Ann. Geophys., 24, 2445–2449, 2006
www.ann-geophys.net/24/2445/2006/
© European Geosciences Union 2006

Annales Geophysicae

Impact of missing sounding reports on mandatory levels and tropopause statistics: a case study

J. C. Antuña[1], **J. A. Añel**[2], **and L. Gimeno**[2,3]

[1]Estación Lídar, Centro Meteorológico de Camagüey, Cuba
[2]Departamento de Física Aplicada, Universidad de Vigo, Ourense, Spain
[3]University of Lisbon, CGUL, IDL, Lisbon, Portugal

Received: 22 July 2006 – Revised: 27 September 2006 – Accepted: 16 October 2006 – Published: 20 October 2006

Abstract. This paper describes the effect of missing sounding reports on temperature and pressure mean values for mandatory levels using the aerological information from the Camagüey Meteorological Centre. Also it is described the effect of missing data on mean temperature and pressure values at the multiple tropopause levels. The case study belongs to one station for a time lag of eight years. Up to the present these types of studies have been conducted using simulated datasets. The present one uses a real inhomogeneous radiosonde dataset. The main reason for missing reports were transmission problems and possible encoding-decoding difficulties. It has been found that profiles of the mean temperature and altitude show little differences between the complete and incomplete datasets. Moreover, no statistical significant differences were found for the mean values of the variables for the complete and incomplete datasets. The most probable reason for those results is that the cause of the missing reports has a random behaviour. Finally we have found that the only two effects noticed on the statistics were slightly higher values of the mean temperatures in the complete dataset and the decrease in the percent of multiple tropopause reports for the incomplete dataset.

Keywords. Meteorology and atmospheric dynamics (Climatology) – Atmospheric composition and structure (Pressure, density, and temperature; Instruments and techniques)

1 Introduction

It is very common in atmospheric sciences to carry out research using incomplete datasets. In such cases scientists are in a crossroad. They can make assumptions about how much the incomplete dataset is representative of the inexistent complete dataset or abandon the study. Fortunately, most scientists assume the risks and conduct the research. The radiosonde datasets are one example of the situation described above.

Radiosonde observations are one of the main atmospheric components of the Global Climate Observation System (GCOS), together with the surface network of meteorological stations and the meteorological satellites. But the global coverage of radiosonde stations is not homogeneous over the globe, with fewer stations in the tropics and the Southern hemisphere. An additional difficulty is related to the incompleteness of radiosonde data series. In particular poor coding of data and missing data are attributed by the GCOS monitoring centre to telecommunications failure (WMO, 2005).

The present case study makes use of the fact that not all the radiosonde measurements conducted at the WMO station 78 355 between 1981 and 1988 are available in the most recent developed global radiosonde database. This situation creates a real, not simulated, temporarily inhomogeneous radiosonde dataset, produced by missing data. It is recognized that one of the several causes of the temporal inhomogeneity of radiosonde data is the missing observations. Several studies have been conducted in the past using simulated series or soundings on missing data and its impact in measures of central tendency, variability, and trends (Kidson and Trenberth, 1988; Gaffen et al., 2000; Seidel and Free, 2006). Although the station 78 355 is not part of the GCOS Upper Air Network (GUAN), it is representative of the tropics and in particular of the Caribbean. Only four stations from this region are compiled in GUAN. This station is also the one holding longest records in Cuba, excluding Guantanamo Bay, operated by the U.S.

The recent development and availability of the Integrated Global Radiosonde Archive (IGRA) dataset is an important contribution for the atmospheric sciences community. This new archive compiles most of the soundings reports available in several former datasets. It is the most comprehensive quality assured and largest radiosonde dataset (Durre et al., 2006).

Correspondence to: L. Gimeno
(l.gimeno@uvigo.es)

One of the authors of this paper worked extensively with the aerological information collected at Camagüey Meteorological Centre (CMC), Cuba (WMO station number 78 355) for the period 1981 to 1988. Upon checking the availability of sounding reports for the station 78 355 in IGRA we realized that at least for the cited period, there is a lack of around 40% of the soundings in the current IGRA dataset. This situation prompted the present case study, having into account that here we deal with the two worse assumptions of missing data as previously found by Kidson and Trenberth (1988), random sampling and missing blocks of data. Our main aim was to evaluate the impact of missing data on mean values of altitude and temperature at mandatory levels for a particular station for an eight years period. We also evaluated the impact of missing data on mean pressure and temperature at the multiple tropopause levels.

2 Description of former studies

In the late eighties a lidar station for stratospheric aerosols studies was installed at CMC (21.4° N, 77.9° W), because of the scientific cooperation between Cuba and the former Soviet Union. The establishment of this facility required the determination of the local mean sounding as well as the mean tropopause features for the processing and analysis of its measurements. Such studies were conducted using aerological information collected by the radio-sounding station located at CMC (Antuña et al., 1991[1]; Antuña et al., 1992[2]). This station has been part of the international exchange of meteorological information reporting to the regional centre in WMO Region IV.

The former two studies used the original 1563 handwritten sounding reports at 1200Z, from 1981 to 1988, archived at the CMC. During that period a Russian made Meteorit 2 tracking system operated at CMC, using RKZ-5 and MARS-2-2 radiosondes (Gaffen, 1993). Processing was conducted semi-manually with the help of specifically designed calculation rules (CAO, 1973).

In the first study (Antuña et al., 1991[1]), the original sounding reports were digitized to derive the mean temperature and altitude values at mandatory levels. Then, all the values of each variable were tested for plausibility, using realistic limits. Values out of such limits were discarded after verification in the written reports for human errors. In a next step a preliminary statistics was calculated. Another quality control was conducted using the mean and standard deviation (σ) for each variable at the mandatory levels. Values out of the limits defined by the variable mean $\pm 2\sigma$ were also verified to avoid human errors. Finally the mean and σ for both variables at every mandatory level were calculated. The study provided a statistical model of the vertical distribution of atmospheric variables at Camagüey, improving the lidar processing and analysis.

In the second study (Antuña et al., 1992[2]) the original sounding reports were reviewed to extract pressure and temperature at each one of the multiple tropopause levels reported. The mean values of both variables at the multiple tropopause levels were quality controlled using a similar procedure than the one described above. In the final step the mean and σ for both variables at every multiple tropopause level were calculated. The results showed a similar seasonal behaviour for the first tropopause variables than other tropical stations in the Pacific Ocean and India. An interesting feature was the fact the first tropopause altitude at Camagüey was higher in winter (lower pressure) and lower in summer (higher pressure) than the reported altitudes (and pressures) from the other tropical stations. Several causes were considered as possible responsible of those results. Among them were that the time periods of the other tropical stations correspond to earlier years than at the WMO station 78 355. Also the different radiosounding instrumentation, the particular physical-geographical conditions of the Caribbean and potential global climate changes were considered as possible causes of that feature.

3 Comparisons: IGRA vs. CMC data

3.1 Available data at IGRA dataset

When reviewing the IGRA dataset for the station 78 355, the first thing we noticed was the complete lack of data for 1981, while a total of 248 soundings were conducted at 1200Z at that site along this year. A total of 886 soundings at 1200Z from 1982 to 1988 are reported in IGRA, representing almost the 60% of the total soundings conducted at the 78 355 station between 1981 and 1988. This is a particular example of incompleteness of IGRA. It was caused mainly by to the lack of transmission and probably difficulties with the encoding-decoding of soundings. The disagreement can also be higher for reported tropopause values, because no all the tropopause values passed the quality control conducted for the generation of the IGRA dataset during this period (I. Durre, personal communication, 2006).

The lack of data in the IGRA dataset for the WMO station 78 355 for the year 1981 is the particular case of missing data caused by missing blocks of data while the lack of soundings between 1982 and 1988 represents the particular case of missing data due to random sampling. Both cases of missing data are considered the two which most affects the

[1] Antuña, J. C., Marin, D., and Aroche, R.: Statistical model of some atmospheric parameters at Camagüey Meteorological Site, (in Spanish, Unpublished Manuscript, available from the Camagüey Meteorological Centre Library), 24 pp., 1991.

[2] Antuña, J. C., Aroche, R., Guaty, A., Morales, C., and Perez, P.: Behaviour of the tropopause at the Camagüey Meteorological Site. Part I: Aerological variables, (in Spanish, Unpublished Manuscript, available from the Camagüey Meteorological Centre Library), 32 pp., 1992.

Table 1. Statistics for the height and temperature at selected mandatory levels from reports at Camagüey radio-sounding station (WMO number 78 355) derived from original handwritten reports and from the current IGRA dataset.

	Antuña et al. (1991)[1]					IGRA Dataset					
	Altitude (m)		T (°C)			Altitude (m)			T (°C)		
Pr(hPa)	Mean	σ	Mean	σ	Cases	Mean	σ	Cases	Mean	σ	Cases
1000	141	16	22.4	1.7	1664	147	15	552	22.0	2.7	587
850	1544	20	16.1	1.9	1664	1545	20	666	15.9	2.5	661
700	3176	27	8.2	1.9	1647	3176	28	656	8.1	2.1	654
500	5876	48	−7.2	2.1	1570	5878	40	608	−7.3	2.1	603
400	7581	56	−18.3	2.3	1500	7583	51	554	−18.5	2.2	542
300	9664	81	−33.4	2.5	1423	9665	63	519	−33.5	2.6	512
250	10925	90	−43.0	2.7	1337	10 923	68	473	−43.0	2.7	462
200	12 392	82	−54.2	2.7	1230	12 386	82	424	−54.2	2.8	416
100	16 595	102	−74.6	3.3	919	16 595	99	293	−74.5	3.7	298
70	18 680	120	−71.8	3.5	777	18 680	103	252	−72.2	5.1	254
50	20 692	129	−64.2	3.1	613	20 697	127	189	−64.4	4.2	193
30	23 871	176	−54.9	2.7	428	23 855	177	125	−55.5	3.6	125
20	26 516	173	−48.4	3.0	246	26 498	220	57	−48.9	3.7	56

homogeneity of radiosonde datasets (Kidson and Trenberth, 1988).

Using the current available information from IGRA for the 78 355 station at 1200Z between 1982 and 1988 we derived the mean values of height and temperature as well as its σ. In order to compare, we selected the same mandatory levels that were used in the former study by Antuña et al. (1991)[1]. Also the statistics for pressure and temperature at the multiple tropopause levels reported in the current IGRA 78 355 dataset were calculated. Soundings conducted at 1200Z were used for such purpose, in the same way that it was done in the previous referred study (Antuña et al., 1992[2]).

Moreover, T-Students tests were used to estimate the statistical significance of the difference of the means from both datasets (results from the original handwritten dataset as the population and the actual ones contained in IGRA as the sample).

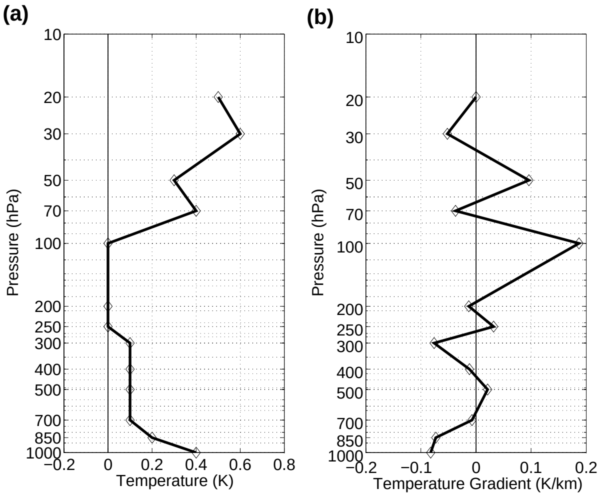

Fig. 1. Profiles of the differences between the population and the sample. **(a)** Differences of the mean temperatures, **(b)** Differences of the vertical temperature gradient.

3.2 Mean sounding

Table 1 shows the height and temperature means and standard deviations for the selected mandatory levels together with the number of cases at each level. The available altitude and temperature data in the current IGRA dataset ranges between 40% at lower levels and 20% at the higher ones in comparison with the total of soundings conducted at the station 78 355 between 1981 and 1988.

A comparison of the altitude mean and σ between the IGRA dataset (sample) and the original handwritten dataset (population) show light differences. We find a similar feature for the temperature. The vertical profiles of the differences of the mean values of temperature and mean temperature gradient between the population and the sample are shown in Fig. 1. In Fig. 1a the differences of the mean temperature show values of 0.4 K at 1000 hPa, decreasing to the middle troposphere. Between 700 and 100 hPa the differences are 0.1 K, with no differences in the layer from 250 to 100 hPa, around the tropopause. Then the mean temperature differences increase above the tropopause reaching its larger values on the top of the profiles. Such relative higher differences values on top of the sounding are associated with the 80% of the missing data in that layers in comparison with the 60% in the lower ones. In general the magnitudes of the differences are below or equal to 0.6 K, with the mean temperatures showing higher values in the population than in the

Table 2. Statistics for the multiple tropopause reports at Camagüey radio-sounding station (WMO number 78 355) derived from original handwritten reports and from the current IGRA dataset.

		Antuña et al. (1992)[2]		IGRA, 78 355	
		Pr (hPa)	T(°C)	Pr (hPa)	T(°C)
1st Tropopause	Mean	112	−74.2	124	−74,3
	Standard Dev.	16	3.8	73	7.7
	Cases	760		314	
2nd Tropopause	Mean	87	−74.5	81	−73.9
	Standard Dev.	15	3.2	20	6,4
	Cases	218		54	
3rd Tropopause	Mean	83	−73.0	67	−69,8
	Standard Dev.	12	3.3	20	10.7
	Cases	25		8	

sample. That is a notable result considering the amount of missing data.

Figure 1b shows the differences of the mean temperature gradient between the population and the sample. In general the absolute value of the differences is lower than 0.1 K/km, except at 100 hPa, where it reaches almost 0.2 K/km. Highest relative values of the temperature gradient differences are located on top of the sounding as in the case of temperature differences. The reason is the same, the increase of missing data from bottom to top. Both negative and positive differences are present. In the case of the altitudes, the differences between the population and the sample are below 20 m, with both positive and negative relative values.

The test of the statistical significance of the differences between the means showed no significant differences at all the levels, both for temperature and altitude. The significance of the test ranged between 99.51 and 99.66% for altitude and between 99.50 and 99.57% for temperature.

The fact that there are no statistical differences between the means of both datasets, although there are between 60 and 80% of missing data at different mandatory levels in the sample is relevant. The explanation is likely related to the normal distribution of the variables together with the fact that the main cause of the missing values (transmission difficulties) has a random origin.

The former results are an indication that although the high amount of missing data in the IGRA dataset for the WMO station 78 355, the existing dataset is statistically representative of the complete dataset. It is also an indicative of the high quality of the IGRA dataset for this particular station.

3.3 Tropopause means features

The 48.6% of the soundings reported tropopauses in the population and only 35.4% in the sample. In the same way, for the population 28.6% and 3.2% reported double and triple tropopauses, while 17.1% and 2.5% reported double and triple tropopauses in the sample.

Although there is a noticeable difference in the percent of multiple tropopause reports, the comparison among the mean values of pressure and temperature for the population and the sample is encouraging.

Table 2 contains the statistics for the pressure and temperature of the multiple tropopauses derived both from the population and the sample. The mean temperature values for the first, second and third tropopauses in the sample are inside the interval defined by the mean temperature value plus minus the standard deviation of the population. Something similar happens with the pressure, except for the third tropopause. But at least the mean pressure value for the third tropopause in the IGRA dataset is inside the interval defined by the mean pressure value plus minus two standard deviation of the population. In that particular case the very few cases of third tropopauses in the IGRA dataset may introduce certain bias in the mean pressure value.

Despite the IGRA dataset contains only 60% of the soundings conducted at the station 78 355 for the period 1981 to 1988, there are generally no significant differences between the mean values for the pressure and temperature of the multiple tropopauses derived from the population and the sample.

4 Conclusions

We have evaluated the effects of missing sounding reports on the temperature and pressure at mandatory and multiple tropopause levels using a real inhomogeneous dataset, instead of a simulated one. The results of this study demonstrated that, for this station although we are in presence of the two worse cases of missing sounding reports, they do not produce statistical significant changes neither in the mean values for temperature and altitude at mandatory levels nor in the pressure and temperature mean values at the multiple tropopause levels. In the present case study the only noticeable effects of the missing soundings are the slight decrease of the mean temperatures in the incomplete dataset and the decrease in the percent of tropopause reports.

The results of this study provide an indication of the statistically representatively and high quality of the data currently stored in the IGRA dataset.

Acknowledgements. Authors thank the NOAA for providing the IGRA dataset as well as I. Durre for her support in the issues related to the dataset. J. C. Antuña acknowledges the support from the "Programa de Promoción de la Investigación Xeral de la Xunta de Galicia" and the Cuban National Climate Program Grant #01303177. This work has been partially funded by the Spanish Science Ministry through TROJET project.

Topical Editor F. D'Andrea thanks W. Lahoz for his help in evaluating this paper.

References

CAO: Handbook of hydrometeorological stations and posts. Issue 4: Upper –air observations, Gydrometeoizdat, Leningrad, 256 pp., 1973.

Durre, I., Vose, R. S., and Wuertz, D. B.: Overview of the Integrated Global Radiosonde Archive, J. Climate, 19, 53–68, 2006.

Gaffen, D. J.: Historical changes in radiosonde instruments and practices, WMO/TD, no. 541, 123 pp., 1993.

Gaffen, D. J., Sargent, M. A., Habermann, R. E., and Lanzante, J. R.: Sensitivity of Tropospheric and Stratospheric Temperature Trends to Radiosonde Data Quality, J. Climate, 13, 1776–1796, 2000.

WMO: Analysis of data exchange problems in global atmospheric and hydrological networks, GCOS-96, WMO/TD-No. 1255, 49 pp., 2005.

Kidson, J. W. and Trenberth, K. E.: Effects of missing data on estimates of monthly mean general circulation statistics, J. Climate, 1, 1261–1275, 1988.

Seidel, D. J. and Free, M.: Measurement Requirements for Climate Monitoring of Upper-Air Temperature Derived from Reanalysis Data, J. Climate, 19, 854–870, 2006.