APPENDIX A

Statistical Issues Regarding Trends

Author: Tom M.L. Wigley, NSF NCAR
With contributions by: Benjamin D. Santer, DOE LLNL; John R. Lanzante, NOAA

Abstract

The purpose of this Appendix is to explain the statistical terms and methods used in this Report. We begin by introducing a number of terms: mean, standard deviation, variance, linear trend, sample, population, signal, and noise. Examples are given of linear trends in surface, tropospheric, and stratospheric temperatures. The least squares method for calculating a best- fit linear trend is described. The method for quantifying the statistical uncertainty in a linear trend is explained, introducing the concepts of standard error, confidence intervals, and significance testing. A method to account for the effects of temporal autocorrelation on confidence intervals and significance tests is described. The issue of comparing two data sets to decide whether differences in their trends could have occurred by chance is discussed. The analysis of trends in state-of-the-art climate model results is a special case because we frequently have an ensemble of simulations for a particular forcing case. The effect of ensemble averaging on confidence intervals is illustrated. Finally, the issue of practical versus statistical significance is discussed. In practice, it is important to consider construction uncertainties as well as statistical uncertainties. An example is given showing that these two sources of trend uncertainty can be of comparable magnitude.

(1) WHY DO WE NEED STATISTICS?

Statistical methods are required to ensure that data are interpreted correctly and that apparent relationships are meaningful (or "significant") and not simply chance occurrences.

A "statistic" is a numerical value that describes some property of a data set. The most commonly used statistics are the average (or "mean") value, and the "standard deviation," which is a measure of the variability within a data set around the mean value. The "variance" is the square of the standard deviation. The linear trend is another example of a data "statistic."

Two important concepts in statistics are the "population" and the "sample." The population is a theoretical concept, an idealized representation of the set of all possible values of some measured quantity. An example would be if we were able to measure temperatures continuously at a single site for all time – the set of all values (which would be infinite in size in this case) would be the population of temperatures for that site. A sample is what we actually see and can measure: *i.e.,* what we have available for statistical analysis, and a necessarily limited subset of the population. In the real world, all we ever have is limited samples, from which we try to estimate the properties of the population.

As an analogy, the population might be an infinite jar of marbles, a certain proportion of which (say 60%) is blue and the rest (40%) are red. We can only draw off a finite number of these marbles (a sample) at a time; and, when we measure the numbers of blue and red marbles in the sample, they need not be in the precise ratio 60:40. The ratio we measure is called a "sample statistic." It is an estimate of some hypothetical underlying population value (the corresponding "population parameter"). The techniques of statistical science allow us to make optimum use of the sample statistic and obtain a best estimate of the population parameter. Statistical science also allows us to quantify the uncertainty in this estimate.

(2) DEFINITION OF A LINEAR TREND

If data show underlying smooth changes with time, we refer to these changes as a trend. The simplest type of change is a linear (or straight line) trend, a continuous increase or decrease over time. For example, the net effect of increasing greenhouse-gas concentrations and other human-induced factors is expected to cause warming at the surface and in the troposphere and cooling in the stratosphere (see Figure 1). Warming corresponds to a positive (or increasing) linear trend, while cooling corresponds to a negative (or decreasing)

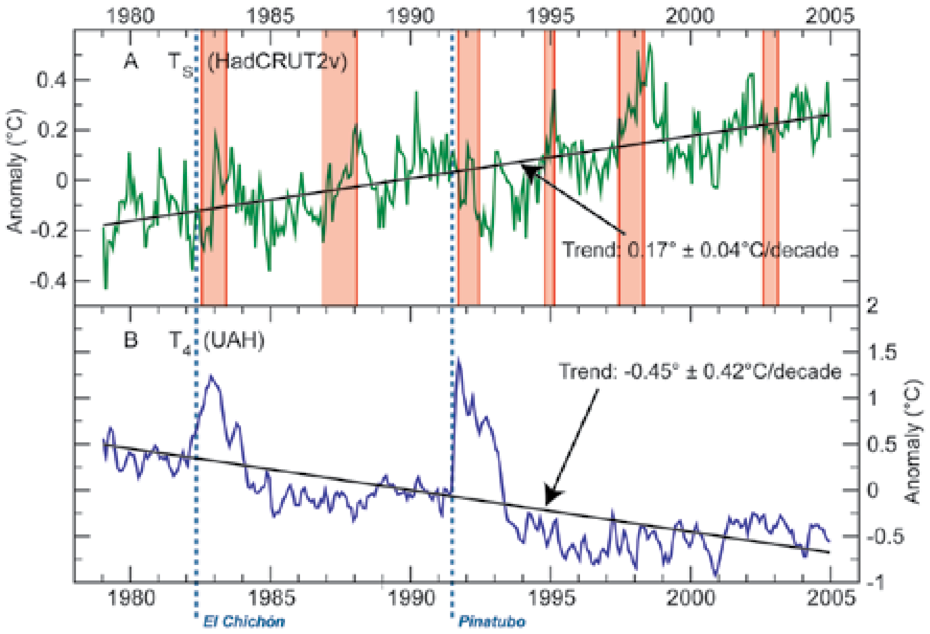

Figure 1: Examples of temperature time series with best-fit (least squares) linear trends: A, global-mean surface temperature from the UKMO Hadley Centre/Climatic Research Unit data set (HadCRUT2v); and B, global-mean MSU channel 4 data (T_4) for the lower stratosphere from the University of Alabama at Huntsville (UAH). Note the much larger temperature scale on the lower panel. Temperature changes are expressed as anomalies relative to the 1979 to 1999 mean (252 months). Dates for the eruptions of El Chichón and Mt. Pinatubo are shown by vertical lines. El Niños are shown by the shaded areas. The trend values are as given in Chapter 3, Table 3.3. The ± values define the 95% confidence intervals for the trends, also from Chapter 3, Table 3.3. The smaller confidence interval for the surface data shows that the straight line fit in this case is better than the straight line fit to the stratospheric data.

trend. Over the present study period (1958 onwards), the expected changes due to "anthropogenic" (human-induced) effects are expected to be approximately linear. In some cases, natural factors have caused substantial deviations from linearity (see, *e.g.,* the lower stratospheric changes in Figure 1B), but the linear trend still provides a simple way of characterizing the overall change and of quantifying its magnitude.

Alternatively, there may be some physical process that causes a rapid switch or change from one mode of behavior to another. In such a case the overall behavior might best be described as a linear trend to the change-point, a step change at this point, followed by a second linear trend portion. Tropospheric temperatures from radiosondes show this type of behavior, with an apparent step increase in temperature occurring around 1976 (see Chapter 3, Figure 3.2a, or Figure 1 in the Executive Summary).

Step changes can lead to apparently contradictory results. For example, a data set that shows an initial cooling trend, followed by a large upward step, followed by a renewed cooling trend could have an overall warming trend. To state simply that the data showed overall warming would misrepresent the true underlying behavior.

A linear trend may therefore be deceptive if the trend number is given in isolation, removed from the original data. Nevertheless, used appropriately, linear trends provide the simplest and most convenient way to describe the overall change over time in a data set, and are widely used.

Linear temperature trends are usually quantified as the temperature change per year or per decade (even when the data are available on a month by month basis). For example, the trend for the surface temperature data shown in Figure 1 is 0.169ºC per decade. (Note that 3 decimals are given here purely for mathematical convenience. The accuracy of these trends is much less, as is shown by the confidence intervals given in the Figure and in the Tables in Chapter 3. Precision should not be confused with accuracy.) Giving trends per decade is a more convenient representation than the trend per month, which, in this case, would be 0.169/120 = 0.00141ºC per month, a very small number. An alternative method is to use the "total change" over the full data period – *i.e.,* the total change for the fitted linear trend line from the start to the end of the record (see Figure 2 in the Executive Summary). In Figure 1 here, the data shown span January 1979 through December 2004 (312 months or 2.6 decades). The total change is therefore 0.169x2.6 = 0.439ºC.

(3) EXPECTED TEMPERATURE CHANGES: SIGNAL AND NOISE

Different physical processes generally cause different spatial and temporal patterns of change. For example, anthropogenic emissions of halocarbons at the surface have led to a reduction in stratospheric ozone and a contribution to stratospheric cooling over the past three or four decades. Now that these chemicals are controlled under the Montreal Protocol, the concentrations of the controlled species are decreasing and there is a trend towards a recovery of the ozone layer. The eventual long-term effect on stratospheric temperatures is expected to be non-linear: a cooling up until the late 1990s followed by a warming as the ozone layer recovers.

This is not the only process affecting stratospheric temperatures. Increasing concentrations of greenhouse gases lead to stratospheric cooling; and explosive volcanic eruptions cause sharp, but relatively short-lived stratospheric warmings (see Figure 1)[1]. There are also natural variations, most notably those associated with the Quasi-Bienniel Oscillation (QBO)[2]. Stratospheric temperature changes (indeed, changes at all levels of the atmosphere) are therefore the combined results of a number of different processes acting across all space and time scales.

In climate science, a primary goal is to identify changes associated with specific physical processes (causal factors) or combinations of processes. Such changes are referred to as "signals." Identification of signals in the climate record is referred to as the "detection and attribution" (D&A) problem. "Detection" is the identification of an unusual change, through the use of statistical techniques like significance testing (see below). "Attribution" is the association of a specific cause or causes with the detected changes in a statistically rigorous way.

The reason why D&A is a difficult and challenging statistical problem is because climate signals do not occur in isolation. In addition to these signals, temperature fluctuations in all parts of the atmosphere occur even in the absence of external driving forces. These internally generated fluctuations represent the "noise" against which we seek to identify specific externally forced signals. All climate records, therefore, are "noisy," with the noise of this natural variability tending to obscure the externally driven changes. Figure 1 illustrates this. At the surface, a primary noise component is the variability associated with ENSO (the El Niño/Southern Oscillation phenomenon), while, in the stratosphere, if our concern is to identify anthropogenic influences, the warmings after the eruptions of El Chichón and Mt. Pinatubo constitute noise.

If the underlying response to external forcing is small relative to the noise, then, by chance, we may see a trend in the data due to random fluctuations purely as a result of the noise. The science of statistics provides methods through which we can decide whether the trend we observe is "real" *(i.e.,* a signal associated with some causal factor) or simply a random fluctuation *(i.e.,* noise).

(4) DERIVING TREND STATISTICS

There are a number of different ways to quantify linear trends. Before doing anything, however, we should always inspect the data visually to see whether a linear trend model is appropriate. For example, in Figure 1, the linear warming trend appears to be a reasonable description for the surface data (top panel), but it is clear that a linear cooling model for the lower stratosphere (lower panel) fails to capture some of the more complex changes that are evident in these data. Nevertheless, the cooling trend line does give a good idea of the magnitude of the overall change.

There are different ways to fit a straight line to the data. Most frequently, a "best-fit" straight line is defined by finding the particular line that minimizes the sum, over all data points, of the squares of deviations about the line (these deviations are generally referred to as "residuals" or "errors"). This is an example of a more general procedure called least squares regression.

In linear regression analysis, a predictand (Y) is expressed as a linear combination of one or more predictors (X_i):

$$Y_{est}=b_0 + b_1X_1 + b_2X_2+ \dots \qquad \dots\dots (1)$$

Where the subscript "est" is used to indicate that this is the estimate of Y that is given by the fitted relationship. Differences between the actual and estimated values of Y, the residuals, are defined by

$$e = Y - Y_{est} \qquad \dots\dots (2)$$

For linear trend analysis of temperature data (T) there is a single predictor, time (t; t = 1,2,3, …). The time points are almost always evenly spaced, month-by-month, year-by-year, etc. – but this is not a necessary restriction. In the linear trend case, the regression equation becomes:

[1] Figure 1 shows a number of interesting features. In the stratosphere, the warmings following the eruptions of El Chichón (April 1982) and Mt Pinatubo (June 1991) are pronounced. For El Chichón, the warming appears to start before the eruption, but this is just a chance natural fluctuation. The overall cooling trend is what is expected to occur due to anthropogenic influences. At the surface, on short time scales, there is a complex combination of effects. There is no clear cooling after El Chichón, primarily because this was offset by the very strong 1982/83 El Niño. Cooling after Pinatubo is more apparent, but this was also partly offset by the El Niño around 1992/93 (which was much weaker than that of 1982/83). El Niño events, characterized by warm temperatures in the tropical Pacific, have a noticeable effect on global-mean temperature, but the effect lags behind the Pacific warming by 3-7 months. This is very clear in the surface temperature changes at and immediately after the 1986/87 and 1997/98 El Niños, also very large events. The most recent El Niños were weak and have no clear signature in the surface temperatures.

[2] The QBO is a quasi-periodic reversal in winds in the tropical stratosphere that leads to alternating warm and cold tropical stratospheric temperatures with a periodicity of 18 to 30 months.

$$T_{est} = a + b\,t \qquad \text{….. (3)}$$

In equation (3), "b" is the slope of the fitted line – *i.e.,* the linear trend value. This is a sample statistic, *i.e.,* it is an estimate of the corresponding underlying population parameter. To distinguish the population parameter from the sample value, the population trend value is denoted ß.

The formula for b is:

$$b = [\Sigma((t - \bar{t})T_t)]/[\Sigma((t - \bar{t})^2)] \qquad \text{….. (4)}$$

Where $\bar{t}$ denotes the mean value, and the summation is over t = 1,2,3, … n *(i.e.,* the sample size is n). T_t denotes the value of temperature, T, at time "t". Equation (4) produces an unbiased estimate[3] of population trend, ß.

For the usual case of evenly spaced time points, $\bar{t} = (n+1)/2$, and

$$\Sigma((t - \bar{t})^2) = n(n^2-1)/12 \qquad \text{….. (5)}$$

When we are examining deviations from the fitted line the sign of the deviation is not important. This is why we consider the squares of the residuals in least squares regression. An important and desirable characteristic of the least squares method is that the average of the residuals is zero.

Estimates of the linear trend are sensitive to points at the start or end of the data set. For example, if the last point, by chance, happened to be unusually high, then the fitted trend might place undue weight on this single value and lead to an estimate of the trend that was too high. This is more of a problem with small sample sizes *(i.e.,* for trends over short time periods). For example, if we considered tropospheric data over 1979 through 1998, because of the unusual warmth in 1998 (associated with the strong 1997/98 El Niño; see Figure 1), the calculated trend may be an overestimate of the true underlying trend.

There are alternative ways to estimate the linear trend that are less sensitive to endpoints. Although we recognize this problem, for the data used in this Report tests using different trend estimators give results that are virtually the same as those based on the standard least-squares trend estimator.

(5) TREND UNCERTAINTIES

Some examples of fitted linear trend lines are shown in Figure 1. This Figure shows monthly temperature data for the surface and for the lower stratosphere (MSU channel 4) over 1979 through 2004 (312 months). In both cases there is a clear trend, but the fit is better for the surface data. The trend values *(i.e.,* the slopes of the best fit straight lines that are shown superimposed on monthly data) are +0.17ºC/decade for the surface and –0.45ºC/decade for the stratosphere. For the stratosphere, although there is a pronounced overall cooling trend, as noted above, describing the change simply as a linear cooling considerably oversimplifies the behavior of the data[1].

A measure of how well the straight line fits the data *(i.e.,* the "goodness of fit") is the average value of the squares of the residuals. The smaller this is, the better is the fit. The simplest way to define this average would be to divide the sum of the squares of the residuals by the sample size *(i.e.,* the number of data points, n). In fact, it is usually considered more correct to divide by n – 2 rather than n, because some information is lost as a result of the fitting process and this loss of information must be accounted for. Dividing by n – 2 is required in order to produce an unbiased estimator[3].

The population parameter we are trying to estimate here is the standard deviation of the trend estimate, or its square, the variance of the distribution of b, which we denote Var(b). The larger the value of Var(b), the more uncertain is b as an estimate of the population value, ß.

The formula for Var(b) is …

$$\mathrm{Var}(b) = [\sigma^2]/[\Sigma((t - \bar{t})^2) \qquad \text{….. (6)}$$

where σ^2 is the population value for the variance of the residuals. Unfortunately, we do not in general know what σ^2 is, so we must use an unbiased sample estimate of σ^2. This estimate is known as the Mean Square Error (MSE), defined by …

$$\mathrm{MSE} = [\Sigma(e^2)]/(n - 2) \qquad \text{….. (7)}$$

Hence, equation (6) becomes

$$\mathrm{Var}(b) = (\mathrm{SE})^2 = \mathrm{MSE}/[\Sigma((t - \bar{t})^2)] \qquad \text{….. (8)}$$

where SE, the square root of Var(b), is called the "standard error" of the trend estimate. The smaller the value of the standard error, the better the fit of the data to the linear change description and the smaller the uncertainty in the sample trend as an estimate of the underlying population

[3] An unbiased estimator is one where, if the same experiment were to be performed over and over again under identical conditions, then the long-run average of the estimator will be equal to the parameter that we are trying to estimate. In contrast, in a biased estimator, there will always be some slight difference between the long-run average and the true parameter value that does not tend to zero no matter how many times the experiment is repeated. Since our goal is to estimate population parameters, it is clear that unbiased estimators are preferred.

trend value. The standard error is the primary measure of trend uncertainty. The standard error will be large if the MSE is large, and the MSE will be large if the data points show large scatter about the fitted line.

There are assumptions made in going from equation (6) to (8): viz. that the residuals have mean zero and common variance, that they are Normally (or "Gaussian") distributed[4], and that they are uncorrelated or statistically independent. In climatological applications, the first two assumptions are generally valid. The third assumption, however, is often not justified. We return to this below.

(6) CONFIDENCE INTERVALS AND SIGNIFICANCE TESTING

In statistics we try to decide whether a trend is an indication of some underlying cause, or merely a chance fluctuation. Even purely random data may show periods of noticeable upward or downward trends, so how do we identify these cases?

There are two common approaches to this problem, through significance testing and by defining confidence intervals. The basis of both methods is the determination of the "sampling distribution" of the trend, *i.e.,* the distribution of trend estimates that would occur if we analyzed data that were randomly scattered about a given straight line with slope ß. This distribution is approximately Gaussian with a mean value equal to ß and a variance (standard deviation squared) given by equation (8). More correctly, the distribution to use is Student's "t" distribution, named after the pseudonym "Student" used by the statistician William Gosset. For large samples, however (n more than about 30), the distribution is very nearly Gaussian.

Confidence Intervals

The larger the standard error of the trend, the more uncertain is the slope of the fitted line. We express this uncertainty probabilistically by defining confidence intervals for the trend associated with different probabilities. If the distribution of trend values were strictly Gaussian, then the range b – SE to b + SE would represent the 68% confidence interval (C.I.) because the probability of a value lying in that range for a Gaussian distribution is 0.68. The range b – 1.645(SE) to b + 1.645(SE) would give the 90% C.I.; the range b – 1.96(SE) to b + 1.96(SE) would give the 95% C.I.; and so on. Quite often, for simplicity, we use b – 2(SE) to b + 2(SE) to represent (to a good approximation) the 95% confidence interval. (This is often called the "two-sigma" confidence interval.) Examples of 95% confidence intervals are given in Figure 1. Here, the smaller value for the surface data compared with the stratospheric data shows that a straight line fits the surface data better than it does the stratospheric data.

Because of the way C.I.s are usually represented graphically, as a bar centered on the best-fit estimate, they are often referred to as "error bars." Confidence intervals may be expressed in two ways, either (as above) as a range, or as a signed error magnitude. The approximate 95% confidence interval, therefore, may be expressed as b ± 2(SE), with appropriate numerical values inserted for b and SE.

As will be explained further below, showing confidence interval for linear temperature trends may be deceptive, because the purely statistical uncertainties that they represent are not the only sources of uncertainty. Such confidence intervals quantify only one aspect of trend uncertainty, that arising from statistical noise in the data set. There are many other sources of uncertainty within any given temperature data set and these may be as or more important than statistical uncertainty. Showing just the statistical uncertainty may therefore provide a false sense of accuracy in the calculated trend.

Significance Testing

An alternative method for assessing trends is hypothesis testing. In practice, it is much easier to disprove rather than prove a hypothesis. Thus, the standard statistical procedure in significance testing is to set up a hypothesis that we would like to disprove; we call this a "null hypothesis." In the linear trend case, we are often interested in trying to decide whether an observed data trend that is noticeably different from zero is sufficiently different that it could not have occurred by chance – or, at least, that the probability that it could have occurred by chance is very small. The appropriate null hypothesis in this case would be that there was no underlying trend (ß = 0). If we disprove *(i.e.,* "reject") the null hypothesis, then we say that the observed trend is "statistically significant" at some level of confidence and we must accept some alternate hypothesis. The usual alternate hypothesis in temperature analyses is that the data show a real, externally forced warming (or cooling) trend. (In cases like this, the statistical analysis is predicated on the assumption that the observed data are reliable, which is not always the case. If a trend were found to be statistically significant, then an alternative possibility might be that the observed data were flawed.)

[4] The "Gaussian" distribution (often called the "Normal" distribution) is the most well-known probability distribution. This has a characteristic symmetrical "bell" shape, and has the property that values near the center (or mean value) of the distribution are much more likely than values far from the center.

An alternative null hypothesis that often arises is when we are comparing an observed trend with some model expectation. Here, the null hypothesis is that the observed trend is equal to the model value. If our results led us to reject this null hypothesis, then (assuming again that the observed data are reliable) we would have to infer that the model result was flawed – either because the external forcing applied to the model was incorrect and/or because of deficiencies in the model itself.

An important factor in significance testing is whether we are concerned about deviations from some hypothesized value in any direction or only in one direction. This leads to two types of significance test, referred to as "one-tailed" (or "one-sided") and "two-tailed" tests. A one-tailed test arises when we expect a trend in a specific direction (such as warming in the troposphere due to increasing greenhouse-gas concentrations). Two-tailed tests arise when we are concerned only with whether the trend is different from zero, with no specification of whether the trend should be positive or negative. In temperature trend analyses we generally know the sign of the expected trend, so one-tailed tests are more common.

The approach we use in significance testing is to determine the probability that the observed trend could have occurred by chance. As with the calculation of confidence intervals, this involves calculating the uncertainty in the fitted trend arising from the scatter of points about the trend line, determined by the standard error of the trend estimate (equation [8]). It is the ratio of the trend to the standard error (b/SE) that determines the probability that a null hypothesis is true or false. A large ratio (greater than 2, for example) would mean that (except for very small samples) the 95% C.I. did not include the zero trend value. In this case, the null hypothesis is unlikely to be true, because the zero trend value, the value assumed under the null hypothesis, lies outside the range of trend values that are likely to have occurred purely by chance.

If the probability that the null hypothesis is true is small, and less than a predetermined threshold level such as 0.05 (5%) or 0.01 (1%), then the null hypothesis is unlikely to be correct. Such a low probability would mean that the observed trend could only have occurred by chance one time in 20 (or one time in 100), a highly unusual and therefore "significant" result. In technical terms we would say that "the null hypothesis is rejected at the prescribed significance level", and declare the result "significant at the 5% (or 1%) level." We would then accept the alternate hypothesis that there was a real deterministic trend and, hence, some underlying causal factor.

Even with rigorous statistical testing, there is always a small probability that we might be wrong in rejecting a null hypothesis. The reverse is also true – we might accept a null hypothesis of no trend even when there is a real trend in the data. This is more likely to happen when the sample size is small. If the real trend is small and the magnitude of variability about the trend is large, it may require a very large sample in order to identify the trend above the background noise.

For the null hypothesis of zero trend, the distribution of trend values has mean zero and standard deviation equal to the standard error. Knowing this, we can calculate the probability that the actual trend value could have exceeded the observed value by chance if the null hypotheses were true (or, if we were using a two-tailed test, the probability that the magnitude of the actual trend value exceeded the magnitude of the observed value). This probability is called the "p-value." For example, a p-value of 0.03 would be judged significant at the 5% level (since 0.03<0.05), but not at the 1% level (since 0.03>0.01).

Since both the calculation of confidence intervals and significance testing employ information about the distribution of trend values, there is a clear link between confidence intervals and significance testing.

A Complication: The Effect of Autocorrelation

The significance of a trend, and its confidence intervals, depend on the standard error of the trend estimate. The formula given above for this standard error (equation [8]) is, however, only correct if the individual data points are unrelated, or statistically independent. This is not the case for most temperature data, where a value at a particular time usually depends on values at previous times; *i.e.,* if it is warm today, then, on average, it is more likely to be warm tomorrow than cold. This dependence is referred to as "temporal autocorrelation" or "serial correlation." When data are auto-correlated *(i.e.,* when successive values are not independent of each other), many statistics behave as if the sample size was less than the number of data points, n.

One way to deal with this is to determine an "effective sample size," which is less than n, and use it instead of n in statistical formulae and calculations. The extent of this reduction from n to an effective sample size depends on how strong the autocorrelation is. Strong autocorrelation means that individual values in the sample are far from being independent, so the effective number of independent values must be much smaller than the sample size. Strong autocorrelation is common in temperature time series. This is accounted for by reducing the divisor "n – 2" in the mean square error term (equation [7]) that is crucial in determining the standard error of the trend (equation [8]).

There are a number of ways that this autocorrelation effect may be quantified. A common and relatively simple method is described in Santer *et al.* (2000). This method makes the assumption that the autocorrelation structure of the temperature data may be adequately described by a "first-order autoregressive" process, an assumption that is a good approximation for most climate data. The lag-1 autocorrelation coefficient $_{(r1)}$ is calculated from the observed data[5], and the effective sample size is determined by

$$n_{eff} = n\ (1- r_1)/(1 + r_1) \qquad \text{….. (9)}$$

There are more sophisticated methods than this, but testing on observed data shows that this method gives results that are very similar to those obtained by more sophisticated methods.

If the effective sample size is noticeably smaller than n, then, from equations (7) and (8) it can be seen that the standard error of the trend estimate may be much larger than one would otherwise expect. Since the width of any confidence interval depends directly on this standard error (larger SE leading to wider confidence intervals), then the effect of autocorrelation is to produce wider confidence intervals and greater uncertainty in the trend estimate. A corollary of this is that results that may show a significant trend if autocorrelation is ignored are frequently found to be non-significant when autocorrelation is accounted for.

(7) COMPARING TRENDS IN TWO DATA SETS

Assessing the magnitude and confidence interval for the linear trend in a given data set is standard procedure in climate data analysis. Frequently, however, we want to compare two data sets and decide whether differences in their trends could have occurred by chance. Some examples are:

(a) comparing data sets that purport to represent the same variable (such as two versions of a satellite data set) – an example is given in Figure 2;

(b) comparing the same variable at different levels in the atmosphere (such as surface and tropospheric data); or

(c) comparing models and observations.

In the first case (Figure 2), we know that the data sets being compared are attempts to measure precisely the same thing, so that differences can arise only as a result of differences in the methods used to create the final data sets from the same "raw" original data. Here, there is a pitfall that some practitioners fall prey to by using what, at first thought, seems to be a reasonable approach. In this naive method, one would first construct C.I.s for the individual trend estimates by applying the single sample methods described above. If the two C.I.s overlapped, then we would conclude that there was no significant difference between the two trends. This approach, however, is seriously flawed.

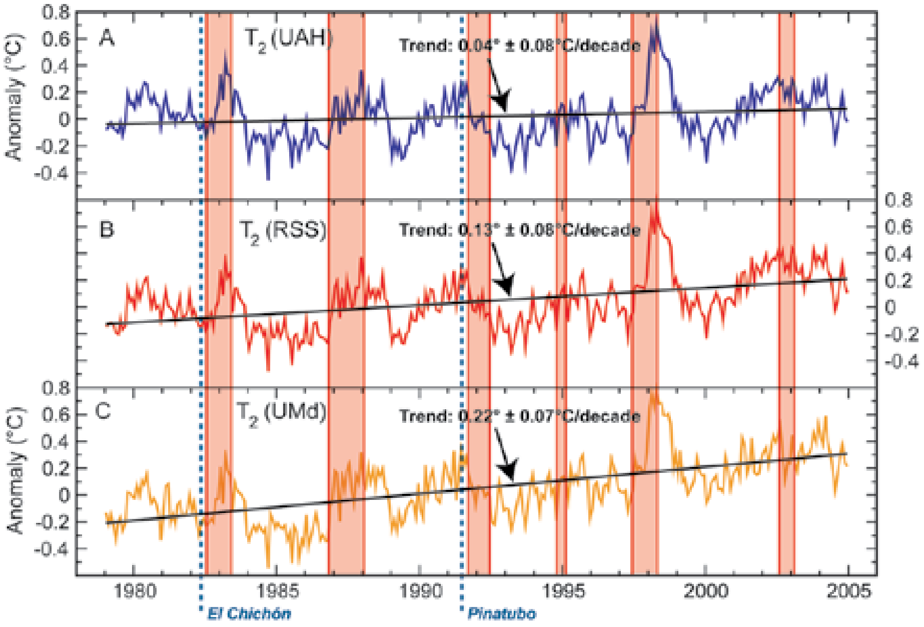

Figure 2: Three estimates of global-mean temperature changes for MSU channel 2 (T_2), expressed as anomalies relative to the 1979 to 1999 mean. Data are from: A, the University of Alabama in Huntsville (UAH); B, Remote Sensing Systems (RSS); and C, the University of Maryland (UMd) The estimates employ the same "raw" satellite data, but make different choices for the adjustments required to merge the various satellite records and to correct for instrument biases. The statistical uncertainty is virtually the same for all three series. Differences between the series give some idea of the magnitude of structural uncertainties. Volcano eruption and El Niño information are as in Figure 1. The trend values are as given in Chapter 3, Table 3.3. The ± values define the 95% confidence intervals for the trends, also from Chapter 3, Table 3.3.

[5] From the time series of residuals about the fitted line.

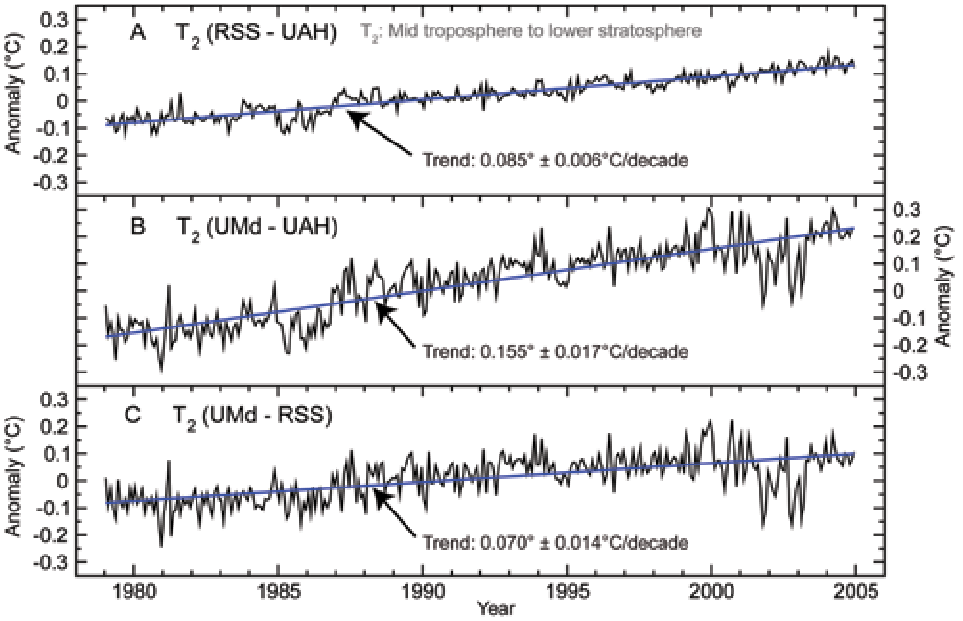

Figure 3: Difference series for the global-mean MSU T_2 series shown in Figure 2. Variability about the trend line is least for the UAH minus RSS series indicating closer correspondence between these two series than between UMd and either UAH or RSS. The trend values are consistent with results given in Chapter 3, Table 3.3, with greater precision given purely for mathematical convenience. The ± values define the 95% confidence intervals for the trends (see also Figure 4).

An analogous problem, comparing two means rather than two trends, discussed by Lanzante (2005), gives some insights. In this case, it is necessary to determine the standard error for the difference between two means. If this standard error is denoted "s", and the individual standard errors are s_1 and s_2, then

$$s^2 = (s_1)^2 + (s_2)^2 \quad …..(10)$$

The new standard error is often called the pooled standard error, and the pooling method is sometimes called "combining standard errors in quadrature." In some cases, when the trends come from data series that are unrelated (as when model and observed data are compared; case (c) above) a similar method may be applied to trends. If the data series are correlated with each other, however (cases (a) and (b)), this procedure is not correct. Here, the correct method is to produce a difference time series by subtracting the first data point in series 1 from the first data point in series 2, the second data points, the third data points, etc. The result of doing this with the microwave sounding unit channel 2 (MSU T_2) data shown in Figure 2 is shown in Figure 3. To assess the significance of trend differences we then apply the same methods used for trend assessment in a single data series to the difference series.

Analyzing differences removes the variability that is common to both data sets and isolates those differences that may be due to differences in data set production methods, temperature measurement methods (as in comparing satellite and radiosonde data), differences in spatial coverage, etc.

Figures 2 and 3 provide a striking example of this. Here, the three series in Figure 2 have very similar volcanic and ENSO signatures. In the individual series, these aspects are noise that obscures the underlying linear trend and inflates the standard error and the trend uncertainty. Since this noise is common to each series, differencing has the effect of canceling out a large fraction of the noise. This is clear from Figure 3, where the variability about the trend lines is substantially reduced. Figure 4 shows the effects on the trend confidence intervals (taking due account of autocorrelation effects). Even though the individual series look very similar in Figure 2, this is largely an artifact of similarities in the noise. It is clear from Figures 3 and 4 that there are, in fact, very significant differences in the trends, reflecting differences in the methods of construction used for the three MSU T_2 data sets.

Comparing model and observed data for a single variable, such as surface temperature, tropospheric temperature, etc., is a different problem. Here, when using data from a state-of-the-art climate model (a coupled Atmosphere/Ocean General Circulation Model[6], or "AOGCM"), there is no reason to expect the background variability to be common to both the model and observations. AOGCMs generate their own internal variability entirely independently of what is going on in the real world. In this case, standard errors for the individual trends can be combined in quadrature (equation [10]). (There are some model/observed data comparison cases where an examination of the difference series may still be appropriate, such as in experiments where an atmospheric GCM is forced by observed sea surface temperature varia-

[6] An AOGCM interactively couples together a three-dimensional Ocean General Circulation Model (OGCM) and an Atmospheric GCM (AGCM). The components are free to interact with one another and they are able to generate their own internal variability in much the same way that the real-world climate system generates its internal variability (internal variability is variability that is unrelated to external forcing). This differs from some other types of model (*e.g.*, an AGCM) where there can be no component of variability arising from the ocean. An AGCM, therefore, cannot generate variability arising from ENSO, which depends on interactions between the atmosphere and ocean.

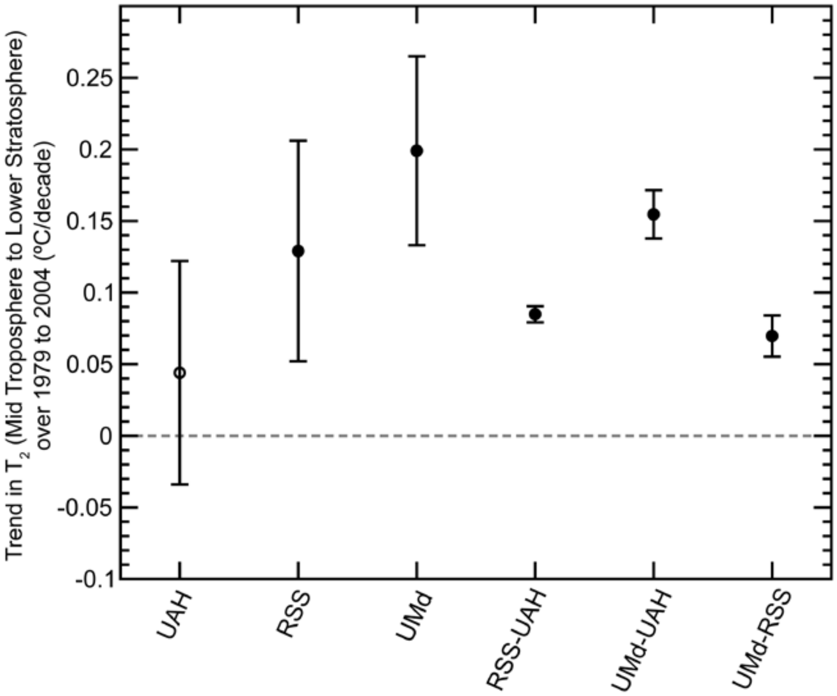

Figure 4: 95% confidence intervals for the three global-mean MSU T_2 series shown in Figure 2 (see Table 3.3 in Chapter 3), and for the three difference series shown in Figure 3.

tions so that ocean-related variability should be common to both the observations and the model.)

For other comparisons, the appropriate test will depend on the degree of similarity between the data sets expected for perfect data. For example, a comparison between MSU T_2 and MSU T_{2LT} produced by a single group should use the difference test – although interpretation of the results may be tricky because differences may arise either from construction methods or may represent real physical differences arising from the different vertical weighting profiles, or both.

There is an important implication of this comparison issue. While it may be common practice to use error bars to illustrate C.I.s for trends of individual time series, when the primary concern (as it is in many parts of this Report) is the comparison of trends, individual C.I.s can be misleading. A clear example of this is given in Figure 4 (based on information in Figures 2 and 3). Individual C.I.s for the three MSU T_2 series overlap, but the C.I.s for the difference series show that there are highly significant differences between the three data sets. Because of this, in some cases in this Report, where it might seem that error bars should be given, we consider the disadvantage of their possible misinterpretation to outweigh their potential usefulness. Individual C.I.s for all trends are, however, given in Tables 3.2, 3.3, 3.4 and 3.5 of Chapter 3; and we also express individual trend uncertainties through the use of significance levels. As noted in Section (9) below, there are other reasons why error bars can be misleading.

(8) MULTIPLE AOGCM SIMULATIONS

Both models and the real world show weather variability and other sources of internal variability that are manifest on all time scales, from daily up to multi-decadal. With AOGCM simulations driven by historical forcing spanning the late-19th and 20th centuries, therefore, a single run with a particular model will show not only the externally forced signal, but also, superimposed on this, underlying internally generated variability that is similar to the variability we see in the real world. In contrast to the real world, however, in the model world we can perturb the model's initial conditions and re-run the same forcing experiment. This will give an entirely different realization of the model's internal variability. In each case, the output from the model is a combination of signal (the response to the forcing) and noise (the internally generated component). Since the noise parts of each run are unrelated, averaging over a number of realizations will tend to cancel out the noise and, hence, enhance the visibility of the signal. It is common practice, therefore, for any particular forcing experiment with an AOGCM, to run multiple realizations of the experiment *(i.e.,* an ensemble of realizations). An example is given in Figure 5, which shows four separate realizations and their ensemble average for a simulation using realistic 20th century forcing (both natural and anthropogenic).

This provides us with two different ways to assess the uncertainties in model results, such as in the model-simulated temperature trend over recent decades. One method is to express uncertainties using the spread of trends across the ensemble members (see, *e.g.,* Figures 3 and 4 in the Executive Summary). Alternatively, the temperature series from the individual ensemble members may be averaged and the trend and its uncertainty calculated using these average data.

Ensemble averaging, however, need not reduce the width of the trend confidence interval compared with an individual realization. This is because of compensating factors: the time series variability will be reduced by the averaging process (as is clear in Figure 5), but, because averaging can inflate

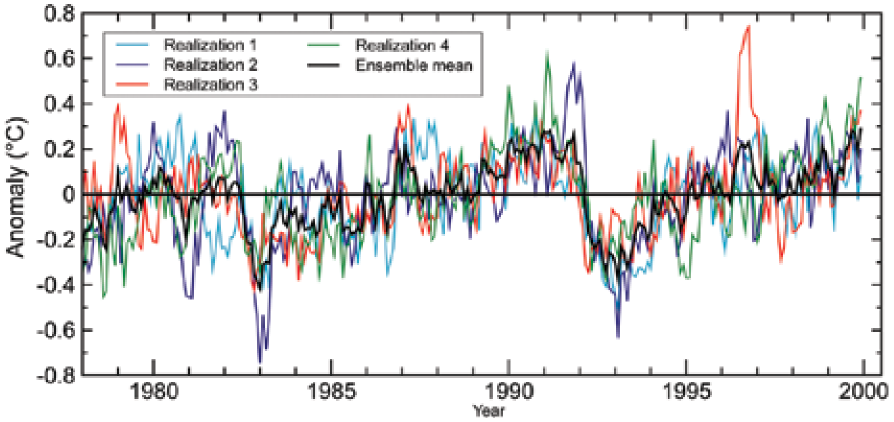

Figure 5: Four separate realizations of model realizations of global-mean MSU channel 2 (T_2) temperature changes, and their ensemble average, for a simulation using realistic 20th Century forcing (both natural and anthropogenic) carried out with one of the National Centre for Atmospheric Research's AOGCMs, the Parallel Climate Model (PCM). The cooling events around 1982/3 and 1991/2 are the result of imposed forcing from the eruptions of El Chichón (1982) and Mt. Pinatubo (1991). Note that the El Chichón cooling in these model simulations is more obvious than in the observed data shown in Figure 1. In the real world, a strong El Niño warming event occurred at the same time as the volcanic cooling, largely masking this cooling. In the four model worlds shown here, the sequences of El Niño events, which necessarily occurred at different times in each simulation, never overlapped with the El Chichón cooling.

the level of autocorrelation, there may be a compensating increase in uncertainty due to a reduction in the effective sample size. This is illustrated in Figure 6.

Averaging across ensemble members, however, does produce a net gain. Although the width of the C.I. about the mean trend may not be reduced relative to individual trend C.I.s, averaging leaves just a single best-fit trend rather than a spread of best-fit trend values.

(9) PRACTICAL VERSUS STATISTICAL SIGNIFICANCE

The Sections above have been concerned primarily with statistical uncertainty, uncertainty arising from random noise in climatological time series – *i.e.,* the uncertainty in how well a data set fits a particular "model" (a straight line in the linear trend case). Statistical noise, however, is not the only source of uncertainty in assessing trends. Indeed, as amply illustrated in this Report, other sources of uncertainty may be more important.

The other sources of uncertainty are the influences of non-climatic factors. These are referred to in this Report as "construction uncertainties." When we construct climate data records that are going to be used for trend (or other statistical) analyses, we attempt to minimize construction uncertainties by removing, as far as possible, non-climatic biases that might vary over time and so impart a spurious trend or trend component – a process referred to as "homogenization."

The need for homogenization arises in part because most observations are made to serve the short-term needs of weather forecasting (where the long-term stability of the observing system is rarely an important consideration). Most records therefore contain the effects of changes in instrumentation, instrument exposure, and observing practices made for a variety of reasons. Such changes generally introduce spurious non-climatic changes into data records that, if not accounted for, can mask (or possibly be mistaken for) an underlying climate signal.

An added problem arises because temperatures are not always measured directly, but through some quantity related to temperature. Adjustments must therefore be made to obtain

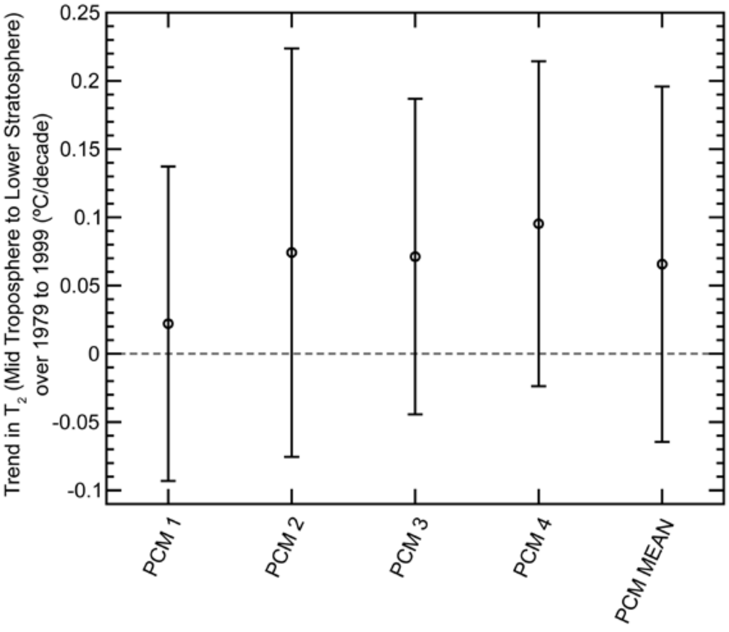

Figure 6: 95% confidence intervals for individual model realizations of global-mean MSU T_2 temperature changes (as shown in Figure 5), compared with the 95% confidence interval for the four-member ensemble average.

temperature information. The satellite-based microwave sounding unit (MSU) data sets provide an important example. For MSU temperature records, the quantity actually measured is the upwelling emission of microwave radiation from oxygen atoms in the atmosphere. MSU data are also affected by numerous changes in instrumentation and instrument exposure associated with the progression of satellites used to make these measurements.

Thorne *et al.* (2005) divide construction uncertainty into two components: “structural uncertainty” and “parametric uncertainty.” Structural uncertainty arises because different investigators may make different plausible choices for the method (or “model”) that they apply to make corrections or “adjustments” to the raw data. Differences in the choice of adjustment model and its structure lead to structural uncertainties. Parametric uncertainties arise because, once an adjustment model has been chosen, the values of the parameters in the model still have to be quantified. Since these values must be determined from a finite amount of data, they will be subject to statistical uncertainties.

Sensitivity studies using different parameter choices may allow us to quantify parametric uncertainty, but this is not always done. Quantifying structural uncertainty is very difficult because it involves consideration of a number of fundamentally different (but all plausible) approaches to data set homogenization, rather than simple parameter “tweaking.” Differences between results from different investigators give us some idea of the magnitude of structural uncertainty, but this is a relatively weak constraint. There are a large number of conceivable approaches to homogenization of any particular data set, from which we are able only to consider a small sample – and this may lead to an under-estimation of structural uncertainty. Equally, if some current homogenization techniques are flawed then the resulting uncertainty estimate will be too large.

An example is given above in Figure 2, showing three different MSU T_2 records with trends of 0.044ºC/decade, 0.129ºC/decade, and 0.199ºC/decade over 1979 through 2004. These differences, ranging from 0.070ºC/decade to 0.155ºC/decade (Figure 3), represent a considerable degree of construction uncertainty. For comparison, the statistical uncertainty in the individual data series, as quantified by the 95% confidence intervals, ranges between ±0.066 and ±0.078ºC/decade; so uncertainties from these two sources are of similar magnitude.

An important implication of this comparison is that statistical and construction uncertainties may be of similar magnitude. For this reason, showing, through confidence intervals, information about statistical uncertainty alone, without giving any information about construction uncertainty, can be misleading.