Copia Lorenzo (Orcid ID: 0000-0002-9552-1196)

Wassenaar Leonard I. (Orcid ID: 0000-0001-5532-0771)

Proficiency testing of 78 international laboratories measuring tritium in environmental waters by decay counting and mass spectrometry for age dating and water resources assessment

Lorenzo Copia*, Leonard I. Wassenaar, Stefan Terzer-Wassmuth, Darren J. Hillegonds[1], Philipp M. Klaus, and Luis J. Araguás-Araguás

International Atomic Energy Agency, Vienna International Centre, A-1400 Vienna, Austria

[1]Present Address: University of Oxford, Department of Earth Sciences, South Parks Road, Oxford OX1 3AN, UK

*Corresponding author: L.Copia@iaea.org

Keywords: water, tritium, HTO, ^{3}H, proficiency test, groundwater, age-dating, monitoring

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/rcm.8832

ABSTRACT

RATIONALE: Tritium (^{3}H) is an important hydrological tracer commonly used for over 60 years to evaluate water residence times and water dynamics in shallow/recent groundwaters, streams, lakes and the ocean. We tested the analytical performance of 78 international laboratories engaged in low-level ^{3}H assays for water age dating and monitoring of environmental waters.

METHODS: Seven test waters were distributed by the IAEA to 78 international tritium laboratories. Set 1 included a tritium-free groundwater plus three ultra-low ^{3}H samples (0.5- 7 TU) for meeting groundwater dating specifications. Set 2 contained three higher ^{3}H-content samples (40-500 TU) suitable for testing of environmental monitoring laboratories.

RESULTS: Seventy of the laboratories used liquid scintillation counting with or without electrolytic enrichment, seven utilized ^{3}He accumulation and mass-spectrometry, and one used gas-proportional counting. Only ~50 % of laboratories demonstrated the ability to generate accurate ^{3}H data that was precise enough for water age dating purposes.

CONCLUSIONS: TRIC2018 helped identify recurrent weaknesses and potential solutions. Strategies for performance improvement of ^{3}H laboratories include: a) improved quantification of ^{3}H detection limits and analytical uncertainty, b) stricter quality control practices in routine operations along with care and recalibration of ^{3}H standards traceable to primary NIST standards, c) annual assessment of enrichment factors and instrumental performance, and d) for water age dating purposes the use of electrolytic enrichment systems having the highest possible ^{3}H enrichment factors (e.g. > 50x).

1 INTRODUCTION

Tritium (^{3}H, ½ -life 4500 ± 8 d [1]) is an important radiotracer of water residence times in streams, lakes and soils, seawater, and is a longstanding isotopic tool for determining the age of recent (<60 years old) water in aquifers for replenishment quantification and for water pollution and groundwater vulnerability studies [2,3]. In the hydrological sciences, tritium contents are expressed as the tritium to protium (^{1}H) ratio or the classical "Tritium Unit" (TU) where 1 TU equals one ^{3}H atom in 10^{18} protium atoms; or otherwise as massic activities where 1 TU = 0.11919 ± 0.00021 Bq/kg[4,5]. ^{3}H is found naturally in rainwater stemming from cosmic spallation of nitrogen to form HTO in the upper atmosphere[6], and tritium in freshwater at considerably higher contents may arise locally around some types of nuclear facilities (e.g. power plants, research reactors, and reprocessing facilities) or dispersed at continental scales by nuclear weapons detonation (e.g. global ^{3}H "bomb peak" of the 20th century). Nevertheless, ^{3}H contents in rain and environmental waters nowadays are very low and have largely subsided towards "pre-bomb levels" of <0.5-30 TU in the absence of local anthropogenic sources[7-9]. Accordingly, precise quantification of contemporary levels of ^{3}H for use in hydrological studies or for groundwater age

dating by liquid scintillation (LSC) or gas proportional counting (GPC) methods requires 15 to 100-fold electrolytic pre-concentration of HTO using 250-2000 mL samples in order to obtain sufficient decay counts for accurate and precise results suitable for reliable hydrological interpretations [10,11]. For environmental monitoring purposes, preparative ^{3}H enrichments before LSC counting are usually not conducted because international and World Health Organization (WHO) ^{3}H drinking water limits for human consumption are high and range from 100 to 76000 Bq/kg[12], depending on the country or jurisdiction, and which nevertheless far exceed natural HTO levels in both fresh and marine waters. More recently, high-sensitivity static noble gas mass-spectrometry has been employed to precisely quantify ^{3}H levels by measuring the amount of ^{3}He created via tritium decay in water samples which have been degassed to remove dissolved helium, isolated, and left to accumulate ^{3}He over a prescribed period of time on as little as 10-150 mL [13,14]. Regardless of the sample preparative method or ^{3}H detection instrumentation used, in all applications accurate tritium assays must be demonstrated to provide fit-for-purpose precision or appropriate to the detection limits of the analytical technology used.

Since the 1960s the International Atomic Energy Agency (IAEA, Vienna, Austria) has conducted international ^{3}H analysis proficiency tests, primarily aimed towards high-precision users in the hydrological and groundwater sciences[15-21] (see Table S1, supporting information). These proficiency tests provide an independent assessment of the quality of tritium data generated by laboratories. In this study, the IAEA tested the performance of 78 international laboratories conducting ^{3}H assays for both hydrogeologic and environmental monitoring purposes (Tritium Inter-Comparison 2018 or TRIC2018). The IAEA tests are aimed at ultra-low-level and monitoring assays, as opposed to commercial proficiency tests for monitoring that utilize far higher ^{3}H activities of >1000 TU (e.g. www.npl.co.uk). The participating laboratories largely comprised conventional LSC decay counting operations, with a smaller proportion using the ^{3}He ingrowth method and mass spectrometry or GPC either with or without pre-enrichment sample processing.

TRIC2018 consisted of two discrete sample sets. Set 1 comprised a ^{3}H-free (dead-water) sample and three water samples having ^{3}H contents ranging between 0.5 and 7 TU. These low levels of ^{3}H typically necessitate pre-distillation and pre-enrichment for both LSC and GPC laboratories (optionally for ingrowth).

Set 2 comprised three higher ^{3}H content (40-500 TU) water samples, a range which fell within direct counting (DC) detection limits of modern LSC instruments and did not require any ^{3}H pre-enrichment.

Here we synthesize the results of TRIC2018 and provide an overview of the state of contemporary tritium assays and operations. We identify common problem areas and suggest areas for improvement to enable ^{3}H laboratories to obtain the best possible analytical accuracy and precision for analyses of natural waters for age dating and monitoring purposes.

2 MATERIALS AND METHODS

In preparation for TRIC2018, an online survey of tritium laboratories was conducted in February 2018 asking which range of tritium contents in environmental waters would be appropriate for most users' operational needs, the instrumentation and sample preparation methods used, and the type and amount of water required for routine tritium analyses. Of 74 replies the majority recommended one set of "groundwater-type" samples (i.e. not pre-distilled water) having typical tritium contents of 0-20 TU for those laboratories

conducting pre-distillation, electrolytic enrichment, and high-precision decay counting by LSC and GPC or by ^{3}He ingrowth, and a second optional set of higher ^{3}H content in the 50-500 TU range for laboratories to measure samples directly without any preparative enrichment, or for laboratories not using sample pre-enrichment procedures. These survey responses informed our final sample test composition.

Online registration for TRIC2018 was conducted from March to April 2018; a total of 93 laboratories from 52 countries registered, acknowledging that performance reports would be confidential, and their results could be used anonymously in this report. Participants optionally provided additional metadata including detailed information about laboratory preparative techniques, measurement instrumentation, and procedures. Each participating laboratory was assigned an ID and given 7 months to complete the test with a due date of January 31, 2019. Of the registered laboratories, 70 reported LSC results, 7 reported ^{3}He ingrowth and mass spectrometry results, and one laboratory reported GPC results, for one or both Sets. The remaining 15 laboratories were unable to complete the TRIC test on time for various reasons such as equipment

or instrumental problems. A few laboratories used both LSC and mass spectrometry or multiple enrichment-level capabilities; multiple reports from those laboratories were evaluated as separate submissions. Forty-five of the participating laboratories used preparative distillation and electrolytic enrichment for Set 1. Individual laboratory performance outcome reports were sent to all participants in March 2019.

2.1 Test Sample Preparation

Seven test water samples (T28–T34) were prepared for TRIC2018. All samples were created using a stock of "^{3}H-free" (dead-water) groundwater and adjusted gravimetrically by adding small aliquots of NIST ^{3}H reference standard to achieve required ^{3}H targets. The diluting dead-water was a 500-L stock of water from Vöslauer Well VI (660 m depth), a mineral water obtained from Bad Vöslau, Austria, supplied in 1.5-L plastic bottles. The total dissolved solids (TDS) of this well water was 678 mg L^{-1}. Radiocarbon dating revealed that this groundwater is over 15,000 years old (www.voeslauer.com/web/at/quelle) and was previously determined to be tritium-free (https://secure.umweltbundesamt.at/webgis-portal/isotopen/map.xhtml). Pre-testing of this sample by pre-distillation and electrolytic enrichment using 500-mL samples and LSC at the IAEA Isotope Hydrology Laboratory in Vienna verified

that the ^{3}H content of this groundwater was below the Detection Limit (0.16 TU)[22]. Furthermore, $\delta^{18}O$ values and TDS of Vöslauer samples were measured on 1 out of 6 bottles (pack size) to check for geochemical consistency. The $\delta^{18}O$ values were lower than that of than modern rainfall in Austria affirming a paleo-water origin and in line with the Austrian Water Isotope Database above. For the purposes of TRIC2018 this groundwater was considered "^{3}H-free" and became sample T28 in addition to functioning as a diluting stock for non-zero ^{3}H-content dilutions. Sample T28 was stored in a pressurized closed system stainless steel container with a siphon dispenser in order to eliminate potential ^{3}H isotope exchange with laboratory air moisture.

The remaining TRIC2018 samples T29-T34 were made up using precise serial gravimetric dilutions of Reference Material NIST SRM 4361C (2.009±0.015 Bq/g; 3 September 1998) along with precise gravimetric aliquots of T28 dead-water. To improve gravimetric accuracy, we conducted a series of stepwise intermediate dilutions. Each prepared test sample batch (100 L each for Set 1: 10 L each for Set 2) of T29-T34 was mechanically agitated for two hours in its drum at room temperature under 0.5 bar N_2 pressure. Each drum had a siphon-and-

valve outlet system which allowed dispensating without opening the drum and exposing the test sample to the atmosphere. Set 1 samples were dispensed into labelled 1-L HDPE bottles, whereas Set 2 samples were dispensed into labelled 50-mL HDPE bottles. Each TRIC sample was dispensed in a single session. The reference date of all TRIC test samples for decay calculation purposes was 1 June 2018. All gravimetric data for ^{3}H determinations and uncertainties of test samples were determined in accordance with established guidelines[23,24].

2.2 Establishing TRIC Sample Reference Values

The reference ^{3}H values for the TRIC test samples were established using a gravimetric approach with full uncertainty propagation as summarized in Table 1. Set 1 (T28-T31) contained one ^{3}H-free sample (T28) plus three low-level samples (0-7 TU) intended to be representative of tritium in modern rainfall, seawater, and surface and groundwaters, and which would require distillation and electrolytic enrichment for LSC and GPC counting or be measured directly by the ^{3}He ingrowth method. Set 2 (T32-34) had higher ^{3}H contents (40-500 TU). The rationale for Set 2 was that these samples were intended for all laboratories, including those employing direct counting without enrichment (e.g.,

environmental or nuclear monitoring facilities). Furthermore, sample T32 had a target ^{3}H content of ~40 TU, which represented a low-end ^{3}H abundance for direct counting using modern LSC instruments whereas T33-T34 (100-500 TU) represented the ^{3}H content and range of what the "low-level" samples (Set 1) would typically be electrolytically enriched to, and represent the typical tritium range of low-level ^{3}H laboratory calibration reference standards. Since no ^{3}H electrolytic enrichment was required for samples T32-T34 this allowed a direct assessment of laboratory counting or ingrowth performance.

2.3 Data Evaluation

For laboratory assessment the gravimetrically determined reference values for T29-34 (Table 1) were used as the target tritium values to be achieved by laboratories following international proficiency testing practices, whereby we established our performance to be appropriate for high-precision groundwater age dating purposes according to ISO practices[24,25]. On this basis, we established a "standard deviation for proficiency assessment" (σ_p) for each test sample which was an expert opinion about relevant ^{3}H uncertainty budgets needed to yield sufficiently precise groundwater residence time determinations, a choice

that was consistent with previous IAEA proficiency testing. These standard deviation targets for laboratory assessments are considered "high-performance" benchmarks, but they nevertheless are reasonable for a modern isotope hydrology community particularly in light of recent advances in ^{3}He ingrowth methods and the global return towards low "pre-bomb" ^{3}H levels which require better ^{3}H detection limits for most modern environmental waters.

For each test sample, to determine whether the reported laboratory ^{3}H measurement was sufficiently close to its known value (i.e. accuracy or trueness) a *z*-score was calculated:

$$z = \frac{x - x_a}{\sigma_p}$$

where x was the laboratory-reported tritium content for each sample, x_a was the assigned ^{3}H reference value (Table 1), and σ_p was the acceptable standard deviation for proficiency assessment (Table 1). Accordingly, *z*-scores were considered *Satisfactory* when $-2 \leq z \leq 2$, *Questionable* when $2 < |z| < 3$, and *Unsatisfactory when* $|z| \geq 3$. A sample *z* - score of 0 implied a perfect match (no bias) from the established reference value.

In addition, a zeta-score (ζ') was calculated for each laboratory ^{3}H measurement, enabling an assessment of the laboratory's reported value while incorporating their reported measurement uncertainty:

$$\zeta' = \frac{x - x_a}{\sqrt{u^2(x) + u^2(x_a)}}$$

where x was the laboratory reported TU content, x_a was the reference value (Table 1), $u(x)$ was the reported one-sigma uncertainty of the laboratory, and $u(x_a)$ was the combined one-sigma uncertainty of the assigned value (Table 1). All ζ' scores were evaluated in a similar way to *z*-scores where *Satisfactory* was $|\zeta'| \leq 2$, *Questionable* when $2 < |\zeta'| < 3$, and *Unsatisfactory* when $|\zeta'| \geq 3$, but only for a subset of submissions where the uncertainty budget of the participating laboratory could be verified (e.g. some laboratories did not provide uncertainties or gave blanket uncertainty). Accordingly, we graded the zeta-scores as *Satisfactory / Questionable / Unsatisfactory* to make general conclusions concerning laboratories reporting statistically reasonable data[25].

To produce a graphical visualization of laboratory performance combining *z*-score (bias) and zeta tests (uncertainty), we created a variation of the PomPlot [26] whereby the scaling factor was the standard deviation for proficiency

assessment, σ_p, instead of the median absolute deviation (MAD) from a reference value, as per the original paper. This allowed us to plot all ^{3}H test sample results on identical plots regardless of their ^{3}H content. The z values were displayed on the horizontal axis, while on the vertical axis the laboratory-reported uncertainty (u) for each sample was shown relative to the acceptable performance uncertainty for that test sample (u/σ_p). The ζ-values ±1, ±2 and ±3 are depicted using solid lines, resulting in a pyramidal structure. A visual interpretation of this PomPlot modification is displayed in Figure 1. Briefly, positive or negative biases from the sample ^{3}H reference value fall on either side of zero along the x-axis. Uncertainty relative to the established acceptable σ_p for each test sample increases down the y-axis. For example, a value of $u/\sigma_p = 1$ in the y-axis meant that the laboratory-reported uncertainty for that sample was equal to the maximum acceptable uncertainty for that ^{3}H sample (see Table 1), and <1 meant that the reported uncertainty was better than the minimum acceptable limit. A y-axis value of 2 and 3 meant twice and triple the maximum acceptable uncertainty for each test sample and so on, bearing in mind the acceptable bounds circumscribed by the ζ lines. Exceptional and acceptable

laboratory performance for groundwater dating purposes lies within the centre top of the graphical pyramid and square to a maximum value of u/σ_p of 2. It was possible to obtain accurate ^3H results (inside the green dashed lines) but with very high analytical uncertainty (e.g. y=7; seven times acceptable SD of Table 1), which would fall in the lower shaded part of the plot. Precise but inaccurate results plotted near the top of the graph but to the right or left side ($z>|2|$) outside the green dashed lines. Notably, accurate results with high uncertainty at these low ^3H levels may be considered acceptable for environmental monitoring purposes, but not for ground water age dating purposes.

Graphical results of laboratory results by technology for samples T29-T34 were depicted using cumulative *z*-scores for all samples to reveal overall laboratory performance per TRIC Set. Conventional S-plots for each TRIC test sample by individual laboratories and their code are found in the online Supporting Information. In all Figures, counting laboratories are depicted with circles (LSC) or diamonds (GPC) and mass spectrometry laboratories with squares. Dashed lines show the *z*-score boundaries of *Satisfactory*, *Questionable* and

Unacceptable performance. For all graphical plots extreme outliers were removed, however, the full TRIC dataset including extreme outliers are found in the online Supporting Information.

3 RESULTS

3.1 Dead-Water Sample

Reported results of "dead-water" sample T28 are summarized in Table 2, with some qualifications. This sample served several simultaneous but not mutually exclusive purposes for evaluation of the laboratory results. Primarily, a ^{3}H-free sample allowed for possible detection of contamination in the laboratory, which could occur at several steps in the sample preparation or electrolytic enrichment process (contaminated labware or other exposure, carryover memory in the enrichment cells, etc.), and therefore served as a laboratory procedural "blank". T28 also allowed us to evaluate how the tritium community reported ^{3}H data and uncertainty for samples falling below a laboratory Detection Limit (DL), but without providing any guidelines on how to report such low or "zero" activity levels. While T28 was presumed ^{3}H-free based on its deep groundwater source, we could not guarantee "zero" ^{3}H activity because of potential for contamination in its extraction, bottling, storage and shipping by the supplier. Furthermore, measurements of T28 at the IAEA could only establish that this sample was below the detection limit of that counting laboratory (0.16 TU)[22]. Hence, it was decided that T28 could not be reliably used for laboratory performance evaluations using *z*-scores and zeta scores.

There was considerable diversity in how laboratories reported their results of dead-water sample T28, as shown in Table 2. While we did not request the tritium DL for each laboratory as a mandatory field in the reporting form, sample T28 gave insights into the

various ways that laboratories reported ^{3}H sample results falling below their own detection limits. In the tabulated results we found 88 % of laboratories reported a numerical value for their ^{3}H detection limit (DL), with 12 % not reporting any DL. Of those reporting T28 as part of Set 1, tabulated in Table 2, 32 % reported a TU value along with their stated DL. Another 20 % reported T28 as being below a specific DL, and 16 % reported that T28 was "zero" or "not detected" and stating a numerical value for their DL. The remaining laboratories were unspecific, for example providing a result but not specifying a DL (24 %) or giving negative TU values (7 %). All reporting approaches that indicated detection limit awareness (except reporting negative TUs) seemed reasonable in the absence of instructions on how to coherently report "non-detectable" ^{3}H values.

We also considered the possibility that T28 was not truly ^{3}H-free during our post-assessment of the laboratory samples. To assess this, we selected a subset of laboratories which provided a specific DL for their laboratory. We then introduced a very low value of 0.03±0.01TU for T28 into our gravimetric dilution calculations and re-processed their laboratory evaluations. The differences were negligible and did not impact the outcomes of T28 or any other test samples.

3.2 Set 1 - Low-Level Samples T29-T31

The results for low-level samples T29 to T31 are summarized in PomPlots in Figure 2 (left panel). From a graphical inspection, it was apparent for T29 that the overall population presented a strong asymmetry of positive bias, i.e. higher massic activities than the reference value for this sample. This could also be seen in TU values for each sample and in the S-plots (Supporting Information). To quantify this asymmetry for T29, we found the percentage of laboratories giving high biased results (z >2: 39 %) was higher than those laboratories that gave low-biased results (z<-2: 5 %). This asymmetry towards positive bias gradually faded

with increasing ^{3}H for samples T30 and T31 (Figure 2). For T31, results were more evenly split with 23 % of laboratories giving high-biased versus 19 % low-biased values.

The most obvious explanation for high bias was found in the laboratory-reported metadata, especially the reported detection limit as well as the enrichment factors of the electrolytic enrichment system, which was in turn closely related to the enrichment cell volume. The ^{3}H activity concentration of T29 was close to the detection limit of a typical configuration of a commercial LSC coupled with a 250-mL volume electrolytic enrichment system (DL ~ 0.6 TU). The majority of LSC laboratories reported results for Set 1 without mentioning a DL, which posed serious questions on if or how a DL was applied, and whether this reporting approach was used routinely in their laboratory results.

Considering Set 1 (Figure 2), it was found that many laboratories reported unacceptably high uncertainties ($u/\sigma_p > 3$) for ground water age dating purposes. The majority of these were the LSC labs: T29 - 30%, T30 - 34%, T31 - 23%. This is was in part due to differences in the purpose of the analyses (e.g. monitoring labs) with counting procedures (mainly counting times) that were inadequate, but also with unclear uncertainty estimations. The single GPC laboratory (diamond) gave satisfactory results for T29-T31.

The high numbers of gross outliers - not depicted in Figure 2 for T29, T30 and T31 - were mainly generated by 7 direct counting labs that inappropriately attempted to measure Set 1. The mass spectrometry ingrowth laboratories produced surprisingly variable results for Set 1 (see below). Particularly for sample T31 (Figure 2) only 2 out of 7 MS laboratory results were accurate with acceptable precision.

3.3 Set 2- Higher-Level Samples T32-T34

Surprisingly, for the higher tritium content samples T32-T34 (41-500 TU) which did not require any pre-enrichment the outcomes were similarly as unexpected as for the LSC laboratories. A significant fraction of LSC laboratories (30 %) produced unacceptably biased results ($|z|>2$) and/or reported unacceptably high uncertainty ($u/\sigma_p>3$). Contrary to Set 1, a symmetric distribution with respect to z values was observed in the PomPlot for the three samples of Set 2. Similarly, a lower fraction of laboratories (12-27 %) reported unacceptable high uncertainty ($u/\sigma_p>3$). The single GPC laboratory produced acceptable results for all Set 2 samples. Most of the MS laboratories did not measure Set 2, stating that the tritium concentrations were too high for their methods or in case of T32 sample loss during the analytical procedure. The few that did produced acceptable results with one exceptional outlier (T34). The range of reported TU values varied widely for Set 2, spanning 18-140 TU for T32 (40 TU), 70-300 TU for T33 (120 TU) and 350-> 700 TU for T34 (500 TU).

These results again demonstrated clear differences in LSC laboratories measuring tritium for different purposes; even if the results were "accurate within the reported uncertainty" (zeta pyramid) they were not acceptable for hydrological purposes. Notably T34 is in the typical range of low-level ^3H laboratory calibration reference standards. Similarly, several extreme outliers not depicted in Figures 2 and 3 for T29, T30 and T31 were mainly LSC laboratories utilizing counting times <100 minutes or using LSC counters that are not suited for this type of application.

3.4 Overall Performance for Set 1 or 2

To assess the overall comparative performance of participant laboratories we plotted and ranked their sum of z-scores ($\Sigma|z|$) for a subset of laboratories which had reported both Set 1 or Set 2, as depicted in Figure 3. The best overall performing laboratories had low $\Sigma|z|$ values, with performance sorted in decreasing order from left to right. At or below the green dashed line denoted overall *Satisfactory* performance $\Sigma|z|<6$ for both Set 1 and Set 2; and any results falling above the orange line were considered *Unsatisfactory*, i.e. $\Sigma|z| \geq 9$ for Set 1 or Set 2, with results falling between these two lines were considered *Questionable*. The sum of the absolute values of the z score ($\sum|z|$) for Set 1 and Set 2 was determined only if the results of all the samples were submitted. For Set 1 (Figure 3, upper panel), 51 labs (86 %) fully reported Set 1, and of these only 45 % were Satisfactory, with 55 % of laboratories producing questionable or unsatisfactory results. Of the 70 laboratories that reported data for Set 2, only 54 % were Satisfactory (Figure 3, lower panel). We acknowledge that the definition of *Acceptable* in this test may not encompass the purpose of each laboratory, and that our criteria are intended to critique and rank participating laboratories in a uniform manner in the context of isotope hydrology and environmental monitoring.

Considering only the larger LSC community, among laboratories that produced unsatisfactory results for Set 1, 15 (around 58 %) were also unsatisfactory for Set 2,

demonstrating that inaccuracies were possibly present at different steps of the entire LSC analytical procedure, from the storage of calibration standards to the counting strategy, which affect overall performance. This was verified later by poorly performing laboratories in a Post-Testing Survey.

3.5 Mass Spectrometry Laboratories

Considering only MS laboratories utilizing the ^{3}He ingrowth method, a positive bias was observed for most samples and was particularly evident for sample T31 (Figure 2). To elaborate further, the results from each MS lab for Set 1 samples were plotted on a single modified PomPlot (Figure 4). Four out of seven MS labs showed a positive bias for all three samples in Set 1. Moreover, there were puzzling results: for example, one lab NYEH (green triangle) obtained satisfactory z scores but had unsatisfactory zeta scores, which suggested under-estimation of the true analytical uncertainty. MVWK and NYEH (red square and green triangle, respectively) showed a linear dependency of *z* and zeta results corresponding to increasing tritium massic activity of the test samples, failing fit-for-purpose requirements in the case of the highest TU sample of the set. It was impossible to extrapolate a general behaviour or a statistically meaningful result from this small specialized group of mass spectrometry analyses, but these results should trigger a deeper reflection on crucial analytical aspects (e.g. blanks) and other potential sources of systematic errors for the ^{3}He ingrowth and mass spectrometry method.

3.6 Post-Testing Survey

A post-TRIC survey conducted a few months after the performance results were sent out resulted in 31 (40 %) responses from participating laboratories (27 LSC, 3MS, 1 GPC). Nineteen of these LSC laboratories reported they had located possible explanations for the unsatisfactory performance, 85 % reported finding issues related to different stages of the counting procedure, such as improperly calibrated laboratory standards or incorrect LSC

counter settings. Four laboratories found problems with their electrolytic enrichment procedures, some in combination with LSC counter issues. One laboratory reported mistakes with the post-processing calculations. One MS lab reported possible issues related to the concentration of ^{3}He accumulated.

4 CONCLUSIONS

The measurement and interpretation of contemporary low levels of ^{3}H in nature, and the importance of tritium monitoring as a legal obligation in many countries, require that laboratories conducting ^{3}H measurements in environmental samples can produce reliable, accurate, and legally defensible results. Inter-comparison tests like TRIC help laboratories to identify prevailing and current weaknesses and potential solutions. The data synthesis and evaluation of TRIC2018 identified three key areas for improving the performance of tritium laboratories, and particularly for those conducting hydrological applications:

1) Improve quantification and reporting of the instrumental Detection Limit as well as analytical uncertainty.
2) Improve overall Quality Control practices by using Control samples in routine operations, proper storage and handling of ^{3}H standards, regular recalibration of laboratory ^{3}H standards traceable to primary NIST standards, and annual assessment of enrichment factors and instrumental performance.

3) To achieve the required performance for age-dating purposes in the ^{3}H range <2 TU, the adoption of electrolytic enrichment systems with the highest possible enrichment factor (e.g. > 50) is likely to provide a substantial performance improvement for water samples in this low-level range.

Ultimately, every laboratory conducting ^{3}H measurements needs to be capable of adequately stating their measurement ranges and limits; awareness needs to be fostered about the correct choice of analytical procedure (enrichment, counting time, etc.) appropriate and fit for purpose.

Acknowledgements

We thank the participating laboratories for their contributions. We thank Ms L Monteiro for assistance with graphics. Tim Chavez assisted with sample shipments. Funding for TRIC2018 was provided by the IAEA.

References

1. Lucas LL, Unterweger MP. Comprehensive review and critical evaluation of the half-life of Tritium. *Journal of research of the National Institute of Standards and Technology.* 2000;105(4):541.
2. Kaufman S, Libby WF. The Natural Distribution of Tritium. *Phys Rev.* 1954;93(6):1337-1344.
3. Gleeson T, Befus KM, Jasechko S, Luijendijk E, Cardenas MB. The global volume and distribution of modern groundwater. *Nat Geosci.* 2016;9(2):161-167.
4. Gröning M, Rozanski K. Uncertainty assessment of environmental tritium measurements in water. *Accreditation and Quality Assurance.* 2003;8(7-8):359-366.
5. Rozanski K, Gröning M. Tritium assay in water samples using electrolytic enrichment and liquid scintillation spectrometry. *Quantifying uncertainty in nuclear analytical measurements IAEA-TECDOC-1401, IAEA, Vienna.* 2004:195-217.
6. Rozanski K, Gonfiantini R, Araguas-Araguas L. Tritium in the global atmosphere: Distribution patterns and recent trends. *Journal of Physics G: Nuclear and Particle Physics.* 1991;17(S):S523.
7. Morgenstern U, Taylor CB. Ultra low-level tritium measurement using electrolytic enrichment and LSC. *Isot Environ Health Stud.* 2009;45(2):96-117.
8. Palcsu L, Morgenstern U, Sültenfuss J, et al. Modulation of Cosmogenic Tritium in Meteoric Precipitation by the 11-year Cycle of Solar Magnetic Field Activity. *Scientific Reports.* 2018;8.
9. GNIP. Global Network of Isotopes in Precipitation. International Atomic Energy Agency, Vienna. 2020; https://nucleus.iaea.org/wiser, 2020.

10. Cameron J, Payne B. Apparatus for concentration and measurement of low tritium activities. Paper presented at: Proc. of the 6th lnt. Conf. on radiocarbon and tritiuln dating. Washington State University, Pullman. WA, USA.1965.
11. Wassenaar LI, Han LF, Schiefer T, Kainz G, Araguas-Araguas L, Aggarwal PK. A simple polymer electrolyte membrane system for enrichment of low-level tritium ((3)H) in environmental water samples. *Isot Environ Health Stud.* 2018;54(3):274-287.
12. Commission CNS. Standards and Guidelines for Tritium in Drinking Water (INFO-0766). 2014.
13. Clarke WB, Jenkins W, Top Z. Determination of tritium by mass spectrometric measurement of 3He. *The international journal of applied radiation and isotopes.* 1976;27(9):515-522.
14. Schlosser P, Stute M, Dörr H, Sonntag C, Münnich KO. Tritium/3He dating of shallow groundwater. *Earth Planet Sci Lett.* 1988;89(3-4):353-362.
15. Cameron J. Survey of systems for concentration and low background counting of tritium in water. Paper presented at: Radioactive dating and methods of low-level counting. Proceedings of a symposium1967.
16. Florkowski T, Payne B, Sauzay G. Interlaboratory comparison of analysis of tritium in natural waters. *The International Journal of Applied Radiation and Isotopes.* 1970;21(8):453-458.
17. Taylor C. Interlaboratory comparison of low-level tritium measurements in water. *The International Journal of Applied Radiation and Isotopes.* 1978;29(1):39-48.
18. Hut G. Intercomparison of low-level tritium measurements in Water. In. Vienna: International Atomic Energy Agency; 1987:1-51.

19. Ostlund G, Scott E, Taylor C. Fifth IAEA intercomparison of low-level tritium measurements in water. *Report, IAEA, Vienna.* 1995.
20. Gröning M, Taylor C, Winckler G, Auer R, Tatzber H. Sixth IAEA intercomparison of low-level tritium measurements in water (tric2000). In. Vienna: International Atomic Energy Agency; 2001.
21. Gröning M, Tatzber H, Trinkl A, Klaus P, Duren Mv. Eighth IAEA international interlaboratory comparison of the determination of low-level tritium activities in water (TRIC2008) In. Vienna: International Atomic Energy Agency; 2008:43.
22. ISO. ISO 11929: Determination of the Characteristic Limits (decision Threshold, Detection Limit and Limits of the Confidence Interval) for Measurements of Ionizing Radiation: Fundamentals and Application. In: International Organization for Standardization; 2010.
23. ISCO/IEC. Guide to the expression of uncertainty in measurement. . In: International Organization for Standardization; 1995:1-101.
24. ISO. ISO/IEC 17043: Conformity assessment–General requirements for proficiency testing. In: International Organization for Standardization; 2010.
25. ISO. ISO 13528: Statistical methods for use in proficiency testing by interlaboratory comparisons. In: International Organization for Standardization; 2015.
26. Spasova Y, Pommé S, Wätjen U. Visualisation of interlaboratory comparison results in PomPlots. *Accreditation and Quality Assurance.* 2007;12(12):623-627.

Table 1. TRIC2018 Reference Values as determined by gravimetry, including the reference date for decay corrections, expanded uncertainties, and the target uncertainty (σ_p) used for evaluating laboratory performance. The reported expanded uncertainty is based on the combined uncertainty multiplied by a coverage factor k = 2, with a confidence level of approximately 95%.

Sample	Reference Date	Reference ^{3}H Content and Uncertainty (TU)	Uncertainty Margin for Evaluation in TU (σ_p)
T28	2018-06-01	0	*
T29	2018-06-01	0.499 ± 0.004	±0.1
T30	2018-06-01	2.001 ± 0.017	±0.1
T31	2018-06-01	7.007 ± 0.060	±0.21
T32	2018-06-01	40.82 ± 0.35	±2.0
T33	2018-06-01	120.0 ± 1.0	±3.6
T34	2018-06-01	499.8 ± 4.1	±7.5

*- not applied

Table 2. Reporting practices for ^{3}H-free (dead-water) sample T28. A total of 59 laboratories (80 %) provided a detection limit, whereas 20 % did not report any laboratory specific detection limit (DL). The breakdown of reporting practices for sample T28 are tabulated below.

Laboratories	Reporting of Dead Water Sample T28
32 %	Reported a numerical TU value as well as a laboratory DL in TU.
20 %	Reported a "< X" with X a numerical TU value.
24 %	Reported a result without stating a laboratory DL.
17%	Reported T28 as "zero TU" or "Below DL" or "Not Detected" with a laboratory DL in TU
7 %	Reported T28 as a negative TU value.

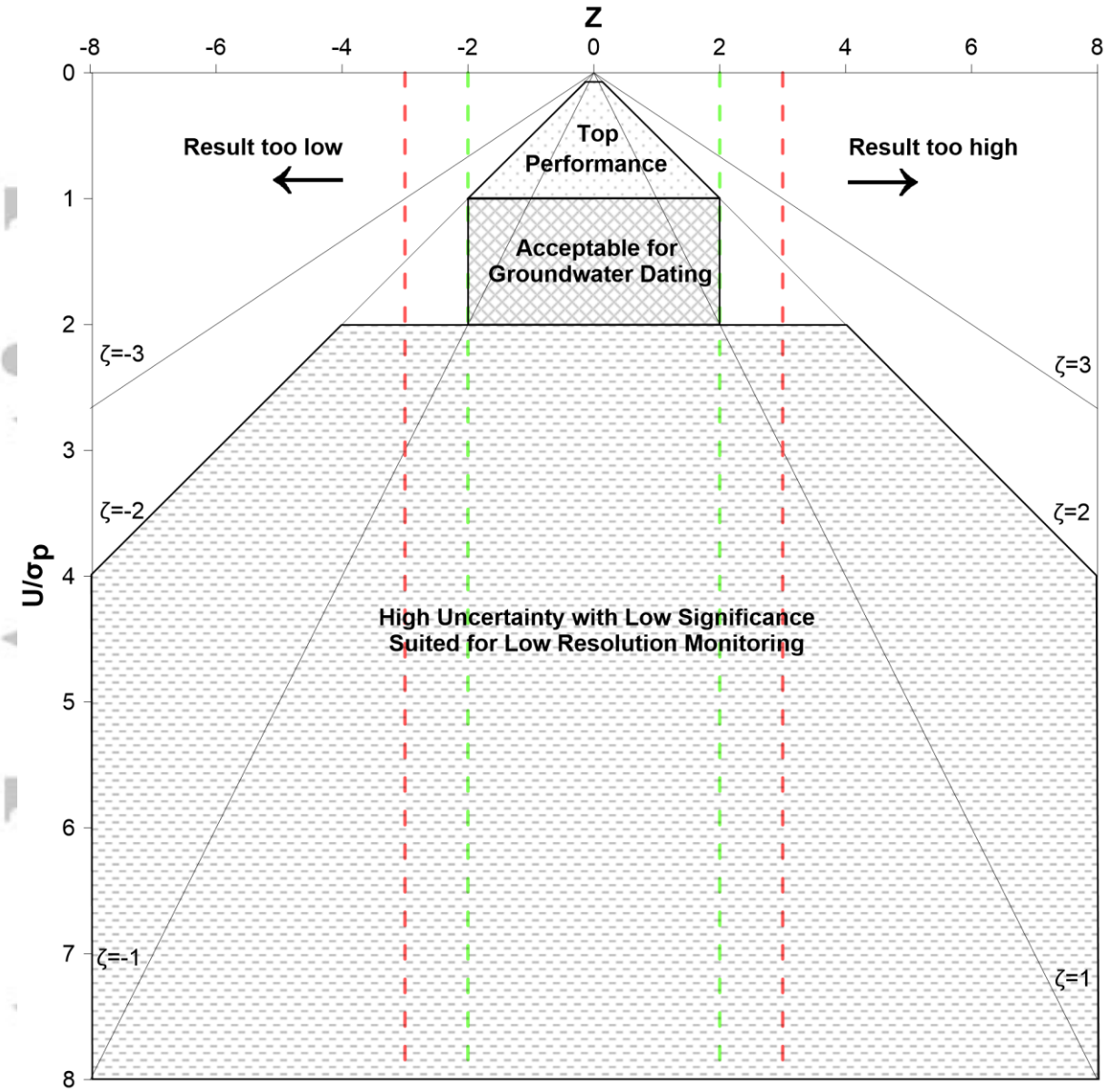

Figure 1. Modified PomPlot plot for evaluating laboratory performance of TRIC samples T29-34 incorporating laboratory reported uncertainty.

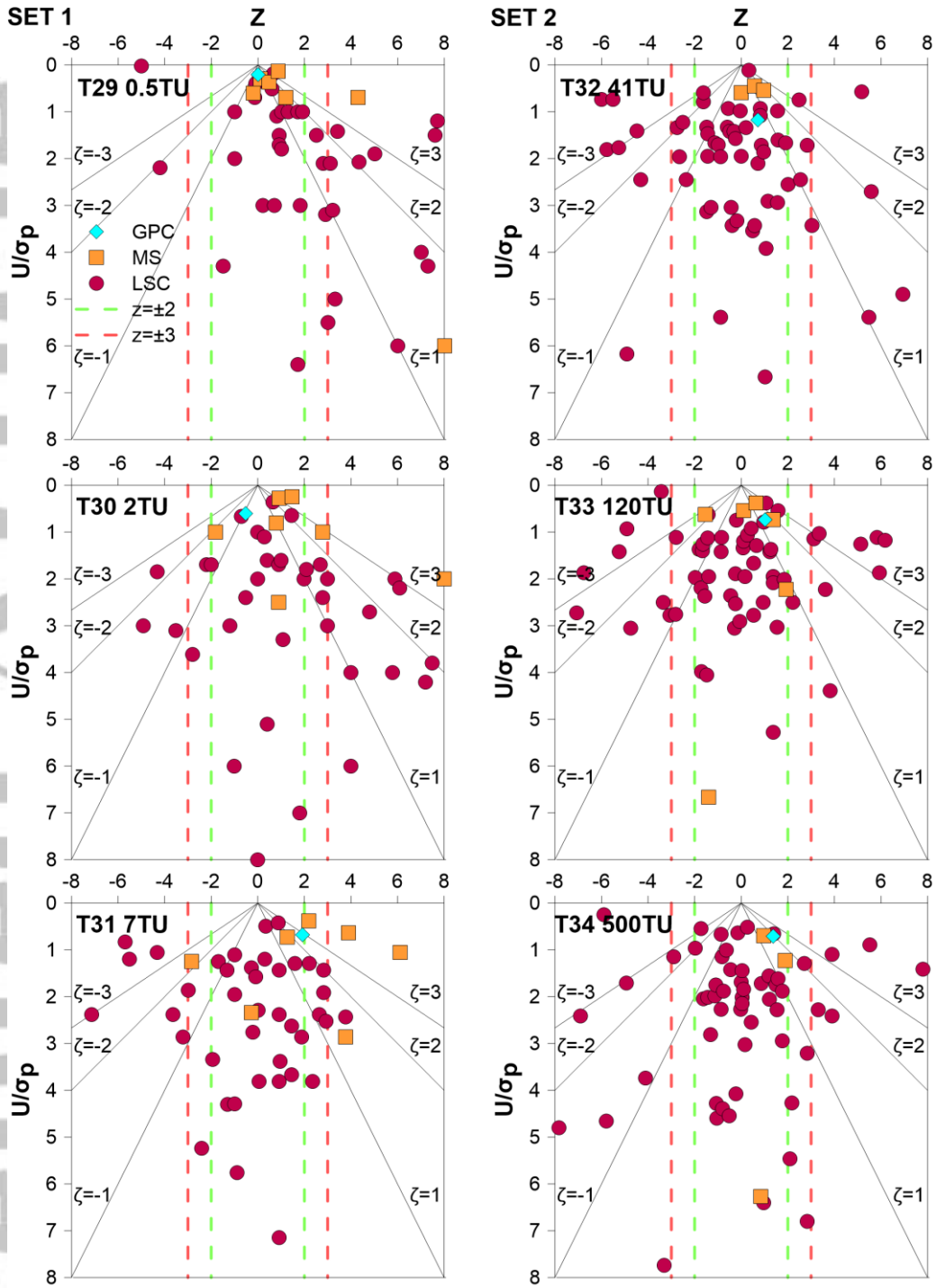

Figure 2. PomPlots for results of TRIC samples for Set 1 (left side) and Set 2 (right side). Symbols denote LSC (circles), MS (squares) and GPC (diamond) laboratories. Extreme outliers not depicted (see Supplemental Materials).

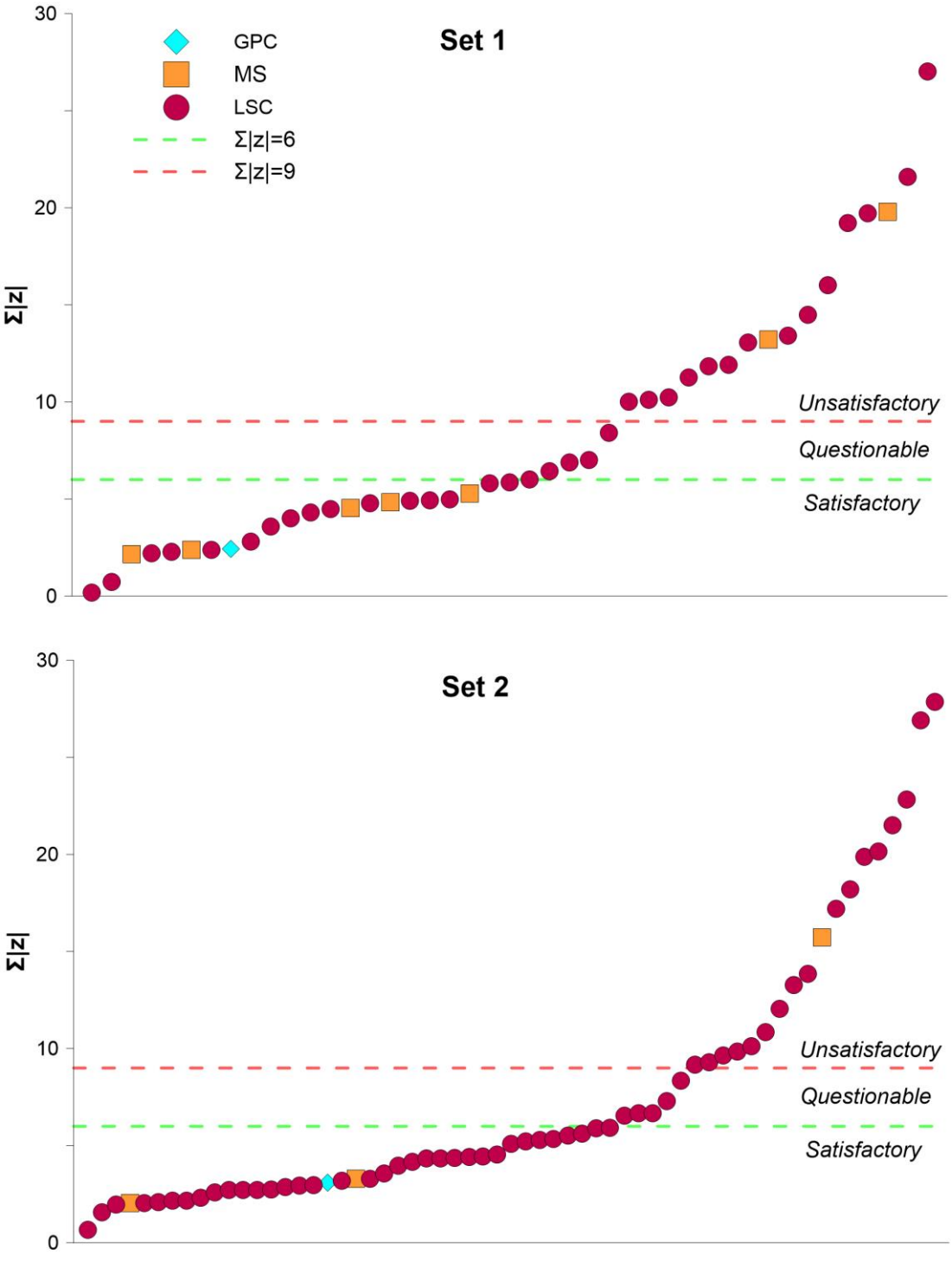

Figure 3. Summed z-scores for TRIC Set 1 (upper) and Set 2 (lower) samples. Symbols denote LSC, MS and GPC laboratories. Extreme outliers not depicted (see Supplemental Materials).

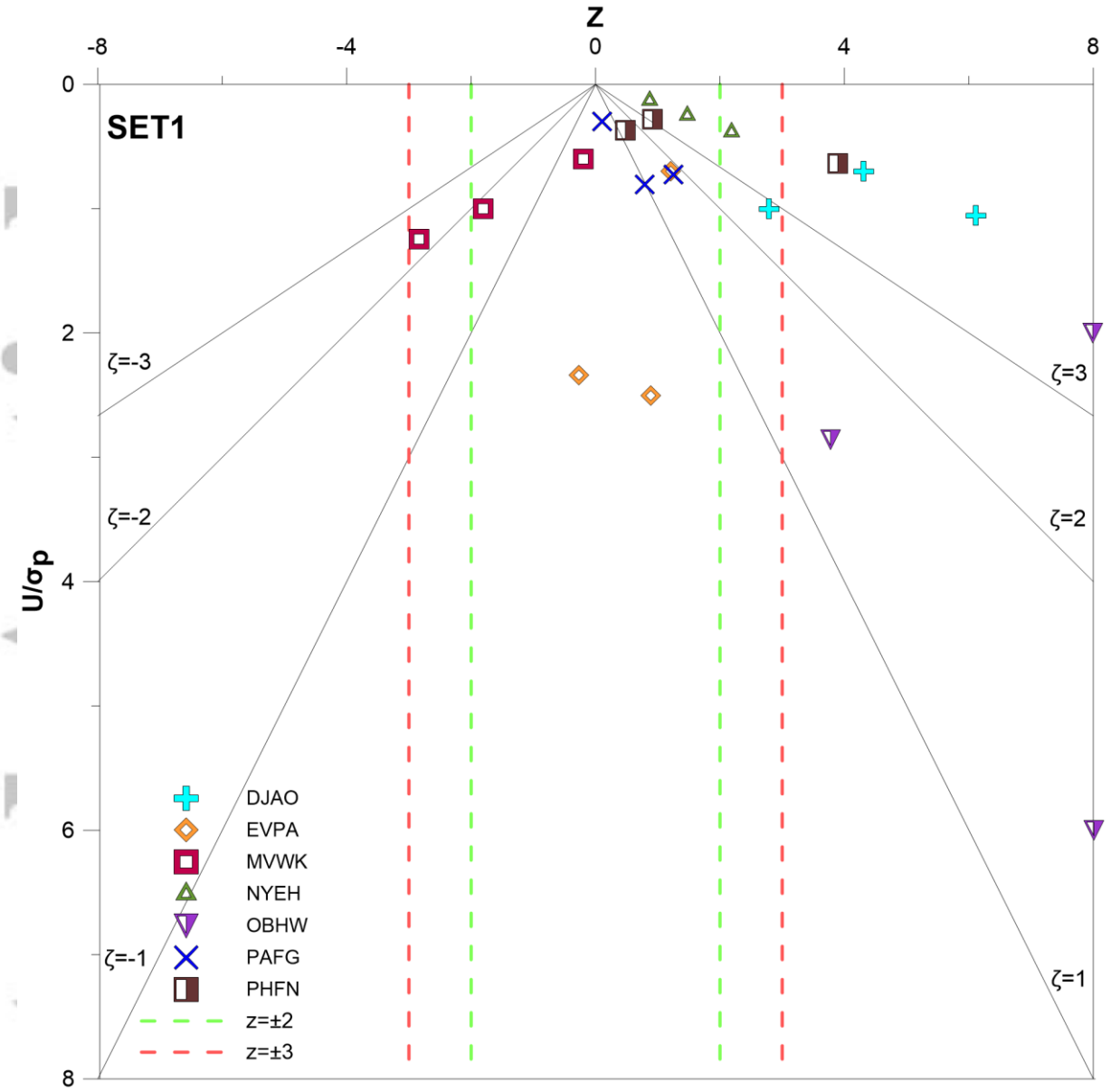

Figure 4. TRIC results for MS labs utilizing 3He ingrowth method for the Set 1 samples. Each symbol denotes an individual MS laboratory.