SnT 2023

CTBT: SCIENCE AND TECHNOLOGY CONFERENCE

HOFBURG PALACE - Vienna and Online

19 TO 23 JUNE

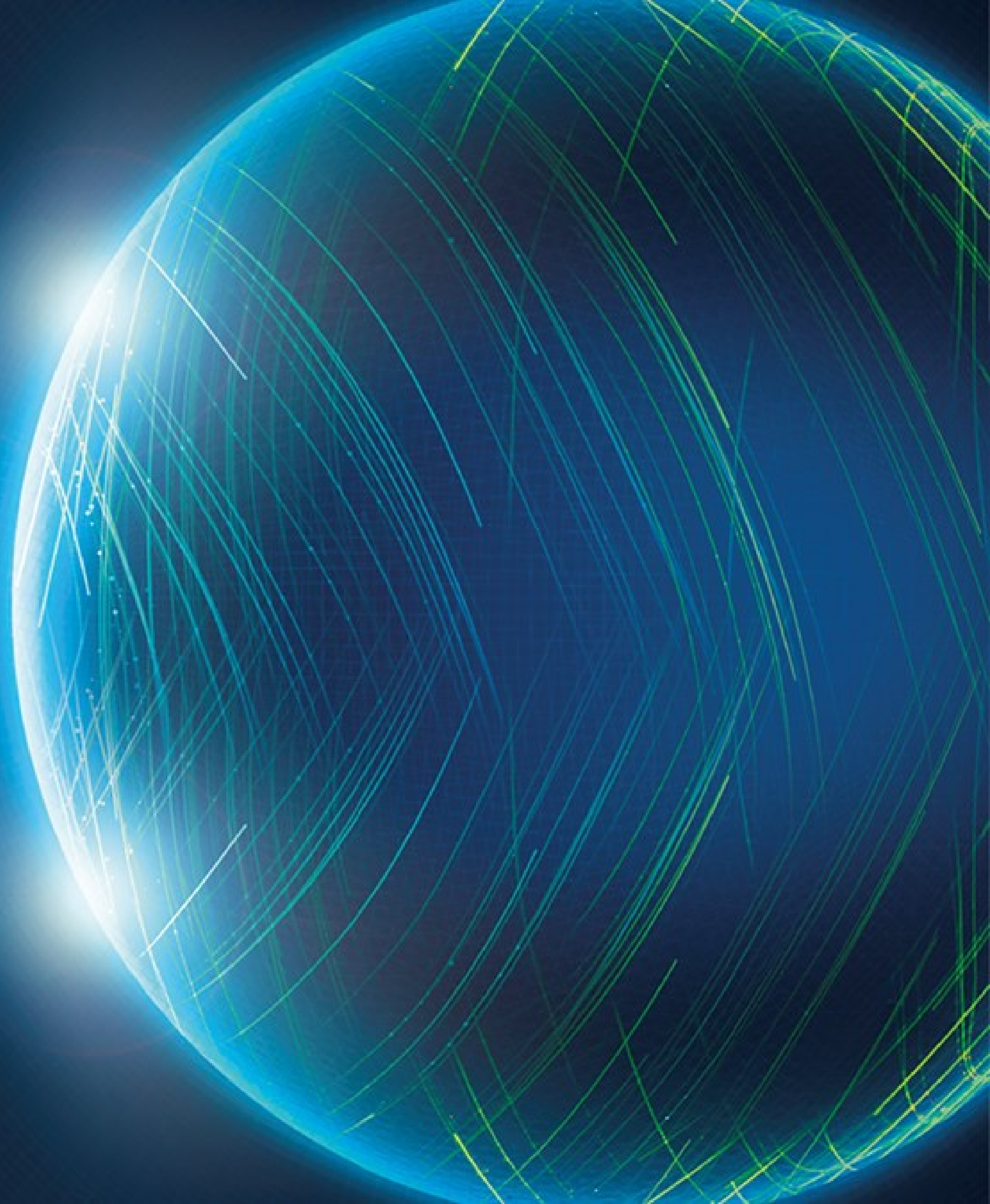
1st Nuclear Explosion Signal Screening Open Inter-Comparison Exercise 2021

C. Maurer[1], B. Liu[2], J. Brioude[3], D. Arnold Arias[1], Y. Kijima[2], J. Kúsmierczyk-Michulec[2], R. Schoemaker[2], A. Tipka[2], and M. Kalinowski[2]

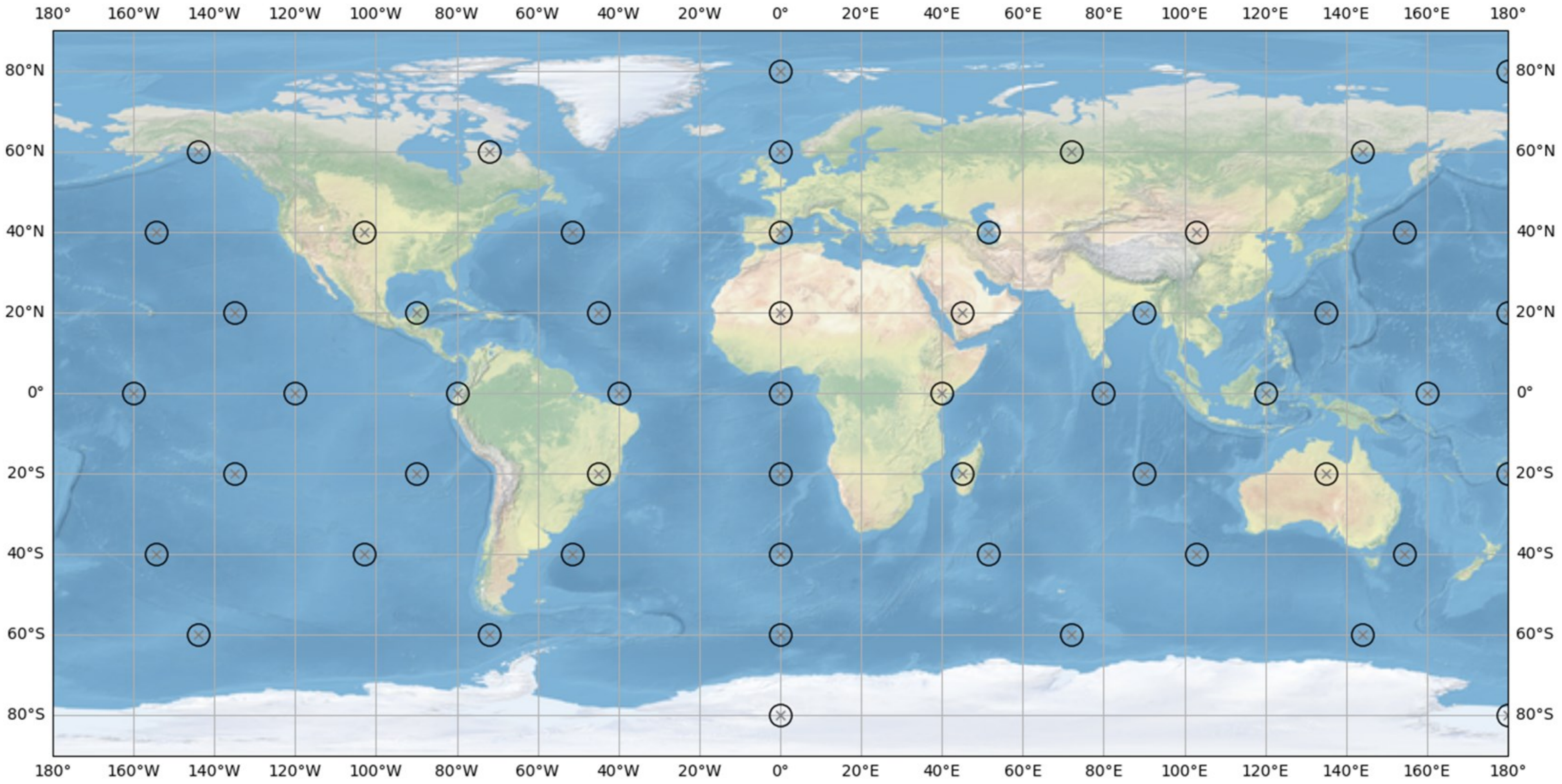
[1]GeoSphere Austria (former ZAMG), [2]CTBTO-IDC, [3]Université de la Réunion

I2.3-046

Presentation Date: 21 06 2023



1. Explosion scenarios



- **424 scenarios: 8 date-times and 53 locations**
- **136 1 kt underground explosions: 24 hours containment, 10% venting** (7.76E14 Bq Xe-133, 6.54E13 Bq Xe-133m, 3.52E11 Bq Xe-131m, 5.84E15 Bq Xe-135 (IDC source term))
- **288 1 kt underwater explosions: prompt 0.92% venting** (3.25E11 Bq Xe-133, 2.32E12 Bq Xe-133m, 2.41E8 Bq Xe-131m, 1.75E14 Bq Xe-135 (*Burnett et al., 2020* source term))
- **23 IMS stations** with data as of 2014, **explosion signals added on top of civil background**

2. Level 1 evaluation

- **Required participants' expertise:** *ATM (of civil sources) only*

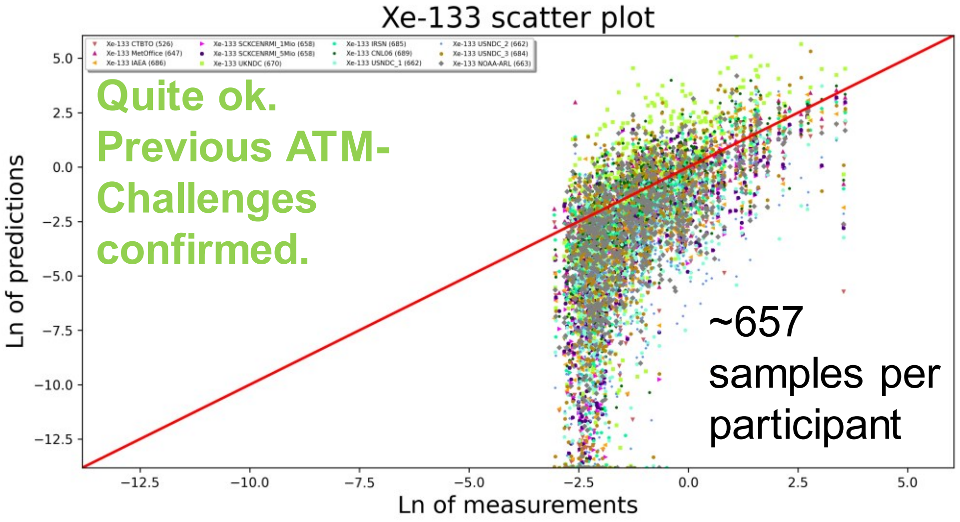
- **Question 1:** *"Is an isotopic measurement an anomaly (regardless of what has caused it)?"* **->**
  1. Filter the test data set according to LC.
  2. Evaluate **distributions 1) of (pseudo-)observations** and **2) of residuals between (pseudo-) observations and participant's predictions based on supplied source terms and participant's ATM method subtracting only a value > 0 for observations >= MDC ("hybrid approach")** per IMS station and scenario in the test data. 1) serves as reference for 2).
  3. Claim a detection if a certain percentile value is exceeded for a sample.
  4. Calculate true positive and false positive rates (TPRs & FPRs) per isotope based on **A) *positives & negatives*** (default) and **B) additionally *excluding positives ("neutrals")* if the mere isotopic test signal is > 0 but < LC.**

- **Question 2:** *„Has an underground or underwater nuclear explosion to be assumed based on isotopic ratios?"* **->**
  Based on all claimed (true and false) multi-isotope positives according to detection power evaluation evaluate TPRs and FPRs for:
  1. three and four radioxeonon isotope discrimination relations (*Kalinowski et al., 2010*)
  2. radioxenon isotope pairs according to Bayesian limits (*Zaehringer and Kirchner, 2008*)

- **Question 3:** *„Can we determine the release time +/- uncertainty within a predefined time window?"* **->**
  Calculate timing success rates based on Bateman equations and single samples which where found to be true positives after detection and screening power evaluation and based on a 10% tolerance criterion.

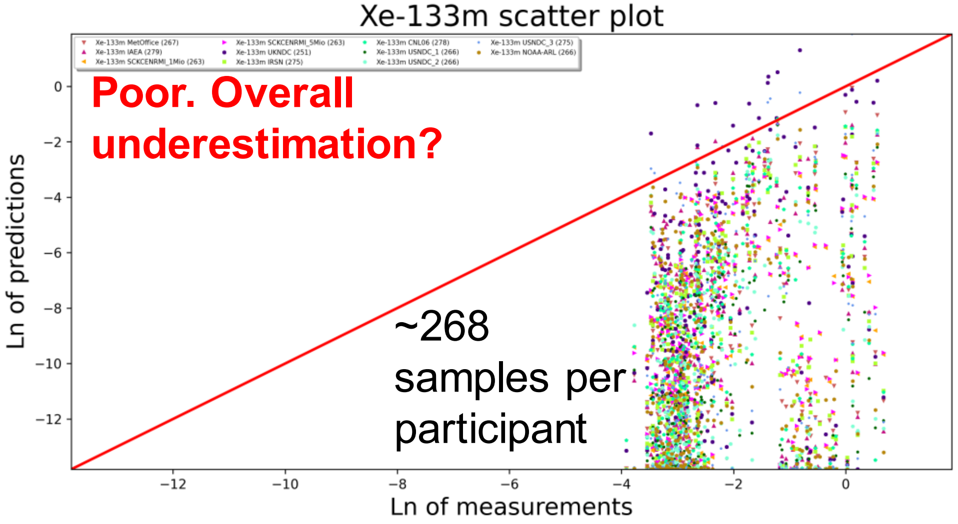
3.1.1. Global isotopic performance of background prediction: Samples >= LC

Thresholding on residuals: Missing (underestimation & data sparsity) and problematic (TPR is sensitive to overestimation) residual impacts…



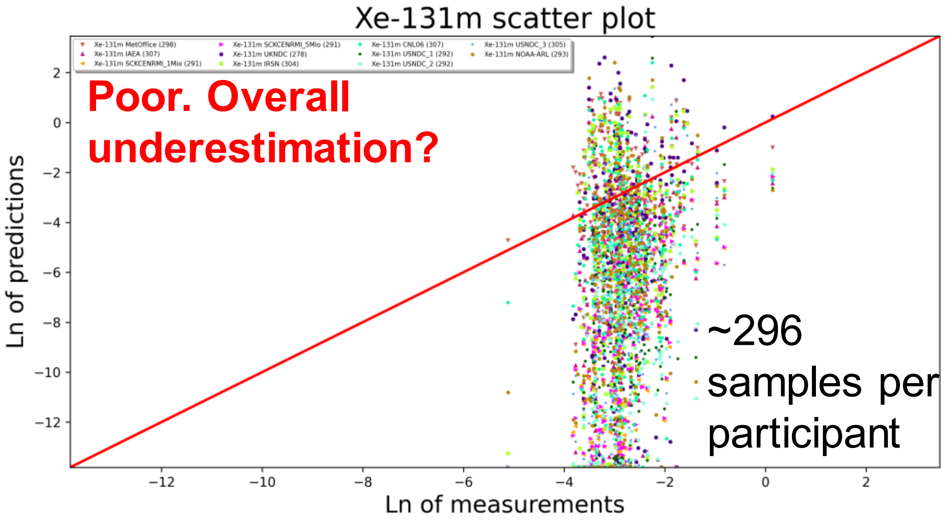
Quite ok. Previous ATM-Challenges confirmed.

~657 samples per participant



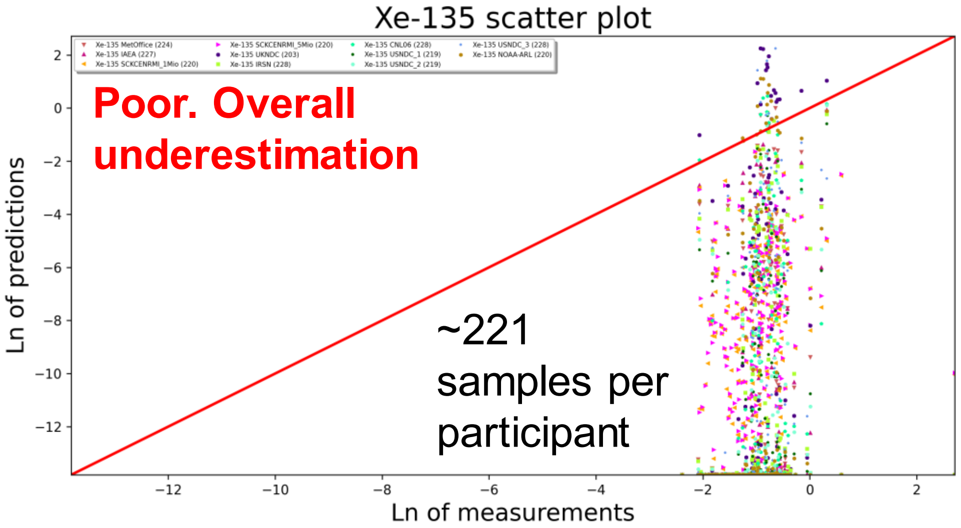
Poor. Overall underestimation?

~268 samples per participant



Poor. Overall underestimation?

~296 samples per participant



Poor. Overall underestimation

~221 samples per participant

IMPORTANT:

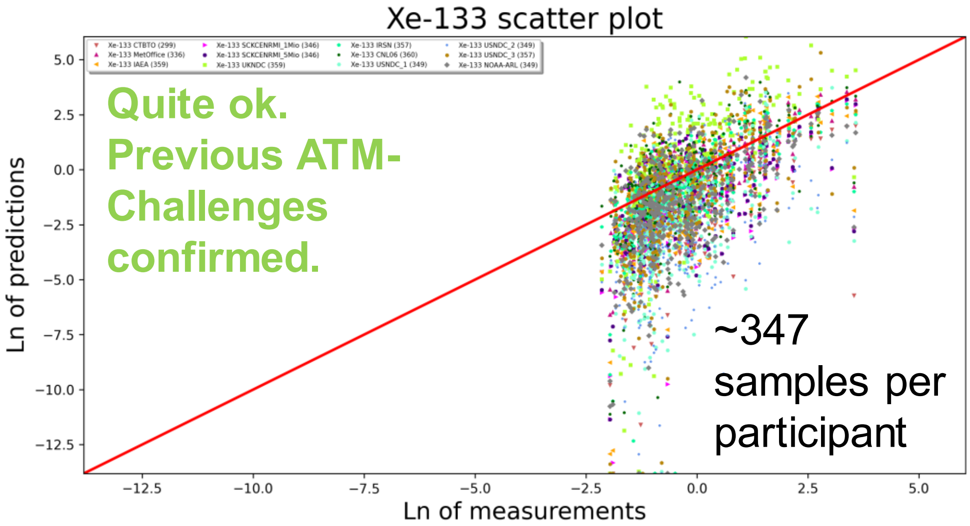
Deficiencies for Xe-133m, Xe-131m and Xe-135 cannot be blamed on ATM. ATM for Xe-133 works quite fine and the difference to other Xe-isotopes in ATM is just half-life.

Rather:

- Underestimated or even unknown (for Xe-135 local) emissions
- For Xe-135, Xe-133m and Xe-131m a lot of values are between the LC and MDC -> more false positives for metastables as of 2014, high measurement uncertainty between LC and MDC

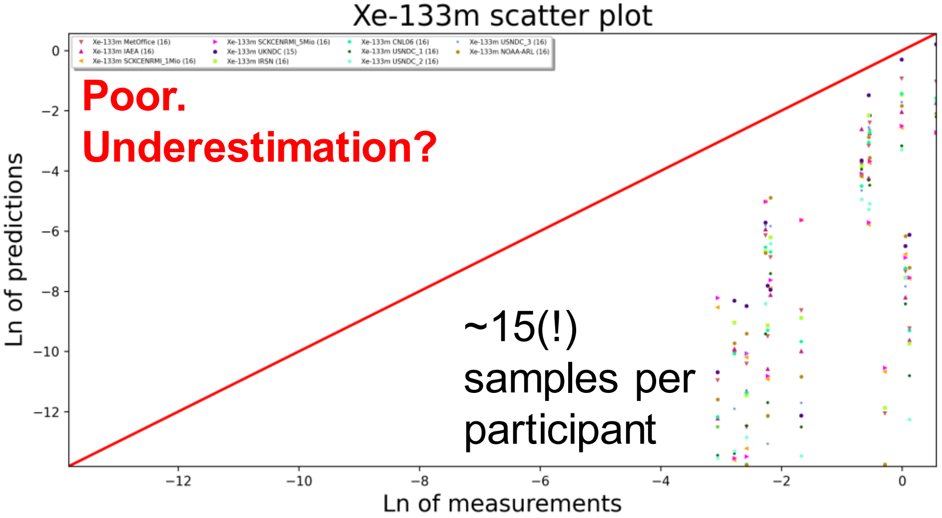
3.1.2. Global isotopic performance of background prediction: Samples >= MDC

Not much data (up to two orders of magnitude) **is left for Xe-135, Xe-133m and Xe-131m if only observed samples >= MDC are considered**



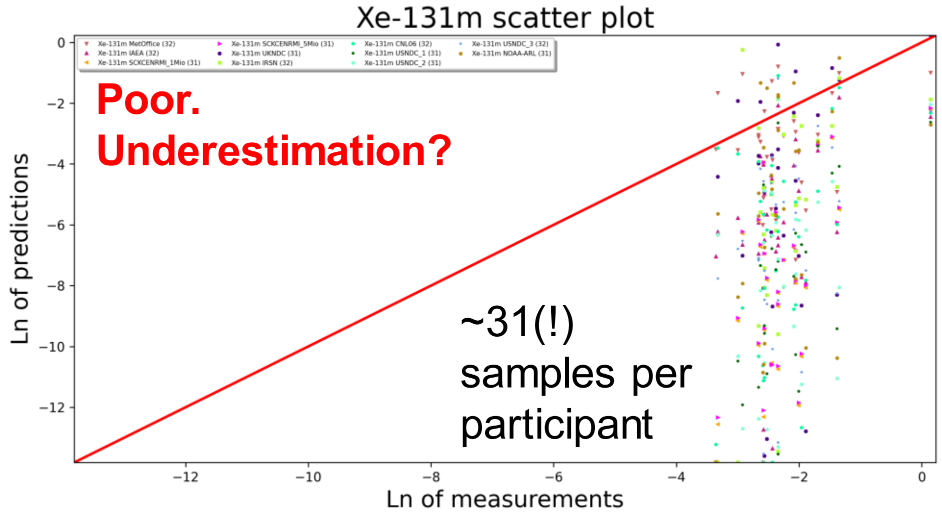
Quite ok. Previous ATM-Challenges confirmed.

~347 samples per participant



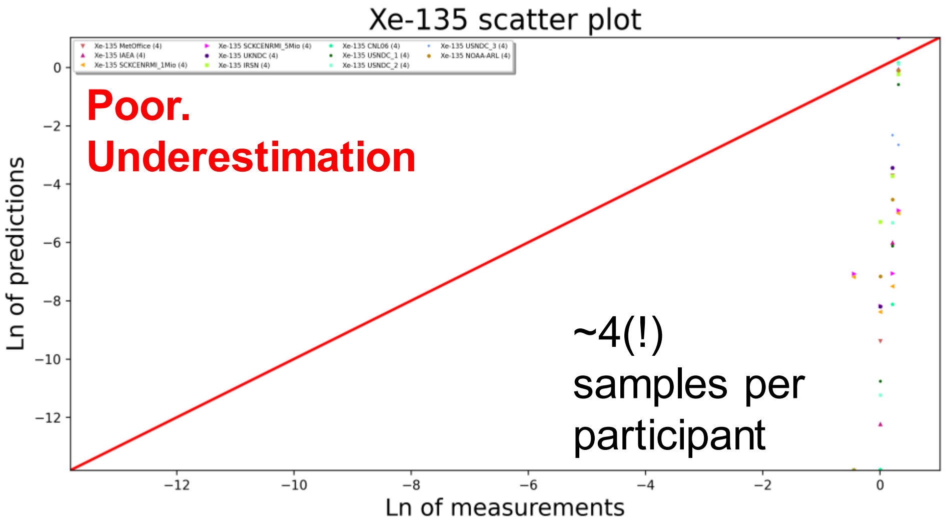
Poor. Underestimation?

~15(!) samples per participant



Poor. Underestimation?

~31(!) samples per participant



Poor. Underestimation

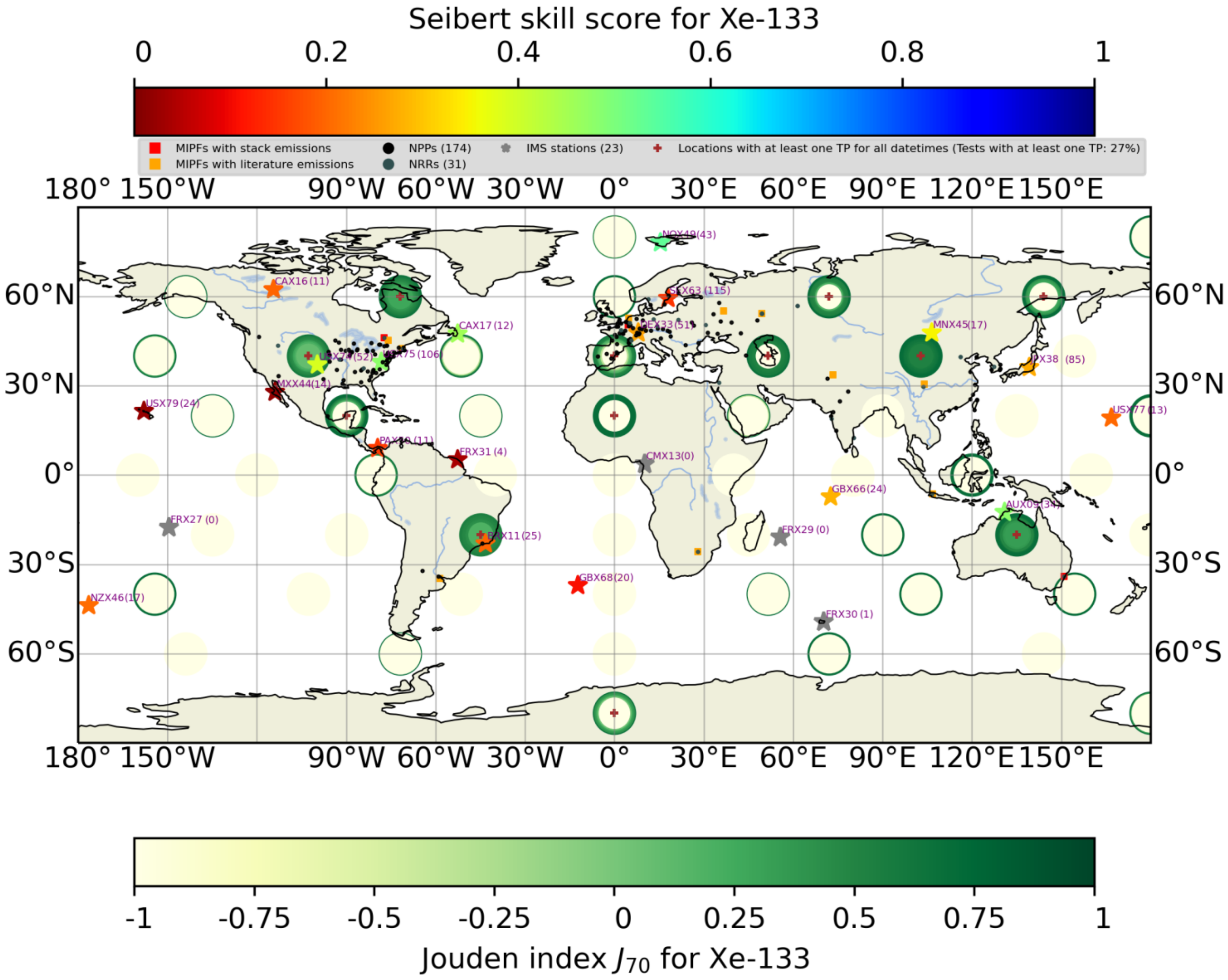
~4(!) samples per participant

IMPORTANT: Deficiencies for Xe-133m, Xe-131m and Xe-135 cannot be blamed on ATM. ATM for Xe-133 works quite fine and the difference to other Xe-isotopes in ATM is just half-life.

Rather:

- Underestimated or even unknown (for Xe-135 local) emissions
- For Xe-135, Xe-133m and Xe-131m a lot of values are between the LC and MDC -> more false positives for metastables as of 2014, high measurement uncertainty between LC and MDC

3.2. Detection power and local performance of background (for >= MDC samples) prediction for Xe-133



Jouden index = Sensitivity (= TPR) + Specificity (= 1-FPR) -1; [-1,1]

- ***Global* median** (over 12 submissions): **J = 0.48 excluding „neutrals“ ; J = 0.13 including them** (factor 4 difference)
- **Best detectability for Northern Hemisphere extratropics underground tests. Low detectability in the Southern Hemisphere** (however, just eight IMS NG systems were operating as of 2014!) **and in the Northern Hemisphere tropics. Excluding (including) „neutrals“ 72% (50%)** of the 424 tests **(mainly underwater) produce no signal >=LC (>0) (J set to -1!)**
- **Highest background prediction skill scores for** IMS stations **USX75** (CNL dominance), **CAX17** (CNL dominance), **NOX49** and **AUX09** (ANSTO dominance),

$$S_r = 2(1+R)\left(\frac{\sigma_m}{\sigma_o} + \frac{\sigma_o}{\sigma_m}\right)^{-2} \qquad S_b = \frac{1}{1 + b\mathrm{FB}^2}$$

Seibert Skill Score:

$$SS = \alpha S_r + (1-\alpha) S_b$$

3.3. Screening and timing power in an ideal world

*Results for **OMITTING** civil background* (as done in many other studies before, no „neutrals“) – *sanity check for screening and timing*:

Underground tests:

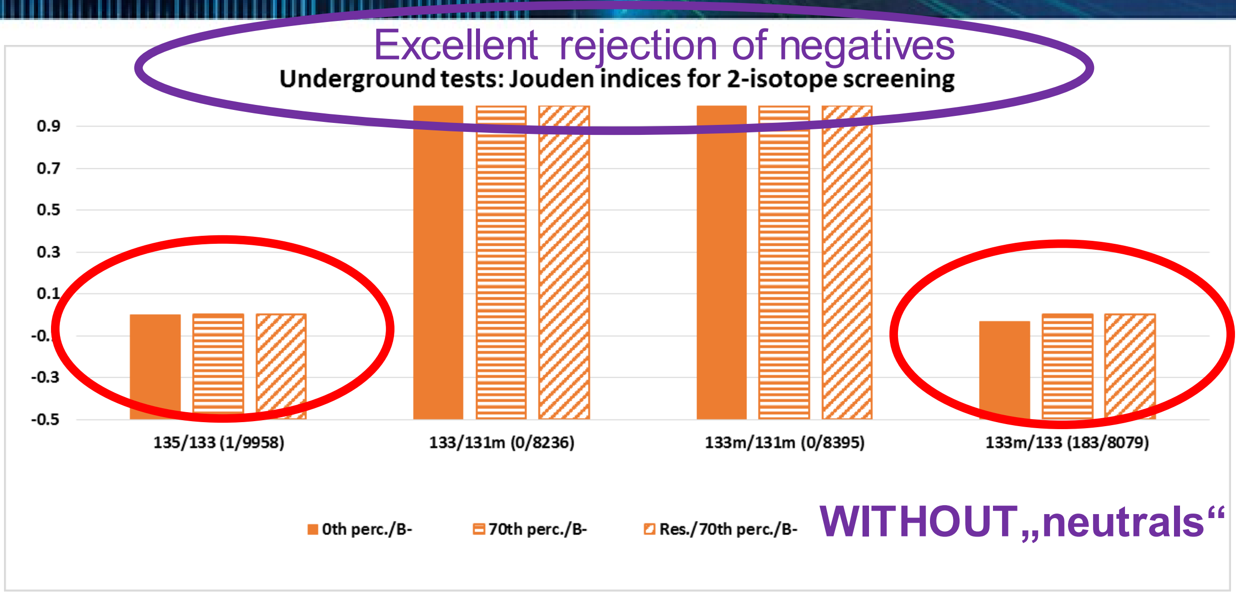
- Cases with **only two isotopes** above the LC confined to ratios **Xe-135/Xe-133** and **Xe-133m/Xe-133**. **NO screening power can be stated -> problem with IDC screening flags for these ratios given the delayed releases??**
- Cases with **only three isotopes** above the LC confined to ratios **Xe-133m/Xe-133--Xe-133m/Xe-131m** and **Xe-135/Xe-133--Xe-133m/Xe-133**. **Maximum attainable screening power is reached.**
- **Maximum attainable screening power is reached with all four isotopes above the LC.**
- Subsequent **timing success rate ranges between 0% for Xe-135/Xe-133, 80% for Xe-133/Xe-131m**, to **100% for Xe-133m/Xe-131m and Xe-133m/Xe-133.** Too strict tolerance criterion for Xe-135/Xe-133?

Undewater tests:

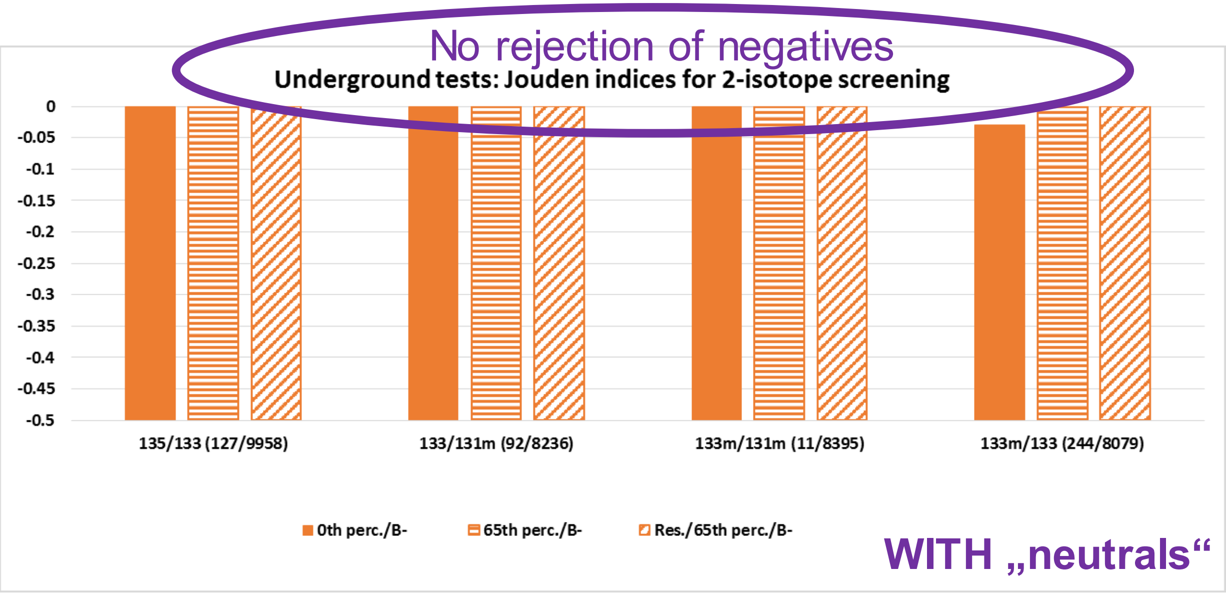
- Cases with **only two isotopes** above the LC confined to ratios **Xe-135/Xe-133** and **Xe-133m/Xe-133** as well. **Maximum attainable screening power for Xe-135/Xe-133 and for Xe-133m/Xe-133!**
- Cases with **only three isotopes** above the LC confined to ratio **Xe-135/Xe-133--Xe-133m/Xe-133. Maximum attainable screening power is reached.**
- **No cases with all four isotopes above the LC. No Xe-131m above the LC.**
- Subsequent **timing success rate ranges between 25% for Xe-135/Xe-133 and 100% for Xe-133m/Xe-133.**

3.4.1. Screening power: Underground tests in a real world

Please check Liu et al. P2.1-681!


Excellent rejection of negatives

WITHOUT „neutrals“


No rejection of negatives

WITH „neutrals“

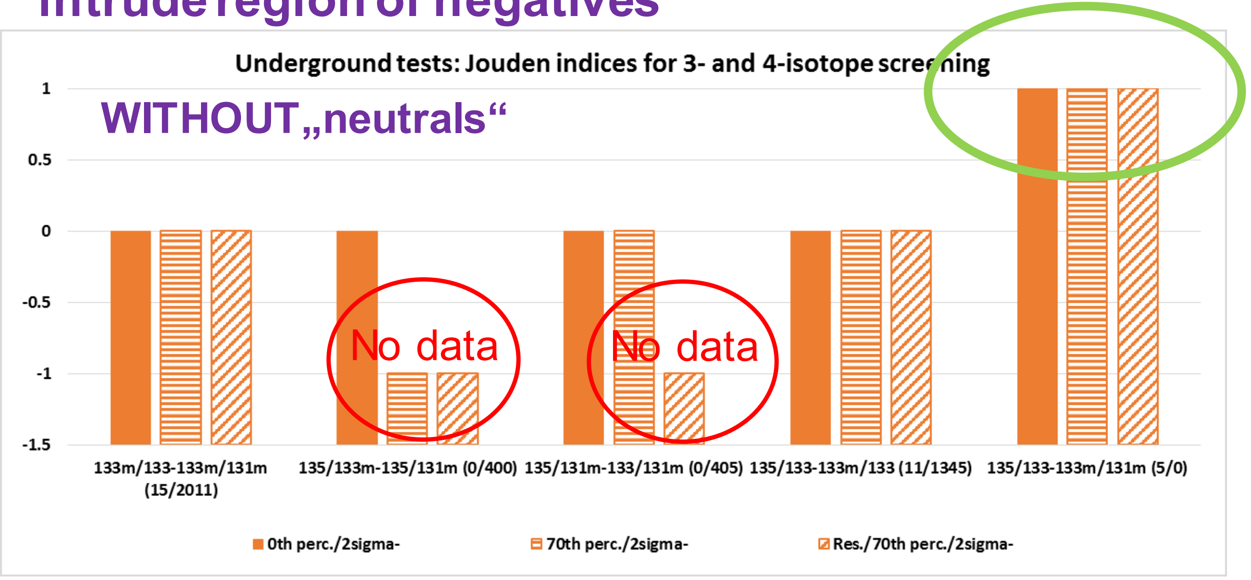
0th perc./B- or 2sigma- : Applying default IDC screening procedures *to data without any selection*

70th perc./B- or 2sigma- : Applying default IDC screening procedures *to data selected based on observation thresholding*

Res./70th perc./B- or 2sigma- : Applying default IDC screening procedures *to data selected based on hybrid observation-residual thresholding*

Screening results:

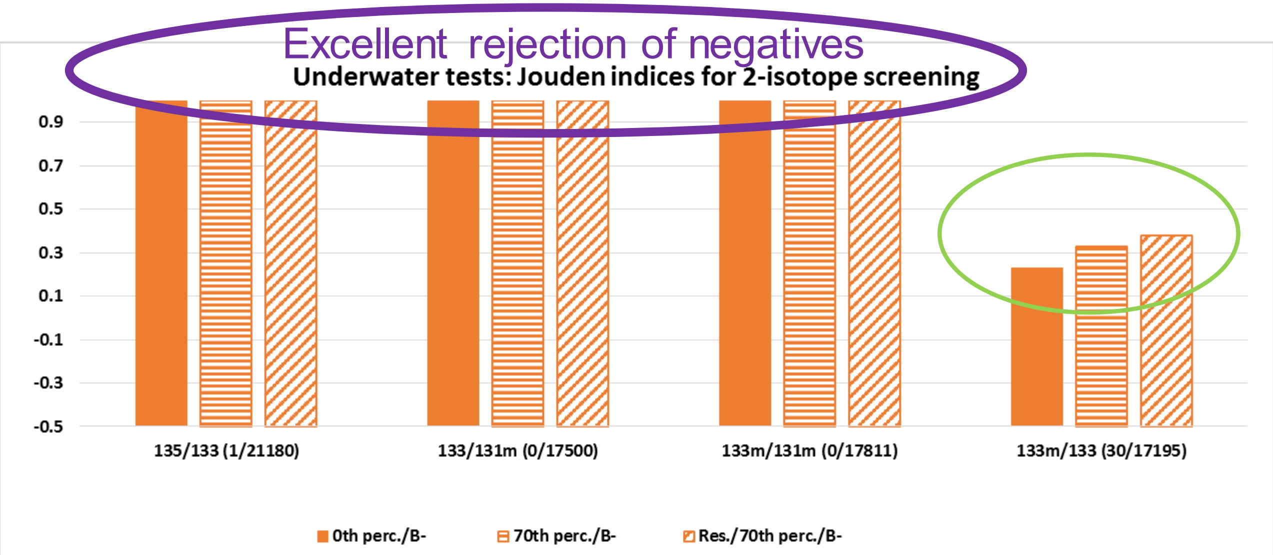
- **Problematic IDC screening flags**
- **Perfect 4-isotope screening – also with „neutrals“!**
- **„Neutrals“ can have a huge impact -> „Neutrals“ intrude region of negatives**


WITHOUT „neutrals“

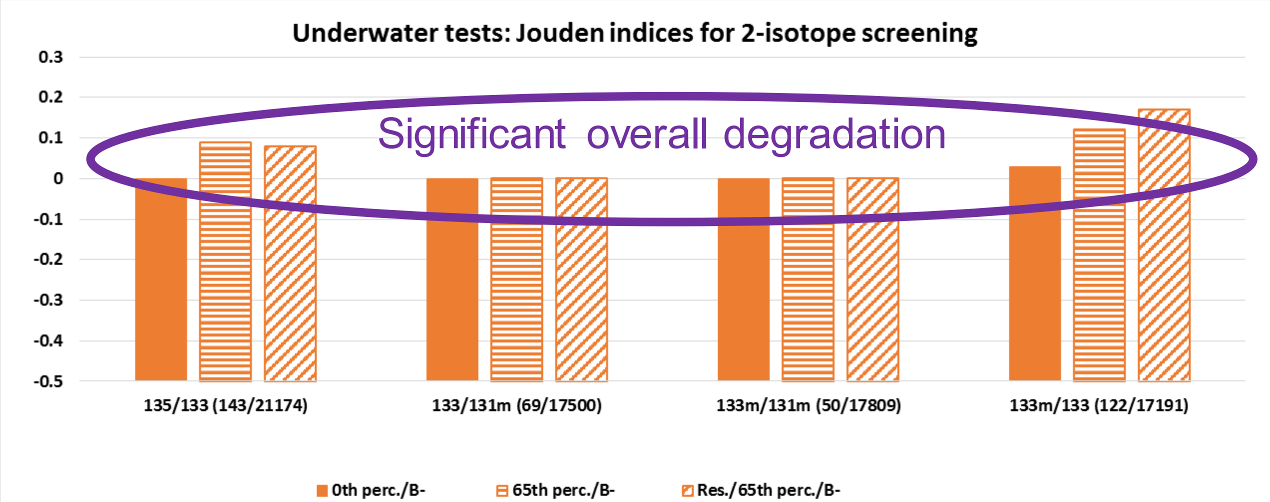
No data

No data

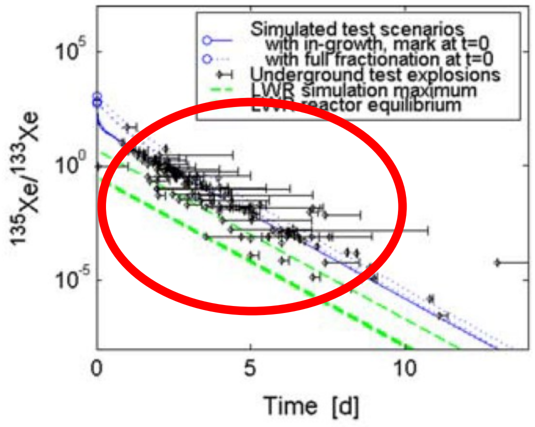
3.4.2. Screening power: Underwater tests in a real world



WITHOUT„neutrals“



WITH„neutrals“

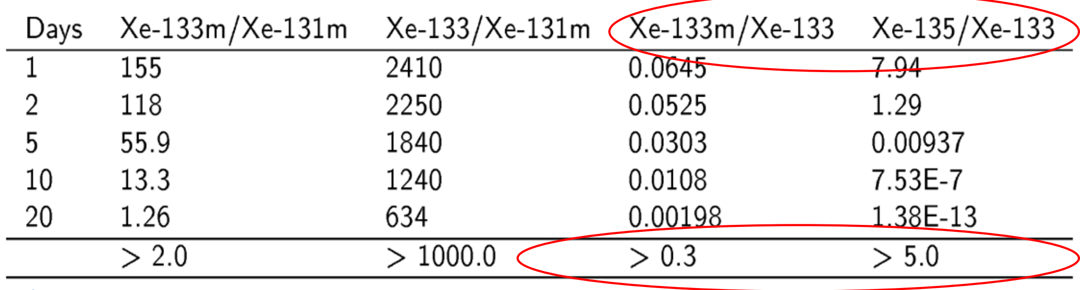


From *Kalinowski et. al (2010)*

Screening results:

- Indication of skill for Xe-133m/Xe-133 due to prompt release despite lower release rates. **Increased skill for residual thresholding for this ratio.** Reconsider IDC screening flags?
- **Huge impact of „neutrals“**

Isotopic activity ratios at different times afterward an UNE (U-235) as well as IDC screening flags.



Days	Xe-133m/Xe-131m	Xe-133/Xe-131m	Xe-133m/Xe-133	Xe-135/Xe-133
1	155	2410	0.0645	7.94
2	118	2250	0.0525	1.29
5	55.9	1840	0.0303	0.00937
10	13.3	1240	0.0108	7.53E-7
20	1.26	634	0.00198	1.38E-13
	> 2.0	> 1000.0	> 0.3	> 5.0

Targeted towards fresh and old test signals

Targeted towards fresh test signals, will not work for delayed releases

4.1. Level 2 evaluation: Tasks and methods

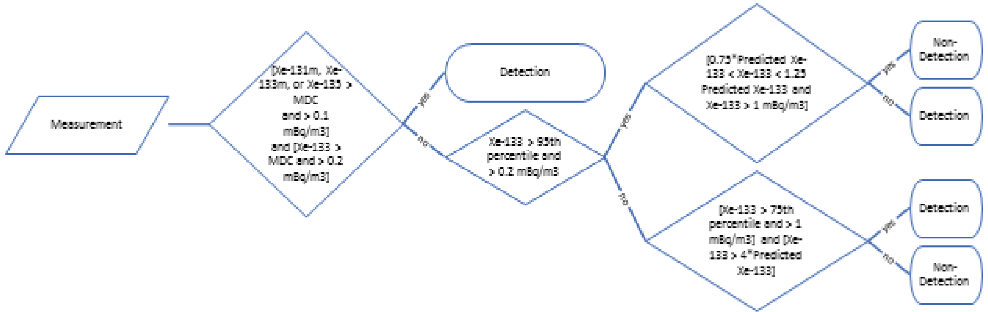
- **Required participants' expertise:** *ATM and/or radionuclide expertise*

- **Task 1:** *State the isotope name(s), station(s) and collection stop time(s) for anomalous activity concentration(s) for one (or several) radioxenon isotope(s) within each of the given time periods and for each of the test scenarios.*

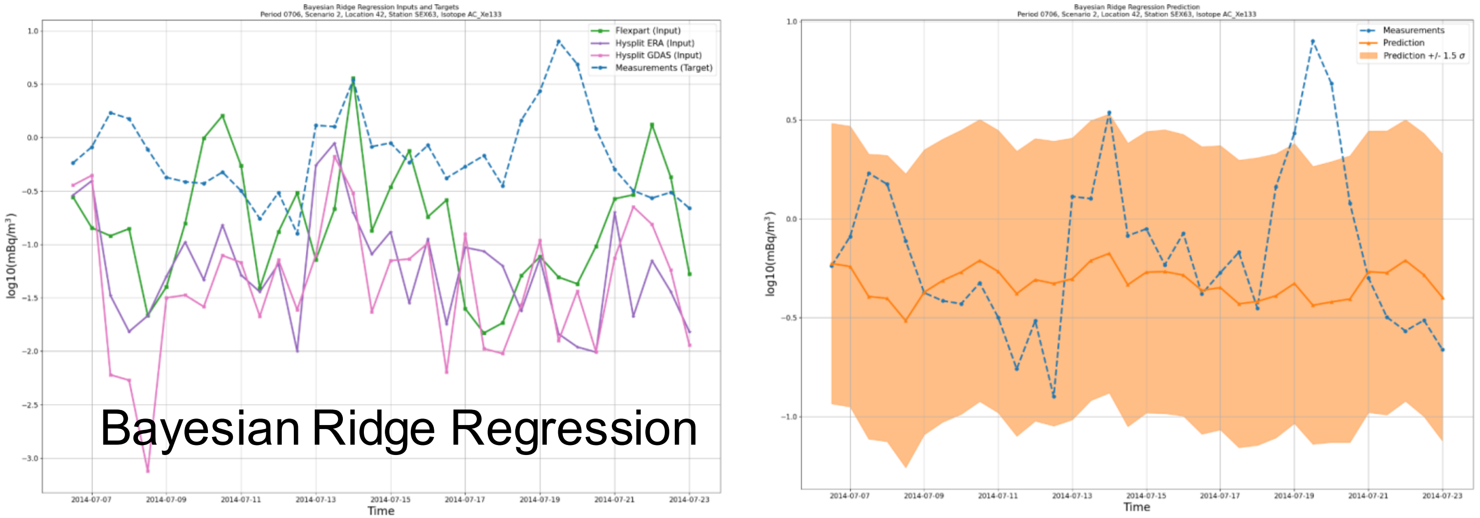
- **Task 2:** *State the isotope names, station(s) and collection stop time(s) for (an) isotopic ratio(s) related to anomalous activity concentrations indicating a military event.*

- **Task 3:** *If an explosion was found, state the time zero, including an uncertainty estimate.*

- **Methods:** ATM, Thresholding, Machine Learning (Isolation Forest and Decision Tree), ATM and others combined, Bayesian Ridge Regression & Lognormal distribution fitting



Thresholding



Bayesian Ridge Regression

4.2. Level 2 evaluation: Results

- **Pretended scenarios** (background only, 103) **were not discarded** by all but one participant (who, however, discarded many actual explosion scenarios).

- **For Xe-133 Thresholding** (up to J=0.61) **yields global results comparable to the Level 1 percentile approach** if a full year of data (CTBTO run) is used for the percentile approach.

- **For Xe-133m** (up to J=0.73)**, Xe-131m** (up to J=0.97) **and Xe-135** (up to J= 0.84) **using isotopic ratio information** (be in in the frame of **thresholding** or in the frame of **machine learning**) **yields the best result.**

- Clear **difference of Level 2 compared to Level 1 detection power analysis: FPR mostly below 5%!** However, **a lower FPR comes at the cost of a lower TPR.**

- **Disadvantage of Level 1 compared to Level 2 detection power analysis: A priori definition of a percentile threshold is needed in real life**, **which may depend on** the given, probably unknown **nuclear source term.**

- Number of isotope ratios in the nuclear explosion domain stated compares quite well if not different approaches are taken. One participant performed only (but successfully) 4-isotope screening using all above zero values.

- Timing success for events is rather poor (few %). However, source terms were not known by participants.

5. Level 3 evaluation

- **Required participants' expertise:** *Higher-level ATM and statistical expertise*

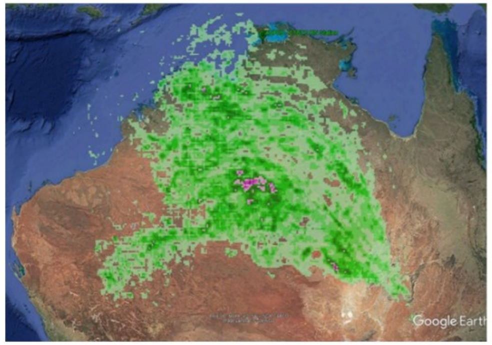
- **Task 1:** *Determine the geographical area (lon/lat) for eight selected nuclear explosions (underground or underwater), including an uncertainty estimate of that area.*

- **Task 2:** *State the total release (in Bq) of the explosion for all four Xe isotopes, including an uncertainty estimate per isotope.*

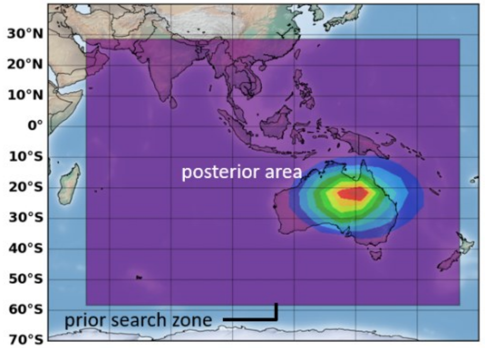
- **Methods***:* Bayesian inference (also combined with Machine Learning), overlap counting and PSR

- **Results:**
  - Source reconstruction using **Bayesian approaches works quite** (given the challenging scenarios) **well if input samples can be identified.**
  - **Probable source regions are quite large for different reasons** (e.g., network sparsity or lack of multiple isotope detections) and frequently add up to several hundreds of kilometers (and sometimes even up to ca. 2500 km).
  - **A lot depends on which samples are selected for source term inversion.**

Algorithm 1



Algorithm 2



- It seems to be easier to estimate release magnitudes correctly although this is only true for Xe-133 and is likely due to the underlying explosion source terms.
- Simple methods (e.g., PSR or overlap counting) can additionally be used to get crude first impressions.

6.1. Conclusions I

- **Adding nulcear explosion signals on top of the civil background creates a special kind of positives („neutrals“) with huge impact on detection and screening power.**

- Using **ATM based residuals alters detection power** compared to direct (pseudo-)observation distribution analysis **depending on the average background and background prediction performance in relation to the nuclear explosion source term magnitude. Noteworthy (positive) influence is only on Xe-133** (up to +15%).

- **There is an area of conflict between the necessity of using all above LC samples for nuclear explosion screening and the uncertain measurements and predictions between the LC and the MDC.**

- **Shortcomings for Xe-133m, Xe-131m and Xe-135 cannot be blamed to ATM. Emission deficiencies and issues with detection and quantification of below MDC IMS measurements** seem to be problems on their own.

- There is **a high fraction** (up to 72% for Xe-133 excluding „neutrals“) **of nuclear tests causing no signal >= LC given the source terms** (extremely weak for underwater) **investigated and the IMS station network as of 2014** (23 stations).

6.2. Conclusions II

- **Screening and timing** (not shown) based on true positive screened samples **for the ratio Xe-133m/Xe-133 can likely be improved by using residuals in case of underwater explosions.** This is likely related to the subtle signals of underwater explosions compared to the substantial (Xe-133) background which is removed (at least partly) via the residual approach.

- **Methods for Level 2 detection power estimation are at least methodically superior to Level 1 methods**. **Level 3 source term estimation strongly depends on finding and selecting appropriate samples**.

- **More knowledge** would be needed **regarding emission inventories of Xe-133m, Xe131m and Xe-135**. As this may be difficult to achieve in entireness even on long-term, **Machine Learning (ML) based approaches for anomaly detection** and/or **nudging ATM simulations towards (IMS) observations** (*Zwaaftink et al. (2018)*) as well as **source term inversion** may be used as remedy to **overcome effects of source term and transport errors**.

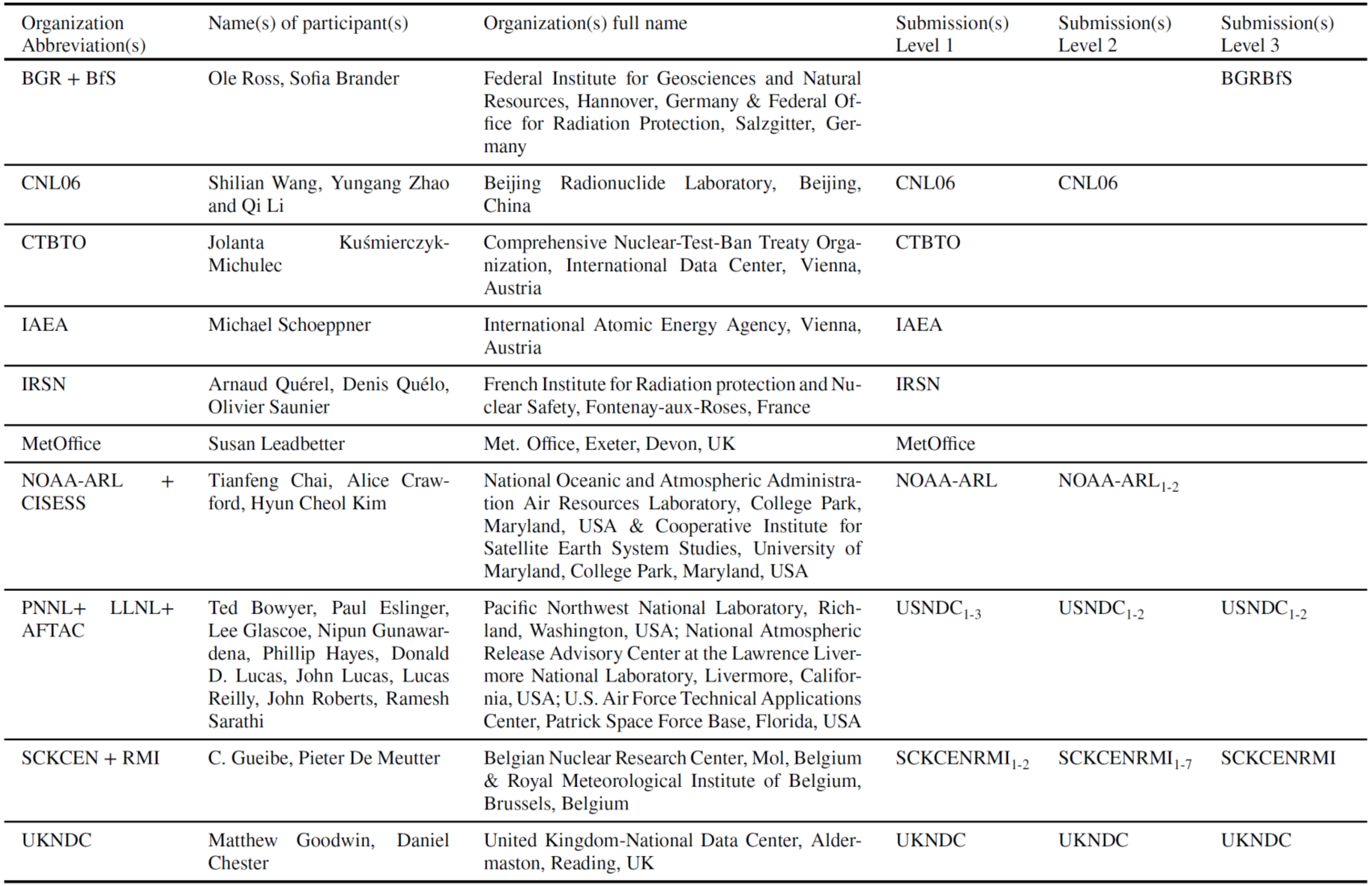
- **Looking forward to NG noble gas measurements from an increasing number of IMS noble gas stations ! -> significant 4-isotope samples more likely**

7. References

- A. Axelsson, A. Ringbom, M. Aldener, T. Fritioff, and A. Mörtsell (2014): The Impact of System Characteristics on Noble Gas Network Verification Capability for CTBT. Report No. FOI-R-3856-SE, ISSN-1650-1942, Stockholm, Sweden.
- J. L. Burnett, P. W. Eslinger, B. D. Milbrath (2019): The detectability of the Wigwam underwater nuclear explosion by the radionuclide stations of the International Monitoring System. *Journal of Environmental Radioactivity* **208–209**, 106030.
- M. B. Kalinowski, A. Axelsson, M. Bean, X. Blanchard, T. W. Bowyer, G. Brachet, S. Hebel, J. I. McIntyre, J. Peters, C. Pistner, M. Raith, A. Ringbom, P. R. J. Saey, C. Schlosser, T. J. Stocki, T. Taffary, and R. K. Ungar (2010): Discrimination of Nuclear Explosions against Civilian Sources Based on Atmospheric Xenon Isotopic Activity Ratios. *Pure and Applied Geophysics* **167**, 517–539.
- vDEC-Virtual Data Exploitation Centre. CTBTO, https://www.ctbto.org/specials/vdec/
- M. Zähringer and G. Kirchner (2008): Nuclide ratios and source identification from high-resolution gamma-ray spectra with Bayesian decision methods. *Nuclear Instruments and Methods in Physics Research A*.

The exercise and ist participants

Table 1: Participants of the 1st Nuclear Explosion Signal Screening Open Inter-Comparison Exercise 2021.



Organization Abbreviation(s)	Name(s) of participant(s)	Organization(s) full name	Submission(s) Level 1	Submission(s) Level 2	Submission(s) Level 3
BGR + BfS	Ole Ross, Sofia Brander	Federal Institute for Geosciences and Natural Resources, Hannover, Germany & Federal Office for Radiation Protection, Salzgitter, Germany			BGRBfS
CNL06	Shilian Wang, Yungang Zhao and Qi Li	Beijing Radionuclide Laboratory, Beijing, China	CNL06	CNL06	
CTBTO	Jolanta Kuśmierczyk-Michulec	Comprehensive Nuclear-Test-Ban Treaty Organization, International Data Center, Vienna, Austria	CTBTO		
IAEA	Michael Schoeppner	International Atomic Energy Agency, Vienna, Austria	IAEA		
IRSN	Arnaud Quérel, Denis Quélo, Olivier Saunier	French Institute for Radiation protection and Nuclear Safety, Fontenay-aux-Roses, France	IRSN		
MetOffice	Susan Leadbetter	Met. Office, Exeter, Devon, UK	MetOffice		
NOAA-ARL + CISESS	Tianfeng Chai, Alice Crawford, Hyun Cheol Kim	National Oceanic and Atmospheric Administration Air Resources Laboratory, College Park, Maryland, USA & Cooperative Institute for Satellite Earth System Studies, University of Maryland, College Park, Maryland, USA	NOAA-ARL	NOAA-ARL$_{1-2}$	
PNNL+ LLNL+ AFTAC	Ted Bowyer, Paul Eslinger, Lee Glascoe, Nipun Gunawardena, Phillip Hayes, Donald D. Lucas, John Lucas, Lucas Reilly, John Roberts, Ramesh Sarathi	Pacific Northwest National Laboratory, Richland, Washington, USA; National Atmospheric Release Advisory Center at the Lawrence Livermore National Laboratory, Livermore, California, USA; U.S. Air Force Technical Applications Center, Patrick Space Force Base, Florida, USA	USNDC$_{1-3}$	USNDC$_{1-2}$	USNDC$_{1-2}$
SCKCEN + RMI	C. Gueibe, Pieter De Meutter	Belgian Nuclear Research Center, Mol, Belgium & Royal Meteorological Institute of Belgium, Brussels, Belgium	SCKCENRMI$_{1-2}$	SCKCENRMI$_{1-7}$	SCKCENRMI
UKNDC	Matthew Goodwin, Daniel Chester	United Kingdom-National Data Center, Aldermaston, Reading, UK	UKNDC	UKNDC	UKNDC

- Date of issue: Dec., 1st, 2021
- Date of (official) closure: June, 30th, 2022 Unofficial closure: October, 10th, 2022
- Ten participating organizations or entities from seven international countries (Belgium, Germany, UK, Austria, China, France and the US)