

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19

APPENDIX: STATISTICAL ISSUES REGARDING TRENDS

Tom M.L. Wigley

Abstract:

21

22 The purpose of this Appendix is to explain the statistical terms and methods used in this Report.

23 We begin by introducing a number of terms: mean, standard deviation, variance, linear trend,

24 sample, population, signal, and noise. Examples are given of linear trends in surface,

25 tropospheric, and stratospheric temperatures. The least squares method for calculating a best fit

26 linear trend is described. The method for quantifying the statistical uncertainty in a linear trend is

27 explained, introducing the concepts of standard error, confidence intervals, and significance

28 testing. A method to account for the effects of temporal autocorrelation on confidence intervals

29 and significance tests is described. The issue of comparing two data sets to decide whether

30 differences in their trends could have occurred by chance is discussed. The analysis of trends in

31 state-of-the-art climate model results is a special case because we frequently have an ensemble of

32 simulations for a particular forcing case. The effect of ensemble averaging on confidence

33 intervals is illustrated. Finally, the issue of practical versus statistical significance is discussed. In

34 practice, it is important to consider construction uncertainties as well as statistical uncertainties.

35 An example is given showing that these two sources of trend uncertainty can be of comparable

36 magnitude.

37

38 **(1) Why do we need statistics?**

39

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

42

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

47

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

56

57 As an analogy, the population might be an infinite jar of marbles, a certain proportion of which
58 (say 60%) is blue and the rest (40%) are red. We can only draw off a finite number of these
59 marbles (a sample) at a time; and, when we measure the numbers of blue and red marbles in the
60 sample, they need not be in the precise ratio 60:40. The ratio we measure is called a “sample

61 statistic”. It is an estimate of some hypothetical underlying population value (the corresponding
62 “population parameter”). The techniques of statistical science allow us to make optimum use of
63 the sample statistic and obtain a best estimate of the population parameter. Statistical science
64 also allows us to quantify the uncertainty in this estimate.

65

(2) Definition of a linear trend

67

68 If data show underlying smooth changes with time, we refer to these changes as a trend. The
69 simplest type of change is a linear (or straight line) trend, a continuous increase or decrease over
70 time. For example, the net effect of increasing greenhouse-gas concentrations and other human-
71 induced factors is expected to cause warming at the surface and in the troposphere and cooling in
72 the stratosphere (see Figure 1). Warming corresponds to a positive (or increasing) linear trend,
73 while cooling corresponds to a negative (or decreasing) trend. These changes are not expected to
74 be strictly linear, but the linear trend provides a simple way of characterizing the change and of
75 quantifying its magnitude.

76

77

78

79 Figure 1: Examples of temperature time series with best-fit (least squares) linear trends: A, global-mean surface
80 temperature from the UKMO Hadley Centre/Climatic Research Unit data set (HadCRUT2v); and B, MSU channel 4
81 data (T_4) for the lower stratosphere from the University of Alabama at Huntsville (UAH). Note the much larger
82 temperature scale on the lower panel. Temperature changes are expressed as anomalies relative to the 1979 to 1999
83 mean (252 months). Dates for the eruptions of El Chichón and Mt Pinatubo are shown by vertical lines. El Niños are
84 shown by the shaded areas.

85

86

87 Alternatively, there may be some physical process that causes a rapid switch or change from one
88 mode of behavior to another. In such a case the overall behavior might best be described as a
89 linear trend to the changepoint, a step change at this point, followed by a second linear trend
90 portion. Many temperature data sets show this type of behavior, arising from a change in the
91 pattern of variability in the Pacific that occurred around 1976 (a switch in a mode of climate
92 variability called the Pacific Decadal Oscillation).

93

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

98

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

102

103 Linear temperature trends are usually quantified as the temperature change per year or per
104 decade (even when the data are available on a month by month basis). For example, the trend for
105 the surface temperature data shown below in Figure 1 is 0.169°C per decade. This is a more
106 convenient representation than the trend per month, which would be $0.169/120 = 0.00141^\circ\text{C}$ per
107 month, a very small number. An alternative method is to use the “total trend” over the full data
108 period – i.e., the total change for the fitted line from the start to the end of the record (see Figure
109 2 in the Executive Summary). In Figure 1, the data shown span January 1979 through December
110 2004 (312 months or 2.6 decades). The total change is therefore $0.169 \times 2.6 = 0.439^\circ\text{C}$.

111

112 **(3) Expected temperature changes: signal and noise**

113

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

121

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

128

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

135
136 The reason why D&A is a difficult and challenging statistical problem is because climate signals
137 do not occur in isolation. In addition to these signals, temperature fluctuations in all parts of the
138 atmosphere occur even in the absence of external driving forces. These internally-driven
139 fluctuations represent the “noise” against which we seek to identify specific externally-forced
140 signals. All climate records, therefore, are “noisy”, with the noise of this natural variability
141 tending to obscure the externally-driven changes. Figure 1 illustrates this. At the surface, a
142 primary noise component is the variability associated with ENSO (the El Niño/Southern
143 Oscillation phenomenon)¹, while, in the stratosphere, if our concern is to identify anthropogenic
144 influences, the warmings after the eruptions of El Chichón and Mt Pinatubo constitute noise.

145
146 If the underlying response to external forcing is small relative to the noise, then, by chance, we
147 may see a trend in the data due to random fluctuations purely as a result of the noise. The science
148 of statistics provides methods through which we can decide whether the trend we observe is
149 “real” (i.e., a signal associated with some causal factor) or simply a random fluctuation (i.e.,
150 noise).

151

152 **(4) Deriving trend statistics**

153

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

161

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

166

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

169

$$170 \quad Y_{\text{est}} = b_0 + b_1 X_1 + b_2 X_2 + \dots \quad \dots (1)$$

171

172 where the subscript ‘est’ is used to indicate that this is the estimate of Y that is given by the fitted
173 relationship. Differences between the actual and estimated values of Y, the residuals, are defined
174 by

175

$$176 \quad e = Y - Y_{\text{est}} \quad \dots (2)$$

177

178 For linear trend analysis of temperature data (T) there is a single predictor, time (t; t = 1,2,3, ...).

179 The time points are almost always evenly spaced, month by month, year by year, etc. – but this is

180 not a necessary restriction. In the linear trend case, the regression equation becomes:

181

$$182 \quad T_{\text{est}} = a + b t \quad \dots (3)$$

183

184 In equ. (3), ‘b’ is the slope of the fitted line – i.e., the linear trend value. This is a sample statistic,

185 i.e., it is an estimate of the corresponding underlying population parameter. To distinguish the

186 population parameter from the sample value, the population trend value is denoted β .

187

188 The formula for b is:

189

$$190 \quad b = [\sum((t - \langle t \rangle)T_t)] / [\sum(t - \langle t \rangle)^2] \quad \dots (4)$$

191

192 where $\langle \dots \rangle$ denotes the mean value, and the summation is over t = 1,2,3, ... n (i.e., the sample

193 size is n). T_t denotes the value of temperature, T, at time ‘t’. Equation (4) produces an unbiased

194 estimate³ of population trend, β .

195

196 For the usual case of evenly spaced time points, $\langle t \rangle = (n+1)/2$, and

197

$$\sum (t - \langle t \rangle)^2 = n(n^2 - 1)/12 \quad \dots (5)$$

199

200 When we are examining deviations from the fitted line the sign of the deviation is not important.

201 This is why we consider the squares of the residuals in least squares regression. An important

202 and desirable characteristic of the least squares method is that the average of the residuals is

203 zero.

204

205 Estimates of the linear trend are sensitive to points at the start or end of the data set. For

206 example, if the last point, by chance, happened to be unusually high, then the fitted trend might

207 place undue weight on this single value and lead to an estimate of the trend that was too high.

208 This is more of a problem with small sample sizes (i.e., for trends over short time periods). For

209 example, if we considered tropospheric data over 1979 through 1998, because of the unusual

210 warmth in 1998 (associated with the strong 1997/98 El Niño; see Figure 1), the calculated trend

211 may be an overestimate of the true underlying trend.

212

213 There are alternative ways to estimate the linear trend that are less sensitive to endpoints.

214 Although we recognize this problem, for the data used in this Report tests using different trend

215 estimators give results that are virtually the same as those based on the standard least-squares

216 trend estimator.

217

218

218 **(5) Trend uncertainties**

219

220 Some examples of fitted linear trend lines are shown in Figure 1. This Figure shows monthly
221 temperature data for the surface and for the lower stratosphere (MSU channel 4) over 1979
222 through 2004 (312 months). In both cases there is a clear trend, but the fit is better for the surface
223 data. The trend values (i.e., the slopes of the best fit straight lines that are shown superimposed
224 on monthly data) are $+0.169^{\circ}\text{C}/\text{decade}$ for the surface and $-0.452^{\circ}\text{C}/\text{decade}$ for the stratosphere.
225 For the stratosphere, although there is a pronounced overall cooling trend, as noted above
226 describing the change simply as a linear cooling considerably oversimplifies the behavior of the
227 data¹.

228

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

236

237 The population parameter we are trying to estimate here is the standard deviation of the trend
238 estimate, or its square, the variance of the distribution of b , which we denote $\text{Var}(b)$. The larger
239 the value of $\text{Var}(b)$, the more uncertain is b as an estimate of the population value, \square .

240

241 The formula for $\text{Var}(b)$ is ...

242

$$243 \quad \text{Var}(b) = [\sigma^2]/[\sum(t - \langle t \rangle)^2] \quad \dots (6)$$

244

245 where σ^2 is the population value for the variance of the residuals. Unfortunately, we do not in

246 general know what σ^2 is, so we must use an unbiased sample estimate of σ^2 . This estimate is

247 known as the Mean Square Error (MSE), defined by ...

248

$$249 \quad \text{MSE} = [\sum(e^2)]/(n - 2) \quad \dots (7)$$

250

251 Hence, equ. (6) becomes

252

$$253 \quad \text{Var}(b) = (\text{SE})^2 = \text{MSE}/[\sum(t - \langle t \rangle)^2] \quad \dots (8)$$

254

255 where SE, the square root of $\text{Var}(b)$, is called is called the “standard error” of the trend estimate.

256 The smaller the value of the standard error, the better the fit of the data to the linear change

257 description and the smaller the uncertainty in the sample trend as an estimate of the underlying

258 population trend value. The standard error is the primary measure of trend uncertainty. The

259 standard error will be large if the MSE is large, and the MSE will be large if the data points show

260 large scatter about the fitted line.

261

262 There are assumptions made in going from equ. (6) to (8): viz. that the residuals have mean zero

263 and common variance, that they are Normally (or “Gaussian”) distributed⁴, and that they are

264 uncorrelated or statistically independent. In climatological applications, the first two are
265 generally valid. The third assumption, however, is often not justified. We return to this below.
266

267 **(6) Confidence intervals and significance testing**

268

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

272

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

281

282 ***Confidence intervals***

283

284 The larger the standard error of the trend, the more uncertain is the slope of the fitted line. We
285 express this uncertainty probabilistically by defining confidence intervals for the trend associated
286 with different probabilities. If the distribution of trend values were strictly Gaussian, then the
287 range $b - SE$ to $b + SE$ would represent the 68% confidence interval (C.I.) because the
288 probability of a value lying in that range for a Gaussian distribution is 0.68. The range $b -$
289 $1.645(SE)$ to $b + 1.645(SE)$ would give the 90% C.I.; the range $b - 1.96(SE)$ to $b + 1.96(SE)$

290 would give the 95% C.I.; and so on. Quite often, for simplicity, we use $b - 2(SE)$ to $b + 2(SE)$ to
291 represent (to a good approximation) the 95% confidence interval.

292

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

298

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

306

307 *Significance testing*

308

309 An alternative method for assessing trends is hypothesis testing. In practice, it is much easier to
310 disprove rather than prove a hypothesis. Thus, the standard statistical procedure in significance
311 testing is to set up a hypothesis that we would like to disprove. This is called a “null hypothesis”.
312 In the linear trend case, we are often interested in trying to decide whether an observed data trend

313 that is noticeably different from zero is sufficiently different that it could not have occurred by
314 chance – or, at least, that the probability that it could have occurred by chance is very small. The
315 appropriate null hypothesis in this case would be that there was no underlying trend ($\square = 0$). If
316 we disprove (i.e., “reject”) the null hypothesis, then we say that the observed trend is
317 “statistically significant” at some level of confidence and we must accept some alternate
318 hypothesis. The usual alternate hypothesis in temperature analyses is that the data show a real,
319 externally-forced warming (or cooling) trend. (In cases like this, the statistical analysis is
320 predicated on the assumption that the observed data are reliable. If a trend were found to be
321 statistically significant, then an alternative possibility might be that the observed data were
322 flawed.)

323

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

330

331 An important factor in significance testing is whether we are concerned about deviations from
332 some hypothesized value in any direction or only in one direction. This leads to two types of
333 significance test, referred to as “one-tailed” (or “one-sided”) and “two-tailed” tests. A one-tailed
334 test arises when we expect a trend in a specific direction (such as warming in the troposphere due
335 to increasing greenhouse-gas concentrations). Two-tailed tests arise when we are concerned only

336 with whether the trend is different from zero, with no specification of whether the trend should
337 be positive or negative. In temperature trend analyses we generally know the sign of the expected
338 trend, so one-tailed tests are more common.

339

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

349

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

357

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

363
364 For the null hypothesis of zero trend, the distribution of trend values has mean zero and standard
365 deviation equal to the standard error. Knowing this, we can calculate the probability that the
366 actual trend value could have exceeded the observed value by chance if the null hypotheses were
367 true (or, if we were using a two-tailed test, the probability that the magnitude of the actual trend
368 value exceeded the magnitude of the observed value). This probability is called the ‘p-value’. For
369 example, a p-value of 0.03 would be judged significant at the 5% level (since $0.03 < 0.05$), but not
370 at the 1% level (since $0.03 > 0.01$).

371
372 Since both the calculation of confidence intervals and significance testing employ information
373 about the distribution of trend values, there is a clear link between confidence intervals and
374 significance testing.

375

376 *A complication; the effect of autocorrelation*

377

378 The significance of a trend, and its confidence intervals, depend on the standard error of the trend
379 estimate. The formula given above for this standard error (equ. (8)) is, however, only correct if
380 the individual data points are unrelated, or statistically independent. This is not the case for most

381 temperature data, where a value at a particular time usually depends on values at previous times;
382 i.e., if it is warm today, then, on average, it is more likely to be warm tomorrow than cold. This
383 dependence is referred to as “temporal autocorrelation” or “serial correlation”. When data are
384 autocorrelated (i.e., when successive values are not independent of each other), many statistics
385 behave as if the sample size was less than the number of data points, n .

386
387 One way to deal with this is to determine an “effective sample size”, which is less than n , and
388 use it instead of n in statistical formulae and calculations. The extent of this reduction from n to
389 an effective sample size depends on how strong the autocorrelation is. Strong autocorrelation
390 means that individual values in the sample are far from being independent, so the effective
391 number of independent values must be much smaller than the sample size. Strong autocorrelation
392 is common in temperature time series. This is accounted for by reducing the divisor ‘ $n - 2$ ’ in the
393 mean square error term (equ. (7)) that is crucial in determining the standard error of the trend
394 (equ. (8)).

395
396 There are a number of ways that this autocorrelation effect may be quantified. A common and
397 relatively simple method is described in Santer et al. (2000). This method makes the assumption
398 that the autocorrelation structure of the temperature data may be adequately described by a “first-
399 order autoregressive” process, an assumption that is a good approximation for most climate data.
400 The lag-1 autocorrelation coefficient (r_1) is calculated from the observed data⁵, and the effective
401 sample size is determined by

402
403
$$n_{\text{eff}} = n (1 - r_1) / (1 + r_1) \quad \dots (9)$$

404

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

407

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

415

416 **(7) Comparing trends in two data sets**

417

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

422

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

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

427 (c) comparing models and observations.

428

429

430

431

432 Figure 2: Three estimates of temperature changes for MSU channel 2 (T_2), expressed as anomalies relative to the
433 1979 to 1999 mean. Data are from: A, the University of Alabama at Huntsville (UAH); B, Remote Sensing Systems
434 (RSS); and C, the University of Maryland (U.Md.) The estimates employ the same ‘raw’ satellite data, but make
435 different choices for the adjustments required to merge the various satellite records and to correct for instrument
436 biases. The statistical uncertainty is virtually the same for all three series. Differences between the series give some
437 idea of the magnitude of structural uncertainties. Volcano eruption and El Niño information are as in Figure 1.
438

439

440 In the first case (Figure 2), we know that the data sets being compared are attempts to measure
441 precisely the same thing, so that differences can arise only as a result of differences in the
442 methods used to create the final data sets from the same ‘raw’ original data. Here, there is a
443 pitfall that some practitioners fall prey to by using what, at first thought, seems to be a
444 reasonable approach. In this naïve method, one would first construct C.I.s for the individual trend
445 estimates by applying the single sample methods described above. If the two C.I.s overlapped,

446 then we would conclude that there was no significant difference between the two trends. This
447 approach, however, is seriously flawed.

448

449 An analogous problem, comparing two means rather than two trends, discussed by Lanzante
450 (2005), gives some insights. In this case, it is necessary to determine the standard error for the
451 difference between two means. If this standard error is denoted ‘s’, and the individual standard
452 errors are s_1 and s_2 , then

453

$$454 \quad s^2 = (s_1)^2 + (s_2)^2 \quad \dots(10)$$

455

456 The new standard error is often called the pooled standard error, and the pooling method is
457 sometimes called “combining standard errors in quadrature”. In some cases, when the trends
458 come from data series that are unrelated (as in the model/observed data comparison case; (c)
459 above) a similar method may be applied to trends. If the data series are correlated with each
460 other, however (cases (a) and (b)), this procedure is not correct. Here, the correct method is to
461 produce a difference time series by subtracting the first data point in series 1 from the first data
462 point in series 2, the second data points, the third data points, etc. The result of doing this with
463 the microwave sounding unit channel 2 (MSU T₂) data shown in Figure 2 is shown in Figure 3.
464 To assess the significance of trend differences we then apply the same methods used for trend
465 assessment in a single data series to the difference series.

466

467

468

469 Figure 3: Difference series for the MSU T₂ series shown in Figure 2. Variability about the trend line is least for the
470 UAH minus RSS series indicating closer correspondence between these two series than between U.Md. and either
471 UAH or RSS.
472

473

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

478

479 Figures 2 and 3 provide a striking example of this. Here, the three series in Figure 2 have very
480 similar volcanic and ENSO signatures. In the individual series, these aspects are noise that
481 obscures the underlying linear trend and inflates the standard error and the trend uncertainty.

482 Since this noise is common to each series, differencing has the effect of canceling out a large
483 fraction of the noise. This is clear from Figure 3, where the variability about the trend lines is
484 substantially reduced. Figure 4 shows the effects on the trend confidence intervals (taking due
485 account of autocorrelation effects). Even though the individual series look very similar in Figure
486 2, this is largely an artifact of similarities in the noise. It is clear from Figures 3 and 4 that there
487 are, in fact, very significant differences in the trends, reflecting differences in their methods of
488 construction.

489

490

491

492 Figure 4: 95% confidence intervals for the three MSU T_2 series shown in Figure 2 (see Table 3.3 in Chapter 3), and
493 for the three difference series shown in Figure 3.
494

495
496 Comparing model and observed data for a single variable, such as surface temperature,
497 tropospheric temperature, etc., is a different problem. Here, when using data from a state-of-the-
498 art climate model (a coupled Atmosphere/Ocean General Circulation Model⁶, or “AOGCM”),
499 there is no reason to expect the background variability to be common to both the model and
500 observations. AOGCMs generate their own internal variability entirely independently of what is
501 going on in the real world. In this case, standard errors for the individual trends can be combined
502 in quadrature (equ. (10)). (There are some model/observed data comparison cases where an
503 examination of the difference series may still be appropriate, such as in experiments where an
504 atmospheric GCM is forced by observed sea surface temperature variations so that ocean-related
505 variability should be common to both the observations and the model.)
506

507 For other comparisons, the appropriate test will depend on the degree of similarity between the
508 data sets expected for perfect data. For example, a comparison between MSU T_2 and MSU T_{2LT}
509 produced by a single group should use the difference test – although interpretation of the results
510 may be tricky because differences may arise either from construction methods or may represent
511 real physical differences arising from the different vertical weighting profiles, or both.
512

513 There is an important implication of this comparison issue. While it may be common practice to
514 use error bars to illustrate C.I.s for trends of individual time series, when the primary concern (as
515 it is in many parts of this Report) is the comparison of trends, individual C.I.s can be quite

516 misleading. In some cases in this Report, therefore, where it might seem that error bars should be
517 given, we consider the disadvantage of their possible misinterpretation to outweigh their
518 potential usefulness. Instead, we have chosen to express individual trend uncertainties through
519 the use of significance levels, which can be represented by a less obtrusive symbol. As noted in
520 Section (9) below, there are other reasons why error bars can be misleading.

521

(8) Multiple AOGCM simulations

523

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

539

540

541

542 Figure 5: Four separate realizations of model realizations of global-mean MSU channel 2 (T_2) temperature changes,
543 and their ensemble average, for a simulation using realistic 20th Century forcing (both natural and anthropogenic)
544 carried out with one of the National Centre for Atmospheric Research's AOGCMs, the Parallel Climate Model
545 (PCM). The cooling events around 1982/3 and 1991/2 are the result of imposed forcing from the eruptions of El
546 Chichón (1982) and Mt. Pinatubo (1991). Note that the El Chichón cooling is more obvious than in the observed
547 data shown in Fig. 1, because, in the model simulations, the ENSO sequences differed from the real world, and from
548 each other.
549

550

551 This provides us with two different ways to assess the uncertainties in model results, such as in
552 the model-simulated temperature trend over recent decades. One method is to express
553 uncertainties using the spread of trends across the ensemble members (see, e.g., Figures 3 and 4
554 in the Executive Summary). Alternatively, the temperature series from the individual ensemble
555 members may be averaged and the trend and its uncertainty calculated using these average data.
556

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

560 averaging can inflate the level of autocorrelation, there may be a compensating increase in
561 uncertainty due to a reduction in the effective sample size. This is illustrated in Figure 6.
562

563

564 Figure 6: 95% confidence intervals for individual model realizations of MSU T₂ temperature changes (as shown in
565 Fig. 5), compared with the 95% confidence interval for the ensemble (n=4) average.
566

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

570 **(9) Practical versus statistical significance**

571

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

577

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

583

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

590

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

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

598

599 Thorne et al. (2005) divide construction uncertainty into two components: “structural
600 uncertainty” and “parametric uncertainty”. Structural uncertainty arises because there is no *a*
601 *priori* knowledge of the correct way to homogenize a given raw data set. Independent
602 investigators given the same raw data will make different seemingly sensible and defensible
603 adjustment choices based on their training, technological options at their disposal, and their
604 understanding of the raw data, amongst other factors. Differences in the choice of adjustment
605 pathway and its structure lead to structural uncertainties. Parametric uncertainty arises because,
606 once an adjustment approach or pathway has been chosen, additional choices may have to be
607 made with regard to specific correction factors or parameters.

608

609 Sensitivity studies using different parameter choices may allow us to quantify parametric
610 uncertainty, but this is not always done. Quantifying structural uncertainty is very difficult
611 because it involves consideration of a number of fundamentally different (but all plausible)
612 approaches to data set homogenization, rather than simple parameter “tweaking”. Differences
613 between results from different investigators give us some idea of the magnitude of structural
614 uncertainty, but this is a relatively weak constraint. There are a large number of conceivable
615 approaches to homogenization of any particular data set, from which we are able only to consider

616 a small sample – and this may lead to an under-estimation of structural uncertainty. Equally, if
617 some current homogenization techniques are flawed then the resulting uncertainty estimate will
618 be too large.

619

620 An example is given above in Figure 2, showing three different MSU T₂ records with trends of
621 0.044°C/decade, 0.129°C/decade, and 0.199°C/decade over 1979 through 2004. These
622 differences, ranging from 0.070°C/decade to 0.155°C/decade, represent a considerable degree of
623 construction uncertainty. For comparison, the statistical uncertainty, which is very similar for
624 each series and which can be quantified by the 95% confidence interval, is ±0.066 to
625 ±0.078°C/decade.

626

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

631 Footnotes

632

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

646

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

650

651 ³ An unbiased estimator is one where, if the same experiment were to be performed over and
652 over again under identical conditions, then the long-run average of the estimator will be equal to
653 the parameter that we are trying to estimate. In contrast, in a biased estimator, there will always

654 be some slight difference between the long-run average and the true parameter value that does
655 not tend to zero no matter how many times the experiment is repeated. Since our goal is to
656 estimate population parameters, it is clear that unbiased estimators are preferred.

657

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

662

663 ⁵ From the time series of residuals about the fitted line.

664

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

673

674 **References:**

675

676 Santer, B.D., Wigley, T.M.L., Boyle, J.S., Gaffen, D.J., Hnilo J.J., Nychka, D., Parker, D.E. and
677 Taylor, K.E., 2000: Statistical significance of trends and trend differences in layer-average
678 temperature time series. *Journal of Geophysical Research* **105**, 7337–7356.

679

680 Thorne, P.W., Parker, D.W., Christy, J.R. and Mears, C.A., 2005: Uncertainties in climate
681 trends: lessons from upper-air temperature records. *Bulletin of the American*
682 *Meteorological Society* **86**, 1437–1442.

683

684 Lanzante, J.R., 2005: A cautionary note on the use of error bars. *Journal of Climate* **18**, 3699–
685 3703.

686