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

wosmip Workshop on Signatures of Man-made Isotope Production

GeoSphere Austria

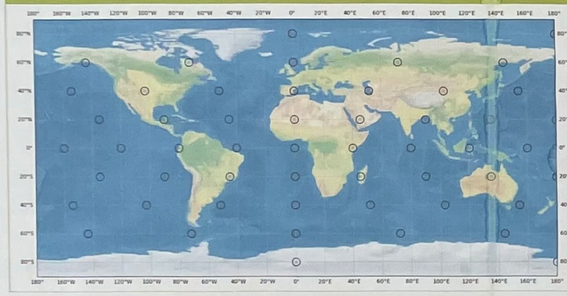
C .Maurer[1], B. Liu[2], J. Brioude[3], D. Arnold Arias[1], Y. Kijima[2], J. Kúsmierczyk-Michulec[2], R. Schoemaker[2], A. Tipka[2], and M. Kalinowski[2]

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

Abstract

The *1st Nuclear Explosion Signal Screening Open Inter-Comparison Exercise 2021* evaluated different screening methods to identify radioxenon detections not consistent with the radioxenon background from nuclear facilities. It was based on a subset of a comprehensive radioxenon test data set produced for the whole year of 2014 with hypothetical nuclear underground and underwater explosion signals added to IMS observations.
The exercise considered three levels of participation requiring different levels of expertise: 1) *Level 1* (basic, ATM expertise only), for which participants provided simulated radioxenon background time series at the 23 IMS stations defined in the test data set to be used as input for screening based on a set of predefined metrics covering detection, screening, and timing powers; 2) *Level 2* (ATM and/or radionuclide expertise), for which, in addition to Level 1, participants provided their own screening methods and results for detection, screening, and timing powers; and 3) *Level 3* (higher-level ATM and statistical expertise), for which, in addition to Level 2, results were provided for location and magnitude estimates for a few selected test cases.
Depending on the specific level the exercise had three to seven participating organizations from seven countries.

1. Explosion scenarios – test data set

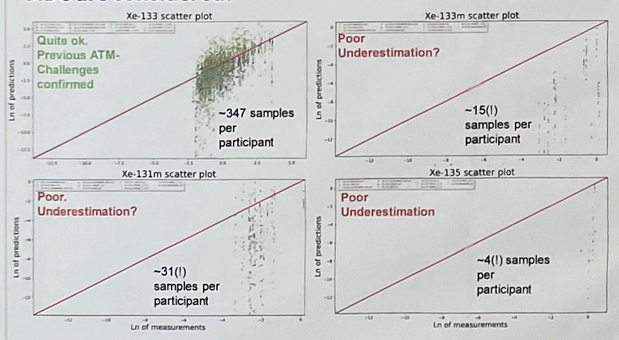


- **424 scenarios: 8 date-times and 53 locations**
- **136 1 kt underground explosions: 24 hours containment, 10% venting** (IDC source term)
- **288 1 kt underwater explosions: prompt 0.92% venting** (Burnett et al., 2020, source term)
- **23 IMS stations** with data as of 2014 (SAUNA + SPALAX), **explosion signals added on top of civil background**

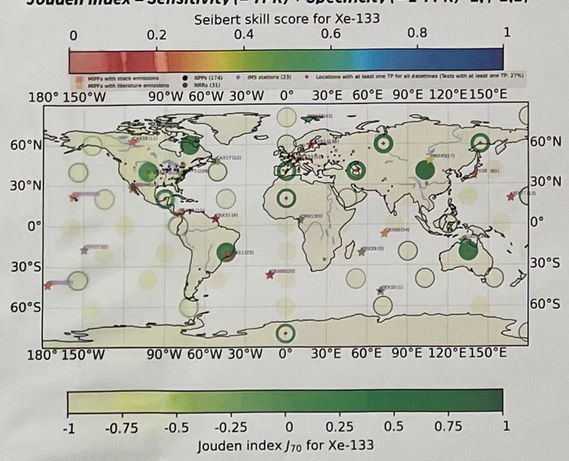
2. L1 evaluation: *"Is an isotopic measurement an anomaly?"*

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 ("Thresholding").
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.**

Thresholding on residuals: Not much data is left for Xe-135, Xe-133m and Xe-131m if only observed samples >= MDC are considered.



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



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.

- Underestimated or even unknown (for Xe-135 local) emissions
- Apart from Xe-133 a lot of values are between the LC and MDC -> too much false positives as of 2014, higher measurement uncertainty between LC and MDC

- ***Global* median** (over 12 submissions): **J = 0.48 excluding „neutrals" ; J = 0.13 including them** (as displayed above; factor 4 difference)
- **Best detectability for NH extratropics underground tests. Low detectability in the SH** (just eight IMS NG systems as of 2014!) **and in the NH 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**, **CAX17** (both CNL dominance), **NOX49** and **AUX09** (ANSTO dominance).

3. L1 evaluation: „*Has a nuclear explosion to be assumed? If yes, can we determine the release time within a predefined time window?*"

Underground tests:

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

- **Two isotopes** >=LC confined to ratios **Xe-135/Xe-133** and **Xe-133m/Xe-133. NO screening power -> problem with IDC screening given the delayed releases!**
- **Three isotopes** >= 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.**
- **Maximum attainable screening power with all four isotopes above the LC.**
- **Timing success rate: 0% for Xe-135/Xe-133, 80% for Xe-133/Xe-131m, 100% for Xe-133m/Xe-131m and Xe-133m/Xe-133.** Too strict criterion for Xe-135/Xe-133?

*2) Results for **INCLUDING** background* (both with and without „neutrals")

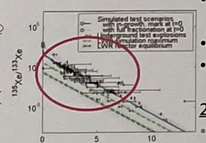
- **Perfect 4-isotope screening – also with „neutrals"!**
- **„Neutrals" can have a huge impact -> „Neutrals" intrude region of civil domain**

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



From *Kalinowski et. al (2010)*

Underwater tests:

*1) Results for **OMITTING** background*

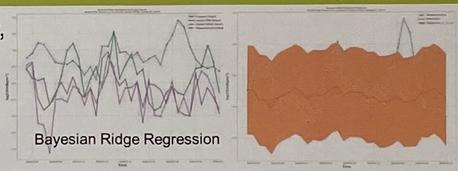
- **Two isotopes** >= LC confined to ratios **Xe-135/Xe-133** and **Xe-133m/Xe-133** as well. **Maximum attainable screening power for Xe-135/Xe-133 and for Xe-133m/Xe-133!**
- **Three isotopes** >= LC confined to ratio **Xe-135/Xe-133--Xe-133m/Xe-133. Maximum attainable screening power.**
- **No cases with all four isotopes >= LC. No Xe-131m >= LC.**
- **Timing success rate 25% for Xe-135/Xe-133 and 100% for Xe-133m/Xe-133.**

*2) Results for **INCLUDING** background* (both with and without „neutrals")

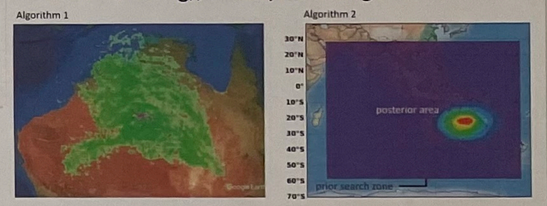
- Indication of some screening power for Xe-133m/Xe-133 due to prompt release despite lower release rates. **Increased skill for residual thresholding for this ratio.**
- **Huge impact of „neutrals"**

4. L2+L3 evaluation

Methods L2: ATM, Thresholding, Machine Learning (Isolation Forest and Decision Tree), ATM and others combined, Bayesian Ridge Regression & Lognormal distribution fitting



Methods L3: Bayesian inference (also combined with Machine Learning), overlap counting and PSR



- **Pretended scenarios** (background only, 103) **were not discarded** by all but one participant (who, however, discarded many actual explosion scenarios).
- **Xe-133 thresholding global results comparable to L1 percentile approach.**
- **For Xe-133m , Xe-131m and Xe-135 adding isotopic ratio information yields the best result.**
- **Difference L2 to L1 detection power analysis: FPR mostly below 5%!** However, **a lower FPR comes at the cost of a lower TPR.**
- **Disadvantage of L1 versus L2 detection power analysis: A priori definition of a percentile threshold is needed**, which **may depend on** the **nuclear source term.**

- **Bayesian approaches work quite 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.
- **A lot depends on which samples are selected.**

5. Conclusions

- **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. These kind of samples do not exist if no background is considered (as done in most previous studies).**
- 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%).
- **Shortcomings for Xe-133m, Xe-131m and Xe-135 cannot be blamed to ATM. Emission deficiencies and issues with detections and quantification of below MDC IMS measurements** seem to be problems on their own.
- **There is a 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.**
- There is **a high fraction of nuclear tests causing no signal >= LC given the specific source terms investigated and the IMS network as of 2014.**
- **Screening and timing** 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 related to the subtle signals of underwater explosions compared to the substantial Xe-133 background.
- **Methods for L2 detection power estimation are** at least **methodically superior to L1 methods. L3 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.** Additionally, **Machine Learning (ML) based approaches for anomaly detection** and/or **nudging ATM simulations towards (IMS) observations** 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 ! -> significant 4-isotope samples more likely**

Contact: Christian Maurer, Department Environmental Meteorology | christian.maurer@geosphere.at | Tel. +43 1 360 26 /2012 | www.geosphere.at