WORLD METEOROLOGICAL ORGANIZATION

INSTRUMENTS AND OBSERVING METHODS

REPORT No. 95

GLOBAL CRITERIA FOR TRACING THE IMPROVEMENTS OF RADIOSONDES OVER THE LAST DECADES

P. Jeannet *(MeteoSwiss, Switzerland),*
C. Bower *(NOAA, USA),*
B. Calpini *(MeteoSwiss, Switzerland)*

WMO/TD-No. 1433

2008

NOTE

FOREWORD

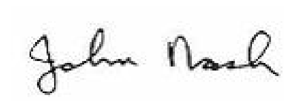
This report represents the deliverable of the action initiated at the First Session of the Expert Team on Upper-Air Systems Intercomparisons (ET-UASI), Geneva, Switzerland, 2004: „To elaborate global criteria for tracing the improvements, based on previous intercomparisons and recent radiosondes development“.

In the report, different approaches are identified with the objective to define these criteria, and the following method was selected as the most appropriate one: Time evolution of selected statistical parameters reported in the different IOM international radiosonde comparisons organized between 1984 and 2005. A set of performance criteria is proposed together with a way how to extract the corresponding values from the IOM reports and to present the results.

The WMO international radiosonde intercomparisons have played a key role in the improvement of these measurements. These results have been used by manufactures with the goal to improve their products. The report enables to monitor these technology improvements.

I wish to express my sincere gratitude and that of CIMO to the authors of this report, namely C. Bower (NOAA, USA), B. Calpini and P. Jeannet (MeteoSwiss, Switzerland) for their remarkable work done in elaborating this report.

I am confident that WMO Members as well as radiosonde designers and manufactures will find this report very useful. It will contribute to further improvement of upper-air measurements.



President
Commission for Instruments
and Methods of Observation

Global Criteria for Tracing the Improvements of Radiosondes over the Last Decades

Pierre Jeannet, Carl Bower, and Bertrand Calpini

TABLE OF CONTENTS

1 Introduction

The accuracy of radiosondes and the homogeneity of the world upper air network was already a matter of concern more than 50 years ago, from the time of the first international radiosonde comparisons. At that time, the former CIMO sub-commission for experimental aerology had difficulty in determining the relative performance of the radiosonde types. During the last decades, the CIMO expert team on upper-air systems intercomparisons (ET on UASI) organized several new international radiosonde comparisons and has been actively promoting the improvements of radiosondes. Its reports on the different international radiosonde comparisons documented the performance of widely used radiosonde systems. Furthermore, they pointed out needed sensor improvements to fulfil the growing requirements of the meteorological community (weather forecast and long-term atmospheric monitoring).

The radiosonde systems compared in the international comparisons of the previous decades have been replaced by new ones. Major technology changes occurred, particularly driven by new sensor technologies, powerful computers, microprocessors, and GPS technology. Within the framework of its reporting on the performance of upper-air observations, the CIMO expert team decided in 2004 to address the following task: "Develop performance measures to demonstrate the continuous improvement in the quality of upper-air observations". The required action was to "elaborate global criteria for tracing the improvements, based on previous intercomparisons and recent radiosonde development, and including remote sensing". The present report constitutes the required deliverable for this action. Upper-air remote sensing is not addressed in this report[1]. As explained in the next sections, the report focuses on the radiosonde improvements in the last two decades.

Simple criteria are proposed which allow tracing the improvements in the quality of radiosonde observations. Some of the criteria are applied in order to determine if they fit the objectives. Finally we present results.

Ideally criteria for tracing radiosonde improvements should be based on data already available in the previous IOM reports. They also should be considered in further CIMO field campaigns as well as national or international intercomparisons to ensure the required continuous improvement tracing.

2 Method

Tracing the instrumental improvements requires an appropriate method and specific criteria. Different methods could be followed for tracing the radiosonde performance improvements, by either: (1) using the previous radiosonde comparison reports, (2) comparing radiosonde measurements with model values, (3) establishing a questionnaire to the attention of NMHSs, or (4) comprehensive review of the open literature. The first one has been selected for the present report and is introduced in section 2.1. The others are briefly commented on in section 2.2.

[1] A similar task is also defined for surface meteorological measurements.

2.1 WMO international radiosonde comparisons

The first international WMO comparisons had been conducted in 1950 and 1956 (World Comparisons of Radiosondes, Payerne, Switzerland), in 1968 (WMO Reference Radiosonde Comparison, Tateno, Japan), and in 1979 (Low-level Intercomparison Experiment, USA). In 1950, six radiosonde types were compared using a first version of the procedure that was used in the WMO campaigns since the 1980s. At that time, the results were limited to the first 16 km (100 hPa). The differences between the data given by any two radiosondes on the same flight were in 95% of the cases for respectively pressure, temperature and relative humidity not greater than 15 hPa, 2 ^{o}C, and 20% up to 700 hPa, and 1% of the height of a given pressure level. In 1956, 14 radiosondes were compared and some soundings reached 10 hPa. A brief excerpt of the results is reproduced in Table 1. One can note the huge pressure measurement problems. The poor accuracy and reproducibility of most measurements at that time strongly hampered the conclusions of the expert group: "*It is, however, impossible to present definitive conclusions, owing to the large random errors which tend to mask the systematic differences, and to the fact that the full programme could not be completed…. The accuracy of the radiosondes is not yet satisfactory*". The CIMO working group recommended the re-establishment of a working group, due to the large number of outstanding problems connected with the instrumental techniques used for the investigation of the atmosphere. Due to the technology used at that time, their results can only be compared in the troposphere with those of the sounding systems emerging in the early 1980s.

Table 1. Excerpt of the results of the international radiosonde comparisons of 1950 and 1956 at Payerne (maximum differences (95%) between the data given by any two sondes on the same flight were not outside those given in the table).

	Pressure (hPa)		**Temperature (oC)**		**Geopotential height (gpm)**		
1950	**p ≥ 300**	**p <300**	**p ≥ 300**	**p <300**	**500 hPa**	**300 hPa**	**200 hPa**
Bias	5	15	1.3	2.2	20	40	60
Std. Dev.	10	16	1.2	1.7			
1956	**p = 500**	**p = 50**	**p = 500**	**p = 50**	**500 hPa**	**100 hPa**	**50 hPa**
Max. Bias	6.1	3.5	0.9	2.3			

In 1984, when the WMO Phase I international radiosonde comparison took place, the participating radiosondes were able to reach the 10 hPa pressure level (31 km) slightly more than 50 percent of the time. The planning of these experiments had been carefully defined by the new CIMO expert group, as well as the data processing and the statistical methods [Hooper, 1983)]. Consequently, their reports are organized in a similar way and most of the key statistical results can be found in all reports. Where needed, the intercomparison guidelines have been improved during successive Phases.

The intercomparisons with a CIMO report took place in 1984 in the UK (Phase I), 1985 in the USA (Phase II), 1989 in the former USSR (Phase III), 1993 in Japan (Phase IV), 2001 in Brazil (Phase V), and 2005 at Mauritius (Phase VI). A few other special international campaigns also took place in this time period, such as in the UK (PREFERS, 1992), and in Russia and the USA (humidity sensors, 1995; see Balagurov et al., 2006). These provide additional results to those of the Phase I to VI intercomparisons.

The following Figures 1 to 4 briefly illustrate some of the WMO Radiosonde Comparisons, from 1956 to 2005. Installing the ground systems, preparing the radiosondes, tracking them and getting the results are nowadays easy tasks compared to the past.

Figure 1. WMO World Comparison of Radiosondes at Payerne, Switzerland, 1956.

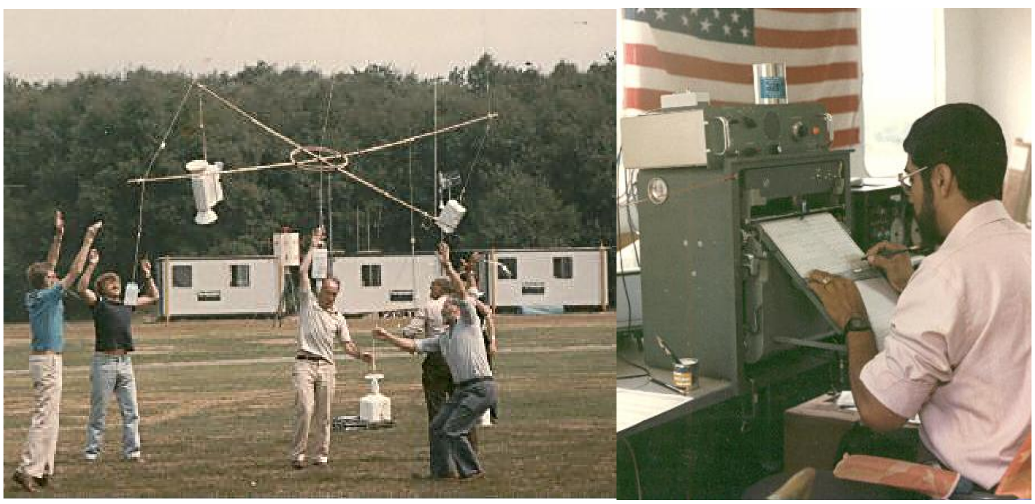


Figure 2. WMO Radiosonde Comparison (Phase I) at Beaufort Park, U.K., 1984.

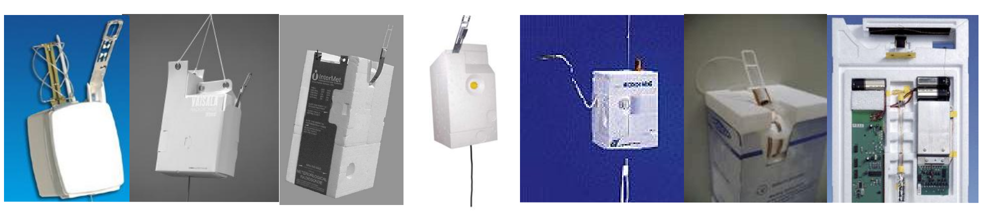


Figure 3. High quality radiosondes involved in the WMO Radiosonde Comparison at Vacaos, Mauritius, 2005.

The statistical parameters (systematic biases, standard deviations, etc) based on differences between the measurements obtained with different types of radiosondes for simultaneous measurements represent a valuable tool for comparison over the last two decades. Each of the Phase I to VI intercomparisons used “link radiosondes” in order to define one reference value (working reference) for comparing it with the measurements of all participating radiosondes. However, the time-continuity of the link radiosondes could not be guaranteed. As an example, the mean differences in the temperature and pressure observations of link radiosondes in Phase III appeared to be larger than in previous Phases (see Tables 5.3-4 in the Phase III report). But later on, in Phase IV, the link radiosondes were found to be compatible with those of the previous phases. The Accurate Temperature Measuring (ATM) three-thermistor radiosondes, which were considered to provide reference quality measurements for the air temperature, participated in the WMO radiosonde comparison for the first time in 1993 (Phase IV). Although the link sonde approach has shortcomings, it allows a straightforward use of the IOM reports.

The comparison of simultaneous measurements delivered by the international radiosonde comparisons over the last decades is used in the present report as the most straightforward method for getting information on the improvement of radiosonde instruments and data processing procedures.

Figure 4. Radiosondes launch during the WMO Radiosonde Comparison at Vacaos, Mauritius, 2005.

2.2 Other possible methods

The following methods have been briefly explored as possible complementary or alternative ways to fulfil our objective. They are summarized here below, but have been evaluated not as promising as the first one, or going beyond the defined objective of this study.

<u>Same as 2.1, but at standard pressure levels</u>

A slightly different method to 2.1 is presented in the IOM radiosonde comparison reports. The comparisons are provided at standard pressure levels, which are useful to users of radiosonde data, especially the numerical weather prediction users. When comparing radiosondes at pressure levels, without considering the measurement simultaneity, the differences between radiosondes are smaller than in the previous case. This analysis has partially been conducted so far. Elms [2003, Table 10] provides a comparison of the estimated reproducibility of geopotential heights between 500 and 10 hPa for the major radiosonde types obtained from the WMO intercomparisons. Nevertheless this approach is based on the assumption that the radiosonde pressure measurements are exact and does not take into account all sources of measurement errors. It will not be followed here.

Time evolution of the comparison between measured and modelled geopotential heights

In this type of analysis the comparison is made between radiosonde measurements and the first guess of NWP models, the latter being considered as a relatively stable working reference [Oakley, 1998]. The CIMO Rapporteur on Radiosonde Compatibility Monitoring produced reports, which cover the years 1988, 1990-1992, 1995-1997, 1998-2001, as well as Excel tables for the very last years [Kitchen, 1989; Oakley, 1993; Oakley, 1998; Elms, 2003]. The 2004 and 2005 statistics can be found on the CIMO/IMOP website for upper-air monitoring statistics and comparisons with the ECMWF First Guess (http://www.wmo.int/web/www/IMOP/monitoring.html) However, the numerical models underwent significant changes in the last decades. Thus criteria based on the comparison between measurement and model results will inherently have the drawback of adding errors induced by the method of observations together with errors induced by the model results. This method nevertheless relies on well defined parameters for these differences: e.g. the bias and standard deviations of the geopotential altitude of the 100 hPa level as well as the height increment from 100 to 30 hPa and its standard deviation, at 00 and 12 UTC.

This analysis gives valuable information about the radiosondes' performances as they are operated by the different NMHSs (quality of operational upper air observations). Oakley compared the monitoring results for the main radiosonde types in section 3.1.3 of his 1998 report [Oakley, 1998]: "In general about 70 percent of the radiosonde stations in the period 1995 to 1997 were producing observations within the suggested quality limits. This compared with 65 percent in 1992". In the 2003 Elms report [Elms, 2003], section 6.3 stated that "The majority of radiosonde types were assessed as of "good" overall performance. However, radiosonde systems with larger systematic errors or poorer reproducibility have shown little improvement since the last Rapporteur's report", and section 6.4 further concluded that "the overall measurement quality of the radiosonde network has continued to improve".

Note that the next IOM reports on the compatibility of radiosonde geopotential measurements should allow a continuous analysis over more than 10 years. If the quarterly statistical parameters can be worked out for a longer period (e.g. since 1988) or recalculated on the basis of ERA40, they would represent a valuable complement to the results obtained from the radiosonde intercomparisons.

Questionnaire to the NHMSs

A detailed questionnaire could be elaborated, in order to highlight the time evolution for radiosonde measurements worldwide. For example, statistical values such as the percent of 00 and 12 UTC soundings reaching the 100, 30 and 10 hPa levels, as well as the temperature and geopotential altitude differences between the 00 and 12 UTC observations at different standard levels, could be helpful for tracing improvements over the years. However, this would require an important contribution from the NMHSs. Furthermore, height improvements are due to other factors than radiosonde improvements, such as balloon improvements, size selection, radiosonde system improvements, lighter weight radiosondes, or national needs.

Compilation of the articles published in the scientific journals

Information from the open literature may be another valuable source of information about global criteria for tracing the improvement of radiosonde performances over the years. A number of intercomparison campaigns and reviews have already been published in the framework of research projects such as ALPEX, TOGA, AWEX, and others. Phase II report gives a list of the radiosondes comparisons that were organized before the 1980s. Gaffen [1993] presents a bibliography on radiosonde performance and comparisons up to 1983. The Mauritius final report gives references to recent publications. For example, the absolute accuracy of water vapour measurements from six operational radiosonde types was studied by Miloshevich and al. [2006]. These research oriented radiosonde intercomparisons provide a valuable accuracy status for a specific time period. However, few include a historical perspective, which has a high priority in the framework of long-term climate observing programmes and trend analyses projects, (see http://www.wmo.int/web/gcos/gcoshome.html, e. g. the GCOS reference upper-air network (WMO/TD No. 1379) and herein references). Since long-term radiosonde records are affected by numerous instrumental and operational changes, time series homogenization is a pre-requisite to trend analyses. The homogeneity adjustments provided by such studies reduce temporally constant biases, i. e. artificial systematic breaks. Recently, Haimberger [2007] homogenized a global set of radiosonde temperature time series using the ECMWF reanalysis (consistent model). He provided on the web the time series of its temperature adjustments for the time period back to 1958 (http://www.univie.ac.at/theoret-met/research/raobcore/). Their statistical analysis would allow tracing the temperature improvements over nearly 50 years, enabling subsets according to different criteria such as country and sounding system. This method would represent a valuable extension of the method followed in this report. It would extend the results of intensive field campaigns organized with a somewhat regular time schedule with results derived from the operational radiosonde network. However, most radiosonde long-term record homogenization projects are limited to one parameter, often temperature. In addition, they do not fully avoid the influences of the model improvements and hardly separate the role played by the different parameters.

3 Quality performance criteria

Absolute references are not or only partially available for upper air measurements even in radiosonde comparisons. The “references” for calculating biases and standard deviations are based on the data obtained using as working references the “link radiosondes” or “high quality radiosondes”. Since the “link radiosondes” were also in some cases showing intrinsic problems, a more general approach would be to use all the data measured under similar conditions in a given field campaign, and find the average and the standard deviation of the parameter under investigation. This approach where the different sondes are directly “inter-compared”, without using link sondes, has been briefly checked. The results are similar to those of the “link radiosondes” approach. As not all Phase reports contain the needed original information, this more general approach is not further applied.

In this section and in the following ones, we will restrict our analysis on the WMO international radiosonde comparisons as defined in 2.1. Our method for radiosonde performance improvement tracing will be based on specific criteria that are defined using the link radiosondes.

Candidate criteria are presented in Table 2. They rely on comparison of simultaneous (so-called time-paired) measurements. Our selection is made with the objective of a small number of criteria in order to trace the improvement of radiosondes over the years with a straightforward data analysis. These criteria are based on mean differences and standard deviations of differences, where the differences are taken as the measurements of the referred sonde minus the simultaneous measurements of the link sonde. Criteria related to mean differences between sondes (bias) correspond to systematic measurement errors. They are more sensitive to radiosonde measurement problems than criteria based on the standard deviation of the differences between radiosondes. They allow determining the origins of radiosonde deficiencies. Standard deviation of the differences between radiosondes complements the information provided by the mean difference between radiosondes. If the standard deviation is smaller than the bias, it allows assigning measurement errors to a systematic problem in the radiosonde design and/or in the data processing. The standard deviation may also help identify error sources in the radiosonde reproducibility. The daytime soundings are generally defined as those around local noon, but can extend to several hours around noon. The night-time soundings correspond to the full darkness time.

Table 2. Candidate criteria for tracing the improvements of radiosondes.

Criteria	Remarks
Temperature difference around*) 10 or 30hPa, @ night/day time	The 10 hPa level is the highest standard level in the TEMP messages. Reaching a high quality standard around this level is a demanding task. Temperature errors are different during night and daytime (noon). A higher data sample is found around 30 hPa than around 10 hPa, particularly in the first Phases.
Standard deviation of the temperature differences around*) 10 or 30 hPa, @ night/day time	
Geopotential difference around*) 10 or 30 hPa, @ night/day time	Geopotential measurements from a radiosonde accumulate errors from other parameters (temperature, pressure, etc.) between surface and this level. Recent advances in GPS positioning have brought major upgrade on this criteria.
Standard deviation of the geopotential differences around*) 10 or 30 hPa, @ night/day time	
Geopotential difference around*) 100 hPa, @ night/day time	The 100 hPa level is the primary level used in the quality control of upper air data based on comparison with numerical model outputs.
Standard deviation of the geopotential differences around*) 100 hPa, @ night/day time	
Pressure difference around*) 100 hPa, @ night/day time	Pressure criteria are very sensitive to the pressure range. The 100 hPa level is proposed as it is a primary pressure level that is used in the operational radiosonde quality monitoring.
Standard deviation of the pressure differences around*) 100 hPa, @ night/day time	
Systematic relative humidity difference in the temperature range between -35 and -45 ^{o}C (only tropospheric values)	Two temperature ranges could be an alternative to the unique range proposed in the left column: between -20 and -30 C as well as between -40 and -50 C. For modern humidity sensors temperature ranges below -50 C and/or in the stratosphere can help documenting their high performance. In order to account for the different humidity ranges, a further selection in 3 classes should be made: e.g. below 25%, 25% – 75%, above 75% RH.
Standard deviation of the humidity differences in the temperature range between -35 and -45 ^{o}C (only tropospheric values)	

*) These approximate pressure levels (“around”) are defined as follow. The first WMO radiosonde comparisons defined 15 pressure categories in the comparison of simultaneous measurements. The 10 hPa category considered all measurements between 8.4 and 11.9 hPa, as defined by the link sondes. The 30 hPa category was more exactly centred at 32 hPa (24.5 – 41.5). The 100 hPa category range was 84 – 119 hPa. This ensured that the statistics were relying on a sufficient number of simultaneous measurements. In the more recent radiosonde comparisons, 2 km wide altitude categories were introduced instead of the previous ones. The altitude category that included the referred pressure level was then used.

4 Illustrative check of the method

As an illustrative check, the systematic temperature differences around 10 hPa at night-time are reported in Annexe I, as well as in Figure 5 below. The two Tables in the Annexe contain bias values extracted from Figures or Tables of the different IOM reports, without any additional processing. Each value is given with the Figure or Table number as well as report number from where it was taken. Each symbol in Figure 5 corresponds exactly to a value reported in the Annexe (bias of one sonde type during one radiosonde comparison). Almost 30 radiosonde types have been intercompared at least once over all comparisons. The radiosondes appear anonymously in this Figure. The aim is not to find out the “best radiosonde”, but to assess the general improvement in the quality of upper-air observations occurred over the last 20 years. The horizontal axis is a time axis covering the last 25 years. On the vertical axis, the span of the bias values is more important than their exact positions in relation to the zero point, as the reference is a relative one. In every comparison an outlier point may strongly increase the span.

Basic statistical parameters can be added to the individual results. Figure 5 contains the envelope of the extreme values (maximum, minimum), the maximum span (difference between maximum and minimum), as well as the standard deviations of the biases. Other statistical parameters could be added, such as the average biases and the mean absolute biases. Nevertheless, one should be aware that all statistical parameters in this study have a limited statistical significance, as less than 10 radiosonde types (including different post-processings for the same sonde) were engaged in each comparison.

A first look at Figure 5 shows an improvement of the quality of the night-time temperature measurements over the 20 last years. The envelope of the individual biases narrows with time approximately by a factor of 2, as best shown by the span reduction. The standard deviations of the individual biases improve also by the same factor.

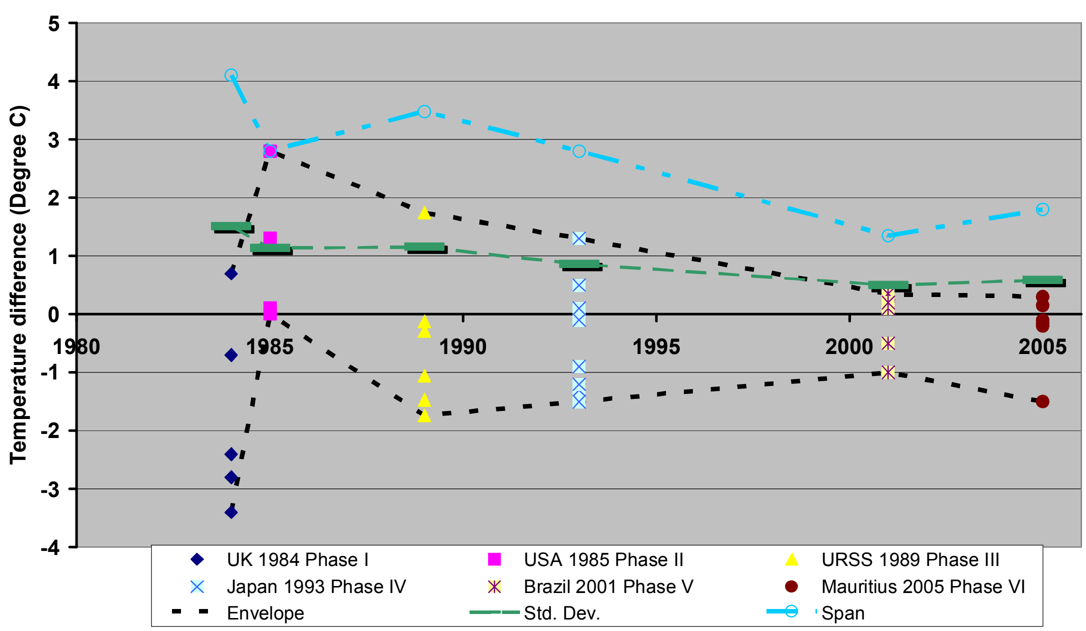


Figure 5. Night-time temperature bias around 10 hPa for the six WMO Radiosonde Comparisons (simultaneous measurements). The two dotted lines represent the envelope of all individual results, which is converted into a span with the dash-dotted line. The

horizontal green bars on the dashed green curve correspond to one standard deviation of the biases of each comparison.

If we consider the first two comparisons as a single one, by taking the minimum of -3.4 °C for Phase I and the maximum of 2.8 °C for Phase II, the span starts with a value of 6.2 °C. Then, the improvement seems to be larger by a factor of 2. However, the combination of Phase I and II night-time temperatures were analyzed in the final joint report of Phases I and II. A span of approximately 4°C was reported in Figure 5.3 of this report, less than the combined value of 6.2°C as found according to the individual reports. An outlier was removed in the final report of Phases I+II. In addition the reference values provided by the link radiosondes were improved. Consequently, this points out the difficulty of properly describing the overall radiosondes performance with a simple statistical parameter, especially for the first international comparisons.

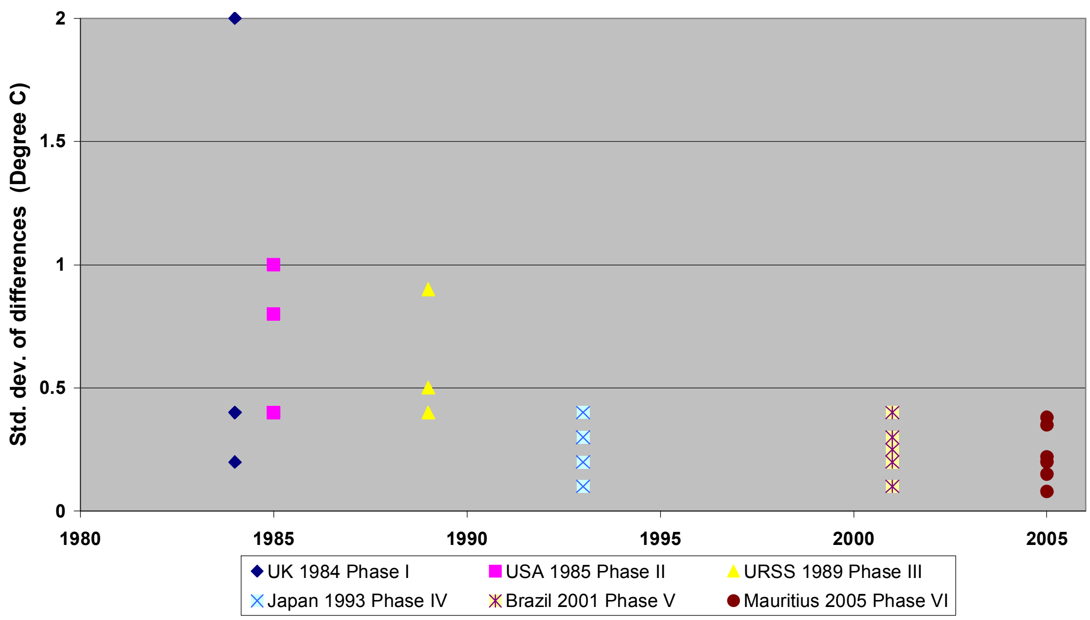


Figure 6. Estimated random errors of the night-time temperatures around 10 hPa for the six WMO Radiosonde Comparisons (simultaneous measurements).

In Figure 5, the value of -1.5°C in the Mauritius experiment was due to the white coating on one radiosondes thermistor. As it was acting as a black body in the infrared, it underwent a strong radiative cooling at this level. This radiosonde thermistor was the only one not using the technology silver/aluminium coating standard. The biases for the five other radiosondes are focused in the range between -0.25 and 0.3 °C. If we do not consider this outlier, the night-temperature measurements around 10 hPa improved by a factor of 8. If we leave out one extreme value in each intercomparison, we get the span time series of 2.8, 1.2, 1.6, 1.8, 0.85, 0.55 °C. This corresponds to a fivefold improvement. Hence, even statistical results of the newest intercomparisons reported in Figure 5 should be weighted by more specific information and have to be accompanied with an interpretation.

The standard deviation or the mean absolute deviation may be appropriate for the small samples of bias values. The time series of the former one is shown with horizontal green bars on the green dashed curve in Figure 5 (1.5, 1.1, 1.2, 0.9, 0.5, 0.6 °C).

According to this statistical parameter a decrease by a factor of 3 might be ascribed to the night-time temperature bias around the 10 hPa level.

According to these two statistical parameters, the improvement of the night time temperatures around the 10 hPa level over the last 20 years lies approximately between a factor of 3 and 5. Worth mentioning is that the night-time conditions are not as challenging as the day-time conditions for radiosonde temperature measurements.

Besides the systematic error, the reproducibility (random error) of the temperature sensors is interesting to analyze. In a similar way as in Figure 5, Figure 6 presents the estimated random errors of the temperatures around 10 hPa for the six WMO Radiosonde Comparisons. In this Figure no statistical parameter is added to the individual results. In the first three campaigns, the values apply both to night and day time conditions, as the estimates of the reproducibility were jointly calculated (see final report of campaigns I and II). For the last campaigns, separate results were available in the reports or in conference presentations. One can conclude that Figure 6 qualitatively illustrates that the random error is another appropriate candidate parameter for our demonstration of performance improvements.

This first illustrative analysis brings valuable results, with the following comments:

- There are no true reference radiosondes, but only link radiosondes or working references. As a consequence, the results of the comparison reports are not absolute. As the performance of the link radiosondes improved during the last decades, the results of the different intercomparisons shifted from strong qualitative to near absolute values.
- The comparison reports for the six Phases present their results in somewhat different forms, although significant efforts have been devoted to their standardization.
- Some comparisons addressed a certain class of parameters and thus do not present results for the other ones.
- There are not only different radiosondes, but different data post processing methods (correction of the radiation error on temperature, data filtering, data smoothing, etc.).
- A single temperature criterion (e.g. 10 hPa bias for night-time conditions) does not allow a comprehensive evaluation for temperature, although it refers to a very challenging altitude level. Other temperature criteria have to be addressed, e.g. for the low stratosphere and the troposphere.
- This type of analysis needs an additional method in order to provide technical and physical explanations, as well as a confirmation of its results. Therefore, in the following sections, some key interpretations and conclusions will be extracted from the comparison reports, and references to available results of process oriented studies will be given.

Despite the above mentioned limitations, the method introduced in this section proves to be valuable for our objectives and will be used in the present report. The following sections present results for geopotential height, temperature, pressure, and relative humidity.

5 Geopotential height

Up to a few years ago, radiosonde geopotential heights were mostly calculated with the hydrostatic equation. This method needs the pressure, temperature and humidity profiles and combines their errors into the calculated geopotential heights. Nowadays, the newer radiosondes use the GPS technology, directly measure geometric height and convert it with a high precision to geopotential height (for mid-latitudes, the difference is approximately 100 m at 30 km altitude). This new technology represents a technology jump. It brings a large improvement for the derived radiosonde pressure and geopotential height measurement accuracy.

Geopotential altitude measurements are highly demanding. A 1 hPa error at 10 hPa corresponds to a 600 m geopotential error. Reporting meteorological parameters at pressure levels (considered as true values) alleviates the errors of radiosonde pressure measurements. However, it only partly reflects the true accuracy of radiosonde measurements. As this report is devoted to the accuracy of radiosonde measurements, it will only compare truly simultaneous measurements performed within the WMO intercomparison experiments.

Figure 7 illustrates these improvements by pointing out in an anonymous manner all radiosonde biases of the geopotential altitude around the 10 hPa level, using the same method as introduced in section 4 for temperature. In Phase II, a high precision radar was used as altitude reference. It demonstrated the real altitude errors the radiosondes were making. In Phase V, for the first time, GPS was introduced for height measurements on two sondes and the comparisons included in Figure 7 use these GPS results (converted to geopotential height) as reference. In Phase VI, there was still sondes measuring height on the basis of pressure sensors, but only the three full GPS radiosondes are documented in Figure 7

The envelope on Figure 7 started with a span of 1000 – 2000 m in the first three comparisons. In Phase IV, better sensors and better calibration curves reduced this span to approximately 500 m. In the Brazil campaign (Phase V), 8 years later, some additional improvement were demonstrated, but fewer balloons reached 10 hPa during this campaign than during the previous ones and the next one and this hampered the comparisons above 30 km. The move to the GPS technology brought a new standard in geopotential measurement accuracy. In 2001, this technology was in an introductory phase. In 2005, it proved having reached its full potential. The GPS radiosondes are nowadays able to measure geopotential altitude at 31 km with an average agreement of about 20 meters (cf. Mauritius report). They reach the same absolute accuracy over their entire altitude range.

Phases I and II reports focused their main conclusions on geopotential altitudes up to 100 hPa. In addition, they gave some information on the higher layers, but fewer radiosondes provided measurements up to the 10 hPa level. Consequently, this hampered the statistical results at this high altitude. The 2005 Mauritius comparison is the first one that definitely demonstrates the accuracy jump brought by the GPS technology.

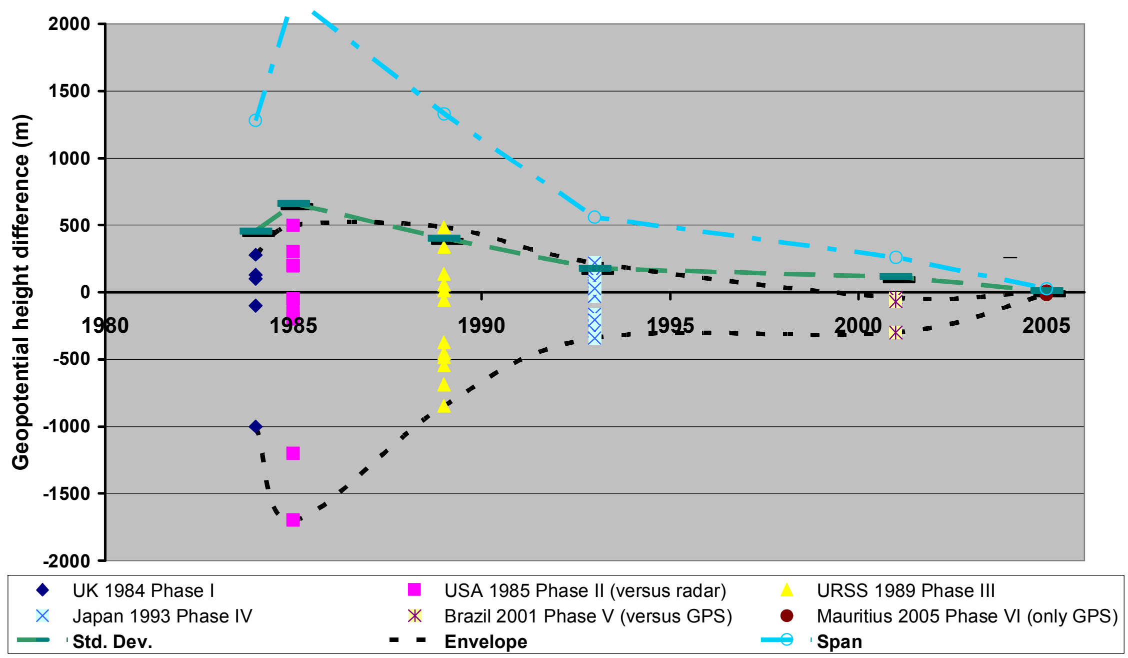


Figure 7. Bias of the geopotential altitude around 10 hPa for the six WMO Radiosonde Comparisons (simultaneous measurements). The two dotted lines represent the envelope of all individual results, which is converted into a span with the dash-dotted line. The horizontal green bars on the dashed green curve correspond to one standard deviation of the biases for each comparison.

Figure 8 complements the results of Figure 7 with the corresponding estimated random errors of the geopotential altitudes around 10 hPa. It demonstrates that the reproducibility of the geopotential measurements improved as much as their accuracy. However, the improvements depict a stepwise behaviour. In the 1980s, radiosonde systems had random errors up to 1000 meters in the altitude range above 30 km. Both the poor accuracy and the poor reproducibility of the radiosondes were responsible for their limited performances in the middle stratosphere. In the 1990s, a noticeable improvement was reached from improvements in the radiosonde technology and the altitude errors have been reduced to 100 – 200 meters near 30 km. In the last years, the GPS technology allowed a new improvement of an order of magnitude in the accuracy and reproducibility of radiosonde geopotential heights. This improvement has led to the additional conclusion of the CIMO upper air group [Nash, 2005]: "*Thus, as long as the temperature sensor provides uninterrupted measurements of good quality, and the measurement locks correctly onto surface pressure and station height at launch, the cost of GPS radiosondes can be reduced by not using a pressure sensor.*" This assumes a high availability of GPS data for each flight to enable geopotential calculations from GPS geometric height data, as well as a high quality thermometer that mitigates both the long and short wave radiation errors (see next section).

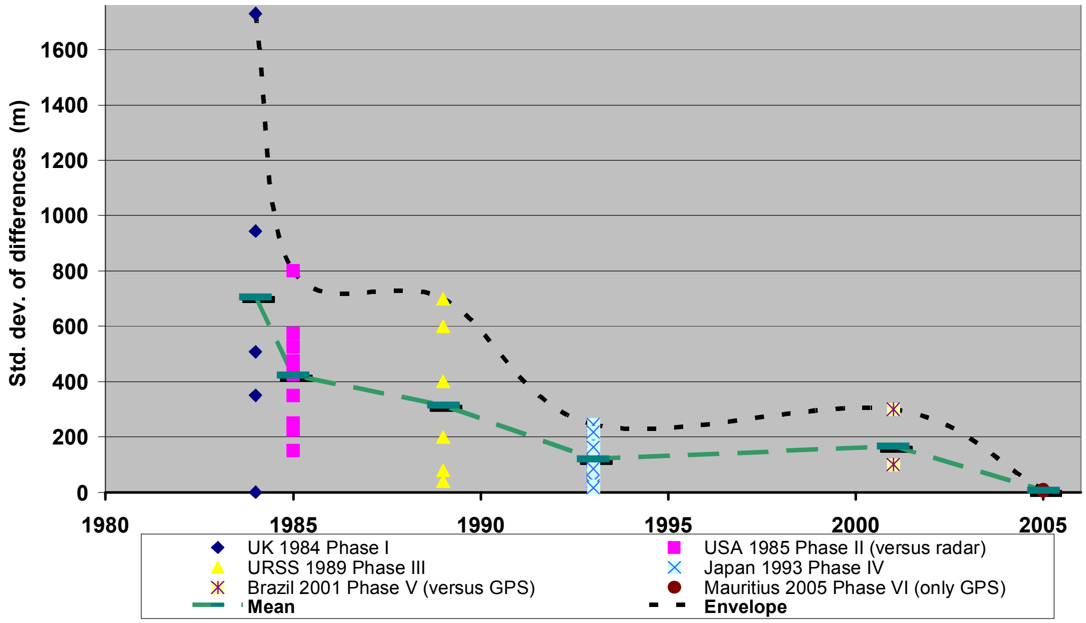


Figure 8. Estimated random errors of the geopotential height measurements around 10 hPa for the six WMO Radiosonde Comparisons (simultaneous measurements). The dotted line represents the envelope of all individual results. The horizontal green bars on the dashed green curve correspond to the mean random errors.

Table 3 illustrates these improvements in another way, by quoting the conclusions of the IOM reports. It complements the present analysis with results for the entire altitude range of the radiosonde measurements. The individual Phase reports focused on their respective results and only marginally compared them with the previous results in order to trace improvement. Besides their specific report, Phases I and II were concluded with a joint report as both Phases were organized in two consecutive years. The Mauritius report brought valuable comparisons with previous campaigns. It is worth noting that the conclusions of these reports often introduced a subset of sondes providing measurements of better quality than the other ones, or excluded one sonde producing outliers either due to systematic sonde deficiencies or to specific problems during the campaign. This reflects some of the difficulties encountered in the overall evaluation of radiosondes performances. In the troposphere, the radiosonde performances were already reasonably good in the 1980s. The largest improvements have been more apparent in the middle stratospheric levels. Table 3 confirms that it is possible to demonstrate a general improvement over several successive radiosonde comparisons. However, no conclusive result can be derived on the basis of only two or three successive radiosonde comparisons, as each campaign had some particularities.

This analysis refers to the comparisons of simultaneous measurements of radiosondes suspended under the same balloon. The alternative method described as the first one in section 2.2 is based on the assumption that the radiosonde pressure measurements are exact and does not take into account all sources of measurement errors. Consequently, it delivers a much better agreement between radiosondes at pre-defined pressure levels, but it does not reflect the true performance of the radiosondes.

Conclusion: Geopotential height measurements around the 10 hPa level underwent a very large improvement over the two last decades. The GPS technology allowed an improvement of an order of magnitude in the quality of radiosonde geopotential heights at 30 km altitude. At Mauritius, all GPS height measurements agreed on average to within ±20 m from the surface to 34 km, whereas the mid-1980 technology provided differences in the order of 500 m at 30 km altitude.

Table 3. Results of the different WMO radiosonde comparisons related to geopotential heights.

Phase	Results
I, 1984	Three radiosondes agree within 30 gpm (geopotential meters) close to the ground, within 50 gpm at 16 km altitude, and within 230 gpm at 30 km. Two radiosondes terminated early in these flights. Variability at 16 km ranges from a standard deviation of ±20 to ±140 gpm.
II, 1985	Agreement between radiosonde derived heights and radar measured heights (reference) was relatively good at levels below 100 hPa but differences became progressively larger above the altitude of the pressure level of 100 hPa (16 km).
I + II (1984-85)	At 100 hPa the geopotential height observations of 3 radiosondes were on average about 50±20 gpm higher than the radar heights. At about 30 hPa the average bias between geopotential height observations of the radiosondes of two countries and the radar heights were always within the range ±100 gpm. At 32 hPa the day-night difference in the comparison of two radiosonde geopotential and radar heights were about 140±70 gpm, and 64±26 gpm respectively.
III, 1989	The simultaneous geopotential height differences between three radiosonde types do not exceed 30-40 gpm below 100 hPa and increase to 100-200 gpm at the 10 hPa level at night and up to 600 gpm (only one radiosonde) at daytime.
IV, 1993	In the nighttimes, differences among the radiosondes were within 150 gpm up to the 10 hPa level, except for two radiosondes. At midday, differences among the radiosondes were within 100 gpm up to the 10 hPa level, except for two radiosondes (due to differences in temperature or pressure).
V, 2001	Pairs of radiosondes showed height differences <100 gpm for the complete sounding. The other pairs showed an increase of the difference with the height, reaching up to 300 gpm.
VI, 2005	All the GPS height measurements agreed on average to within ±20 gpm from the surface to 34 km. The best GPS systems had random errors in height measurements of around 4 gpm or better, with the less precise GPS height systems still less than 15 gpm at most heights. Thus, GPS heights are suitable to replace geopotential from pressure sensors at all heights, i.e. a pressure sensor is no longer a necessity for a best quality radiosonde. The reproducibility of the GPS geopotential heights at 32 km was an order of magnitude better than the reproducibility of the heights from best pressure sensors. The removable of a pressure sensor from a radiosonde is premised on the fact that availability of the GPS has to be sufficient to provide the height and pressure data.

6 Temperature

In the 1980s temperature sensors were still of varying types: bimetal element with mechanical link, coiled tungsten wire (resistive element), thermistor rod, thermistor bead, thermocapacitive bead, etc. The first two types were no more used in the Phase I to VI WMO comparisons and new sensors appeared. Modern temperature sensors are often much smaller than those in use in earlier years and they are placed outside of the radiosonde box.

Due to the different night- and day-time behaviour of the radiosonde temperature sensors from the sunlight and infrared radiation, the radiosonde intercomparisons were systematically performed under both conditions (night and day). As far as possible, they also captured the daily cycle with soundings at different solar elevations.

Figure 9 illustrates these improvements by pointing out in an anonymous manner all individual radiosonde biases around the 10 hPa level. Both midnight and daytime results are presented on the same graph, in order to emphasize the long-term evolution of the envelope of the negative and positive bias. The differences between Figures 5 and 9 are due to the inclusion of daytime results: values only present in Figure 9 belong to daytime.

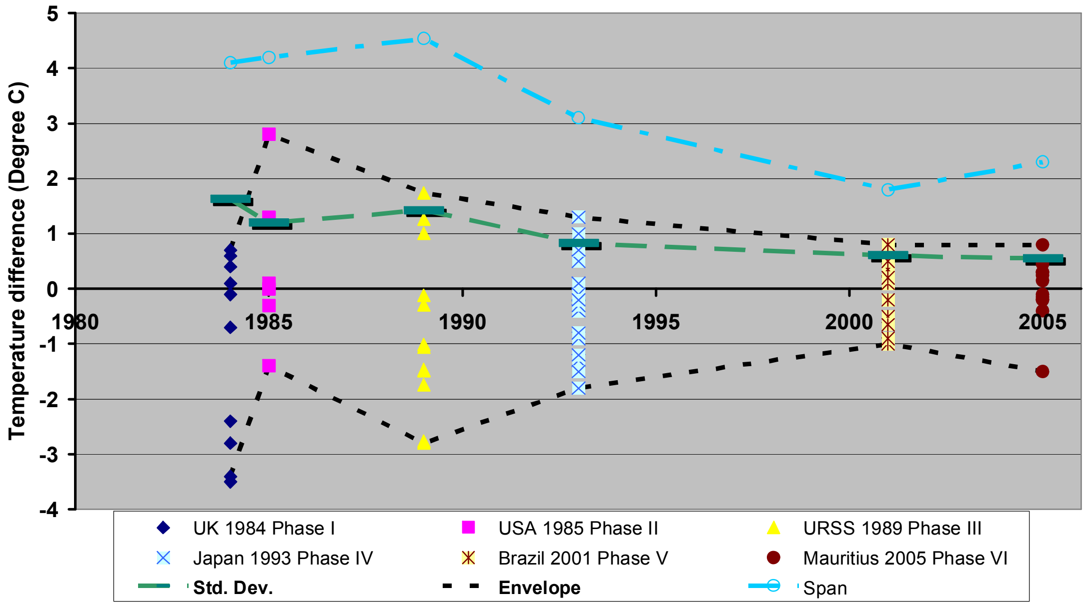


Figure 9. Night and day time temperature bias around 10 hPa for the six WMO Radiosonde Comparisons (simultaneous measurements). The two dotted lines represent the envelope of all individual results, which is converted into a span with the dash-dotted line. The horizontal green bars on the green dashed curve correspond to one standard deviation of the biases of each comparison.

Within the last 20 years, the technology evolution of the temperature sensors as well as more sophisticated data processing (improved sensor coatings, radiation correction algorithms, improved sensor booms, calibration fits, removal of statistical bias, etc.) has allowed large improvements. Nevertheless, they are not as large as in the case of the geopotential measurements. An improvement by a factor of roughly 3 emerges from Figure 9 that is similar to what can be seen in Figure 5. The direct sun radiation generally has a larger influence than the IR radiation balance at night, although it only heats – and never cools – the sensor. The Mauritius results in Figure 9 depict a broader span than in Figure 5. The negative outlier that is explained in section 4 remains in Figure 9 (thermometer with white coating producing a radiative cooling at night time) and larger positive biases are introduced at noon by the sun. In the previous campaigns the daytime biases only marginally extend the span on the positive side.

Figure 10 completes the results of Figure 9 with the corresponding estimated random errors of the temperature around 10 hPa (31 km). It shows a large improvement between the 1980s and the 1990s. Later on, the results from Phases IV to VI are nearly the same. However, this does not reflect the general results, as the radiosondes agreed more closely together in the low stratosphere and in the troposphere in the Mauritius campaign than in Brazil (see Table 4). At Mauritius, two of the radiosondes had daytime random errors less than 0.2 K at heights up to 30 km, whereas the other ones did reach this performance only up to 16 km. A redesign of their temperature sensor mount would minimize the fluctuations from air that has passes over surrounding sensor support structures.

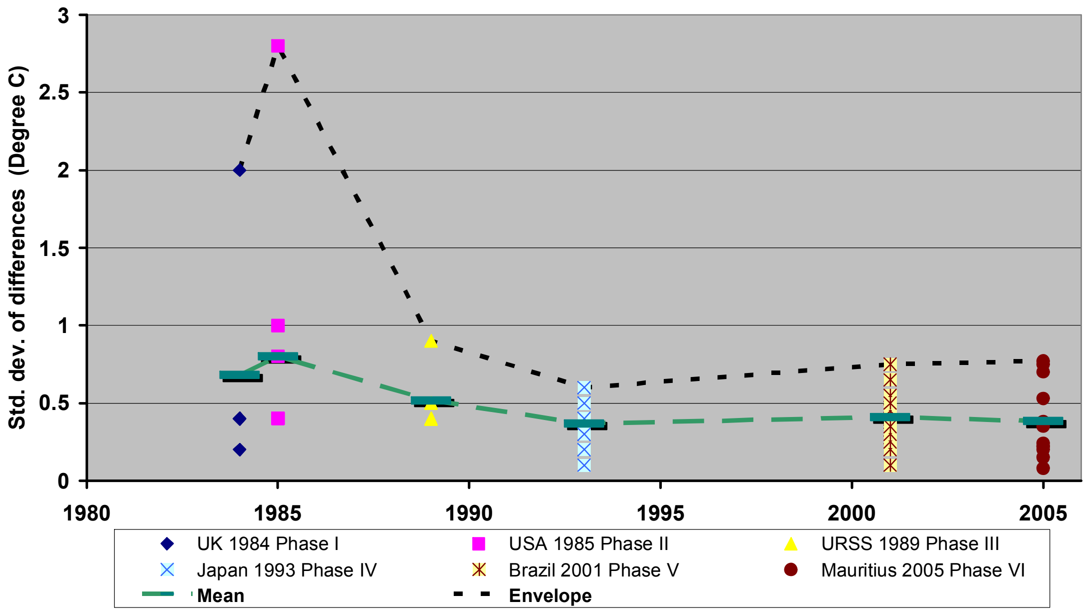


Figure 10. Estimated random errors of the temperature measurements around 10 hPa for the six WMO Radiosonde Comparisons (simultaneous measurements). The dotted line represents the envelope of all individual results. The horizontal green bars on the dashed green curve correspond to the mean random errors.

Figure 11 [J. Nash, 2004] illustrates the influence of some changes or problems in temperature measurements, on the basis of the example of the Vaisala RS80 radiosonde. Similar process analyses represent the only way to explain how the improvements of radiosonde performance have been reached or could be reached. The use of meta data in this way is appropriate to document and re-evaluate historical time series on a physical basis.

Table 4 traces the evolution of the temperature radiosonde performances from the different intercomparisons, by quoting key statements of their reports. Contrary to Figures 9 and 10, this Table presents their tropospheric and stratospheric results and highlights some of the main evolutions of the temperature measurements. As in Table 3, it is worth recalling that the first comparisons provided rather few results in the middle stratosphere.

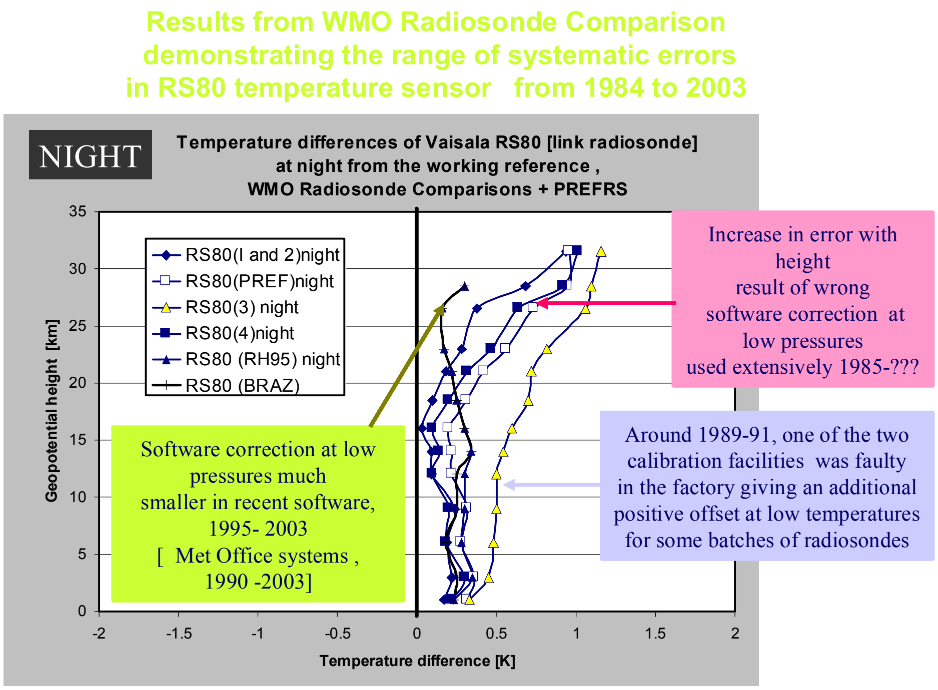


Figure 11. Range of systematic errors in RS80 temperature measurements from 1984 to 2003 [Nash, 2004].

This analysis traces the large improvements in the quality of upper-air temperature measurements. An improvement by a factor of 3 at 30 km altitude is reported. The best high quality radiosondes performed very well in the last experiment in 2005. At night most of the temperature measurements fell on average within ±0.2 K of the chosen reference. The range of the temperature measurements was similar in daytime measurements in the troposphere, but daytime measurements in the stratosphere were within ± 0.5 K of the chosen reference. Detailed studies were necessary in the last decades in order to improve these sensors. The expert group made additional recommendations for further improvements of some of the radiosondes having participated to the Mauritius experiment. This experiment indicated which high quality radiosondes are closest to the requirements of climate studies [GCOS, 2002; GCOS, 2007], i.e. producing temperature measurements of accuracy between 0.1 and 0.2 K.

The conclusions of Phases VI report pose other interesting questions, e.g. what are the performances (accuracy, reproducibility) of temperature sensors of the newest generation radiosondes? These conclusions are reproduced in 4 (see bottom lines). The conclusions previously given by J. Nash [2004] on “accuracy of best radiosonde temperature measurements” are reproduced here:

- *Accuracy depends on temperature sensor error and also the error in the height (pressure) assigned to the temperature.*
- *Temperature sensor errors are smaller at night, as long as sensor coating has low emissivity in the infrared.*
- *Solar heating introduces significant systematic errors, difficult to correct, at pressure lower than 100 hPa.*
- *Random errors in temperature are less than 0.2 K at night and less than 0.3 K at daytime in the troposphere and lower stratosphere.”*

Table 4. Results of the different WMO radiosonde comparisons related to temperature.

Phase	Results
I, 1984	In the darkness the five radiosondes provide mean temperatures through the troposphere within 0.5 °C of each other.
II, 1985	During the daytime the range of differences experienced by all of the sensors between surface and 100 hPa was less than 0.5 °C. At night the range of differences increased to about 1.0 °C.
I + II (1984-85)	The temperature at a given flight time could be measured to a reproducibility (1σ level) of at least 0.2 °C by most of the radiosondes. The observed temperature differences between radiosonde measurements were as large at night as in daytime conditions. At levels above 20 hPa the radiosondes with ducted temperature sensors show marked low biases at night, as also did the radiosondes using rod thermistor temperature sensors coated with lead carbonate. Certain discrepancies remain to be resolved (e.g., the range of 0.6 °C of the mean differences of the night time temperature measurements at levels below 50 hPa).
III, 1989	Observed differences between simultaneous temperature radiosonde measurements were within the range of 0.5 °C, and in many cases 0.1-0.2 °C, in the layers below the 50 hPa pressure level. At levels above the 30 hPa pressure level, mean bias increased up to 1-2 °C.
IV, 1993	At night-time, temperatures measured simultaneously agreed within 0.3 °C up to the altitude of the 70 hPa level. Above the 70 hPa level, the differences among them increased to a maximum of 3.2 °C, but temperatures obtained from aluminium-coated thermistors without infrared correction agreed well with those obtained from the 3-thermistor radiosondes within about 0.5 °C. At daytime, temperature obtained after the solar radiation correction agreed within 0.3 °C up to the 70 hPa level, and their differences from those obtained from 3-thermistor radiosondes were within 0.5 °C. Above the 70 hPa level, the differences among the temperatures obtained after the solar radiation correction increased and exceeded 1.0 °C, while the difference between the temperatures obtained from 3-thermistor radiosondes reached 0.7°C.
V, 2001	For daytime conditions, up to an altitude with a temperature of -30 °C, the absolute differences among the radiosondes are within the range from -0.5 to +0.5 °C. For temperatures lower than -35 °C, the differences between radiosondes increased substantially. For nigh-time periods, the differences are mostly reduced to a value close to -0.3 °C up to the height of temperature -30 °C. Above this height, the differences are still lower compared to daytime conditions and are mostly in the range from -0.5 to +0.5 °C.
VI, 2005	The temperature measurements of the seven radiosonde types agreed more closely together than in the previous intercomparisons in Brazil. This was true for both daytime and nigh time ascents. At night most of the temperature measurements fell on average within ±0.2 K of the chosen reference. The range of the temperature measurements was similar in daytime measurements in the troposphere, but daytime measurements in the stratosphere were within ± 0.5 K of the chosen reference. Two systems offer best performance night and day over the complete altitude range (most suitable radiosondes for climate monitoring up to 5 hPa). All

Phase	Results
	participating systems are acceptable for good quality operational use for flights up to 30 hPa. It is doubtful whether "reference" quality radiosondes can provide more consistent measurements than the best operational radiosondes.

7 Pressure

The pressure sensors in the early 1980s were based on the measure of the deflection of the plates of an aneroid cell. However, the methods applied for transferring pressure change with respect to altitude change were quite different. Hypsometers were mainly used with ozonesondes; only one operational radiosonde presently uses a water hypsometer. Nowadays, pressure sensors are mainly represented by silicon chips or replaced by GPS systems.

As contrasted to the previous sections where the criterion is applied to an altitude range around the 10 hPa level, the altitude range around 100 hPa (16 km) is introduced for the pressure. Figure 12 illustrates the improvements by pointing out in an anonymous manner all individual radiosonde biases of the pressure measurements around the 100 hPa level, using the method described in section 4. For the first time, Phase V intercomparison involved 2 sondes measuring pressure through GPS, which are used as reference in this section. Most radiosondes in Phase VI were of this new generation and only those are considered here. Compared to the results in the 1980s, the results from 1993 in Japan already showed a large improvement (factor 3). The span of the Brazil campaign increased then slightly, but not the standard deviation. Similarly to the geopotential altitude, the GPS technology jump brought the largest improvement. Figure 13 is devoted to the corresponding estimated random error. It confirms the improvement in the 1990s compared to the 1980s. According to the Mauritius results (Mauritius report, p. 27), "the GPS height measurements gave geopotential heights that were more accurate than the best pressure sensors at all heights above 16 km and were of similar accuracy to pressure sensor measurements at heights below 16 km".

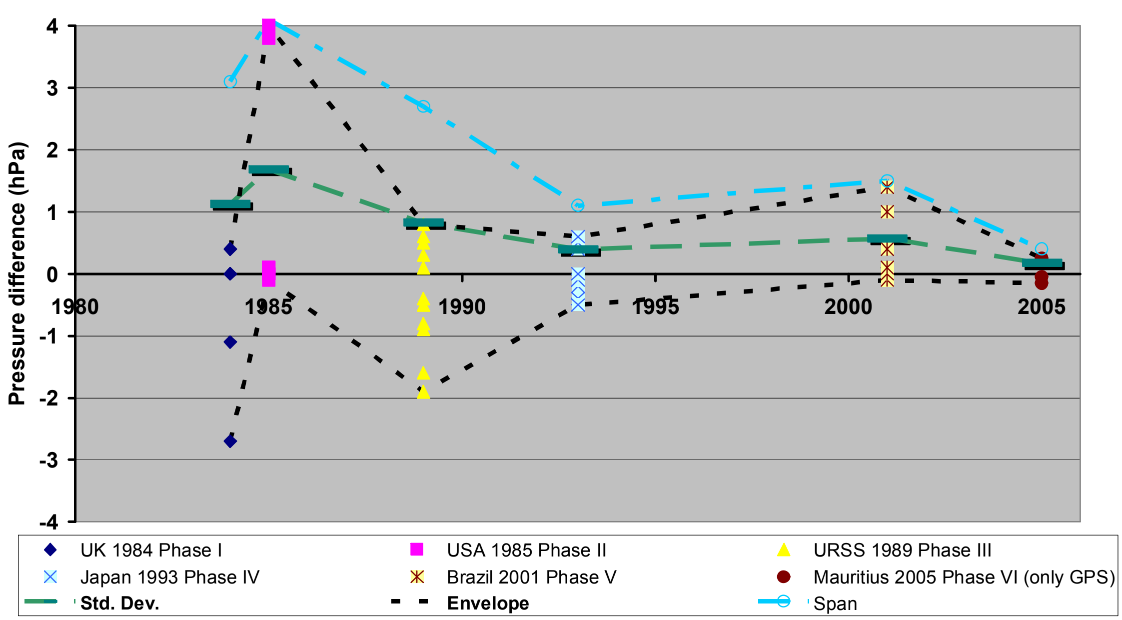


Figure 12. Bias of the pressure measurements and/or derivations around 100 hPa for the six WMO Radiosonde Comparisons (simultaneous measurements). The two dotted lines represent the envelope of all individual results, which is converted into a span with the dash-dotted line. The horizontal green bars on the dashed green curve correspond to one standard deviation of the biases for each comparison.

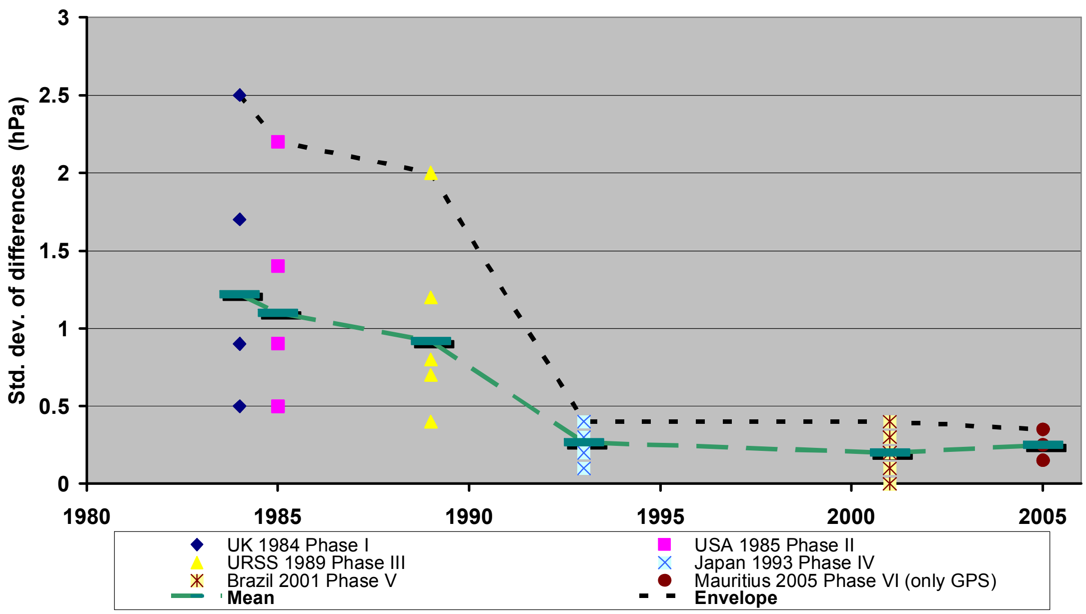


Figure 13. Estimated random errors of the pressure measurements and or derivations around 100 hPa for the six WMO Radiosonde Comparisons (simultaneous measurements). The dotted line represents the envelope of all individual results. The horizontal green bars on the dashed green curve correspond to the mean random errors.

Table 5. Results of the different WMO radiosonde comparisons related to pressure.

Phase	Results
I, 1984	Three radiosonde designs provided pressure data within, on average, one hPa of each other in the troposphere, reducing to 0.1 hPa of each other at the 10 hPa level. Two other designs differed substantially from that of the first three designs. The pressure variability (standard deviation) of the best two designs was less than ± 0.75 hPa in the lower troposphere, reducing up through the stratosphere to within ± 0.35 hPa.
II, 1985	Above 500 hPa, comparisons show the mean pressure difference to be less than 0.5 hPa.
I + II	Significant (greater than 2 hPa) mean bias errors existed in the pressure measurements of three of the eight participating radiosondes.
III, 1989	The range of values indicated by different systems was 2-3 hPa at 500 hPa decreasing to 0.3-0.5 hPa at high levels. It was noticed that pressure sensors were better at low levels, but the radar method was more reliable in the high atmosphere.
IV, 1993	The biases of most radiosondes were within 1.5 hPa up to the 200 hPa level.
V, 2001	In general, differences between pairs of radiosondes are in the range of -2.7 up to +2.7 hPa between surface and 8 km, reducing this to -0.9 up to 1.1 hPa above 8 km.

Phase	Results
VI, 2005	See 3 for pressure results derived from the GPS measurements: e. g. all GPS height measurements agreed on average to within ±20 gpm from the surface to 34 km. At 30 km, pressure sensors were in error by values between -70 gpm up to +120 gpm. The pressure sensors considered here were of extremely good quality compared to earlier generations of sensors, but were unable to provide very reliable heights at pressure < 10 hPa.

Table 5 documents these improvements by quoting the conclusions of the IOM reports. It complements the present analysis with results for the whole altitude range of the radiosonde measurements. As noted in the previous section, the conclusions of these reports provide some additional insight into the radiosonde improvements. As pressure diminishes logarithmically with altitude, the comparison of the performance of pressure measurements in different altitudes ranges should also consider their relative errors. Relative errors of pressure sensors were in the range of ±1‰ in the first 5 km at Mauritius. They grew to ±3‰ at 100 hPa (16 km), and finally to ±1-2% at 10 hPa (31 km). Opposite to the pressure sensors, the GPS technology did not experience an increase of the relative error of pressure above 16 km.

Conclusion: Large improvements have also been achieved for pressure sensors (e. g. by more than a factor 3 at the 100 hPa level), but the GPS technology constitutes a better way to improve the accuracy of pressure measurements in the stratosphere.

8 Relative humidity

In the early 1980s, goldbeater's skin, hair, and Lithium Chloride sensors were in use in some radiosondes, while others had already moved to carbon film or organic polymer film hygristors. In the mid 1980s carbon resistive hygristors were in widespread use with good results in the low troposphere. Further developments on thin film capacitor sensors have brought them to the most suitable operational humidity sensor for the lower and upper troposphere. Nowadays this technique is used in all modern operational radiosondes. The chilled mirror hygrometer remains a working reference sensor for research and development studies

The humidity measurements underwent large improvements in the last decades, but accuracy and reproducibility are still challenging goals for the low temperatures encountered in the upper troposphere. The difficulty of upper air humidity measurements is well illustrated by the fact that they have been limited for many years to the low and middle troposphere, for temperatures above and slightly below 0 ^{o}C. Phase I and II radiosonde intercomparisons restricted the humidity performance analysis to the 500 or 400 hPa levels. Later on, some of the intercomparisons were dedicated to relative humidity and additional intercomparisons were organized for this parameter, e. g. at Wallops Flight Facility in Sept. 1995. Phase V intercomparison in Brazil and Phase VI in Mauritius were also focusing on humidity sensors.

It is known that accurate measurements of relative humidity are influenced by pressure, temperature, solar radiation, and contamination from both water droplets on the sensor as well as volatile organic compounds. Moreover, the relative humidity can change rapidly, especially when the radiosonde passes through a cloud. The relative humidity is a measure that is referenced to saturation (100 % RH). As the latter strongly depends on temperature, a 50 % RH value at 20 ^{o}C corresponds to much higher water vapour content (approx. 10 g/kg at 700 hPa) than at -50 ^{o}C (less than 0.03 g/kg at 700 hPa). This wide range of water vapour content in the atmosphere implies that the relative humidity performance analyses should also include a parameter such as the specific humidity or the mixing ratio.

The WMO radiosonde comparison reports do not present humidity results in a uniform way that would allow the extraction of comparable results for all Phases. Systematic relative humidity difference in the temperature range between -35 and -45 ^{o}C (only tropospheric values) can be found in some of the reports. The classification changed with time in these reports. Pressure or altitude classes were used in the first ones, which are not appropriate for humidity (saturation depends on temperature, not on pressure). Alternatively, the analysis was restricted to one class with temperatures between 0 and +20 ^{o}C. The classification has been improved with the different reports. Nevertheless, one should reanalyze the datasets and recalculate statistics according to a single scheme. Therefore in the present study, no attempt has been made to plot the humidity performance improvements along the Phase I to VI intercomparisons.

Table 6. Results of the different WMO radiosonde comparisons related to relative humidity.

Phase	Results
I, 1984	At high relative humidity the standard deviation of all five designs is typically 2.5% RH, increasing to 6% RH for humidity below about 50% RH (measurements were only analyzed between surface and 500 hPa).
II, 1985	Based on a repeatability of 2% RH for the carbon hygristor, the repeatability appears to be 4-6% for the capacitive sensor, and 10% for the LiCl hygristor (measurements were only analyzed between surface and 400 hPa).
I + II (1984-85)	The carbon hygristor sensor has a typical reproducibility of about 3.5 % RH, but a poor resolution at RH below 20 %. The thin film capacitance sensor measures too low near saturation in low level clouds, but is considered more reliable than the carbon hygristors at the dry end of the humidity scale. Goldbeater's skin, hair and Lithium Chloride sensors have more limited capabilities than the carbon resistor and thin film capacitor sensors.
III, 1989	The thin film capacitor sensor had a better time response at lower pressure than the other sensors. However, it did not prove the same reliability under pressure significantly lower than 200 hPa.
IV, 1993	Large humidity differences were observed in the low humidity range, according to the type of sensor (capacitive film or carbon hygristor).
1995	None of the sensors reported identical humidity profiles. Final and very important conclusion is that it is doubtful that the sensor measurements can be accepted at temperatures lower then -40 ^{o}C.
V, 2001	In the troposphere up to around 8000 m, where the mixing ratio is large, the radiosonde measurements presented a low dispersion. At higher altitudes the measurements were highly dispersed.

VI, 2005	Estimating a suitable working reference is most difficult for relative humidity. At night the two most reliable relative humidity sensors agreed on average within ±2 percent relative humidity from the surface to 14 km (-70 °C) over the full range of relative humidity encountered in the intercomparison. This performance represents a large improvement over any relative humidity sensing system in previous WMO Radiosonde intercomparisons. Large systematic biases in relative humidity measurements occurred in night-time measurements as well as in daytime measurements. At temperatures higher than -40°C, maximum bias from the chosen reference at night was + 10 %. In the daytime, many radiosonde types had systematic biases in the range -10 to -20 % relative humidity for temperatures lower than -40°C. Standard deviations of the differences between different relative humidity sensors were usually relatively small (less than 5 per cent) at temperatures higher than -40°C, so the random errors in relative humidity were usually much smaller than the large systematic biases. This suggests that many of the large systematic biases could be resolved by improved sensor mounting and exposure, plus improved estimation/measurement of the relative humidity sensor temperature.

Alternatively, the main results can be extracted from the conclusions of the WMO comparison reports (Table 6), where the results of the WMO radiosonde relative humidity sensor intercomparison of 1995 are included). Due to the lack of a suitable working reference, the results of the first radiosonde comparisons mainly focused on the reproducibility of the humidity sensors. As pointed out in the conclusion of the 1995 experiment, "*it was doubtful that the sensor measurements can be accepted at temperatures lower then -40 °C*". Six years later, in the Brazil comparison mainly dedicated to GPS and humidity measurements, the results were more encouraging.

According to Sapucci et al., 2005, "*the Brazil results showed that the humidity measurements achieved by different sensors were quite similar in the low troposphere below 3 km (the bias median value regarding RS80 was around 1.8 %) and were quite dispersed in the superior layers (the median rms regarding the RS80 was around 14.9 %)*". For the first time, the Mauritius intercomparisons registered the large improvement in the humidity measurements achieved by the working references and the best capacitive hygristors (see last row of Table 6).

9 Summary and conclusions

This report represents an answer to the action initiated at the 2004 CIMO expert team meeting on upper-air systems intercomparisons (ET on UASI-1, Geneva, 17-20.03.2004): "Elaborate global criteria for tracing the improvements, based on previous intercomparisons and recent radiosonde development". It identifies different possible approaches in order to define these criteria, and selects the following method as the most appropriate one:

- Time evolution of selected statistical parameters reported in the different IOM international radiosonde comparisons organized between 1984 and 2005.

A list of candidate performance criteria is proposed (Table 2) as well as the way to extract the corresponding values from the IOM reports and to present the results. They rely on comparisons of simultaneous (time-paired) measurements. A small number of criteria are proposed in order to trace the improvement of radiosondes over the years with a straightforward data analysis. They are related to systematic and random measurement errors. These criteria can be applied to further radiosondes intercomparisons.

An illustrative check of the selected method is presented, whose promising results justify this choice. Some comments on the possibilities and limits of the method are included in this section.

Geopotential height around the 10 hPa level is the first selected criterion with very large improvements over the two last decades (Figures 7 and 8). The GPS technology allowed **an improvement of an order of magnitude in the quality of radiosonde geopotential heights at 30 km altitude**. At Mauritius, all the GPS height measurements agreed on average to within ±20 m from the surface to 34 km, whereas the mid-1980 technology provided differences in the order of 500 m at 30 km altitude. The pressure sensors compared at Mauritius were of extremely good quality compared to earlier generations of sensors, but were unable to provide very reliable heights at pressure < 10 hPa.

Large improvements have been achieved for **temperature** (Figures 9 and 10): **an improvement by a factor of 3 at 30 km altitude is reported**. The best high quality radiosondes performed very well in the last experiment in 2005: a) At night most of the temperature measurements fell on average within ±0.2 K of the chosen reference. b) The range of the temperature measurements was similar in daytime measurements in the troposphere, but daytime measurements in the stratosphere were within ± 0.5 K of the chosen reference. Detailed studies were necessary in the last decades in order to obtain these improvements. The expert group made additional recommendations for further improvements of some of the radiosondes having participated to the Mauritius experiment. This experiment indicated which high quality radiosondes are closest to the requirements of climate studies [GCOS, 2002; GCOS, 2007], i.e. producing temperature measurements of accuracy between 0.1 and 0.2 K.

Large improvements have also been achieved for **pressure sensors**, but the **GPS technology constitutes a better way to improve the accuracy of pressure measurements in the stratosphere**. Figures 12 and 13 highlight the very large improvements (by more than a factor 3) around the 100 hPa level (16 km)

Upper-air relative **humidity** measurements are most challenging. Relative humidity has been only partially treated in this report, due to their limited performance in the middle troposphere at the time of the first radiosonde comparisons. New systematic calculations on the basis of the original data sets would also be necessary in order to apply the same method as for temperature. **The Mauritius results document a large improvement over any relative humidity sensing system in previous WMO Radiosonde intercomparisons, especially for very negative temperatures encountered in the middle and upper troposphere.**

Wind has not been studied in this report, as it would require more specific criteria. However, it is well recognized that large improvements have been achieved during the last 20 years. The Mauritius report concludes "*that the new generation of GPS radiosondes should be capable of very accurate wind measurements in tropical locations, with missing data normally 5 percent or less*". All the GPS radiosondes in the Mauritius intercomparison can measure winds accurately enough to any height to satisfy climatological requirements, given that percentage of missing data is low.

Some final remarks complement these results:

- Although large improvements in the quality of the radiosondes have been achieved in the last two decades, it is not easy to quantify the overall improvements in a simple and comprehensive way. The method used in this report focuses on a limited number of criteria and provides for them a clear demonstration of large improvements in the performance of radiosondes in the last 20 years.

- The WMO international radiosonde comparisons, as well as all the other similar experiments, played a key role in the improvement of these measurements. Their results and the technology improvements have been extensively used by the manufacturers in order to successfully improve their products.
- The large improvements in radiosonde accuracy demonstrated by the WMO intercomparisons over the last 20 years represent the best achievable improvements at the operational upper-air stations provided they moved to the newest radiosonde technology.

10 References

10.1 CIMO international radiosonde intercomparisons

Hooper, A.H., 1986: WMO international radiosonde comparison, Phase I, Beaufort Park, U.K., 1984, WMO/TD-No. 174, Instruments and Observing Methods Report No. 28, World Meteorological Organization, Geneva, 118 pp.

Schmidlin, F.J., 1988: WMO international radiosonde comparison, Phase II, Wallops Island, USA, 1985, WMO/TD-No. 312, Instruments and Observing Methods Report No. 29, World Meteorological Organization, Geneva, 113 pp.

Nash, J., F.J. Schmidlin, 1987: WMO international radiosonde comparison, (U.K. 1984, U.S.A. 1985) Final report, WMO/TD-No. 195, Instruments and Observing Methods Report No. 30, World Meteorological Organization, Geneva, 103 pp.

Ivanov, A., A. Kats, S. Kumosenko, J. Nash, and N. Zaitseva, 1991: WMO international radiosonde intercomparison phase III (Dzhambul, USSR, 1989) final report. WMO/TD-No. 451, Instruments and Observing Methods Report No. 40, World Meteorological Organization, Geneva, 135 pp.

Yagi, S., A. Mita, and N. Inoue, 1996: WMO international radiosonde intercomparison phase IV (Tsukuba, Japan, 1993) final report. WMO/TD-No. 742, Instruments and Observing Methods Report No. 59, World Meteorological Organization, Geneva, 130 pp.

Balagurov, A., A. Kats, N. Krestyannikova, F. Schmidlin, 2006: WMO radiosonde humidity sensor intercomparison, 1995. WMO/TD-No. 1305, Instruments and Observing Methods Report No. 88, World Meteorological Organization, Geneva, 88 pp. http://www.wmo.int/pages/prog/www/IMOP/publications/IOM-85_RSO-RH-Phase_I-II/IOM-85_RsoHumiditySensors_Phase-I-II.pdf

da Silveira, R., G. Fisch, L. Machado, A. Dall'Antonia, L. Sapucci, D. Fernandes, and R. Marques, 2006: WMO intercomparison of GPS radiosondes, Alcantara, Brazil, 2001. WMO/TD-No. 1314, Instruments and Observing Methods Report No. 90, World Meteorological Organization, Geneva, 65 pp. http://www.wmo.int/pages/prog/www/IMOP/publications/IOM-90_RSO-Brazil/IOM-90_RSO_EMA_Alcantara2001.pdf

Nash, J., R. Smout, T. Oakley, B. Pathack, S. Kurnosenko, 2006: WMO Intercomparison of High Quality Radiosonde Systems. Vacaos, Mauritius, 2005. Final report. World Meteorological Organization, Geneva, 118 pp. http://www.wmo.int/pages/prog/www/IMOP/intercomparisons/RSO-2005/RSO-IC-2005_Final_Report.pdf

10.2 Other CIMO references

Elms, J., 2003: WMO Catalogue of Radiosondes and Upper-Air Wind Systems in use by Members in 2002; and Compatibility of Radiosonde Geopotential Measurements for the period 1998 to 2001, WMO TD-No. 1197. Instruments and Observing Methods Report No. 80, World Meteorological Organization, Geneva, 53 pp + Tables.

Gaffen, D.J., 1993: Historical changes in radiosonde instruments and practices. Final report, WMO TD-No. 541, Instruments and Observing Methods Report No. 50, World Meteorological Organization, Geneva.

Hooper, A.H., J.F. Ponting, 1983: Analysis Schemes For Treating Data from Radiosonde Intercomparisons International Organizing Committee for Radiosonde Intercomparison 1984, Report of the First Session, Annex I.

Hooper, A.H., 1986: Algorithms for Automatic Aerological Soundings. WMO TD-No. 175. Instruments and Observing Methods Report No. 21, World Meteorological Organization, Geneva, 50 pp.

Oakley, T., 1993, Report by the Rapporteur on Radiosonde Compatibility Monitoring, Part A: WMO Catalogue of Radiosondes and Upper-air Wind Systems in Use by Members (1993), Part B: Compatibility of Radiosonde Geopotential Measurements 1990, 1991 and 1992, WMO TD-No. 587. Instruments and Observing Methods Report No. 56, World Meteorological Organization, Geneva, 90 pp.

Oakley, T., 1998, Report by the Rapporteur on Radiosonde Compatibility Monitoring, Part A: WMO Catalogue of Radiosondes and Upper-air Wind Systems in Use by Members (1998), Part B: Compatibility of Radiosonde Geopotential Measurements 1995, 1996 and 1997, WMO TD-No. 886. Instruments and Observing Methods Report No. 72, World Meteorological Organization, Geneva, 112 pp.

WMO, 2006: Guide to Meteorological Instruments and Methods of Observation, WMO No. 8, seventh edition.
http://www.wmo.int/web/www/IMOP/publications/WMO-8-Guide-contents.html

10.3 Other references

GCOS, 2002: Manual on the GCOS Surface and Upper-Air Networks: GSN and GUAN, GCOS-73, WMO/TD-No. 1106, April 2002. http://www.wmo.int/web/gcos/gcoshome.html

GCOS, 2007: GCOS Reference Upper-Air Network (GRUAN): Justification, requirements, siting and instrumentation options, GCOS-112, WMO/TD-No. 1379, April 2007.

Haimberger, L., 2007: Homogenization of Radiosonde Temperature Time Series Using Innovation Statistics, Journal of Climate, Volume 20, Issue 7 (April 2007) pp. 1377–1403, DOI:10.1175/JCLI4050.1.

Miloshevich, L. M., H. Vömel, D. N. Whiteman, B. M. Lesht, F. J. Schmidlin, and F. Russo (2006), Absolute accuracy of water vapor measurements from six operational radiosonde types launched during AWEX-G and implications for AIRS validation, J. Geophys. Res., 111, D09S10, doi:10.1029/2005JD006083.

Sapucci, Luiz F., Luiz A. T. Machado, Reinaldo B. da Silveira, Gilberto Fisch and João F. G. Monico, 2005: Analysis of Relative Humidity Sensors at the WMO Radiosonde Intercomparison Experiment in Brazil. J. of Atmospheric and Oceanic Technology: Vol. 22, No. 6, pp. 664–678.

11 Annexe

Tables.

Example of Excel sheet used for the extraction of the quality performance results of the Phases I to VI WMO radiosonde comparisons. This Table presents the systematic temperature differences at 10 hPa, at night time, in Degree Celsius. The results are presented in the columns under the Phase labels. The source of each result is referenced with a figure number in the next column to the right. In addition, the reference used in the intercomparisons, the report and figure are given. In the first reports, the graphs do not present the corresponding values: a direct “analogue” reading of the values on the plots is then specified (in which case the absolute value reported may already be affected by this step!).

Systematic temperature difference at 10 hPa, day time in Degree Celsius	Phase -		UK 1984 Phase I		USA 1985 Phase II		URSS 1989 Phase III		Japan 1993 Phase IV		Brazil 2001 Phase V		Mauritius 2005 Phase VI	
			1984		1985		1989		1993		2001		2005	
Radiosonde														
OCAN 1524-511			-3.5	4d										
RS 3 (UK)			-0.1	4d										
RS4 MK3(Aus)					-0.3	5.5								
MK-III (India)					1.3	5.5								
Graw 78 C (D)			0.4	4d										
Graw DFM97 (D)											-0.65	ppt	0.45	9.13
SMA-TC-1 (SMT) China														
SMA-GZZ (SMG) China							-2.76	5.10						
MARS-2 (MRS) URSS							-1.49	5.10						
MRZ-3A (MRZ) URSS							-2.8	5.10						
Meisei RS2-80 (JP1)									-1.8	2.2a				
Meisei RS2-91 (JP2)									0.1	2.2a				
Meisei RS2-01G													0.8	9.13
Vaisala RS80-15N (FN1)			0.1	4d	-1.4	5.5	-1.01	5.10	-0.8	2.2a				
Vaisala RS80-15N (FN3)										2.2.c				
Vaisala RS80-15LH (FN2)											-1	ppt		
Vaisala RS90-…											-0.2	ppt		
Vaisala RS92-…													-0.2	9.13
AIR IS-4A- (AR1)							1.26	5.10	0.7	2.2a				
AIR IS-4A- (AR2)									-0.4	2.2a				
AIR IS-4A- (AR3)									-0.3	2.2.c				
VIZ 1392 (VIZ0)			0.6	4d	0	5.5	1.01	5.10						
VIZ Mark II (VIZ)									1	2.2a				
VIZ Mark II (VZ2)									-0.1	2.2.c				
VIZ Mark II (VZ3)									-0.2	2.2.c	0.8	ppt		
Sippican Chip (prototype)											0.5	ppt		
Sippican LMS-5													0.25	9.13
GL-98 (MODEM)											-0.9	ppt	0.3	9.13
SRS-C34													-0.4	9.13
Statistics														
Minimum			-3.5		-1.4		-2.8		-1.8		-1.0		-0.4	
Maximum			0.6		1.3		1.3		1.0		0.8		0.8	
Span			4.1		2.7		4.1		2.8		1.8		1.2	
Mean			-0.5		-0.1		-1.0		-0.2		-0.2		0.2	
Mean Abs. Dev.			1.2		0.8		1.4		0.6		0.6		0.3	
Sigma			1.5		1.0		1.6		0.8		0.7		0.4	
Mean - 1*sigma			-2.0		-1.1		-2.6		-1.0		-0.9		-0.2	
Mean + 1 sigma			1.0		0.9		0.7		0.6		0.4		0.6	
RMSD			1.6		1.0		1.9		0.8		0.7		0.4	
Reference used:			Mean(FIN,UK)		VIZ0		Mean(FN1,VIZ0)		Mean(FN1,AR1)		id. Mauritius		Mean(Me, Si, Va, SRSadj)	
Table used:			I:4d		II: 5.5 (1400UTC)		III: 5.10		IV: 2.2a, 2.2.c		ppt Nash		Fig. 9.13	
Type of reading			Analogic		Analogic		Digital		Digital		Analogic		Analogic	

Same as previous Table, but for the daytime and with some statistical parameters.

Systematic temperature difference at 10 hPa, day time in Degree Celsius	Phase -		UK 1984 Phase I		USA 1985 Phase II		URSS 1989 Phase III		Japan 1993 Phase IV		Brazil 2001 Phase V		Mauritius 2005 Phase VI	
			1984		**1985**		**1989**		**1993**		**2001**		**2005**	
Radiosonde														
OCAN 1524-511			**-3.5**	**4d**										
RS 3 (UK)			**-0.1**	**4d**										
RS4 MK3(Aus)					**-0.3**	**5.5**								
MK-III (India)					**1.3**	**5.5**								
Graw 78 C (D)			**0.4**	**4d**										
Graw DFM97 (D)											**-0.65**	**ppt**	**0.45**	**9.13**
SMA-TC-1 (SMT) China														
SMA-GZZ (SMG) China							**-2.76**	**5.10**						
MARS-2 (MRS) URSS							**-1.49**	**5.10**						
MRZ-3A (MRZ) URSS							**-2.8**	**5.10**						
Meisei RS2-80 (JP1)									**-1.8**	**2.2a**				
Meisei RS2-91 (JP2)									**0.1**	**2.2a**				
Meisei RS2-01G													**0.8**	**9.13**
Vaisala RS80-15N (FN1)			**0.1**	**4d**	**-1.4**	**5.5**	**-1.01**	**5.10**	**-0.8**	**2.2a**				
Vaisala RS80-15N (FN3)										**2.2.c**				
Vaisala RS80-15LH (FN2)											**-1**	**ppt**		
Vaisala RS90-…											**-0.2**	**ppt**		
Vaisala RS92-…													**-0.2**	**9.13**
AIR IS-4A- (AR1)							**1.26**	**5.10**	**0.7**	**2.2a**				
AIR IS-4A- (AR2)									**-0.4**	**2.2a**				
AIR IS-4A- (AR3)									**-0.3**	**2.2.c**				
VIZ 1392 (VIZ0)			**0.6**	**4d**	**0**	**5.5**	**1.01**	**5.10**						
VIZ Mark II (VIZ)									**1**	**2.2a**				
VIZ Mark II (VZ2)									**-0.1**	**2.2.c**				
VIZ Mark II (VZ3)									**-0.2**	**2.2.c**	**0.8**	**ppt**		
Sippican Chip (prototype)											**0.5**	**ppt**		
Sippican LMS-5													**0.25**	**9.13**
GL-98 (MODEM)											**-0.9**	**ppt**	**0.3**	**9.13**
SRS-C34													**-0.4**	**9.13**
Statistics														
Minimum			**-3.5**		**-1.4**		**-2.8**		**-1.8**		**-1.0**		**-0.4**	
Maximum			**0.6**		**1.3**		**1.3**		**1.0**		**0.8**		**0.8**	
Span			**4.1**		**2.7**		**4.1**		**2.8**		**1.8**		**1.2**	
Mean			**-0.5**		**-0.1**		**-1.0**		**-0.2**		**-0.2**		**0.2**	
Mean Abs. Dev.			**1.2**		**0.8**		**1.4**		**0.6**		**0.6**		**0.3**	
Sigma			**1.5**		**1.0**		**1.6**		**0.8**		**0.7**		**0.4**	
Mean - 1*sigma			**-2.0**		**-1.1**		**-2.6**		**-1.0**		**-0.9**		**-0.2**	
Mean + 1 sigma			**1.0**		**0.9**		**0.7**		**0.6**		**0.4**		**0.6**	
RMSD			**1.6**		**1.0**		**1.9**		**0.8**		**0.7**		**0.4**	
Reference used:			Mean(FIN,UK)		VIZ0		Mean(FN1,VIZ0)		Mean(FN1,AR1)		id. Mauritius		Mean(Me, Si, Va, SRSadj)	
Table used:			I:4d		II: 5.5 (1400UTC)		III: 5.10		IV: 2.2a, 2.2.c		ppt Nash		Fig. 9.13	
Type of reading			Analogic		Analogic		Digital		Digital		Analogic		Analogic	