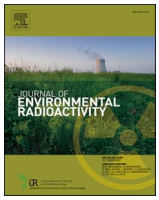
Journal of Environmental Radioactivity 255 (2022) 106968

Contents lists available at ScienceDirect

Journal of Environmental Radioactivity

journal homepage: www.elsevier.com/locate/jenvrad



Third international challenge to model the medium- to long-range transport of radioxenon to four Comprehensive Nuclear-Test-Ban Treaty monitoring stations☆

C. Maurer [a,*], S. Galmarini [b], E. Solazzo [b], J. Kuśmierczyk-Michulec [c], J. Baré [c], M. Kalinowski [c], M. Schoeppner [c], P. Bourgouin [c], A. Crawford [d], A. Stein [d], T. Chai [d], F. Ngan [d], A. Malo [e], P. Seibert [f], A. Axelsson [g], A. Ringbom [g], R. Britton [h], A. Davies [h], M. Goodwin [h], P.W. Eslinger [i], T.W. Bowyer [i], L.G. Glascoe [j], D.D. Lucas [j], S. Cicchi [j], P. Vogt [j], Y. Kijima [k], A. Furuno [k], P.K. Long [l], B. Orr [m], A. Wain [n], K. Park [o], K.-S. Suh [o], A. Quérel [p], O. Saunier [p], D. Quélo [p]

[a] Zentralanstalt für Meteorologie und Geodynamik (ZAMG), Vienna, Austria
[b] European Commission - Joint Research Center (JRC), Ispra VA, Italy
[c] Comprehensive Nuclear-Test-Ban Treaty Organization (CTBTO), Vienna, Austria
[d] National Oceanic and Atmospheric Administration Air Resources Laboratory (NOAA-ARL), College Park, MD, USA
[e] Environment and Climate Change Canada (ECCC), Meteorological Service of Canada, Canadian Meteorological Centre (CMC), Environmental Emergency Response Section, RSMC Montréal, Dorval, Québec, Canada
[f] University of Natural Resources and Life Sciences (BOKU), Institute of Meteorology and Climatology, Vienna, Austria
[g] Swedish Defence Research Agency (FOI), Stockholm, Sweden
[h] Atomic Weapons Establishment/United Kingdom-National Data Center (AWE/UK-NDC), Aldermaston, Reading, United Kingdom
[i] Pacific Northwest National Laboratory (PNNL), Richland, WA, USA
[j] National Atmospheric Release Advisory Center (NARAC) at the Lawrence Livermore National Laboratory (LLNL), Livermore, CA, USA
[k] Japan Atomic Energy Agency (JAEA), Tokai, Ibaraki, Japan
[l] Vietnam Atomic Energy Institute (VINATOM), Hanoi, Vietnam
[m] Australian Radiation Protection and Nuclear Safety Agency (ARPANSA), Yallambie/Miranda, Australia
[n] Bureau of Meteorology (BOM), Melbourne, Australia
[o] Korea Atomic Energy Research Institute (KAERI), Daejeon, Republic of Korea
[p] French Institute for Radiation Protection and Nuclear Safety (IRSN), Fontenay-aux-Roses, France

ARTICLE INFO

Keywords:
CTBTO
Atmospheric transport modeling
Radioxenon background

ABSTRACT

In 2015 and 2016, atmospheric transport modeling challenges were conducted in the context of the Comprehensive Nuclear-Test-Ban Treaty (CTBT) verification, however, with a more limited scope with respect to emission inventories, simulation period and number of relevant samples (i.e., those above the Minimum Detectable Concentration (MDC)) involved. Therefore, a more comprehensive atmospheric transport modeling challenge was organized in 2019. Stack release data of Xe-133 were provided by the Institut National des Radioéléments/IRE (Belgium) and the Canadian Nuclear Laboratories/CNL (Canada) and accounted for in the simulations over a three (mandatory) or six (optional) months period. Best estimate emissions of additional facilities (radiopharmaceutical production and nuclear research facilities, commercial reactors or relevant research reactors) of the Northern Hemisphere were included as well. Model results were compared with observed atmospheric activity concentrations at four International Monitoring System (IMS) stations located in Europe and North America with overall considerable influence of IRE and/or CNL emissions for evaluation of the participants' runs. Participants were prompted to work with controlled and harmonized model set-ups to make runs more comparable, but also to increase diversity. It was found that using the stack emissions of IRE and CNL

☆ This document is the result of a research project funded by the Austrian Federal Ministry of European and International Affairs as well the Austrian Federal Ministry of Education, Science and Research. The sponsors were not involved in the study design, in the collection, analysis and interpretation of data, in the writing of the report and in the decision to submit the article for publication.

* Corresponding author.
E-mail address: christian.maurer@zamg.ac.at (C. Maurer).
URL: https://www.zamg.ac.at (C. Maurer).

https://doi.org/10.1016/j.jenvrad.2022.106968
Received 21 December 2021; Received in revised form 8 July 2022; Accepted 15 July 2022
Available online 20 September 2022

with daily resolution does not lead to better results than disaggregating annual emissions of these two facilities taken from the literature if an overall score for all stations covering all valid observed samples is considered. A moderate benefit of roughly 10% is visible in statistical scores for samples influenced by IRE and/or CNL to at least 50% and there can be considerable benefit for individual samples. Effects of transport errors, not properly characterized remaining emitters and long IMS sampling times (12–24 h) undoubtedly are in contrast to and reduce the benefit of high-quality IRE and CNL stack data. Complementary best estimates for remaining emitters push the scores up by 18% compared to just considering IRE and CNL emissions alone. Despite the efforts undertaken the full multi-model ensemble built is highly redundant. An ensemble based on a few arbitrary runs is sufficient to model the Xe-133 background at the stations investigated. The effective ensemble size is below five. An optimized ensemble at each station has on average slightly higher skill compared to the full ensemble. However, the improvement (maximum of 20% and minimum of 3% in RMSE) in skill is likely being too small for being exploited for an independent period.

1. Introduction

Atmospheric transport modeling (ATM) is an essential tool for linking radionuclide measurements of the International Monitoring System (IMS, Ctbto Preparatory Commission (2019)) of the Comprehensive Nuclear-Test-Ban Treaty Organization (CTBTO) to possible sources in order to enable verification of compliance with the Comprehensive Nuclear-Test-Ban Treaty (CTBT, Ctbto (1996); Wotawa et al. (2003); Matthews and De Geer (2004); Ctbto (2020b)). However, possible underground nuclear explosions like those conducted by the Democratic People's Republic of Korea (DPRK) in the recent years (e.g., Ringbom et al. (2009) or Ringbom et al. (2014)) are especially challenging to verify. If the explosion cavity is well contained only small fractions of noble gases will escape into the atmosphere, in a quantity that is comparable with or smaller than releases from medical isotope production facilities (MIPFs) or nuclear reactors (see the review article Bowyer (2020) and references therein). Properly simulating the industrial radioxenon background via ATM is of utmost importance and has attracted increasing attention over the last years (e.g., most recently Goodwin et al. (2021a) or Goodwin et al. (2021b)). Emission inventories – accounting above all for completeness in terms of the number of existing industrial emitters (not only, but in the first place for MIPFs) and for rough estimates of their daily emissions – are one important ingredient for successfully modeling the civil radioxenon background for the purpose of CTBT verification (see review article Bowyer (2020) and references therein).

An analysis by the International Data Center (IDC) of CTBTO performed for the year 2014 revealed that meanwhile enough historic MIPF emission data have been collected to perform a more comprehensive study. In this context it has to be mentioned that the IDC has started receiving measured release data following WOSMIP V (Workshop on the Signature of Man-Made Isotope Production, https://www.wosmip.org/). The study presented makes use of this evolving emission data pool.

Since ATM is connected with inherent uncertainties international multi-model exercises – so called "challenges" – to model the radioxenon background under different settings have been conducted in the past to tackle the issue of civil radioxenon background analysis. After successfully performing two ATM challenges (2015 & 2016, Eslinger et al. (2016) and Maurer et al. (2018)) in the past years with some limitations in terms of emission inventories, simulation period and number of relevant samples (i.e., those above the Minimum Detectable Concentration (MDC)) involved, a first step in the transition from merely scientific case studies towards the practical use of modeling radioxenon background at selected IMS stations had been envisioned for a third challenge. Questions regarding the required temporal emission resolution for medical isotope production sites needed for ATM and tolerable emission uncertainty could be answered by the last two challenges and other recent publications. De Meutter et al. (2018) were the first to state that daily emission data resolution should be sufficient for most cases, at least for the current IMS sampling time intervals (12–24 h). Recently, Generoso et al. (2022) demonstrated that the benefit from more detailed emission data gathered directly at the stack of a facility (therefore usually called stack data in brief) decreases with increasing distance from the source. Analysis of one year simulation time series with/-without stack data shows that in case of the distance of 250 km between the Institut National des Radioéléments (IRE) in Fleurus (Belgium) and Paris (France) an added value is clearly visible. Nevertheless, deficiencies in depicting meteorological processes can play an important role for the overall performance of dispersion modeling. At IMS station Spitzbergen (NOX49) and beyond (1000–1500 km) in any case no difference between the use of stack data and the use of disaggregated best estimate literature data can be found. Goodwin et al. (2021b) report an imperfect agreement between simulated and observed activity concentration at the IMS radionuclide laboratory GBL15 (UK) likely due to a combination of atmospheric transport uncertainty and the mixing of fresh plumes with plumes related to older emissions, possibly from other locations than IRE. According to these authors Europe has a substantial (and variable) radioxenon background and therefore the likelihood of measuring a single emission from a single facility (i.e., without any contamination from others) is very low. In general detections are stated to be underestimated by Goodwin et al. (2021b) when comparing measurement data with model simulations using stack emission data from IRE as input only. Consistently, Kuśmierczyk-Michulec et al. (2019) illustrated the higher added value of knowing stack data from a single emitter in the Southern Hemisphere compared to the Northern Hemisphere due to the different emitter density in the two hemispheres.

In the Third ATM-Challenge it was systematically investigated to what extent actual daily emission data from two main emitters provide an added value in simulations for four IMS stations in Europe as well as North America. Results were compared to those based on average emission values known from literature. The analysis covered runs submitted by 14 participating organizations, an extended period of investigation of three or six months and incorporated as many emitters as are known. Further, to treat the inherent uncertainties of ATM, a multi-input multi-model approach known from air-quality modeling (Kioutsioukis and Galmarini, 2014) was applied within the frame of the Third ATM-Challenge. It relies on the concept that a group of diverse, but at the same time reasonably accurate models, can be optimally combined in a multi-model ensemble approach, thus giving more accurate simulations. An optimal combination of models can be defined as one where a sub-group of all available models (or ideally the ensemble based on all models) outperforms the best individual run. According to Kioutsioukis and Galmarini (2014) there is the possibility to train an optimal ensemble per station if sufficient observations and model data are available. The training approach was another goal of the Third ATM-Challenge. For this purpose it is important to include as many above MDC values and as many known sources as possible as was the case for the Third ATM-Challenge. Involving many above MDC values implies that one should consider many possible sources which could have caused them.

The ultimate goal of any efforts related to ATM in the context of CTBT verification has always been to finally provide an (ensemble) analysis of radioxenon background levels at IMS stations frequently

affected by industrial emissions. Analysts at the National Data Centers (NDCs) should in the end ideally be able to immediately sort out suspicious events which can be clearly traced back to radiopharmaceutical isotope production facilities and/or nuclear reactors. Focus in this exercise is solely on Xe-133 as it is the most common radioxenon isotope in ambient air samples. Due to its half-life of 5.24 days and its large cumulative fission yields of up to about 7%, this radioxenon isotope has a high detectability. It is almost continuously present in the atmosphere in regions of the globe with active nuclear facilities (Achim et al., 2016). Results or concepts from this study can be easily adopted for the other xenon isotopes, bearing in mind the different half-lifes.

Finally, another aspect – as for the last challenges – was to compare results of the current challenge in terms of individual model performance. New statistical metrics were implemented and some parameters (e.g., model output resolution) were tried to set uniformly in order to enhance comparability. However, this time the model inter-comparison was only a minor purpose of the exercise.

Section 2 describes the scenario, radioxenon inventories used, IMS measurements, set-up of the atmospheric transport and dispersion models as well as the statistical analyses applied. Results and a discussion are provided in section 3. The paper is completed by conclusions drawn in section 4 covering the relevance and implications of the results found for radioxenon background modeling. Notes on individual model performance can be extracted from Appendix A.

2. Material and methods

The exercise was aimed at modeling Xe-133 activity concentrations at four IMS stations frequently affected by industrial radioxenon emissions from IRE (Belgium) and/or the Canadian Nuclear Laboratories/CNL (Canada) radiopharmaceutical plants, i.e., St. John's/CAX17 (Newfoundland and Labrador, Canada), Schauinsland/DEX33 (Freiburg, Germany), Stockholm/SEX63 (Sweden) and Charlottsville/USX75 (Virginia, United States of America) for up to six months (June to November 2014). The modeling incorporated compulsively daily emissions from these two main sources. In addition, recent publicly available annual and quarterly emission estimates for commercial nuclear power plants (NPPs) distributed over the Northern Hemisphere and broken down to daily emissions on the basis of facility operating factors as well as the 24 nuclear research reactors (NRRs) of the Northern Hemisphere with the highest estimated radioxenon releases were considered depending on the participants' own decision. The same applies to the annual emissions from additional medical isotope production and research facilities, i.e., the Mallinckrodt facility (The Netherlands), the NIIAR facility (Russia), the Karpov Institute (Russia), HFETR (China), PINSTECH (Pakistan), Lantheus Medical Imaging (Canada) and Nordion (Canada) which were disaggregated and also supplied to optionally refine predictions. Participants used well established modeling chains of their institutions adhering, however, to specified predefined output formats.

2.1. Challenge scenario

The unique opportunity of having access to radioxenon stack emission data from IRE and CNL for 2014 motivated us to set up the scenario for the Third ATM-Challenge in the Northern Hemisphere. In addition, in 2014, out of 23 noble gas IMS stations which were in operation, 15 were located in the Northern Hemisphere. Using the Continuously Emitting Sources (CES) functionality of the CTBTO/IDC software WEBGRAPE (Web connected Graphics Engine, Ctbto (2020b)), the emissions in Table 1 as well as those from commercial nuclear reactors (see **subsection 2.3**) and the operationally produced SRS (source-receptor-sensitivity) backward fields (Wotawa et al., 2003; Kuśmierczyk-Michulec et al., 2021) from 2014, it was possible to select stations for which the influence of emissions from IRE and/or CNL is dominant and results in frequent above MDC Xe-133 detections. This was to ensure that the contribution from not-well defined sources would be minimized. Following this pre-study, stations CAX17, DEX33, SEX63 and USX75 were selected. For the purpose of this exercise, it was estimated that the period with the largest number of above MDC Xe-133 measurements at the above-mentioned stations is between June and November. Taking into account that some participants of this ATM-Challenge had resources to model just a three month period, July to September was proposed as the core period with most abundant above MDC samples; the remaining months served as an optional period.

In terms of percentage contributions of individual emitters to the overall modeled signal it is evident from Fig. 1 that there is a big difference between Europe and North America on one side and between individual months on the other. At the European IMS stations DEX33 and SEX63, IRE contributes on average below 40% to the total signal. November 2014 is exceptional in so far as the IRE (located to the northwest of DEX33) contributions due to persistent south-westerlies converge to zero and commercial nuclear power plants are the dominating emitters. At the North American IMS stations CAX17 and USX75 the CNL influence adds up to nearly 100% for the whole period of investigation. According to Gueibe et al. (2017) average annual emissions of CNL with a total activity of 1.5E16 Bq were one order of magnitude larger than those of IRE with a total activity of 2E15 Bq. CNL, however, is no longer producing since October 2016.

2.2. Emission data for IRE, CNL and other radiopharmaceutical production and nuclear research facilities of the Northern Hemisphere

IRE and CNL MIPFs have given the permission to use the original stack data within this challenge. Annual estimates broken down to daily values for Curium, former Mallinckrodt, are taken from Saey (2009); for IRE, CNL, NIIAR and Karpov Institute from Kalinowski (2022); for HFETR and PINSTECH PARR-1 from Achim et al. (2016); and for Lantheus and Nordion from Pnnl (2017). An overview of the nine involved medical isotope production and research facilities together with their emissions is provided in Table 1. Their geographical location is shown in Fig. 2.

The original CNL emission file included recordings with a frequency

Table 1
Name, geographical location, stack height and Xe-133 emissons for radiopharmaceutical production and nuclear research facilities of the Northern Hemisphere considered in the Third ATM-Challenge. For IRE and CNL both actual daily stack release data as well as disaggregated annual values were used. [a]Estimates.

MIPF or institution name	Longitude [°E]	Latitude [°N]	Elevation [m a.s.l.]	Stack height [m]	Emission/day [Bq]
Institut National des Radioéléments-IRE	4.54	50.45	183	26	variable or 1.52E12
Canadian Nuclear Laboratories-CNL	−77.37	46.05	159	61	variable or 2.14E13
Mallinckrodt	4.68	52.78	not available	25[a]	2.00E09
NIIAR	49.48	54.19	not available	25[a]	5.51E12
Karpov Institute	36.57	55.08	not available	25[a]	8.25E11
HFETR	104.03	30.51	not available	25[a]	1.00E12
PINSTECH PARR-1	73.26	33.65	not available	25[a]	1.00E12
Lantheus Medical Imaging	−71.31	42.55	not available	25[a]	3.15E07
Nordion	−75.92	45.34	not available	25[a]	4.11E10

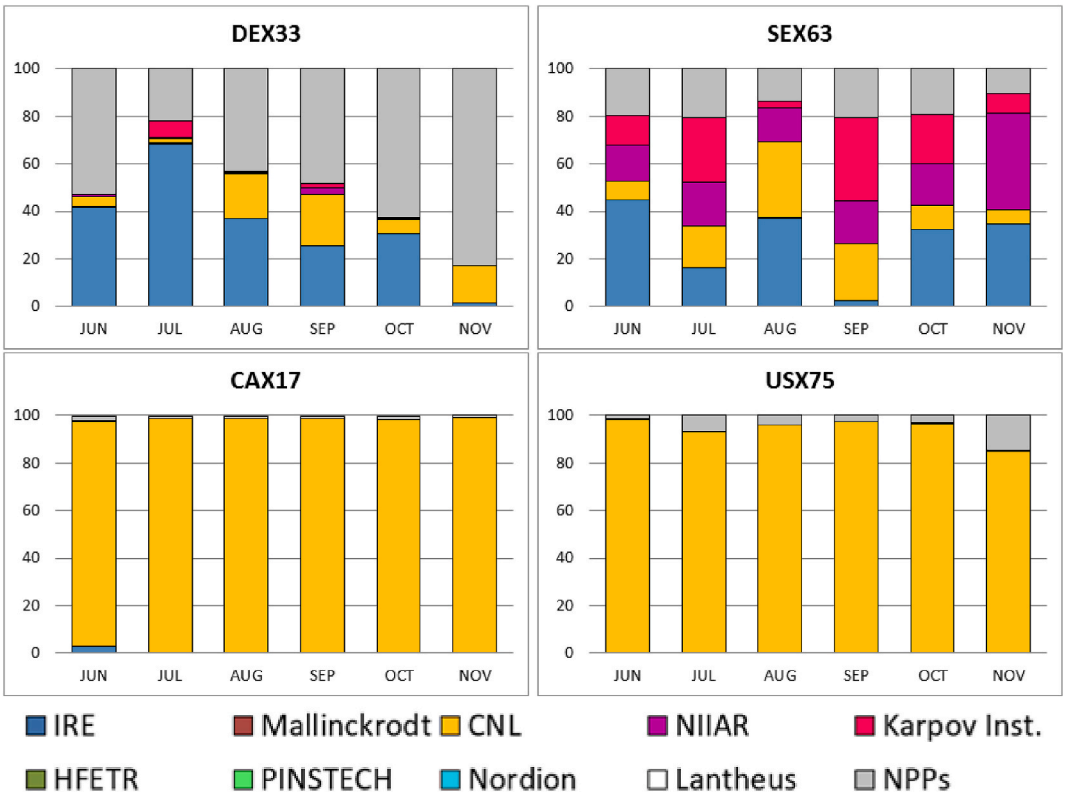


Fig. 1. Overview of different emitters' influence at the four selected IMS stations and for the six selected months of the Third ATM-Challenge. Numbers on the y-axis depict percentage contributions from individual emitters to total modeled activity concentrations based on all emitters for a given month.

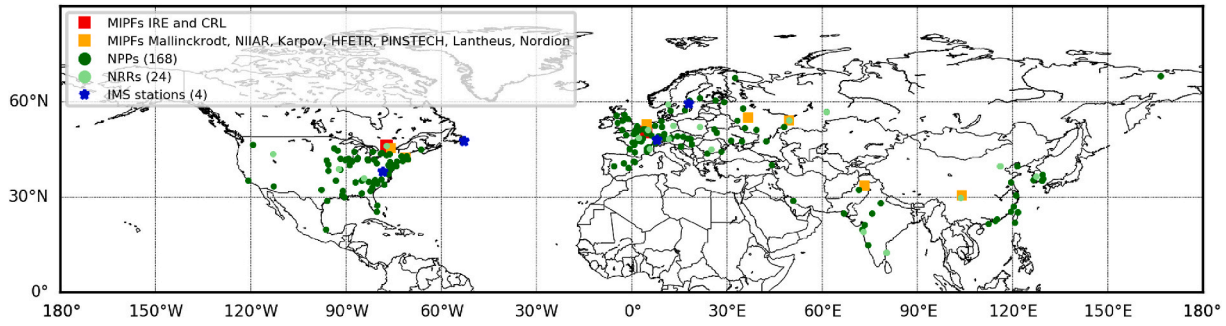


Fig. 2. Overview of the locations of different emitters and the four selected IMS stations of the Third ATM-Challenge.

of 30 s, which for the purpose of the ATM-Challenge was not suitable, and therefore had to be aggregated to daily values. However, before doing that, the quality of the data was carefully checked, and all outliers as well as missing records related to periods of calibration and technical maintenance were identified. A similar procedure was applied to stack emission data from the IRE facility which come at 15 min temporal resolution.

2.3. Emission data for NPPs and NRRs of the Northern Hemisphere

The emissions of NPPs are taken from Kalinowski and Tatlisu (2020a) and Kalinowski and Tatlisu (2020b). The daily emissions of NPPs range over many orders of magnitude from 1E5 to more than 1E10 Bq/day. The emissions of NRRs are taken from Kalinowski et al. (2021). Since only the NRRs with the highest emissions are relevant in comparison to NPPs with the lowest release rates, a cut was made at 1E7 Bq/day. There were 24 NRRs in the Northern Hemisphere that exceeded the latter release rate. They are listed in Table 2. The geographical location of the NPPs and selected NRRs is as well shown in Fig. 2.

2.4. Radioxenon IMS measurements at CAX17, DEX33, SEX63 and USX75

Measurement data used in this exercise were collected by two different noble gas systems employed in the IMS network (Table 3). The SPALAX system (Fontaine et al., 2004) at CAX17 and DEX33 collects air samples of about 80 m^3 at ambient temperatures during cycles of 24 h. Each sample is dried, concentrated and purified to produce a final stable xenon volume of about 7.5 ml/air sample. The spectrum acquisition is automatically performed by high resolution γ-spectrometry (HPGe detector). The SAUNA system (Ringbom et al., 2003) at SEX63 and USX75 collects air samples of about 7 m^3 at ambient temperatures during cycles of 12 h (by combining two samples of 3.5 m^3, each one being collected during 6 h). From each sample, the system extracts a unique stable xenon volume of about 1.0 ml/air sample. The spectrum acquisition is performed by beta-gamma coincidence detection technique (BC404

Table 2
Country, facility name, geographical location, and Xe-133 emissions for the 24 NRRs of the Northern Hemisphere considered in the Third ATM-Challenge.

Country	Facility Name	Longitude [°E]	Latitude [°N]	Emission/ day [Bq]
United States of America	ATR	−112.97	43.59	3.85E08
Russian Federation	VK-50	49.48	54.19	2.43E08
India	DHRUVA	72.90	19.01	1.54E08
Russian Federation	SM-3	49.48	54.19	1.07E08
Russian Federation	MIR.M1	49.48	54.19	9.48E07
United States of America	HFIR	−84.30	35.92	6.05E07
Belgium	BR-2	5.10	51.22	5.93E07
France	ILL High Flux Reactor	5.69	45.21	5.53E07
Russian Federation	BOR-60	49.52	54.20	5.51E07
China	HFETR	103.55	29.73	4.45E07
India	FBTR	80.17	12.55	2.96E07
Canada	NRU	−77.36	46.05	2.75E07
Poland	MARIA	21.34	52.12	2.54E07
Korea, Republic of	HANARO	127.37	36.43	2.49E07
Germany	FRM II	11.68	48.27	2.02E07
Russian Federation	IVV-2M	61.32	56.84	1.96E07
Norway	HBWR	11.40	59.13	1.66E07
Romania	TRIGA II PITESTI - SS CORE	24.98	44.95	1.66E07
Germany	BER-II	13.13	52.41	1.54E07
France	OSIRIS	2.15	48.73	1.43E07
China	CEFR	116.03	39.74	1.33E07
United States of America	MURR Univ. of Missouri-Columbia	−92.34	38.93	1.32E07
China	CARR	116.14	39.75	1.22E07
France	Orphee	2.15	48.73	1.16E07

plastic scintillator combined with a NaI detector). Both detection systems have MDCs below 1 mBq/m^3 for Xe-133. Over the period under investigation for this exercise, the observed maximum average MDC of all four IMS stations occurs at DEX33 with 0.33 mBq/m^3. After spectrum acquisition at the stations, spectra are automatically sent to the IDC of CTBTO where they are processed, quality-controlled and analysed. Analysis reports are then provided to State Signatories. Like for the Second ATM-Challenge samples falling below the MDC were set to zero for the evaluation process because they cannot be reasonably quantified based on the radioxenon measurement. The geographical location of the different stations is shown in Fig. 2. Relevant meta data information on the selected IMS noble gas measurement systems is provided in Table 3.

2.5. The participants and their ATM set-ups

14 organizations from 10 countries (Australia, Austria, Canada, France, Japan, Republic of Korea, Sweden, United Kingdom, USA and Vietnam) took part in the Third ATM-Challenge (Table 4). An overview of the models used is provided in Table 5 together with some key parameters of the model set-ups. Six Lagrangian models as well as one Eulerian model (LdX-C3X) and one mixed model (HYSPLIT-GEM) were used. ATM models employed comprise HYSPLIT(-GEM) (six participants & nine submissions; Stein et al. (2015)), FLEXPART (five participants & six submissions; Stohl et al. (1998, 2005); Pisso et al. (2019)), MLDP (one participant & three submissions; D'Amours et al. (2010, 2015)), LdX-C3X (one participant & one submission; Groëll et al. (2014); Tombette et al. (2014)), LODI (one participant & one submission; Ermak and Nasstrom (2000); Larson and Nasstrom (2002)), GEARN (one participant & one submission; Terada and Chino (2008); Terada et al. (2013)) and LADAS (one participant & eight submissions; Suh et al. (2006)).

Meteorological input data (or any other kind of input data apart from emissions) were gathered by the participants themselves. The list of meteorological drivers comprises largely operational or re-analysis European Center for Medium-Range Weather Forecasts (ECMWF; Simmons et al. (1989); Ecmwf (2018)) and U.S. National Oceanic and Atmospheric Administration's (NOAA's) National Weather Service's National Centers for Environmental Prediction (NCEP; Ncep (2003); Saha et al. (2014); Noaa (2020)) products, but also the Australian Community Climate and Earth-System Simulator-Global (ACCESS-G) version APS1 (Puri et al., 2013), the Canadian Meteorological Center (CMC) Global Deterministic Prediction System (GDPS; Buehner et al. (2013, 2015); Charron et al. (2012)), the Action de Recherche Petit Echelle Grand Echelle (ARPEGE; Déqué et al. (1994); Déqué and Piedelievre (1995)), the Weather Research and Forecasting Model (WRF; Skamarock et al. (2008)) driven with the Grid Point Value global model from the Japan Meteorological Agency (GPV; Jma (2019)) and the Unified Model Global Data Assimilation and Prediction System (UM-GDAPS; Davies et al. (2005); Kma (2018, 2021); Metoffice (2021)) products.

The ATM and meteorological driver mix is similar to the one of the second challenge (Maurer et al., 2018) with a slight increase in diversity. Whereas NAME (Jones et al. (2007); driven with UM-GDAPS) is not part of the ensemble in the Third ATM-Challenge, GEARN and LADAS could be added to the pool of models with corresponding meteorological drivers WRF-GPV and UM-GDAPS (Unified Model (UM) Global Data Assimilation and Prediction System (GDAPS) product by the Korea Meteorological Administration (KMA) under license from the UK MetOffice). One HYSPLIT run was driven with ACCESS-G APS1 of the Australian Bureau of Meteorology instead with NCEP-GFS as in the second challenge and two other ones with the recently available ECMWF-ERA5 (Ecmwf, 2020) re-analysis substituting the previous ERA-Interim ECMWF product. ACCESS-G version APS1 is also based on UM-GDAPS, thus incorporating the model in fact twice in the challenge. An aim was that not all FLEXPART runs are driven by ECMWF input data and not all HYSPLIT runs by NCEP data as was largely the case for the Second ATM-Challenge.

Participants were encouraged to submit multiple runs, but were asked to submit at least one run with *three-hourly meteorological input resolution* and *0.5° horizontal output resolution* spanning the *time frame July to September 2014*. If computationally possible the full six months period June to November 2014 was considered (managed by 13 out of the 16 participants). Particle aging (i.e., tracking the particles of a specific release just for a pre-defined amount of time) was employed to partly overcome computational limitations. Participants applied different meteorological drivers in different spatial resolutions (however, mostly 1.0° and 0.5°). There were up to eight contributions per

Table 3
Relevant information on the selected IMS noble gas measurement systems.[1] Average collection stop. Collection stops often vary somehow, especially over extended time periods.

Station code	Longitude [°E]	Latitude [°N]	Elevation [m a.s.l.]	System	Collection stop (UTC) [HH:MM][1]	Collection time [h]
CAX17	−52.74	47.59	205	SPALAX	13:30	24
DEX33	7.92	47.92	1208	SPALAX	06:00	24
SEX63	17.95	59.40	51	SAUNA	00:00	12
USX75	−78.40	38.00	104	SAUNA	04:00	12

Table 4
Participants of the Third ATM-Challenge. Organizations participating in the second challenge are printed **bold**. *Involved in drafting the challenge.

Organization Abbreviation	Name(s) of participant(s)	Organization full name	Submission (s)
ARPANSA/ BOM	Blake Orr, Alan Wain	Australian Radiation Protection and Nuclear Safety Agency, Yallambie/Miranda, Australia and Bureau of Meteorology, Melbourne, Australia	$ARPANSA_{1-2}$
AWE/UK-NDC	Rich Britton, Ashley Davies, Matthew Goodwin	Atomic Weapons Establishment/United Kingdom-National Data Center, Aldermaston, Reading, UK	AWE_{1-2}
BOKU	Petra Seibert	University of Natural Resources and Life Sciences, Institute of Meteorology and Climatology, Vienna, Austria	$BOKU_{1-2}$
CTBTO	Michael Schoeppner	Comprehensive Nuclear-Test-Ban Treaty Organization, Onsite Inspection Division, Vienna, Austria	CTBTO-1
CTBTO*	Jolanta Kuśmierczyk-Michulec	Comprehensive Nuclear-Test-Ban Treaty Organization, International Data Center, Vienna, Austria	$CTBTO\text{-}2_{1-2}$
ECCC-CMC	Alain Malo	Environment and Climate Change Canada, Meteorological Service of Canada, Canadian Meteorological Center, Environmental Emergency Response Section, RSMC Montréal, Dorval, Québec, Canada	CMC_{1-3}
FOI	Anders Axelsson, Anders Ringbom	Swedish Defence Research Agency, Stockholm, Sweden	FOI
IRSN	Arnaud Quérel, Olivier Saunier, Denis Quélo	French Institute for Radiation protection and Nuclear Safety, Fontenay-aux-Roses, France	IRSN
JAEA	Yuichi Kijima	Japan Atomic Energy Agency, Tokai, Ibaraki, Japan	JAEA-1
JAEA	Akiko Furuno	Japan Atomic Energy Agency, Tokai, Ibaraki, Japan	JAEA-2
KAERI	Kihyun Park, Kyung-Suk Suh	Korea Atomic Energy Research Institute, Daejeon, Republic of Korea	$KAERI_{1-8}$
LLNL	Lee G. Glascoe, Donald D. Lucas, Sara Cicchi, Phil Vogt	National Atmospheric Release Advisory Center at the Lawrence Livermore National Laboratory, Livermore, California, USA	LLNL
NOAA-ARL	Alice Crawford, Ariel Stein, Tianfeng Chai, Fong Ngan	National Oceanic and Atmospheric Administration Air Resources Laboratory, College Park, Maryland, USA	$NOAA\text{-}ARL_{1-3}$
PNNL	Paul W. Eslinger	Pacific Northwest National Laboratory, Richland, Washington, USA	PNNL
VINATOM	Pham Kim Long		VINATOM
ZAMG*	Christian Maurer	Vietnam Atomic Energy Institute, Hanoi, Vietnam Zentralanstalt für Meteorologie und Geodynamik, Vienna, Austria	ZAMG

organization. Depending on the IMS station overall 29 to 31 modeled time series were gathered.

All participants modeled daily IRE & CNL releases based on a date-time template and unit emissions (183 per facility, 1 Bq per daily release) tracked individually as it was the case for the Second ATM-Challenge for ANSTO emissions (see Maurer et al. (2018), specifically **subsection 2.7** therein). Individual SRS values had to be scaled and summed for each station and collection period during the evaluation. A unit emission approach had already been applied by Maurer et al. (2018). It has several advantages: 1) Most importantly it enabled participation in the challenge without the absolute need to sign a vDEC agreement (Ctbto, 2020a) in order to access confidential emission data, 2) different kinds of emissions can be taken into account even after the runs are completed, and 3) it was hardly possible to tune the runs towards IMS measurements which most of the participants have access to. However, due to experiences recently gathered within the CTBT community (see the introduction) *only daily resolution of stack data* was considered. Higher temporal resolutions were no longer taken into account for the Third ATM-Challenge.

Nuclear reactors and remaining radiopharmaceutical production and nuclear research facilities were added by most of the participants (for all but six runs) employing daily (NPPs) or disaggregated annual (NRRs and remaining radiopharmaceutical production and nuclear research facilities) releases. However, including NPPs (168), NRRs (24) and additional radiopharmaceutical facilities was optional. These releases – if taken into account – did not need to be tracked individually as no scaling of releases was intended. For all those participants that simulated only IRE and CNL emissions (ARPANSA, IRSN, JAEA-1 and JAEA-2) ZAMG's results for nuclear reactors and the remaining radiopharmaceutical facilities were added for major parts of the evaluation in agreement with participants. In case of NOAA-ARL HYSPLIT-GEM results for nuclear reactors and remaining radiopharmaceutical facilities were added to NOAA-ARL HYSPLIT results.

Since IMS radioxenon measurement systems report activity concentrations referenced to a standard atmosphere (0° C, 1013.25 hPa; Ctbto (2008)), modeled ambient activity concentrations at DEX33 - with the exception of BOKU's and VINATOM's submissions which had already been accounting for this effect - were corrected by multiplying the provided time series values with the quotient of standard density (ρ = 1.2754 kg/m^3) and average output layer ambient density. This procedure improved scores for those participants (e.g., ZAMG or CMC) who had sampled a model output layer at Mount Schauinsland several 100 m above the surface as it is depicted in a smoothed model topography.

As already stated above one aim of the current exercise was to increase model diversity as much as possible under the given constraints (i.e., participating institutions being largely confined to well-established atmospheric transport models like FLEXPART and HYSPLIT and prevalent meteorological data sources like ECMWF and NCEP). Therefore, set-up parameters of the dispersion runs were controlled as much as possible in the current challenge. Before starting their runs, participants were asked to give feedback to the scenario team regarding their planned set-up (see also Table 5) within two weeks after the launch of the challenge and to tell the scenario team how flexible they are in possibly changing the set-up. The scenario team finally asked some participants for a

Table 5
Models and set-ups used. Columns from left to right indicate the ID of the submission, the name of the atmospheric transport model (ATM), the meteorological model providing the meteorological input (NWP), the horizontal and vertical resolution (with number of model levels below 2.5 km including the surface level) of the meteorological input (Meteorological input resolution Δx & Δz), adaptive time step used yes or no (with a constant time step or the factor by which the time step must be smaller than the Lagrangian time scale provided), the bottom and top height(s) of the ATM output layer(s) starting at the surface (SFC) used for activity concentration averaging and temporal output resolution (Output resolution z_b - z_t & Δt), the ATM simulation direction (forward in time from the source/FWD or backward in time from a receptor/BWD), IRE and CNL considered as emitters only yes or no, the total number of particles released per hour (Total #particles/hour released, k = 1000, M = 1 million), the simulation period and the number of days an individual release was tracked (Release cut-off threshold).

ID	ATM	NWP	Meteorological input resolution		Adaptive time step (seconds or factor)	Output resolution		Simulation direction	IRE & CNL only	Total #particles/ hour released	Simulation period	Release cut-off threshold [days]
			Δx (°)	Δz (m) (# levels below 2.5 km)		z_b - z_t (m)	Δt (h)					
$ARPANSA_{1,2}$	HYSPLIT 4.2.0 MPI	ACCESS-G APS1	0.375 x 0.5625	20-7870 (19)	Yes	SFC-100 or 100-500[a]	1	FWD	Yes	28.8 k	June–November	14
AWE_1	HYSPLIT 4.2.0	NCEP-GFS-operational	1.0	200-400 (8)	Yes	SFC-200	12/24	FWD	No	2 M	June–November	simulation duration
AWE_2	HYSPLIT 4.2.0	NCEP-GFS-operational	0.5	10-540 (23)	Yes	SFC-200	12/24	FWD	No	2 M	June–November	simulation duration
$BOKU_1$	FLEXPART 10.4	ECMWF-IFS-operational	0.5	10-5700 (30)	Yes (3)	Receptor output	0.5	FWD	No	187.5 k[b]	June–November	14
$BOKU_2$	FLEXPART 10.4	ECMWF-IFS-operational	0.5	10-5700 (30)	Yes (3)	SFC-100[c]	24	BWD	No	166.7 k (41.6 k per station)	June–November	14
CMC_1	MLDP	GDPS-analysis[d]	0.22x0.35	40-1500 (15)	No (600 s)	SFC-500	1	FWD	No	6.6 M[e]	June–November	21
CMC_2	MLDP	GDPS-analysis[d]	0.22x0.35	40-1500 (15)	No (600 s)	500–1000	1	FWD	No	6.6 M[e]	June–November	21
CMC_3	MLDP	GDPS-analysis[d]	0.22x0.35	40-1500 (15)	No (600 s)	SFC-100	1	FWD	No	6.6 M[e]	June–November	21
CTBTO-1	FLEXPART 9.3.2	ECMWF-IFS-operational	0.5	10-5700 (30)	No (900 s)	SFC-150	1	BWD	No	283.33 k	July–September	14
CTBTO-$2_{1,2}$[f]	FLEXPART 9.3.2	ECMWF-IFS-operational	0.5	10-5700 (30)	No (900 s)	SFC-150 or SFC-700[g]	1	FWD	No	283.33 k[h]	June–November	14
FOI	HYSPLIT 4 rev.761 MPI	NCEP-GDAS	1.0	111-4000 (9)	Yes	SFC-100	3	FWD	No	150 k	June–November	14
IRSN	LdX-C3X	ARPEGE	0.5	40-2200 (9)	No (600 s)	SFC-40	1	FWD	Yes	Eulerian model	June–November	simulation duration
JAEA-1	HYSPLIT 4.2.0	NCEP-GDAS	0.5	40-3400 (18)	Yes	SFC-100	1	FWD	Yes	15 k	July–September	14
JAEA-2	GEARN	GPV-WRF	0.5	50-750 (12)	No (30 s)	SFC-20	2	FWD	Yes	104.2 k	July–September	20
$KAERI_{1-8}$[i]	LADAS	UM-GDAPS (KMA)	0.35 x 0.23	100-3500 (5)	No (300 s)	SFC-100 for SEX63 and USX75, 100–550 or 100–200 for CAX17, 550–800 or 700–800 for DEX33	1	FWD	No	1.5 M[j]	June–November	30
LLNL	LODI	NCEP-GFS-operational-ADAPT	0.5	15-1200 (30)	Yes	SFC-20/34 (terrain dependent)[k]	1	FWD	No	166.7 k	June–November	30
$NOAA\text{-}ARL_1$	HYSPLIT-GEM 4.2.0	NCEP-GDAS	1.0	200-400 (8)	No (600 s)	SFC-100	12/24	FWD	No	Eulerian model	June–November	31
$NOAA\text{-}ARL_2$	HYSPLIT 4.2.0	ECMWF-ERA5	0.5	200 (11)	No (600 s)	SFC-200	12/24	FWD	Yes	1 M	June–November	31
$NOAA\text{-}ARL_3$	HYSPLIT 4.2.0	ECMWF-ERA5	0.5	200 (11)	No (600 s)	SFC-500	12/24	FWD	Yes	1 M	June–November	31
PNNL	HYSPLIT 4 svn:951	NCEP-GDAS	0.5	10-540 (23)	Yes	SFC-100	1	FWD	No	980 k[l]	June–November	24
VINATOM	FLEXPART 10.4	NCEP-CFSv2	0.5	111-25908 (9)	Yes (3)	Receptor output	0.5	FWD	No	187.5 k[b]	June–November	14
ZAMG	FLEXPART 10.3-beta	ECMWF-IFS-operational	0.5	10-5700 (30)	No (900 s)	SFC-100 or 500–600[m]	1	FWD	No	222.2 k[n]	June–November	15

[a] for DEX33 and $ARPANSA_2$, second submission for DEX33 only.

[b] 62.5 k for IRE, CNL and all remaining emitters together each.
[c] 0.08°x0.05° resolution around IRE and CNL, 0.5° for the remaining hemisphere.
[d] 6-hourly resolution only.
[e] for each emitter (subgroup) between 416.7 k to 583.3 k.
[f] for CAX17 and CTBTO-2_2 longitude rounded to −53.0 instead of −52.5, second submission for CAX17 and DEX33 only.
[g] for DEX33 and CTBTO-2_2.
[h] 100 k for IRE & CNL each and 83.33 for the remaining emitters all together.
[i] $KAERI_1$: 0.35°x0.23° output resolution, horizontal and vertical diffusion constants $k_h = 2.5 \times 10^4$ m^2/s, $k_v = 1.0$ m^2/s, sampling the layer 100–550 m a.g.l. for CAX17 as well as 550–800 m a.g.l. for DEX33; $KAERI_2$: same as $KAERI_1$, but with 0.5°x0.5° output grid cells centered on the IMS stations and sampling the layer 100–200 m a.g.l. for CAX17 as well as 700–800 m a.g.l. for DEX33; $KAERI_3$: same as $KAERI_2$, but with 0.35°x0.23° output grid cells centered on the IMS stations; $KAERI_4$: same as $KAERI_2$, but with 0.1°x0.1° output grid cells centered on the IMS stations; $KAERI_{5-8}$: same as $KAERI_{1-4}$, but with $k_h = 50$ m^2/s, $k_v = 0.1$ m^2/s.
[j] 120 k for IRE and CNL each, 12 k for every additional radiopharmaceutical facility, 6 k for every NPP or NRR.
[k] 25 × 25 km horizontal output resolution (ca. 0.25°).
[l] 300 k for IRE and CNL, 420 k for NPPs, 120 k for NRRs and 140 k for additional radiopharmaceutical facilities.
[m] for DEX33.
[n] 111.1 k for IRE and CNL together and 111.1 k for all remaining emitters together.

change in set-up parameters in order to try to enhance diversity. For the three FLEXPART forward runs based on ECMWF data, two ($BOKU_1$ & ZAMG) were made with convection and subgrid-scale terrain effects on boundary layer height enabled and one (CTBTO-2) with just the latter parameterization activated. Two ($BOKU_1$ & VINATOM) of the three FLEXPART forward runs with both parameterizations turned on used actual receptor output. Finally, FLEXPART runs are different in terms of meteorological input (NCEP-CFSv2 for VINATOM versus ECMWF-IFS for the rest), the simulation direction (backward for $BOKU_2$ and for CTBTO-1 versus forward for the rest), the output grid resolution ($BOKU_2$ employed a grid with 0.08°x0.05° resolution around IRE and CNL as well as a hemispheric grid with 0.5° resolution) and the internal time step (constant for CTBTO-1, CTBTO-2 and ZAMG; adaptive for $BOKU_1$, $BOKU_2$ and VINATOM).

There was also some coordination among the groups running HYSPLIT to try to create runs which were different from each other. Four different meteorological inputs were utilized (NCEP-GFS, NCEP-GDAS, ECMWF-ERA5 and ACCESS-G APS1). Additionally, the HYSPLIT-GEM (Global Eulerian Model) was used for one run. HYSPLIT has several (physics) options which can be varied. These include different options for computing the boundary layer stability, variance of the turbulent velocity distribution, and mixing layer depth. For the most part, default parameters were used. JAEA utilized the option for estimating the mixing layer depth from the temperature profile instead of using the Numerical Weather Prediction (NWP) model input. NOAA-ARL also used wind and temperature profiles to calculate boundary layer stability rather than the default which is to use heat and momentum fluxes. The default time step is a variable time step with a minimum of 1 min. PNNL utilized the variable time step with a minimum of 5 min and NOAA-ARL used a constant 10 min time step.

2.6. Statistical methods

The statistical methods cover two different aspects: 1) run performance evaluation and 2) optimum ensemble construction.

2.6.1. Run performance evaluation

In order to make the results of the Third and Second ATM-Challenge as comparable as possible, largely analogous scores (details can be found in Maurer et al. (2018)) were evaluated and again four of them (determination coefficient (R^2), fractional bias (FB), fraction within a factor of 5 (F5) and accuracy (ACC)) combined into an equally-weighted rank number spanning the interval [0,4]:

$$\text{Rank} = R^2 + \left(1 - \frac{|\text{FB}|}{2}\right) + \text{F5} + \text{ACC} \tag{1}$$

To take additionally into account the similarity or dissimilarity in the distribution of observed and modeled sample activity concentrations, the rank was amended by the Kolmogorov-Smirnov Parameter (KSP):

$$\text{Rank}_{\text{KS}} = \text{Rank} + (1 - \text{KSP}/100) \tag{2}$$

which is provided in addition to the original rank metric for the current challenge. Rank_{KS} takes values in the interval [0,5]. The KSP is defined as:

$$\text{KSP} = Max|D(M_k) - D(O_k)|100\% \tag{3}$$

where D is the cumulative distribution of the modeled and observed concentrations over the range of k values such that D is the probability that the concentration will not exceed a modeled M_k or observed O_k. The score measures the ability of the model to reproduce the observed concentration distribution regardless of space and time. The maximum difference between any two distributions cannot be more than 100%.

Further, the bias-corrected root mean square error BC_{RMSE} was introduced to replace the root mean square error (RMSE) as employed in the Second ATM-Challenge:

$$BC_{RMSE} = \sqrt{\frac{1}{N}\sum[(M_k - \overline{M}) - (O_k - \overline{O})]^2} \quad (4)$$

with $\overline{M}$ and $\overline{O}$ being the average modeled and observed values. BC_{RMSE} is subsequently involved in another skill score definition.

Alternatively compared to **equations (1) and (2)** and following Taylor (2001) the bias-corrected root mean square error, BC_{RMSE}, considered implicitly via observed and modeled standard deviations in combination with the correlation coefficient, can be used as basis to define a skill score S_r, however, attributing higher weight to the ratio of standard deviations compared to R:

$$S_r = 2(1+R)\left(\frac{\sigma_m}{\sigma_o} + \frac{\sigma_o}{\sigma_m}\right)^{-2} \quad (5)$$

with σ_m and σ_o being the standard deviations of modeled and observed values. The definition ensues the idea that for any given variance, S_r should increase monotonically with increasing correlation and for any given correlation the score should increase as the modeled variance approaches the observed variance. As σ_m -> σ_o and R -> 1 (equivalent to BC_{RMSE} -> 0), the skill score reaches unity (interval [0,1]). However, at lower correlation values, models with too little variability are penalized in **equation (5)** despite a possible reduction in the BC_{RMSE}.

If one wants to include also the fractional bias FB into the skill score, Seibert (2004) suggests a formulation to transform the fractional bias into a skill score S_b spanning the interval (0,1]:

$$S_b = \frac{1}{1 + b\mathrm{FB}^2} \quad (6)$$

The parameter b is introduced in order to yield a skill score converging to 0 for large values of FB^2 (upper limit of FB^2 is 4). A value of $b = 10$ appears to give a relationship fulfilling Seibert (2004)'s subjective idea about such a skill score. Finally, both individual skill scores can be combined into a total skill score SS:

$$SS = \alpha S_r + (1-\alpha)S_b \quad (7)$$

The value of α is rather arbitrary and depends on the application. $\alpha = 0.5$ as adopted in the current study might be acceptable. An additive and not a multiplicative combination of the two scores is suggested because a model that has skill either in reproducing the mean (expressed via S_b) or in reproducing the pattern (expressed via S_r) should be attributed some total skill; the product of the two scores would be zero with one of the factors being zero.

2.6.2. Optimum ensemble construction

The ensemble analysis is performed according to the sequence of steps listed below. The various steps have the scope of selecting the ensemble features that allow *to harness the best accuracy while preserving the variability*. As demonstrated by, e.g., Solazzo et al. (2013) and references therein, this is achieved by promoting diversity among the members of the ensemble and thus discarding any redundant information retaining the highest level of independence among the remaining models. It is important that the pool of available members is numerous for the feature selection algorithms presented hereafter to be successful. The analysis is performed for the period of June to November 2014 independently at the four locations where measurements are available, namely CAX17, DEX33, SEX63 and USX75, and limited to above MDC samples.

- *Rank histogram*: The first analysis performed with the available ensembles is to construct rank histograms or Talagrand diagrams (Talagrand et al., 1997) where the number of bins equals the number of ensemble members +1 to equally subdivide the range of all ranked simulations. Subsequently, the observations are distributed in the corresponding bins. A balanced ensemble will result in a flat histogram as bins within the range of modeled values have equal probability of accommodating observations.
- *Analysis of the correlation matrix*: In the presence of large ensembles one should investigate the occurrence of repeating information that would not add any new element to the ensemble analysis by inspecting the correlation matrix of all model results and exclude those that correlate above a predefined threshold value (in this work 0.95). This enables skipping redundant data sets in part of the subsequent calculations which are numerically intensive.
- *Determination of the effective number of models*: This analysis provides an estimate of the level of independence of the ensemble members and the minimum number of degrees of freedom (each model is intended as a degree of freedom, as it would theoretically bring new, unexplored information to the ensemble) sufficient to reproduce the variance of the observations. From Bretherton et al. (1999) we calculate the number of effective models sufficient to reproduce the variability of the full ensemble N (the multi-model ensemble generated with all available members) as:

$$\mathrm{Neff} = \frac{\left(\sum_{n=1}^{N}\lambda_k\right)^2}{\sum_{n=1}^{N}\lambda_k^2} \quad (8)$$

with λ_k as eigenvalues of the $corr(d_i, d_j)$ matrix, the matrix of the linear correlation coefficients between any pair d_i, $d_j(i, j = 1 \ldots N)$. d is a metric defined according to Pennel and Reichler (2011):

$$d_m = e_m - Re_M \quad (9)$$

where the index m identifies the model, e_M is the multi-model error (the average over all individual models' errors) and R is the Pearson correlation coefficient between e_m, the error of model m, and e_M. The removal of e_M in **equation (9)** makes model errors more dissimilar and uncovers "hidden" features that are otherwise outweighed by overarching commonalities. The aim of the metric d_m is to establish similarities among models beyond the more obvious ones induced by shared inputs and/or common parameterisations. Only if all eigenvalues were equal to unity, **equation (8)** would return $N_{eff} = N$, which corresponds to the situation where all eigenvector directions are present and equally relevant. On the other hand, if all error fields were similar, all eigenvectors would collapse onto one and $N_{eff} = 1$.

- *Determination of statistical indicators for all possible ensemble linear combinations*: For all combinations of n ensemble members, where n goes from one to the total number of models, Root Mean Square Errors (RMSEs) and Pearson Correlation coefficients (Rs) are calculated. In case of redundancy, the optimal combination (min. RMSE and max. R) is not reached for the full-member ensemble but rather for a subset of it. The smaller the subset, the higher the redundancy.
- *Promoting diversity by using weighted ensemble combinations:* This analysis allows to determine the level of proximity of the various ensemble members and the level of clustering of the results. It is based on a selection of thresholds of the values d_m and the clustering is represented in terms of bifurcations in a dendrogram whenever a threshold level is passed. The level of proximity also determines the weight of each member: the higher the associativity (lower d_m), the lower the weight and vice versa. The optimal weights applied to the modeled time series are those that minimize the RMSE of the ensemble. The approach generates a multi-model ensemble that penalizes commonalities without discarding any member. Weights are assigned to individual runs according to i) the optimal number of clusters, and ii) the number of models composing each cluster. The algorithm starts with a minimum of two clusters and assigns a weight of $\frac{1}{2}$ to each (the total weight to be assigned sums up to unity). If there are, for example, 20 members in total and, for example, the first cluster is composed of fifteen members and the second cluster of five, then to each member of the first cluster a weight of $\frac{1}{30}$ ($\frac{\frac{1}{2}}{15}$) is assigned

while to the members of the second cluster a weight of $\frac{1}{10}$ ($\frac{1}{5}$). The next loop of the algorithm then checks whether a larger number of clusters generates a weighted ensemble mean with lower RMSE. The looping stops when the RMSE cannot be minimized any further.

3. Results and discussion

3.1. Evaluating the added value of IRE and CNL stack emission data versus literature emission estimates

In the following subsections the added value of IRE and CNL stack emission data is inspected under different perspectives.

3.1.1. Analysis including all valid model-observation pairs

In a first step it was investigated what the average benefit of using actual historic daily stack emission compared to disaggregated literature emission data for IRE and CNL radiopharmaceutical plants would be. For this purpose, statistics were averaged over all four selected IMS stations, over all available and valid model-observation pairs between June, 8th, to November, 30th, or July, 8th, to September 30th, 2014 and over all submitted runs. Thus, a spin-up period for civil radioxenon background establishment of seven days was warranted. The latter two six and three month periods constitute the two analysis periods for the Third ATM-Challenge. In terms of Rank, $Rank_{KS}$ and SS (see Table 6) it has to be concluded that there is no average benefit from using daily IRE & CNL stack emission data over all possible valid model-observation pairs independent of the score employed. Table 6 rather slightly indicates the opposite with only R_{max}, SS and SS_{max} for stack data outperforming the respective numbers for literature estimates in case all emitters (not just IRE & CNL) are considered. However, the latter finding must in any case be considered case specific and can be due to compensating errors of different kinds of emissions as well as of emissions and transport modeling. Conclusions to be drawn are the same for both time periods. Moreover, median scores are very similar indicating no relevant seasonal influence on overall scores. If all emitters and related emissions were to be known perfectly and ATM was perfect too, stack emissions would undoubtedly always outperform literature estimates significantly. The results found simply reflect the imbalance between two emitters perfectly characterized on one side and many emission and transport related errors on the other.

Nevertheless, despite the remaining known emitters, i.e. emitters apart from IRE and CNL, may be poorly characterized, there is an added value of approximately 18% of including these roughly estimated emissions in ATM runs. This finding is supported invariably by all scores in Table 6 and is in agreement with a finding from the First ATM-Challenge (Eslinger et al., 2016) where best results for DEX33 predictions were found if additional source estimates were included apart from IRE stack emission data. However, the picture differs considerably for CAX17 and USX75 on one side and DEX33 and SEX63 on the other. The reader is referred for related details to subsection A.1 in the appendix.

Table 6
Statistical scores (median [min., max.]) over all four selected IMS stations, over all available and valid model-observation pairs from June, July, respectively, 8th, to September, November, respectively, 30th, 2014 and over all submitted runs depending on the type of emissions included in ATM runs.

	Rank	$Rank_{KS}$	SS
All emitters, stack emissions, Jun.–Nov.	2.32 [1.31, 2.78]	3.10 [1.85, 3.65]	0.49 [0.18, 0.75]
All emitters, literature emissions, Jun.–Nov.	2.51 [1.36, 2.76]	3.32 [1.90, 3.65]	0.48 [0.19, 0.59]
IRE & CNL stack emissions only, Jun.–Nov.	2.02 [0.90, 2.44]	2.67 [1.50, 3.19]	0.37 [0.07, 0.55]
IRE & CNL literature emissions only, Jun.–Nov.	2.20 [1.07, 2.59]	2.91 [1.54, 3.38]	0.41 [0.09, 0.57]
All emitters, stack emissions, Jul.–Sep.	2.39 [1.47, 2.91]	3.13 [2.02, 3.73]	0.51 [0.16, 0.71]
All emitters, literature emissions, Jul.–Sep.	2.55 [1.48, 2.93]	3.36 [2.03, 3.78]	0.48 [0.18, 0.66]
IRE & CNL stack emissions only, Jul.–Sep.	1.96 [0.96, 2.41]	2.58 [1.53, 3.10]	0.43 [0.12, 0.56]
IRE & CNL literature emissions only, Jul.–Sep.	2.24 [1.05, 2.59]	2.96 [1.64, 3.33]	0.45 [0.12, 0.65]

3.1.2. Analysis including samples with major influences of IRE & CNL

In a next analysis step it was postulated that the statistical benefit of stack emission data depends on the samples selected: The higher the influence of MIPFs IRE and/or CNL on a given sample the higher the benefit of using stack emission data should be. The necessary selection process is bound to two important aspects: 1) The influence on any observed sample can only be properly quantified if its value is greater than or equal to the MDC. 2) As a matter of fact any qualitative or quantitative influence of a certain emitter on an observed sample can only be stated based on ATM. Because of this inherent shortcoming different approaches and a step-wise procedure were applied:

1. *Selecting samples according to the MDC:* As stated in **section 2.4** below MDC samples were set to zero for the purpose of model evaluation. This inevitably leads to small values being overpredicted by models. Nevertheless, the approach does not lead to a distinct influence on Rank, $Rank_{KS}$ and SS. Averaged over all four stations scores do not change substantially when switching from all valid observed samples to those reaching at least the MDC as displayed in Fig. 3. At least 65% of samples per IMS station are retained (compare sample numbers in the caption of Fig. 3) if omitting below MDC values.
2. *Determining MIPF impact based on different ATM realizations:* MIPFs' impact using stack emission data was calculated based on (A) FLEXPART V9 backward runs or a (B) FLEXPART V9-CTBTO forward run (see runs from CTBTO-2 in Table 5) and (A) 1° or (B) 0.5° meteorological input and output resolution (operational CTBTO/IDC set-up as of 2014 or set-up for the Third ATM-Challenge 2019, respectively).
3. *Calculating MIPF impact ratios:* Finally, the following ratios (*modes 1*, *1a* and *2*) were formed. Alternatively to using observed samples (*modes 1* and *1a*), the CTBTO forward run (CTBTO-2) was also used to evaluate the ratio tagged as *mode 2* based entirely on simulations:

$$\frac{|\text{measurement} - \text{MIPFs}' \text{ contributions}|}{\text{measurement}} \quad \text{(mode 1)}$$

$$\frac{\text{measurement} - \text{MIPFs}' \text{ contributions}}{\text{measurement}} \quad \text{(mode 1a)}$$

$$\frac{\text{NPPs}' + \text{NRRs}' + \text{other radiopharmaceutical facilities}' \text{ contributions}}{\text{total simulated value}} \quad \text{(mode 2)}$$

Samples were only selected if the above quotients stayed below 50% or even 20% (equivalent to at least 50% and 80% IRE and CNL impact). *Mode 1* was chosen in order to demonstrate the effect of implictly removing sample pairs from the data set which clearly suffer from transport errors, i.e., from an overestimation by at least 20% or 50% of the measurement by just considering IRE and/or CNL emissions.

Figs. 3 and 4 reflect the selection effect based on selection criterion A1 (FLEXPART V9 bwd runs in combination with *mode 1*). Including all valid sample pairs or only sample pairs where the observed value is equal to or exceeds the MDC into the analysis yields a rather fuzzy picture with literature emissions outperforming stack emissions for SEX63 based on all scores, stack emissions outperforming literature emissions for USX75 based again on all scores and results depending on the score for CAX17 and DEX33. Selecting only samples with at least 50% or even 80% influence of IRE and/or CNL emissions leads to the expected result of stack emissions overall outperforming literature

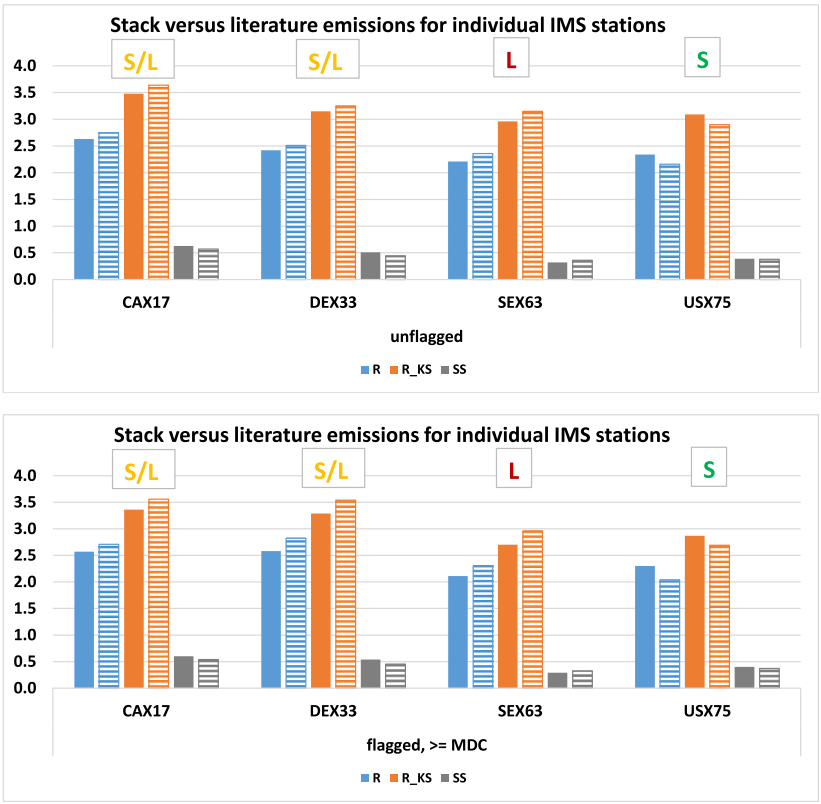


Fig. 3. Performance of stack emission versus literature emission data from IRE and CNL with no sample selection applied (upper panel) and for samples reaching at least the MDC (lower panel). Median score values per IMS station are displayed. All emitters are included. Full color bars indicate the use of stack emission data, dashed color bars the use of literature emission data. "S": stack emission data outperform literature emission data based on all three scores, "L": literature emission data outperform stack emission data based on all three scores, "S/L": stack emission data outperform literature emission data based on either Rank and $Rank_{KS}$ or based on SS. Number of involved samples per IMS station with no selection applied: CAX17: 155, DEX33: 136, SEX63: 345, USX75: 333. Number of involved samples per IMS station reaching at least the MDC: CAX17: 118, DEX33: 88, SEX63: 241, USX75: 236. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

emissions. However, differences in performance are rather small. For the 50% influence selection average improvements over all four stations vary between 5% for Rank and $Rank_{KS}$ and 20% for SS, for the 80% influence selection over three stations (excluding DEX33 due to the limited amount of involved samples) between 5% for SS and 8% for Rank and $Rank_{KS}$. One might expect a better performance for the 80% influence selection, but it is important to note that scores for both influence selections are based on quite different numbers of underlying samples. Samples are reduced to at least 1/3 if switching from the 50% to the 80% influence selection. A higher influence of IRE and/or CNL does not guarantee that a specific activity concentration can be predicted more successfully with ATM. Further, influence estimates themselves are affected by ATM uncertainties.

Not surprisingly results from approach A1 or B1 (not plotted) are most in favor of stack emission data due to the selection of samples influenced at least to 50% or 80% by IRE and/or CNL and suppressing at the same time samples with evident ATM errors. Weakest evidence for the benefit of stack data comes from approach B2 (also not plotted) where just modeled values are involved.

For an overall quantitative assessment median scores were first averaged over all four stations for stack as well as literature emissions separately for the different data sets (≥MDC, A1, B1, B1a and B2). Next, the relative score improvements for literature and stack emission data separately when switching from all above MDC samples to a specific 50% or 80% selection or from literature to stack emission data within a specific 50% or 80% influence data selection were averaged over the $Rank_{KS}$ and the SS metrics. The Rank was skipped in this averaging process because it is too similar to the $Rank_{KS}$. Finally, averaging was done over all four selection approaches (A1, B1, B1a and B2). The analysis gives an impression of the impact of sample selection versus emission data selection. In fact selecting samples according to IRE and/or CNL influence, but using literature emission data only has a bigger positive impact than switching from literature to stack emission data (Table 7). Roughly two third of the overall improvement is achieved when samples are selected according to a considerable IRE and/or CNL influence and only one third when switching from literature to stack emission data. Encompassing all individual data selections three maximum relative improvements in Table 7 are allotted to A1, two to B1 and one to B1a. On the other side three minimum relative improvements in Table 7 are allotted to B2, two to B1 and one to B1a. Whether one uses the 50% or 80% selection criterion does not matter.

3.1.3. Analysis including samples related to outstanding emissions of IRE or CNL with regard to the average

Finally, a few outstanding daily stack emissions (outstanding with respect to the mean daily value as deduced from disaggregating the annual sum) and related samples modeled by the CTBTO forward run (CTBTO-2) were selected subjectively. The aim was to bias the subsequent outcome in terms of an added value of stack emission data. This approach does not raise the claim to be a full statistical evaluation. Rather it shall exemplify the interaction of different emission sources contributing to a modeled sample as well as the entanglement of possible emission and atmospheric transport errors. Fig. 5 displays the daily stack emission profiles for IRE and CNL. As can be extracted from the figure the disaggregated annual literature estimates and the average daily values based on stack emission data agree remarkably well, especially for CNL. Starting with a large daily deviation from the average it was checked whether this specific emission is related to an above MDC

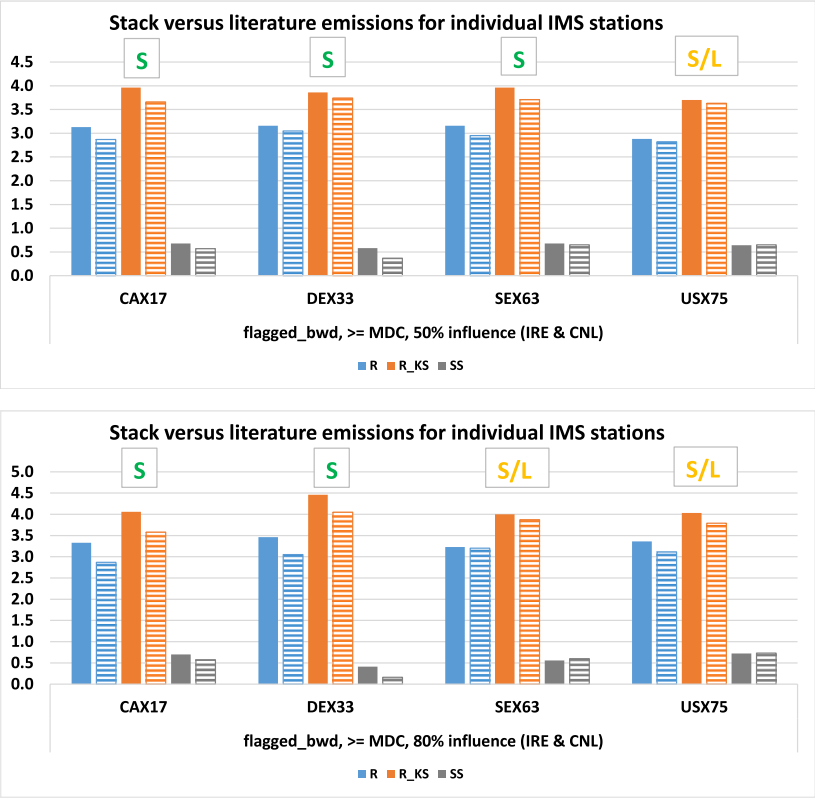


Fig. 4. Performance of stack emission versus literature emission data from IRE and CNL for samples with at least 50% (upper panel) and 80% (lower panel) influence of IRE and/or CNL. Median score values per IMS station are displayed. All emitters are included. Full color bars indicate the use of stack emission data, dashed color bars the use of literature emission data. "S": stack emission data outperform literature emission data based on all three scores, "L": literature emission data outperform stack emission data based on all three scores, "S/L": stack emission data outperform literature emission data based on either Rank and $Rank_{KS}$ or based on SS. Number of involved samples per IMS station with at least 50% influence of IRE and/or CNL: CAX17: 45, DEX33: 19, SEX63: 43, USX75: 64. Number of involved samples per IMS station with at least 80% influence of IRE and/or CNL: CAX17: 12, DEX33: 3, SEX63: 15, USX75: 21. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 7
Average relative statistical improvements [%] including minimum and maximum over all four IMS stations and all four ratio formulations evaluated (A1, B1, B1a and B2). Row labels indicate the type of emission data used (stack or literature) and the sample selection criterion (≥MDC, 50% and 80% IRE and/or CNL influence).

Data sets compared	Relative improvement [%]
Stack 50% vs. Stack ≥ MDC	29 [14, 36]
Literature 50% vs. Literature ≥ MDC	20 [14, 24]
Stack 50% vs. Literature 50%	10 [0.2, 14]
Stack 80% vs. Stack ≥ MDC	27 [22, 38]
Literature 80% vs. Literature ≥ MDC	20 [11, 34]
Stack 80% vs. Literature 80%	9 [0.5, 13]

sample. In case of the IRE emission profile samples at SEX63 were consulted, in case of the CNL emission profile samples at USX75. These two stations were chosen because of their 12-hourly collection periods which in theory should be better suited to demonstrate the added value of stack emission data compared to the 24-hourly collection periods of the two other stations investigated in the frame of the challenge. If a sample above the MDC was found related to a large daily deviation from the average emission the full footprint of it, i.e. all daily emission chunks with contribute to the sample, was marked on the respective emission profile. The footprints cover two to eight daily emission chunks. Thus, it becomes clear that – depending on the station sensitivity with respect to an emission chunk – positive and negative deviations from the average line may easily conteract.

In case of IRE and SEX63 stack emission data are beneficial in 2/3 of the selected cases (compare second and last two columns of Table 8). However, deviations are minor for four cases (in the 0.01 mBq/m^3 range of magnitude) and only add up to roughly a factor of two in two cases. It is only for the latter two samples (with collection starts at 2014090300 and 2014101300) that switching between the two types of MIPF emission data results in substantial differences with noteworthy overall simulation improvements occurring alternately for stack and literature data. Not unexpectedly, for all six selected cases the IRE (together with CNL in the case of the sample with collection start 2014100300) stack emission based contributions alone do not explain the station signal.

Table 8
Comparison of the effect of using daily stack emissions versus a disaggregated annual literature emission estimate on selected samples at SEX63. Columns from left to right with all activity concentrations in mBq/m^3 are: Collection start yyyymmddhh [UTC], observed value, IRE contributions to the sample based on daily stack emission data, IRE contributions to the sample based on disaggregated literature emission data, favored IRE emission estimate (stack "S" or literature "L") based on just considering IRE's contributions, NPPs'+ NRRs'+ other radiopharmaceutical facilities' contributions to the sample (including CNL for sample with collection start on 2014100300), favored IRE emission estimate (stack "S" or literature "L") based on considering all emitters' contributions, total modeled sample value including IRE daily stack emission data and total modeled sample value including IRE disaggregated literature emission data.

Coll. start	Obs.	IRE stack	IRE lit.	Fav. emiss. est.	NPPs'+NRRs'+ other fac.	Fav. emiss. est.	Sum stack	Sum lit.
2014070700	1.71	0.16	0.09	S?	0.11	S	0.27	0.20
2014072700	0.60	0.05	0.08	L?	0.54	S	0.60	0.62
2014090300	0.41	0.12	0.41	S?	0.29	S	0.42	0.70
2014100300	1.01	0.30	0.34	L?	1.52	S	1.82	1.86
2014101300	0.72	0.19	0.48	L?	0.15	L	0.33	0.62
2014102400	1.36	0.02	0.03	L?	1.02	L	1.04	1.05

Table 9
Same as Table 8 but for CNL and USX75.

Coll. start	Obs.	CNL stack	CNL lit.	Fav. emiss. est.	NPPs'+NRRs'+ other fac.	Fav. emiss. est.	Sum stack	Sum lit.
2014062016	5.63	4.82	3.57	S?	0.13	S	4.95	3.71
2014080216	18.51	3.92	1.07	S?	0.09	S	4.01	1.16
2014082016	0.80	0.17	0.23	L?	0.10	L	0.27	0.33
2014100116	4.80	6.85	2.61	L!	0.67	L!	7.52	3.28
2014101416	0.95	0.38	0.15	S?	0.05	S	0.43	0.20
2014110216	0.53	0.88	20.70	S!	0.36	S!	1.24	21.06

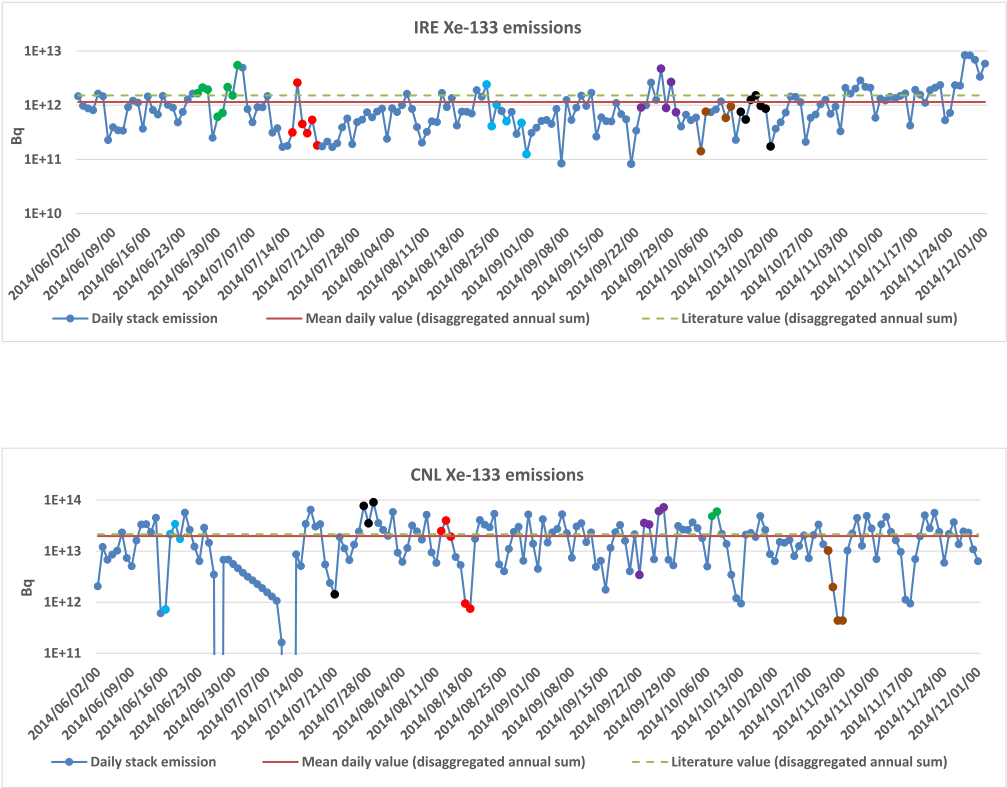


Fig. 5. IRE (upper panel) and CNL (lower panel) daily stack emission profiles together with the mean daily values as well as the disaggregated annual literature estimates. Filled, colored circles along the profiles indicate emission chunks releated to specific samples observed at SEX63 and USX75 as listed in Tables 8 and 9. Different colors are used for different samples. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Missing additional emitters or emissions are evident. Even more, NPPs', NRRs' and other radiopharmaceutical facilities' summed contributions are in 2/3 of the cases and up to two orders of magnitude larger than IRE stack emission based contributions. Thus, IRE is not the driving force for these samples. This is in agreement with the upper panel of Fig. 1.

For CNL and USX75 the picture is quite different although again in 2/3 of the cases stack data outperform literature emission data (compare second and last two columns of Table 9). Deviations in the overall modeled values resulting from employing the two different emission data sets are no longer minor reaching for four samples at least a factor of two. There is even one sample (with collection start 2014110216) with a factor of 20. For 2/3 of the selected cases the CNL stack emission based contributions alone do not explain the signal although CNL was by far the dominating emitter in the North American domain in 2014. For the sake of completeness it should be noted that Xe-133m is not present in the simulation but respective stack data is also not available for CNL. If Xe-133m was released as a substantial fraction of the Xe-133 (as it might be from a MIPF) not accounting for Xe-133m and related parent-daughter decay might result in underpredicting Xe-133. However, the underprediction may as well be an effect of errors in the simulation of atmospheric transport. NPPs, NRRs and other radiopharmaceutical facilities' summed contributions are always and up to one order of magnitude smaller than CNL stack emission based contributions. Thus, CNL is indeed the driving force for these samples. The sample with collection start 2014110216 on one side impressively demonstrates the added value stack data can have for individual samples. On the other side an ATM error becomes evident for this sample as well and - to an even larger extent - for the sample with collection start 2014100116. If simulations based on just including very well known stack emission data already exceed the observed value these simulations cannot be accurate.

3.2. Evaluating the added value of employing a multi-input multi-model ensemble

The analysis of the rank histograms (Fig. 6) shows deviations from flatness and reveals skewed ensembles. Solazzo and Galmarini (2015) discussed the relationship between flatness of the ranked histogram and redundancy. The relatively high population of bins on the far right of the histograms (USX75, SEX73) is an indication of the tendency to under-predict observations and is in agreement with results from **subsection 3.1.3**. Missing emissions and/or emitters at SEX63 and USX75 are very likely.

According to subsection 2.6.2, the data set has been reduced for computational reasons because one of the listed procedures is based on the combination of all results configurations toward the minimization or maximization of the statistical parameters RMSE and *R*. An ensemble of 26–28 members implies an unmanageable number of combinations (on the order of 2^{28}) of which not all are necessary. Many of the results are obtained from versions of the same underlying model (especially AWE, BOKU, CMC, KAERI, NOAA-ARL) which combinations would likely be "variations on the theme" of the parent model. In order to verify this hypothesis, the correlation matrix of all model results for the full ensemble has been calculated and is displayed in Fig. 7. A cut off value is introduced at $R = 0.95$ thus excluding all models that correlate above this R value and only one representative of any correlation group is retained. E.g., for SEX63 $KAERI_{4,8}$ are retained as representatives of the highly correlated $KAERI_{1,2,3,4}$ and $KAERI_{5,6,7,8}$. The obvious redundancy provided by the correlation matrix reduced the ensembles as presented in Table A.9. Starting from the reduced ensembles the values for the maximum, average and minimum RMSE where calculated for all possible combinations of members. Fig. 8 shows the results. It clearly appears that at all stations the ensembles are behaving very regularly regardless of the number of models that compose them.

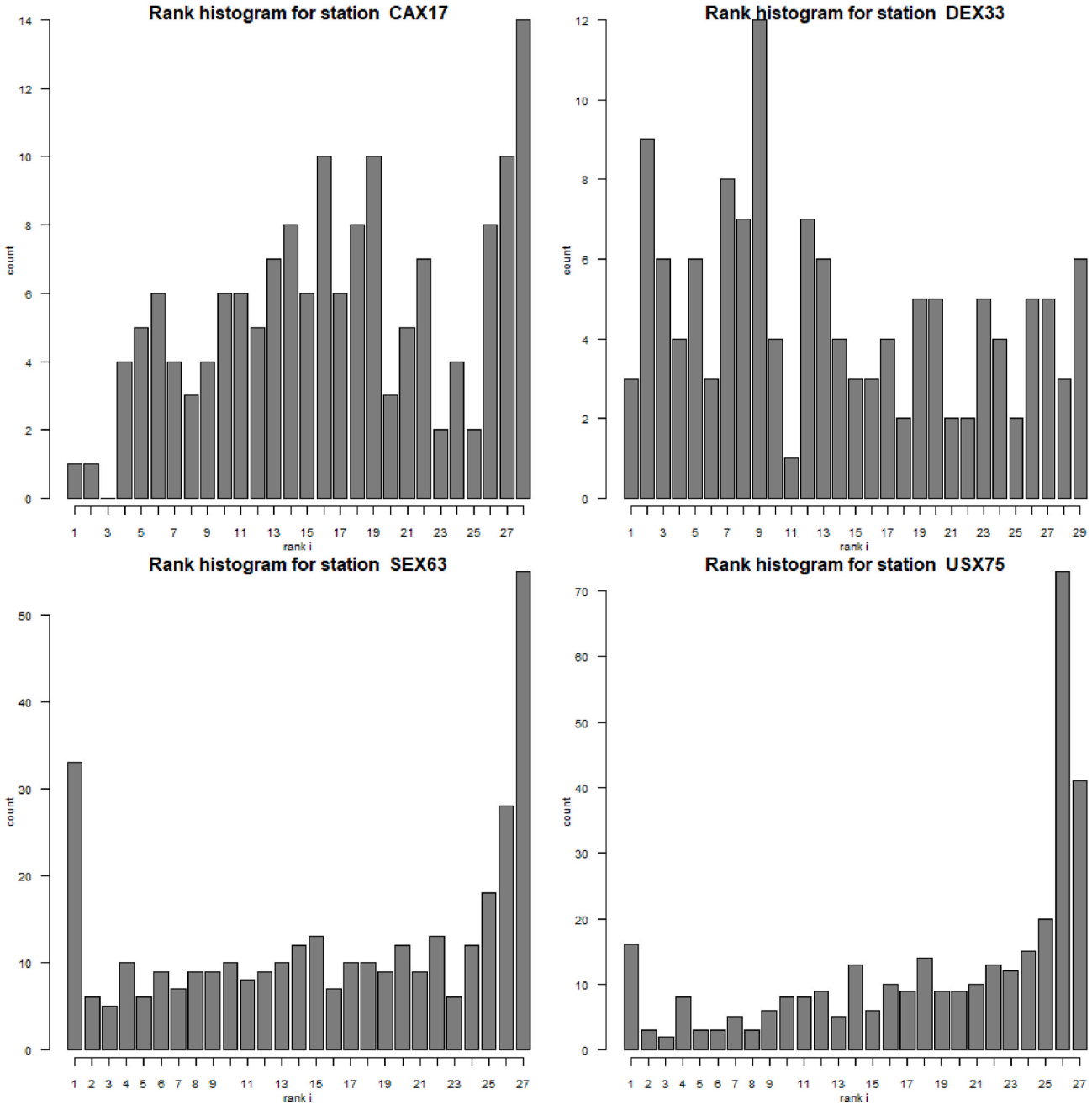


Fig. 6. Rank (or Talagrand) histograms for all four involved IMS stations and the period June to November 2014 including all samples equal to or above the MDC.

Fig. 7. Correlation matrix calculated for all model results available at the four locations for the period June to November 2014 including all samples equal to or above the MDC. The correlation of results is indicated by the color and the size of the symbol. The larger the symbol the higher the correlation. All models showing a correlation larger than or equal to 0.95 are indicated with a cross through the symbol. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 10 provides the condensed results of the performance of the reduced ensemble, the ensemble that minimizes the RMSE/maximizes the *R* and the performance of the weighted ensemble; in addition to that also the number of effective models is obtained. Several are the considerations that can be deduced from these results. The small effective number of models points toward a high level of redundancy in the original ensemble. In fact out of 16–21 models constituting the reduced ensembles at the four stations averagely 3.5 models are sufficient to reflect the actual diversity of the ensemble. The rest of the models are adding no element of independent information that could contribute to the improvement of the ensemble result. This can be attributed to two major causes: (1) Models are conceptually and structurally not too different from one another; (2) in spite of the different approaches for the specific case analysed many tend to agree in the prediction. According to Table 10 the general lack of diversity is also apparent from the comparison of the RMSE of the reduced ensemble and the best performing grouping. The differences are present but not particularly striking especially for two (SEX63 & USX75) out of four stations. This

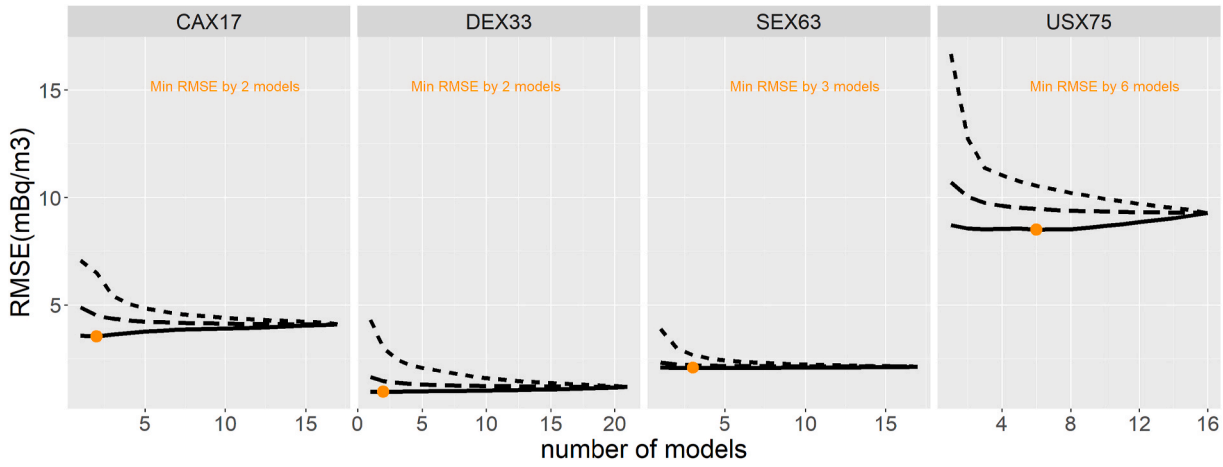


Fig. 8. Min., max. and average RMSE calculated for all combinations of models. The x-axis represents the number of models constituting the ensemble. For example, a value of 3 on the x-axis indicates that the min., max. and average values of RMSE are obtained based on all possible combinations of the available model results in groups of three. The dot laying on the minimum RMSE curve indicates the absolute minimum value obtained and the corresponding ensemble size that produces it.

Table 10
Summary table for the ensemble analysis. N: full ensemble size, $N_{r,\ R>0.95}$: reduced ensemble size following the correlation analysis, $N_{RMSEmin}$: number of models for which RMSE is minimized, AC_{imp}: improved accuracy with respect to the reduced ensemble and according to the RMSE, N_{eff}: effective ensemble size, $RMSE_{min}$: minimum RMSE obtained by optimal combination of the members of the reduced ensemble, $RMSE_r$: RMSE of the reduced (i.e., after correlation analysis) ensemble, R_{max}: maximum Pearson Correlation Coefficient obtained by optimal combination of the members of the reduced ensemble, R_r: Pearson Correlation Coefficient of the reduced ensemble, $RMSE_w$: RMSE of the weighted-diversity ensemble (based on d_m), R_w: Pearson Correlation Coefficient of the weighted diversity ensemble (based on d_m). For an explanation of the weighting procedure see subsection 2.6.2.

	N	$N_{r,\ R>0.95}$	$N_{RMSEmin}$	AC_{imp} [%]	N_{eff}	$RMSE_{min}$	$RMSE_r$	R_{max}	R_r	$RMSE_w$	R_w
CAX17	27	17	2	15	3.9	3.5	4.1	0.74	0.63	3.7	0.72
DEX33	28	21	2	20	3.5	0.95	1.2	0.74	0.53	1.1	0.71
SEX63	26	17	3	3	2.8	2.07	2.14	0.27	0.22	2.07	0.24
USX75	26	16	6	8	3.7	8.6	9.3	0.78	0.68	8.9	0.67

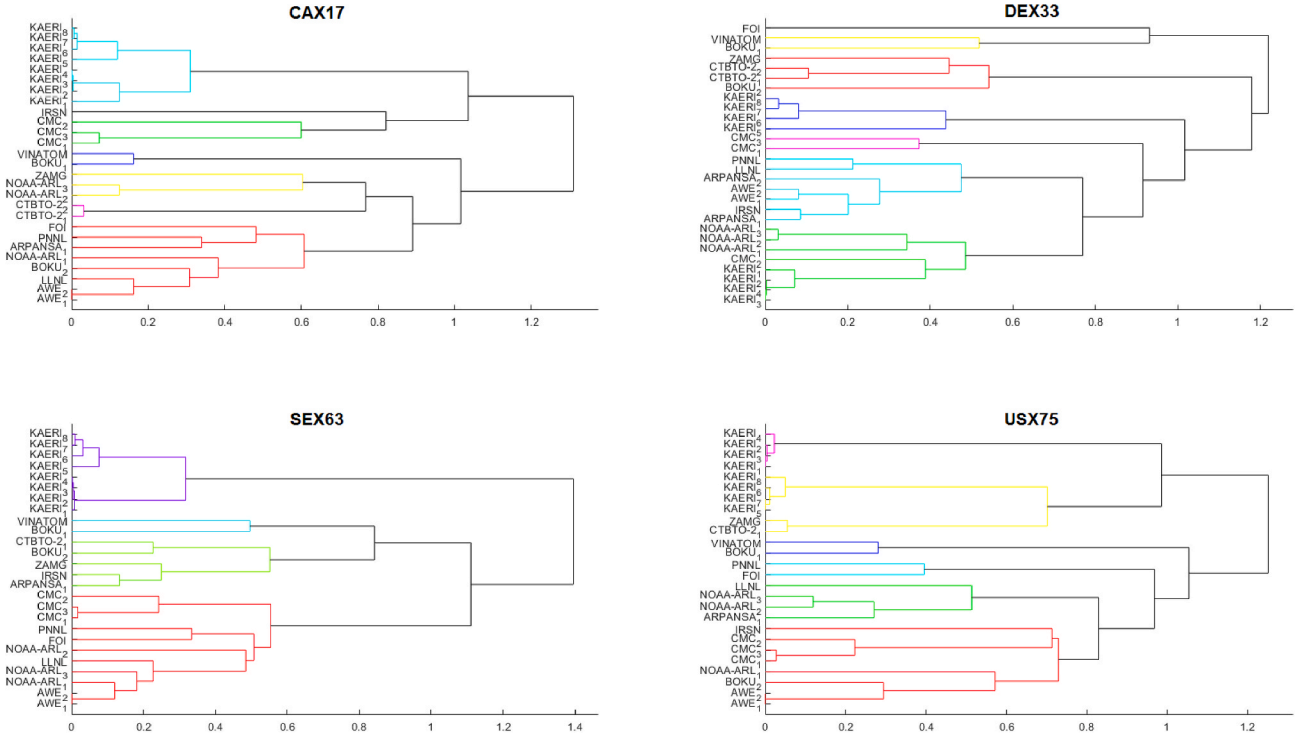


Fig. 9. Dendrograms calculated for all four sampling locations. See text for details.

indicates that the model results do not contain a sufficient level of diversity to produce large improvements even when reducing the ensemble size to enhance diversity. The more relevant improvement for DEX33 does not come as a surprise due to the challenging orography of this site and model outputs sampled partly at considerably different output layers. The fact that the optimum ensemble configuration comprises two runs sampling an elevated model output layer above ground level underpins the importance of accounting for poorly resolved model orography. The apparent discrepancy between the number of members of the RMSE-minimizing ensemble and the effective number of models simply depends on the nature of the two analyses. In one case all possible combinations of possible groupings of models have been used thus pinpointing the grouping that produces the minimum RMSE, in the second case a statistical analysis of how the model errors are distributed and represent the degrees of freedom with which all results are represented is made.

Having identified the presence of scarce levels of differentiation among the models it is important to determine whether some grouping occurs within the ensembles that would allow one to identify macroscopic clusters of model behavior. By transforming the model correlations into the distance d_m, a dendrogram can be produced as presented in Fig. 9. This time again the full original ensemble has been considered.

Let us analyse station CAX17 as an example. From the dendogram the presence of three major groupings appears clearly. The KAERI family cluster constitutes the first one and a central group is made up of four small sub-clusters: one made up of NOAA-ARL runs that also group with ZAMG, one made up of BOKU and VINATOM, one made up of the CTBTO-2 family, and another one of the CMC family. Finally, a large cluster in red is made up of the largest variety of participating organizations. Such diverse clusters of models from different organizations are normally very interesting to analyse since behind the names may lay a clear diversity that is beneficial to ensemble treatments. As shown earlier, however, the diversity in organizations does not translate in diversity in the model results and therefore a detail scrutiny of the model workings may be required before asserting a truly different model behavior. Similar considerations can be drawn from the other dendrograms.

Models are all numerical representations of the physical processes related to ATM. If we would know exactly how to represent mathematically those processes there would not be any need for multi-model ensembles. Any numerical representation of the analytical description would produce almost identical results. In the absence of analytical models different modeling approaches to the representation of the same physical phenomena exist and it is therefore a good reason to opt for a statistical combination of model results. Combing fully independent contributions would yield the best result, by far better than any individual model result. Unfortunately model results are not independent and therefore one has to make sure that the contributions to the ensemble are original contributions. As was widely demonstrated models belong to genealogies (Knutti et al., 2013), they can be the evolutions of common ancestors (be it full models or sets of modules) which may produce a similarity in the result that is more connected to the similarity of the tools used than to an agreement in determining the same prediction. The analysis presented aims specifically at avoiding the presence of redundant results in the ensemble and extracting only the information that is contributing to an effective improvement of the ensemble result.

The selection of the models that originally contribute to the ensemble can be performed only in the presence of experimental evidence. One aspect that is very important to consider is that the behavior of models can change from one case to another. There are situations in which a model can produce results that are replicating other models, and other cases in which it may contribute independently to the final result. One should always consult large pools of models results in order to increase the likelihood of extracting meaningful ones. Training an ensemble against experimental evidence is the best way to assess whether some models behave systematically like others and therefore should be weighted less or even skipped in an ensemble. In the absence of a such training the combination of model results can only be managed statistically therefore opting for the mean or the median results. However, this approach is accompanied by the obvious drawback of diluting an ensemble of model results with those model results with few outliers and thus of discarding relevant outliers.

In a recent study the use of Ensemble Prediction System (EPS) based multi-input atmospheric transport ensembles was inspected by Maurer et al. (2021) and even recommended to the CTBTO by the latter authors. Such an approach should probably be favored for the purpose of radioxenon ensemble modeling in case computational power is available. Any redundancy can be precluded a priori and the approach seems also easier and better controllable on an organizational level (collection of runs in the proper format in due time) in case of a potential violation of the CTBT.

4. Conclusions

After performing multi-model exercises in 2015 and 2016 an even more comprehensive Xe-133 atmospheric transport modeling challenge was organized in 2019. Previous challenges suffered from limitations in terms of emission inventories, simulation period and number of above MDC samples involved, all which could be substantially improved for the current challenge.

Already from Fig. 1 it becomes clear at first glance that knowing exact IRE (and CNL) emissions is clearly not enough for accurately predicting samples at DEX33 and SEX63. Although the situation looks different for CAX17 and USX75 the impression is elusive because the analysis just covers known emitters and known emission quantities. From the rank histogram for USX75 (Fig. 6) missing emissions and/or emitters have to be anticipated too. The benefit of knowing the actual high-quality IRE and CNL emission profiles is in general evidently easily offset by poorly characterized or even unknown other emitters as well as atmospheric transport errors and long IMS sampling times (12–24 h). The detailed findings of this paper can be summarized as follows:

- To demonstrate an average moderate added value of stack emissions from IRE and/or CNL compared to literature emissions for ATM and the four IMS stations investigated, the selected samples must be partly or predominantly influenced by these emitters. However, there can be considerable benefit from these stack data for individual samples.
- It is interesting to note that the mere selection of samples partly (at least to 50%) or predominantly (at least to 80%) influenced by IRE and/or CNL pushes the scores up most. The relative increase in scores on average adds up to ca. 20% when switching from all above MDC samples to those with 50% or 80% IRE and/or CNL influence using literature emissions compared to ca. 10% when additionally switching from literature to stack emissions for 50% or 80% influence samples. This demonstrates that 1) for the four IMS stations considered knowing a large emitter and its location as well as 2) a proper average emission rate per day is more important than knowing the exact emission profile – at least as long as remaining relevant emitters are not properly characterized. Implicitly suppressing samples with overprediction >50% or 20% in the sample selection process (via an absolute difference metric) can further enhance the scores which demonstrates the effects of the transport error on scores.
- Simulating the radioxenon background at CAX17 without selecting samples according to CNL influence nevertheless seems to be promising since CAX17 is a remote station with (at the time of 2014) dominating CNL influence. The latter feature is also reflected by the largest fractions of samples retained by the different sample selection processes applied and matches findings for CAX16 (Yellowknife, Canada) from Saey et al. (2007) in the context of DPRK's nuclear test

in 2006. CAX17 scores are highest for all metrics investigated and no sample selection applied. There is, however, no significant difference whether stack emissions or literature emissions are employed.
- It seems thus to be very important to gain more knowledge about non IRE- and CNL-related emissions. The related sample contributions may be small individually (but can also be large, see MIPF NIIAR or Karpov institute for SEX63), but in any case their sum – depending on the predominant synoptic situation (e.g., November 2014 for DEX33) and the magnitude of the major emitter – can be a decisive factor in accurately predicting the radioxenon background at IMS stations.

The Third ATM-Challenge presented here is in addition a practical application of what was described by Galmarini et al. (2004). Therein the possible combinations of meteorological input, emission conditions and dispersions models with the scope of creating an ensemble of dispersion predictions was presented. The Third ATM-Challenge represents the ultimate configuration hypothesized by Galmarini et al. (2004). It includes multiple meteorological input fields driving a community of dispersion models with the scope of creating the largest possible diversity in dispersion fields starting from all plausible fields and models. However, the current analysis reflects that:

- The existent, full ensemble is highly redundant. An ensemble based on a few arbitrary submissions after a quick correlation analysis to discard members with very high correlation is good enough to model the Xe-133 background. The effective ensemble size is below five. This finding is in agreement with findings from the First and Second ATM-Challenge. The characteristic is very likely related to the fact that radioxenons - treated as tracers with just some decay - are species which are modeled rather uniformly by all models. If other atmospheric processes, e.g., wet scavenging would need to be included it could lead to a larger variability between models. Further, meteorological input consists of analyses (in most cases concatenated with short-term forecasts) thus reducing forecast uncertainty. Finally, there are dominating transport models (FLEXPART & HYSPLIT) in combination with dominant meteorological drivers (ECMWF-IFS, NCEP-GFS).
- The study shows that an optimized ensemble at each station has slightly higher skill compared to the reduced ensemble after correlation analysis. But the improvement (max. of 20% in RMSE for DEX33 but only 3% for SEX63) in skill is likely being too small for being exploited, especially for an independent period.
- Assembling a multitude of dispersion calculations is a necessary but insufficient condition toward obtaining an optimal ensemble. Potential pitfalls loom behind such a construction and the real diversity of the ensemble has to be proven. Unfortunately model results are often not independent. An EPS-based ensemble may be a better alternative compared to the ensemble gathered in the frame of the Third ATM-Challenge.

Disclaimer

The views expressed herein are those of the authors and not necessarily reflect the views of the CTBTO Preparatory Commission or authors' affiliated institutions.

CRediT authorship contribution statement

C. Maurer: Conceptualization of this study, Methodology, Software, Data curation, Writing - Original draft preparation. **S. Galmarini:** Conceptualization of this study, Methodology, Software, Data curation, Writing - Original draft preparation. **E. Solazzo:** Methodology, Software, Data curation, Writing - Original draft preparation. **J. Kuśmierczyk-Michulec:** Conceptualization of this study, Methodology, Data curation, Writing - Original draft preparation. **J. Baré:** Data curation, Writing - Original draft preparation. **M. Kalinowski:** Conceptualization of this study, Data curation, Writing - Original draft preparation. **M. Schoeppner:** Data curation. P. Bourgouin: Data curation. **A. Crawford:** Conceptualization of this study, Data curation. **A. Stein:** Data curation. **T. Chai:** Data curation. **F. Ngan:** Data curation. **A. Malo:** Conceptualization of this study, Data curation. **P. Seibert:** Methodology, Data curation. **A. Axelsson:** Data curation. **A. Ringbom:** Data curation. **R. Britton:** Data curation. **A. Davies:** Data curation. **M. Goodwin:** Data curation. **P. W. Eslinger:** Data curation. **T. W. Bowyer:** Methodology. **L. G. Glascoe:** Data curation. **D. D. Lucas:** Data curation. **S. Cicchi:** Data curation. **P. Vogt:** Data curation. **Y. Kijima:** Data curation. **A. Furuno:** Data curation. **P. K. Long:** Data curation. **B. Orr:** Data curation. **A. Wain:** Data curation. **K. Park:** Data curation. **K.-S. Suh:** Data curation. **A. Quérel:** Data curation. **O. Saunier:** Data curation. **D. Quélo:** Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We want to thank Ian Hoffman (Health Canada) for assisting in getting access to the stack release data from CNL, Canada, and Benoît Deconninck for providing stack release data from IRE, Belgium. We acknowledge all CTBTO colleagues who contributed to the IMS monitoring system or the IDC analysis. Further, we express our gratitude to CTBTO for making available the virtual Data Exploitation Center (vDEC, https://www.ctbto.org/specials/vdec/) for distributing stack emission and IMS data to the participants of the Third ATM-Challenge.

Appendix

A.1. Individual participants' results

A minor aspect of the Third ATM-Challenge was to do another model inter-comparison. Although of general interest for the participating organizations the added value of such an inter-comparison from a practical point of view is limited for CTBT verification, civil radioxenon background modeling respectively, since modifying (possibly multi-purpose) modeling chains based on single case studies with different settings is hardly justified. The purpose of involving runs from many different participants in international exercises has since the First ATM-Challenge to a much larger degree been focused on building optimized ensembles or definitely answering questions regarding the requirements for emission data (e.g., temporal resolution needed or emission accuracy).

As demonstrated in all three ATM-Challenges individual model performance, model ranking respectively, varies case study by case study (despite employing the very same rank metric for the Third and Second ATM-Challenge), among target IMS stations for the very same case study (see Tables A.1 and A.2, Tables A.3 and A.4, Tables A.5 and A.6 and Tables A.7 and A.8) and even between different time frames for a specific IMS station for the very same case study as shown also below. Further, any overall metrics of the three ATM-Challenges should only be compared restrainedly. To some degree

participants, but above all model set-ups (e.g., largely uniform versus non-uniform output grids) and likely model versions are different. Seasonal coverages, completeness of emission inventories, partly the considered hemispheres and IMS stations as well as the number of samples above the MDC (very low for the Second ATM-Challenge) are different too.

If at all one can compare the results from the First ATM-Challenge (DEX33 was the only IMS station involved) with the results for DEX33 of the present study. As stated in the caption of Table 3 in Eslinger et al. (2016) the average metrics in the last line were calculated based on ensemble predictions made up of all submissions. Thus, the same approach was taken in this study for a comparison. Whereas scores in Eslinger et al. (2016) are valid for the period November, 10th, to December 10th, 2013 and rely on 26 members, the current statistics are based on the time frame July, 8th, to September 30th, 2014 and rely on 31 members. The values for correlation are similar (0.69 in Eslinger et al. (2016) versus 0.61 in the present study), those for the fractional biases are similar according to the absolute value, but different according to the sign (0.27 for Eslinger et al. (2016) versus −0.12 in the present study). The F5 is lower for Eslinger et al. (2016) (61%) than for the present study (85%) and the KS is considerably higher for Eslinger et al. (2016) (42%) than for the present study (8%). Computing the rank following the definition in Eslinger et al. (2016) (see **section 3** therein) yields 3.09 versus 2.53 in the former study. However, modeling the atmospheric transport to DEX33, located on a mountain next to the Rhine valley, will be more challenging in winter than in summer and the runs in Eslinger et al. (2016) with one exemption did not include any other emitters than just IRE.

Table A.1, Table A.3, Table A.5 and Table A.7 show the results based on stack emission data per individual run and averaged over all runs of a specific participant for IMS stations CAX17, DEX33, SEX63 and USX75. For strict comparability reasons the tables only cover the months July to September which were modeled without exception by all participants. Apart from $Rank_{KS}$, the Seibert Score SS, the Kolmogorov-Smirnov parameter KS and the (normalized) bias corrected root mean square error BC_{RMSE} (BC_{NMSE}) as defined in subsection 2.6.1, as well as the ratio of observed and modeled standard deviations replacing the cross correlation coefficient, scores (*R*, FB, F5, ACC, NAAD and Rank) are indentical to those employed and defined in Maurer et al. (2018). Runs are sorted according to metric Rank as for the Second ATM-Challenge. Highest overall performance (see last but third line in the tables) is reached for CAX17 and Rank, $Rank_{KS}$, respectively with median values of 2.62, 3.49, respectively, and for DEX33 and SS with a value of 0.58. Rank values range from 2.31 (USX75) to 2.62 (CAX17, 13% difference between the best and the worst Rank value), $Rank_{KS}$ values from 3.02 (USX75) to 3.49 (CAX17, 16% difference between the best and the worst $Rank_{KS}$ value) and SS values from 0.40 (SEX63) to 0.58 (DEX33, 45% difference between the best and the worst SS value). Comparing the full (June–November, tables not shown) with the reduced (July-Septmeber) evaluation period one can find some differences per IMS station although averaged over all four stations results are similar (see Table 6). For CAX17 median [minimum, maximum] add up to 2.63 [1.22, 3.16] for Rank, to 3.48 [1.75, 4.07] for $Rank_{KS}$ and 0.63 [0.12, 0.86] for SS with two of the three best performing runs being identical for both evaluation periods. For DEX33 analogous figures are 2.42 [1.70, 3.04], 3.15 [2.39, 3.92] and 0.51 [0.17, 0.81] again with two of the three best performing runs being identical for both evaluation periods. For SEX63 we find 2.21 [0.75, 2.58], 2.96 [1.14, 3.49] and 0.32 [0.03, 0.76] and one of the three best performing runs being identical for both evaluation periods. And, finally, USX75 yields 2.34 [1.06, 2.86], 3.09 [1.58, 3.69] and 0.39 [0.09, 0.75] again with two of the three best performing runs being identical for both evaluation periods. Both evaluation periods reflect better predictability for CAX17 and DEX33 compared to SEX63 and USX75.

KAERI runs 2 to 4 and 6 to 8 offer the unique opportunity to compare the effect of different output resolutions (0.5° × 0.5°, 0.35° × 0.23° and 0.1° × 0.1°) under otherwise unchanged settings. Under KAERI standard settings (runs 2 to 4) for turbulent diffusion constants, i.e., $k_h = 2.5 \times 10^4$ m^2/s and $k_v = 1.0$ m^2/s, there is virtually no difference between the three output resolutions. If at all, noteworthy differences are confined to the experimental setting (runs 6 to 8) for turbulent diffusion constants, i.e., $k_h = 50$ m^2/s and $k_v = 0.1$ m^2/s. Largest discrepancies can be found for DEX33 with the 0.5° × 0.5° outperforming the 0.1° × 0.1° run. However, a coarse resolution run outperforming a high resolution run does not come as a surprise at IMS station DEX33 and was discussed in **section 4. Conclusions** of Maurer et al. (2018) in detail.

Finally, Table A.2, Table A.4, Table A.6 and Table A.8 show analogous results compared to Table A.1, Table A.3, Table A.5 and Table A.7 with the only exception that emitters apart from IRE and CNL were omitted thus avoiding any bias in the analysis possibly introduced by adding ZAMG's results for remaining emitters to the IRE and CNL only simulations of ARPANSA, IRSN, JAEA-1 and JAEA-2. In fact Table A.2, Table A.4, Table A.6 and Table A.8 should be used when the only aim is to compare run performances relatively to each other. In most cases ranking is improving for those four participants if ZAMG's results are not added. Ranking differences for ARPANSA (averaged in case of DEX33 over two individual runs), IRSN, JAEA-1 and JAEA-2 reflect the different impact of IRE and/or CNL on the four IMS stations and the difference between emitters IRE and CNL themselves with regard to release amounts. The maximum change in ranking adds up to two positions for CAX17 (compare Tables A.1 and A.2), followed by five positions for USX75 (compare Tables A.7 and A.8), six for DEX33 (compare Tables A.3 and A.4) and nine for SEX63 (compare Tables A.7 and A.8). Findings of **section 3.1.3** get corroborated: Not a lot run performance can be gained when adding remaining emitters for IMS stations CAX17 and USX75 with dominant CNL influence. The radioxenon background at CAX17 can especially be well simulated without taking into consideration any other emitters.

Ultimately, it is important to note that ranking runs with respect to one metric does not necessarily do justice to the submissions. However, the highest Ranks (differences between Rank and $Rank_{KS}$ are minor) tend to come with high SSs.

Table A.1
Statistics for CAX17 and time frame July, 8th, to September 30th, using stack emissions for IRE & CNL and including all remaining emitters. Results are shown per organization for individual run IDs (organization IDs plus subcripts) and over all run-IDs per participating organization ordered by rank. Bottom lines show metrics over all organizations after averaging over all run-IDs per organization. Number of valid sample pairs involved including observed samples below MDC: 82.

Organizations' runs	R	FB [mBq/m^3]	F5 [%]	BC_{RMSE} [mBq/m^3]	BC_{NMSE}	KS [%]	ACC [%]	NAAD [%]	Ratio stdvs	**Rank**	**$Rank_{KS}$**	**SS**
AWE_2	0.62	−1.67	13	4.99	39.21	47	51	93	10.11	**1.20**	**1.73**	**0.03**
AWE	0.62	−1.67	13	4.99	39.12	47	51	93	10.10	**1.20**	**1.73**	**0.03**
AWE_1	0.62	−1.67	13	4.99	39.04	47	51	93	10.09	**1.20**	**1.73**	**0.03**
LLNL	0.46	−0.97	68	4.74	9.10	29	82	72	2.65	**2.22**	**2.93**	**0.21**
JAEA-2	0.19	0.15	56	6.17	4.64	10	71	115	1.22	**2.23**	**3.13**	**0.70**
FOI	0.46	−0.38	67	4.84	4.89	20	82	69	1.50	**2.50**	**3.30**	**0.51**
CTBTO-1	0.54	−0.59	75	4.46	5.16	21	79	63	2.14	**2.55**	**3.34**	**0.34**
$KAERI_4$	0.41	0.40	73	6.22	3.65	15	85	125	0.87	**2.55**	**3.40**	**0.54**
$KAERI_2$	0.42	0.42	75	6.17	3.51	16	85	126	0.87	**2.57**	**3.41**	**0.53**

(*continued on next page*)

Table A.1 (*continued*)

Organizations' runs	R	FB [mBq/m^3]	F5 [%]	$BC_{RMSE\ [mBq/m^3]}$	BC_{NMSE}	KS [%]	ACC [%]	NAAD [%]	Ratio stdvs	**Rank**	**$Rank_{KS}$**	**SS**
$BOKU_1$	0.28	−0.23	78	6.33	7.11	11	83	83	1.01	**2.57**	**3.46**	**0.65**
$KAERI_3$	0.41	0.41	76	6.18	3.58	15	85	125	0.87	**2.58**	**3.43**	**0.54**
VINATOM	0.30	−0.17	75	5.95	5.94	11	83	86	1.12	**2.59**	**3.48**	**0.71**
JAEA-1	0.54	0.10	62	4.77	2.90	10	76	85	1.17	**2.62**	**3.52**	**0.83**
BOKU	0.48	−0.46	75	5.18	5.94	15	82	73	1.57	**2.62**	**3.48**	**0.49**
PNNL	0.73	−0.70	65	3.93	4.54	22	79	58	2.26	**2.63**	**3.41**	**0.32**
$KAERI_6$	0.52	0.45	73	6.02	3.24	16	87	117	0.79	**2.64**	**3.48**	**0.52**
KAERI	0.49	0.38	74	5.82	3.25	15	86	114	0.88	**2.65**	**3.50**	**0.57**
$KAERI_8$	0.52	0.36	71	5.73	3.22	15	85	109	0.84	**2.66**	**3.51**	**0.58**
$BOKU_2$	0.69	−0.70	73	4.03	4.76	18	82	62	2.13	**2.67**	**3.49**	**0.33**
$KAERI_5$	0.55	0.39	71	5.54	2.94	14	87	105	0.85	**2.68**	**3.54**	**0.58**
$KAERI_7$	0.52	0.40	75	5.82	3.20	16	87	112	0.82	**2.69**	**3.53**	**0.56**
ARPANSA	0.67	−0.58	73	4.00	4.09	19	80	64	1.92	**2.70**	**3.51**	**0.40**
ZAMG	0.64	−0.42	80	4.09	3.63	20	79	60	1.85	**2.79**	**3.59**	**0.47**
$NOAA\text{-}ARL_1$	0.71	0.05	61	3.88	2.03	33	72	77	1.93	**2.81**	**3.48**	**0.78**
$KAERI_1$	0.53	0.23	79	4.90	2.69	13	87	95	1.11	**2.82**	**3.69**	**0.70**
CMC_3	0.57	0.15	83	5.26	3.37	10	88	76	0.88	**2.96**	**3.86**	**0.80**
CMC_2	0.56	−0.15	84	4.55	3.42	9	88	62	1.28	**2.96**	**3.87**	**0.77**
$NOAA\text{-}ARL_3$	0.69	−0.28	77	3.86	2.81	13	85	58	1.56	**2.96**	**3.83**	**0.63**
NOAA-ARL	0.71	−0.12	72	3.77	2.29	19	80	64	1.57	**2.96**	**3.77**	**0.75**
CMC	0.59	0.03	84	4.80	3.18	9	89	69	1.05	**3.01**	**3.92**	**0.81**
IRSN	0.64	−0.10	81	4.10	2.63	8	88	64	1.36	**3.05**	**3.97**	**0.83**
CMC_1	0.63	0.09	86	4.59	2.74	8	90	69	1.00	**3.11**	**4.03**	**0.87**
$NOAA\text{-}ARL_2$	0.74	−0.11	77	3.58	2.03	11	84	57	1.23	**3.11**	**4.00**	**0.86**
$CTBTO\text{-}2_1$	0.91	−0.16	77	2.26	0.85	11	84	46	1.17	**3.35**	**4.24**	**0.86**
CTBTO-2	0.90	−0.14	77	2.28	0.85	11	85	46	1.16	**3.36**	**4.25**	**0.88**
$CTBTO\text{-}2_2$	0.90	−0.13	77	2.30	0.85	11	85	47	1.15	**3.37**	**4.26**	**0.90**
Average over all organizations	**0.56**	**−0.35**	**69**	**4.62**	**6.38**	**18**	**80**	**75**	**2.09**	**2.60**	**3.43**	**0.55**
Median over all organizations	**0.57**	**−0.28**	**74**	**4.75**	**4.32**	**17**	**81**	**69**	**1.53**	**2.62**	**3.49**	**0.54**
Maximum over all organizations	**0.90**	**0.38**	**84**	**6.17**	**39.12**	**47**	**89**	**115**	**10.10**	**3.36**	**4.25**	**0.88**
Minimum over all organizations	**0.19**	**−1.67**	**13**	**2.28**	**0.85**	**8**	**51**	**46**	**0.88**	**1.20**	**1.73**	**0.03**

Table A.2
Same as Table A.1, but with IRE and CNL stack emissions only.

Organizations' runs	R	FB [mBq/m^3]	F5 [%]	$BC_{RMSE\ [mBq/m^3]}$	BC_{NMSE}	KS [%]	ACC [%]	NAAD [%]	Ratio stdvs	**Rank**	**$Rank_{KS}$**	**SS**
AWE_2	0.62	−1.67	13	4.99	39.73	48	51	93	10.18	**1.20**	**1.72**	**0.03**
AWE	0.62	−1.67	13	4.99	39.73	48	51	93	10.18	**1.20**	**1.72**	**0.03**
AWE_1	0.62	−1.67	13	4.99	39.73	48	51	93	10.18	**1.20**	**1.72**	**0.03**
LLNL	0.45	−0.97	66	4.74	9.22	30	82	72	2.67	**2.20**	**2.90**	**0.21**
JAEA-2	0.19	0.14	56	6.17	4.66	10	71	115	1.22	**2.23**	**3.13**	**0.70**
FOI	0.46	−0.39	67	4.84	4.94	20	82	69	1.51	**2.50**	**3.30**	**0.51**
CTBTO-1	0.54	−0.59	75	4.47	5.22	21	79	64	2.15	**2.54**	**3.33**	**0.34**
$KAERI_4$	0.41	0.39	73	6.21	3.67	15	85	124	0.88	**2.55**	**3.40**	**0.54**
$BOKU_1$	0.28	−0.24	78	6.33	7.19	13	83	84	1.01	**2.57**	**3.44**	**0.64**
VINATOM	0.30	−0.18	75	5.95	6.02	13	83	86	1.13	**2.58**	**3.45**	**0.70**
$KAERI_2$	0.42	0.41	76	6.15	3.53	16	87	125	0.87	**2.59**	**3.43**	**0.53**
PNNL	0.73	−0.71	65	3.94	4.61	22	78	58	2.28	**2.61**	**3.39**	**0.32**
$KAERI_3$	0.41	0.40	77	6.16	3.59	15	87	124	0.88	**2.61**	**3.46**	**0.54**
BOKU	0.48	−0.47	75	5.18	5.98	16	82	73	1.57	**2.62**	**3.46**	**0.49**
JAEA-1	0.54	0.10	62	4.77	2.92	10	76	85	1.17	**2.62**	**3.52**	**0.83**
$KAERI_6$	0.52	0.44	73	6.00	3.26	16	87	116	0.79	**2.64**	**3.48**	**0.53**
KAERI	0.49	0.37	74	5.81	3.27	15	86	113	0.88	**2.66**	**3.51**	**0.57**
$KAERI_8$	0.52	0.36	71	5.72	3.24	14	85	108	0.85	**2.66**	**3.52**	**0.59**
$BOKU_2$	0.69	−0.70	73	4.03	4.77	18	82	62	2.13	**2.67**	**3.49**	**0.33**
ARPANSA	0.67	−0.58	71	4.00	4.14	19	80	65	1.93	**2.68**	**3.49**	**0.39**
$KAERI_5$	0.55	0.38	71	5.53	2.96	14	87	104	0.85	**2.69**	**3.55**	**0.58**
$KAERI_7$	0.52	0.39	75	5.80	3.22	16	87	111	0.83	**2.69**	**3.53**	**0.57**
ZAMG	0.64	−0.43	80	4.09	3.66	20	79	60	1.86	**2.79**	**3.59**	**0.46**
$NOAA\text{-}ARL_1$	0.71	0.04	61	3.88	2.06	33	72	77	1.93	**2.81**	**3.48**	**0.78**
$KAERI_1$	0.53	0.22	79	4.89	2.71	13	87	94	1.12	**2.82**	**3.69**	**0.71**
$NOAA\text{-}ARL_3$	0.69	−0.30	76	3.86	2.84	14	84	58	1.57	**2.92**	**3.78**	**0.61**
NOAA-ARL	0.71	−0.13	71	3.78	2.32	19	79	64	1.58	**2.93**	**3.74**	**0.75**
CMC_3	0.57	0.14	83	5.25	3.39	11	88	76	0.89	**2.96**	**3.85**	**0.80**
CMC_2	0.56	−0.16	86	4.55	3.46	9	89	62	1.29	**2.98**	**3.89**	**0.76**
CMC	0.58	0.02	84	4.80	3.20	9	89	69	1.06	**3.01**	**3.91**	**0.81**
IRSN	0.64	−0.10	79	4.10	2.65	8	87	64	1.36	**3.01**	**3.93**	**0.83**
$NOAA\text{-}ARL_2$	0.74	−0.12	76	3.59	2.06	11	82	57	1.23	**3.06**	**3.95**	**0.85**

(*continued on next page*)

Table A.2 (*continued*)

Organizations' runs	R	FB [mBq/ m^3]	F5 [%]	$BC_{RMSE\ [mBq/m^3]}$	BC_{NMSE}	KS [%]	ACC [%]	NAAD [%]	Ratio stdvs	**Rank**	**$Rank_{KS}$**	**SS**
CMC_1	0.62	0.08	84	4.59	2.76	8	89	69	1.00	**3.08**	**4.00**	**0.88**
CTBTO-2_2	0.90	−0.16	75	2.29	0.87	12	84	47	1.16	**3.33**	**4.21**	**0.86**
CTBTO-2	0.91	−0.18	76	2.27	0.88	13	84	46	1.17	**3.33**	**4.21**	**0.84**
CTBTO-2_1	0.91	−0.20	77	2.25	0.88	13	84	46	1.18	**3.33**	**4.20**	**0.82**
Average over all organizations	**0.56**	**−0.36**	**68**	**4.62**	**6.46**	**18**	**79**	**75**	**2.11**	**2.59**	**3.41**	**0.55**
Median over all organizations	**0.56**	**−0.29**	**73**	**4.75**	**4.38**	**17**	**81**	**69**	**1.54**	**2.62**	**3.48**	**0.54**
Maximum over all organizations	**0.91**	**0.37**	**84**	**6.17**	**39.73**	**48**	**89**	**115**	**10.18**	**3.33**	**4.21**	**0.84**
Minimum over all organizations	**0.19**	**−1.67**	**13**	**2.27**	**0.88**	**8**	**51**	**46**	**0.88**	**1.20**	**1.72**	**0.03**

Table A.3
Statistics for DEX33 and time frame July, 8th, to September 30th, using stack emissions for IRE & CNL and including all remaining emitters. Results are shown per organization for individual run IDs (organization IDs plus subcripts) and over all run-IDs per participating organization ordered by rank. Bottom lines show metrics over all organizations after averaging over all run-IDs per organization. Number of valid sample pairs involved including observed samples below MDC: 79.

Organizations' runs	R	FB [mBq/ m^3]	F5 [%]	$BC_{RMSE\ [mBq/m^3]}$	BC_{NMSE}	KS [%]	ACC [%]	NAAD [%]	Ratio stdvs	**Rank**	**$Rank_{KS}$**	**SS**
JAEA-1	0.45	−0.72	65	1.52	2.96	38	61	70	2.16	**2.11**	**2.73**	**0.29**
FOI	0.26	0.21	56	2.96	4.31	26	63	133	0.59	**2.16**	**2.90**	**0.59**
$ARPANSA_1$	0.63	−0.92	74	1.44	3.38	39	56	70	3.32	**2.23**	**2.84**	**0.18**
AWE_2	0.16	−0.64	78	1.69	3.33	25	85	74	4.71	**2.33**	**3.08**	**0.15**
ARPANSA	0.66	−0.87	76	1.40	3.02	40	57	67	3.07	**2.33**	**2.93**	**0.21**
AWE	0.09	−0.53	76	1.73	3.09	20	84	77	4.05	**2.34**	**3.15**	**0.20**
AWE_1	0.02	−0.41	73	1.77	2.86	14	82	80	3.39	**2.35**	**3.21**	**0.26**
CTBTO-2_1	0.29	0.76	85	5.79	9.13	19	81	164	0.28	**2.37**	**3.18**	**0.16**
VINATOM	0.45	0.58	76	4.44	6.54	20	75	143	0.35	**2.42**	**3.22**	**0.25**
$ARPANSA_2$	0.69	−0.82	78	1.36	2.66	41	58	64	2.82	**2.43**	**3.02**	**0.23**
$BOKU_2$	0.54	−0.23	70	1.86	2.64	44	57	76	0.80	**2.45**	**3.01**	**0.69**
ZAMG	0.45	−0.38	78	1.64	2.39	32	71	70	1.24	**2.50**	**3.18**	**0.55**
CTBTO-2	0.32	0.36	83	4.05	6.22	24	77	123	0.52	**2.50**	**3.26**	**0.48**
BOKU	0.48	−0.18	77	1.92	2.65	35	63	77	0.85	**2.54**	**3.20**	**0.74**
PNNL	0.25	0.23	76	1.75	1.48	41	85	99	1.68	**2.56**	**3.15**	**0.57**
CTBTO-1	0.35	−0.14	79	1.70	1.99	22	73	83	1.47	**2.57**	**3.35**	**0.71**
LLNL	0.22	−0.06	77	1.75	1.96	19	78	88	1.82	**2.58**	**3.39**	**0.70**
CTBTO-2_2	0.34	−0.03	81	2.30	3.31	29	72	82	0.76	**2.63**	**3.34**	**0.81**
$BOKU_1$	0.41	−0.13	85	1.97	2.66	25	68	77	0.89	**2.64**	**3.39**	**0.78**
$KAERI_8$	0.58	−0.29	83	1.48	1.76	34	67	64	1.16	**2.69**	**3.35**	**0.66**
IRSN	0.62	−0.56	86	1.40	2.10	22	77	59	2.56	**2.73**	**3.51**	**0.31**
$NOAA\text{-}ARL_2$	0.51	−0.20	78	1.47	1.59	13	81	67	1.85	**2.75**	**3.62**	**0.62**
$NOAA\text{-}ARL_3$	0.51	−0.23	80	1.46	1.63	14	82	66	1.90	**2.77**	**3.63**	**0.58**
CMC_3	0.51	0.26	80	1.50	1.05	40	85	82	1.47	**2.77**	**3.37**	**0.63**
$KAERI_7$	0.60	−0.27	87	1.45	1.66	32	71	59	1.12	**2.80**	**3.48**	**0.68**
NOAA-ARL	0.52	−0.12	81	1.46	1.48	23	83	69	2.05	**2.82**	**3.59**	**0.63**
JAEA-2	0.60	−0.22	86	1.40	1.47	23	71	62	1.27	**2.82**	**3.59**	**0.71**
CMC	0.52	0.03	83	1.48	1.31	29	86	73	1.53	**2.85**	**3.56**	**0.66**
CMC_1	0.51	0.12	79	1.51	1.22	29	86	75	1.41	**2.85**	**3.56**	**0.77**
CMC_2	0.54	−0.28	90	1.44	1.67	18	87	63	1.72	**2.92**	**3.74**	**0.57**
$NOAA\text{-}ARL_1$	0.53	0.06	84	1.46	1.21	41	85	74	2.40	**2.94**	**3.53**	**0.68**
$KAERI_6$	0.67	−0.21	88	1.31	1.28	27	78	53	1.17	**3.01**	**3.74**	**0.75**
KAERI	0.67	−0.24	91	1.31	1.32	21	80	53	1.40	**3.04**	**3.84**	**0.68**
$KAERI_5$	0.65	−0.12	94	1.40	1.35	18	82	57	1.04	**3.12**	**3.94**	**0.85**
$KAERI_3$	0.72	−0.29	93	1.21	1.18	16	84	47	1.73	**3.14**	**3.98**	**0.60**
$KAERI_4$	0.72	−0.29	93	1.21	1.18	15	84	48	1.73	**3.14**	**3.99**	**0.59**
$KAERI_2$	0.72	−0.27	93	1.21	1.15	12	86	47	1.70	**3.17**	**4.05**	**0.62**
$KAERI_1$	0.72	−0.18	94	1.18	1.01	10	90	47	1.53	**3.27**	**4.17**	**0.74**
Average over all organizations	**0.43**	**−0.16**	**78**	**1.99**	**2.77**	**27**	**74**	**84**	**1.66**	**2.55**	**3.28**	**0.52**
Median over all organizations	**0.45**	**−0.16**	**78**	**1.67**	**2.24**	**24**	**76**	**75**	**1.50**	**2.55**	**3.24**	**0.58**
Maximum over all organizations	**0.67**	**0.58**	**91**	**4.44**	**6.54**	**41**	**86**	**143**	**4.05**	**3.04**	**3.84**	**0.74**
Minimum over all organizations	**0.09**	**−0.87**	**56**	**1.31**	**1.31**	**19**	**57**	**53**	**0.35**	**2.11**	**2.73**	**0.20**

Table A.4
Same as Table A.3, but with IRE and CNL stack emissions only.

Organizations' runs	R	FB [mBq/m^3]	F5 [%]	$BC_{RMSE\ [mBq/m^3]}$	BC_{NMSE}	KS [%]	ACC [%]	NAAD [%]	Ratio stdvs	**Rank**	**$Rank_{KS}$**	**SS**
AWE_1	0.34	−1.66	9	1.63	17.02	77	24	93	5.41	**0.61**	**0.84**	**0.06**
AWE	0.34	−1.66	9	1.63	17.01	77	24	93	5.41	**0.61**	**0.84**	**0.06**
AWE_2	0.34	−1.66	9	1.63	17.00	77	24	93	5.41	**0.61**	**0.84**	**0.06**
LLNL	0.53	−1.29	34	1.48	6.12	69	32	80	2.82	**1.29**	**1.60**	**0.18**
$ARPANSA_1$	0.62	−1.32	30	1.47	6.38	59	43	80	3.86	**1.45**	**1.86**	**0.12**
JAEA-1	0.41	−1.07	39	1.56	4.82	57	44	80	2.25	**1.46**	**1.89**	**0.23**
$NOAA\text{-}ARL_3$	0.46	−1.05	36	1.52	4.44	59	43	81	2.01	**1.48**	**1.89**	**0.27**
$NOAA\text{-}ARL_2$	0.45	−0.96	36	1.52	4.01	58	46	81	1.96	**1.54**	**1.96**	**0.29**
ARPANSA	0.65	−1.28	33	1.44	5.70	57	44	78	3.56	**1.56**	**1.99**	**0.14**
CTBTO-1	0.57	−1.05	40	1.40	3.82	59	39	75	2.03	**1.60**	**2.01**	**0.29**
PNNL	0.47	−0.94	44	1.51	3.78	53	44	77	2.08	**1.64**	**2.11**	**0.27**
$ARPANSA_2$	0.69	−1.24	36	1.40	5.01	55	46	77	3.26	**1.67**	**2.12**	**0.16**
ZAMG	0.42	−0.69	45	1.67	3.44	54	48	80	1.29	**1.76**	**2.22**	**0.42**
NOAA-ARL	0.46	−0.84	50	1.52	3.59	45	52	78	2.19	**1.81**	**2.36**	**0.29**
$KAERI_8$	0.53	−0.72	49	1.51	2.91	56	41	74	1.27	**1.81**	**2.25**	**0.44**
FOI	0.36	−0.42	42	2.58	6.14	54	48	104	0.65	**1.82**	**2.28**	**0.46**
$BOKU_1$	0.39	−0.49	49	1.96	3.84	47	48	83	0.92	**1.88**	**2.41**	**0.49**
$KAERI_7$	0.56	−0.68	53	1.49	2.70	56	42	72	1.21	**1.92**	**2.36**	**0.47**
$CTBTO2_2$	0.25	−0.40	54	2.35	5.00	47	54	91	0.81	**1.95**	**2.48**	**0.49**
BOKU	0.46	−0.44	49	1.90	3.44	51	47	83	0.87	**1.96**	**2.45**	**0.53**
CTBTO-2	0.26	0.12	63	4.09	7.82	41	59	125	0.55	**2.03**	**2.62**	**0.34**
$BOKU_2$	0.53	−0.39	48	1.84	3.03	55	47	82	0.83	**2.04**	**2.49**	**0.57**
$KAERI_6$	0.64	−0.64	56	1.33	2.06	54	43	65	1.27	**2.08**	**2.54**	**0.49**
$CTBTO2_1$	0.26	0.63	72	5.82	10.64	35	65	159	0.28	**2.11**	**2.76**	**0.19**
CMC_3	0.52	−0.59	62	1.47	2.41	47	54	73	1.58	**2.13**	**2.66**	**0.42**
IRSN	0.58	−0.87	70	1.44	3.19	48	56	68	2.75	**2.16**	**2.68**	**0.22**
CMC_2	0.48	−0.66	70	1.50	2.70	45	57	74	1.83	**2.17**	**2.72**	**0.36**
VINATOM	0.46	0.13	51	4.30	9.77	45	52	118	0.35	**2.17**	**2.72**	**0.57**
CMC	0.50	−0.60	67	1.50	2.50	45	57	73	1.63	**2.18**	**2.73**	**0.41**
$KAERI_5$	0.65	−0.57	59	1.35	1.98	52	46	64	1.12	**2.19**	**2.67**	**0.53**
KAERI	0.65	−0.65	60	1.33	2.13	49	50	64	1.51	**2.20**	**2.71**	**0.44**
CMC_1	0.49	−0.53	69	1.52	2.39	43	58	74	1.49	**2.24**	**2.81**	**0.45**
JAEA-2	0.58	−0.45	66	1.42	1.92	41	58	68	1.31	**2.36**	**2.95**	**0.53**
$KAERI_3$	0.69	−0.68	66	1.26	1.94	43	56	61	1.87	**2.36**	**2.93**	**0.38**
$KAERI_4$	0.70	−0.68	66	1.25	1.94	44	57	60	1.88	**2.37**	**2.93**	**0.38**
$KAERI_2$	0.69	−0.67	68	1.26	1.90	43	57	60	1.85	**2.39**	**2.96**	**0.39**
$NOAA\text{-}ARL_1$	0.47	−0.51	78	1.51	2.33	18	67	71	2.58	**2.42**	**3.24**	**0.31**
$KAERI_1$	0.71	−0.60	68	1.21	1.64	44	57	57	1.64	**2.45**	**3.01**	**0.45**
Average over all organizations	**0.48**	**−0.75**	**48**	**1.92**	**5.32**	**53**	**47**	**84**	**1.95**	**1.79**	**2.26**	**0.34**
Median over all organizations	**0.47**	**−0.76**	**47**	**1.51**	**3.80**	**52**	**48**	**79**	**1.83**	**1.81**	**2.32**	**0.31**
Maximum over all organizations	**0.65**	**0.13**	**70**	**4.30**	**17.01**	**77**	**59**	**125**	**5.41**	**2.36**	**2.95**	**0.57**
Minimum over all organizations	**0.26**	**−1.66**	**9**	**1.33**	**1.92**	**41**	**24**	**64**	**0.35**	**0.61**	**0.84**	**0.06**

Table A.5
Statistics for SEX63 and time frame July, 8th, to September 30th, using stack emissions for IRE & CNL and including all remaining emitters. Results are shown per organization for individual run IDs (organization IDs plus subcripts) and over all run-IDs per participating organization ordered by rank. Bottom lines show metrics over all organizations after averaging over all run-IDs per organization. Number of valid sample pairs involved including observed samples below MDC: 168.

Organizations' runs	R	FB [mBq/m^3]	F5 [%]	$BC_{RMSE\ [mBq/m^3]}$	BC_{NMSE}	KS [%]	ACC [%]	NAAD [%]	Ratio stdvs	**Rank**	**$Rank_{KS}$**	**SS**
AWE_2	0.50	−1.66	15	0.90	15.27	68	32	91	6.07	**0.88**	**1.20**	**0.06**
AWE	0.52	−1.64	16	0.89	14.20	68	32	91	6.00	**0.93**	**1.25**	**0.06**
AWE_1	0.54	−1.62	17	0.89	13.13	67	32	90	5.93	**0.98**	**1.31**	**0.06**
CTBTO-1	0.51	−0.89	63	0.83	3.17	41	52	73	2.07	**1.96**	**2.55**	**0.29**
LLNL	0.55	−0.88	64	0.81	2.98	41	52	72	2.15	**2.02**	**2.61**	**0.28**
FOI	0.64	−0.89	62	0.76	2.66	40	57	67	2.18	**2.16**	**2.76**	**0.29**
PNNL	0.45	−0.52	64	0.97	2.83	30	60	73	1.10	**2.18**	**2.88**	**0.49**
JAEA-1	0.38	−0.40	65	0.97	2.48	25	60	80	1.29	**2.20**	**2.95**	**0.52**
ZAMG	0.54	−0.75	72	0.82	2.60	37	58	66	1.56	**2.21**	**2.84**	**0.39**
ARPANSA	0.62	−0.88	76	0.79	2.82	37	58	69	2.55	**2.28**	**2.91**	**0.24**
VINATOM	0.44	−0.51	75	0.88	2.31	23	68	72	1.68	**2.37**	**3.14**	**0.42**
$BOKU_1$	0.51	−0.55	74	0.85	2.23	25	65	71	1.46	**2.38**	**3.13**	**0.45**
BOKU	0.61	−0.58	71	0.77	1.90	31	63	64	1.50	**2.43**	**3.12**	**0.46**
$BOKU_2$	0.71	−0.62	68	0.69	1.57	37	61	58	1.53	**2.48**	**3.11**	**0.46**
JAEA-2	0.50	−0.46	80	0.86	2.08	20	68	65	1.46	**2.50**	**3.30**	**0.49**
IRSN	0.61	−0.62	82	0.77	2.00	24	71	61	1.97	**2.59**	**3.35**	**0.37**
$NOAA\text{-}ARL_1$	0.55	−0.38	80	0.84	1.84	19	76	67	3.30	**2.66**	**3.47**	**0.32**
CMC_3	0.56	−0.33	83	0.83	1.69	16	76	61	1.28	**2.73**	**3.57**	**0.60**

(*continued on next page*)

Table A.5 (*continued*)

Organizations' runs	R	FB [mBq/m^3]	F5 [%]	$BC_{RMSE\ [mBq/m^3]}$	BC_{NMSE}	KS [%]	ACC [%]	NAAD [%]	Ratio stdvs	**Rank**	**$Rank_{KS}$**	**SS**
CMC_2	0.59	−0.34	84	0.81	1.64	17	73	59	1.22	**2.74**	**3.57**	**0.61**
NOAA-ARL	0.63	−0.48	83	0.78	1.73	20	76	62	2.55	**2.75**	**3.55**	**0.36**
CMC	0.57	−0.33	84	0.82	1.66	16	75	60	1.25	**2.75**	**3.59**	**0.62**
CMC_1	0.57	−0.31	84	0.83	1.65	16	77	60	1.24	**2.78**	**3.62**	**0.63**
$NOAA\text{-}ARL_3$	0.68	−0.56	85	0.74	1.73	21	76	60	2.28	**2.79**	**3.58**	**0.35**
$NOAA\text{-}ARL_2$	0.67	−0.50	85	0.74	1.62	19	76	59	2.06	**2.80**	**3.61**	**0.40**
$KAERI_5$	0.53	0.01	80	1.20	2.51	13	74	71	0.69	**2.83**	**3.70**	**0.84**
$KAERI_7$	0.53	−0.00	79	1.18	2.43	14	75	70	0.71	**2.83**	**3.69**	**0.84**
$KAERI_8$	0.53	−0.01	80	1.19	2.51	16	77	71	0.70	**2.85**	**3.69**	**0.84**
$KAERI_6$	0.54	−0.02	81	1.12	2.27	12	76	68	0.75	**2.85**	**3.73**	**0.85**
KAERI	0.53	−0.01	83	1.06	2.02	13	80	67	0.84	**2.90**	**3.78**	**0.86**
$KAERI_3$	0.53	−0.02	86	0.95	1.61	13	83	64	0.97	**2.97**	**3.84**	**0.88**
$KAERI_4$	0.53	−0.02	86	0.95	1.61	12	83	64	0.97	**2.97**	**3.85**	**0.88**
$KAERI_1$	0.54	−0.01	85	0.96	1.63	12	83	63	0.95	**2.97**	**3.85**	**0.88**
$KAERI_2$	0.54	−0.02	86	0.94	1.57	11	84	63	0.99	**2.98**	**3.87**	**0.88**
CTBTO-2	0.65	−0.02	82	0.78	1.09	9	81	56	1.07	**3.05**	**3.96**	**0.91**
Average over all organizations	**0.55**	**−0.62**	**70**	**0.85**	**3.03**	**30**	**63**	**69**	**1.95**	**2.33**	**3.03**	**0.44**
Median over all organizations	**0.55**	**−0.55**	**73**	**0.82**	**2.39**	**28**	**62**	**67**	**1.62**	**2.32**	**3.03**	**0.40**
Maximum over all organizations	**0.65**	**−0.01**	**84**	**1.06**	**14.20**	**68**	**81**	**91**	**6.00**	**3.05**	**3.96**	**0.91**
Minimum over all organizations	**0.38**	**−1.64**	**16**	**0.76**	**1.09**	**9**	**32**	**56**	**0.84**	**0.93**	**1.25**	**0.06**

Table A.6
Same as Table A.5, but with IRE and CNL stack emissions only.

Organizations' runs	R	FB [mBq/m^3]	F5 [%]	$BC_{RMSE\ [mBq/m^3]}$	BC_{NMSE}	KS [%]	ACC [%]	NAAD [%]	Ratio stdvs	**Rank**	**$Rank_{KS}$**	**SS**
AWE_2	0.44	−1.78	6	0.91	25.02	0	30	94	6.16	**0.67**	**1.67**	**0.05**
AWE	0.44	−1.78	6	0.91	25.02	0	30	94	6.16	**0.67**	**1.67**	**0.05**
AWE_1	0.44	−1.78	6	0.91	25.02	0	30	94	6.16	**0.67**	**1.67**	**0.05**
CTBTO-1	0.29	−1.32	23	0.94	7.62	62	34	88	2.20	**0.99**	**1.37**	**0.21**
$BOKU_2$	0.37	−1.34	22	0.91	7.32	64	36	86	1.89	**1.05**	**1.41**	**0.26**
BOKU	0.31	−1.21	24	0.96	6.66	59	37	89	1.73	**1.11**	**1.52**	**0.28**
FOI	0.39	−1.34	26	0.89	7.03	62	38	83	2.36	**1.12**	**1.50**	**0.21**
ARPANSA	0.38	−1.35	29	0.89	7.22	61	37	85	2.71	**1.13**	**1.52**	**0.17**
LLNL	0.35	−1.32	32	0.91	7.06	61	35	85	2.31	**1.14**	**1.53**	**0.21**
$CTBTO\text{-}2_1$	0.38	−1.27	27	0.90	6.40	60	38	82	1.97	**1.16**	**1.56**	**0.25**
$BOKU_1$	0.25	−1.08	27	1.01	6.01	54	38	92	1.57	**1.17**	**1.63**	**0.30**
ZAMG	0.38	−1.17	29	0.92	5.77	60	37	81	1.58	**1.22**	**1.62**	**0.32**
$NOAA\text{-}ARL_3$	0.44	−1.29	31	0.87	6.11	59	38	82	2.47	**1.23**	**1.64**	**0.20**
VINATOM	0.19	−1.05	34	1.01	5.76	51	40	90	1.84	**1.26**	**1.75**	**0.25**
NOAA-ARL	0.35	−1.16	39	0.90	5.44	51	40	81	2.75	**1.35**	**1.84**	**0.19**
$NOAA\text{-}ARL_2$	0.44	−1.20	36	0.87	5.28	55	42	79	2.18	**1.37**	**1.82**	**0.24**
JAEA-1	0.25	−0.72	31	1.07	4.28	47	42	94	1.27	**1.43**	**1.96**	**0.38**
$NOAA\text{-}ARL_1$	0.16	−1.01	49	0.96	4.91	38	42	83	3.60	**1.44**	**2.06**	**0.12**
CMC_3	0.39	−0.91	40	0.96	4.33	50	42	81	1.32	**1.52**	**2.02**	**0.38**
CMC	0.39	−0.89	41	0.97	4.29	49	43	81	1.29	**1.54**	**2.05**	**0.38**
CMC_1	0.39	−0.88	42	0.97	4.25	49	43	81	1.28	**1.55**	**2.06**	**0.38**
CMC_2	0.38	−0.88	42	0.97	4.29	49	43	81	1.28	**1.56**	**2.07**	**0.38**
JAEA-2	0.35	−0.80	43	0.96	3.81	42	46	77	1.45	**1.61**	**2.19**	**0.36**
PNNL	0.34	−0.81	43	1.05	4.61	45	49	84	1.13	**1.64**	**2.19**	**0.40**
IRSN	0.42	−1.00	47	0.88	4.12	47	49	75	2.02	**1.64**	**2.17**	**0.27**
$KAERI_5$	0.31	−0.58	43	1.32	5.60	44	46	93	0.77	**1.71**	**2.27**	**0.42**
$KAERI_6$	0.31	−0.63	46	1.26	5.34	44	47	90	0.83	**1.71**	**2.27**	**0.42**
$KAERI_8$	0.32	−0.58	46	1.31	5.49	45	47	92	0.78	**1.74**	**2.29**	**0.42**
$KAERI_7$	0.32	−0.59	48	1.30	5.47	44	49	92	0.79	**1.77**	**2.33**	**0.42**
KAERI	0.31	−0.60	50	1.20	4.76	39	50	87	0.92	**1.79**	**2.40**	**0.43**
$KAERI_1$	0.31	−0.60	54	1.12	4.10	33	52	82	1.03	**1.85**	**2.52**	**0.44**
$KAERI_3$	0.31	−0.61	55	1.11	4.06	34	52	82	1.06	**1.86**	**2.52**	**0.43**
$KAERI_2$	0.31	−0.62	56	1.10	4.01	34	52	82	1.07	**1.86**	**2.52**	**0.43**
$KAERI_4$	0.31	−0.60	55	1.11	4.04	35	52	83	1.05	**1.86**	**2.51**	**0.43**
Average over all organizations	**0.35**	**−1.11**	**33**	**0.96**	**6.87**	**50**	**40**	**85**	**2.11**	**1.30**	**1.80**	**0.27**
Median over all organizations	**0.35**	**−1.17**	**32**	**0.93**	**5.77**	**51**	**39**	**84**	**1.91**	**1.24**	**1.71**	**0.26**
Maximum over all organizations	**0.44**	**−0.60**	**50**	**1.20**	**25.02**	**62**	**50**	**94**	**6.16**	**1.79**	**2.40**	**0.43**
Minimum over all organizations	**0.19**	**−1.78**	**6**	**0.88**	**3.81**	**0**	**30**	**75**	**0.92**	**0.67**	**1.37**	**0.05**

Table A.7
Statistics for USX75 and time frame July, 8th, to September 30th, using stack emissions for IRE & CNL and including all remaining emitters. Results are shown per organization for individual run IDs (organization IDs plus subcripts) and over all run-IDs per participating organization ordered by rank. Bottom lines show metrics over all organizations after averaging over all run-IDs per organization. Number of valid sample pairs involved including observed samples below MDC: 166.

Organization's runs	R	FB [mBq/m^3]	F5 [%]	BC_{RMSE} [mBq/m^3]	BC_{NMSE}	KS [%]	ACC [%]	NAAD [%]	Ratio stdvs	**Rank**	**$Rank_{KS}$**	**SS**
AWE_2	0.16	−0.77	28	4.63	14.16	47	51	102	0.88	**1.42**	**1.95**	**0.36**
AWE	0.16	−0.77	28	4.63	14.12	47	51	102	0.88	**1.42**	**1.96**	**0.36**
AWE_1	0.16	−0.77	28	4.63	14.09	46	51	102	0.88	**1.42**	**1.96**	**0.36**
$BOKU_2$	0.28	−0.67	29	3.80	8.51	45	54	96	1.13	**1.58**	**2.13**	**0.41**
VINATOM	0.19	0.97	60	18.87	36.02	17	74	272	0.17	**1.89**	**2.72**	**0.08**
$NOAA\text{-}ARL_1$	0.26	0.94	57	7.53	5.98	33	77	215	0.43	**1.94**	**2.61**	**0.22**
BOKU	0.31	−0.19	47	5.69	10.48	32	66	118	0.77	**1.98**	**2.66**	**0.42**
JAEA-1	0.35	0.67	60	9.16	12.20	25	72	170	0.34	**2.11**	**2.86**	**0.22**
ARPANSA	0.44	−0.50	55	3.17	4.90	31	66	83	1.33	**2.15**	**2.84**	**0.48**
$KAERI_8$	0.36	0.39	57	7.74	11.85	28	69	140	0.40	**2.19**	**2.91**	**0.36**
CTBTO-1	0.54	−0.49	52	3.00	4.35	34	63	78	1.16	**2.19**	**2.85**	**0.52**
PNNL	0.43	0.44	56	6.96	9.12	31	67	137	0.43	**2.20**	**2.89**	**0.36**
$KAERI_6$	0.36	0.44	58	7.62	10.83	26	73	147	0.41	**2.22**	**2.96**	**0.34**
$KAERI_7$	0.38	0.41	59	7.45	10.74	27	70	142	0.42	**2.23**	**2.96**	**0.36**
$KAERI_5$	0.37	0.42	60	7.51	10.80	27	71	143	0.41	**2.24**	**2.97**	**0.35**
ZAMG	0.45	−0.32	57	3.52	5.02	31	67	88	0.99	**2.28**	**2.97**	**0.61**
NOAA-ARL	0.43	0.17	64	4.50	4.17	28	73	123	1.05	**2.34**	**3.07**	**0.53**
FOI	0.51	0.17	52	5.06	6.33	30	67	98	0.57	**2.37**	**3.07**	**0.67**
JAEA-2	0.56	0.94	67	6.73	4.82	24	86	197	0.42	**2.38**	**3.14**	**0.25**
$BOKU_1$	0.34	0.30	65	7.57	12.46	19	77	139	0.42	**2.39**	**3.20**	**0.43**
CTBTO-2	0.48	−0.33	60	3.21	4.22	26	73	82	1.16	**2.39**	**3.13**	**0.60**
KAERI	0.40	0.38	70	6.22	7.99	22	75	123	0.52	**2.43**	**3.22**	**0.44**
$NOAA\text{-}ARL_2$	0.46	−0.17	66	3.17	3.48	25	72	81	1.27	**2.50**	**3.25**	**0.73**
LLNL	0.55	−0.16	60	3.43	4.03	26	69	84	0.88	**2.52**	**3.26**	**0.78**
$NOAA\text{-}ARL_3$	0.56	−0.28	69	2.81	3.05	25	72	72	1.44	**2.59**	**3.34**	**0.63**
$KAERI_2$	0.43	0.39	83	5.12	5.18	15	80	107	0.60	**2.62**	**3.47**	**0.48**
$KAERI_3$	0.44	0.35	82	4.92	5.00	16	80	104	0.62	**2.64**	**3.48**	**0.52**
$KAERI_1$	0.44	0.35	83	4.95	5.02	16	80	104	0.62	**2.64**	**3.48**	**0.51**
$KAERI_4$	0.45	0.27	81	4.48	4.48	17	80	98	0.69	**2.67**	**3.50**	**0.60**
IRSN	0.65	0.01	72	2.87	2.40	26	75	74	0.96	**2.88**	**3.62**	**0.91**
CMC_2	0.76	0.35	75	3.04	1.90	16	82	81	0.72	**2.98**	**3.82**	**0.62**
CMC	0.76	0.40	80	3.25	2.08	16	84	84	0.67	**3.02**	**3.86**	**0.57**
CMC_1	0.77	0.43	81	3.26	2.00	16	85	86	0.66	**3.04**	**3.88**	**0.55**
CMC_3	0.75	0.40	83	3.47	2.34	15	85	85	0.64	**3.04**	**3.89**	**0.55**
Average over all organizations	**0.45**	**0.09**	**59**	**5.64**	**8.27**	**28**	**71**	**120**	**0.77**	2.29	3.01	0.49
Median over all organizations	**0.45**	**0.09**	**60**	**4.56**	**4.96**	**27**	**71**	**100**	**0.82**	2.31	3.02	0.50
Maximum over all organizations	**0.76**	**0.97**	**80**	**18.87**	**36.02**	**47**	**86**	**272**	**1.33**	3.02	3.86	0.91
Minimum over all organizations	**0.16**	**−0.77**	**28**	**2.87**	**2.08**	**16**	**51**	**74**	**0.17**	1.42	1.96	0.08

Table A.8
Same as Table A.7, but with IRE and CNL stack emissions only.

Organizations' runs	R	FB [mBq/m^3]	F5 [%]	BC_{RMSE} [mBq/m^3]	BC_{NMSE}	KS [%]	ACC [%]	NAAD [%]	Ratio stdvs	**Rank**	**$Rank_{KS}$**	**SS**
AWE_2	0.16	−0.79	23	4.63	14.49	48	50	103	0.88	**1.35**	**1.87**	**0.35**
AWE	0.16	−0.79	23	4.63	14.49	48	50	103	0.88	**1.35**	**1.87**	**0.35**
AWE_1	0.16	−0.79	23	4.63	14.49	48	50	103	0.88	**1.35**	**1.87**	**0.35**
$BOKU_2$	0.28	−0.67	29	3.80	8.51	45	54	96	1.13	**1.58**	**2.13**	**0.41**
VINATOM	0.19	0.96	49	18.87	36.54	23	72	272	0.17	**1.77**	**2.54**	**0.08**
BOKU	0.31	−0.20	42	5.69	10.68	35	63	119	0.77	**1.91**	**2.57**	**0.43**
$NOAA\text{-}ARL_1$	0.26	0.93	56	7.51	6.06	32	79	212	0.44	**1.96**	**2.64**	**0.22**
ARPANSA	0.44	−0.54	44	3.18	5.17	34	63	84	1.33	**1.99**	**2.65**	**0.46**
$KAERI_8$	0.36	0.36	43	7.74	12.25	35	63	141	0.40	**2.01**	**2.66**	**0.38**
$KAERI_5$	0.37	0.39	45	7.50	11.18	32	64	144	0.41	**2.03**	**2.71**	**0.37**
$KAERI_7$	0.37	0.38	45	7.45	11.12	33	64	143	0.42	**2.04**	**2.71**	**0.38**
$KAERI_6$	0.36	0.41	45	7.62	11.22	31	66	149	0.41	**2.04**	**2.73**	**0.35**
JAEA-1	0.35	0.66	55	9.15	12.36	26	71	170	0.34	**2.05**	**2.79**	**0.22**
PNNL	0.43	0.42	46	6.94	9.25	34	64	137	0.43	**2.08**	**2.74**	**0.37**
CTBTO-2	0.47	−0.43	45	3.22	4.71	35	64	85	1.18	**2.09**	**2.74**	**0.53**
CTBTO-1	0.54	−0.53	46	3.01	4.58	37	62	79	1.17	**2.11**	**2.74**	**0.51**
ZAMG	0.45	−0.36	46	3.52	5.24	34	65	89	0.99	**2.13**	**2.79**	**0.58**
NOAA-ARL	0.43	0.12	56	4.50	4.33	30	71	122	1.05	**2.22**	**2.92**	**0.49**
$BOKU_1$	0.34	0.27	54	7.57	12.85	24	72	141	0.42	**2.25**	**3.01**	**0.46**
KAERI	0.40	0.34	56	6.22	8.27	26	71	124	0.52	**2.26**	**2.99**	**0.46**
FOI	0.51	0.14	47	5.04	6.44	32	64	97	0.57	**2.31**	**2.99**	**0.69**
$NOAA\text{-}ARL_2$	0.46	−0.22	55	3.17	3.69	28	69	82	1.27	**2.33**	**3.05**	**0.68**

(*continued on next page*)

Table A.8 (*continued*)

Organizations' runs	R	FB [mBq/m^3]	F5 [%]	BC_{RMSE} [mBq/m^3]	BC_{NMSE}	KS [%]	ACC [%]	NAAD [%]	Ratio stdvs	**Rank**	**$Rank_{KS}$**	**SS**
JAEA-2	0.56	0.93	63	6.71	4.84	24	86	196	0.42	**2.34**	**3.10**	**0.25**
$NOAA\text{-}ARL_3$	0.56	−0.34	56	2.81	3.25	30	67	73	1.45	**2.37**	**3.07**	**0.57**
LLNL	0.55	−0.18	54	3.42	4.10	28	68	85	0.88	**2.43**	**3.15**	**0.76**
$KAERI_3$	0.44	0.31	66	4.93	5.18	20	78	104	0.62	**2.47**	**3.27**	**0.54**
$KAERI_2$	0.43	0.36	68	5.12	5.36	18	78	108	0.60	**2.47**	**3.29**	**0.50**
$KAERI_1$	0.44	0.32	67	4.95	5.20	20	78	105	0.62	**2.48**	**3.28**	**0.53**
$KAERI_4$	0.45	0.24	66	4.48	4.64	20	76	99	0.69	**2.50**	**3.30**	**0.64**
IRSN	0.64	−0.02	61	2.87	2.48	31	69	75	0.97	**2.71**	**3.40**	**0.91**
CMC_3	0.75	0.38	71	3.46	2.39	19	80	85	0.64	**2.89**	**3.70**	**0.57**
CMC_2	0.76	0.32	69	3.03	1.94	20	78	80	0.72	**2.89**	**3.69**	**0.64**
CMC	0.76	0.37	70	3.24	2.13	19	80	83	0.68	**2.90**	**3.70**	**0.59**
CMC_1	0.77	0.41	71	3.24	2.04	19	81	86	0.66	**2.92**	**3.73**	**0.56**
Average over all organizations	**0.45**	**0.06**	**50**	**5.64**	**8.47**	**31**	**68**	**120**	**0.77**	**2.16**	**2.85**	**0.48**
Median over all organizations	**0.44**	**0.05**	**48**	**4.56**	**5.20**	**32**	**67**	**100**	**0.83**	**2.12**	**2.79**	**0.48**
Maximum over all organizations	**0.76**	**0.96**	**70**	**18.87**	**36.54**	**48**	**86**	**272**	**1.33**	**2.90**	**3.70**	**0.91**
Minimum over all organizations	**0.16**	**−0.79**	**23**	**2.87**	**2.13**	**19**	**50**	**75**	**0.17**	**1.35**	**1.87**	**0.08**

A.2. Results of the correlation analysis in order to reduce the full ensemble

Table A.9
List of models retained as part of the reduced ensemble and those discarded because of being affected by high correlation (>0.95) with other model sets.

	CAX17	DEX33	SEX63	USX75
Number of original/retained members	27/17	28/21	26/17	26/16
Models excluded	AWE_1	$CTBTO\text{-}2_1$	ARPANSA	AWE_1
	CMC_1	$KAERI_1$	AWE_1	CMC_1
	$CTBTO\text{-}2_1$	$KAERI_2$	CMC_1	CTBTO-2
	$KAERI_1$	$KAERI_3$	$KAERI_1$	$KAERI_1$
	$KAERI_2$	$KAERI_6$	$KAERI_2$	$KAERI_2$
	$KAERI_3$	$KAERI_7$	$KAERI_3$	$KAERI_3$
	$KAERI_5$	$NOAA\text{-}ARL_2$	$KAERI_5$	$KAERI_5$
	$KAERI_6$		$KAERI_6$	$KAERI_6$
	$KAERI_7$		$KAERI_7$	$KAERI_7$
	$NOAA\text{-}ARL_2$			$NOAA\text{-}ARL_2$
Models included	ARPANSA	$ARPANSA_1$	AWE_2	ARPANSA
	AWE_2	$ARPANSA_2$	$BOKU_1$	AWE_2
	CMC_2	AWE_1	$BOKU_2$	$BOKU_1$
	CMC_3	AWE_2	CMC_2	$BOKU_2$
	$CTBTO\text{-}2_2$	$BOKU_1$	CMC_3	CMC_2
	$KAERI_4$	$BOKU_2$	CTBTO-2	CMC_3
	$NOAA\text{-}ARL_1$	CMC_1	FOI	FOI
	$NOAA\text{-}ARL_3$	CMC_2	IRSN	IRSN
	$BOKU_1$	CMC_3	$KAERI_4$	$KAERI_4$
	$BOKU_2$	$CTBTO\text{-}2_2$	$KAERI_8$	$KAERI_8$
	FOI	FOI	LLNL	LLNL
	IRSN	IRSN	$NOAA\text{-}ARL_1$	$NOAA\text{-}ARL_1$
	$KAERI_8$	$KAERI_4$	$NOAA\text{-}ARL_2$	$NOAA\text{-}ARL_3$
	VINATOM	$KAERI_5$	$NOAA\text{-}ARL_3$	PNNL
	PNNL	$KAERI_8$	PNNL	VINATOM
	ZAMG	LLNL	VINATOM	ZAMG
	LLNL	$NOAA\text{-}ARL_1$	ZAMG	
		$NOAA\text{-}ARL_3$		
		PNNL		
		VINATOM		
		ZAMG		

References

Achim, P., Generoso, S., Morin, M., Gross, P., Le Petit, G., Moulin, C., 2016. Characterization of Xe-133 global atmospheric background: Implications for the International Monitoring System of the Comprehensive Nuclear-Test-Ban-Treaty. J. Geophys. Res. Atmos. 121, 4951–4966. https://doi.org/10.1002/2016JD024872.

Bowyer, T.W., 2020. A Review of Global Radioxenon Background Research and Issues. Pure Appl. Geophys. https://doi.org/10.1007/s00024-020-02440-0.

Bretherton, C.S., Widmann, M., Dymnikov, V.P., Wallace, J.M., Bladé, I., 1999. The effective number of spatial degrees of freedom of a time-varying field. J. Clim. 12, 1990–2009.

Buehner, M., McTaggart-Cowan, R., Beaulne, A., Charette, C., Garand, L., Heilliette, S., Lapalme, E., Laroche, S., MacPherson, S.R., Morneau, J., Zadra, A., 2015. Implementation of Deterministic Weather Forecasting Systems Based on Ensemble-Variational Data Assimilation at Environment Canada. Part I: The Global System. Mon. Weather Rev. 143, 2532–2559. https://doi.org/10.1175/MWR-D-14-00354.1.

Buehner, M., Morneau, J., Charette, C., 2013. Four-dimensional ensemble-variational data assimilation for global deterministic weather prediction. Nonlinear Process Geophys. 20, 669–682. https://doi.org/10.5194/npg-20-669-2013.

Charron, M., Polavarapu, S., Buehner, M., Vaillancourt, P.A., Charette, C., Roch, M., Morneau, J., Garand, L., Aparicio, J.M., MacPherson, S., Pellerin, S., St-James, J., Heilliette, S., 2012. The Stratospheric Extension of the Canadian Global Deterministic Medium-Range Weather Forecasting System and Its Impact on Tropospheric Forecasts. Mon. Weather Rev. 140, 1924–1944. https://doi.org/10.1175/MWR-D-11-00097.1.

Ctbto, 1996. Text of the Comprehensive Nuclear-Test-Ban Treaty. URL: http://www.ctbto.org/the-treaty/treaty-text/. online. accessed January, 24th, 2017.
Ctbto, 2008. Certification of Noble Gas Equipment at IMS Radionuclide Stations (With Guidelines for Station Installation). Technical Report CTBT/PTS/INF.921/Rev.3. Online on CTBTO's secure webportal accessed Oct., 15th, 2021.
Ctbto, 2020a. VDEC request for access. https://www.ctbto.org/specials/vdec/vdec-request-for-access/.online accessed July, 15th, 2020.
Ctbto, 2020b. WEBGRAPE 1.8.6. CTBTO Preparatory Commission, International Data Center (IDC). Technical Report.
Ctbto Preparatory Commission, 2019. Verification regime. Technical report. URL: http://www.ctbto.org/verification-regime/. online. accessed Feb., 27th, 2019.
D'Amours, R., Malo, A., Flesch, T., Wilson, J., Gauthier, J.P., Servranckx, R., 2015. The Canadian Meteorological Centre's atmospheric transport and dispersion modelling suite. Atmos.-Ocean 53, 176–199. https://doi.org/10.1080/07055900.2014.1000260.
D'Amours, R., Malo, A., Servranckx, R., Bensimon, D., Trudel, S., Gauthier-Bilodeau, J.P., 2010. Application of the atmospheric Lagrangian particle dispersion model MLDP0 to the 2008 eruptions of Okmok and Kasatochi volcanoes. J. Geophys. Res. Atmos. 115, 1–11. https://doi.org/10.1029/2009JD013602.
Davies, T., Cullen, M.J.P., Malcolm, A.J., Mawson, M.H., Staniforth, A., White, A.A., Wood, N., 2005. A new dynamical core for the Met Office's global and regional modelling of the atmosphere. Q. J. R. Meteorol. Soc. 131, 1759–1782. https://doi.org/10.1256/qj.04.101.
De Meutter, P., Camps, J., Delcloo, A., Deconninck, B., Termonia, P., 2018. Time resolution requirements for civilian radioxenon emission data for the CTBT verification regime. J. Environ. Radioact. 182, 117–127. https://doi.org/10.1016/j.jenvrad.2017.11.027.
Déqué, M., Dreveton, C., Braun, A., Cariolle, D., 1994. The ARPEGE/IFS atmosphere model: a contribution to the French community climate modelling. Clim. Dynam. 10, 249–266. https://doi.org/10.1007/BF00208992.
Déqué, M., Piedelievre, J.P., 1995. High resolution climate simulation over Europe. Clim. Dynam. 11, 321–339. https://doi.org/10.1007/BF00215735.
Ecmwf, 2018. ECMWF official homepage. URL: https://www.ecmwf.int/.online. accessed June, 27th, 2020.
Ecmwf, 2020. ERA5 documentation. URL: https://confluence.ecmwf.int/display/CKB/ERA5%3A+data+documentation.online. accessed July, 15th, 2020.
Ermak, D.L., Nasstrom, J.S., 2000. A Lagrangian stochastic diffusion method for inhomogeneous turbulence. Atmos. Environ. 34, 1059–1068. https://doi.org/10.1016/S1352-2310(99)00379-9.
Eslinger, P.W., Bowyer, T.W., Achim, P., Chai, T., Deconninck, B., Freeman, K., Generoso, S., Hayes, P., Heidmann, V., Hoffman, I., Kijima, Y., Krysta, M., Malo, A., Maurer, C., Ngan, F., Robins, P., Ross, J.O., Saunier, O., Schlosser, C., Schoeppner, M., Schrom, B.T., Seibert, P., Stein, A.F., Ungar, K., Yi, J., 2016. International challenge to predict the impact of radioxenon releases from medical isotope production on a comprehensive nuclear test ban treaty sampling station. J. Environ. Radioact. 157, 41–51. https://doi.org/10.1016/j.jenvrad.2016.03.001.
Fontaine, J.P., Pointurier, F., Blanchard, X., Taffary, T., 2004. Atmospheric xenon radioactive isotope monitoring. J. Environ. Radioact. 72, 129–135.
Galmarini, S., Bianconi, R., Klug, W., Mikkelsen, T., Addis, R., Andronopoulos, S., Astrup, P., Baklanov, A., Bartniki, J., Bartzis, J.C., et al., 2004. Ensemble dispersion forecasting – Part I: concept, approach and indicators. Atmos. Environ. 38, 4607–4617.
Generoso, S., Achim, P., Morin, M., Gross, P., Douysset, G., 2022. Use of STAX data in global-scale simulation of 133Xe atmospheric background. J. Environ. Radioactiv. Submitted.
Goodwin, M.A., Britton, R., Davies, A.V., 2021a. A Consideration of Radioxenon Detections Around the Korean Peninsula. Pure Appl. Geophys. 178, 2651–2664. https://doi.org/10.1007/s00024-020-02500-5.
Goodwin, M.A., Davies, A.V., Britton, R., 2021b. Analysis of environmental radioxenon detections in the UK. J. Environ. Radioact. 234 https://doi.org/10.1016/j.jenvrad.2021.106629.
Groëll, J., Quélo, D., Mathieu, A., 2014. Sensitivity analysis of the modelled deposition of 137 Cs on the Japanese land following the Fukushima accident. Int. J. Environ. Pollut. 55, 67–75. https://doi.org/10.1504/ijep.2014.065906.
Gueibe, C., Kalinowski, M.B., Baré, J., Gheddou, A., Krysta, M., Kuśmierczyk-Michulec, J., 2017. Setting the baseline for estimated background observations at IMS systems of four radioxenon isotopes in 2014. J. Environ. Radioact. 178–179, 297–314.
Jma, 2019. Outline of the Operational Numerical Weather Prediction at the Japan Meteorological Agency. URL: https://www.jma.go.jp/jma/jma-eng/jma-center/nwp/outline2019-nwp/index.htm. online; accessed Nov., 3rd, 2021.
Jones, A., Thomson, D., Hort, M., Devenish, B., 2007. The U.K. Met Office's Next-Generation Atmospheric Dispersion Model, NAME III. Springer, pp. 580–589.
Kalinowski, M.B., 2022. Global emission inventory of 131mXe, 133Xe, 133mXe, and 135Xe from all kinds of nuclear facilities for the reference year 2014. J. Environ. Radioactiv. Submitted.
Kalinowski, M.B., Tatlisu, H., 2020a. Global radioxenon emission inventory from nuclear power plants for the calendar year 2014. Pure Appl. Geophys. 178, 2695–2708. https://doi.org/10.1007/s00024-020-02579-w.
Kalinowski, M.B., Tatlisu, H., 2020b. Global radioxenon emission inventory from nuclear power plants for the calendar year 2014. Correction. Pure Appl. Geophys. 178, 2709–2710. https://doi.org/10.1007/s00024-021-02812-0.
Kalinowski, M.B., Tayyebi, P., Lechermann, M., Tatlisu, H., 2021. Global radioxenon emission inventory from nuclear research reactors. Pure Appl. Geophys. 178, 2711–2739. https://doi.org/10.1007/s00024-021-02719-w.
Kioutsioukis, I., Galmarini, S., 2014. De praeceptis ferendis: good practices in multi-model ensembles. Atmos. Chem. Phys. 14, 11791–11815. https://doi.org/10.5194/acp-14-11791-2014.
Kma, 2018. Annual Joint WMO Technical Progress Report on the Global Data Processing and Forecasting System and Numerical Weather Prediction Research Activities for 2018. Technical Report. Korea Meteorological Administration, Republic of Korea.
Kma, 2021. Weather Forecast/Numerical Weather Prediction (NWP). URL: http://web.kma.go.kr/eng/biz/forecast_02.jsp. online; accessed August, 31st, 2021.
Knutti, R., Masson, D., Gettelman, A., 2013. Climate model genealogy: Generation CMIP5 and how we got there. Geophys. Res. Lett. 40, 1194–1199. https://doi.org/10.1002/grl.50256.
Kuśmierczyk-Michulec, J., Becker, A., Wotawa, G., Krysta, M., Bourgouin, P., Tipka, A., Kalinowski, M.B., 2021. Advancements in atmospheric transport modelling (ATM) at the CTBTO PTS during the past two decades and plans for the future. In: CTBT Science and Technology Conference 2021. URL: https://conferences.ctbto.org/event/7/contributions/1367/.online. accessed December, 20th, 2021.
Kuśmierczyk-Michulec, J., Deconninck, B., Kalinowski, M., Hoffmann, E., 2019. Quantifying uncertainties in the atmospheric modelling (ATM) simulations resulting from different emission time resolution. In: CTBT Science and Technology Conference 2019 (SnT2019). URL: https://ctnw.ctbto.org/ctnw/abstract/32825.online. accessed December, 20th, 2021.
Larson, D.J., Nasstrom, J.S., 2002. Shared- and distributed-memory parallelization of a Lagrangian atmospheric dispersion model. Atmos. Environ. 36, 1559–1564. https://doi.org/10.1016/S1352-2310(01)00540-4.
Matthews, K.M., De Geer, L.E., 2004. Processing of data from a global atmospheric radioactivity monitoring network for CTBT verification purposes. J. Radioanal. Nucl. Chem. 263, 235–240.
Maurer, C., Arnold Arias, D., Brioude, J., Haselsteiner, M., Weidle, F., Haimberger, L., Skomorowski, P., Bourgouin, P., 2021. Evaluating the added value of multi-input atmospheric transport ensemble modeling for applications of the Comprehensive Nuclear Test-Ban Treaty organization (CTBTO). J. Environ. Radioact. 237, 106649 https://doi.org/10.1016/j.jenvrad.2021.106649.
Maurer, C., Baré, J., Kuśmierczyk-Michulec, J., Crawford, A., Eslinger, P.W., Seibert, P., Orr, B., Philipp, A., Ross, O., Generoso, S., Achim, P., Schoeppner, M., Malo, A., Ringbom, A., Saunier, O., Quélo, D., Mathieu, A., Kijima, Y., Stein, A., Chai, T., Ngan, F., Leadbetter, S.J., De Meutter, P., Delcloo, A., Britton, R., Davies, A., Glascoe, L.G., Lucas, D.D., Simpson, M.D., Vogt, P., Kalinowski, M., Bowyer, T.W., 2018. International challenge to model the long-range transport of radioxenon released from medical isotope production to six Comprehensive Nuclear- Test-Ban Treaty monitoring stations. J. Environ. Radioact. 192, 667–686. https://doi.org/10.1016/j.jenvrad.2018.01.030.
Metoffice, 2021. Unified model partnership. URL: https://www.metoffice.gov.uk/research/approach/collaboration/unified-model/partnership. online. accessed August, 16th, 2021.
Ncep, 2003. Environmental Modeling Center: The GFS Atmospheric Model. NOAA/NCEP, Environmental Modeling Center Office Note 442. Technical Report. URL: http://www.emc.ncep.noaa.gov/officenotes/newernotes/on442.pdf. online; accessed June, 12th, 2017.
Noaa, 2020. GFS/GDAS changes since 1991 – history of recent modifications to the global forecast/analysis system. URL: https://www.emc.ncep.noaa.gov/gmb/STATS/html/model_changes.html. online. accessed August, 5th, 2020.
Pennel, C., Reichler, T., 2011. On the effective numbers of climate models. J. Clim. 24, 2358–2367.
Pisso, I., Sollum, E., Grythe, H., Kristiansen, N.I., Cassiani, M., Eckhardt, S., Arnold, D., Morton, D., Thompson, R.L., Groot Zwaaftink, C.D., Evangeliou, N., Sodemann, H., Haimberger, L., Henne, S., Brunner, D., Burkhart, J.F., Fouilloux, A., Brioude, J., Philipp, A., Seibert, P., Stohl, A., 2019. The Lagrangian particle dispersion model FLEXPART version 10.4. Geosci. Model Dev. (GMD) 12, 4955–4997. https://doi.org/10.5194/gmd-12-4955-2019.
Pnnl, 2017. WOSMIP VI - Workshop on Signatures of Man-Made Isotope Production. URL: https://www.wosmip.org/sites/default/files/documents/wosmipVI2017.pdf. online; accessed December, 1st, 2021.
Puri, K., Dietachmayer, G., Steinle, P., Dix, M., Rikus, L., Logan, L., Naughton, M., Tingwell, C., Xiao, Y., Barras, V., Bermous, I., Bowen, R., Deschamps, L., Franklin, C., Fraser, J., Glowacki, T., Harris, B., Lee, J., Le, T., Roff, G., Sulaiman, A., Sims, H., Sun, X., Sun, Z., Zhu, H., Chattopadhyay, M., Engel, C., 2013. Implementation of the initial ACCESS numerical weather prediction system. Aust. Met. Oc. Journal 63, 265–284.
Ringbom, A., Axelsson, A., Aldener, M., Auer, M., Bowyer, T.W., Fritioff, T., Hoffman, I., Khrustalev, K., Nikkinen, M., Popov, V., Popov, Y., Ungar, K., Wotawa, G., 2014. Radioxenon detections in the CTBT international monitoring system likely related to the announced nuclear test in North Korea on February 12, 2013. J. Environ. Radioact. 128, 47–63. https://doi.org/10.1016/j.jenvrad.2013.10.027.
Ringbom, A., Elmgren, K., Lindh, K., Peterson, J., Bowyer, T.W., Hayes, J.C., McIntyre, J. I., Panisko, M., Williams, R., 2009. Measurements of radioxenon in ground level air in South Korea following the claimed nuclear test in North Korea on October 9, 2006. J. Radioanal. Nucl. Chem. 282, 773–779.
Ringbom, A., Larson, T., Axelsson, A., Elmgren, K., Johansson, C., 2003. SAUNA – a system for automatic sampling, processing, and analysis of radioactive xenon. Nucl. Instrum. Methods A 508, 542–553.
Saey, P.R., 2009. The influence of radiopharmaceutical isotope production on the global radioxenon background. J. Environ. Radioact. 100, 396–406. https://doi.org/10.1016/j.jenvrad.2009.01.004.
Saey, P.R.J., Bean, M., Becker, A., Coyne, J., d'Amours, R., De Geer, L.E., Hogue, R., Stocki, T.J., Ungar, R.K., Wotawa, G., 2007. A long distance measurement of

radioxenon in Yellowknife, Canada, in late October 2006. Geophys. Res. Lett. 34, L20802 https://doi.org/10.1029/2007GL030611.
Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., Behringer, D., Hou, Y.T., Chuang, H.Y., Iredell, M., Ek, M., Meng, J., Yang, R., Mendez, M.P., van den Dool, H., Zhang, Q., Wang, W., Chen, M., Becker, E., 2014. The NCEP Climate Forecast System Version 2. J. Clim. 27, 2185–2208.
Seibert, P., 2004. Remarks on Statistical Parameters for Model Evaluation. Institute of Meteorology, BOKU, Vienna, Austria.
Simmons, A.J., Burridge, D.M., Jarraud, M., Girard, C., Wergen, W., 1989. The ECMWF medium-range prediction models development of the numerical formulations and the impact of increased resolution. Meteorol. Atmos. Phys. 40, 28–60. https://doi.org/10.1007/BF01027467.
Skamarock, W., Klemp, J., Dudhia, J., Gill, D.O., Barker, D., Duda, M.G., Huang, X., Wang, W., 2008. A Description of the Advanced Research WRF Version 3. Report NCAR/TN-475+STR. National Center for Atmospheric Research. https://doi.org/10.5065/D68S4MVH.
Solazzo, E., Galmarini, S., 2015. A science-based use of ensembles of opportunities for assessment and scenario study. Atmos. Chem. Phys. Discuss. 15, 2535–2544.
Solazzo, E., Riccio, A., Kioutsioukis, I., Galmarini, S., 2013. Pauci ex tanto numero: reduce redundancy in multi-model ensembles. Atmos. Chem. Phys. 13, 8315–8333. URL: https://www.atmos-chem-phys.net/13/8315/2013/. doi:10.5194/acp-13-8315-2013.
Stein, A.F., Draxler, R.R., Rolph, G.D., Stunder, B.J.B., Cohen, M.D., Ngan, F., 2015. NOAA's HYSPLIT atmospheric transport and dispersion modeling system. Bull. Am. Meteorol. Soc. 96, 2059–2077. https://doi.org/10.1175/BAMS-D-14-00110.1.
Stohl, A., Forster, C., Frank, A., Seibert, P., Wotawa, G., 2005. Technical note: the Lagrangian particle dispersion model FLEXPART version 6.2. Atmos. Chem. Phys. 5, 2461–2474. https://doi.org/10.5194/acp-5-2461-2005.
Stohl, A., Hittenberger, M., Wotawa, G., 1998. Validation of the Lagrangian particle dispersion model FLEXPART against large-scale tracer experiment data. Atmos. Environ. 32, 4245–4264. https://doi.org/10.1016/s1352-2310(98)00184-8.
Suh, K.S., Jeong, H.J., Kim, E.H., Hwang, W.T., Han, M.H., 2006. Verification of the Lagrangian particle model using the ETEX experiment. Ann. Nucl. Energy 33, 1159–1163.
Talagrand, O., Vautard, R., Strauss, B., 1997. Evaluation of probabilistic prediction systems. ECMWF, Workshop on Predictability, pp. 20–22. October 1997.
Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. J. Geophys. Res. 106, 7183–7192.
Terada, H., Chino, M., 2008. Development of an atmospheric dispersion model for accidental discharge of radionuclides with the function of simultaneous prediction for multiple domains and its evaluation by application to the Chernobyl nuclear accident. J. Nucl. Sci. Technol. 45, 920–931. https://doi.org/10.3327/jnst.45.920.
Terada, H., Nagai, H., Yamazawa, H., 2013. Validation of a Lagrangian atmospheric dispersion model against middle-range scale measurements of Kr-85 concentration in Japan. J. Nucl. Sci. Technol. 50, 1198–1212. https://doi.org/10.1080/00223131.2013.840545.
Tombette, M., Quentric, E., Quélo, D., Benoit, J.P., Mathieu, A., Korsakissok, I., Didier, D., 2014. C3X: a software platform for assessing the consequences of an accidental release of radioactivity into the atmosphere. In: Posters Presented at Fourth European IRPA Congress. June 2014, Geneva, pp. 23–27.
Wotawa, G., De Geer, L.E., Denier, P., Kalinowski, M., Toivonen, H., D'Amours, R., Desiato, F., Issartel, J.P., Langer, M., Seibert, P., Frank, A., Sloani, C., Yamazawa, H., 2003. Atmospheric transport modelling in support of CTBT verification - overview and basic concepts. Atmos. Environ. 37, 2529–2537.