Identifying weekly cycles in meteorological variables: The importance of an appropriate statistical analysis

The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.
Identifying weekly cycles in meteorological variables: The importance of an appropriate statistical analysis

J. S. Daniel, R. W. Portmann, S. Solomon, and D. M. Murphy

Received 3 February 2012; revised 30 April 2012; accepted 9 May 2012; published 4 July 2012.

[1] Weekly cycles in several meteorological parameters have been previously reported. Yet the extent to which these cycles are caused by anthropogenic activity remains unclear. Some of the complications associated with establishing this link are discussed here. Specifically, we highlight and quantify some common errors that have been made in the application of statistical techniques to this problem. Some errors, including the inappropriate use of the Student t test, have been significant enough to affect the conclusions of previous studies. A resampling technique that can properly account for both temporal and spatial correlation is evaluated and is shown to be accurate for determining the statistical significance of weekly cycles at the station level and for evaluating total field significance. We demonstrate that this resampling approach performs comparably to a Fourier analysis that evaluates the significance of the power at a seven-day period. Regardless of the analysis technique used, an understanding of the behavior of and uncertainties associated with the statistical analysis is critical to arriving at a justifiable conclusion regarding a human influence on weekly cycles and for putting results in context with other studies. We also discuss some general errors that can be made in weekly cycle analysis. These include selection of an analysis region after identifying where weekly cycles are significant, acceptance of a physical explanation for the hypothesized link that has not been properly tested given its large number of degrees of freedom, and ignoring the correlation among meteorological parameters.

Citation: J. Geophys. Res., 117, D13203, doi:10.1029/2012JD017574.

1. Introduction

[2] The potential relationships between human activities and weekly variations in weather have been investigated for over eight decades [Ashworth, 1929]. In earlier studies, motivation included simple curiosity and a desire to understand if the weather were really different on weekends when people had more time to engage in outdoor activities. More recent studies are often motivated by the idea that an understanding of human-caused weekly cycles in meteorological variables could lead to additional insight into the role of aerosols and other anthropogenic effects on climate. The expectation is that any consistent variability on time scales as short as a week must be caused by substances or processes with similarly short atmospheric lifetimes. Thus, aerosol direct and indirect effects are thought to be likely candidates. If the impact of human-induced weekly variations in aerosol concentrations on surface temperature (T) and precipitation were to be quantified, for example, this information might improve the understanding of these interactions at the process level as well as to constrain models in their attempt to explain past and future aerosol forcing of climate and the climatic response to this forcing. Quaas et al. [2009] have taken some initial steps in this direction, finding that when forcing their models with weekly cycles in anthropogenic emissions, they could simulate a weekly cycle in cloud droplet number, for example, but not in other meteorological variables. In addition to the possible role of aerosols, it has also been hypothesized that some weekly cycles in urban areas may be caused by weekly variations in the direct release of heat from human activities [Fujibe, 1987; Simmonds and Kaval, 1986; Simmonds and Keay, 1997].

[3] The earliest studies of weekly cycles focused exclusively on precipitation, with T, daily temperature range (DTR), and other meteorological parameters investigated later. In several papers, Ashworth [1929, 1933, 1944] suggested that the weekly variations in smoke emissions from factories led to weekly precipitation variability in Rochdale, England. He attributed the drier Sundays to either the hot emissions during the week that caused instability and rain, or to an aerosol effect on clouds that led to precipitation. A comment by Napier Shaw found at the end of Ashworth

---

1Chemical Sciences Division, ESRL, NOAA, Boulder, Colorado, USA.
2Department of Earth, Atmospheric and Planetary Sciences, MIT, Cambridge, Massachusetts, USA.

Corresponding author: J. S. Daniel, Chemical Sciences Division, 3A116 ESRL, NOAA, Boulder, CO 80305, USA. (john.s.daniel@noaa.gov)

This paper is not subject to U.S. copyright.
Published in 2012 by the American Geophysical Union.
While Ashworth did not investigate whether Laux and Stjern is a human construct, it does not follow that the Rosenfeld and Bell Gordon Walshaw Forster and Solomon Nicholson Hendricks Franssen, 2006, 2007; Definitions of the weekly cycle magnitude and ever, most of the subsequent researchers recognized that cycles [this problem [recognized [importance of such statistical testing was beginning to be with natural variations before suggesting that the cycle is the presence of a weekly cycle; rather, it is necessary to statistic-ies were leading to a weekly cycle in aerosols, as evidenced by the Royal Air Force altering its training schedule because of the high industrial emission levels of Rochdale, and other papers during this period also addressed weekly cycles in precipitation, but most of these considered much shorter time periods and provided little assessment of the statistical significance of the cycles [Carter, 1931; Met Office, 1932]. However, many people were convinced that human activities were leading to a weekly cycle in aerosols, as evidenced by the Royal Air Force altering its training schedule because of the hygroscopic nuclei rather than the high temperatures of the gases.” While Ashworth did not investigate whether the weekly cycles were statistically significant, he did show that hourly relationships between precipitation and emissions over a two-year period were consistent with his hypothesis that emissions led to more rain. He also showed that precipitation on Sundays became more similar to precipitation on other days of the week during the World War I years, when factories were in operation seven days a week. And finally, he showed that data at Stonyhurst, an area without the high industrial emission levels of Rochdale, were characterized by much smaller weekly cycles. Other papers during this period also addressed weekly cycles in precipitation, but most of these considered much shorter time periods and provided little assessment of the statistical significance of the cycles [Carter, 1931; Met Office, 1932]. However, many people were convinced that human activities were leading to a weekly cycle in aerosols, as evidenced by the Royal Air Force altering its training schedule because of the weekly impact of anthropogenic activity on visibility [Ratcliffe, 1953].

In general, it is not enough to simply identify the presence of a weekly cycle; rather, it is necessary to statistically determine whether that cycle is likely inconsistent with natural variations before suggesting that the cycle is the result of some anthropogenic forcing. By the 1960s, the importance of such statistical testing was beginning to be recognized [Hartley-Russell, 1964; Walshaw, 1963]. There was initially some debate over the importance of statistics to this problem [Scorer, 1964, 1969], and some work continued to simply demonstrate the presence of nonzero weekly cycles [Glasspoole, 1969; Nicholson, 1965, 1969]. However, most of the subsequent researchers recognized that meaningful results depended on detailed consideration of statistics to ensure that the results were not due to chance. It has recently been further suggested that even the presence of statistically significant weekly cycles may not imply an anthropogenic cause, but may only be indicative of the natural periodicities of atmospheric Rossby waves [Kim et al., 2010]. Nevertheless, identifying a statistically significant weekly cycle that is inconsistent with random variability is a necessary though not sufficient condition for establishing a link to anthropogenic behavior.

Despite the use of statistical tests over several decades, a consistent picture of weekly cycles and their causes has remained elusive, with many apparently contradictory results. Some data have shown Thursdays to be wettest [Nicholson, 1965, 1969], while another found Thursdays to be driest [Norgate, 1974], and another identified no significant cycle at all [Cehak, 1982]. The confusion continued with many researchers finding significant weekly cycles in various meteorological parameters [Bell et al., 2006; Bell et al., 2008; Cerveny and Balling, 1998; Forster and Solomon, 2003; Gong et al., 2006, 2007; Gordon, 1994; Laux and Kunstmann, 2008; Rosenfeld and Bell, 2011; Shutters and Balling, 2006; Simmonds and Keay, 1997] while others found that the observed cycles are consistent with natural variability [DeLisi et al., 2001; Hendricks Franssen, 2008; Hendricks Franssen et al., 2009; Stjern, 2011]. It has been suggested that disparate results are to be expected because the interactions with aerosols and clouds will vary with time and location [Bell and Rosenfeld, 2008]. It has also been suggested that apparently contradictory results are due to a lack of consistency [Sanchez-Lorenzo et al., 2009] or appropriateness [Bell and Rosenfeld, 2008] in the statistical methods used to analyze weekly cycles. Sanchez-Lorenzo et al. [2012] provide a current review of the weekly cycle literature and also discuss the impact of the varied statistical approaches on our understanding of weekly cycles.

Here, we use simulated T data to examine the performance of several of the statistical tests that have been used to evaluate the significance of weekly cycles. The list of tests examined is not exhaustive, but some lessons regarding weekly cycle analyses can be derived. After assessing these statistical tests, we present the general attributes of a successful statistical analysis of weekly cycles, including analysis of individual stations and a group of stations. We do not analyze meteorological observations here, but we are currently doing this by applying some of the lessons learned from these simulations.

We begin by describing our method for simulating data in section 2. Several statistical tests that have been used in the literature are described in this section as well. The performance of these tests is discussed in section 3. Section 3 also includes a discussion of some of the other critical attributes of a successful weekly cycle analysis beyond the statistical analysis. Our conclusions are found in section 4.

2. Motivation and Methods

It is frequently stated that the existence of a weekly cycle in a meteorological parameter, such as T, implies an anthropogenic influence. While it is true that the “work week” is a human construct, it does not follow that the presence of any weekly cycle must represent evidence of a

![Figure 1. Definitions of the weekly cycle magnitude and the weekend effect magnitude over an eight-day period. The weekly cycle is calculated by subtracting the coolest temperature of the week from the warmest. The weekend effect is calculated by subtracting the average weekday value from the average weekend value. Here the weekend is defined as Saturday through Monday, while the weekday period is defined as Wednesday through Friday.](image-url)
human influence. Natural variations will cause some magnitude of weekly cycle. The definitions used for the weekly cycle magnitude and the weekend effect magnitude in this work are illustrated in Figure 1. The weekly cycle magnitude is defined as the difference between the highest value of the week and the lowest. The weekend effect is defined here as the difference between the mean value over Saturday through Monday and the mean value over Wednesday through Friday. Other equally valid definitions are also possible for defining weekdays and weekends. Both weekly cycle and weekend effect definitions can also be applied for multiple weeks of observations by averaging the values over the entire time period for each day of the week.

Figure 2. Weekly cycle magnitudes arising from random variability in AR(1) time series. (top and middle) The first ten weeks of yearlong time series simulations are shown for lag-1 autocorrelations of 0.1 and 0.6; these particular time series are the ones that are characterized by weekly cycle magnitudes at the 99.9th percentile over these 10 weeks out of 10,000 generated time series. The average weekly cycles over these 10-week periods are superimposed as the red curves. (bottom) The average weekly cycle and weekend effect magnitudes for full years with different levels of prescribed AR(1) temporal autocorrelation. Maximum (minimum) magnitudes represent the 95th (5th) percentile levels and are shown as dotted lines. The purple (red) lines represent the magnitude above which the Student t test would suggest that the weekly cycle or weekend effect is not consistent with random variability at the $p = 0.10$ ($p = 0.05$) confidence level. A one-tailed $t$ test is used for the weekly cycle magnitude and a two-tailed test for the weekend effect. All time series are prescribed to have a unit standard deviation. Note that the ordinate represents the prescribed standard deviation; as the autocorrelation gets close to 1, the sample variance begins to fall significantly below this value (e.g., for a lag-1 autocorrelation coefficient of 0.95, the average sample standard deviation is 0.94).

[9] Figure 2 (first and second panels) shows examples of weekly cycles for random time series. In the first panel, the time series is prescribed to have a lag-1 autocorrelation of 0.1, while the time series in the second panel has an autocorrelation of 0.6. The degree of daily variation both in the time series and in the 10-week averaged weekly cycle is evident, with the time series with the high autocorrelation demonstrating slower changes from day to day. Time series with higher autocorrelation also tend to have smaller weekly cycle magnitudes due to this reduced daily variability. These two panels demonstrate that random time series can, at times, appear to have strong and persistent weekly periodicities even though no physical process is driving the apparent weekly cycle. The weekly cycle can even appear to be related to some consistent weekly forcing. The two bottom panels show the average weekly cycle and weekend effect magnitudes and their 5th and 95th percentile values for single years as a function of temporal autocorrelation. The weekend effect has an average of 0 as expected, because for a random time series it is just as likely that the weekend will be warmer than the weekdays as it is that it will be cooler. On the other hand, by definition the weekly cycle magnitude cannot be negative, and unless all days have the identical value, the weekly cycle magnitude cannot be 0. Because nonzero weekly cycle magnitudes always result from random time series, it is clear that for an analysis of weekly cycles to be most useful, an evaluation must be made...
regarding the likelihood that the observed cycles result from random processes. This evaluation is accomplished through statistical hypothesis testing. The null hypothesis is assumed to be true if there is not convincing evidence to the contrary. For weekly cycle analyses, the null hypothesis is that the observed value of the metric is consistent with random processes, or noise. The hypothesis test then determines the likelihood that the null hypothesis should be rejected. Superimposed on the bottom panels of Figure 2 are hypothesis test threshold levels generated by the Student t test. The purple and red curves represent the threshold magnitudes that, if reached, would imply that the weekly cycle or weekend effect magnitude is inconsistent with random variability at the 90% and 95% confidence level (often referred to as the $p = 0.10$ and $p = 0.05$ levels, respectively).

To calculate these thresholds for the weekly cycle test, first the variances of the warmest and coolest days of the week for each yearlong simulation are determined at each autocorrelation value. If, for example, Tuesday is the warmest day of one simulation and Friday the coolest, the variance of the temperature values for every Tuesday and the variance of the temperature values for every Friday are recorded for that simulation. The same is done for each of the other simulations, with the mean variances over all simulations for that autocorrelation used to determine the threshold temperature levels shown in Figure 2 directly from the definition of the t test statistic. The same is done for the weekend effect hypothesis test thresholds except that the variances of the weekend and weekday temperature values are used.

Figure 2 (bottom left) shows that for autocorrelation values less than 0.65, even the average weekly cycle magnitude for the random time series is larger than the 90% threshold. In other words when using the t test in this way it is typical, not unusual, for the weekly cycle magnitude to appear to be statistically significant and not consistent with random variability. Another way to understand the permissiveness of this result is that over a period of 7 days, there are 21 unique pairs of days that could be tested with the t test to see if they are significantly different. The pair demonstrating the greatest difference is generally the one representing the weekly cycle magnitude test, and this is the only t test that is considered. So it is not surprising that this pair is often found to be statistically different. Thus, the use of the t test frequently leads to extremely inaccurate and misleading results. In sharp contrast to the weekly cycle, the t test for the weekend effect behaves as expected in cases of low temporal autocorrelation (Figure 2, bottom right). When using the two-tailed test, appropriate when there is no a priori knowledge about which sample should be bigger than the other, 5% of the random time series have weekend effects larger than the upper 90% threshold and 5% have weekend effects smaller than the lower threshold. As the autocorrelation is increased, fewer simulations have weekly cycle magnitudes larger than the thresholds, implying that the t test becomes conservative for the weekend effect, very different behavior from the permissive nature of the t test when analyzing the weekly cycle magnitude.

Ideally, one would not rely only on a statistical analysis, but would incorporate a process-based understanding of aerosol-cloud interactions, for example, that could lead to a physically based expectation of how meteorological parameters would vary on a weekly time scale [Bell and Rosenfeld, 2008]. However, the complicated nature of aerosol-cloud interactions currently makes such an approach impractical in many cases. In the absence of a well-understood physical mechanism, the reliance on a robust statistical approach is necessary for a convincing argument for anthropogenically caused weekly cycles or weekend effects. While it is possible to conduct a purely statistical hypothesis test to determine if a weekly cycle or weekend effect could be due to natural variability, it is not simple and as shown in Figure 2, inappropriate tests can lead to misleading results. Many types of statistical tests have been used to assess the significance of weekly cycles. Some have been used appropriately and some have not. Several commonly used approaches are discussed in the next section, including the Student t test, analysis of variance, a non-parametric test, and bootstrap analyses. If weekly cycles from multiple stations are analyzed, it is further necessary to determine whether the level of statistical significance across all the individual stations is likely consistent with random variability or whether it could imply an anthropogenic forcing process. As discussed below, this analysis is generally complicated by spatial correlation among stations.

2.1. Time Series Simulations

We use simulated T time series to evaluate the strengths and weaknesses of several statistical approaches that have been used to analyze the significance of weekly cycles in meteorological data. An AR(1) autoregressive model is used with a constant prescribed variance, mean, and lag-1 autocorrelation value to simulate the random component of time series at individual locations. Other models of noise could also have been considered, for example higher order autoregressive models or autoregressive moving average (ARMA) models, but it is not expected that they would qualitatively alter our conclusions. The AR(1) lag-1 autocorrelation value is varied to explore the sensitivity of the various methods to this factor. An autocorrelation lag-1 value of 0 implies no temporal autocorrelation while a value near 1 implies a very strong dependence of a value at some time on the value at the preceding time step. For reference, we find that observations of daily mean T across the United States (U.S.) are characterized by lag-1 autocorrelation values between 0 and 0.8, with an U.S. average annual value of about 0.6, consistent with autocorrelation values at selected stations shown in Vinnikov et al. [2008]. We do not include a seasonal cycle in our simulations, nor do we include any potential seasonal change in variance. When analyzing observational data, one could remove the seasonal cycle with various approaches, including high pass digital filtering, some other type of smoothing, for example the LOWESS (locally weighted scatterplot smoothing) approach [Cleveland, 1994], or one as described in Cerveny and Coakley [2002]. One could also reduce the lag-1 autocorrelation of the time series using the equation

$$x'_i = x_i - rx_{i-1}$$

(1)

where $x'$ is the time series with autocorrelation reduced, $x$ is the original time series, $i$ is the time step increment, and $r$ is the lag-1 autocorrelation value. We do not attempt to reduce the autocorrelation here, however, because we prefer that the analyzed time series be as relevant to the actual time series as possible. While it is necessary to remove low-frequency
variations over, for example, seasonal and annual time scales [see, e.g., Cerveny and Coakley, 2002], reducing the auto-
correlation can lead to weekly cycles that differ noticeably
from the actual cycle, thus making the physical interpreta-
tion of the results less straightforward.

[13] Many studies have analyzed time series from multiple
stations when searching for regional or urban influences on
weekly cycles. To evaluate the statistical methods used to
determine whether a group of stations demonstrates signifi-
cant weekly cycles (referred to as field significance), we
generate multiple time series that are characterized by both
temporal autocorrelation as well as spatial correlation. These
time series are generated using the approach described in
Khalili et al. [2007].

2.2. Statistical Analyses Assessed

[14] In order to evaluate the significance of a weekly cycle
or weekend effect magnitude, hypothesis testing is per-
formed. If the evidence is so convincing that there is
believed to be a very small chance (less than 5%, for
example) that the observed behavior could result from a
random time series, the null hypothesis is rejected, and the
weekly cycle is said to be significant at the $\alpha = 0.05$ level, or
the 95% confidence level, or is said to be significant with a
$p$-value of 0.05. The level at which processes are generally
considered significant, and the null hypothesis rejected,
varies, but is typically at the $p = 0.01$ or $p = 0.05$ level.
Several researchers have also used the 0.10 level. The choice
of significance level represents a balance between the desire
to have it permissive enough to try to capture as many “real”
weekly cycles as possible while having a value small enough
to reduce the likelihood of experiencing a false positive
identification caused by random variability. This type of
hypothesis testing must be performed for each individual
station, and for all stations as a group if multiple stations are
considered. It should be noted that the failure to reject the
null hypothesis does not generally imply that there is no
anthropogenic process that is leading to a weekly cycle in
some meteorological variable. It only proves that the
response is not large enough to identify it as statistically
significant; it may or may not exist. Similarly, rejection of
the null hypothesis does not assure that the cycle is not
caused by random variability.

[15] In addition to the analysis of weekly cycle magni-
tudes and the weekend effect, a third approach could be
considered that would involve determining which day of
each week has the highest or lowest $T$ for many weeks in a
time series. It could then be determined if any particular day
of the week was more or less likely to have an extreme
value. Luckily, this approach is almost never used, as it has
been shown that a biased number of ‘hits’ at the beginning
and end of each week will result when the original time
series is temporally autocorrelated [Coakley, 2000]; the
effect exists for any choice of the day on which the week
begins and is due to the increased likelihood of experiencing
an increasing or decreasing trend across entire weeks com-
pared with a time series with no autocorrelation.

2.2.1. Local Significance

[16] One common test used to evaluate local significance
is the Student $t$ test, often referred to as simply the $t$ test. This
test can be used to determine whether two independent
variables have means that are different. It requires that the
two means tested originate from independent data and
technically that both distributions be normally distributed;
however, it can often also be successfully applied to distri-
butions that do not deviate too dramatically from normal.
The $t$ test has been used to evaluate weekly cycle magnitudes
and the weekend effect and can be applied as a one- or two-
tailed test depending on whether or not it is not known a
priori which one of the variables should be bigger. If it is
known that one should be bigger, a one-tailed test can be
used. However, the improper use of a one-tailed test when a
two-tailed test should be used will result in assigning too
much significance to the difference between the two vari-
able. If a one-tail test is used, as in Gordon [1994], explicit
justification of its applicability should be provided.

[17] Analysis of variance (ANOVA) is another approach
to determining whether there is a significant weekly cycle.
It tests the null hypothesis that the meteorological quantity
for every day of the week is taken from the same distribu-
tion. A rejection of the null hypothesis implies that the value
on at least one day is statistically different from the others
at some level of confidence. As with the $t$ test, the analysis
of variance approach requires that the data for each day of
the week be independent. Temporal autocorrelation, as found
in most meteorological data, leads to a violation of this
requirement, with the magnitude of the error induced partly
determined by the degree of autocorrelation.

[18] The resampling technique (also referred to as the
bootstrap technique [Efron, 1979]) has been frequently used
as a way to account for temporal autocorrelation as well as
with data that are not normally distributed. “Resampling”
refers to the process of generating new “random” time series
directly from the original time series. This allows the resam-
pling approach to approximately account for temporal auto-
correlation and to reproduce the distribution of the original
time series, allowing the approach to be useful in more situa-
tions than many standard statistical approaches. One common
implementation, which is used here, involves choosing a
block length, $n$, and then randomly picking the starting loca-
tion for that particular block in the original series. The $n$
points are taken and placed (as an entire block) in the resampled
series. This process is repeated until the resampled time series
has the same length as the initial time series. It is performed
“with replacement,” allowing the value at a particular time
in the original time series to possibly appear multiple times or
not at all in the resampled time series. The purpose of the
resampling is to scramble any weekly coherence that might
have existed in the original series so that the new time series
are characterized by magnitudes of weekly cycle and weekend
effects that would be caused by random variability. The dis-
tribution of the weekly metrics (e.g., weekly cycle magnitude,
weekend effect) can be obtained from the large number of
resampled time series, and threshold values can be determined
de that define the range in the metric that encompasses some
large fraction of the resampled cases; specifically that fraction
is $1-p$, where $p$ is the hypothesis test level (section 2.2). If the
metric in the original time series falls outside this range,
the null hypothesis is rejected and it is considered unlikely that
the weekly cycle, for example, was caused by natural vari-
ability. A more complete discussion of the resampling
approach can be found in Wilks [1997].

[19] Fourier analysis can also be applied to weekly time
series. If there is a significant weekly periodicity, one would
expect that the power spectrum would show significant power at a period of 7 days. This will be discussed further in section 3.1.3 in comparing with the resampling results. It should be noted that the Fourier approach used here is different from that applied in Bell et al. [2008] and Bell et al. [2009] both in terms of identifying the magnitude of the cycle with a periodicity of seven days and of applying the hypothesis test.

### 2.2.2. Field Significance

When analyzing data from multiple stations, it is necessary to determine whether the results imply that there is a significant effect when considering the entire group of stations. While a single station in the group may exhibit a weekly cycle that is significant to a very high confidence level considered alone, if there are many stations, this behavior is more likely to occur simply from random chance. For example, if 1000 independent stations are evaluated, there is a 37% chance that one station would be significant at the $p = 0.001$ level, a level of very high significance if a single station were analyzed. The larger the number of independent stations, the more likely it is to have one or a few stations that are characterized by what would be considered a significant weekly cycle if analyzed alone. It is therefore necessary to account for the number of observations and the spatial correlation among observations to evaluate the field significance. The null hypothesis for a field significance test is that all the stations considered have weekly cycles consistent with random variation in their time series. If the null hypothesis is rejected, the weekly cycle in at least one station is so large that it is very unlikely that it could be caused by natural variability. It is also important to recognize that the relevant spatial correlation referred to here is the high frequency (e.g., 1/7 d$^{-1}$) correlation. This may be quite different from the correlation of temperature data that has not been filtered to remove, for example, the seasonal cycle.

We consider two types of field significance tests here. One is called the “false discovery rate” test (FDR) [Benjamini and Hochberg, 1995; Wilks, 2006] and the other has been referred to as the “counting” test [Livezey and Chen, 1983]. We consider both of these because of their complementary nature. Both tests require an evaluation of the statistical significance of the weekly cycles at each site. We perform this using a resampling technique discussed above.

In the counting approach, the number of stations that have a rejected null hypothesis at the local level is determined. If the number is higher than the determined threshold level, the field is significant, implying there is evidence that a true weekly cycle exists at some station in the field at that level of confidence [Livezey and Chen, 1983]. The significance level chosen for the field significance test does not need to be the same as that for the local significance tests. For example, the field significance could be tested at the $p = 0.05$ level when the individual stations are tested at the $p = 0.10$ level. The field threshold is determined from the resampled time series generated for the local significance test. The identical scrambling sequence is used at every station to ensure that the spatial correlation of each resampled set of time series is consistent with the original time series. From these series, a distribution of the number of significant stations in the field is built and a threshold can be found where some large fraction of the resampled fields had fewer significant stations than that threshold value. The counting approach is effective at successfully rejecting the null hypotheses when there are many stations with even small weekly cycle signals relative to noise levels. It is less effective at identifying statistical significance at a very few stations even if those few stations have highly significant cycles. Thus, this test would be expected to be a particularly useful test if anthropogenic behavior influences meteorological variables over large areas and not just in a few isolated locations.

If, on the other hand, human impacts occur at only a few locations, for example near the largest urban areas, the FDR approach might be preferred [Wilks, 2006]. It is able to discern whether only even a single station has a significant weekly cycle or weekend effect when considering the total number of stations analyzed. The FDR implementation we use involves examining the significance of the weekly cycles at each station and determining whether the distribution of $p$ values (significance levels) taken from all stations is such that the null hypothesis should be rejected [see Wilks, 2006, equation 7b]. Weekly cycles have been suggested to occur over broad spatial ranges by some researchers [e.g., Bäumer and Vogel, 2007; Bell et al., 2009], and also at particular urban locations [e.g., Shutters and Balling, 2006; Simmonds and Keay, 1997].

### 3. Results

#### 3.1. Local Significance

##### 3.1.1. Student $t$ Test

The $t$ test can be an effective approach for determining whether two randomly selected independent means are different. However, as shown in Figure 2, a potential problem with its implementation in analyzing weekly cycle magnitudes is that by choosing the highest and lowest values of the week to compare, the $t$ test results can be strongly biased. Figure 3 shows the percentage of rejected null hypotheses when analyzing weekly cycles from random AR(1) time series with various levels of autocorrelation at a significance level of $p = 0.05$. If the test behaved as desired and were unbiased, the rejection rate would be 5% for all autocorrelation values. However, as shown in the figure this is not the case; the $t$ test identifies over 20% of the random cases as significant when the autocorrelation is less than 0.65 and over 60% as the autocorrelation approaches zero. Thus, using the $t$ test in this way will result in the frequent identification of significant cycles that are, in fact, not significant at all, but simply arise from random fluctuations in the time series.

If the two days compared are chosen randomly, such as always comparing Monday with Thursday, the figure shows that the $t$ test works as expected as long as autocorrelation is low. The test becomes conservative as the autocorrelation increases because the variations within individual weeks decrease for a given variance of the time series. In other words, in random time series the test will incorrectly identify weekly cycle magnitudes as statistically significant less frequently than the test significance level would suggest. The results for the weekend effect test are similar. Although it performs as expected for time series with low autocorrelation, it quickly identifies too few cycles as being significant due to random variations. Rather than identifying 5% of
the random cases as significant, as the lag-1 autocorrelation coefficient increases above 0.5, as it does frequently in meteorological data, the number of significant cases drops below 2%. The conservative bias when evaluating the weekend effect occurs for the same reason as it did for the weekly cycle test of randomly chosen days.

Figure 4 shows the level of permissiveness of the $t$ test when the autocorrelation is 0 for various hypothesis test levels. If the test performs as desired, the fraction of significant random cases would equal the $p$ value of the test, with the results following the 1:1 line. The figure shows that if the $t$ test is used to compare temperatures from two fixed days of the week (e.g., Monday and Thursday), it performs as expected, just as shown in Figure 3 for the 0 autocorrelation case. If the test is performed for the $p = 0.01$ significance level, for example, 1% of the random cases exhibit a significant difference between the two days. When evaluating weekly cycle magnitudes the bias is essentially independent of the time series length for time series longer than about half a year. As shown, the bias is greater than a factor of 10 for hypothesis tests with a $p$ value less than about 0.05. A two-tailed test is used when the two days are chosen randomly because either day could be warmer than the other, and a one-tailed test is used when comparing the warmest and coldest days of the week; this choice of a one-tailed test leads to a larger bias than if a two-tailed test were used.

3.1.2. Analysis of Variance (ANOVA)

Figure 3 shows that the ANOVA technique performs as desired when data are characterized by little temporal autocorrelation, but becomes more conservative as autocorrelation increases. The behavior is qualitatively similar to that of the weekend effect test and the weekly cycle magnitude analysis when comparing randomly chosen days.

A conservative test may often be preferred to a permissive test so that the rejection of the null hypothesis implies that the true significance is actually higher than inferred from the test and not lower. However, if a cycle is not found, care must be taken in interpreting the meaning of not rejecting the null hypothesis. As with the $t$ test, there are implications for comparing to previously published work and for determining what the significance results imply when the actual testing level of significance is not what is expected.

3.1.3. Resampling

The resampling technique can often perform better than other approaches because it can account for much of the temporal autocorrelation in the initial time series. For most accurate testing, however, the block length must be appropriate, with the best choice depending on the temporal autocorrelation of the data. Figure 5 shows the fraction of AR(1) random time series that are determined to have
significant weekly cycles at the \( p = 0.10 \) level when the original time series has prescribed lag-1 autocorrelation coefficients ranging from 0.00 to 0.75. These results are given as a function of block length. Results for four prescribed values of temporal autocorrelation are shown. These values are shown in the top panel and are color-coded the same as the corresponding curves. Vertical dashed lines represent the block length suggestion for the prescribed autocorrelation and time series length from Wilks [1997].

\[
\nu = \frac{1 + \alpha}{1 - \alpha}
\]

where \( \alpha \) is the lag-1 autocorrelation coefficient prescribed in generating the AR(1) time series.

Another important metric for evaluating a hypothesis test is the test power. This describes the ability of the test to identify a real signal and to successfully reject the null hypothesis. We have examined the power of the resampling test described here for an idealized case of an AR(1) time series prescribed to have an unchanging variance and temporal autocorrelation with a sinusoidal weekly oscillation added to it. Figure 6 shows the results of our hypothesis tests as the weekly cycle amplitude varies from 0.001 to 0.15 of the time series standard deviation. The top panel shows the fraction of cases in which the weekly cycle and weekend effect are determined to be significant for time series of 10 y and 100 y in length. The bottom panel shows the effect of changing autocorrelation. Each of these hypothesis tests is performed with the resampling approach discussed in section 2.2.1. As an example, successful rejection of the null hypothesis occurs in the weekly cycle magnitude test about half the time for a 10 y time series when the weekly cycle

**Figure 5.** Fraction of random AR(1) 10 y (520 weeks) time series demonstrating statistically significant weekly cycles and weekend effects using a hypothesis test at the \( p = 0.1 \) significance level for a range of block lengths. Shaded region represents 5–95% confidence interval around the desired result of 0.10, given 20,000 time series used in generating each point. Results for four prescribed values of temporal autocorrelation are shown. These values are shown in the top panel and are color-coded the same as the corresponding curves. Vertical dashed lines represent the block length suggestion for the prescribed autocorrelation and time series length from Wilks [1997].

**Figure 6.** Power of resampling test for detecting significant weekly cycles, weekend effects, and 7-day spectral power. Temperature simulations are generated from AR(1) time series with lag-1 autocorrelation coefficients of 0.0 and 0.5 with an added weekly sinusoidal term. The amplitude of the weekly term is represented on the abscissa in units of the standard deviation of the AR(1) component. Time series are 10 years and 100 years in length. The power is represented by the fraction of cases (out of 10 000 here) in which the weekly cycle is found to be statistically significant at the \( p = 0.10 \) level.

\[
I = (n - l + 1) \frac{(l-1)/\nu}{(l-1/\nu)}
\]

where \( n \) is the number of data points in the time series, \( l \) is the block length, and \( \nu \) is given by

\[
\nu = \frac{1 + \alpha}{1 - \alpha}
\]
amplitude is about 5% of the time series standard deviation and the autocorrelation is 0.5. This implies that if the standard deviation of a random time series is 5 K, a sinusoidal weekly cycle with an amplitude of 0.25 K superimposed on it will be successfully detected about half the time. Use of a longer time series reduces the required amplitude of the sinusoidal oscillation by the square root of the ratio of time series lengths. For example, going from a time series length of 10 y to 100 y allows for a weekly cycle detection that has an amplitude about 3.2 times smaller. The autocorrelation of the time series only slightly affects this result (bottom panel). The figure also shows that the weekend effect magnitude is detected with slightly more power than the weekly cycle magnitude.

In addition to weekly cycle and weekend effect analysis, one could use Fourier analysis to calculate the power spectrum in order to determine whether a 7-day cycle is present. However, as with the previous analyses, the key question is whether the power present at 7 days is statistically inconsistent with random variability. A complete Fourier analysis is fraught with complications and details that go beyond the scope of our goals here. Our intent is to use an extremely simple approach to demonstrate a few of the difficulties that would be associated with such an approach and to gain a general understanding about how such an analysis might compare with the other techniques applied here. In our Fourier approach, we use our simulated spectra with various weekly cycle magnitudes. From many simulations, we determine the threshold level of power significance using the power distribution between 6.75 and 7.25 d, not including 7 d. This range of periods is somewhat arbitrary; we have made this choice because the range is large enough to build up statistical power distribution functions without too many time series, but it is narrow enough that the magnitude distribution appears to be nearly unchanged across the range. From the distribution, we define a threshold at the 90th percentile power level. If the power at 7 days is greater than this threshold, it is determined to be statistically significant at the $p = 0.10$ level.

Three sample spectra and power distribution functions are shown for three weekly cycle magnitudes in Figure 7. As the 7-day power distribution function moves further to the right (right panels), the detection of the weekly cycle becomes more frequent. For example, in the third case (amplitude = 0.10), almost 90% of the 7-day power distribution is above the threshold level. The power of the spectral results is also shown in the top panel of Figure 7 and is comparable to the power of the weekly cycle and weekend effect tests. It is worth noting that when analyzing the difference between the weekend days and weekdays, analysis results are highly dependent on the phase of the existing cycle. For example, if the phase were shifted by 1.4 days compared with the cycles analyzed in Figure 7, the weekend effect would be close to zero. This suggests that the weekly cycle test using a weekend effect analysis if the phase of the weekly cycle is not known.

While resampling can be a powerful method to identify weekly cycles, depending on how the resampling is performed, it can also lead to some misleading results. Laux and Kunstmann [2008] (hereinafter referred to as LK08) used an alternative resampling approach to provide evidence of a weekly cycle in T. They argue that peaks in the magnitude of weekly cycles when using block lengths that are
perfectly divisible by 7 days implies the presence of a weekly cycle. While a weekly cycle would lead to this type of behavior, this behavior does not necessarily imply the presence of a weekly cycle. The approach of dividing the time series into sequential, non-overlapping blocks, and then resampling these blocks with replacement qualitatively leads to peaks at blocks lengths that are integral multiples of seven even for random time series with no prescribed weekly cycle. Figure 8 is similar to Figure 3 of LK08, showing the results from a random AR(1) time series. The peaks at multiples of seven days occur because resampled time series derived from block lengths of 7, 14, 21, etc. days sometimes have multiple 7-day weeks in the time series that are identical. The duplication of the weekly cycle in those weeks that arises from random processes leads to less cancellation of the random weekly cycle when averaging over all weeks considered. This then leads to a larger weekly cycle in these time series than for the other block lengths, where duplicate blocks only fall on the same seven-day cycle on average 1/7th as frequently (for lengths larger than 7 days). Quantitatively, the magnitude of the differences between the weekly cycles from integral block lengths of seven days with other block lengths is not as large in Figure 8 as it was in LK08. More examination is required to determine whether there are some fundamental differences between our simulated random time series and their observed time series or whether there is truly a weekly cycle in the LK08 time series that makes the difference larger than expected from a random time series.

3.2. Field Significance

As stated in section 2.2.2, the null hypothesis for testing field significance is that all the stations are characterized by weekly cycles that are consistent with random variability. This implies that if a field is determined to be significant due to the rejection of the null hypothesis, at least one station is characterized by a weekly cycle that is unlikely to arise from random variability in the time series.

The counting test uses the number of stations exhibiting significant weekly cycles in the field to determine field significance. If there is no spatial correlation among analyzed stations, it is a simple matter to determine the number of individual stations that are expected to be significant at a given level by using the binomial distribution. In Figure 9, the binomial distribution is used to generate the cumulative distribution function (cdf) of the number of significant stations for the $p = 0.2$ significance level where there are 49 independent stations. The cdf shows that if at least 14 stations are significant at the $p = 0.2$ level, the uncorrelated field would be significant at the $p = 0.05$ level. More (less) stringent levels of station significance evaluations will simply decrease (increase) the number of individual stations that need to be significant for the field to be significant at some level.

If there is spatial correlation, it becomes more likely that very few or very many stations will be significant, leading to a broader distribution. An example of such a distribution is also shown in Figure 9 for a field that has some level of spatial correlation; this field was calculated using the technique described in Khalili et al. [2007]. In this particular case, adjacent stations are characterized by a correlation of 0.8. As stated in section 2.2.2, this spatial correlation has been accounted for in the analysis of field significance by resampling the time series at every station in the field using exactly the same indices. In other words, if the first value in the resampled time series for one station is taken from, for example, the 43rd value of the original time series of that station, the resampled time series for every station has a first value that is equal to the 43rd value in that station’s original time series. If resamplings for each station were generated with different indices, the spatial correlation would be lost and results would be similar to those of using the binomial distribution. For this figure, we generate 10,000 resampled time series; from these we are able to generate the correlated distribution shown in Figure 9. If the spatial correlation is neglected, it is clear by comparing the curves for spatially correlated and uncorrelated data that significant
The power of the FDR and counting tests are shown in Figure 10. The ordinate represents the amplitude of the weekly cycle in units of standard deviation of the underlying random time series. The abscissa represents the number of stations that have this added weekly cycle. For example, if 1 station has a weekly cycle with an amplitude that is 10% of the underlying standard deviation, the FDR approach rejects the null hypothesis at the $\alpha = 0.2$ level about 50% of the time while the counting approach rejects it less than 30% of the time. We choose the $\alpha = 0.2$ level not because we recommend such a liberal significance level, but in order to improve the statistics of our comparisons. The lack of power of the counting test when few stations have a real weekly cycle is understood and has been discussed in Wilks [2006]. Even a very large weekly cycle at only a few stations will help bring the total number of significant stations closer to the threshold number of significant stations needed to reject the null hypothesis under the counting approach, but it will not guarantee that the entire field will pass this threshold.

On the other hand, if 25 stations have prescribed weekly cycles with amplitudes of 3% of the noise level, the FDR rejects the null hypothesis 40% of the time while the counting approach rejects it over 60% of the time. The implication of these results is that the FDR approach can be more powerful at identifying weekly cycles if they appear in only a few stations in the domain, while the counting approach is somewhat more powerful if many stations exhibit a relatively small weekly cycle. The determination of which test is preferable clearly depends on the spatial nature of the weekly cycle. If it is expected that the weekly cycle is widespread (local), it would be preferable to use the counting (FDR approach). If there is no a priori expectation, a combination of both approaches could allow for the detection of small regional weekly cycles that are present in many stations as well as larger weekly cycles that might be present at a very few sites. However, in cases in which both tests are applied and one test identifies statistical significance and the other does not, one would have to determine the true significance level given the dual hypothesis test that was performed.

3.3. Other Considerations

As discussed in sections 3.1 and 3.2 the use of an appropriate statistical technique is necessary to arrive at a justifiable conclusion regarding the potential impact of anthropogenic behavior on meteorological variables when using weekly cycle analyses. However, even with a proper statistical technique, there are several other errors that can invalidate the entire weekly cycle conclusion.

3.3.1. Elaborate Hypothesis

Many studies arguing that observed cycles are significant do not find a consistent day of the week that is the warmest or wettest, even in polluted areas. Differing mechanisms are suggested for these weekly cycles, many of which seem reasonable. But frequently, the explanations are developed after the weekly cycle nature is identified, albeit perhaps using an inappropriate statistical method, to of course be consistent with those observations. The potential problem with this approach is that if a hypothesis has enough degrees of freedom, it can be adjusted to match a wide range of observations even though the suggested process actually has no causal relationship with the observed variability.
weekly cycle. A good discussion of the problems associated with the elaborate, or post hoc, hypothesis is given in Pitlock [1978, section 3e]. Equipped with an elaborate hypothesis and with a multiplicity of analyses (see section 3.3.3) one can be quite successful at arriving at a physically based explanation for an event that is actually caused by random variability. This is certainly a danger to be aware of in weekly cycle analysis. One approach to surmounting this problem is to analyze additional data that were not used in forming the hypothesis, while another is to conduct detailed modeling calculations that accurately account for the suggested physical processes.

3.3.2. Correlated Meteorological Data

It is generally not sufficient to use the correlation among weekly cycles of various meteorological variables as additional evidence for an anthropogenic effect. Meteorological variables are linked by the physics of the atmosphere and are therefore generally correlated, so if any one exhibits a weekly cycle due to random processes, it is likely that several will exhibit such cycles. Although consistent results for multiple variables do provide evidence that the effect is not an artifact of the data collection process, correlated results should not be considered to be independent pieces of information in determining the likelihood of a weekly cycle being anthropogenic in nature. For example, since cloud cover is frequently correlated with T, if there is a weekly cycle in one that is due to random variability, it is likely that it will occur in the other as well. Such a result thus does not provide significantly more support that the observed weekly cycle is a result of anthropogenic behavior than does having a single variable demonstrating a significant weekly cycle. This must be considered in determining whether the observations are consistent with random variability.

3.3.3. Multiplicity

Another point that should be considered relates to the multiplicity of tests. It is common to perform weekly cycle analyses over subintervals and/or subregions of the available time series. These include seasonal and decadal analyses, for example. There are frequently good physical reasons that would motivate such an approach, such as inhomogeneous distributions of various types of aerosols, different interactions between aerosols and clouds expected during different seasons, temporal trends in aerosol amount, etc. The multiplicity complication arises in determining how to interpret one or a few significant weekly cycles in certain regions and over certain time periods. For example, if 5 decades of data are analyzed for each of the 4 seasons at the 10% significance level, there are 20 nearly independent data sets if the null hypothesis is true, and on average it is expected that 2 decadal averaged seasons will demonstrate significance at this level. But there is still a 19% and 9% chance that 3 and 4 seasons will be significant, respectively. There is a 14% chance of having the same season demonstrate significance in consecutive decades. It could then be tempting to suggest a physical process that could explain that phenomenon when, in reality, there is no effect to explain at all. If all other significance testing is done accurately and multiplicity is not considered appropriately, the conclusion can still be erroneous. Just as it is more likely to find a single station that exhibits a particularly large weekly cycle when a larger number of stations is analyzed, it is more likely that the null hypothesis will be rejected due to random variability in one or several time periods when tests are performed over a larger number of periods.

3.3.4. Post Hoc Domain Selection

Another potential pitfall is to analyze all the data for weekly cycles and then, after that analysis, to divide the field significance analysis into geographic regions. An example of this would be to analyze weekly cycles in T for the entire United States, and then perform a field significance test on only a smaller region that contains a large number of stations that demonstrate significant weekly cycles. As with analyzing multiple time periods, an examination at the regional level could be well justified. It might be expected, for example, that a region having higher levels of aerosols might demonstrate large weekly cycles. However, to avoid a biased result, the regions should be determined before performing the local analysis. It is also preferred that any domain limitation be made as a result of some physical insight [Bell and Rosenfeld, 2008]. If some region in a domain is seen to contain a large number of significant stations, it is certainly inappropriate to apply the field significance test over just this limited domain without a physically based justification. Because of the spatial correlation in meteorological variables, it might be quite likely that some region will exist in the overall domain that has significant weekly cycles at many stations. If a region is identified to have many weekly cycles and a hypothesized physical cause is proposed, as in the case of an “elaborate hypothesis,” it would be necessary to assess the field significance of the identified sub-domain with data independent of that used to develop the hypothesis.

Considering multiplicity and a post hoc domain selection, one could imagine that almost any random meteorological field could be analyzed in a way so that some subset of that field will show significant weekly cycles if analyzed alone. If the entire analysis procedure were not described, and only the significant region were shown, it could be impossible for a reader to recognize that there was anything wrong with the analysis. This is but one way that the reader must rely completely on the researcher to be forthright in his/her explanation of the analysis procedure and to possess a sufficient understanding of statistics.

4. Conclusions

We have shown that the presence of weekly cycles or weekend effects in meteorological data does not necessarily imply an anthropogenic impact. Natural variability can lead to the appearance of these cycles. Thus, the goal of any weekly cycle analysis must be to prove that it is unlikely to some level of confidence that the observed cycles are simply the result of natural variability. Our purpose has not been to suggest that significant weekly cycles in meteorological variables are absent. We have only pointed out that in many instances the evidence for these cycles should be revisited with more appropriate statistical approaches. We have focused on temperature although the results generally apply to precipitation and other meteorological fields.

A key aspect of weekly cycle analysis is the statistical test that is used to determine the likelihood that an observed cycle is not caused by natural variability. Many different tests have been used in the literature, with some likely performing as intended, some being permissive, and some being conservative. The Student t test is an example of
highly permissive test when evaluating weekly cycle magnitudes of time series with lag-1 autocorrelation values less than about 0.8. In fact, the \( t \) test will identify weekly cycles as being significant more than 60% of the time for completely random time series with no autocorrelation. Such a result would appear to suggest an anthropogenic link, when in fact, the weekly cycle was caused by natural variability. It is critical to understand whether a statistical test rejects the null hypothesis too frequently or requires too much evidence to reject it both to understand the actual implications of the results and to put the results into the proper context when comparing with other studies. The wide variety of tests used and some misapplications of these tests are partly responsible for the confusion in the literature regarding the presence of weekly cycles. By analyzing synthetic data, we have evaluated a resampling test that performs well both for station and field significance tests as long as the block length is appropriately chosen. One major advantage of the resampling test is that it can account for both temporal autocorrelation and spatial correlation.

Even if the statistical test is appropriate and it is used correctly, there are numerous ways to invalidate the ultimate weekly cycle conclusion regarding potential anthropogenic impacts. For example, care must be taken in appropriately accounting for correlation among meteorological parameters if weekly cycles in multiple parameters are used to test for an anthropogenic impact. It can also be problematic to select, post hoc, particular areas in an analyzed domain on which to focus field significance analyses, because even randomly timed weather patterns combined with spatial correlation can cause some areas to have a disproportionate number of stations characterized by significant weekly cycles. This must not be confused for a regional anthropogenic impact. Finally, when multiple time periods or regions are analyzed, the presence of statistically significant results in some periods or regions must be interpreted with the understanding that some fraction of the analyzed domain is expected to exhibit statistically significant weekly cycles due to random chance. The significance of a weekly cycle exhibited by a particular region over a particular time must be interpreted not in isolation but in context of the entire observational domain and of all the time periods analyzed.

Acknowledgments. We thank Benjamin Santer for helpful discussions. We also thank Thomas Bell, Graham Weedon, and one anonymous reviewer for their helpful comments and suggestions in the review process.

References
Ashworth, J. (1933), Rainfall and atmospheric pollution, *Nature*, 132, 443, doi:10.1038/132443a0.
Ashworth, J. (1944), Smoke and rain, *Nature*, 154(3902), 213–214, doi:10.1038/154213b0.