Version of Record: https://www.sciencedirect.com/science/article/pii/S0265931X16300510
Manuscript_bf019b99d64db035f374650715cc407f

International Challenge to Predict the Impact of Radioxenon Releases from Medical Isotope Production on a Comprehensive Nuclear Test Ban Treaty Sampling Station

Paul W. Eslinger [a], Ted W. Bowyer [a], Pascal Achim[b], Tianfeng Chai[c], Benoit Deconninck[d], Katie Freeman[e], Sylvia Generoso[b], Philip Hayes[f], Verena Heidmann[g], Ian Hoffman[h], Yuichi Kijima[i], Monika Krysta[j], Alain Malo[k], Christian Maurer[l], Fantine Ngan[c], Peter Robins[e], J. Ole Ross[m], Olivier Saunier[n], Clemens Schlosser[g], Michael Schöppner[o], Brian T. Schrom[a], Petra Seibert[p], Ariel F. Stein[c], Kurt Ungar[h], Jing Yi[h]

Author Affiliations:

[a] Pacific Northwest National Laboratory, 902 Battelle Boulevard, Richland, WA, 99352 USA
[b] Commissariat à l'Energie Atomique, CEA, DAM, DIF, 91297 Arpajon, France
[c] NOAA/Air Resources Laboratory, College Park, Maryland
[d] Institut des Radioéléments, Fleurus, Belgium
[e] AWE, Aldermaston, Reading, RG7 4PR, United Kingdom
[f] Air Force Technical Applications Center, Patrick Air Force Base, Florida, USA
[g] Federal Office for Radiation Protection (*Bundesamt für Strahlenschutz, Bfs*), Frieburg, Germany
[h] Health Canada, Radiation Protection Bureau, Ottawa, Canada
[i] Japan Atomic Energy Agency, Tokai, Ibaraki, Japan
[j] Comprehensive Test Ban Treaty Organization (CTBTO), International Data Center, Vienna, Austria
[k] Environment Canada, Canadian Meteorological Centre, Dorval, Canada
[l] Zentralanstalt für Meteorologie und Geodynamik, Vienna, Austria
[m] Federal Institute for Geosciences and Natural Resources (BGR), Hannover, Germany
[n] French Institute for Radiation Protection and Nuclear Safety, Fontenay-aux-Roses, France
[o] Program on Science and Global Security, Princeton University, Princeton New Jersey USA
[p] University of Natural Resources and Life Sciences, Institute of Meteorology and University of Vienna, Department of Meteorology and Geophysics, Vienna, Austria

Corresponding author:

Paul W. Eslinger
Tel.: +1 509 372 4392
E-mail address: paul.w.eslinger@pnnl.gov
Pacific Northwest National Laboratory, MSIN K7-76, 902 Battelle Boulevard, P.O. Box 999, Richland, Washington, USA

Author Email Addresses:

Paul W. Eslinger paul.w.eslinger@pnnl.gov
Ted. W. Bowyer ted.bowyer@pnnl.gov
Brian T. Schrom brian.schrom@pnnl.gov
Pascal Achim pascal.achim@cea.fr
Tianfeng Chai tianfeng.chai@noaa.gov
Benoit Deconninck Benoit.Deconninck@ire-elit.eu
Sylvia Generoso sylvia.generoso@cea.fr
Philip Hayes philip.hayes.2@us.af.mil
Ian Hoffman ian.hoffman@hc-sc.gc.ca
Yuichi Kijima kijima.yuichi@jaea.go.jp

Monika Krysta monika.krysta@ctbto.org
Christian Maurer christian.maurer@zamg.ac.at
Fantine Ngan fantine.ngan@noaa.gov
Peter Robins peter.robins@awe.co.uk
Ole Ross ole.ross@bgr.de
Olivier Saunier olivier.saunier@irsn.fr
Clemens Schlosser cschlosser@bfs.de
Michael Schoeppner schoeppner@princeton.edu
Petra Seibert petra.seibert@univie.ac.at
Ariel F. Stein ariel.stein@noaa.gov
Kurt Ungar Kurt.Ungar@hc-sc.gc.ca

International Challenge to Predict the Impact of Radioxenon Releases from Medical Isotope Production on a Comprehensive Nuclear Test Ban Treaty Sampling Station

Abstract

The International Monitoring System (IMS) is part of the verification regime for the Comprehensive Nuclear-Test-Ban-Treaty Organization (CTBTO). At entry-into-force, half of the 80 radionuclide stations will be able to measure concentrations of several radioactive xenon isotopes produced in nuclear explosions, and then the full network may be populated with xenon monitoring afterward. An understanding of natural and man-made radionuclide backgrounds can be used in accordance with the provisions of the treaty (such as event screening criteria in Annex 2 to the Protocol of Treaty) for the effective implementation of the verification regime.

Fission-based production of ^{99}Mo for medical purposes also generates nuisance radioxenon isotopes that are usually vented to the atmosphere. One of the ways to account for the effect emissions from medical isotope production has on radionuclide samples from the IMS is to use stack monitoring data, if they are available, and atmospheric transport modeling. Recently, individuals from seven nations participated in a challenge exercise that used atmospheric transport modeling to predict the time-history of ^{133}Xe concentration measurements at the IMS radionuclide station in Germany using stack monitoring data from a medical isotope production facility in Belgium. Participants received only stack monitoring data and used the atmospheric transport model and meteorological data of their choice.

Some of the models predicted the highest measured concentrations quite well. A model comparison rank and ensemble analysis suggests that combining multiple models may provide more accurate predicted concentrations than any single model. None of the submissions based only on the stack monitoring data predicted the small measured concentrations very well. Modeling of sources by other nuclear facilities

with smaller releases than medical isotope production facilities may be important in understanding how to discriminate those releases from releases from a nuclear explosion.

Keywords

Medical isotope production; ^{133}Xe; source-term estimation; atmospheric modeling; CTBTO

1. Introduction

The International Monitoring System (IMS) is part of the verification regime for the Comprehensive Nuclear-Test-Ban-Treaty Organization (CTBTO, 2014). The verification regime is designed to detect nuclear explosions no matter where they occur on the earth. When complete, 80 of the IMS stations will have aerosol measurement systems sensitive enough to detect releases from nuclear explosions at great distances. At entry-into-force, half of the 80 stations will also have equipment that measures concentrations of four radioactive xenon isotopes (^{131m}Xe, ^{133}Xe, ^{133m}Xe, and ^{135}Xe) produced in a nuclear explosion, and following entry-into-force, a plan to add xenon monitoring capabilities to the other 40 stations will be reviewed (Comprehensive Nuclear-Test-Ban Treaty, 1996). An understanding of natural and man-made radionuclide backgrounds can also be used in accordance with the provisions of the treaty (such as event screening criteria in Annex 2 to the Protocol of Treaty) for the effective implementation of the verification regime.

A number of studies of the release and transport of radioxenon from nuclear explosions, nuclear power plants, and medical isotope production facilities have been published (Becker et al., 2010; Eslinger et al., 2014; Hoffman et al., 2009; Kalinowski et al., 2008; Saey et al., 2010b; Wotawa et al., 2010; Wotawa et al., 2003; Zähringer et al., 2009). These studies confirm that fission-based production of ^{99}Mo for medical purposes is the largest routine contributor of radioxenon to worldwide background levels. The ^{99}Mo (half-life of 66 hours) decays into ^{99m}Tc (half-life of 6 hours) and the resulting ^{99m}Tc is used in approximately

30-40 million medical procedures per year (Peykov and Cameron, 2014) and the demand is expected to increase in the future.

A reduction in radioxenon releases to relatively low levels (Bowyer et al., 2013) has the potential to reduce background radioxenon to levels that don't significantly impact treaty verification activities. However, medical isotope production facilities meet regulatory release requirements and their releases don't pose public health risks, thus the operators have no financial incentive to reduce releases. Another way of mitigating the impact on treaty verification activities is to use stack monitoring data, if they are available, and atmospheric transport modeling. In the modeling context, one could attempt to model background sources accurately enough to subtract a background contribution from any sampled value. Given the uncertainties (source terms, modeling), simulated peaks may not accurately represent reality. Thus, alternately, when a xenon peak is observed, one could check whether the simulated background increases during the same period (synchronization in time). If that is the case, the observed peak could be linked to the rise of the radioxenon background.

Unfortunately, the details of the stack monitoring data needed, such as the time resolution, the accuracy, and whether or not local weather data are needed is not well known. There have been questions about whether stack data would be useful in a practical way at all, depending on the type of data made available and when it could be made available from a producer. To date, only one published study (Schöppner et al., 2013) has addressed the impacts the time resolution of stack monitoring data have on predicted concentrations at an IMS station location. The minimum source term resolution considered in that study was one day. Atmospheric modeling studies using inert tracers have been conducted since the early 1980s (Ferber et al., 1986; Gudiksen et al., 1984). This study addresses the difficult nuance of whether atmospheric models currently in wide use can yield information on the accuracy and timing of the source term data needed to faithfully reproduce sampling data.

This paper describes a challenge exercise formulated to start to answer some of these questions. Namely, to ascertain the level of agreement that can be achieved between atmospheric transport models using stack monitoring data and xenon isotopic concentration measurements at IMS stations. An evaluation criterion is used to measure the level of agreement. However, the real value of the exercise is in discussions resulting from the challenge without over-analyzing the evaluation criterion. The challenge is expected to spark discussions on what techniques are best, what gaps exist in our knowledge, and what type of data fidelity is needed from stack monitors. In general, this challenge will help inform the international treaty verification community of the status of the current capability.

The general approach of the exercise was to challenge atmospheric transport modeling groups to reproduce the time-history of ^{133}Xe measurements at an IMS station using stack monitoring data from a medical isotope production facility. Participants received stack monitoring data that included the location, UTC date and time of releases, the measured activity concentrations of ^{133}Xe in Bq m^{-3}, an average stack flow rate (80,000 m^{3} hr^{-1}), and the height (m above ground level) of the release. All other data were gathered by the participants. Each participant used the atmospheric transport model and the associated meteorological data of their choice. The individuals participating in the challenge are identified in Table 1. Participants were asked not to use the IMS sampling data, if they had access to them, until after completing the modeling exercise.

2. Atmospheric transport models and meteorological data

The participants used several transport codes and several different sources for meteorological data. Several participants submitted results for more than one model. Some of the submissions were averages of other models or low and high resolution runs for the same model. Model metadata are provided in Table 2. Although the analysis considers all twenty six submissions, a subset of the submissions was selected to discuss common model characteristics. The reduced set of submissions is identified in the last column of Table 2. Some submissions are not specifically identified in Table 2. The submission Hof 3 was an average of the submissions Hof1 and Hof2. Submissions Sei4, Sei5 and Sei6 were slight variations, including different release height assumptions, on submissions Sei 1, Sei2, and Sei3. Ros2 was a low resolution (smaller number of particles) version of submission Ros1 and Mau1 was a low resolution version of Mau3.

The participants used five different atmospheric transport models. The models, in order of the number of uses by participants are the following: FLEXPART (Stohl et al., 2005; Stohl et al., 1998), a Lagrangian particle dispersion model; HYSPLIT (Draxler and Hess, 1998, 2010) a hybrid single particle Lagrangian integrated trajectory model; Eulerian ldX (Tombette et al., 2014) which is part of IRSN's (French Institute for Radiation protection and Nuclear Safety) C3X operational platform; the Weather Research and Forecasting (WRF) model (Done et al., 2004; Michalakes et al., 2001) and MLDP0 (D'Amours et al., 2015; D'Amours et al., 2010) a Lagrangian particle dispersion model designed for long-range problems associated with events of regional, continental and global consequences.

The participants used six different meteorological data sets, some of which are available in different spatial and time resolutions. Meteorological analysis data are created by assimilation of a forecast model to observational data. Reanalysis data (i.e. GDAS) are produced later to have a consistent standardized gridded product of past weather patterns.

Thirteen of the submissions used global analysis data from the European Centre for Medium-Range Weather Forecasts (ECMWF) (Simmons et al., 1989). The U.S. National Oceanic and Atmospheric Administration's (NOAA) National Weather Service's National Centers for Environmental Prediction (NCEP) (Environmental Modeling Center, 2003) produces operational forecasts and a series of computer analyses. NCEP's Global Forecast System (GFS) produces pressure level data that can be used in FLEXPART (NCEP tag in Table 2). It also produces the GDAS (Global Data Assimilation System) reanalysis data which can be used in HYSPLIT (Kanamitsu et al., 1991). Five submissions used NCEP data and three submissions used GDAS data. Two submissions used the Weather Research and

Forecasting (WRF) model (Done et al., 2004; Michalakes et al., 2001; Skamarock et al., 2008). One participant used the global model ARPEGE (Action de Recherche Petite Echelle Grande Echelle) from the French meteorological office (Météo-France) (Déqué et al., 1994; Déqué and Piedelievre, 1995). One participant used the global meteorological analyses provided by the Canadian Meteorological Centre (CMC). CMC runs operationally a complete integrated suite of numerical weather prediction (NWP) models under an infrastructure called the Global Environmental Multiscale (GEM) system (Côté et al., 1998). The GEM system executed in a global configuration is called the GDPS: Global Deterministic Prediction System (Buehner et al., 2015; Buehner et al., 2013; Charron et al., 2012). The GDPS includes a 4D vibrational data assimilation system and is run twice a day (00 and 12 UTC) with a horizontal grid mesh defined at ~25 km (0.23° horizontal resolution). This global meteorological analyses database is used to drive MLDP0.

The spatial resolution of the meteorological grids in Table 2 is typically expressed in units of degrees. A 1° grid for meteorological data in this region of the world has a north-south spacing of approximately 111 km and an east-west spacing of 78 km. Similarly, a 0.5° grid has a spacing of 55 and 39 km, and a 0.2° grid has a spacing of about 22 and 16 km.

3. Comparison measures

The purpose of this challenge was to ascertain the level of agreement one can achieve between simulated concentrations and IMS measurements using only the stack data and an atmospheric transport model, as might be expected for situations in which there was a detection of radioxenon at an IMS station and very little other information. Concentration estimates from this modeling exercise are expected to be quite variable (Draxler et al., 2015), thus it is useful to explore the general characteristics of the models with the closest agreement with the sampled data. Researchers have proposed a number of different performance measures for comparing the outputs of atmospheric transport models. For purposes of this analysis, five statistical measures described by other researchers (Chang and Hanna, 2004; Draxler, 2006) are used.

A brief introduction of each statistical measure is provided here. Additional information is given in the Appendix. The fractional bias (FB) is a measure of the bias between measured and predicted values. The correlation coefficient R is used to represent the linear relationship between measured and predicted values. The fraction of predicted values within a factor of five of the measured value (F5) is also used. The Kolmogorov–Smirnov (KS) statistic quantifies the differences between the distribution of unpaired measured and predicted values. The normalized mean square error (NMSE) is a measure of the difference between paired measured and predicted values.

The five statistical model comparison measures implicitly assume that all of the ^{133}Xe measured at the IMS sampling station in originated from the IRE facility. Although IRE is the largest emitter of ^{133}Xe in the region, it is not the only one. Nuclear power plants emit low levels of ^{133}Xe (Kalinowski and Tuma, 2009; Saey, 2009) and a number of nuclear power plants in Europe were in operation during this time period. Another medical isotope production facility in the Netherlands (Tyco Healthcare) releases about 0.1% of the amount of ^{133}Xe (Saey, 2009) as released from IRE on an annual basis. The medical isotope production facility in Chalk River, Canada, annually releases from three to four times as much ^{133}Xe (Saey, 2009) as IRE and under suitable meteorological conditions, may produce a measurable contribution to the ^{133}Xe levels across Europe. In spite of these other sources, this is a realistic test case when data are only available from a single facility. In other words, for real world scenarios, we are testing the hypothesis that a single larger emitter may dominate the concentrations observed at an IMS facility.

Based on approaches suggested by other researchers (Chang and Hanna, 2004; Draxler, 2006), we combine four of the statistics into a single model ranking parameter as follows:

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

The model rank ranges from 0 (a model with no predictive ability) to 4 (a perfect model).

It is desirable to have contributors to an overall rank that measure different aspects of disparity. For example, a data set could have an R^2 value of 1.0 but have a large magnitude of FB. There is some concern that FB and F5 measure similar aspects of disparity. However, for this data set, other than the four submissions with the lowest F5, the values for F5 and FB do not seem to be correlated.

4. Release and detection data

Participants in the modeling challenge received ^{133}Xe stack emission data from the Institut des Radioéléments (IRE) radiopharmaceutical plant in Fleurus, Belgium. Releases from IRE have a measurable influence on ^{133}Xe concentrations collected at DEX33 (Saey et al., 2010a) which is located 376 km from the IRE stack. The emission data covered the period 10 Nov 2013 through 8 Dec 2013. The measured concentration values for the stack data are based only on the 81 keV decay energy and have an uncertainty (one sigma) of approximately 10% of the measured values. The stack air flow rate was 8×10^4 $m^3\ h^{-1}$, without any uncertainty estimate. The concentrations of ^{133}Xe in the exhaust stack air were provided for 2784 contiguous 15-min release periods. The amount released (concentration multiplied by the air flow rate) in each 15 minute period is shown in Fig. 1. Release quantities may vary by as much as two orders of magnitude for different 15-min duration periods in the same day.

The German national authority Bundesamt für Strahlenschutz (BfS) provided the ^{133}Xe activity concentration data collected at the IMS noble gas sampler at Radionuclide Station RN33 (DEX33) at mount Schauinsland, Germany for the challenge. This sampling station is located at 1205 m above sea level on a mountain in the Black Forest. Surrounding low-level terrain ranges in elevation from 200 to 600 m. The SPALAX™ system (Fontaine et al., 2004) at this station uses a sample collection period of 24 hours. The time tag for each sample is the beginning of the sample collection period and the reported concentration is an average value decay-corrected to the beginning of the sample collection period. The measured data at DEX33 and their uncertainties (one sigma) are shown in Fig. 2. The uncertainties range from 2.3% of the largest measured value to approximately 40% of the smallest values.

5. Model comparison results

Thirteen participants submitted 26 solutions containing modeled concentrations of ^{133}Xe at the sampler (DEX33) in Germany on the time periods used by the sampler. A plot of modeled concentrations for all 26 submissions and the concentrations at the sampler (black dots connected by a dotted line) is provided in Fig. 3. One submission had two predicted concentration values larger than 100 mBq m^{-3}, but the upper limit on this plot partially obscures that fact. Some of the values were zero, thus they cannot be represented on a log plot and the lines for adjacent nonzero values give the appearance of discontinuous data. However, the data were discrete values for each day and the lines on this plot are provided to aid in tracing of the time sequence of individual submissions.

The measured concentrations show five peaks separated in time and most modeled concentrations also show five peaks separated in time. There are three time periods (Nov. 17-19, Nov. 26-27 and Dec. 8-9) where most or all of the modeled concentrations are smaller than the measured concentrations. Data collected at DEX33 when IRE was not operating (Saey et al., 2010a) show that approximately 90% of the historical samples have concentrations above 0.1 mBq m^{-3}. Thus, it is reasonable to expect detectable background concentrations of ^{133}Xe at this sampler from other sources even when the wind is blowing releases from IRE in a different direction.

Although the measured concentrations are influenced by releases from IRE, the highest concentrations in the plume often bypassed the sampling station during the time period shown in Fig. 3. The sample collection period of the first sample from DEX33 used in this study starts only 6 h after the first IRE release data, but it is 15 h before the first large release. Earlier simulations suggest that releases from IRE in the previous 3 d move to the northeast and almost all of the plume bypasses the sampler. An example modeled ^{133}Xe plume using the HYSPLIT computer code and GDAS data (3 h temporal resolution, 1° spatial resolution) corresponding to the time of the sample with collection start at 0600 UTC on November 14 is shown in Fig. 4. The plume is truncated on the south in Fig. 4 to minimize the output file size. This particular model run slightly underestimates the sampler concentration for this time period but it still illustrates the sharp gradients on the edges of the main body of the plume. As a consequence, relatively small discrepancies in the direction of movement between the modeled plume and the real plumes can lead to large concentration discrepancies at sampling locations.

5.1 Statistical performance measures

The values of the individual statistics and the ranking parameter are provided in Table 3 for every submitted solution. The entries in the table are sorted by descending rank. The best values for the individual performance measure are highlighted in bold text. The mean square error (MSE) between the modeled and predicted values is also provided because it is used in the ensemble calculation in the next section.

The only difference between Mau1 and Mau3 is that Mau3 used 4×10^7 particles while Mau1 used 3×10^6 particles. The accuracy of predictions improved significantly using more particles. The submission with the largest rank (Sch) used background source estimates (average releases from other medical isotope production facilities and nuclear power plants) in addition to the releases from IRE in the calculation. This submission illustrates the effect additional sources can have on the KS statistic, because it is highly influenced by the additional sources (fewer predicted concentrations are near zero). The F5 statistic is influenced by the additional sources to a lesser extent.

5.2 Ensemble performance measures

Rather than comparing the results of individual models, one can attempt to combine them in an optimal way to provide a better prediction. A number of researchers (Kolczynski et al., 2009; Solazzo and

Galmarini, 2015) have started using ensembles of the individual models in an effort to produce better modeled concentrations. One of the justifications for using ensembles is to overcome the high sensitivity to the direction of plume movement illustrated in Fig. 4.

An ensemble reduction technique based on minimizing the mean square error between the measured and predicted concentrations is now available (Stein et al., 2015) in the HYSPLIT suite of codes. Using this approach, we calculate the average of all possible model combinations composed by increasing the number of ensemble members from 1 to 25 and estimate their MSE. The combination with the minimum MSE is then selected. In other words, we combine the 25 model outputs in 300 pairs, 2300 trios, etc., and determine which combination provides the minimum MSE. Fig. 5 shows the minimum MSE obtained as a function of the number of submissions in the reduced ensembles. The curve has a minimum at two ensemble members. In addition, the best ensembles with two, three or four members all have lower MSE than the single best model. This means that including more than about four members in the ensemble will produce a less accurate result.

The MSE of an average of several submissions used to select the ensemble members is different than the performance measures shown in Table 3. The ensemble of four members yields an average value that has KS=0.42, R=0.98, FB=-0.25, F5=0.61, Rank = 3.03, NMSE=0.81 and MSE=2.74. As a comparison, the ensemble with only two members (Hof4 and Mau3) has KS=0.42, R=0.97, FB=0.01, F5=0.58, Rank=3.10, NMSE=0.31 and MSE=1.34. The rank for the two member ensemble is better than the rank of the best submission and the rank of the four member ensemble is about equal to the rank of the best submission. The correlation (R) of the four member ensemble is higher than for the single best submission, but the fractional bias (FB) is worse. The modeled ^{133}Xe concentrations for the ensemble members and the ensemble average for the minimum MSE ensemble of four members is provided in Fig. 6. Two of the ensemble members used releases varying every 15 min while the other two used sources varying every 3 hr. These four models use four different meteorological data sets and two different computer codes, implying independence between the four ensemble members. Independence among ensemble members is a necessary but not sufficient condition for building accurate ensembles (Kioutsioukis and Galmarini, 2014).

This study, and historical sampling data from DEX33 when IRE was not operating (Saey et al., 2010a), suggests that the largest sample values are heavily dominated by releases from IRE. A comparison of measured and predicted concentrations are provided in Table 4 for the five largest sampled values for the submissions that scored the highest on individual statistical performance measures. The ensemble with four members is also included for comparison. The percentage values are the relative difference of the predicted and measured concentrations, and a negative value means the predicted value is smaller than the measured value. The Hof2 submission had a high correlation (0.97) between the sampled and measured concentrations, but also a large fractional bias. Some of the submissions predicted the largest concentrations to within 15%. The submission (Sau) did not have the best score on any specific statistical measure, but it was one of the four members of the minimum MSE ensemble and it has the smallest maximum relative error on the five largest measured concentrations.

5.3 Comparisons using grouped submissions

Ranks were calculated for several different combinations of the suite of submissions in addition to the minimum MSE ensemble approach. The ranks provided in Fig. 7 are based on the seventeen submissions identified in Table 2. Except for the single submission with the highest rank, the ranks were calculated using the average of each member of the group. The average of all the submissions has a lower rank than the average from the ensemble with four members. The rank for the group of HYSPLIT models is lower than the ranks for the FLEXPART and other models. Most of the FLEXPART models used ECMWF meteorological data while most of the HYSPLIT models used GDAS data. Thus, it is not surprising that the lower ranks using the HYSPLIT model correspond to the lower ranks for GDAS data as compared to other data sets. Although the governing equations generally are time reversible, the implementations yield slightly different concentration estimates depending on the time direction. The average of the forwards time runs had a slightly higher rank than the average of the backwards runs. The average of model runs using meteorological data with finer spatial resolution than 0.5° had higher rank than those using 0.5° resolution data. The average of model runs using 1.0° resolution meteorological data had a rank about equal to the average of finer resolution model runs, however, the normalized MSE for the 1.0° spatial resolution runs was 5.09 while that of the finer spatial resolution runs was 2.89. Those models that incorporated the source term on a 15-min timing basis had higher ranks than models using sources using longer source term aggregation periods.

5.4 Additional sources

The modeling exercise was formulated to consider the hypothesis that a single larger emitter may dominate the concentrations observed at an IMS facility. However, one submission (Sch) included annual average emission rates for nuclear power plants and other medical isotope production facilities as an additional source term. The Sch results are compared to the four member ensemble average in Fig. 8. This submission suggests that the other releases are also influencing the sampler, and this result is consistent with historical data (Saey et al., 2010a). The transport runs done for submission Hof4 yielded effective atmospheric dilution factors that indicate releases from the medical isotope production facility in Chalk River, Canada, could potentially influence 18 of the 30 DEX33 samples. No Chalk River source was introduced in the Hof4 submittal even though releases from the facility seem to have influenced some of the measured data at DEX33.

6. Discussion

The ranking and ensemble analysis in this paper suggests that combining multiple models may provide more accurate predicted concentrations than almost any single model. One ensemble selection technique was used in this paper. Further research is needed to identify optimal methods for selecting ensemble members, and those methods may depend on the nature of the transport problem. Although this exercise only addressed release and transport of a nondepositing noble gas, other radionuclides of interest to the treaty monitoring community (such as ^{137}Cs and ^{131}I) deposit on the ground during transport, and models that work best for predicting air concentrations may not fare as well when predicting deposition on the ground (Draxler et al., 2015).

Participants in this challenge predicted measured concentrations at a sampling station using only releases from one medical isotope production facility. Some of the models predicted the highest measured concentrations quite well (high rank or small MSE); however none predicted the small measured (background) concentrations very well. The one submission that included average release estimates from other nuclear facilities matched the small concentrations much better. If expected releases from future nuclear tests are small, such as releases from the 2013 test by the Democratic People's Republic of Korea (Ringbom et al., 2014), then modeling of sources from nuclear facilities with smaller releases than medical isotope production facilities may also be important.

The grouped model comparisons shown in Fig. 7 categorize prediction performance relative to several of the choices available to modelers. For this exercise, the ranks for submissions using FLEXPART were higher than the ranks for submissions using FLEXPART. However, most HYSPLIT runs used GDAS data while FLEXPART used other meteorological data. Interpretation of the results must recognize that most of the categories are confounded with each other. For example, all of the HYSPLIT model runs in comparisons in Fig. 7 did runs that were forwards in time. In addition, the sampler at DEX33 used a collection interval of 24 h, and 24 h may be long enough to average out some of the differences in the time resolution of the source term. The release data from IRE were provided with a time resolution of 15 min. Two of the models in the four member minimum MSE ensemble used 15 min release data, but the other two aggregated releases to a 3 h basis. The average predicted concentrations for the models that incorporated the source term on a 15-min timing basis had a higher rank than models using longer release periods. However, models using 3 h source averaging had a higher rank than those using 1 h averaging.

Other operational radioxenon samplers in the IMS use a shorter sample collection interval of 12 h (Prelovskii et al., 2007; Ringbom et al., 2003) and new generation radioxenon samplers under development (Hayes et al., 2013; Le Petit et al., 2015) can use collection periods of 6 or 8 h. These shorter collection periods may show more sensitivity to the time resolution of a highly time-variable source term than the current sampler.

Finally, the results of this single exercise indicate that the use of stack monitoring data to determine radionuclide concentrations at a distance of nearly 400 km can yield predicted large concentrations within $\pm$40% of the measured concentrations if an ensemble is used. Individual models have a larger spread than the ensemble results. The uncertainties in the stack data do not appear to dominate the uncertainties in the modeled results. However, the uncertainty in the air flow rate in the stack is not known, so the uncertainty in the release values may be significantly larger than the 10% uncertainty in the isotope concentration data in the stack. More work will be needed to determine the achievable accuracy in other conditions, such as for larger source-receptor distances. We anticipate more exercises of this nature could help to define methods to understand the effect of emissions from fission-based medical isotope production on IMS sampling data.

Acknowledgments

Participants in the atmospheric transport modeling challenge received ^{133}Xe emission data from the Institut des Radioéléments (IRE) radiopharmaceutical plant in Fleurus, Belgium. IRE granted permission to use the data for the challenge.

The German national authority *Bundesamt für Strahlenschutz* (BfS) granted permission to use the ^{133}Xe concentration data collected at the IMS noble gas sampler (DEX33) in Schauinsland, Germany for the challenge. Clemens Schlosser and Verena Heidmann of Bfs manually analyzed the spectra to obtain the ^{133}Xe concentration data and associated error bars.

Some of the authors wish to acknowledge the funding support of the U.S. Department of State and the U.S. Defense Threat Reduction Agency.

Appendix

In the following descriptions, let P denote predicted concentrations, M denote measured concentrations, an overbar denote an average over the data set, and i denote an index of the N sample values. The fractional bias (FB) is measure of the bias between measured and predicted values. The FB is normalized to the range -2 to 2 and positive values indicate predictions are larger than measured values. Small concentrations attributable to releases from facilities other than IRE have a small effect on this performance measure. The fractional bias is defined as:

$$FB = 2\frac{(\overline{P}-\overline{M})}{(\overline{P}+\overline{M})} \qquad (1)$$

The correlation coefficient R is used to represent the linear relationship between measured and predicted values where the summation is taken over all samples. Possible values for R range from -1 to 1. The correlation coefficient is calculated from:

$$R = \frac{\Sigma(M_i-\overline{M})(P_i-\overline{P})}{\sqrt{\Sigma(M_i-\overline{M})^2(P_i-\overline{P})^2}} \qquad (2)$$

The fraction of predicted values within a certain factor of the measured value is often used in model comparisons. This statistic can be heavily influenced if some modeled values are near zero while nuisance sources cause the measured values to be at or just above a detection limit. We define the factor of five (F5) statistic as the fraction of sample values that satisfy:

$$0.2 \leq \frac{P_i}{M_i} \leq 5.0 \qquad (3)$$

The Kolmogorov–Smirnov (KS) statistic (Stephens, 1970) quantifies the differences between the distribution of unpaired measured and predicted values. The values are considered as samples from two different statistical distributions and KS is defined as the maximum difference between two cumulative distributions when $M_k=P_k$, where

$$KS = Max|D(M_k) - D(P_k)|. \tag{4}$$

In this case, D is the cumulative distribution of the measured and predicted concentrations over the range of k values such that D is the probability that the concentration will not exceed M_k or P_k. It measures the ability of the model to reproduce the measured concentration distribution regardless of when or where it occurred. The maximum difference between any two distributions cannot be more than 100%. This statistic can be heavily influenced if some modeled values are near zero while nuisance sources cause the measured values to be at or just above a detection limit.

The normalized mean square error (NMSE) is a measure of the difference between paired measured and predicted values. The normalized mean square error is calculated from:

$$NMSE = \frac{MSE}{\overline{M}\,\overline{P}} \tag{5}$$

where MSE is the mean square error defined as:

$$MSE = \frac{1}{N}\sum(M_i - P_i)^2 \tag{6}$$

References

Becker, A., Wotawa, G., Ringbom, A., Saey, P.R.J., 2010. Backtracking of Noble Gas Measurements Taken in the Aftermath of the Announced October 2006 Event in North Korea by Means of PTS Methods in Nuclear Source Estimation and Reconstruction. Pure Appl. Geophys. 167(4), 581-599. doi:10.1007/s00024-009-0025-0

Bowyer, T.W., Kephart, R., Eslinger, P.W., Friese, J.I., Miley, H.S., Saey, P.R.J., 2013. Maximum reasonable radioxenon releases from medical isotope production facilities and their effect on monitoring nuclear explosions. J. Environ. Radioact. 115(1), 192-200. doi:10.1016/j.jenvrad.2012.07.018

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. Monthly Weather Review 143(7), 2532-2559. doi: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. Nonlin. Processes Geophys. 20(5), 669-682. doi:10.5194/npg-20-669-2013

Chang, J.C., Hanna, S.R., 2004. Air quality model performance evaluation. Meteorology and Atmospheric Physics 87(1-3), 167-196. doi:10.1007/s00703-003-0070-7

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. Monthly Weather Review 140(6), 1924-1944. doi:10.1175/MWR-D-11-00097.1

Comprehensive Nuclear-Test-Ban Treaty. 1996. *Text of the Comprehensive Nuclear-Test-Ban Treaty.* United Nations Office for Disarmament Affairs (UNODA), Status of Multilateral Arms Regulation and Disarmament Agreements, CTBT. Accessed on September 20, 2012, at http://www.ctbto.org/the-treaty/treaty-text/

Côté, J., Desmarais, J.-G., Gravel, S., Méthot, A., Patoine, A., Roch, M., Staniforth, A., 1998. The Operational CMC–MRB Global Environmental Multiscale (GEM) Model. Part II: Results. Monthly Weather Review 126(6), 1397-1418. doi:10.1175/1520-0493(1998)126<1397:TOCMGE>2.0.CO;2

CTBTO. 2014. *Verification Regime*. Accessed on October 13, 2014, at http://www.ctbto.org/verification-regime/monitoring-technologies-how-they-work/radionuclide-monitoring/page-5/

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. Atmosphere-Ocean 53(2), 176-199. doi: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. Journal of Geophysical Research-Atmospheres 115(D00L11), 1-11. doi:10.1029/2009JD013602

Déqué, M., Dreveton, C., Braun, A., Cariolle, D., 1994. The ARPEGE/IFS atmosphere model: a contribution to the French community climate modelling. Climate Dynamics 10(4-5), 249-266. doi:10.1007/BF00208992

Déqué, M., Piedelievre, J.P., 1995. High resolution climate simulation over Europe. Climate Dynamics 11(6), 321-339. doi:10.1007/BF00215735

Done, J., Davis, C.A., Weisman, M., 2004. The next generation of NWP: explicit forecasts of convection using the weather research and forecasting (WRF) model. Atmospheric Science Letters 5(6), 110-117. doi:10.1002/asl.72

Draxler, R., Arnold, D., Chino, M., Galmarini, S., Hort, M., Jones, A., Leadbetter, S., Malo, A., Maurer, C., Rolph, G., Saito, K., Servranckx, R., Shimbori, T., Solazzo, E., Wotawa, G., 2015. World Meteorological Organization's model simulations of the radionuclide dispersion and deposition from the Fukushima Daiichi nuclear power plant accident. J. Environ. Radioact. 139(0), 172-184. doi:10.1016/j.jenvrad.2013.09.014

Draxler, R.R., 2006. The Use of Global and Mesoscale Meteorological Model Data to Predict the Transport and Dispersion of Tracer Plumes over Washington, D.C. Weather Forecast 21(3), 383-394. doi:10.1175/WAF926.1

Draxler, R.R., Hess, G.D., 1998. An overview of the HYSPLIT_4 modeling system of trajectories, dispersion, and deposition. Aust. Meteorol. Mag. 47, 295-308

Draxler, R.R., Hess, G.D., 2010. Description of the HYSPLIT_4 Modeling System, ARL-224, Air Resources Laboratory, National Oceanic and Atmospheric Administration (NOAA), Silver Springs, Maryland

Eslinger, P.W., Friese, J.I., Lowrey, J.D., McIntyre, J.I., Miley, H.S., Schrom, B.T., 2014. Estimates of radioxenon released from Southern Hemisphere medical isotope production facilities using measured air concentrations and atmospheric transport modeling. J. Environ. Radioact. 135(2014), 94-99. doi:10.1016/j.jenvrad.2014.04.006

Ferber, G.J., Heffter, J.L., Draxler, R.R., Lagomarsino, R.J., Thomas, F.L., Deitz, R.N., Benkovitz, C.M., 1986. Cross-Appalachian Tracer Experiment (CAPTEX -83) final report., NOAATech. Memo. ERL ARL-142, NOAA/Air Resources Laboratory, Silver Spring, Maryland

Fontaine, J.P., Pointurier, F., Blanchard, X., Taffary, T., 2004. Atmospheric xenon radioactive isotope monitoring. J. Environ. Radioact. 72(1–2), 129-135. doi:10.1016/S0265-931X(03)00194-2

Gudiksen, P.H., Ferber, G.J., Fowler, M.M., Eberhard, W.L., Fosberg, M.A., Knuth, W.R., 1984. Field studies of transport and dispersion of atmospheric tracers in nocturnal drainage flows. Atmospheric Environment (1967) 18(4), 713-731. doi:http://dx.doi.org/10.1016/0004-6981(84)90257-9

Hayes, J.C., Ely, J.H., Haas, D.A., Harper, W.W., Heimbigner, T.R., Hubbard, C.W., Humble, P.H., Madison, J.C., Morris, S.J., Panisko, M.E., Ripplinger, M.D., Stewart, T.L., 2013. Requirements for Xenon International, PNNL-22227 Rev.1, Pacific Northwest National Laboratory, Richland, Washington. doi:10.2172/1122330

Hoffman, I., Ungar, K., Bean, M., Yi, J., Servranckx, R., Zaganescu, C., Ek, N., Blanchard, X., Le Petit, G., Brachet, G., Achim, P., Taffary, T., 2009. Changes in Radioxenon Observations in Canada and Europe during Medical Isotope Production Facility Shut Down in 2008. Journal of Radioanalytical and Nuclear Chemestry 282, 767-772. doi:10.1007/s10967-009-0235-z

Kalinowski, M.B., Becker, A., Saey, P.R.J., Tuma, M.P., Wotawa, G., 2008. The Complexity of CTBT Verification. Taking Noble Gas Monitoring as an Example. Complexity 14(1), 89-99. doi:10.1002/cplx.20228

Kalinowski, M.B., Tuma, M.P., 2009. Global radioxenon emission inventory based on nuclear power reactor reports. J. Environ. Radioact. 100(1), 58-70. doi:10.1016/j.jenvrad.2008.10.015

Kanamitsu, M., Alpert, J.C., Campana, K.A., Caplan, P.M., Deaven, D.G., Iredell, M., Katz, B., Pan, H.L., Sela, J., White, G.H., 1991. Recent Changes Implemented into the Global Forecast System at NMC. Weather Forecast 6(3), 425-435. doi:10.1175/1520-0434(1991)006<0425:RCIITG>2.0.CO;2

Kioutsioukis, I., Galmarini, S., 2014. De praeceptis ferendis: good practice in multi-model ensembles. Atmos. Chem. Phys. 14(21), 11791-11815. doi:10.5194/acp-14-11791-2014

Kolczynski, W.C., Stauffer, D.R., Haupt, S.E., Deng, A., 2009. Ensemble Variance Calibration for Representing Meteorological Uncertainty for Atmospheric Transport and Dispersion Modeling. Journal of Applied Meteorology and Climatology 48(10), 2001-2021. doi:10.1175/2009JAMC2059.1

Le Petit, G., Cagniant, A., Gross, P., Douysset, G., Topin, S., Fontaine, J.P., Taffary, T., Moulin, C., 2015. Spalax™ new generation: A sensitive and selective noble gas system for nuclear explosion monitoring. Appl. Radiat. Isot. 103(0), 102-114. doi:10.1016/j.apradiso.2015.05.019

Michalakes, J., Chen, S., Dudhia, J., Hart, L., Klemp, J., Middlecoff, J., Skamarock, W., 2001. Development of a next generation regional weather research and forecast model, in: Zwieflhofer, W., Kreitz, N. (Eds.), Developments in Teracomputing: Proceedings of the Ninth ECMWF Workshop on the Use of High Performance Computing in Meteorology. World Scientific Publishing, Singapore, pp. 269–276.

Peykov, P., Cameron, R., 2014. Medical Isotope Supply in the Future: Production Capacity and Demand Forecast for the ^{99}Mo/^{99m}Tc Market, 2015-2020, NEA/SEN/HLGMR(2014)2, Organisation for Economic Co-Operation and Development, Nuclear Energy Agency, Issy-les-Moulineaux, France

Prelovskii, V.V., Kazarinov, N.M., Donets, A.Y., Popov, V.Y., Popov, I.Y., Skirda, N.V., 2007. The ARIX-03F mobile semiautomatic facility for measuring low concentrations of radioactive xenon isotopes in air and subsoil gas. Instrum Exp Tech 50(3), 393-397. doi:10.1134/S0020441207030165

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(0), 47-63. doi:10.1016/j.jenvrad.2013.10.027

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 Meth A 508(3), 542-553. doi:10.1016/s0168-9002(03)01657-7

Saey, P.R.J., 2009. The influence of radiopharmaceutical isotope production on the global radioxenon background. J. Environ. Radioact. 100(5), 396-406. doi:10.1016/j.jenvrad.2009.01.004

Saey, P.R.J., Auer, M., Becker, A., Hoffmann, E., Nikkinen, M., Ringbom, A., Tinker, R., Schlosser, C., Sonck, M., 2010a. The influence on the radioxenon background during the temporary suspension of operations of three major medical isotope production facilities in the Northern Hemisphere and during the start-up of another facility in the Southern Hemisphere. J. Environ. Radioact. 101(9), 730-738. doi:10.1016/j.jenvrad.2010.04.016

Saey, P.R.J., Schlosser, C., Achim, P., Auer, M., Axelsson, A., Becker, A., Blanchard, X., Brachet, G., Cella, L., De Geer, L.-E., Kalinowski, M.B., Le Petit, G., Peterson, J., Popov, V., Popov, Y., et al., 2010b. Environmental Radioxenon Levels in Europe: a Comprehensive Overview. Journal of Pure and Applied Geophysics 167(4-5), 499-515. doi:10.1007/s00024-009-0034-z

Schöppner, M., Plastino, W., Hermanspahn, N., Hoffmann, E., Kalinowski, M., Orr, B., Tinker, R., 2013. Atmospheric transport modelling of time resolved ^{133}Xe emissions from the isotope production facility ANSTO, Australia. J. Environ. Radioact. 126(2013), 1-7. doi:10.1016/j.jenvrad.2013.07.003

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. Meteorology and Atmospheric Physics 40(1-3), 28-60. doi:10.1007/BF01027467

Skamarock, W., Klemp, J.B., Dudhia, J., Gill, D.O., Barker, D., Duda, M.G., Huang, X., Wang, W., 2008. A Description of the Advanced Research WRF Version 3, NCAR/TN-475+STR, National Center for Atmospheric Research, Boulder, Colorado, USA. doi:10.5065/D68S4MVH

Solazzo, E., Galmarini, S., 2015. The Fukushima-137Cs deposition case study: properties of the multi-model ensemble. J. Environ. Radioact. 139, 226-233. doi:10.1016/j.jenvrad.2014.02.017

Stein, A.F., Ngan, F., Draxler, R.R., Chai, T., 2015. Potential Use of Transport and Dispersion Model Ensembles for Forecasting Applications. Weather Forecast 30(3), 639-655. doi:10.1175/WAF-D-14-00153.1

Stephens, M.A., 1970. Use of the Kolmogorov-Smirnov, Cramér-Von Mises and related statistics without extensive tables. Journal of the Royal Statistical Society. Series B (Methodological), 115-122

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(9), 2461-2474. doi: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(24), 4245-4264. doi:10.1016/s1352-2310(98)00184-8

Tombette, M., Quentric, E., Quelo, 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, Poster presented at Fourth European IRPA Congress, 23-27 June 2014, Geneva.

Wotawa, G., Becker, A., Kalinowski, M., Saey, P., Tuma, M., Zähringer, M., 2010. Computation and Analysis of the Global Distribution of the Radioxenon Isotope ^{133}Xe based on Emissions from Nuclear Power Plants and Radioisotope Production Facilities and its Relevance for the Verification of the Nuclear-Test-Ban Treaty. Pure Appl. Geophys. 167(4-5), 541-557. doi:10.1007/s00024-009-0033-0

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., Sloan, C., Yamazawa, H., 2003. Atmospheric transport modelling in support of CTBT verification—overview and basic concepts. Atmos. Environ. 37(18), 2529-2537. doi:10.1016/s1352-2310(03)00154-7

Zähringer, M., Becker, A., Nikkinen, M., Saey, P., Wotawa, G., 2009. CTBT radioxenon monitoring for verification: today's challenges. J Radioanal Nucl Chem 282(3), 737-742. doi:10.1007/s10967-009-0207-3

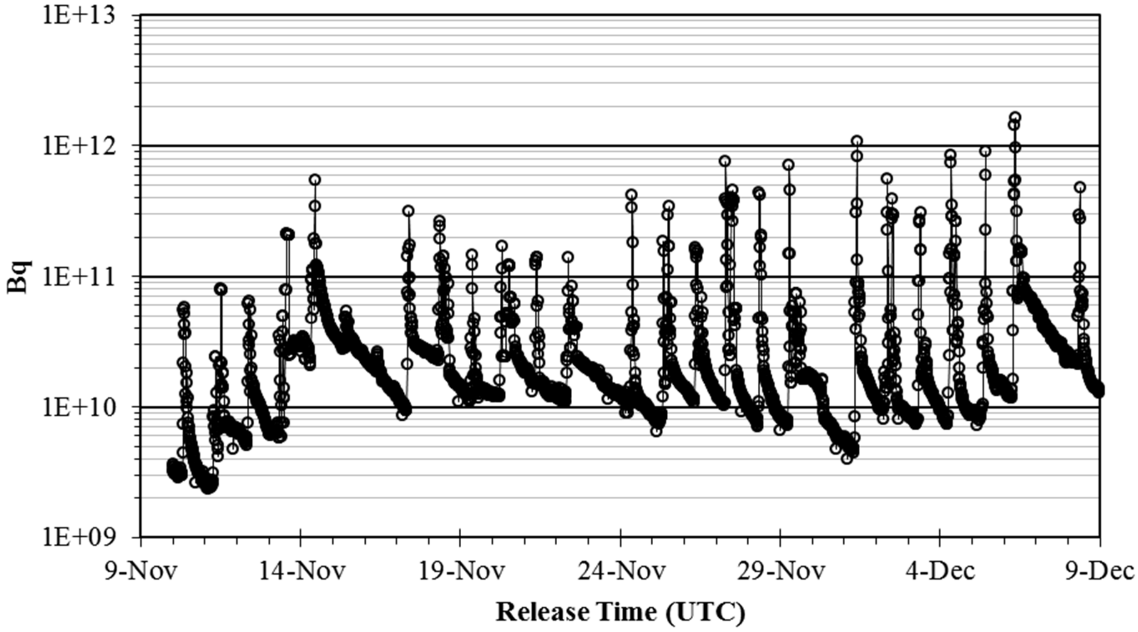

Fig. 1. Releases of ^{133}Xe (Bq) in contiguous 15 minute intervals from the exhaust stack at the Institut des Radioéléments (IRE) radiopharmaceutical plant in Fleurus, Belgium.

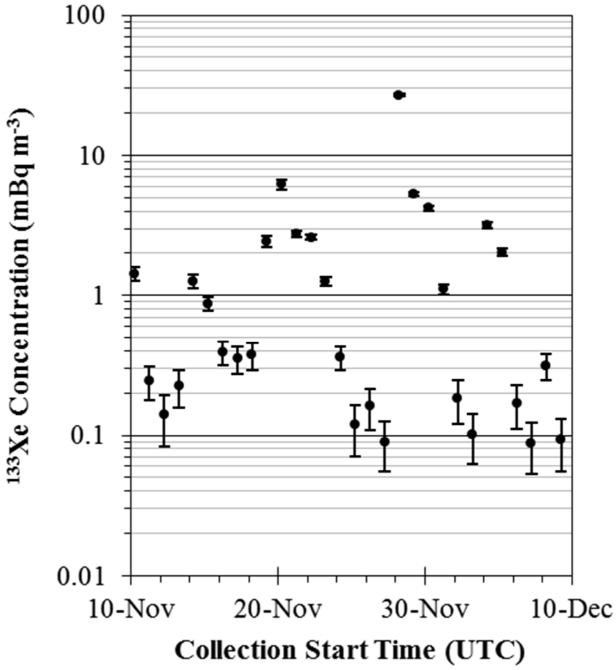

Fig. 2. Measured ^{133}Xe activity concentrations at DEX33. The error bars represent one sigma uncertainties.

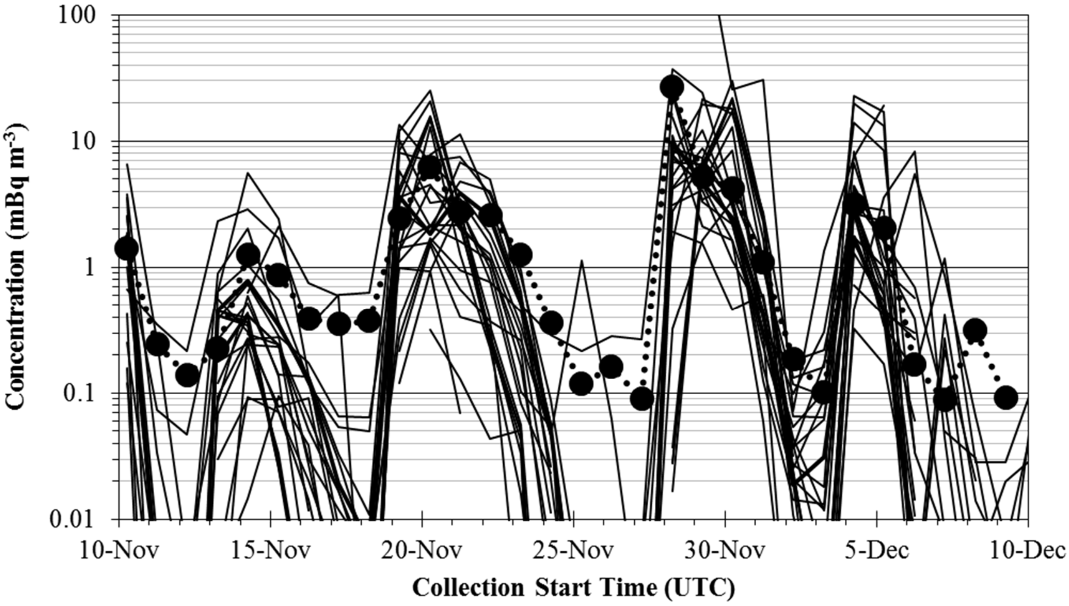

Fig. 3. Modeled ^{133}Xe concentrations for all submissions (solid lines) and measured concentrations at the sampler (large black dots connected by dotted lines).

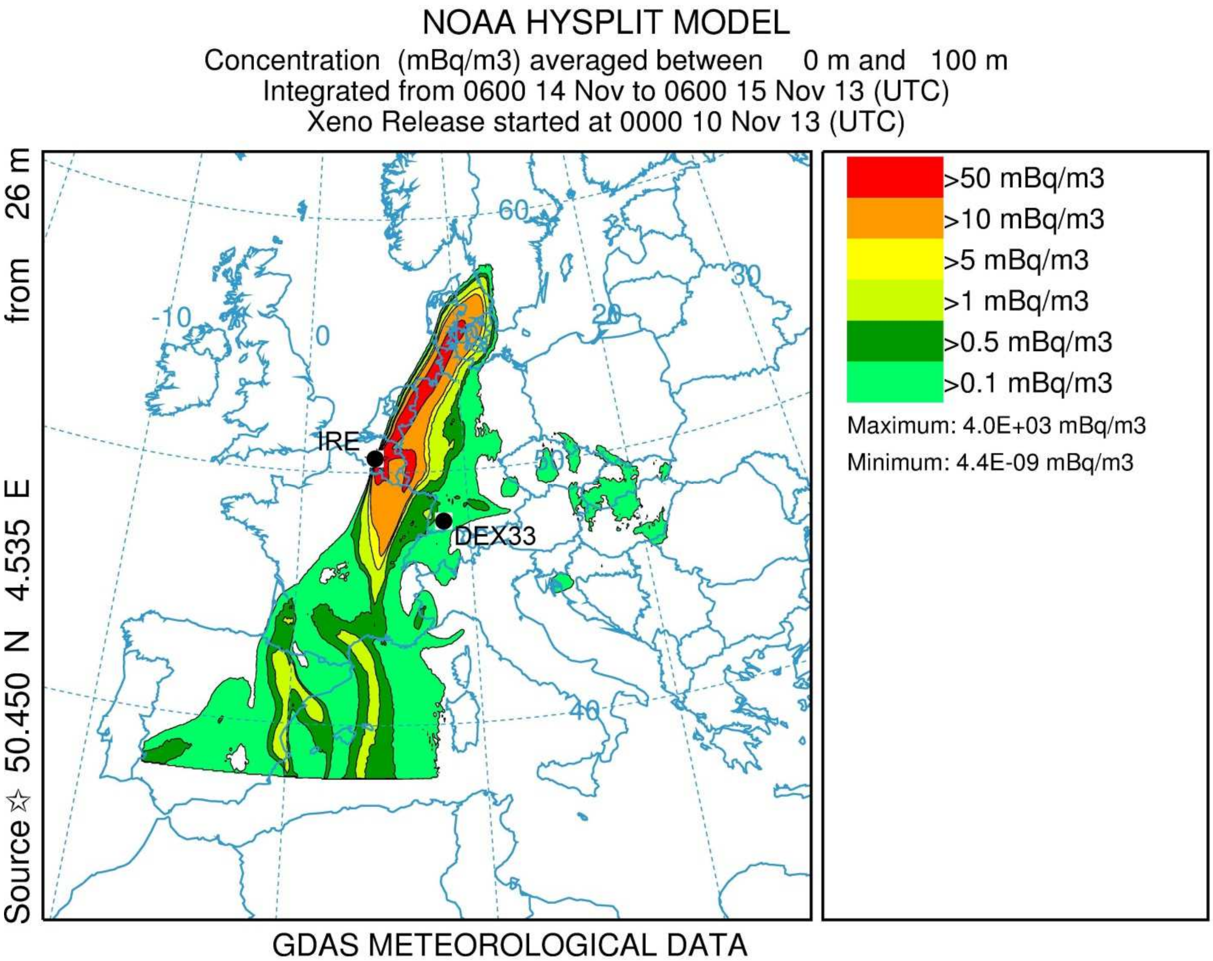

Fig. 4. Modeled ^{133}Xe concentrations using the HYSPLIT computer code and GDAS data corresponding to the DEX33 sample with collection start at 0600 UTC on November 14.

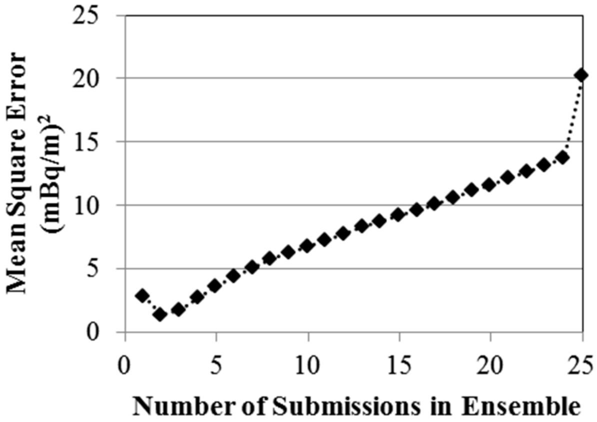

Fig. 5. Minimum MSE as a function of the number of submissions in the ensemble.

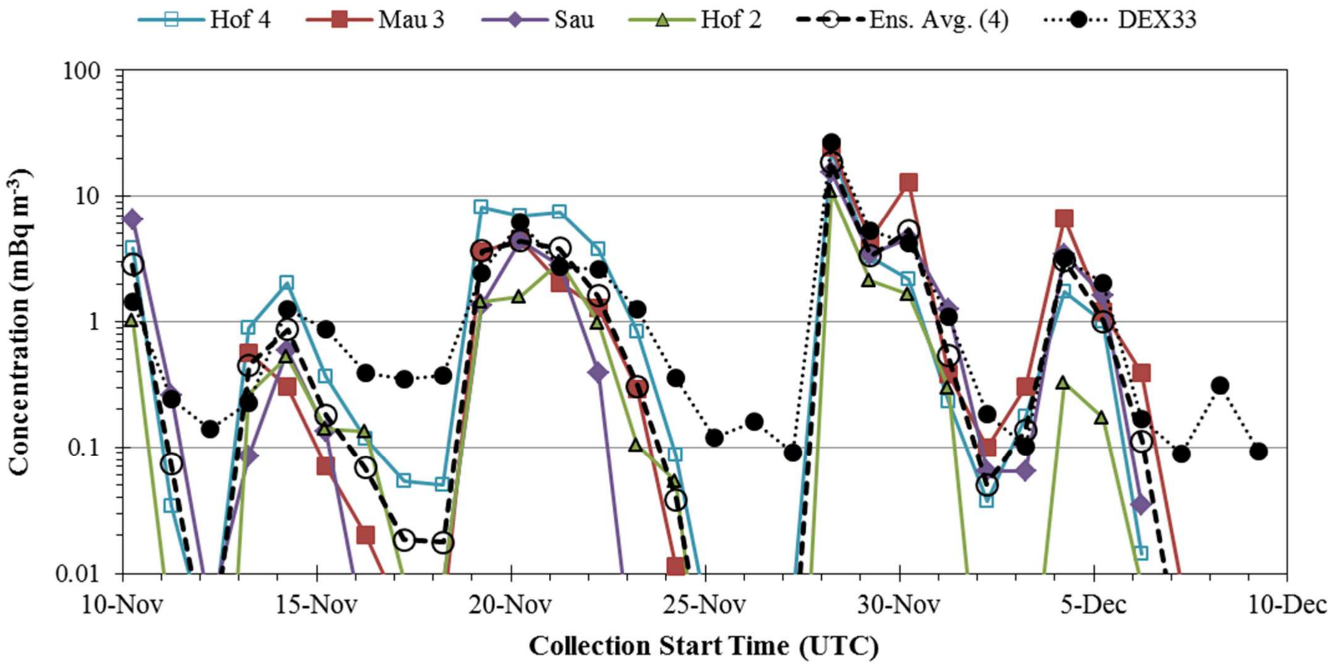

Fig. 6. Modeled ^{133}Xe concentrations for the individual submissions and the ensemble average for the minimum MSE ensemble of four members.

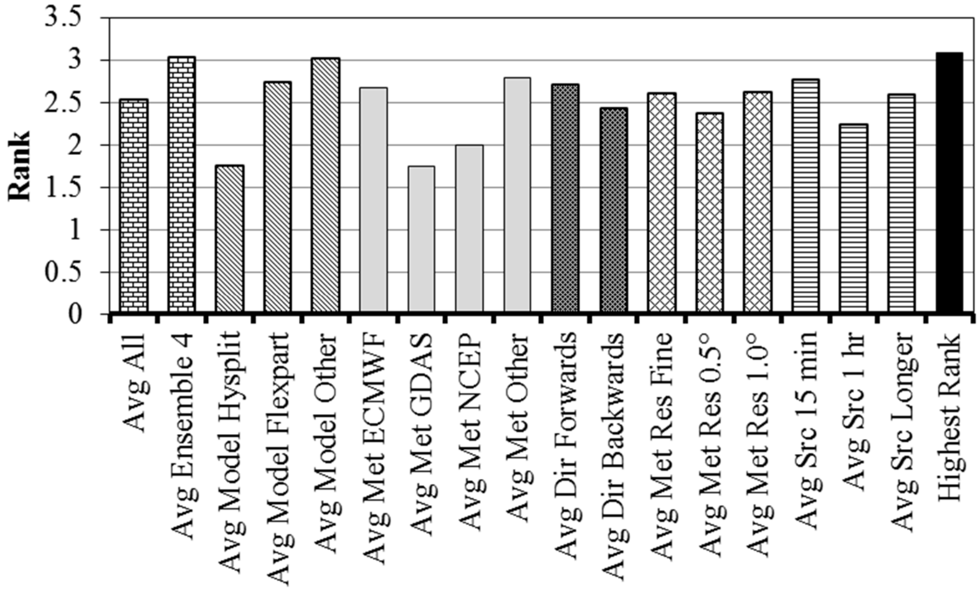

Fig. 7. Rank parameters for grouped model comparisons.

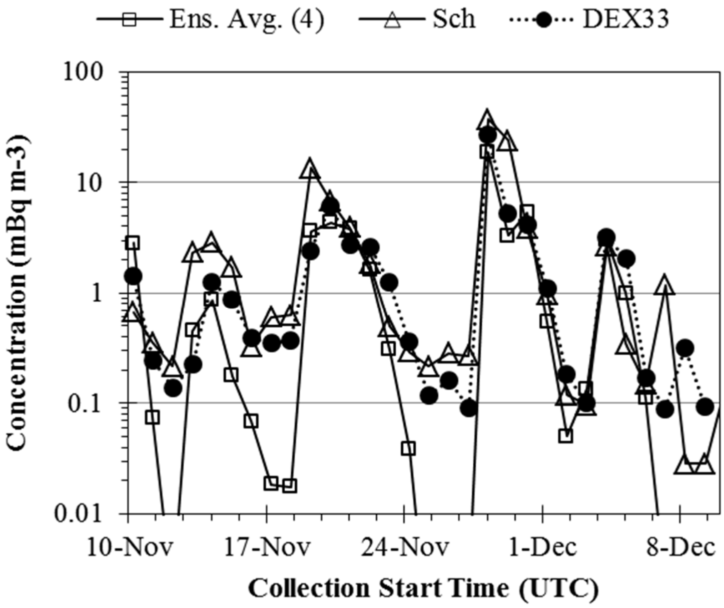

Fig. 8. Modeled ^{133}Xe concentrations for the average of the minimum MSE ensemble of four members and a submission (Sch) that includes emissions from nuclear power plants.

Table 1

Participants in the challenge exercise

ID	Name	Organization
Cha	Tianfeng Chai Fong Ngan Ariel Stein Roland Draxler	National Oceanic and Atmospheric Administration (NOAA) Air Resources Laboratory, College Park, Maryland, USA
Esl	Paul W. Eslinger Ted Bowyer Brian Schrom	Pacific Northwest National Laboratory, Richland, Washington, USA
Gen	Pascal Achim Sylvia Generoso	Commissariat à l'Energie Atomique, CEA, DAM, DIF, 91297 Arpajon, France
Hay	Philip Hayes	Air Force Technical Applications Center, Patrick Air Force Base, Florida, USA
Hof	Ian Hoffman Jing Yi Kurt Ungar	Health Canada, Radiation Protection Bureau, Ottawa, Canada
	Alain Malo	Environment Canada, Canadian Meteorological Centre, Dorval, Canada
Kij	Yuichi Kijima	Japan Atomic Energy Agency, Tokai, Ibaraki, Japan
Kry	Monika Krysta	Comprehensive Test Ban Treaty Organization (CTBTO), International Data Center, Vienna, Austria
Mau	Christian Maurer	Zentralanstalt für Meteorologie und Geodynamik, Vienna, Austria
Rob	Peter Robins Verena Heidmann	Atomic Weapons Establishment (AWE), Aldermaston, Reading, RG7 4PR, United Kingdom
Ros	Jens Ole Ross	Federal Institute for Geosciences and Natural Resources (BGR), Hannover, Germany
Sau	Olivier Saunier	French Institute for Radiation protection and Nuclear Safety, Fontenay-aux-Roses, France
Sch	Michael Schoeppner	Program on Science and Global Security, Princeton University, Princeton, New Jersey USA
Sei	Petra Seibert	University of Natural Resources and Life Sciences, Institute of Meteorology and University of Vienna, Faculty of Earth Sciences, Vienna, Austria

Table 2
Metadata for models used to explore the effects of common characteristics (see text for definitions of the acronyms)

ID	Code	Met. Data Source	Met. Time Resolution (h)	Met. Spatial Resolution (°)	Model Time Direction	Release Length (h)	Include
Cha	HYSPLIT	WRF	1	27/9 km	Forwards	0.25	Yes
Esl	HYSPLIT	NCEP (GDAS)	3	0.5	Forwards	1	Yes
Gen	FLEXPART	NCEP	6	0.5	Forwards	2	Yes
Hay[a]	WRF HYSPLIT	WRF	Ensemble	18/6/2 km	Forwards	0.25	Yes
Hof 1	FLEXPART	ECMWF	3	1	Backwards	3	Yes
Hof 2	FLEXPART	NCEP	3	1	Backwards	3	Yes
Hof 4	MLDP0	CMC	6	0.5	Backwards	3	Yes
Kij	HYSPLIT	NCEP (GDAS)	3	0.5	Forwards	6	Yes
Kry 1	FLEXPART	ECMWF	3	1.0	Backwards	3	Yes
Kry 2	FLEXPART	NCEP	6	1.0	Backwards	6	Yes
Mau 2	FLEXPART	ECMWF	3	0.2	Forwards	0.25	Yes
Mau 3	FLEXPART	NCEP	3	0.5	Forwards	0.25	Yes
Rob	FLEXPART	ECMWF	3	1.0	Backwards	0.25	Yes
Ros 1	HYSPLIT	ECMWF	6	0.2	Forwards	0.25	Yes
Ros 3	HYSPLIT	NCEP (GDAS)	3	0.5	Forwards	0.25	Yes
Sau	Eulerian ldX	ARPEGE	1	0.1	Forwards	0.25	Yes
Sch	FLEXPART	NCEP	1	0.5	Backwards	3	No
Sei 1	FLEXPART	ECMWF	3[b]	0.2	Backwards	1.25[c]	Yes
Sei 2	FLEXPART	ECMWF	3	0.2	Backwards	1.25[c]	No
Sei 3	FLEXPART	ECMWF	1	0.125	Backwards	1.25[c]	No

a. This submission was the mean of an 85 member ensemble
b. Forecasts up to 23 hours are used
c. Five-sample moving average in time

Table 3
Values of the individual statistics and the model rank parameter (Rank) for every model submission. Statistics include the Kolmogorov-Smirnov parameter (KS), Pearson correlation (R), fractional bias (FB), factor of five parameter (F5), normalized mean square error (NMSE) and the mean square error (MSE). Bold values indicate the best score on each statistic.

Model	KS	R	FB	F5	Rank	NMSE	MSE
Sch[a]	0.10	0.89	0.50	0.81	3.25	2.63	19.2
Hof 4	0.39	0.94	0.03	0.61	**3.09**	**0.63**	18.3
Mau 3	0.45	0.93	**-0.02**	0.52	2.92	0.81	**3.50**
Sau	0.52	0.92	-0.33	0.52	2.68	1.77	5.60
Hof 3	0.45	0.90	-0.58	0.55	2.62	4.25	36.5
Hof 1	0.45	0.75	-0.32	0.58	2.53	3.79	25.9
Hof 2	0.45	**0.97**	-0.89	0.39	2.43	5.87	25.0
Rob	**0.29**	0.35	-0.19	**0.68**	2.41	5.72	20.8
Ros 2	0.52	0.81	-0.56	0.39	2.24	4.87	11.9
Mau 1	0.58	0.79	-0.36	0.35	2.22	3.24	9.90
Ros 1	0.52	0.73	-0.56	0.45	2.18	5.42	13.3
Kry 1	0.42	0.47	-0.42	0.58	2.17	6.41	16.2
Sei 1	0.52	0.46	0.13	0.45	2.08	5.45	25.0
Gen	0.39	0.23	0.36	0.58	2.06	6.56	20.5
Esl	0.45	0.30	-0.08	0.35	1.95	7.62	41.4
Sei 2	0.55	0.43	-0.07	0.35	1.95	6.14	37.5
Kry 2	0.52	0.61	-0.67	0.35	1.87	7.40	27.3
Kij	0.45	0.17	-0.13	0.35	1.87	9.80	40.0
Sei 3	0.58	0.20	-0.03	0.35	1.80	8.89	36.6
Sei 7	0.55	0.19	-0.10	0.35	1.79	9.27	35.7
Sei 8	0.55	0.19	-0.13	0.35	1.78	9.29	59.7
Sei 9	0.58	0.19	0.28	0.32	1.64	10.3	25.5
Hay	0.65	0.71	-1.41	0.16	1.31	26.9	25.3
Cha	0.71	0.83	-1.69	0.06	1.20	62.7	23.2
Mau 2	0.58	0.59	1.75	0.23	1.12	192.	12400
Ros 3	0.55	0.18	-1.17	0.23	1.12	21.5	24.5
Average[b]	0.42	0.69	0.27	0.61	2.53	3.52	19.6

[a] This submission used other sources in addition to the releases from IRE. The statistical performance measures for this submission should not be compared directly with those of other submissions.

[b] The Average row is calculated by averaging all of the modeled values for each sample period and treating the averaged values as atmospheric transport model output.

Table 4
Comparison of measured and predicted concentrations (mBq m^{-3}) for the five samples with the highest concentrations and the five submissions with highest values of the individual statistics. Statistics include the Pearson correlation (R), model rank (Rank), Kolmogorov-Smirnov parameter (KS) and fractional bias (FB). The Sau submission was a member of the best ensemble with four members

DEX33	Hof 2 (R)	Hof 4 (Rank)	Rob (KS)	Mau 3 (FB)	Sau (Ensemble)	Best 4 Ensemble
6.19	1.58 (-75%)	6.91 (12%)	3.26 (-47%)	4.56 (-26%)	4.38 (-29%)	4.36 (-30%)
26.8	11.1 (-59%)	23.4 (-13%)	4.18 (-84%)	24.5 (-9%)	15.4 (-42%)	18.6 (-31%)
5.28	2.11 (-60%)	3.29 (-38%)	6.21 (18%)	4.48 (-15%)	3.43 (-35%)	3.33 (-37%)

4.18	1.65 (-61%)	2.20 (-47%)	2.34 (-44%)	12.9 (208%)	4.56 (9%)	5.33 (27%)
3.17	0.32 (-90%)	1.75 (-45%)	2.82 (-11%)	6.65 (110%)	3.44 (9%)	3.04 (-4%)