# ENHANCEMENT AND QUALITY CONTROL OF CRU MONTHLYTEMPERATURE STATION DATA SET

*Ian Harris, CRU - 18 May 2005*

## 0. INTRODUCTION

The CRU Archive of monthly temperature stations exists as three separate files:

- A Worldwide station series (ordered by WMO number) running from inception to approximately 1990 (sometimes later);
- A companion set of series running from 1991 to 2004/5, (not necessarily all of the above), continuously updated;
- A file of additional US and Canadian temperature station data, most of which are not WMO stations and are not updated.

Only the first of these is principally considered here; the second was checked for outliers and the third was classed as having a low priority due to the denisty of stations over these two countries.

The work (and this summary) is separated into the following areas:

1. New stations
2. Error correction
3. Process improvement

## 1. NEW STATIONS

New stations and/or additional station data were added/replaced for Austria, Mali, Switzerland and the Democratic Republic of Congo.

### 1.1 Austria

A total of 16 Austrian stations had their data completely replaced with new data. The main improvements were in the elimination of outliers and a rehomogenisation of the stations.

### 1.2 Mali

A total of 29 Mali series were affected: 5 had partial new data, 8 had completely new data, and 16 were new stations. One WMO ID needed to be corrected, and several altitudes were corrected.

### 1.3 Switzerland

Five Swiss stations were updated with new data for the period 1864 - 2001. Certain systemic offsets were found to have been introduced; these were found using 'early period' and 'late period' means, and corrected where necessary.

### 1.4 Democratic Republic of Congo

Data for 33 Congolese stations was incorporated as follows: 2 matched with existing 'Congo' data; 17 matched with existing 'Zaire' data; 13 were new stations; and one was found to identify with two existing stations (a situation eventually resolved with the deletion of one station).

## 2. ERROR CORRECTION

Quality-control procedures were applied, to address several error types: administrative, such as incorrect dates or formats; erroneous values, for instance outliers; and potentially-erroneous stations, principally duplicates.

### 2.1 Administrative

Some data codes (two-digit markers in the station headers) were found to be set to '0' – these were reset to 31.

A few cases were found where the First Reliable Year was earlier than the true start year of the data – these were reset to the true start year.

Other administrative issues, such as header and data formats, were investigated but no errors were identified.

### 2.2 erroneous Individual Values

After some basic statistical work, it was decided that the data should be visually inspected for outliers using an interactive program. An initial run-through took several days, and produced 885 potential exceptions. These were reexamined and a final list of 281 exceptions produced. A second interactive program allowed each exception to be assessed alongside data from up to 10 of the nearest stations. Additionally, four decades' worth of US Dept of Commerce/NOAA World Weather Records was consulted as necessary. Common mis-keyings were also identified (for instance, 1 instead of 7). Over 270 individual values were either corrected or deleted.

### 2.3 Duplicate Stations

An examination of the data file identified over 1200 pairs of identical data lines (ie, year plus 12 values all in agreement). A second (programmatic) examination found over 250 sites sharing common segments of data at least five consecutive values in length. A graphical program then allowed the duplicate sections to be compared in context. This exercise resulted in the removal of 53 stations in their entirety, and the additional deletion of much duplicate data. Many of these stations arose in the US and resulted from the incorporation of different datasets over time. Stations often had slightly differing names, co-ordinates and WMO IDs.

An exercise looking for abnormal distribution of values identified several unrealistic instances; in addition to a few corrections, one station was deleted.

### 2.4 Zero Values

Temperature values of zero appear regularly in the dataset when scanned by the eye. Since this could be an erroneous missing code, a check was made to see if zero occurs more regularly than, say, 0.1, 0.2, -0.1, -0.2, etc. The distribution of numbers showed no significant bias towards zero: counts of numbers between -1.0°C and 1.0°C were:

| °C/10 | Count | °C/10 | Count | °C/10 | Count |
|---|---|---|---|---|---|
| -1.0 | 3736 | -3 | 4391 | 4 | 4759 |
| -9 | 4109 | -2 | 4401 | 5 | 4035 |
| -8 | 4266 | -1 | 4969 | 6 | 5009 |
| -7 | 4077 | 0 | 3589 | 7 | 4767 |
| -6 | 4308 | 1 | 4433 | 8 | 4922 |
| -5 | 3798 | 2 | 4613 | 9 | 4796 |
| -4 | 4472 | 3 | 4635 | 10 | 4505 |

## 3. PROCESS IMPROVEMENT

The construction of the gridded output series rests on the availability of good normals and standard deviations for the stations. The existing approach was to use 1961-1990 normals and 1921-1990 standard deviations. This left a considerable shortfall in normals, given that the criteria for generation were that a minimum of 20 values should be present, with at least four in each of the three decades. These generated normals were therefore 'topped up' with normals derived either from WMO records, or inferred from neighbouring stations using alternative time periods.

With the aim of improving the 1961-1990 gridded anomalies, investigations were made into the sources of the major problems these 'additional' normals might produce. These revealed that, unsurprisingly, most issues resulted from the additional normals - particularly the WMO Normals.

After some experimentation, the rule for calculating normals was altered such that the 'n-per-decade' rule was omitted, and the minimum count was reduced to 15 values. This effectively added 617 calculated normals. A number of WMO normals were 'set to missing' when there was little data in the station series. Using these normals, with the 'top up' normals only used for the remainder, gave the following hemispherical means for 1961-1990 anomalies:

Northern Hemisphere (Land): 0.0026

Southern Hemisphere (Land): 0.0032

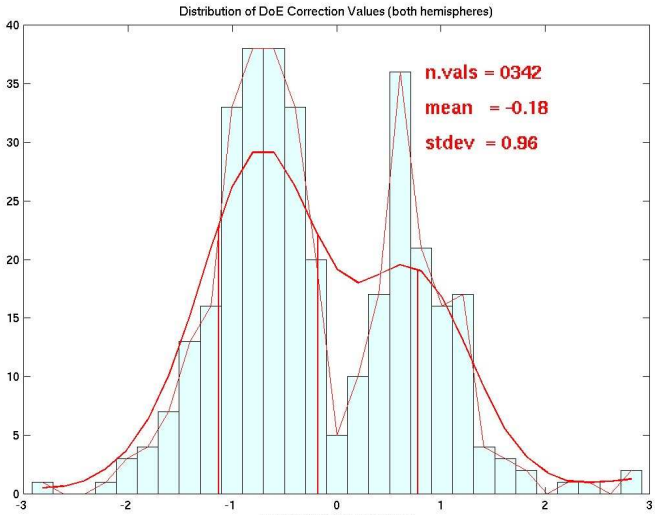
## 4. ADDITIONAL INFORMATION

Adjustment periods were measured; for corrections to the CRU data (DoE 'Yellow Book' corrections) and corrections to the Canadian stations (made in Canada). The lengths of corrected and uncorrected data segments were analysed, giving following information:

| Segments | Mean Length (months) | Std Dev (months) | Segment Count |
|---|---|---|---|
| DoE Corrected | 472.4 | 413.3 | 342 |
| DoE Uncorrected | 450.7 | 344.3 | 389 |
| DoE All | 460.8 | 620.1 | 731 |
| Canada Corrected | 312.7 | 252,8 | 421 |
| Canada Uncorrected | 306.2 | 296.2 | 438 |
| Canada All | 309.4 | 275.7 | 859 |

The mean (ie, yearly mean) corrections for each station in the two sets were also analysed, with the following results:

**DoE Corrections (°C)**
Mean -0.18
Std Dev 0.96
Count 342




**Canadian Corrections (°C)**
Mean -0.13
Std Dev 1.18
Count 421