Non-parametric error distribution analysis from the laboratory calibration of various rainfall intensity gauges

L. G. Lanza and L. Stagi

ABSTRACT

The analysis of counting and catching errors of both catching and non-catching types of rain intensity gauges was recently possible over a wide variety of measuring principles and instrument design solutions, based on the work performed during the recent Field Intercomparison of Rainfall Intensity Gauges promoted by World Meteorological Organization (WMO). The analysis reported here concerns the assessment of accuracy and precision of various types of instruments based on extensive calibration tests performed in the laboratory during the first phase of this WMO Intercomparison. The non-parametric analysis of relative errors allowed us to conclude that the accuracy of the investigated RI gauges is generally high, after assuming that it should be at least contained within the limits set forth by WMO in this respect. The measuring principle exploited by the instrument is generally not very decisive in obtaining such good results in the laboratory. Rather, the attention paid by the manufacturer to suitably accounting and correcting for systematic errors and time-constant related effects was demonstrated to be influential. The analysis of precision showed that the observed frequency distribution of relative errors around their mean value is not indicative of an underlying Gaussian population, being much more peaked in most cases than can be expected from samples extracted from a Gaussian distribution. The analysis of variance (one-way ANOVA), assuming the instrument model as the only potentially affecting factor, does not confirm the hypothesis of a single common underlying distribution for all instruments. Pair-wise multiple comparison analysis revealed cases in which significant differences could be observed.

Key words | accuracy, precision, rainfall intensity, rain gauges

L. G. Lanza (corresponding author)
University of Genova,
Department of Construction,
Environmental and Architectural Engineering,
1 Montallegro, 16145,
Genoa
Italy
E-mail: *luca.lanza@unige.it*

L. G. Lanza
L. Stagi
WMO/CIMO Lead Centre 'B. Castelli' on
Precipitation Intensity,
Italy

INTRODUCTION

When dealing with rainfall intensity (RI) measurements, the quality of traditional ground-based observing systems is often addressed in terms of consistency checks and the related acquisition and reporting errors. These issues are relevant for operational networks and require in-depth analysis and careful treatment of historic records (see e.g. Koutsoyiannis 2010). Inherent inaccuracies of RI measurements, which also affect most recorded rain series, are however often understated, and corrections are rarely applied to account for either the well known systematic mechanical errors of tipping-bucket rain gauges (TBRGs) or the time constant effect of Weighing Gauges (WGs).

The lack of consistency in the underlying accuracy of rainfall measurements performed at various intensity levels affects the most common statistics derived from long-term rain series (Molini *et al.* 2001; La Barbera *et al.* 2002), with not negligible consequences on many practical applications. While the historic rain records are unavoidably affected by these shortcomings, due to understated accuracy problems in rainfall measurements performed in the past, the same level of inaccuracy is no longer acceptable today (Lanza & Stagi 2008). It appears quite unreasonable that one of the most relevant atmospheric variables, whose quantification is essential to the understanding of the land phase processes of the hydrological cycle and many other related processes – i.e. rainfall intensity – is still widely measured today at a much lower accuracy than the present knowledge and technology would permit.

The WMO (World Meteorological Organization) recognised these emerging needs in 2001, and launched two specific instrument intercomparison initiatives in order to contribute to filling this gap in documented instrument

doi: 10.2166/wst.2012.075

performance (Sevruk *et al.* 2009). Thorough analysis of counting and catching errors of both catching and non-catching types of rain gauges was recently possible based on the work performed during the WMO Field Intercomparison of RI Gauges over a variety of measuring principles and instrument design solutions (Vuerich *et al.* 2009). Compliance with the WMO accuracy specifications and the comparison of performance were addressed in previously published papers (Lanza & Stagi 2009; Lanza & Vuerich 2009). The analysis reported here addresses a further step towards the assessment of accuracy and precision of various types of instruments based on the same set of calibration tests performed in the laboratory during the first phase of the WMO Intercomparison (see Lanza & Stagi 2009 for details on the methods and procedures used). Performance of the same gauges in real world conditions at the field test site and the related catching and operational types of errors are addressed in Lanza & Vuerich (2012).

RATIONALE AND METHOD

Following the International Vocabulary of Metrology (JCGM 2008), accuracy is measured by the bias (or trueness) – i.e. the closeness of agreement between the average of an infinite number of measured quantity values and a reference quantity value. The intermediate measurement precision (synthetically 'precision' in this paper) is on the contrary the closeness of agreement between measured quantity values obtained by replicate measurements under a set of conditions of measurement (including the same measurement procedure and replicated measurements over an extended period of time).

In simple terms, the precision of an instrument can be associated with the random dispersion of deviations around the mean value of the measured variable, while accuracy is connected to systematic biases in the measurements with respect to the true value of the measure variable. The biases can be easily recognised and corrected after proper calibration, although this situation is not often the case for operational RI gauges, while precision can be difficult to improve due to uncontrolled factors, which may affect the instrument performance.

Further insights into the behaviour of the errors obtained from laboratory test results are sought in the following sections by investigating their frequency distribution and the associated bias with respect to a reference (true) value of RI, generated in the form of a constant water flow rate under controlled laboratory conditions.

The results of the laboratory tests performed on a number of catching type RI gauges are available from the recent WMO Intercomparison as contemporary values of the reference and measured intensities at a resolution of one minute in time (Lanza & Stagi 2009). The available data were elaborated to derive the one-minute error figures, intended as the relative deviations from the reference intensity. The percentage relative error e_{rel} [%] was therefore calculated as follows:

$$e_{rel}[\%] = \frac{I_{meas} - I_{ref}}{I_{ref}} \times 100$$

with I_{meas} [mm·h^{-1}] the RI measured by the instrument and I_{ref} [mm·h^{-1}] the equivalent reference (true) RI (calculated from the generated reference flow rate).

Per each catching type rain gauge tested in the laboratory, a generally high, although variable number of one-minute error figures is available, obtained for a set of predetermined reference intensity values. These data constitute a sample of error figures that is suitable to investigate accuracy and precision, and to compare the behaviour of different rain gauges.

RESULTS AND DISCUSSION

The accuracy of catching type RI gauges, expressed in quantitative terms through the bias observed with respect to the true value of the measurand variable (the reference intensity), was presented by Lanza & Stagi (2009). Overall results of this analysis are recalled in Figure 1, in which a box-plot representation is used to compare the performance of the various RI gauges (the instrument model and manufacturer are reported on the x-axis). The values obtained for the mean (solid line), median (thin line), 25–75th percentiles (box limits), 10–90th percentiles (whisker caps) and 5–95th percentiles (black circles) per each series of one-minute percentage relative errors obtained at the individual reference intensities are reported. Grey shaded vertical bars in the upper part of the graph indicate the sample size (number of minutes available to calculate statistics of the errors) according to the scale reported on the right hand side of the graph. The ±5% limits for the requested accuracy of RI gauges as from WMO (2008) are also indicated by the horizontal dashed lines for reference purposes.

The bias observed for most of the investigated RI gauges (expressed here by the sample mean values) is contained

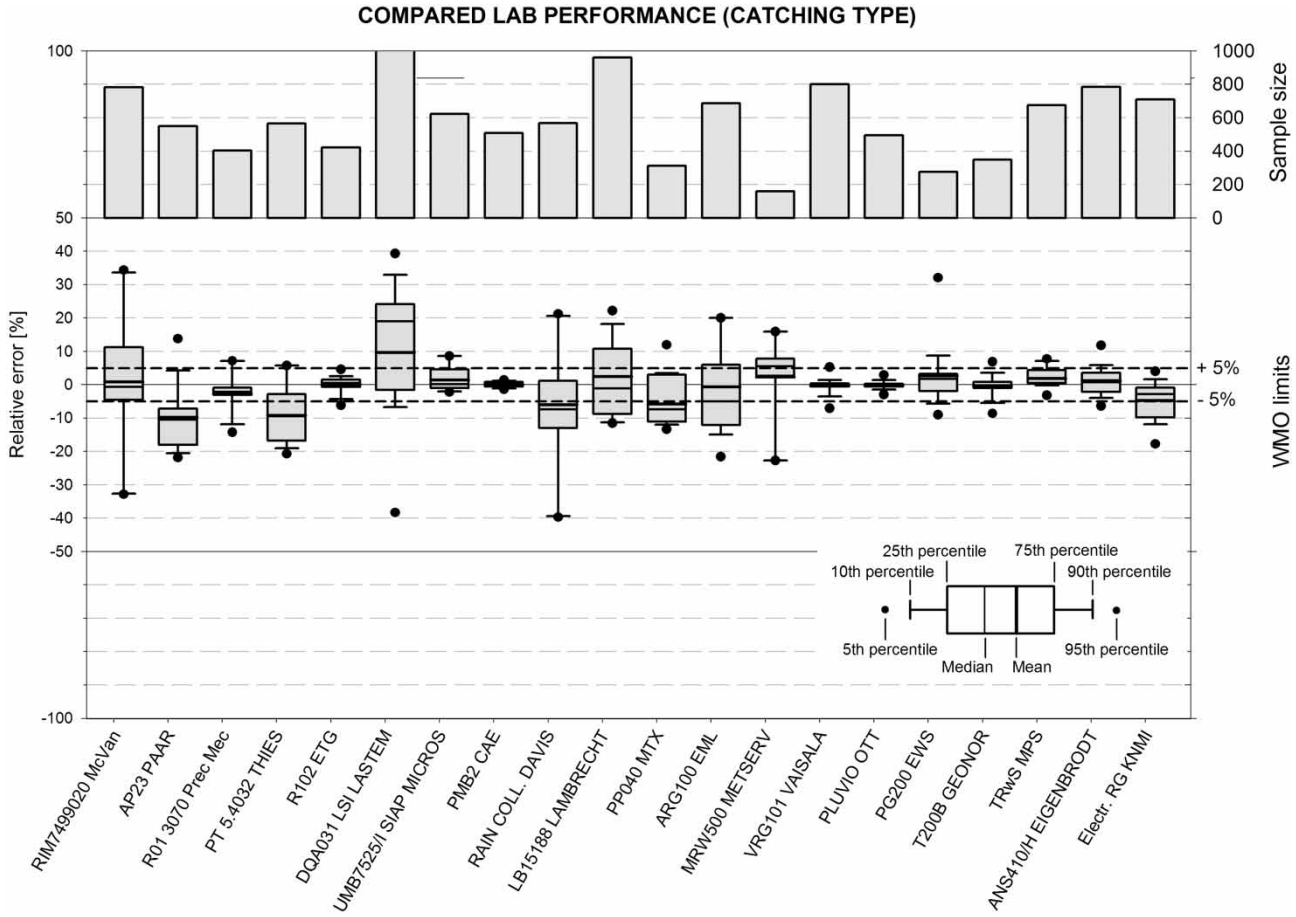

Figure 1 | Overview of the non-parametric analysis of accuracy for the catching type RI gauges investigated in the laboratory intercomparison. The graph is constructed using data from the whole set of tests performed, therefore involving all rainfall intensities in the range from 2 to 2,000 mm/h.

within the limits specified by WMO. This limit must be understood as the general capability of most instruments to provide, on average, acceptable performance under laboratory tests. Note that these tests are performed under steady-state conditions (constant reference flow rate) for a sufficient period of time, so that the instrument can reach stable measurement conditions in which the mean and median values of the errors are facilitated to approach the reference value. In many cases the mean and the median are fairly coincident, while some asymmetric behaviour is observed for some of the investigated instruments.

The dynamic behaviour of the gauges is however much more relevant here, as the real world RI during rain events is far from being constant over even quite short intervals in time. Therefore the one-minute variability is of interest too, even under laboratory steady state conditions. The number of instruments, out of the whole set of catching type gauges, with observed 10–90th error percentiles (whisker caps) and outliers (black circles) that are contained within the WMO limits for the requested RI measurement accuracy is much lower (about 25%) than was observed for the sample mean and median. In those cases the measure can be assumed to be only partially accurate, with some non-negligible dispersion of the errors around their sample mean and median values.

Further analysis of this dispersion, namely the intermediate measurement precision, was carried out in this work with the objective of deriving useful information by investigating the frequency distribution of the errors. The standard variate was therefore calculated for each RI gauge from the series of percentage relative errors and tested for normality using parametric and non-parametric classic statistical tools.

The results are first reported in Figure 2 in the form of standard Q–Q plots, in which the sample quantiles of the frequency distribution of standardised relative errors are plotted against the theoretical quantiles of a Gaussian

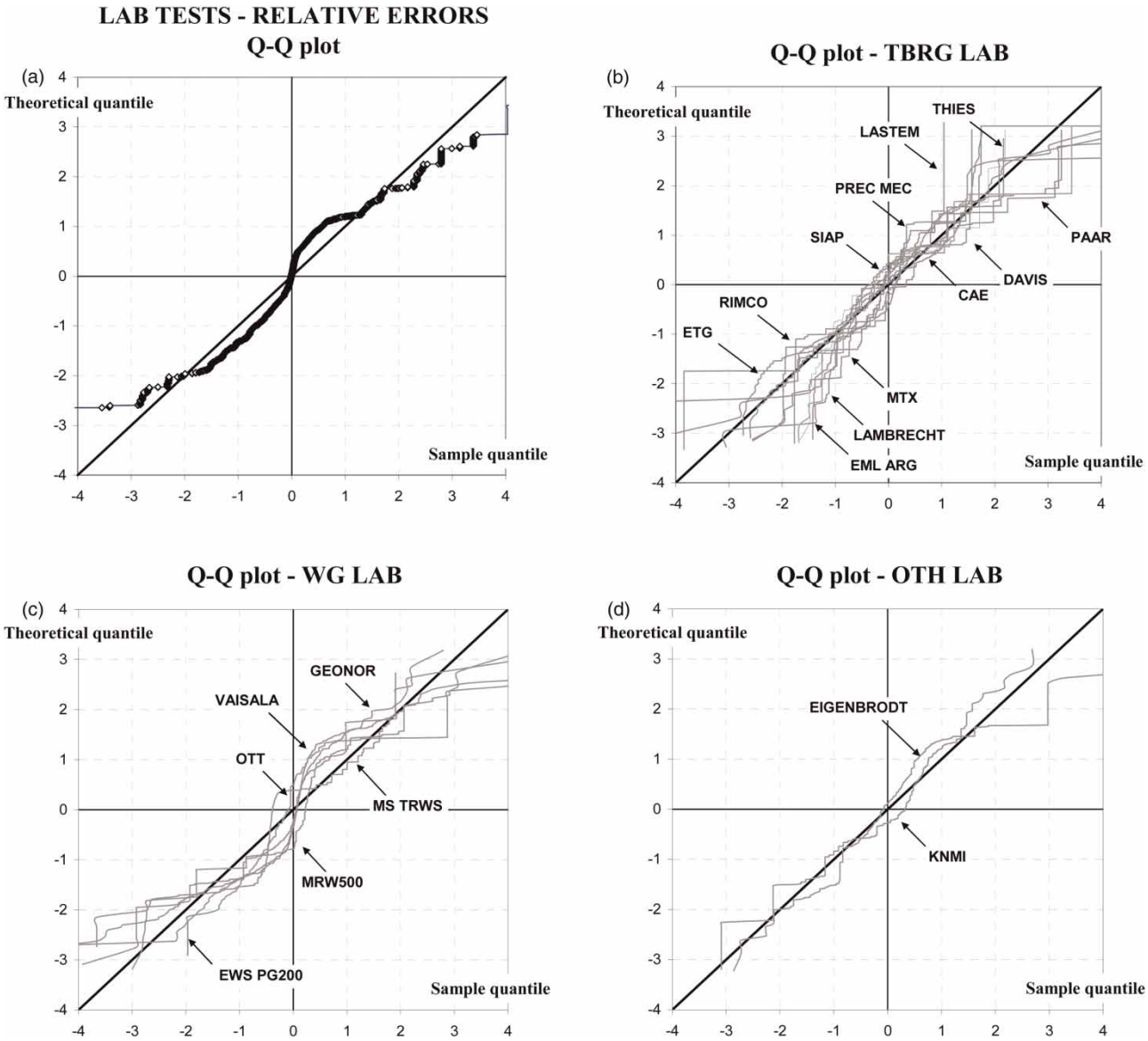

Figure 2 | Q–Q plots of the frequency distribution of relative errors for (a) the whole set of instruments investigated and for (b) tipping-bucket rain gauges (TBRGs), (c) weighing gauges (WGs) and (d) other types of gauges (OTH).

probability distribution $N(0,1)$. One curve is drawn per each single gauge (the manufacturer is reported in the associated label) although results are grouped according to the measuring principle of each instrument in the graphs presented in Figure 2(b, c, d). The three groups are respectively the TBRGs, WGs and other types of gauges – basically level measuring gauges (OTH); the overall distribution is also represented in Figure 2(a).

Relative errors calculated from the laboratory performance of the various instruments show a typical step-wise pattern in the quantile-based representation, due to the finite number of discrete reference intensity values used to perform steady-state constant flow rate tests in the laboratory. Note that TBRGs perform quite differently from each other, depending on the type and nature of the correction employed to account for systematic mechanical errors (including none), although they are all superimposed in this graph.

The TBRGs, being also affected by the sampling error due to the finite volume of the counting volume (the bucket) and due to the discrete nature of the set of reference flow rates generated in the laboratory, show a more accentuated step-wise Gaussian-like distribution of the errors. The normality test is not successful for all the investigated

gauges. The frequency distribution is however quite symmetric, and the 3rd and 4th order sample moments of the distribution are not far from their theoretical Gaussian figures. The Gaussian-like behaviour can be assumed to be an indication of a fairly random distribution of the errors around the bias (or trueness) value, although not necessarily demonstrating a high measurement precision.

Sub-categories of TBRGs were also investigated separately in order to account for the different correction techniques employed. These include no correction, software based correction, pulse based correction and mechanical correction (see Lanza *et al.* 2005 for a description of such correction methods). The associated Q–Q plots are reported in Figure 3 revealing that no special behaviour can be associated with the various types of correction in terms of precision, while accuracy is quite different as already illustrated in Figure 1.

As for the WGs, departures from the Gaussian distribution are significant especially in the vicinity of the sample mean and median values, which are noticeably close to each other. The high value of the kurtosis indicates a much less dispersed distribution of the errors with respect to the Gaussian case. This situation means that both accuracy and precision are fairly good in the laboratory for such gauges (they are 'better than Gaussian'), as the slope

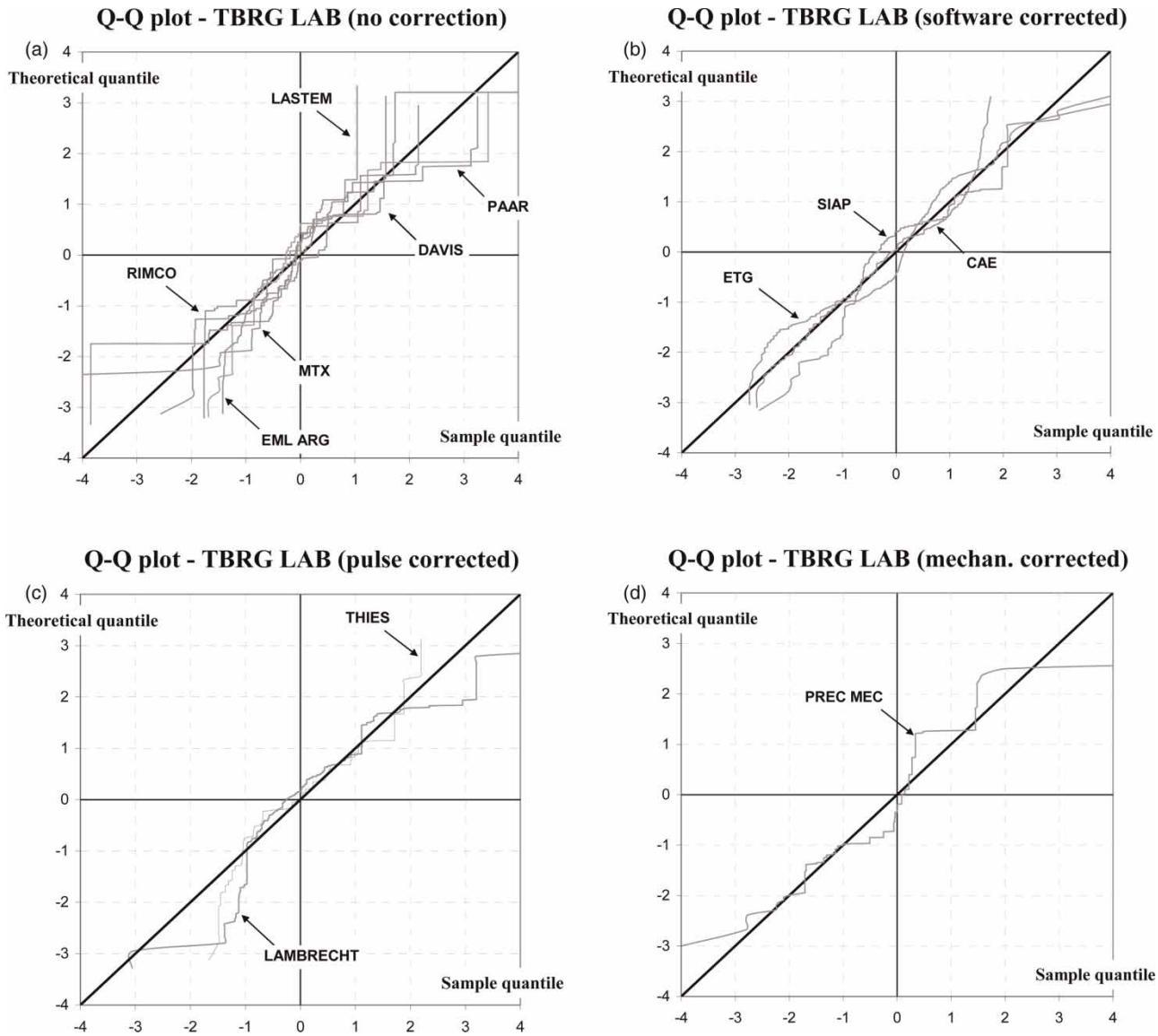

Figure 3 | Q–Q plots of the frequency distribution of relative errors for tipping-bucket rain gauges (TBRGs) with (a) no correction employed, (b) software based correction, (c) pulse based correction and (d) mechanical correction.

of their error distribution approaches the vertical near to the origin. It is therefore an expected result, in this case, that testing for normality is not successful for all WGs.

The third category is that of level measuring gauges, with a behaviour in the laboratory that is similar to the TBRGs with a step-wise Gaussian-like frequency distribution of the errors, again partly due to the discrete nature of the reference flow rates and sampling issues.

The calculated *p*-value determines the probability of being incorrect in concluding that the data are not normally distributed (*p*-value is the risk of falsely rejecting the null hypothesis that the data are normally distributed). Not one of the series of relative errors passed the test for normality (the Wilk–Shapiro test was used in this work).

Further statistical analysis of the series was performed using group comparison techniques, specifically ANOVA (ANalysis Of VAriance) and a few related procedures (see e.g. Howell 2002). Group comparison is used to test two or more different groups for a significant difference in the mean or median values beyond what can be attributed to random sampling variation. One-way ANOVA, is a parametric test that compares the effect of a single factor (the gauge model and manufacturer here) on the mean of two or more groups. The standard variate of relative errors was used in this test, and the null hypothesis is that there is no difference among the populations from which the samples were drawn.

Actually, parametric tests assume that samples were drawn from normally distributed populations with the same variances (or standard deviations). They are based on estimates of the population means and standard deviations, the parameters of a normal distribution, derived from sample statistics. Non-parametric tests do not assume that the samples were drawn from a normal population. Instead, they perform a comparison on ranks of the observations. Rank sum tests automatically rank numeric data, then compare the ranks rather than the original values.

As samples are taken here from populations with apparently non-normal distribution and/or unequal variance, the Kruskal–Wallis ANOVA on ranks was used (Kruskal & Wallis 1952), which is the non-parametric analogue of the one-way ANOVA.

Kruskal–Wallis ANOVA is a non-parametric method for testing whether samples originate from the same distribution. It is used for comparing more than two samples that are independent, or not related. The parametric equivalence of the Kruskal–Wallis test is the one-way ANOVA. The factual null hypothesis is that the populations from which the samples originate, have the same median. When the Kruskal–Wallis test leads to significant results, then at least one of the samples is different from the other samples. The test does not identify where the differences occur or how many differences actually occur.

The Kruskal–Wallis ANOVA on ranks arranges the data into sets of rankings, then performs an ANOVA based on these ranks, rather than directly on the data, so it does not require assuming normality and equal variance. However, the test does assume an identically shaped and scaled distribution for each group, except for any difference in medians. The null hypothesis is that there is no difference in the distribution of values between the different groups.

The advantages of parametric ANOVA on the Kruskal–Wallis non-parametric method are that, when the normality and equal variance assumptions are met, they are slightly more sensitive than the analysis based on ranks. When the assumptions are not met, however, the Kruskall–Wallis ANOVA on ranks is more reliable. The Wilk–Shapiro test for normality (Shapiro & Wilk 1965) and the Levene median test for assessing the equal variance assumption (Brown & Forsythe 1974) were used in this work.

After application of the Kruskall–Wallis ANOVA on ranks for the 20 series of standardised relative errors obtained from the various rain gauges, it can be concluded that the differences in the median values among the series are greater than would be expected by chance; there is a statistically significant difference ($p < 0.001$). Note that ANOVA techniques (both parametric and non-parametric) test the hypothesis of no differences between the groups, but do not indicate what the differences are. Multiple comparison procedures can be used to isolate these differences, and the pair-wise multiple comparison procedure based on Dunn's method (see e.g. Sheskin 2007) was applied to this aim. The Dunn test compares each pair of groups and highlights which of these pair-wise differences are significant. Results are synthetically reported in Figure 4, in which symbols indicate the pairs of instruments that are affected by significant differences in the distribution of the associated relative errors.

It is evident from Figure 4 that in most cases, the null hypothesis that data samples from pairs of different instruments are extracted from the same population can be accepted with reasonable confidence ($p < 0.05$), while in a certain number of cases (15%) the difference is significant beyond what could be expected by chance. Therefore, in those cases, precision can be assumed to be significantly different between the two instruments investigated.

It can be noted that, when considering the TBRGs, the EML ARG, ETG and LASTEM gauges show statistically significant differences when compared with many other instruments of the same family and of the WG type.

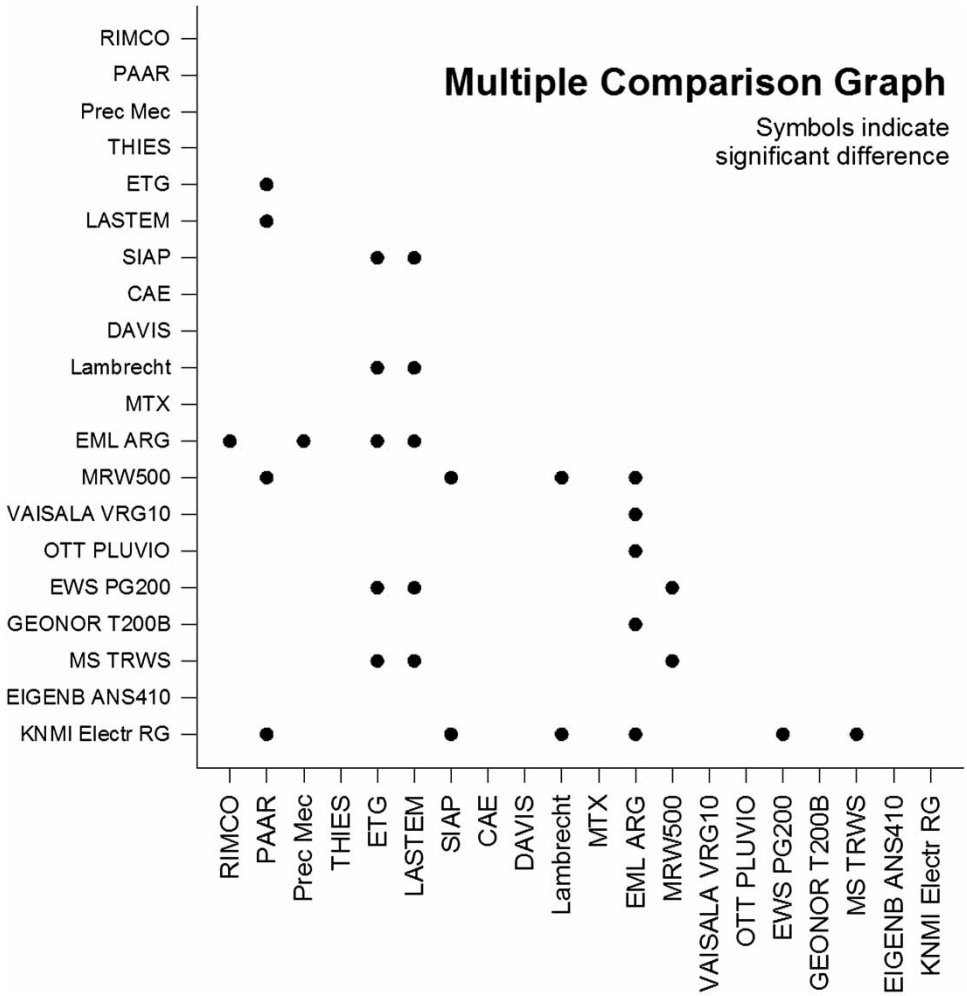

Figure 4 | Synthetic representation of the results of the pair-wise multiple comparison procedure based on Dunn's method for the set of RI catching type gauges investigated.

For the EML ARG and LASTEM this situation could be due to mechanical reasons, as they do not use software correction, while in the case of the ETG gauge the difference could be ascribed to the specific software correction. Note that this does not mean that these instruments have lower performance than the others, as can be easily checked from Figure 1. Regarding the weighing type gauges, it seems that the MRW500 has a different behaviour than many TBRGs and WGs, but this effect could be ascribed to the limited sample size investigated, as reported in Figure 1.

The information will be useful to identify the opportunities to improve the performance of RI measurement instruments and to provide hints for technical or technological developments in those directions. For instance, problems with the software algorithm used in the data-logger of the instrument to elaborate on the raw data to provide RI as output information could lead to a non-symmetric distribution of the deviations of the errors obtained from laboratory tests. A widely spread but Gaussian distribution around an accurate mean value would instead indicate a mere problem of instrument precision, which may be due to the concept used to calculate RI, especially in the case of TBRGs with a low instrument resolution (large bucket volume). Further problems can be identified by performing a one-to-one investigation of each single instrument, based on the relevant instrument technology, design and data processing characteristics, which is however far beyond the scope of the present paper.

CONCLUSIONS

The reported analysis of precision and accuracy of a large number (20) of catching type RI gauges demonstrated the

different performance of the various measuring principles and instrument design characteristics (as far as their counting errors are concerned) under controlled laboratory conditions and constant flow rate tests.

From the performed non-parametric analysis of relative errors it can be concluded that the accuracy of the investigated RI gauges is generally high, after assuming that it should be at least contained within the limits set forth by WMO in this respect. The measuring principle exploited by the instrument is generally not very decisive in obtaining such good results in the laboratory. Rather, the attention paid by the manufacturer in suitably accounting and correcting for systematic errors and time constant related effects was demonstrated to be influential.

The analysis of precision revealed that the observed frequency distribution of relative errors around their mean value is not indicative of an underlying Gaussian population, being much more peaked in most cases than can be expected from samples extracted from a Gaussian distribution. The analysis of variance (one-way ANOVA), assuming the instrument model as the only potentially affecting factor, did not confirm the hypothesis of a single common underlying distribution for all samples. Pair-wise multiple comparison analysis revealed the cases in which significant differences could be observed.

The results of the intercomparison of RI gauges in terms of their accuracy and precision could support both manufacturers and final users in improving the performance of the observing systems deployed on the territory. In the first case, they can provide hints and indications on how to improve the performance of the instrument in measuring RI (directing when investments should be made), as far as is possible using the selected measuring principle. In the second case, they can support the development of technical standards in the field of RI measurements with the aim of providing the final user with some quantitative assessment of the quality and expected performance of the various types of instruments available (set of reference values to be requested in tenders, etc.).

REFERENCES

Brown, M. B. & Forsythe, A. B. 1974 Robust tests for equality of variances. *Journal of the American Statistical Association* **69**, 364–367.

Howell, D. 2002 *Statistical Methods for Psychology*. Duxbury/Thomson Learning, Pacific Grove, CA, pp. 324–325.

JCGM 200:2008 *International Vocabulary of Metrology – Basic and General Concepts and Associated Terms (VIM)*, 3rd edition. Joint Committee for Guides in Metrology, Paris (also available from: http://www.bipm.org/utils/common/documents/jcgm/JCGM_200_2008.pdf).

Koutsoyiannis, D. 2010 A note of caution for consistency checking and correcting methods of point precipitation records. IPC10, *10th International Precipitation Conference*, 23–25 June 2010, Coimbra, Portugal.

Kruskal, W. & Wallis, W. A. 1952 Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association* **47** (260), 583–621.

La Barbera, P., Lanza, L. G. & Stagi, L. 2002 Tipping bucket mechanical errors and their influence on rainfall statistics and extremes. *Water Science and Technology* **45** (2), 1–10.

Lanza, L. G. & Stagi, L. 2008 Certified accuracy of rainfall data as a standard requirement in scientific investigations. *Advances in Geosciences* **16**, 43–48.

Lanza, L. G. & Stagi, L. 2009 High resolution performance of catching type rain gauges from the laboratory phase of the WMO field intercomparison of rain intensity gauges. *Atmospheric Research* **94** (4), 555–563.

Lanza, L. G. & Vuerich, E. 2009 The WMO field intercomparison of rain intensity gauges. *Atmospheric Research* **94** (4), 534–543.

Lanza, L. G. & Vuerich, E. 2012 Non-parametric analysis of one-minute rain intensity measurements from the WMO field intercomparison. *Atmospheric Research* **103**, 52–59.

Lanza, L., Leroy, M., Alexandropoulos, C., Stagi, L. & Wauben, W. 2005 Laboratory intercomparison of rainfall intensity gauges. World Meteorological Organization – Instruments and Observing Methods Report No. 84, WMO/TD No. 1304.

Molini, A., La Barbera, P., Lanza, L. G. & Stagi, L. 2001 Rainfall intermittency and the sampling error of tipping-bucket rain gauges. *Physics and Chemistry of the Earth* **26** (10–12), 737–742.

Sevruk, B., Ondrás, M. & Chvíla, B. 2009 The WMO precipitation measurements intercomparisons. *Atmospheric Research* **92** (3), 376–380.

Shapiro, S. S. & Wilk, M. B., 1965 An analysis of variance test for normality (complete samples). *Biometrika* **52** (3–4), 591–611.

Sheskin, D. J. 2007 *Handbook of Parametric and Nonparametric Statistical Procedures*, 4th edition. Chapman & Hall/CRC, Boca Raton FL, 1736 pp.

Vuerich, E., Monesi, C., Lanza, L. G., Stagi, L. & Lanzinger, E. 2009 *WMO Field Intercomparison of RI Gauges. World Meteorological Organisation – Instruments and Observing Methods Rep. No. 99*, WMO/TD No. 1504, 286 p. (also available from: http://www.wmo.int/pages/prog/www/IMOP/publications-IOM-series.html).

WMO 2008 *Guide to Meteorological Instruments and Methods of Observation*, WMO-No. 8, 7th edition. World Meteorological Organization, Geneva, Switzerland, 593 pp.

First received 4 July 2011; accepted in revised form 9 January 2012