wilks (2006) to try to avoid the problem of the "false discovery rate." As the authors Sanchez-Lorenzo et al. discuss later, when a large number of statistical tests on data are carried out, some of the statistical tests will yield significance levels ("*p*-values") whose *p*-values are extraordinarily low, even though the system being tested in fact satisfies the null hypothesis. The tests that yield unusually low *p*-values are consequences of the expected "false discovery rate." A proper statistical analysis of a problem requiring many independent statistical tests (as occurs when one examines multi-station data one station at a time for evidence of a weekly cycle) has to take account of this effect. Many research papers on the weekly cycle may have been susceptible to this problem, and their estimates of the statistical significance of what they find are consequently highly suspect.

We appreciate the reviewer's comment regarding the "false discovery rate" suggested by Wilks (2006) and correctly applied by Tuttle and Carbone (2011). In order to point out the reliability of their analysis we have slightly changed the sentence: "Very recently, Tuttle and Carbone (2011) analysed radar-estimated summer rainfall during the period 1996-2007 over the U.S., and found a significant WC with weekday maxima in western Pennsylvania after evaluating the field significance in the data (e.g. Wilks, 2006)."

Equally, as suggested by the reviewer, we have included in the revised manuscript (including in Section 3) more comments about the need of a control of the "false discovery rate" problem.

12) P. 10, ll. 9 –12: The paper by Lacke et al. (2009) pursues some very interesting questions, but, as best I can tell, the statistical tests may not be strong enough to assure us of the conclusions the authors reach. The tests are carried out over thousands of grid points, the authors didn't control for the false discovery rate, and so their statistical conclusions are difficult to evaluate. As the review authors point out, Lacke et al. may not have had enough data to make solid statistical inferences possible -- but they are asking good questions. The study needs to be repeated with more data and statistical analyses that include the effects of the false discovery rate.

We agree with the reviewer and we have clarified that Lacke et al. (2009) did not control for the false discovery rate: "Their analysis is based on summer data during the 2003-2004 period and without a control of the FDR, which limits their conclusions."

13) P. 10, ll. 21 –26: [Admittedly I am not unbiased in commenting on the paper by Schultz et al. (2007), since I have published a Comment, Bell and Rosenfeld 2008, on the paper.] Schultz et al. performed statistical tests at 219 gauge sites spanning the continental U.S. and all seasons. The statistical question they asked was, in effect, "Do the statistics at any site (among the 219 sites) for some day of the week differ from the site's statistics for another day of the week?" Since the statistical question being asked is so unfocussed, it isn't surprising that nothing significant was found: It would take an extraordinarily strong weekly cycle to stand out relative to all the "noise" from the other sites and days of the week. It is well established (e.g., see the Hasselmann 1979 reference in the Comment by Bell and Rosenfeld 2009) that when the statistics at multiple sites are examined to see if there is a detectable change in the mean at any one of the sites, the investigator is likely to obtain a null result, even though the mean may have changed at all the sites by a small amount. The description of the statistical techniques by Schultz et al. in their paper makes it difficult to determine if the authors properly controlled for the false discovery rate, but I believe they may have done so. Their approach was not appropriate, however, for deciding whether there is a large-scale weekly cycle in the data, and the fact that their test failed to show anything tells us only that no one site has an extremely strong weekly cycle. The paper specifically claims to be finding evidence that contradicts what is reported in Bell et al. (2008) (which was unpublished at the time Schultz et al. submitted their paper), but a careful reading of Schultz et al. indicates that their methods were not suitable for answering the questions posed by Bell et al. (2008): Bell et al. (2008) looked for a weekly cycle in large-scale averages instead of at single sites; the fact that the years and geographical locations studied by Schultz et al. (2007) were different from what was studied by Bell et al. (2008) made their apparent criticism even less relevant.

We agree with the reviewer's comment. In fact, in the submitted manuscript we already specified that Schultz et al. (2007) used a different time period as compared with Bell et al. (2008), which emphasized the main weaknesses of Schultz et al. (2007) conclusions. Nevertheless, we have slightly modified the sentence in order to include some details suggested by the reviewer: "…They also claimed to find a non-significant weekly signal during the summer period, in

contrast to the results reported by Bell et al. (2008), which was unpublished at the time Schultz et al. (2007) submitted their paper, for the south-eastern U.S. Bell and Rosenfeld (2008) published a comment on the paper of Schultz et al. (2007) indicating that the two studies by Bell et al. (2008) and Schultz et al. (2007) used different time periods and are thus not comparable, as well as the former was designed to study weekly cycles in large-scales averages instead of single sites as in the latter one.".

14) P. 13, ll. 6–10: The paper by Kim and Roh (2010) raises some interesting possibilities based on their (unfortunately quite complex) CSEOF analysis. It should be noted that the near-7-day cycle they discover with their analysis is not an exact 7-day cycle. The peak in the spectrum of their second CSEOF (which explains relatively little of the variability in the data) occurs at a frequency of around 1/(7.5 days). An oscillation at this frequency reverses its phase every 4–5 weeks – in other words, this cycle could be distinguished from a true weekly cycle after a month has passed. The lesson I would prefer to take away from the paper by Kim and Roh (2010) is that there is more "natural" variability in 7-day cycles than one might be tempted to assume, which alerts us to the possibility that testing for the statistical significance of a weekly cycle should be done using methods that capture this extra variance. This in itself is a valuable observation. I believe that methods that break the data into weeklong chunks are implicitly satisfying this requirement. It's my opinion, however, that the analysis of Kim and Roh (2010) does not support the statement that there is a "natural" weekly cycle in the data big enough to cause confusion over the reality of an anthropogenic weekly cycle.

We have modified the last sentence in order to include the reviewer's comment: "…They found a near-7-day oscillation pattern which they claimed to be the result of "natural" weekly cycles (see Section 4.1 for more details), although the robustness of the method needs further scrutiny as has been previously pointed out."

15) P. 14, ll. 8 –15: The paper by Bäumer and Vogel (2007) seems to be using a statistical approach that is very dangerous: they first examine their day-of-the -week averages to find the most interesting anomalous difference among the averages, and only then do they try to estimate the statistical significance of the chosen anomaly. This is a classic pitfall afflicting many statistical meteorological studies: the researcher first scans his/her data to look for something interesting, and then, a posteriori, estimates its statistical significance. In reality the researcher has carried out innumerable statistical tests (in a seat -of -the-pants, informal way, perhaps), and reports only the significance of a single one of the tests -- the one that is guaranteed to be most striking. Such an approach is obviously subject to the problem of the false discovery rate.

Many papers on the weekly cycle first get 7 day-of -the-week averages, find the 2 averages among the 7 that are furthest apart, and then evaluate the statistical significance of the biggest difference. In reality, though, the researcher has examined 21 pairs of days and reports the statistical significance only for the pair most likely to test out as significant. The real statistical significance of the difference is much less than what is calculated for the selected anomaly. Wilks (2006) describes this problem in considerable detail, suggesting some methods for taking into account the effect of the false discovery rate on the estimate of the "global" statistical significance.

We agree with the reviewer's comment. In fact, Section 3 points out some of the deficiencies that are explained by the reviewer. Regarding the work by Bäumer and Vogel (2007) we slightly modified the text in former p. 14, ll. 8 –15, which reads as follows: "Their analyses, which are only based on applying a one-tailed t-test to the days with the larger anomalies in their day-of-the-week averages, showed a general tendency towards …".

16) P. 15, l. 3: I t is true that station-to-station spatial correlations can be small, suggesting that the number of independent samples is about equal to the number of stations. I believe this kind of argument is faulty. Because one is examining the weekly cycle at each station, this is in effect applying a time filter to the data that emphasizes data variability in the neighborhood of frequencies of 1/(7 days). Thus it is not the spatial correlation of the raw, daily data that matters in inferring the amount of spatial dependence of the data. It is the spatial correlation of the time-filtered data that matters, and the spatial correlation of the "slow" components of the data is almost always much larger than the correlation of data with all time scales present. For an extreme version of this, see the paper by Morrissey (1991), who shows that the monthly averages of Pacific atoll rain-gauge data are correlated over hundreds of kilometers, whereas daily data are not.

We have specified that Bäumer and Vogel (2008) refer to the daily precipitation in their reply. Regarding the interesting reviewer's comment concerning the interest of the spatial correlation of the time-filtered data (not the raw daily series) in the neighborhood of frequencies of 7 days, we have added a new sentence in Section 5: "Nevertheless, it is crucial to distinguish the correlation distances of daily series from the much large correlation distances if weekly time scales is considered".

17) P. 15, ll. 4 –20: The paper by Laux and Kunstmann (2008) seems to use post-hoc selection of which anomalies to test, which can be dangerous, as discussed in item (12) above. It's possible that their Monte Carlo estimation of statistical significance may have saved them from this problem; I'm not sure. It is perhaps because of this that they tend not to find cases with statistically significant differences.

We partially agree with the reviewer, but we did not included any new comment in this paragraph as we think that the Monte Carlo method may have saved, at least in some way as the reviewer suggested, the analysis from the problems explained in item number 12.

18) P. 15, ll. 21 –25: Barmet et al. (2009) carry out an interesting analysis using a Kruskal-Wallis (K-W) test for the presence of anomalous differences in day-of -the-week averages. I believe this type of analysis avoids the trap of examining the day-of -the-week averages first and choosing post -hoc the pair to be tested. Because the test makes minimal assumptions about what the anomaly might be, however, it is also weak, in that it requires a particularly strong anomalous difference to be present in order to pass the threshold set by the K-W test. PM weekly variations are fortunately strong enough and regular enough that their cycle is revealed even by the K-W test. Barmet et al. (2009) also include a nice illustration of how the *post-hoc t*-testing of differences produces far more (spuriously) significant anomalies than the K-W test.

We have changed the paragraph in order to not overemphasize the robustness of the K-W test, but clarifying that Barmet et al. (2009) shows how the K-W produce less spuriously results than the *t*-testing: "They also introduced different statistical techniques for significance testing of WCs such as the non-parametric Kruskall-Wallis test, showing how this test produces less spuriously significant results than parametric *t*-test."

19) P. 16, ll. 1 –7: The paper by Quaas et al. (20009) is an interesting attempt to investigate what models can tell us about the detectability of weekly cycles, but I'm not sure at this stage whether GCM's have sufficiently good representations yet of the physical effects of aerosols on clouds nor of the emission and transport properties needed to emulate real aerosol distributions. Theirs is a good effort, however, helping us to focus on what problems still need solutions. It also contains the first published account that I am aware of about detecting the weekly cycle in aerosol concentration using MODIS satellite data—something quite remarkable in itself, given the limitations of the MODIS algorithms over land.

We have included a sentence in Section 5 about the limitations of the current GCM in the parameterization of the indirect aerosol effects: "Nevertheless, it must be noted that current state of the art GCM lack of a good representation of the indirect aerosol effects (e.g. Solomon et al., 2007)".

Regarding the use of MODIS data we think that the first study using this product was published by Xia et al. (2008) as commented in Section 4.2.2.

20) P. 17, l. 21 to P. 18, l. 16: There is another problem with the use of the t-test that should be mentioned. Many authors who use this test are implicitly examining 21 pairs of day-of -the-week averages and picking the pair with the largest difference in the averages. They then, a posteriori, carry out the t-test on that pair. They are therefore subject to the effects of the false discovery rate (FDR), and should have adjusted the significance level they ascribe to their finding to take the FDR into account. This problem is further compounded when many sites are studied, since the investigator may be selecting sites based on whether they will show "significant" weekly cycles.

We have mentioned the reviewer's comment in this paragraph of the revised manuscript: "Equally, most studies that use the t-test are implicitly examining 21 pairs of day-of-the-week averages, and then they perform a posteriori t-test on the pair with the largest differences in the averages. This approach suffers of the effects of the FDR, which should be taken into account to evaluate of the significance level of their findings."

21) P. 18, ll. 20 –26: I'm not sure I agree with this analysis. One does not need to prove that the average for day 1 is significantly different from zero and the average for day 2 is significantly different from zero and the difference of day 1 from day 2 is significantly different from zero. All one needs to prove is that the difference is significant. The proper way to do this is to use something like a *t*-test for the difference time series. If the data for the two days were independent and long enough, one could estimate the expected variance of the day-1-day-2 difference as $\sigma^2 (1 - 2) = \sigma_1^2 + \sigma_2^2$. A difference would be significant at the 5% level if it is larger

than 1.96 σ(1 –2). However, this only works if one has chosen a priori which pair of days to compare. Otherwise one is still subject to the FDR problem.

Following the reviewer's suggestion, and in order to clarify the text, we have deleted this item 2 in Section 3.

22) P. 19, item (3): To me this is an excellent opportunity to point out that the results by Gong et al. (2006) could be seriously affected by the FDR problem. They select which sites to test for significance post hoc, and more subtly but perhaps worse, they examine the size of the weekly cycle in all seasons and then, post hoc, choose to concentrate on the significance of the summer data alone. As the review authors say, too, the fact that the weekly cycle seems to be concentrated in portions of China with some spatial coherence does not lend any extra credibility to the results Gong et al. obtain, because the spatial correlation will tend to make "significant" anomalies cluster like this. It should also be noted, as was noted before, that it is not the spatial correlation of the daily data that matters – it is the spatial correlation of the variations at the time scale of 1 week that should be used to evaluate how much spatial correlation there is.

In this item 3 we have added a sentence in order to include the reviewer's comment regarding the study by Gong et al. (2006): "This approach could seriously affect the analysis through the FDR problem as the sites and/or seasons used to examine the significance are selected post hoc."

23) P. 20, ll. 21 –22: It appears that the statement that the K-W test is a test for equality of the medians is incorrect. Rather, the test detects whether the mean ranks of the different sets are the same. See the example given in the "Handbook of Biologic Statistics" at this URL:
http://udel.edu/~mcdonald/statkruskalwallis.html

We have corrected the error.

24) p. 21, ll. 10 –13: This observation may need additional qualifications, since if one tests for cycles at different frequencies one is inviting the problems of FDR into the picture. As one enlarges the scope of tests on the data, one may have to adjust the significance testing to compensate for the FDR.

We have added a sentence at the end of the paragraph that reads as: "…., although it is worth noting that testing for cycles at different frequencies should imply an adjustment of the significance test to compensate for the FDR."

25) P. 21, l. l4: Perhaps this is a good time to remind the readers of this review, and to remind authors to remind their readers in their papers, that a null statistical result does not "prove a negative." It only tells us that the particular statistical test could not distinguish the signal it was designed to look for from natural variability. As Hasselmann (1997) pointed out, a poorly designed statistical test for climate change can produce a null result when a better -designed test might have revealed something. See also Bell (1986) for a discussion of this point.

We have added a final sentence in Section 3 including the reviewer's suggestion: "Finally, it is important to remind that a null statistical result does not prove the non-existence of a significant

WC, as a lack of well-designed statistical test can produce a null result when a better-designed might have revealed a significant one (e.g. see Bell, 1986; Hasselmann, 1979)"

26) P. 22, ll. 16 –20: Quasi-cyclic behavior on a weekly time scale of the atmosphere at synoptic scales just implies that there is more randomness or "noise" in the weekly cycle than one might guess from the amount of day-to-day variability. With a long enough dataset this extra "noise" should be capable of being captured and significance tests will naturally include it. For example, most of the analyses described in papers by Bell et al. break the data up into weeklong chunks, and, I believe, the variances they estimate probably capture a lot of the "natural" variability of the weekly cycle. (I'm not referring here to the Monte Carlo studies in those papers, however.)

We completely agree with the reviewer's comments, but in this sentence we only summarized one of the possible causes suggested by Forster and Solomon (2003), which on the other hand they did not consider as plausible for their analysis. The reviewer's concerns about the possible "natural" weekly cycles and the extra randomness or "noise" in the weekly cycles have been included in the discussion of the results published by Kim et al. (2010) and Kim and Roh (2010).

27) P. 22, l. 20 to p. 23, l. 10: As I mentioned earlier, I have considerable misgivings about the way the CSEOF technique is used by Kim et al. to support their conclusion that there are significant amounts of Rossby waves with an exact 7-day period that can be confused with an anthropogenic cycle.

We have modified the text according the reviewer's suggestion, as in Section 2 when discussed Kim et al. (2010) and Kim and Roh (2010) papers.

28) P. 23, ll. 11–12: I should mention that in the paper Bell et al. (2009), in its Supplementary Section, we tested for the presence of weekly cycles in synoptic-scale variability and could detect nothing significant. But I would not claim that the tests of this possibility were exhaustive.

We have added a new sentence in this section: "On the other hand, Bell et al. (2009a, Supplementary Material) did not find any evidence of weekly cycles in the synoptic-scale variability over North America."

29) P. 24, ll. 15 –17: It's my impression that the "indirect effect" of aerosols is usually meant to refer to the changes in the radiative interaction of the polluted cloud. I don't believe that it includes the thermodynamic effect, though perhaps the research field has enlarged its definition of the "indirect effect" to include this.

We consider that the thermodynamic effects of aerosols in clouds are considered as an indirect effect of aerosols. We have modified the sentence that now reads as: "In fact, the direct and, especially, indirect effects of anthropogenic aerosols have been suggested as the main cause for WC´s, with some studies showing evidences of this connection, i.e. mainly linked to the indirect effects of aerosols including the thermodynamic effects (e.g., Bell et al. 2008, Rosenfeld and Bell, 2011).

30) P. 30, l. 2: I would say, "… are in general weak at small scales (i.e., at individual sites)."

We have added the reviewer's suggestion.

31) P. 30, l. 7: Note that Tuttle and Carbone (2011) have done just this.

In the revised manuscript we have pointed out that Tuttle and Carbone (2011) already used radar data in their analyses.

32) P. 30, ll. 13 –14: Note, however, that one needs to distinguish the correlation distances of daily precipitation from the much larger correlation distances of precipitation at weekly time scales.

We have included the suggestion in this paragraph that now reads as: "However, for precipitation there is still some potential to explore datasets of regions with a very dense network of rain gauges, as well as its shows a weaker spatial autocorrelation than other meteorological variables. Nevertheless, it is crucial to distinguish the correlation distances of daily series from the much large correlation distances if weekly time scales is considered."

33) P. 30, ll. 14 –18: Studies attempting to detect weekly cycles on a site-by-site basis need to take into account the effects of the FDR as well. The remark, "taking into account the non - normality of the data," is probably an implicit recommendation to move from the t-test to the Kruskal -Wallis test, but it should be noted that when enough data are available the t-test can become insensitive to the non-normality of the data and can legitimately be used – unfortunately, there do not appear to be well-developed statistical guidelines for this. Also, the K-W test is weaker than the t-test at detecting anomalies, so one should be sure that one is satisfied with giving up the statistical power of the t-test. Use of Monte Carlo methods is highly advisable if they're feasible, but the user must be alert to the problem of resampling the data without destroying the important correlations.

Following the reviewer's comments, we have modified and shortened the paragraph in order to clarify the text: "For such regional studies that consider measurement data from several observation points, it is important that statistical testing is done taking into account the effects of the FDR, as well as the spatial and temporal autocorrelation in the data".

34) P. 31, ll. 1 –4: Following this suggestion requires that the researcher figure out what an "unexpectedly large number of grid points" might be. Wilks (2006) provides some suggestions about this. Some readers may find the analysis in Bell et al. (2009a) (in its Supplement) interesting, because the number of grid points significant at various levels is compared with the expected number, thereby providing some guidance about what fraction of the grid points that pass the naïve "significance" test might be real anomalies.

We have included the suggestion in this paragraph that now reads as: "Therefore it is important to extend the study for large spatial areas and to verify that for an "unexpectedly" large number of grid points a significant WC would be found, which will really support the existence of large-scale WCs. In Bell et al. (2009a, section 2 in the Supplementary Material) there is a clear example of this approach in order to differentiate between "real" weekly cycles and noise in a

gridded dataset. They compared at different grid resolutions the number of grid points significant at different levels with the expected number."

35) P. 31, ll. 5–12: It's obvious, but we also need much more modeling on the scale of individual storms (mesoscale modeling) and the development of good parameterizations for the effect of aerosols on cloud dynamics. And we need more microphysical studies as part of field campaigns to try to disentangle the complex feedback processes in aerosol and cloud dynamics.

We have added a sentence following the reviewer's suggestion: "Equally, there is a need of a development of the modelling on the scale of individual storms (mesoscale modelling) and improvement of the parameterizations of the effect aerosols on cloud dynamics and microphysics."

36) P. 32, ll. 3 –6: As I mentioned earlier, I think there are legitimate concerns that some of the studies of data (Gong et al. 2006; Choi et al. 2008; Ho et al. 2009) that have been done using data over China may be less conclusive than they seem at first because they have not taken into account the FDR. The paper by Gong et al. (2007) appears to me to be on solid ground, however.

We have added a new sentence in order to highlight the concerns raised by the reviewer: "Nevertheless, some concerns remain regarding the statistical analysis used in some of these studies: for example, some of them have not taken into account the FDR (e.g. Gong et al. 2006; Choi et al. 2008b; Ho et al. 2009)."

37) P. 32, ll. 15 –20: I would add that statistical methods should be used that are better designed to detect weekly cycles in large-scale averages, and papers that discuss results on a site by site or season by season basis without specifying choices a priori should correct their significance testing for the effects of the false-discovery -rate (FDR) problem.

Done, the new paragraph reads as: "The major weaknesses in the studies focusing on the WCs topic are linked to 1) neglecting spatial autocorrelation of the data, which if considered will reduce the number of independent series and the degrees of freedom; 2) the need of a better design of the methods in order to detect weekly cycles in large-scale averages; 3) the assumption of the normality in the series, which it is not always present in the climate data, especially when dealing with daily resolution; 4) the necessity of correcting the significance testing for the effects of the FDR problem if the analysis are performed on site by site or season by season basis without specifying choices a priori; and 5) the publication bias towards papers reporting significant results, which are more likely to be published than papers showing non-significant WCs."

38) P. 41, Table 2: Cloud lifetime effect: If clouds persist longer, shouldn't that result in increased cloudiness?

The reviewer is right but the stable states what we expect on weekends and not on weekdays. Consequently, we keep the "-" sign in the revised manuscript.

39) P. 41, Table 2: Although there's much to be worked out, I believe that at present the thermodynamic effect is believed to produce larger anvil clouds and perhaps more storms, so midweek cloudiness is expected to increase (and this shows up in the MODIS data, as mentioned in Bell et al. 2009b). I'm not sure what the impact on surface solar radiation, surface temperature, or diurnal temperature range might be, since it depends more subtly on how the cloud structure changes.

We agree with the reviewer's suggestion, and consequently a "-" symbol (expected increase on weekdays and decrease on weekends) had replaced the former "o" (neutral effect) in Table 2 for the weekly cycle expected for the thermodynamic effect on cloudiness.

Minor comments

40) p. 9, l. 11: I would say "suggested" rather than "argued".

Done.

41) P. 11, l. 7: Should be "Forster"?

Yes. The error has been corrected.

42) P. 12, l. 16: Define "asl".

We have replaced "asl" by "above sea level".

43) P. 13, l. 7: Should be "Roh"?

Yes. The error has been corrected.

44) p. 13, l. 9: Define "EOF": "Empirical Orthogonal Function".

Done.

45) P. 18, l. 14: "extent"?

Yes. The error has been corrected.

46) P. 22, l. 3: How about saying, "There is no known natural process …"?

Yes, we have rewritten the sentence using your suggestion.

47) p. 24, l. 13: More precise than "East coast of North America" might be "coastal Atlantic near North America."

Ok. Done.

48) P. 32, l. 3: “a large number of evidences have” —> “a large amount of evidence has”.

Done.

49) P. 32, l. 24: Add an apostrophe: “ aerosols’ ”.

Done.

References

1. Bell, T. L. (1986): Theory of optimal weighting of data to detect climatic change. J. Atmos. Sci., 43, 1694–1710.
2. Hasselmann, K. (1979), On the signal-to-noise problem in atmospheric response studies, in Meteorology Over the Tropical Oceans, edited by D. B. Shaw, pp. 251 –259, R. Meteorol. Soc.

3. Morrissey, M. L. (1991), Using sparse raingages to test satellite -based rainfall algorithms, J. Geophys. Res., 96 D, 18,561–18,571.

4. Wilks, D.S. (2006), On “field significance” and the false discovery rate, J. Appl. Meteor. 655 Climatology, 45, 1181-1189.

References 1, 2, and 4 have been added in the revised manuscript.