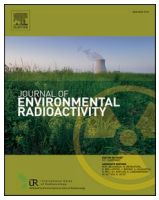
Journal of Environmental Radioactivity 237 (2021) 106649

Contents lists available at ScienceDirect

Journal of Environmental Radioactivity

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



Evaluating the added value of multi-input atmospheric transport ensemble modeling for applications of the Comprehensive Nuclear Test-Ban Treaty organization (CTBTO)☆

C. Maurer [a,*], D. Arnold Arias [a,b], J. Brioude [c], M. Haselsteiner [a], F. Weidle [a], L. Haimberger [d], P. Skomorowski [a], P. Bourgouin [e]

[a] Zentralanstalt für Meteorologie und Geodynamik, Hohe Warte 38, 1190, Vienna, Austria
[b] Arnold Scientific Consulting, Libertat 46, 08243, Manresa, Spain
[c] Atmosphere and Cyclone Lab (LACy - UMR8105), University de La Réunion, Avenue René Cassin 15, 97744, Saint-Denis, La Réunion, France
[d] Institut for Meteorology and Geophysics, University of Vienna, Althanstrasse 14, 1090, Vienna, Austria
[e] Comprehensive Nuclear Test-Ban Treaty Organization, Wagramerstrasse 5, 1400, Vienna, Austria

ARTICLE INFO

Keywords:
CTBTO
Atmospheric transport modeling
ECMWF Ensemble prediction system
Quantifying uncertainty

ABSTRACT

The Comprehensive Nuclear Test-Ban Treaty Organization (CTBTO) runs to date operationally an atmospheric transport modeling chain in backward mode based on operational deterministic European Centre for Medium-Range Weather Forecasts-Integrated Forecasting System (ECMWF-IFS) and on National Centers for Environmental Prediction-Global Forecast System (NCEP-GFS) input data. Meanwhile, ensemble dispersion modeling is becoming more and more widespread due to the ever increasing computational power and storage capacities. The potential benefit of this approach for current and possible future CTBTO applications was investigated using data from the ECMWF-Ensemble Prediction System (EPS). Five different test cases - among which are the ETEX-I experiment and the Fukushima accident - were run in backward or forward mode and - in the light of a future operational application - special emphasis was put on the performance of an arbitrarily selected 10- versus the full 51-member ensemble. For those test cases run in backward mode and based on a puff release it became evident that Possible Source Regions (PSRs) can be meaningfully reduced in size compared to results based solely on the deterministic run by applying minimum and probability of exceedance ensemble metrics. It was further demonstrated that a given puff release of 4E10 Bq of Se-75 can be reproduced within the meteorological uncertainty range [1.9E9 Bq,1.7E13 Bq] including a probability for not exceeding an assumed upper limit source term using simple scaling of a measurement with the corresponding ensemble metrics of backward fields. For the test cases run in forward mode it was found that the control run as well as 10- and 51-member medians all exhibit similar performance in time series evaluation. Maximum rank difference adds up to less than 10% with reference to possible rank values [0,4]. The maximum difference in the Brier score for both ensembles is less than 3%. The main added value of the ensemble lies in producing meteorologically induced concentration uncertainties and thus explaining observed measurements at specific sites. Depending on the specific test case and on the ensemble size between 27 and 74% of samples all lie within concentration ranges derived from the different meteorological fields used. In the future uncertainty information per sample could be used in a full source term inversion to account for the meteorological uncertainty in a proper way. It can be concluded that a 10-member meteorological ensemble is good enough to already benefit from useful ensemble properties. Meteorological uncertainty to a large degree is covered by the 10-member subset because forecast uncertainty is largely suppressed due to concatenating analyses and short term forecasts, as required in the operational CTBTO procedure, on which this study focuses. Besides, members from different analyses times are on average unrelated. It was recommended to Working Group B of CTBTO to implement the ensemble system software in the near future.

☆ This document is the result of a research project funded by the European Union Council Decision VII. The sponsor was not involved in 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).

https://doi.org/10.1016/j.jenvrad.2021.106649
Received 2 September 2020; Received in revised form 5 May 2021; Accepted 10 May 2021
Available online 9 June 2021

1. Introduction

The *Comprehensive Nuclear Test-Ban Treaty* (CTBT), an international agreement to ban all nuclear tests, includes the building of a global network of 321 monitoring stations and 16 laboratories for verification purposes (CTBT, 1996), the *International Monitoring System* (IMS, CTBTO Preparatory Comission (2019)). Stations therein monitor seismic, hydroacoustic, infrasound and radionuclide signatures of different origins with the aim to discriminate non-nuclear natural or civil man-made events against violations of the CTBT. The radionuclide component – to be considered in this paper – shall comprise measurements of radioactivity on aerosols at 80 locations at entry into force of the treaty. Half of the 80 stations shall have additional equipment to measure ambient air concentrations of four radioactive xenon isotopes (Xe-131m, Xe-133, Xe-133m, and Xe-135) produced in nuclear explosions.

In order to link radionuclide measurements with potential sources, the *Comprehensive Nuclear Test-Ban Treaty Organization* (CTBTO) runs operationally an atmospheric transport modeling (ATM) chain in backward mode (Matthews and De Geer, 2004) based on the Lagrangian Particle Transport and Dispersion Model FLEXPART, version 9.3.2 (Pisso et al., 2019; Stohl et al., 2005; 1998; ZAMG, 2018) and the operational deterministic ECMWF-IFS run (since June 30th, 2020, cycle *47r1*) with 0.1° native and 0.5° extracted horizontal resolution and with 137 vertical hybrid levels (ECMWF, 2018a) as well as the operational deterministic NCEP-GFS run (since December 6th, 2019, version *15.1*) with the same native and extracted horizontal resolution, but with 64 native vertical hybrid and 31 extracted pressure levels (NOAA, 2020a). Standard output products are *Source-Receptor Sensitivities* (SRSs), called *Field of Regard* (FOR) when an individual layer and one time step are considered, for the individual radionuclide IMS stations. *Possible Source Regions* (PSRs) are generated on demand in a continuative analysis using WEBGRAPE (Web connected Graphics Engine, CTBTO (2016)) combining several FORs and the corresponding measurements in a linear regression approach (Wotawa et al., 2003). Thus, a PSR is obtained by calculating the correlation between the FOR fields and the vector of corresponding observed concentrations for each geotemporal grid cell (Ringbom et al., 2014). Whereas one or two FOR fields have to be interpreted directly in terms of a possible source region (an unknown source term can be also estimated via scaling measurements with corresponding SRS values), the more advanced PSR method is the standard CTBT verification method of choice if a synopsis of many FORs is necessary (as is the case for all the backward test cases presented in this paper).

So far ensemble atmospheric transport modeling at CTBTO has been confined to multi-model ensemble modeling in case of exceptional particulate measurements (Level-5 measurements, see the definition in Matthews and De Geer (2004)). On the occasion of a Level-5 event so called *Regional Specialized Meteorological Centers* (RSMCs) as designated by the *World Meteorological Organization* (WMO) run their atmospheric transport and dispersion models and provide SRSs to CTBTO and state signatories (WMO, 2019).

Since the beginning of the 1990s (Palmer et al., 1993), ensemble prediction has become a widely-known approach to pragmatically treat uncertainty in numerical weather prediction (NWP) and air quality modeling, being still considered state of the art and subject of continuous research with yet increasing popularity (Bauer et al., 2015; Leutbecher and Palmer, 2008; Shemyakin and Haario, 2018; Wilks, 2011). If computationally affordable, it is the method of choice to take uncertainties of models and measurement data into account and create probabilistic forecasts, even in operational environments.

Because atmospheric transport models like FLEXPART take independent meteorological fields as input driving data during runtime (called "offline" models), they depend also on the accuracy of the output of NWP models and underlying measurements. Consequently, similar issues concerning predictability of ATM parameters arise as they do for the driving models. It is then intuitive to apply ensemble forecasting to those atmospheric transport models as well. Since the first attempt by Straume et al. (1998), the ensemble approach for atmospheric dispersion modeling has increased in popularity. As an example, a platform (European Commission – Joint Research Center, 2020) was founded to inter-link the community and provide a tool to develop and assess ensemble modeling studies including corresponding evaluation.

Galmarini et al. (2004) provide an overview of ensemble types and the research on them. The two most promising approaches for ensemble atmospheric transport modeling were later described and compared by Galmarini et al. (2010):

1. Multi-model ensemble: It uses different atmospheric dispersion models for each run. In most cases the different atmospheric dispersion models are also coupled with data from differing meteorological models.
2. EPS (ensemble prediction system) ensemble: It uses meteorological input data from an ensemble prediction system and each run is driven by a different member. This type of ensemble is tackled in the present research.

Galmarini et al. (2010) compared those two ensemble modes for FLEXPART and four other models. The study in case of EPS-based ensemble modeling was performed with a full 51-member ECMWF ensemble as input data and this EPS-ensemble showed promising results. Multi-model median and the mean of an EPS-based ensemble produced the best guidance when compared against measurements. More recently, in Galmarini et al. (2018) the benefits of a multi-scale (so-called hybrid) ensemble approach were investigated, whereby global and regional scales are combined.

Ensemble modeling is usually based on a large number of ideally independent models or model runs and this comes at a computational and resources expense. Authors dealing with EPS based ensemble dispersion modeling tend to use the full ensemble or, to reduce resources, draw a subset (as for example done in De Meutter et al. (2016), where just 10 ensemble members were considered when re-calculating the ensemble). A most recent work applying multi-input ensemble modeling with FLEXPART using the full 51 ensemble members from the ECMWF ensemble data assimilation (EDA) system is that of De Meutter et al. (2018). The use of ECMWF-EPS ensemble members for operational applications in air dispersion modeling is clearly complicated from a computational point of view. However, unlike with multi-model ensembles (Kioutsioukis and Galmarini, 2014) no redundancy has to be expected due to the design of the NWP ensemble system. Initial conditions are independent/uncorrelated for the 51 members and all the individual members statistically have equal chance to perform best.

Reducing the ensemble size in a sensible way, is often needed to be able to run ensemble forecasts on provided, and often limited, computational facilities. Ensemble runs based on a meteorological ensemble input data set reduced via clustering for forward modeling with FLEXPART for aviation purposes were described and tested in the framework of a Master's thesis by Klonner (2013). The method uses the horizontal wind field forcing at a specific vertical level as clustering variable and demonstrated utility for volcanic events reaching up to typical cruising level altitude (roughly 10 km). However, given that horizontal forcing is lowest in the boundary layer, this approach is not useful for the current study where focus is on the surface layer and where IMS measurements occur. Besides, transport times are as well longer (usually hemispheric or even global transport is to be considered), leading to much more vertical mixing. Klonner also investigated the use of multiple layers in the clustering, but results became fuzzy instead of improving. In addition, if the aim is to reduce the ensemble to a few representativ members, clustering is only reasonable on a continental scale, e.g. Europe (Ferranti and Corti, 2011), and for a limited amount of variables (with special focus on the geopotential). Members being similar to each other over, e.g., Europe do not need to be similar to each other over another or even over several other continents or regions of the earth. CTBTO's

atmospheric transport modeling system, however, covers the whole globe and it is not only horizontal wind components which govern dispersion in the surface layer. Therefore clustering would not lead to a better selection of ensemble members than randomly sampling an ensemble subset.

Principal Component Analysis (PCA) selection procedures require that the ensemble members are distinguishable. This is the case for multi-model or multi-physics ensembles, because each member employs a specific setting or model. The perturbation of initial conditions and the perturbation during the forecasts in ECMWF-EPS on the other hand are designed such that the members are interchangeable.

Most publications deal with ensemble forward modeling of atmospheric dispersion. However, in the work of Becker et al. (2007) ensemble backward atmospheric dispersion calculations are presented, using no EPS based but a multi-model ensemble, consisting of different particle transport models, including FLEXPART. The results of this work show, that also when using atmospheric transport models in backward mode, ensemble methods are superior to single deterministic runs. According to these authors the ensemble PSR average outperforms the median PSR if time and location of an event are unknown.

In order to investigate the added value of ECMWF-EPS based atmospheric transport modeling five test cases (three real cases, two synthetic cases) were performed, three of which constitute typical CTBTO applications in terms of transport ranges:

- *Hypothetical puff release* of 1E15 Bq Xe-133 (default standard emission in CTBTO's operational runs) *at the DPRK test site* near Punggye-ri (129.0° E and 41.3° N) on Dec., 1st, 00:00-01:00 UTC, 2018 traced back from selected receptors for 14 days (backward mode). This test case covers north hemispheric winter time conditions. With a puff-like release this event constitutes a good example in terms of release characteristics to further test the added value of the ensemble PSR-approach.
- *Hypothetical puff release* of 1E15 Bq Xe-133 *at the ANSTO radiopharmaceutical production site* (151.0° E and 34.1° S) on Dec., 1st, 00:00-01:00 UTC, 2018 traced back from selected receptors for 14 days (backward mode). This test case covers south hemispheric summer time conditions. With a puff-like release also this event constitutes a good example in terms of release characteristics to further test the added value of the ensemble PSR-approach.
- *Real, temporally extended and vertically structured releases* of Cs-137 and Xe-133 as available from the literature (Stohl et al., 2012) *for the Fukushima-Daiichi NPP* (141.0° E and 37.4° N) *accident* starting on March 11th, 2011, followed for 14 days (forward mode). It is probably the only test case in the CTBT context where abundant IMS measurements in combination with a best estimate source term are available. Above all, this test case demonstrates the capability of the ECMWF-EPS based ensemble to explain measurements on typical transport scales considered in the CTBT context taking into account meteorological uncertainty provided the source term is known.
- The *May 2019 Selenium-75 puff release from the Belgian research BR2 reactor* (5.1° E and 51.2° N) on the premises of SCKCEN (IRSN, 2019) run in backward mode for 6.5 days. With a puff-like release this event constitutes a good example in terms of release characteristics to further test the added value of the ensemble PSR-approach. In addition, the actual source term is also known with little error so estimating the source term based on simple scaling of a measurement with ensemble metrics of the SRS value at the known source location and emission time can be performed. However, the transport scale (confined to Europe) is not typical for CTBTO applications.
- The *ETEX-I release from Monterfil (France)* (2.0° W and 48.1° N) in forward mode simulated for 4 days, with the possibility to use a lot of measurements at high temporal resolution at various sites all over Europe and an extremely well defined release amount. The ETEX-I and II cases have been widely used by the atmospheric transport modeling community for performance testing (e.g., Galmarini et al. (2010), Graziani et al. (1998), Potempski et al. (2008) or Straume et al. (1998)). Also this test case demonstrates the capability of the ECMWF-EPS based ensemble to explain given measurements. However, the transport scale (confined to Europe) is again not typical for CTBTO applications.

Section 2 describes the set-up of the atmospheric transport and dispersion model FLEXPART together with the ECMWF input data and the (synthetic) measurements used for individual test cases. Results and a discussion thereof are provided in section 3. From the wealth of material created during the project in terms of ensemble FOR and PSR plots as well as box plots of time series and Taylor plots, for selected stations examples will be shown. Special emphasis will be on contrasting the full 51-member ensemble to a random 10-member subset, thereby testing the ensemble approach under possible operational constraints. The paper is completed by conclusions drawn in section 4 covering the relevance and implications of the results found for typical CTBTO applications.

2. Material and methods

For the five test cases, two types of ECMWF ensemble data sets were gathered, where the type depends on their occurrence back in time. Whenever possible – in terms of necessary input variables to run FLEXPART – existing ECMWF products were selected. Extraction from the ECMWF data archive and pre-processing of ECMWF fields for ingesting them into the FLEXPART model was done with the most recent version of a standard software package (Tipka et al., 2020). However, no existing data was available for the Fukushima test case dating back to 2011 and the ETEX-I test case dating even back to 1994, where an effort was undertaken to hindcast the ensembles. ERA5 (ECMWF, 2020a) data, providing global reanalyses from 1950 on till present, although in principle also useable since it includes 10 ensemble members, could not be considered because it is delayed in time by 5 days to the current date and thus would not be available in quasi near-real time as required by the operational needs of CTBTO. Whereas not important from the scientific perspective, for the operational aim of CTBTO one crucial aspect of this work was to test the approach assuming operational conditions. In order to neglect the forecast error which is not relevant to CTBTO's forensic post-event treaty verification tasks, ECMWF analyses were concatenated with short term forecasts rather than using one long-term forecast. Synthetic and real measurements were acquired, selected and pre-processed if necessary. Specific information related to measurement data can be found in corresponding tables in Appendix A. FLEXPART was set up uniformly for all five test cases except for output grid horizontal resolution.

2.1. 1st and 2nd test case: Hypothetical puff releases at the DPRK test site and the ANSTO radiopharmaceutical production site

Global ensemble analyses including the control run (in total 26 different input data sets, available at 00/06/12/18 UTC) were retrieved for Dec., 1st to 15th, 2018 from ECMWF's Meteorological Archival Retrieval System (MARS) archive (ECMWF, 2020b) based on the most recent IFS model cycle at that time, i.e., *45r1* (ECMWF, 2018b). This output of ECMWF's Ensemble of Data Assimilations (EDA) system is an ensemble of 4D-Var data assimilations (Isaksen et al., 2010; Lang et al., 2019) that reflects uncertainties in observations, atmospheric boundary conditions (such as sea-surface temperature) and the model physics by stochastic perturbation of parameterized tendencies. 25 additional symmetric members were created by subtracting 2x the difference between the 25 ensemble members and the control run from each member. Control run and deterministic run are both unperturbed and only differ by resolution. Native horizontal resolution of the EDA product adds up to $T_{CO}639/O640$ corresponding to 0.2°. Subscript "CO" stands for "cubic octahedral" and subscript "O" for octahedral reduced Gaussian grid.

These terms refer to the representation of the shortest wave and to the Gaussian grid characteristics. A switch from "linear" to "cubic" in the wave representation for the spectral IFS model and from the original reduced Gaussian to the octahedral reduced Gaussian grid was introduced in early 2016 (Malardel et al., 2016), which improved horizontal resolution. To end up with 3-hourly time resolution, the intermediate time steps between analysis times 06 and 18 UTC (all parameters to run FLEXPART are only available for forecasts launched at these two analysis times, but not for forecasts launched at 00 and 12 UTC) at a given day had to be filled with forecasts (steps 3, 6, and 9) started from analyses at 18 UTC of the previous day, at 06 UTC and at 18 UTC of the current day. This global meteorological data set with extracted 0.5° horizontal resolution and 137 vertical levels could be used for both synthetic cases as they cover the same time period. For the 4DVAR/-deterministic forecast native resolution adds up to $T_{CO}1279/O1280$ corresponding to 0.1°. Related unperturbed analyses (4DVAR) with interspersed deterministic forecasts based on cycle *45r1* at the very same extracted resolution of 0.5° and with 137 vertical levels were gathered as well. The spatial and temporal resolutions of the data sets used match the ones currently in place at CTBTO for atmospheric transport modeling. Not using the full spatial and temporal resolution is due to computational and storage limitations.

A NCEP-GFS (NOAA, 2020b) based FLEXPART version 8.2.3 run was used to generate pseudo Xe-133 measurements for the two synthetic test cases. Xe-133 is the most abundant xenon isotope measured by the IMS (Achim et al., 2016) and expected to be released by an underground nuclear test (Perkins and Casey, 1996; Sun and Carrigan, 2012) and - above all - during radiopharmaceutical isotope production (Bowyer et al., 2013; Kalinowski et al., 2014; Saey, 2009). NCEP-GFS meteorological input data used as well comprise a mixture of analyses with short-term forecasts on a 0.5° grid. However, vertical resolution is quite different (26 pressure levels for NCEP-GFS versus 137 hybrid model levels for ECMWF-IFS). Synthetic samples from the NCEP-driven FLEXPART version 8.2.3 run were selected based on the following criteria: 1) An average concentration of at least 1 mBq/m^3 over a collection period (12 or 24 h) had to be reached as this is the Minimum Detectable Concentration (MDC) required by the Preparatory Commission for a certified noble gas sampling device (Czyz et al., 2018). 2) Pseudo-samples had to stem either from an operational noble gas IMS station (country code, plus "X" and number, e.g., JPX38) or from a test bed IMS station ("XE" plus number, e.g., XE058; see CTBTO (2020a)). Both criteria ensured working with a realistic scenario in terms of data availability.

Since the Xe-133 plume quickly disappears from the surface model layer right after the release at the ANSTO site, only samples with collection stop times from Dec., 9th, onward could be selected. In terms of data availability this test case constitutes an unfavorable case due to the specific synoptic situation as depicted in the NCEP driving meteorological data. Xe-133 data for both synthetic cases can be found in Tables A1 and A2 in Appendix A.

2.2. 3rd test case: Real, temporally extended and vertically structured releases for the Fukushima-Daiichi NPP accident

North hemispheric hindcast ensembles were produced for the period March, 11th to 25th, 2011, because model level data (needed to run FLEXPART) was no longer available for the operational ensemble product of that time. ECMWF's IFS ensemble for the Fukushima test case was hindcasted based on the IFS-model cycle version *45r1* with native resolution of $T_{CO}639/O640$, corresponding to 0.2°, and 91 vertical levels. Perturbed short-range forecasts with initialization time of 00 and 12 UTC were retrieved on a 0.5° grid and 3-hourly time resolution and concatenated with analyses. The ensembles consist of 51 members, where 50 members were initialized with perturbed initial conditions retrieved from operational ensemble data analysis members and calculation of singular vectors was performed with a modified rescale factor (Vitart et al., 2019).

In addition, the 4DVAR product ($T_L1279/N640$ corresponding to 0.2° native resolution) with interspersed deterministic forecasts at the very same resolution was gathered as well. Subscript "L" stands for "linear" and subscript "N" for original reduced Gaussian grid. Both types of input data (deterministic and ensemble) were interpolated to a 0.5° grid in the frame of data pre-processing and both types of data exhibit 91 vertical levels. However, whereas the 4DVAR and the deterministic forecast product were retrieved from the ECMWF MARS archive and are based on IFS cycle *36r4* (ECMWF, 2010) introduced in November 2010, the ensemble including the control run was hindcasted based on IFS cycle *45r1*. Thus, the ensemble performance has to be evaluated against the control run of the hindcasted ensemble. Nevertheless, the dispersion results based on the 4DVAR product with interspersed deterministic forecasts are useful to demonstrate what the deterministic results would have looked like right after the accident and how ECMWF forecasts have potentially improved in the course of the last decade.

Upper level west and north-westerly currents led to a fast spread of the radioactive material towards the Pacific and to notable measurements above all at the US-IMS noble gas and particulate stations Ashland (USX74/USP74), Charlottesville (USX75/USP75), Wake Island (USX77/USP77) and Oahu (USX79/USP79). Although in an ideal position, Takasaki (JPX38/JPP38) station data, immediately located to the south west of the accident site, was not used in the study due to well-known data quality issues related to the high level of radioactivity (Stohl et al., 2012). Time series for Xe-133 and Cs-137 were selected based on the largest measurements corresponding to IMS stations that had measured significant amounts of both radionuclides within the target time period. Although this approach does not consider the absolute largest measurements, the collocation criteria allows to consider the complementary information (short-lived, non-depositing noble gas versus long-lived, depositing radionuclide attached to particulate matter) provided by the selected measurements in the subsequent evaluation of ECMWF-EPS based ensemble runs. Aerosol bound I-131 was not considered, because it is treated - apart from the half-life correction - exactly in the same way as Cs-137 in FLEXPART. The in total seven CTBTO IMS stations used together with mean and maximum activity concentrations are listed in Table A3. Particulate data are available online in the attachment B-1 to the corresponding *United Nations Scientific Committee on the Effects of Atomic Radiation* (UNSCEAR) Report (UNSCEAR, 2013). Permission to use all radionuclide data was granted to the contractor via contract number 2018-0655. IMS radionuclide data is in principle only available to state signatories or to researchers after signing a contract (CTBTO, 2020b) with CTBTO.

2.3. 4th test case: May 2019 Selenium-75 puff release from the Belgian research BR2 reactor

In May 2019 a release of Selenium (Se-75) occurred accidentally from the Belgian research reactor BR2 on the premises of SCKCEN in Mol (IRSN, 2019). The incident constitutes an ideal test case to evaluate the added value of EPS-based PSRs & FORs due to the specific characteristics of the release. Roughly 4E10 Bq of Se-75 were discharged between 13:15 and 14:00 UTC on May 15th, 2019. The release duration matches almost exactly the hourly output interval used in the FLEXPART calculations. Although a comprehensive data set comprising samples in several European countries is still to be expected only four detections and three non-detections from seven IRSN measurement sites in France were available for the present study (see Table A4 in Appendix A). Thus, the number of available detections (four) is more realistic in terms of the number of detections usually at hand for a CTBT relevant event. However, the distances between the emitter (in Belgium) and the sampling sites (in Northern France) are clearly smaller than in the usual CTBT context.

The Selenium case was evaluated on a European domain (30° W to 30° E and 30° N to 70° N) with 0.2° extracted meteorological input

horizontal resolution. Native horizontal resolution of the EDA product again adds up to $T_{CO}639/O640$, corresponding to 0.2°. However, no additional members had to be generated in a pre-processing step, because ECMWF upgraded its system in 2019 in order to produce 50 members instead of only 25 generated before the upgrade. The corresponding 4DVAR and deterministic forecast data set of ECMWF with native resolution of $T_{CO}1279/O1280$, corresponding to 0.1°, and with 0.2° extracted resolution was used as well. All data come with a vertical resolution of 137 model levels and are based on the very same IFS cycle *45r1*. Anaylses and forecasts were once again concatenated.

Since most collection starts fall before the release they were adjusted (see 5th column of Table A4 in Appendix A) based on a simulation performed by the Canadian Meteorological Service (CMC - ECCC, 2019) in order to avoid gathering of additional meteorological input data before the time of the release. Consequently, measurements - assuming a constant flow rate - had to be scaled by the quotient of original sampling duration and adapted sampling duration resulting in an increase of activity concentrations (see last but one and last column of Table A4). Given a half-life of Se-75 of roughly 120 days the error introduced by this approach is negligible. Measurements falling below the MDC were set to zero.

2.4. 5th test case: ETEX-I release from Monterfil (France)

The ETEX-I case was evaluated on the same European domain (30° W to 30° E and 30° N to 70° N) as the Selenium case and again with 0.2° extracted meteorological input horizontal resolution. ECMWF's IFS ensemble for the ETEX-I scenario was hindcasted based on the previous operational model cycle *46r1* with native resolution of $T_{CO}639/O640$, corresponding to 0.2°, and 91 vertical levels (ECMWF, 2019). 50 members were again initialized with perturbed initial conditions retrieved from ERA5 (ECMWF, 2020a) ensemble data analysis members. For the period from October, 23rd to 28th, 1994, the hindcasts were started twice per day at 00 and 12 UTC with a forecast range of up to 24 h and with hourly output (however, meteorological input data was only used with 3-hourly resolution and for up to 9 h forecast range for running FLEXPART).

In March 1994 ECMWF had switched to cycle *11r7* (ECMWF, 1994) which had only a horizonal resolution of $T_L213/N80$, corresponding to 1°, and 31 vertical levels. So for this test case the 4DVAR/deterministic forecast runs would not only be different in terms of model physics, but also very different in terms of resolution reversing the relationship to the ensemble in terms of resolution. Therefore 4DVAR and deterministic forecast input data dating back to 1994 were not considered in the evaluations.

340 kg of perfluorcarbon (PMCH) were released at Monterfil in France and corresponding measurement data are available at 168 European sites for 30 intervals of 3 h starting with the first sampling interval on Oct., 23rd, 15–18 UTC (Graziani et al., 1998). The ambient background concentration was already subtracted by the data provider. Thus, the concentrations, in ng/m^3, represent only the tracer released. All valid samples (including the categories *valid sample, no tracer found*; *valid sample, tracer found*; *concentration within 2 standard deviations of background variation* and *concentration given or higher (possible saturation effect)*) were gathered and the number of above zero samples was identified. Those stations which fulfilled the criterion of having more than 12 samples above zero were selected. The procedure yielded 15 stations which is similar to the number of particulate and nobel gas sampling stations considered for the Fukushima case. The 15 stations are listed in Table A5 in Appendix A.

2.5. FLEXPART Set-up

FLEXPART version 8.2.3 was used for test cases 1 and 2. For test cases 3 to 5, version 9.3.2 was employed which was specifically adapted to the operational needs of CTBTO. According to the authors' experience, version 9 onwards perform considerably better in the near field, thus being preferred for the Se-75 and the ETEX-I test cases. Undesirable features of version 8.2.3 include lack of dispersion in the horizontal and vertical, unrealistically small, secondary maxima and a plume passage which is too fast. For all test cases FLEXPART was run with a 900 s synchronisation interval and output sampling rate as well as a 3600 s averaging and output time interval. Subgrid terrain effect and convection parameterization were enabled. The FLEXPART output contains a single output layer with an upper height of 150 m a.g.l. for all test cases. The output domains and their spatial resolutions were chosen to be identical to the ones of the computational domains (the meteorological domains, respectively).

3. Results and Discussion

For forward and backward test cases different ensemble products were evaluated. Due to the very different users of CTBTO products, with a variety of backgrounds, not always atmospheric transport ensemble modeling, products together with evaluation statistics were required to be straightforward and easy to understand. Since in the CTBT context there is usually information on the timing and position of an event (via the complementary techniques infrasound, hydroacoustics or seismics) the focus is on linking the radionuclide measurements to the known time and location of a suspicious event. For the three test cases run in backward mode minimum, maximum, median PSR fields and probability of exceeding a threshold PSR value were investigated in detail. The field averages were also checked, but found to be of no added value compared to the medians. Minimum PSR fields can be of added value in case two relevant sites (e.g., civil radiopharmaceutical facility and known military test site with a corresponding waveform signal) fall within the deterministic PSR at the same time, but only one of them falls into the minimum PSR field. Maximum PSR fields are hardly of any value based on the three test cases investigated, because they are considerably enlarged and may no longer allow to distinguish between the likelihood of possible sources in different places. For the Fukushima test case and the ETEX-I case box plots of time series, Taylor diagrams as well as a Rank metric (combining correlation coefficient, fractional bias, Kolmogorov Smirnov parameter and accuracy) and the Brier score were employed. Taylor diagrams combine correlation coefficient, root mean square error and standard deviation and are especially useful for displaying ensemble performance because performance of every single member or deduced metric can be depicted by a (colored) symbol. For the Se-75 test case PSR fields as well as minimum, maximum, median, probability of exceedance and normalized variance of a selected FOR were analyzed. Ensemble results for 10 and 51 members were always contrasted against the control or/and - if reasonable - the deterministic run. Detailed information on the definitions of statistical metrics can be found in Appendix A.

3.1. 1st test case: Hypothetical puff release at the DPRK test site

Fig. 1 shows the PSR field based on deterministic ECMWF input for time step 20181201 00:00-01:00 UTC, corresponding to the time of the puff release of Xe-133. For a proper judgement of performance the whole hemisphere affected has to be plotted, the inset displays the zoom on the Korean peninsula. As it can be learnt from Fig. 1 location accuracy of the synthetic DPRK event is surprisingly high considering the discrepancy between the NCEP deterministic run (used to produce the synthetic measurements) and the ECMWF control run driven FLEXPART run (see Table A1 in Appendix A). The highest PSR values occur right in DPRK (dark blue area south-west of the red cross which indicates the test site). Results are not compromised by the fact that maximum PSR values stay below 0.4. High PSR values are often an artefact of linear regression analysis when just based on a few measurements (like it is the case for the ANSTO synthetic event, see section 3.2 below) and linear regression is dominated by a single pair of high values. If synoptic conditions are

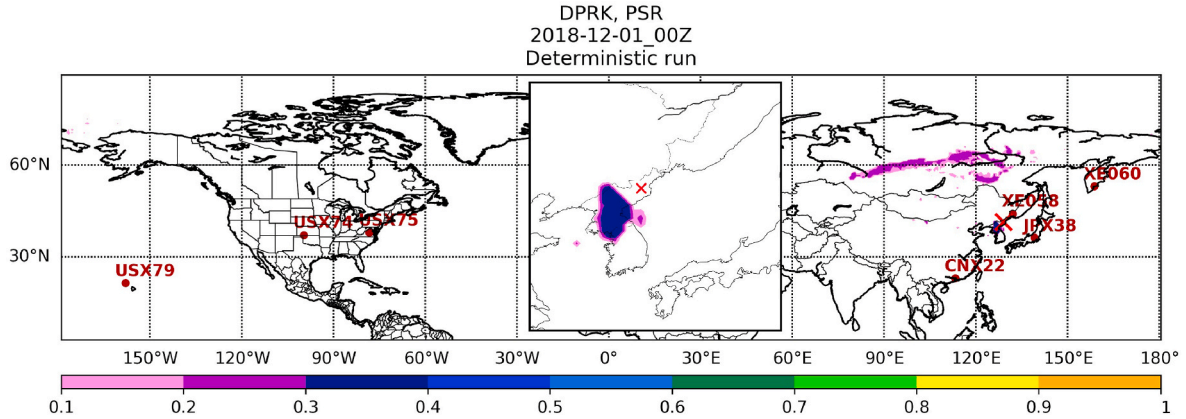


Fig. 1. PSR field based on the FLEXPART run driven with deterministic ECMWF input for time step 20181201 00:00-01:00 UTC. Zoom on DPRK in inset. Red cross indicates source location. IMS stations used to calculate the PSR field are labeled. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

favorable and many stations are hit by the plume under investigation location accuracy can be high. However, the exact match of the FLEXPART modeling output time step and the release duration of the hypothetical release is in favor of a more accurate result.

Figs. 2 and 3 show the outcome based on an arbitrarily drawn 10-member ensemble subset and the full 51-member ensemble for the very same time step. It becomes evident that much of the PSR field ($\geq$ 0.1) over Siberia disappears whereas PSR values of 0.3–0.4 can still be found in DPRK. Thus, already the minimum of the 10-member ensemble constitutes an added value compared to just the deterministic run. However, the minimum of the full ensemble adds further benefit compared to the reduced ensemble with PSR field values $\geq$ 0.2 and $\geq$ 0.3, retreating completely to DPRK, the actual source region. The minimum metric acts as a filter at every grid point, filtering all but the most significant parts of the PSR field. Importantly it does not diminish the correlation values uniformly. It rather enhances the contrasts between higher and lower values.

The median (not shown) provides hardly any added value compared to the deterministic run. The probability metric for exceeding a PSR value of 0.27 (Figs. 4 and 5) performs in a similar way as the ensemble minimum. DPRK is clearly labeled as source of the release with roughly 100% probability south-west of the DPRK test site and roughly up to 30% further off. However, the selection of the threshold is not straightforward. Probability of exceedance applied to PSR fields suffers from using a fixed threshold. Guided by experience throughout the present work the threshold was finally based on the 90% percentile of the whole ensemble over all time steps and grid points. More specifically, it was based empirically on the maximum of individual members' 90% percentiles after discarding those PSR values below 0.1. The feature of selecting a threshold rather subjectively is a clear weakness of the probability metric. This is unlike the situation with activity concentration values, where a generally useful threshold is easier to define (e.g., Operational Intervention Levels - OILs). In essence thresholds need to be redefined for every case based on those values deduced from the ensemble.

The average Figure of Overlap of each member with all the others (see equation A.4 in Appendix A) adds up to roughly 33% for both the reduced ensemble and for the full ensemble for the time step considered, that of the Average Agreement of each member with all the others (see equations A.8 to A.11 in Appendix A) to roughly 69%.These numbers (both to be found in the figure titles) indicate hardly any difference between the reduced and the full ensemble according to the two metrics.

Interestingly neither the deterministic run nor any of the ensemble metrics yields a match in time. The actual source location (red cross) is only covered by the area of highest PSR values, probabilities of exceedances, respectively, around 14 h later compared to the time to the puff release. However, in general the temporal agreement of a waveform event with a PSR field should not be nailed down to a certain hour interval due to the possibility of (considerably) delayed releases (e.g., DPRK nuclear test in 2013, see Ringbom et al. (2014)). This is in addition true for zero times determined based on isotopic ratios which may be uncertain by a day or more due to uncertainties in activity

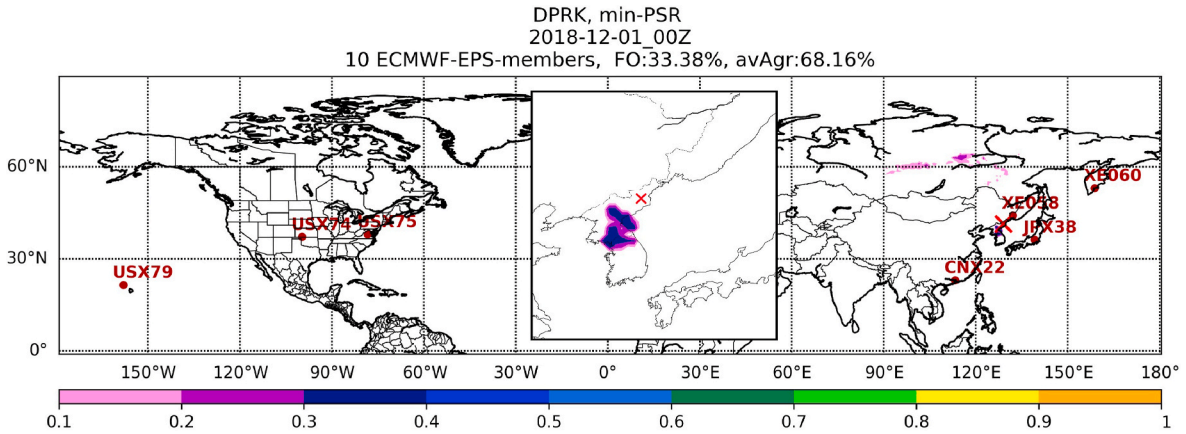


Fig. 2. Minimum PSR field based on the FLEXPART runs driven with a 10-member ECMWF-EPS ensemble for time step 20181201 00:00-01:00 UTC. Zoom on DPRK in inset. Red cross indicates source location. IMS stations used to calculate the PSR fields are labeled. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

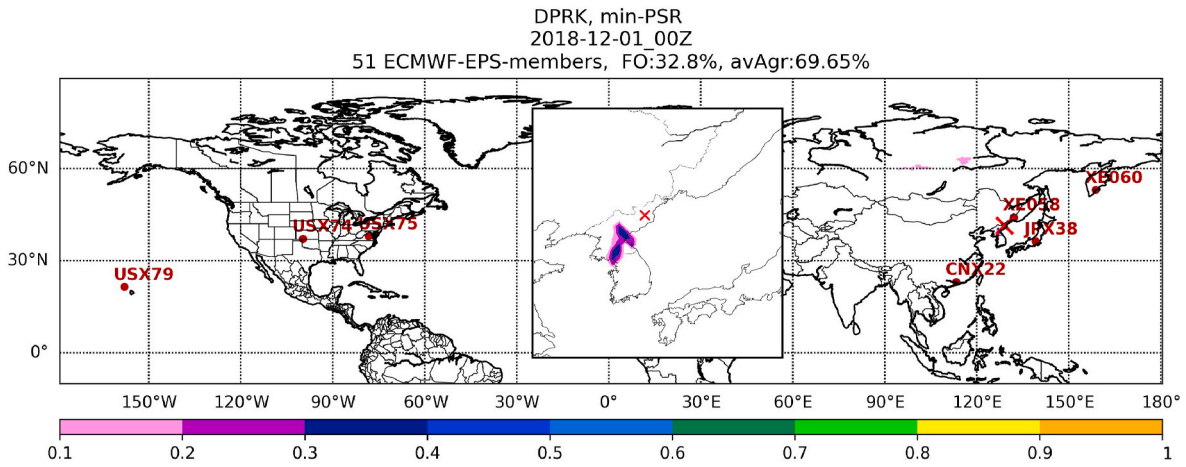


Fig. 3. Minimum PSR field based on the FLEXPART runs driven with the full ECMWF-EPS ensemble for time step 20181201 00:00-01:00 UTC. Zoom on DPRK in inset. Red cross indicates source location. IMS stations used to calculate the PSR fields are labeled. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

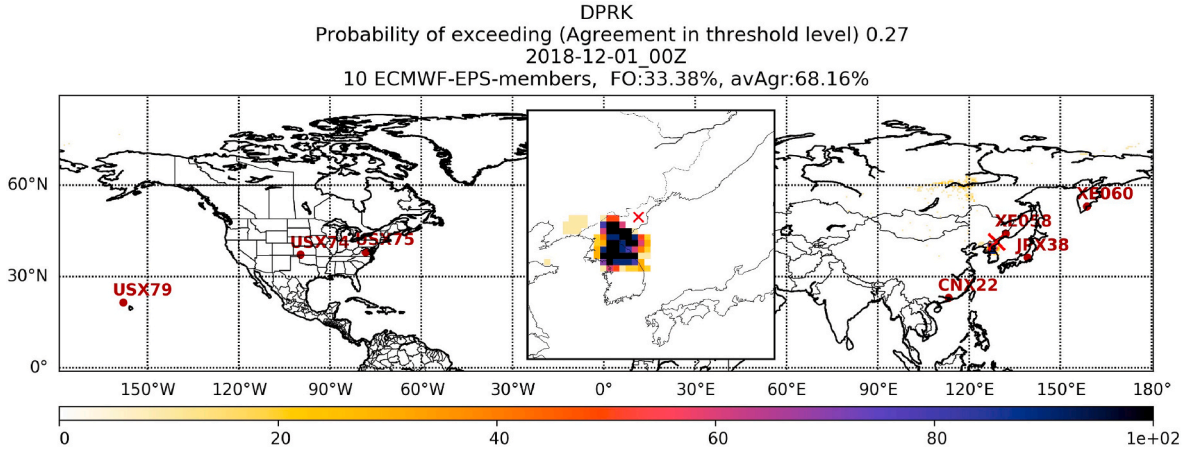


Fig. 4. Probability of exceeding a PSR value of 0.27 based on the FLEXPART runs driven with a 10-member ECMWF-EPS ensemble for time step 20181201 00:00-01:00 UTC. Zoom on DPRK in inset. Red cross indicates source location. IMS stations used to calculate the PSR fields are labeled. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

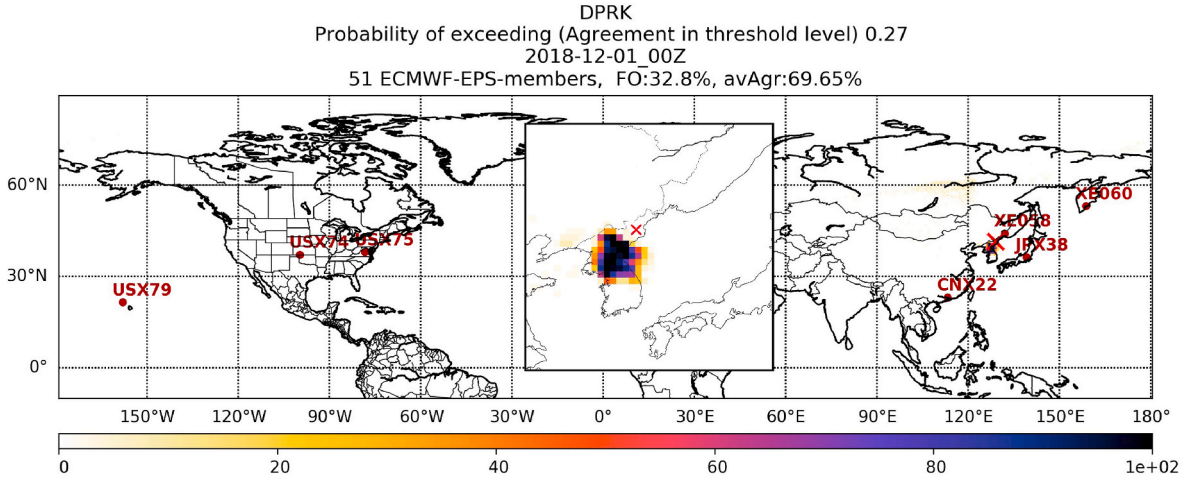


Fig. 5. Probability of exceeding a PSR value of 0.27 based on the FLEXPART runs driven with the full ECMWF-EPS ensemble for time step 20181201 00:00-01:00 UTC. Zoom on DPRK in inset. Red cross indicates source location. IMS stations used to calculate the PSR fields are labeled. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

measurements and different scenarios for ingrowth of precursors (Ringbom et al., 2014) and/or may also be affected by a delay between the date of origin (e.g, generation in a nuclear explosion or a reactor core) and the actual release of radionuclides to the atmosphere.

3.2. 2nd test case: Hypothetical puff release at the ANSTO radiopharmaceutical production site

Fig. 6 shows the PSR field based on the deterministic ECMWF input for time step 20181201 00:00-01:00 UTC, which is exactly the time of the puff release of Xe-133. Fig. 6 demonstrates that the PSR method fails for this very challenging event. This test case was challenging in two ways: 1) considerable discrepancies in the lowest model layer between the NCEP and ECMWF driven FLEXPART run occurred (see Table A2 in Appendix A) and 2) - at least according to NCEP data - the exceptional synoptic situation with only two IMS stations being hit by a reasonable amount ($\geq$ 1 mBq/m^3) of Xe-133 within 14 days after the release. As a consequence the PSR method locates the possible source location wrongly.

Nevertheless, the minimum of the reduced ensemble (Fig. 7) entails an interesting feature compared to Fig. 6 as PSR maxima (0.4–0.5) get confined to a region 1000 km southeast and 3000 km east of the actual source site with only a thin branch (PSR values $<$ 0.4) stretching over Antarctica. The minimum of the full ensemble (Fig. 8) reflects an even broader reduction of PSR size. The part of the PSR field stretching over Antarctica is further diminished (to small patches of values $<$ 0.2) compared to the reduced ensemble (Fig. 7) and nearly disappears completely. Just small parts east and southeast of Australia closest to the actual source location are left (with maxima in the range 0.3–0.4). Only a very small portion of the deterministic PSR field finally remains. The median (not shown) does not outperform the deterministic run. Performance of the probability of exceedance (also not shown) is - like for the DPRK test case - similar to that of the ensemble minimum. Although one might expect a benefit from the maximum of the (full) ensemble specifically for this test case, no benefit can be demonstrated (Fig. 9).

The Figure of Overlap varies between 55% for the reduced ensemble and 50% for the full ensemble, that of the Average Agreement between 71% and 67%. The difference is thus bigger than for the DPRK test case, also reflected by a bigger difference between PSR products for the reduced and the full ensemble. A reduced agreement among all the ensemble members becomes naturally more probable with increasing ensemble size.

3.3. 3rd test case: Real, temporally extended and vertically structured releases for the Fukushima-Daiichi NPP accident

Box plots of time series (including minimum, 1st quartile, median, 2nd quartile and maximum) together with the deterministic run, the control run and the measurements (including the MDC) for Xe-133 are displayed in Figs. 10 and 11 for IMS stations USX75 on the US east coast and USX79 in the Pacific for the full ensemble. These stations were chosen to be displayed because they comprise the largest numbers of available measurements within the simulation period. All measurements with an appropriate quality rating up to the simulation end (March, 25th, 21:00 UTC) are considered (therefore time series date-times may not be continuous). A logarithmic scale was used due to the range of activity values encountered.

Table 1 gives an overview for the reduced 10-member and for the full 51-member ensemble regarding performance to capture measurements for all selected individual time series ("X" refers to noble gas sampling stations, "P" to particulate sampling stations) applying again a threshold of 1 mBq/m^3 to Xe-133 observations and the MDC as threshold to Cs-137 observations. The reason for applying the 1 mBq/m^3 threshold also here is that especially concentrations at or below this level are more likely to be due to the normal radioxenon background (Achim et al., 2016) rather than being related to the Fukushima accident. On average 53% and 74% of the selected Xe-133 samples are covered by the 10- and 51-member ensemble.

A larger number of samples clearly reveals big discrepancies between FLEXPART calculations based on the deterministic run and the control run (see also subsection 2.2 and Table 2). It has to be stated that model concentrations are generally highly underpredicted for Cs-137, also based on improved ECMWF input data produced in the hindcast. However, Stohl et al. (2012) stated that *"for the FLEXPART run driven with ECMWF data wet scavenging of Cs-137 was much stronger compared to the NCEP driven run causing a strong underestimation of Cs-137 concentrations at sites in North America and Europe. Because the agreement of model results (both using a priori and a posteriori emissions* [as derived by Stohl et al., previous to and after the inversion evaluation]*) with measurement data was better with NCEP data than with ECMWF data also for Xe-133, results based on ECMWF data were largely discarded"*. Of course, Stohl's analysis only refers to and can only refer to the IFS cycle of that time, i.e. *36r4*.

Most importantly, model performance is similar for the control run, the reduced and the full ensemble median when evaluating Xe-133 time series in detail. This is especially true for both median values. There is only a slight benefit when switching from the control run to ensemble medians in terms of metrics. These features are confirmed by the two example Taylor diagramms in Figs. 12 and 13 as well as ranks (rank ranges from 0 to 4, see Appendix A, equation A.8) and the Brier scores (calculated for the reduced and the full ensemble based on the MDC, see equation A.3 in Appendix A) in Table 2. It is remarkable that the deterministic run (based on the IFS model version in place in 2011) behaves like an outlier for station USX75 (see Fig. 12). With perfect performance for BS = 0, the ensemble does a very good job in forecasting concentrations above the MDC for three of the six stations (BS $\leq$ 0.25) and a good job for two of the remaining three stations (BS $\leq$ 0.44).

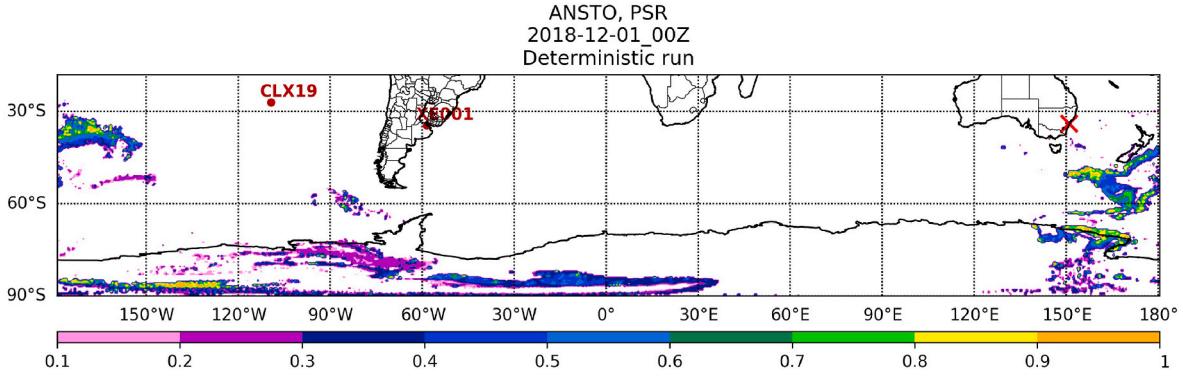


Fig. 6. PSR field based on the FLEXPART run driven with deterministic ECMWF input for time step 20181201 00:00-01:00 UTC. Red cross indicates source location. IMS stations used to calculate the PSR field are labeled. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

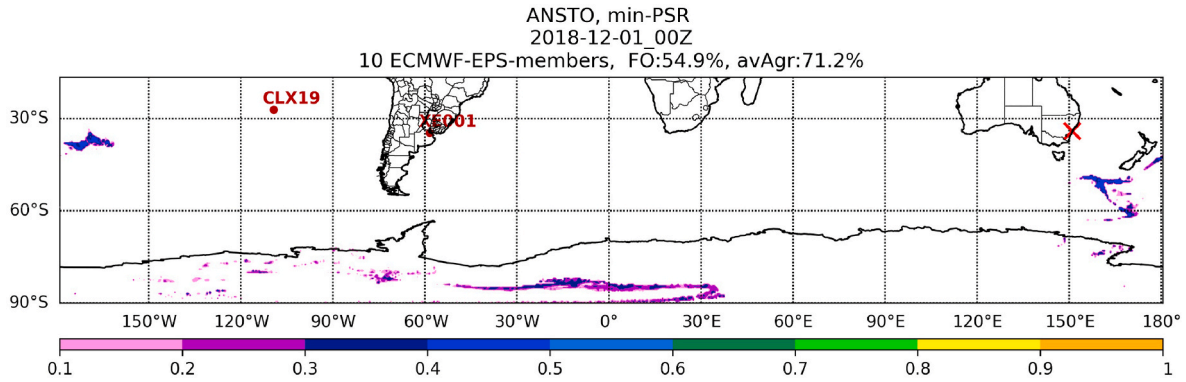


Fig. 7. Minimum PSR field based on the FLEXPART runs driven with a 10-member ECMWF-EPS ensemble for time step 20181201 00:00-01:00 UTC. Red cross indicates source location. IMS stations used to calculate the PSR fields are labeled. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

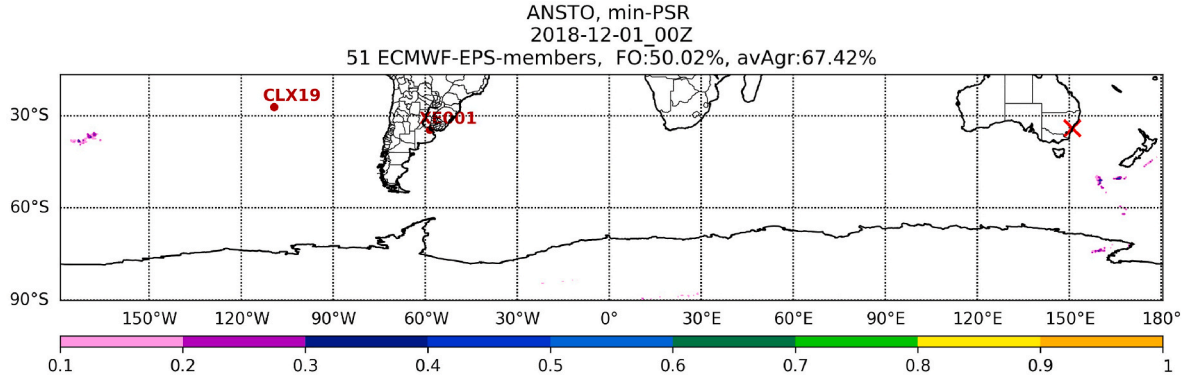


Fig. 8. Minimum PSR field based on the FLEXPART runs driven with the full ECMWF-EPS ensemble for time step 20181201 00:00-01:00 UTC. Red cross indicates source location. IMS stations used to calculate the PSR fields are labeled. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

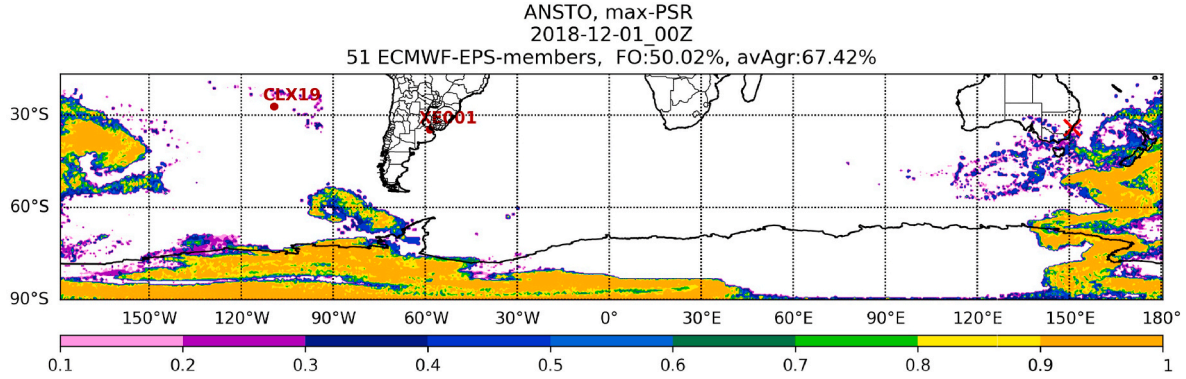


Fig. 9. Maximum PSR field based on the FLEXPART runs driven with the full ECMWF-EPS ensemble for time step 20181201 00:00-01:00 UTC. Red cross indicates source location. IMS stations used to calculate the PSR fields are labeled. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Stations featuring the highest BS values (e.g., station USX75 or SEX63) are those for which there are a number of measurements around or even above 1 mBq/m^3 which are probably not related to the Fukushima accident (see also Achim et al. (2016)). However, no definite statement is possible about the provenance of these samples. Maximum rank difference between the control run and the 10- and 51-member medians adds up to less than 10% with reference to possible rank values [0,4]. The maximum difference in the Brier score for both ensembles is less than 3%. The added value of the ensemble is rather in giving an indication of possible concentration spreads (which often include the actual measurements) than yielding a more accurate forecast (via calculation of the median).

Finally, it has to be stated that all statistical metrics presented in this subsection suffer from the comparatively small number of measurements (with a maximum of 23 samples for USX75) per station. This is also related to the fact that not all elevated measured samples could be modeled due to being forced to cut the simulation on March, 25th, 21 UTC because of storage demands for ECMWF ensemble data (hemispheric ECMWF full ensemble data at 0.5° resolution and with 91 vertical levels requires 1.1 TB of disk space). Cs-137 simulations are

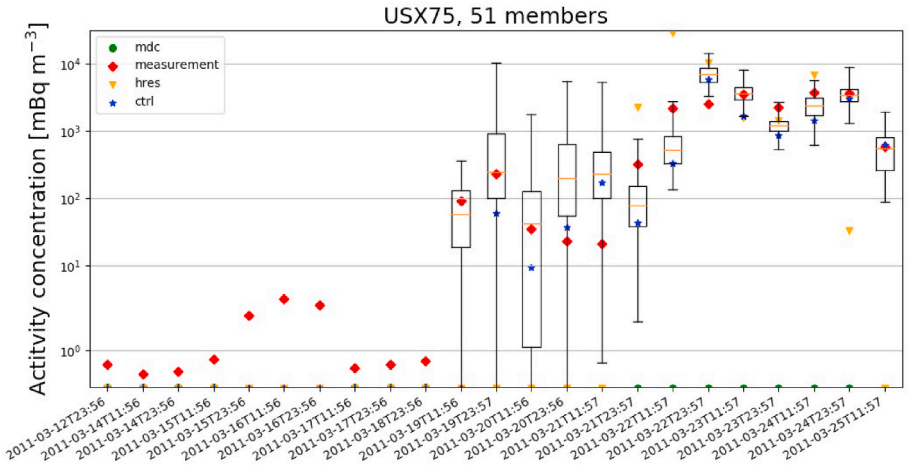


Fig. 10. Box plots of ensemble time series, deterministic run (hres), control run (crtl) and measurements including MDCs for IMS station USX75 for the full ensemble and for Xe-133. x-axis indicates collection stop times.

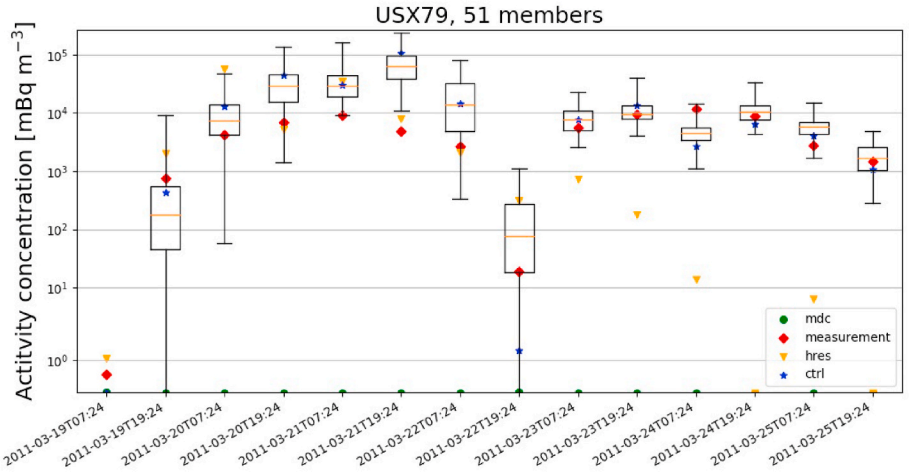


Fig. 11. Box plots of ensemble time series, deterministic run (hres), control run (crtl) and measurements including MDCs for IMS station USX79 for the full ensemble and for Xe-133. x-axis indicates collection stop times.

especially affected since major emissions ended several days after those of Xe-133. However, a sparse number of data pairs will usually be the case when working with CTBTO IMS data.

3.4. 4th test case: May 2019 Selenium-75 puff release from the Belgian research BR2 reactor

Fig. 14 shows the PSR field based on the deterministic ECMWF input for time step 20190515 13:00-14:00 UTC, roughly corresponding to the time of the puff release of Se-75. For a proper judgement of performance again the whole domain containing non-zero PSR values has to be plotted. As could be expected given the close proximity between the actual source location (in Belgium) and the measurement sites (in France) as well as the puff-like nature of the release the PSR method works quite well. Unlike for the DPRK test case (see subsection 3.1) the highest PSR field values include the actual source location right at the time of the actual release (13:00-14:00 UTC).

From Figs. 15–17, it becomes evident that there is a considerable benefit from using the ensemble minimum and the probability of exceedance (the latter is just shown for 10 members) metrics due to an increased contrast between more likely and less likely source regions. The minimum PSR field based on the full ensemble just displays three small patches of correlation values ≥ 0.9 and one of them is right at the actual source. There is a slight added value from using the median ensemble metric in terms of better confining the area with highest correlation values (not shown). The Figure of Overlap and the Average Agreement add up to 65 and 80% for the reduced ensemble and to 67 and 80% for the full ensemble and are thus again very similar for both ensemble sizes. The concept of using the 90% ensemble PSR percentile as threshold for the probability of exceedance gets confirmed via this test case.

Given the reported 4E10 Bq of the puff release, an attempt is made to estimate the release based on ensemble SRS values (sometimes also called "poor man's inversion"). For the detection from the most distant sampling site (Omonville, 0.0011 mBq/m^3, collected between 20190513 00:00 UTC and 20190520 00:00 UTC) Table 3 shows the SRS value and its ensemble metrics interpolated to the location of the BR2-reactor for the reduced and the full ensemble for the time interval 13:00-14:00 UTC, roughly corresponding to the time of the puff release of Se-75. Simple scaling of the (scaled) measured concentration with the SRS values as extracted from the table adds up to roughly 1.26E10 Bq for the deterministic run, to roughly 1.52E10 Bq for the median, 8.97E11 Bq for the minimum and 3.08E9 Bq for the maximum of the reduced ensemble. Source term estimates are inversely proportional to SRS values. The best source term estimates can be deduced from the ensemble median and the deterministic run. More importantly, the

Table 1
Number of samples N greater than or equal to selected threshold value and number of explained samples N for the reduced and the full ensemble for the six selected IMS sampling sites time series for each of the two isotopes.

IMS station	N ≥ thres.	N expl. -10/51 mem.	Comment on Cs-137 time series
CAX16 (Yellowknife, Canada)	3	2/3	
CAP17 (St. John's, Canada)	2	0/0	Modeled Cs-137 too diluted when reaching the station
SEX63 (Stockholm, Sweden)	9	3/5	
SEP63 (Stockholm, Sweden)	2	1/1	Timing of the plume arrival reproduced
USX74 (Ashland, US)	13	5/9	
USP74 (Ashland, US)	6	1/5	For 5/1 samples the ensemble maximum comes close to the measurements
USX75 (Charlottesville, US)	16	10/12	
USP75 (Charlottesville, US)	4	1/3	Timing of the plume arrival reproduced
USX77 (Wake Island, US)	6	3/3	
USP77 (Wake Island, US)	3	0/0	
USX79 (Ohau, US)	13	9/12	
USP79 (Ohau, US)	6	0/1	Timing of the plume arrival reproduced

Table 2
Number of involved sample pairs N, rank and Brier score for the Fukushima test case.

IMS station	N	Rank	BS
CAX16-deterministic	5	1.81	–
CAX16-control	5	2.65	–
CAX16-10 members	5	2.97	0.41
CAX16-full ensemble	5	3.00	0.38
SEX63-deterministic	21	0.71	–
SEX63-control	21	1.97	–
SEX63-10 members	21	2.20	0.66
SEX63-full ensemble	21	2.26	0.66
USX74-deterministic	14	2.24	–
USX74-control	14	3.02	–
USX74-10 members	14	3.27	0.10
USX74-full ensemble	14	3.21	0.11
USX75-deterministic	23	1.86	–
USX75-control	23	2.68	–
USX75-10 members	23	3.07	0.44
USX75-full ensemble	23	3.01	0.44
USX77-deterministic	7	1.13	–
USX77-control	7	2.18	–
USX77-10 members	7	2.34	0.16
USX77-full ensemble	7	2.39	0.17
USX79-deterministic	14	2.21	–
USX79-control	14	1.96	–
USX79-10 members	14	2.16	0.07
USX79-full ensemble	14	2.19	0.07

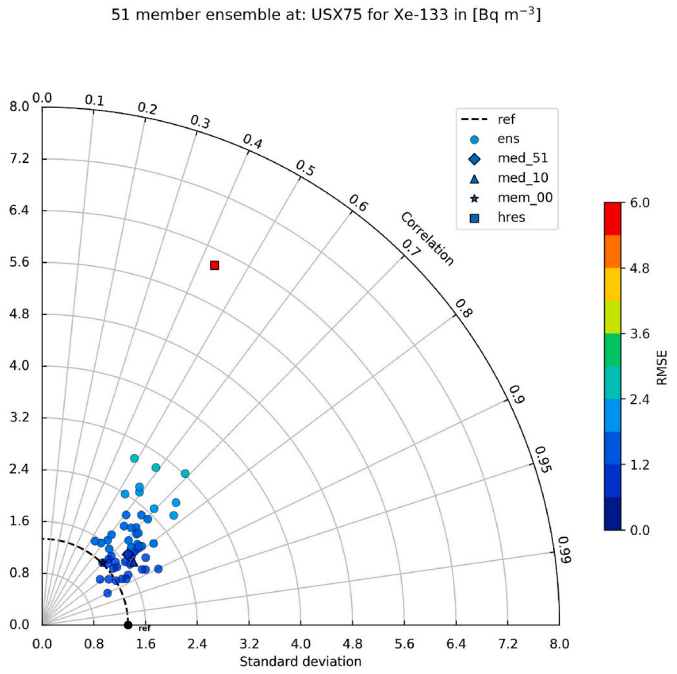


Fig. 12. Taylor diagram including the individual members (ens), the medians (med_10 and med_51–10 and 51 members), the control (mem_00) and the deterministic (hres) run for IMS station USX75. The reference point is displayed as black dot.

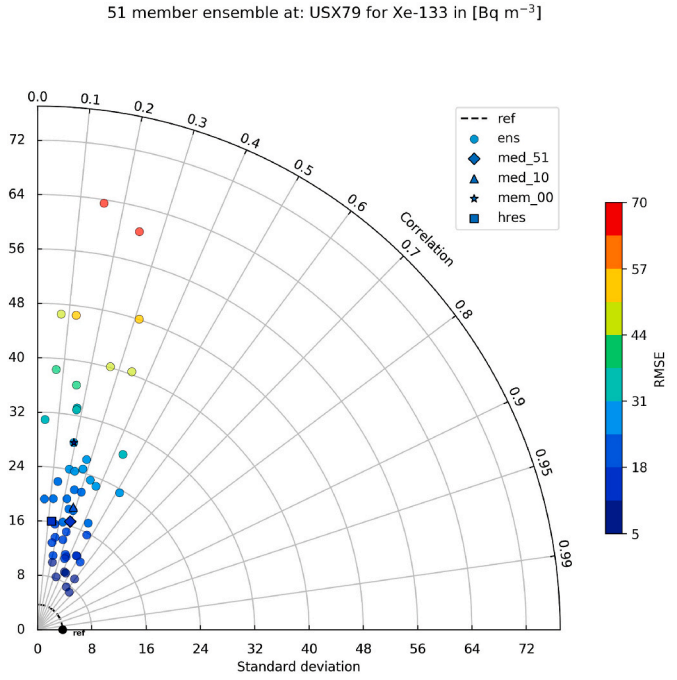


Fig. 13. Taylor diagram including the individual members (ens), the medians (med_10 and med_51–10 and 51 members), the control (mem_00) and the deterministic (hres) run for IMS station USX79. The reference point is displayed as black dot.

actual source term is covered by the ensemble of modeled source terms. The values in the 3rd and 7th row in Table 3 imply that there is a 94–100% probability that the source term falls below 2.00E12 Bq. The concentration threshold of 1E-18 Bq/m^3 was chosen for both the probability of exceedance and the normalized variance as it corresponds to 1 mBq/m^3 if the standard CTBTO source term of 1E15 Bq would have been used. The degree of uncertainty in source term estimates related to ECMWF input data based on median, minimum and maximum SRS values is reflected by a medium to high normalized variance in Fig. 18. The full ensemble has in fact no advantage over the reduced ensemble for estimating the release amount and its uncertainty. SRS values are very similar apart from the minimum and maximum. This behaviour is again underpinned by the Figure of Overlap and the Average Agreement.

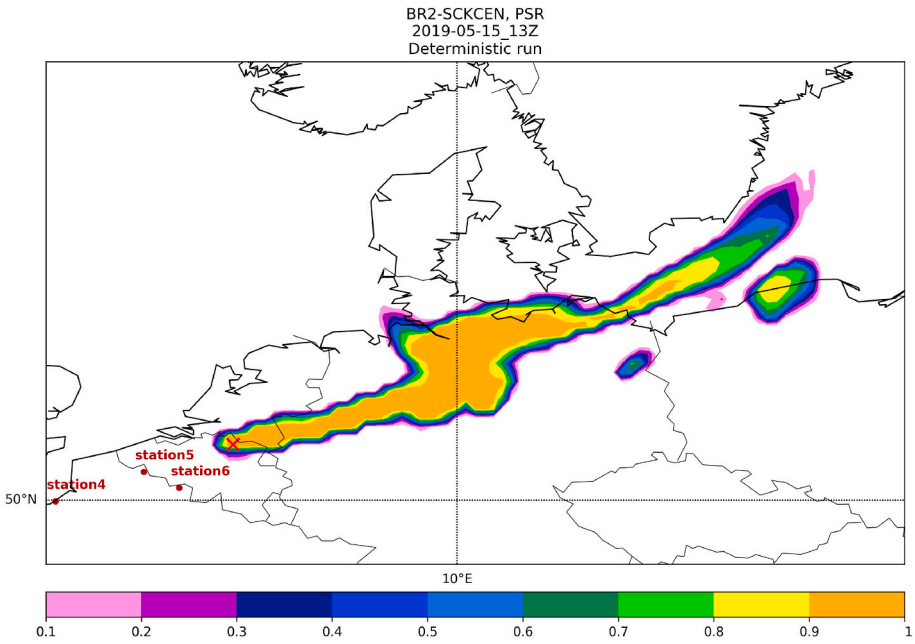


Fig. 14. PSR field based on the FLEXPART run driven with deterministic ECMWF input for time step 20190515 13:00-14:00 UTC. Red cross indicates the actual source location. Stations used to calculate the PSR field are labeled. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

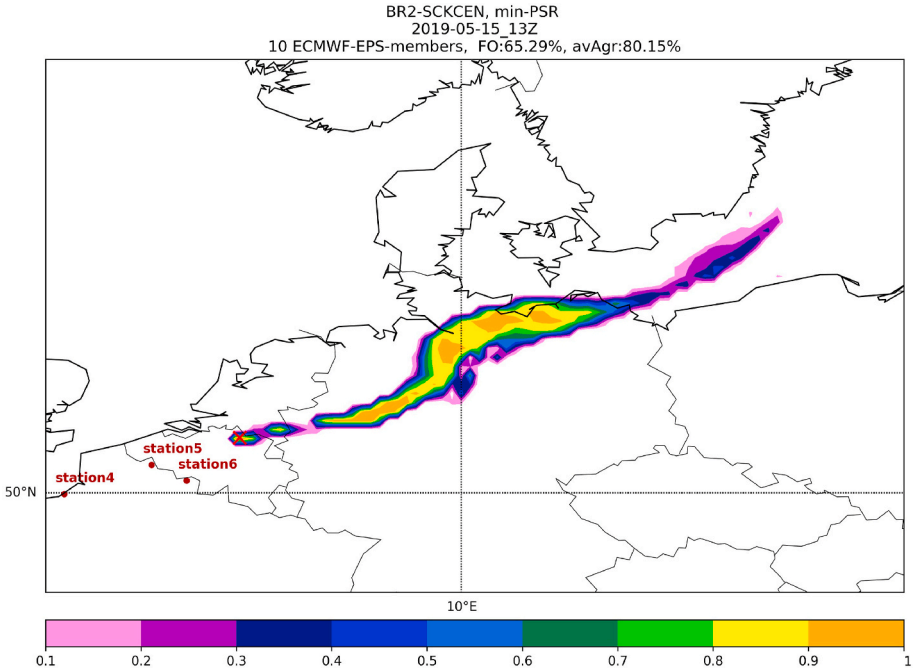


Fig. 15. Minimum PSR field based on the FLEXPART runs driven with a 10-member ECMWF-EPS ensemble for time step 20190515 13:00-14:00 UTC. Red cross indicates the actual source location. Stations used to calculate the PSR fields are labeled. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

These scores add up to 73 and 93% for the reduced ensemble and to 72 and 92% for the full ensemble. The values are higher than for both synthetic test cases (especially for the Figure of Overlap) which has to be expected due to the shorter transport ranges.

3.5. 5th test case: ETEX-I release from Monterfil (France)

Overall, the plume passages are well captured and ranks are comparable to or even higher than the ones of the Fukushima test case, but the ensemble has less skill in covering actual measurements (see Table 4) based on the uncertainty in meteorological input fields. However, it should be recalled that for the Fukushima case the IMS sampling times were much longer. Also, it is well known from model-measurements inter-comparisons that higher measurements are easier to predict than smaller ones (Arnold et al., 2015). For the Fukushima test case it was reasonable to exclude Xe-133 measurements below a threshold for the analogous evaluation.

The differences between the control run, the arbitrarily selected 10-

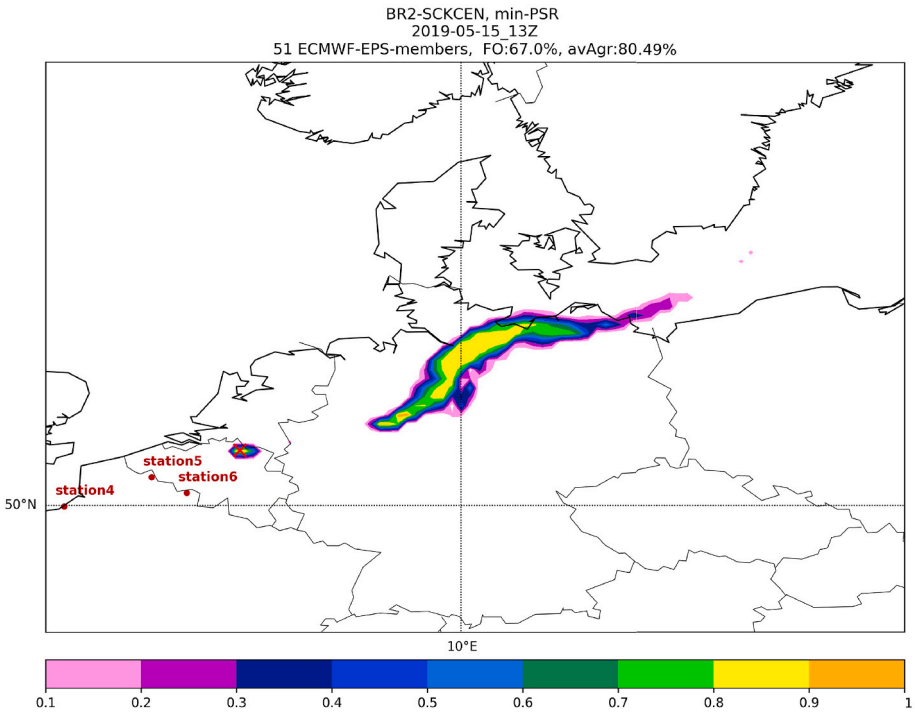


Fig. 16. Minimum PSR field based on the FLEXPART runs driven with a 51-member ECMWF-EPS ensemble for time step 20190515 13:00-14:00 UTC. Red cross indicates the actual source location. Stations used to calculate the PSR fields are labeled. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

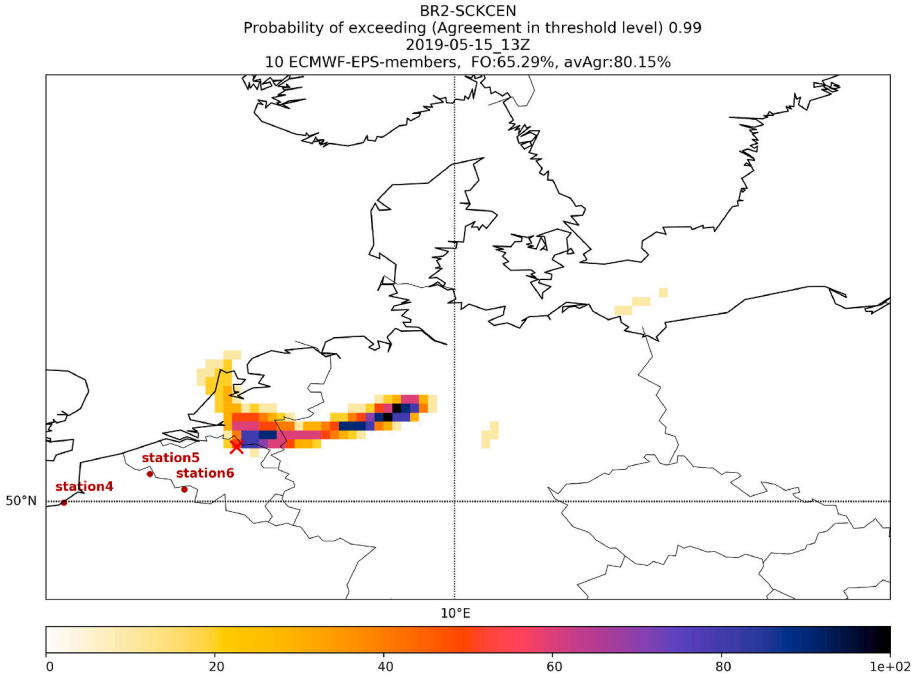


Fig. 17. Probability of exceeding a PSR value of 0.99 based on the FLEXPART runs driven with a 10-member ECMWF-EPS ensemble for time step 20190515 13:00-14:00 UTC. Red cross indicates the actual source location. Stations used to calculate the PSR fields are labeled. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

member and the full ensemble median are very small which is confirmed by corresponding ranks in Table 4. This is again especially true for both medians based on different ensemble sizes. Plume arrival is properly indicated by the ensemble for all but one station (DK06), at least via an above zero ensemble maximum. On average 27% of the samples are covered by the reduced and 37% by the full ensemble. Contrasting these two numbers with the equivalent numbers of the Fukushima test case the added value of the full ensemble is only half as large. The ability to cover measurements varies considerably between the individual stations. Whereas only one measured sample falls into the range of possible modeled concentrations for DK05 and DK10 for both the reduced and the full ensemble, 12 out of 14 samples are captured by the full ensemble

Table 3
(Reduced and full ensemble) metrics names, corresponding SRS values interpolated to 51.2° N and 5.0° S (grid point closest to the BR2 reactor) and estimated source term resulting from scaling with the scaled Omonville measurement.

(Ensemble) metric	SRS value [1/m^3] or probability [%]	Source estimate [Bq]
Deterministic run	1.59E-16	1.26E10
Probability of exceeding 1E-18 1/m^3 – reduced ensemble	100	2.00E12
Minimum – reduced ensemble	2.23E-18	8.97E11
Median – reduced ensemble	1.32E-16	1.52E10
Maximum – reduced ensemble	6.50E-16	3.08E9
Probability of exceeding 1E-18 1/m^3 – full ensemble	94	2.00E12
Minimum – full ensemble	1.15E-19	1.74E13
Median – full ensemble	1.59E-16	1.26E10
Maximum – full ensemble	1.08E-15	1.85E9

for N07 (see Fig. 19 and Table 4).

Box plots of ensemble time series together with the control run and the measurements for species PMCH for the full ensemble are displayed for two ETEX-I sampling stations in Figs. 19 and 20, namely for N07 and PL02.

4. Conclusions

The present work demonstrates that analyzing FLEXPART based ensemble SRS, FOR and PSR products in addition to results based on the deterministic ECMWF run is of added value when (puff) release amounts and possible source regions have to be determined. For the FOR fields, the comprehensive list of ensemble products investigated and that are considered suitable to reflect meteorological uncertainty in a meaningful way comprises the median, minimum, maximum, probability of exceedance and normalized variance. The minimum, median and maximum products as a function of space and time can indicate the possible spread in estimated source terms, whereas a high probability of exceedance given a specific FOR threshold value allows to estimate the upper bound of the source terms under investigation. Since a source term can be preliminarily estimated based on scaling of a measurement with the SRS value at the assumed source location and emission time its variation implicitly gives also information about the variation of the estimated source term. The normalized variance, specifically developed within this project, gives an impression of the expected degree of FOR variation. For the PSR fields the selected metrics include minimum and probability of exceedance, which were demonstrated to better constrain a possible source region via enhancing contrasts between less likely and more likely source regions. Finally, ensemble time series can help explaining measured samples of the IMS if the underlying source term is known.

The test cases confirmed that an arbitrarily selected 10-member ensemble is sufficient (and probably even mandatory under operational CTBTO-constraints) in order to benefit to a large degree from desirable ensemble properties, i.e., more constrained PSR fields as well as uncertainty estimates for FOR fields and modeled time series which capture the actual measurements. The gain of using the full 51-member ensemble is rather small compared to the computational efforts. The relation between skill and ensemble size was found to depend heavily on the used measures in the past, but in general skill is expected to increase with the ensemble size (see, e.g., Buizza and Palmer (1998)) if the ensemble members are chosen randomly. Model performance when trying to predict measured samples, however, is similar for the control run when compared to the ensemble medians (in agreement with the work of De Meutter et al. (2016)), and even more for the reduced and the full ensemble medians in the present study, which may reflect the fact that forecast uncertainty was largely suppressed by concatenating analyses and short term forecast (forecast range is always smaller than or equal to 9 h). Besides, e.g., member #1 started from the analysis at 12

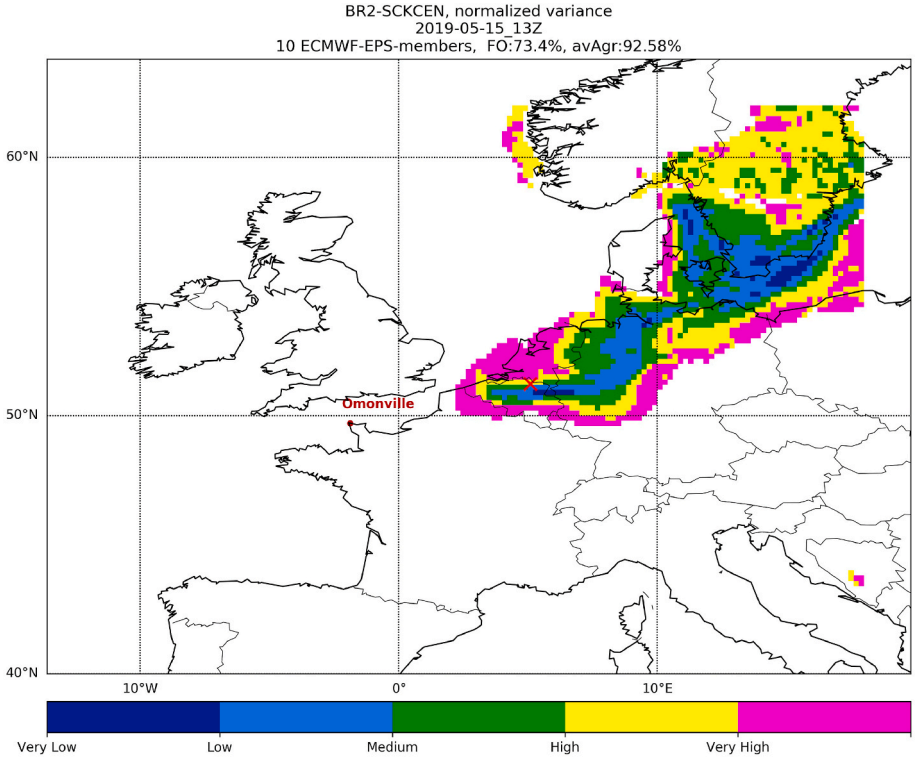


Fig. 18. Normalized variance of FOR fields for the Omonville sample based on the FLEXPART runs driven with a 10-member ECMWF-EPS ensemble for time step 20190515 13:00-14:00 UTC. Red cross indicates the actual source location. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 4
Number of samples N > 0.0, number of explained samples N > 0.0 for the reduced and the full ensemble and rank for the control run as well as the reduced and the full ensemble medians for the 15 selected ETEX-I sampling sites time series.

Sampling station	N > 0.0	N > 0.0 expl. – 10/ 51 mem.	Rank control/10-mem med./ 51-mem med.
A02	20	4/5	1.62/1.60/1.64
CR01	18	3/7	2.99/3.11/3.05
DK01	13	3/4	2.86/2.90/2.89
DK05	16	1/1	2.40/2.40/2.40
DK06	14	4/5	2.25/2.23/2.28
DK10	14	1/1	2.72/2.74/2.72
F02	9	3/3	3.19/3.18/3.14
H01	14	3/3	3.19/3.24/3.18
H02	13	5/10	2.95/3.04/3.06
N07	14	11/12	3.43/3.46/3.44
PL02	13	6/9	3.59/3.59/3.62
PL03	16	5/7	2.89/2.90/2.89
PL08	15	1/3	2.44/2.41/2.46
SR01	15	3/5	3.01/2.90/2.95
SR03	14	3/4	3.01/3.12/3.07

UTC on average does not ”know” the history of member #1 started at 00 UTC due to the randomness of perturbations introduced during data assimilation. This may contribute to the fact that a lot of uncertainty is already covered by a 10-member subset. In principle (although computationally hardly feasible) the ensemble could be enlarged via permutations (i.e., e.g., continuing the first 12 h based on member #1 with member #10 for the next 12 h and so on). There is - on average - no use in ”tagging” ECMWF ensemble members. This is unlike the situation with, e.g., a multi-physics ensemble (e.g., Evans et al. (2012)). The added value of the ensemble is rather in giving an indication of possible concentration ranges (which - as demonstrated - often include the actual measurements) and stating their likelihood rather than yielding a more accurate forecast.

Probably the most important impact under a future perspective will be estimating uncertainties of modeled time series for the full inversion (e.g., Stohl et al. (2012)) of a complex (time variable) source term. Apart from using a specific uncertainty estimate for each modeled sample, using FOR and PSR ensemble metrics may also help to link (a) given measured sample(s) more accurately to a known source location, thus selecting only those samples for source term inversion, where a sole influence of a single emitter under question can be assured. This is an important aspect in the light of future projects planned by CTBTO (i.e., projects dealing with radioxenon source term and background estimates).

Additional real test cases - if proper data is available - should be performed with the ensemble (post-processing) software developed within this project. Transport ranges should preferentially correspond to

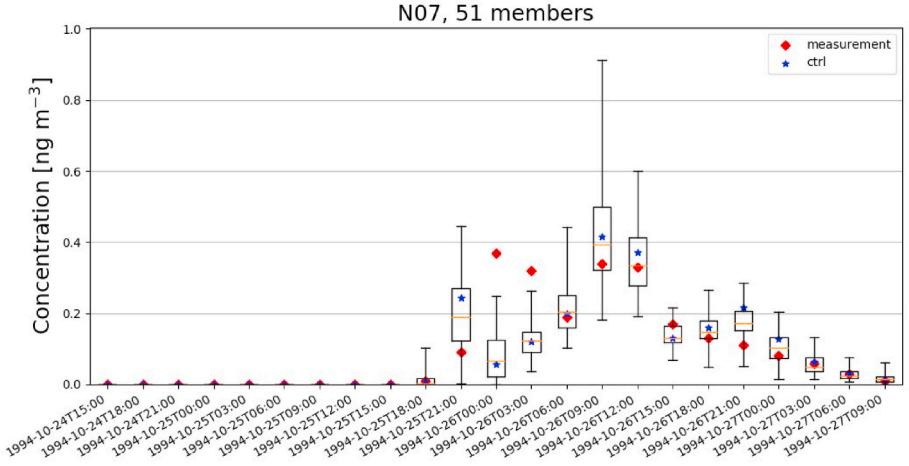


Fig. 19. Box plots of ensemble time series, control run (ctrl) and measurements for ETEX-I sampling site N07 for the full ensemble. x-axis indicates collection stop times.

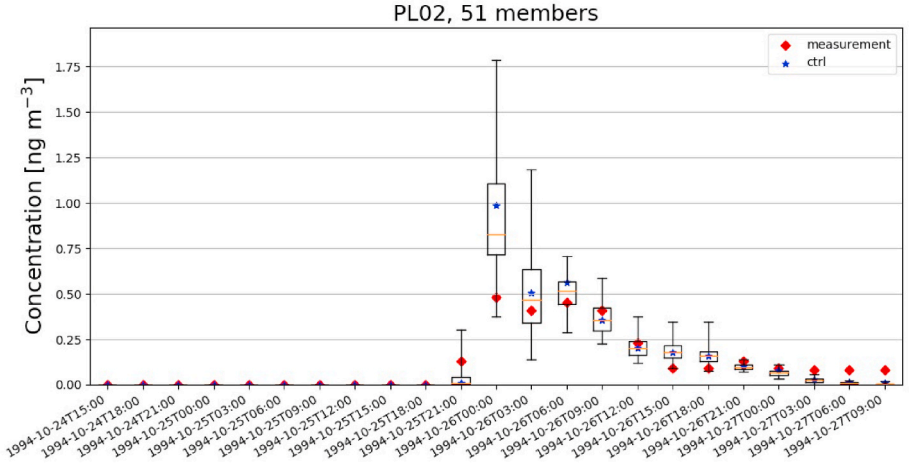


Fig. 20. Box plots of ensemble time series, control run (ctrl) and measurements for ETEX-I sampling site PL02 for the full ensemble. x-axis indicates collection stop times.

those usually considered under CTBT verification tasks. The ensemble hindcast of the ETEX-I case was highly appreciated by the pertinent community. CTBTO and ECMWF agreed on a procedure to keep these data accessible, i.e. extractable from ECMWF's MARS archive (CLASS = AT, TYPE = PF, STREAM = ENFO, EXPVER = b03h) for any public ECMWF user, on the long term for scientific purposes.

Based on the pertinent literature review performed, on the five test cases evaluated as well as on feedback from the *Joint Expert Group* of *Working Group B* of CTBTO, it was recommended to CTBTO to run the ensemble software created in the frame of the work presented operationally or at least quasi-operationally on an ad-hoc basis. If computational and storage constraints are a limiting factor runs should be performed with 10 ECMWF-EPS members only for the time being.

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 would like to thank ECMWF for giving support in hindcasting the ensembles for the Fukushima and ETEX-I test cases. The work was performed under the CTBTO awarded contract for "Provision of Software Engineering Services for Atmospheric Transport Modelling, Data Acquisition, Processing and Dissemination" under funding from the European Union Council Decision VII.

Appendix A

A.1. Test cases data

Tables A1, A2, A3, A4 and A5 list stations with corresponding samples or sample statistics used in the five test cases described in sections 1, 2 and 3.

For the synthetic test cases, the NCEP based FLEXPART version 8.2.3 run used to generate pseudo-measurements was checked regarding consistency with a FLEXPART version 8.2.3 run based on a pertinent ECMWF-EPS control run. Distinct differences could be found as illustrated in Tables A1 and A2 (5th and 6th column). Concentrations for the selected samples differ at least by one order of magnitude being much lower for the ECMWF based run for the DPRK test case, with 11 samples equal to zero. According to the NCEP driven run for the synthetic DPRK test case especially IMS station XE058 is much more affected one day after the release. Also, averaged over all non-zero predictions the ECMWF-EPS control run based FLEXPART run clearly underestimates concentrations compared to the NCEP driven run (0.26 mBq/m^3 vs. 1.01 mBq/m^3). For the synthetic ANSTO test case the situation is the other way round in terms of magnitudes averaged over all non-zero predictions (1.59 mBq/m^3 vs. 0.44 mBq/m^3). The plume according to the NCEP data driven FLEXPART run evidently gets sucked up by an intensive low pressure system (with a reduced core pressure of 967 hPa according to the Australian Bureau of Meteorology, http://www.bom.gov.au/) south of Australia very quickly and only reappears (at concentrations $\geq$ 1 mBq/m^3) several days later in the surface layer over the Pacific ocean.

A comparison with results from the most recent FLEXPART version 10.4 (7th column) proves that these differences are evidently due to the meteorological input data and their digestion by the atmospheric transport model for both test cases. Results between version 8.2.3 and version 10.4 are only different by a few tenths of mBq/m^3 for the very same meteorological driving data. FLEXPART version 10.4 was officially released in 2019 by the Norwegian Institute for Air Research (NILU) and can be considered a major update of the FLEXPART software (Pisso et al., 2019).

Table A.1
List of selected synthetic Xe-133 samples for the DPRK test case. Columns from left to right: IMS station ID, geographical position, collection start [YYYYMMDD HHMMSS], collection stop [YYYYMMDD HHMMSS], FLEXPART V8 average activity concentration [Bq/m^3] based on NCEP input, FLEXPART V8 average activity concentration [Bq/m^3] based on ECMWF control run input and FLEXPART V10 average activity concentration [Bq/m^3] based on NCEP input.

Station ID	Geograph. position	Collection start	Collection stop	Act. conc. NCEP V8 [Bq/m^3]	Act. conc. ECMWF V8 [Bq/m^3]	Act. conc. NCEP V10 [Bq/m^3]
CNX22	113.30° E, 23.10° N	20181211 150000	20181212 150000	0.161E-02	0.0	0.165E-02
CNX22	113.30° E, 23.10° N	20181212 150000	20181213 150000	0.527E-02	0.0	0.520E-02
CNX22	113.30° E, 23.10° N	20181213 150000	20181214 150000	0.149E-02	0.0	0.139E-02
JPX38	139.08° E, 36.30° N	20181204 180000	20181205 060000	0.144E-02	0.0	0.145E-02
JPX38	139.08° E, 36.30° N	20181209 180000	20181210 060000	0.815E-02	0.0	0.822E-02
JPX38	139.08° E, 36.30° N	20181210 060000	20181210 180000	0.527E-02	0.320E-04	0.552E-02
JPX38	139.08° E, 36.30° N	20181210 180000	20181211 060000	0.360E-02	0.594E-05	0.408E-02
JPX38	139.08° E, 36.30° N	20181213 060000	20181213 180000	0.136E-02	0.577E-06	0.139E-02
XE058	132.00° E, 44.15° N	20181201 120000	20181202 000000	0.718E-02	0.0	0.718E-02
XE058	132.00° E, 44.15° N	20181202 000000	20181202 120000	0.144E-01	0.0	0.145E-01
XE058	132.00° E, 44.15° N	20181203 000000	20181203 120000	0.385E-02	0.0	0.391E-02
XE058	132.00° E, 44.15° N	20181203 120000	20181204 000000	0.848E-02	0.0	0.855E-02
XE058	132.00° E, 44.15° N	20181212 000000	20181212 120000	0.136E-02	0.732E-05	0.154E-02
XE058	132.00° E, 44.15° N	20181212 120000	20181213 000000	0.166E-02	0.191E-04	0.196E-02
XE058	132.00° E, 44.15° N	20181213 000000	20181213 120000	0.228E-02	0.147E-05	0.246E-02
XE060	158.78° E, 53.05° N	20181206 000000	20181206 120000	0.132E-01	0.550E-06	0.131E-01
XE060	158.78° E, 53.05° N	20181206 120000	20181207 000000	0.345E-02	0.663E-05	0.357E-02
XE060	158.78° E, 53.05° N	20181209 120000	20181210 000000	0.184E-02	0.276E-04	0.177E-02
XE060	158.78° E, 53.05° N	20181210 000000	20181210 120000	0.142E-02	0.474E-06	0.126E-02
USX74	99.77° W, 37.17° N	20181213 000000	20181213 120000	0.256E-02	0.814E-05	0.265E-02
USX74	99.77° W, 37.17° N	20181213 120000	20181214 000000	0.547E-02	0.694E-05	0.528E-02
USX74	99.77° W, 37.17° N	20181214 000000	20181214 120000	0.230E-02	0.0	0.222E-02
USX74	99.77° W, 37.17° N	20181214 120000	20181215 000000	0.254E-02	0.602E-07	0.253E-02
USX75	78.40° W, 38.00° N	20181211 120000	20181212 000000	0.126E-02	0.0	0.123E-02
USX75	78.40° W, 38.00° N	20181212 000000	20181212 120000	0.350E-02	0.125E-03	0.320E-02
USX75	78.40° W, 38.00° N	20181212 120000	20181213 000000	0.146E-02	0.317E-04	0.147E-02
USX79	158.00° W, 21.52° N	20181213 030000	20181213 150000	0.149E-02	0.300E-05	0.152E-02
USX79	158.00° W, 21.52° N	20181213 150000	20181214 030000	0.140E-02	0.644E-04	0.143E-02

Table A.2
List of selected synthetic Xe-133 samples for the ANSTO test case. Columns from left to right: IMS station ID, geographical position, collection start [YYYYMMDD HHMMSS], collection stop [YYYYMMDD HHMMSS], FLEXPART V8 average activity concentration [Bq/m^3] based on NCEP input, FLEXPART V8 average activity concentration [Bq/m^3] based on ECMWF control run input and FLEXPART V10 average activity concentration [Bq/m^3] based on NCEP input.

Station ID	Geograph. position	Collection start	Collection stop	Act. conc. NCEP V8 [Bq/m^3]	Act. conc. ECMWF V8 [Bq/m^3]	Act. conc. NCEP V10 [Bq/m^3]
XE001	58.47° W, 34.54° S	20181213 000000	20181213 120000	0.114E-02	0.171E-04	0.115E-02
XE001	58.47° W, 34.54° S	20181213 120000	20181214 000000	0.311E-02	0.656E-04	0.287E-02
XE001	58.47° W, 34.54° S	20181214 000000	20181214 120000	0.220E-02	0.902E-06	0.204E-02
CLX19	109.35° W, 27.13° S	20181208 090000	20181209 090000	0.171E-02	0.201E-03	0.182E-02
CLX19	109.35° W, 27.13° S	20181209 090000	20181210 090000	0.743E-02	0.135E-04	0.733E-02
CLX19	109.35° W, 27.13° S	20181210 090000	20181211 090000	0.845E-02	0.156E-04	0.836E-02
CLX19	109.35° W, 27.13° S	20181213 090000	20181214 090000	0.161E-02	0.403E-04	0.159E-02

Table A.3
Selected CTBTO IMS sites used for the Fukushima test case and their statistical parameters (mean and maximum of Xe-133 and Cs-137 activity concentrations). Number of valid samples refers to the time period March, 12th, 00 UTC to 25th, 21 UTC, 2011, where modeled samples lay above zero.

Station	ID	Geograph. position	# valid samples	Mean act. conc. [mBq/m^3]	Max. act. conc. [mBq/m^3]
Yellowknife (Canada)	CAX16	114.47° W, 62.48° N	5	640	2179
St. John's (Canada)	CAP17	52.74° W, 47.59° N	13	0.0004	0.0032
Stockholm (Sweden)	SEX63/SEP63	17.95° E, 59.41° N	21/13	229/0.0011	3306/0.0110
Ashland (US)	USX74/USP74	99.77° W, 37.17° N	14/12	3409/0.0164	16380/0.0560
Charlottesville (US)	USX75/USP75	78.40° W, 38.00° N	23/13	839/0.0461	3775/0.4300
Wake Island (US)	USX77/USP77	166.61° E, 19.29° N	7/12	478/0.0425	1330/0.2300
Ohau (US)	USX79/USP79	158.00° W, 21.52° N	14/12	4891/0.5417	11740/3.1000

Table A.4
Table of IRSN sites with measured and modified Se-75 activity concentrations used in the evaluation of the Se-75 test case. Data basis as of January 2020. 6th column lists the scaling factors which were applied to original activity concentrations (4th column) to accomodate them to adapted collections starts (5th column).

Station	Geograph. position	Original collection period	Original act. conc. ± uncertainty [mBq/m^3]	Modified collection start	Scaling factor	Modified act. conc. [mBq/m^3]
Villeneuve d'Ascq	3.14° E, 50.62° N	20190514 00:00 UTC - 20190517 00:00 UTC	0.096 ± 0.009	20190515 19:00 UTC	72/29	0.238
Villeneuve d'Ascq	3.14° E, 50.62° N	20190517 00:00 UTC - 20190521 00:00 UTC	< MDC	20190517 00:00 UTC	0	0.0
Maubeuge	3.92° E, 50.27° N	20190513 00:00 UTC - 20190516 00:00 UTC	< MDC	20190515 19:00 UTC	0	0.0
Penly	1.21° E, 49.98° N	20190513 00:00 UTC - 20190520 00:00 UTC	0.0053 ± 0.0007	20190516 00:00 UTC	168/96	0.0093
Paluel	0.63° E, 49.86° N	20190513 00:00 UTC - 20190520 00:00 UTC	0.0027 ± 0.0004	20190516 01:00 UTC	168/95	0.0048
Omonville-la-Petite	1.88° W, 49.70° N	20190513 00:00 UTC - 20190520 00:00 UTC	0.0011 ± 0.0002	20190516 05:00 UTC	168/91	0.002
Flamanville	1.87° W, 49.55° N	20190513 00:00 UTC - 20190520 00:00 UTC	< MDC	20190516 05:00 UTC	0	0.0

Table A.5
Selected sites used for the ETEX-I test case and their statistical parameters (including mean and maximum concentrations of PMCH).

Station	ID	Geograph. position	# valid samples	Mean conc. [ng/m^3]	Max. conc. [ng/m^3]
Feuerkogel	A02	13.73° E, 47.82° N	24	0.02	0.07
Cervena	CR01	17.55° E, 49.77° N	23	0.15	0.5
Aalborg Airport	DK01	9.87° E, 57.10° N	16	0.48	3.04
Hvide Sande	DK05	8.13° E, 56.00° N	24	0.26	2.01
Jaegersborg	DK06	12.53° E, 55.77° N	22	0.37	4.34
Skrydstrup Airport	DK10	9.27° E, 55.23° N	21	0.26	1.55
Alencon	F02	0.10° E, 48.45° N	22	0.39	3.23
Budapest/Lorinc	H01	19.18° E, 47.43° N	20	0.21	1.2
Gyor	H02	17.68° E, 47.70° N	23	0.07	0.52
Stavanger/Sola	N07	5.63° E, 58.88° N	23	0.10	0.37
Kielce	PL02	20.70° E, 50.82° N	23	0.12	0.48
Klodzko	PL03	16.65° E, 50.43° N	23	0.13	0.72
Zielona Gora	PL08	15.53° E, 51.93° N	23	0.45	1.97
Jaslovske Bohunice	SR01	17.67° E, 48.48° N	22	0.26	0.82
Lucenec	SR03	19.77° E, 48.33° N	23	0.08	0.24

A.2. Metrics definition

The *normalized variance for FOR fields* with a cut-off threshold (1E-18 Bq/m^3) for the underlying concentrations and a logarithmic color scale was introduced for the purpose of displaying uncertainty related to meteorological ECMWF input. At every grid point and for every time step and with $i = 1…N$ ensemble members the normalized variance norm_var is defined as the quotient of the variance and the squared average:

$$\text{norm\_var} = \frac{\frac{1}{N}\sum_{i=1}^{N}\left(\text{cmod}_i - \overline{\text{cmod}}\right)^2}{\overline{\text{cmod}}^2} \tag{A.1}$$

with cmod_i being the individual ensemble member concentrations and $\overline{\text{cmod}}$ being the average ensemble concentration. The grid point variance is normalized with the square of the mean grid box value since plotting non-normalized values may be misleading. E.g., a 5% variation in absolute numbers around the mean value will be very different for 0.1 compared to 10 Bq/m^3. Also, due the variable range of possible normalized variance values depending on the ensemble size N (maximum possible value equal to N-1) qualitative color bar labels are used instead of quantitative ones and thus it is only possible to extract qualitative information on meteorological uncertainty at every grid point and for every time step.

At every grid point and for every time step and for N ensemble members the *probability of exceedance* prob_ex(T), sometimes also called *Agreement in Threshold Level* (Galmarini et al., 2004), is defined as percentage of ensemble member sensitivities (or correlations) SRS_i (or R_i) exceeding a predefined value T:

$$\text{prob\_ex}(T) = \frac{\sum_{i=1}^{N} n_T}{N} * 100\% \tag{A.2}$$

with $n_T = |\{x \in \text{SRS} : x > T\}|$

The higher the percentage value the higher is the likelihood that at least a defined reference SRS (or R) value is exceeded.

The *Brier score* (BS) (e.g., used by Galmarini et al. (2010)) for a given decision threshold is defined as the mean square error of a probability forecast:

$$\text{BS} = 1 / N \sum_{i=1}^{N} (F_i - O_i)^2 \tag{A.3}$$

where N is the number of forecasts, F_i is the forecast probability on occasion i and O_i is the observation (0 or 1) on occasion i. The score weights larger errors more than smaller ones. In the context of the present paper the Brier score reflects the ability of the ensemble to correctly forecast the probability of above or below MDC samples.

The spatial coverage of a simulated cloud compared with a monitored one (or another simulated one) is well represented by the *Figure of Merit in Space* (FMS) or *Figure of Overlap* (FO) (see, e.g., Galmarini et al. (2010) and Kioutsioukis and Galmarini (2014)) defined as:

$$\text{FMS} = 100\% * |\{x : M(x) > T \cap O(x) > T\}| \,/\, |\{x : M(x) > T \cup O(x) > T\}| \tag{A.4}$$

for some threshold T, where $M(x)$ and $O(x)$ are predicted and observed (or some other predicted) values, respectively, at point x. T was set to 1E-18 Bq/m^3.

The *accuracy* ACC (see, e.g., Galmarini et al. (2010) and Kioutsioukis and Galmarini (2014)) is defined as:

$$\text{ACC} = 100\% * |\{t : M(t) \geq T | O(t) \geq T \cup M(t) < T | O(t) < T\}| \,/\, |\{t : O(t)\}| \tag{A.5}$$

for some threshold T, where $M(t)$ and $O(t)$ are predicted and observed (or some other predicted) values, respectively, at time step t.

The *fractional bias* (FB) (used inter alia by Solazzo and Galmarini (2015)) is defined as:

$$\text{FB} = 2 * \frac{\overline{O} - \overline{M}}{\overline{O} + \overline{M}} \tag{A.6}$$

$\overline{M}$ and $\overline{O}$ denote the corresponding averages over time or space (depending on the application).

The *Kolmogorov-Smirnov parameter* (KSP) is defined as:

$$\text{KSP} = Max|D(M_k) - D(O_k)| * 100\% \tag{A.7}$$

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

In order to partially mitigate the deficiencies of independent statistical measures, an approach whereby a set of statistical scores are combined into one rank measure can be taken. Hegarthy et al. (2013) used the rank defined in equation A.8 for comparing two-dimensional fields:

$$\text{Rank} = R^2 + 1 - |\text{FB} / 2| + \text{FMS} / 100 + (1 - \text{KSP} / 100) \tag{A.8}$$

Becker et al. (2007) combined the explained variance R^2, FB and FMS into a rank measure (however, the KSP may have been skipped for an improper reason; personal communication). For the calculation of the case specific so-called *Agreement of Model p* with all others, cAgreement_p, from the case specific rank value of model p, cRNKp, these authors prefer to give the percentage of the maximum cRNK value (3.0) while excluding the

trivial auto-correlation result as follows:

$$\mathrm{cAgreement_p} = \frac{100\%}{3(N-1)} \sum_{i=1}^{N} \varepsilon_{ip}\mathrm{cRNK_p} \tag{A.9}$$

with N being the total number of experiments' participants and $\varepsilon_{ip} = 1$ for $i \neq p$ and 0 for $i = p$. For the purpose of the present research the rank was amended by the Kolmogorov-Smirnov parameter and used as defined in equation A.8.

The belonging case and model specific anomaly, $\mathrm{cAnomaly_p}$, to the case specific across participants' *Average Agreement* (cAV) is calculated as:

$$\mathrm{cAnomaly_p} = \mathrm{cAgreement_p} - \mathrm{cAV} \tag{A.10}$$

with

$$\mathrm{cAV} = \frac{1}{N} \sum_{p=1}^{N} \mathrm{cAgreement_p} \tag{A.11}$$

The Average Agreement tends to be larger than the Figure of Overlap for a given time step, since the former is based on a rank measure which comprises R^2, the Kolmogorov-Smirnov parameter and the fractional bias in addition to the space sensitive Figure of Overlap. The Kolmogorov-Smirnov parameter in contrast to the other metrics is not at all sensitive to spatial disagreements and can thus positively affect the Average Agreement even in the case of an unsatisfactory Figure of Overlap.

For time series analysis the Figure of Merit in Space in equation A.8 was replaced by the accuracy (equation A.5) as in Maurer et al. (2018). T was set to either the MDC for the Fukushima test case or to zero (background had already been subtracted by the data providers) for the ETEX-I test case. This is motivated by the fact that the ability to discriminate between above and below MDC values is of high importance for CTBT verification.

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.

Arnold, D., Maurer, C., Wotawa, G., Draxler, R., Saito, K., Seibert, P., 2015. Influence of the meteorological input on the atmospheric transport modelling with FLEXPART of radionuclides from the Fukushima Daiichi nuclear accident. J. Environ. Radioact. 139, 212–225. https://doi.org/10.1016/j.jenvrad.2014.02.013.

Bauer, P., Thorpe, A., Brunet, G., 2015. The quiet revolution of numerical weather prediction. Nature 525, 47–55.

Becker, A., Wotawa, G., De Geer, L.E., Seibert, P., Draxler, R.R., Sloan, C., D'Amours, R., Hort, M., Glaab, H., Heinrich, P., et al., 2007. Global backtracking of anthropogenic radionuclides by means of a receptor oriented ensemble dispersion modelling system in support of Nuclear-Test-Ban Treaty verification. Atmos. Environ. 41, 4520–4534.

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, 192–200. https://doi.org/10.1016/j.jenvrad.2012.07.018.

Buizza, R., Palmer, T.N., 1998. Impact of ensemble size on ensemble prediction. Mon. Weather Rev. 126, 2503–2518.

CMC - ECCC, 2019. 1-h average Se-75 near-surface concentrations. https://eer.cmc.ec.gc.ca/people/Alain/eer/case_studies/e5G8JlWc4N8X18SwpKuy76hs56CB/SCK-CEN_Se-75_Emission/anim_CV/Animation.html. (Accessed 15 July 2020).

CTBT, 1996. Text of the Comprehensive Nuclear-Test-Ban Treaty. http://www.ctbto.org/the-treaty/treaty-text/. (Accessed 24 January 2017).

CTBTO, 2016. WEB-GRAPE 1.8.2. Technical Report. CTBTO Preparatory Commission, International Data Center (IDC).

CTBTO, 2020a. International Monitoring System. https://www.ctbto.org/map/. (Accessed 15 July 2020).

CTBTO, 2020b. VDEC request for access. https://www.ctbto.org/specials/vdec/vdec-request-for-access/. (Accessed 15 July 2020).

CTBTO Preparatory Comission, 2019. Verification Regime. Technical report. http://www.ctbto.org/verification-regime/. (Accessed 27 February 2019).

Czyz, S.A., Farsoni, A.T., Ranjbar, L., 2018. A prototype detection system for atmospheric monitoring of xenon radioisotopes. Nucl. Instrum. Methods A. 884, 64–69. https://doi.org/10.1016/j.nima.2017.10.044.

De Meutter, P., Camps, J., Delcloo, A., Deconninck, B., Termonia, P., 2016. On the capability to model the background and its uncertainty of CTBT-relevant radioxenon isotopes in Europe by using ensemble dispersion modeling. J. Environ. Radioact. 164, 280–290. https://doi.org/10.1016/j.jenvrad.2016.07.033.

De Meutter, P., Camps, J., Delcloo, A., Termonia, P., 2018. Source localisation and its uncertainty quantification after the third DPRK nuclear test. Sci. Rep.-UK 8, 1–10.

ECMWF, 1994. 1994 summary of changes. https://www.ecmwf.int/en/forecasts/documentation-and-support/evolution-ifs/cycle-archived/1994-summary-changes. (Accessed 15 July 2020).

ECMWF, 2010. Summary of cycle 36r4. https://www.ecmwf.int/en/forecasts/documentation-and-support/evolution-ifs/cycles/cycle-36r4-summary-changes. (Accessed 15 July 2020).

ECMWF, 2018a. ECMWF official homepage. https://www.ecmwf.int/. (Accessed 27 June 2020).

ECMWF, 2018b. Summary of cycle 45r1. https://www.ecmwf.int/en/forecasts/documentation-and-support/evolution-ifs/cycles/summary-cycle-45r1. (Accessed 15 July 2020).

ECMWF, 2019. Summary of cycle 46r1. https://www.ecmwf.int/en/forecasts/documentation/evolution-ifs/cycles/summary-cycle-46r1. (Accessed 15 July 2020).

ECMWF, 2020a. ERA5 documentation. https://confluence.ecmwf.int/display/CKB/ERA5%3A+data+documentation. (Accessed 15 July 2020).

ECMWF, 2020b. MARS-ECMWF's meteorological archive. https://www.ecmwf.int/assets/elearning/mars/mars1/story_html5.html. (Accessed 15 July 2020).

European Commission - Joint Research Center, 2020. ENSEMBLE: model evaluation and ensemble analysis. http://ensemble.jrc.ec.europa.eu/. (Accessed 14 July 2020).

Evans, J.P., Ekström, M., Ji, F., 2012. Evaluating the performance of a WRF physics ensemble over South-East Australia. Clim. Dynam. 39, 1241–1258. https://doi.org/10.1007/s00382-011-1244-5.

Ferranti, L., Corti, S., 2011. New clustering products. ECMWF newsletter number 127. https://doi.org/10.21957/lr3bcise. https://www.ecmwf.int/en/about/news-centre/media-resources.

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.

Galmarini, S., Bonnardot, F., Jones, A., Potempski, S., Robertson, L., Martet, M., 2010. Multi-model vs. EPS-based ensemble atmospheric dispersion simulations: a quantitative assessment on the ETEX-1 tracer experiment case. Atmos. Environ. 44, 3558–3567.

Galmarini, S., Kioutsioukis, I., Solazzo, E., Alyuz, U., Balzarini, A., Bellasio, R., Benedictow, A.M.K., Bianconi, R., Bieser, J., Brandt, J., Christensen, J.H., Colette, A., Curci, G., Davila, Y., Dong, X., Flemming, J., Francis, X., Fraser, A., Fu, J., Henze, D. K., Hogrefe, C., Im, U., Garcia Vivanco, M., Jiménez-Guerrero, P., Jonson, J.E., Kitwiroon, N., Manders, A., Mathur, R., Palacios-Peña, L., Pirovano, G., Pozzoli, L., Prank, M., Schultz, M., Sokhi, R.S., Sudo, K., Tuccella, P., Takemura, T., Sekiya, T., Unal, A., 2018. Two-scale multi-model ensemble: is a hybrid ensemble of opportunity telling us more? Atmos. Chem. Phys. 18, 8727–8744 doi:10.5194/acp-18-8727-2018. https://www.atmos-chem-phys.net/18/8727/2018/.

Graziani, G., Klug, W., Mosca, S., 1998. Real-time Long-Range Dispersion Model Evalution of the ETEX First Release. Luxembourg Office for Official Publications of the European Communities.

Hegarthy, J., Draxler, R.R., Stein, A.F., Brioude, J., Mountain, M., Eluszkiewicz, J., Nehrkorn, T., Ngan, F., Andrews, A., 2013. Evaluation of Lagrangian particle dispersion models with measurements from controlled tracer releases. J. Appl. Meteorol. Clim. 52, 2623–2637.

IRSN, 2019. Information report from IRSN following the incident at the SCK-CEN facilities in Mol (Belgium). Technical Report. https://www.irsn.fr/EN/newsroom/News/Pages/20190528_Information-report-incident-SCK-CEN-facilities-in-Mol-Belgium.aspx. (Accessed 8 May 2020).

Isaksen, L., Bonavita, M., Buizza, R., Fisher, M., Haseler, J., Leutbecher, M., Raynaud, L., 2010. Ensemble of data assimilations at ECMWF. Technical Report. https://www.ecmwf.int/en/elibrary/10125-ensemble-data-assimilations-ecmwf. (Accessed 5 August 2020).

Kalinowski, M.B., Grosch, M., Hebel, S., 2014. Global xenon-133 emission inventory caused by medical isotope production and derived from the worldwide Technetium-99m demand. Pure Appl. Geophys. 171, 707–716. https://doi.org/10.1007/s00024-0.13-0687-5.

Kioutsioukis, I., Galmarini, S., 2014. De praeceptis ferendis: good practice in multi-model ensembles. Atmos. Chem. Phys. 14, 11791–11815.
Klonner, R., 2013. Clustering ECMWF ENS Ensemble Predictions to Optimise FLEXPART Plume Dispersion Ensembles. Ph.D. thesis. University of Vienna.
Lang, S., Holm, E., Bonavita, M., Tremolet, Y., 2019. A 50-member ensemble of data assimilations. ECMWF Newsletter Number 158. https://www.ecmwf.int/node/18821. (Accessed 5 August 2020).
Leutbecher, M., Palmer, T.N., 2008. Ensemble forecasting. J. Comput. Phys. 227, 3515–3539.
Malardel, S., Wedi, N., Deconinck, W., Diamantakis, M., Kühnlein, C., Mozdzynski, G., Hamrud, M., Smolarkiewicz, P., 2016. A new grid for the IFS. ECMWF Newsletter 146, pp. 23–28. https://www.ecmwf.int/sites/default/files/elibrary/2016/17262-new-grid-ifs.pdf. (Accessed 5 August 2020).
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., Baré, J., Kusmierczyk-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., Quèlo, 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.
NOAA, 2020a. GFS/GDAS CHANGES SINCE 1991 - history of recent modifications to the global forecast/analysis system. https://www.emc.ncep.noaa.gov/gmb/STATS/html/model_changes.html. (Accessed 5 August 2020).
NOAA, 2020b. National Centers for Environmental Information: Global Forecast System. https://www.ncdc.noaa.gov/data-access/model-data/model-datasets/global-forcast-system-gfs. (Accessed 13 May 2020).
Palmer, T.N., Molteni, F., Mureau, R., Buizza, R., Chapelet, P., Tribbia, J., 1993. Ensemble prediction. In: Proceedings of the ECMWF Seminar on Validation of Models over Europe. Shinfield Parks, Reading, UK online; accessed July, 14th, 2020. https://www.ecmwf.int/sites/default/files/elibrary/1992/11488-ensemble-prediction.pdf.
Perkins, R., Casey, L., 1996. Radioxenons: Their Role in Monitoring a Comprehensive Test-Ban Treaty. Technical Report. Pac. Northwest Natl. Lab., Richland, Washington, USA.
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.
Potempski, S., Galmarini, S., Addis, R., Astrup, P., Bader, S., Bellasio, R., Bianconi, R., Bonnardot, F., Buckley, R., D'Amours, R., van Dijk, A., Geertsema, G., Jones, A., Kaufmann, P., Pechinger, U., Persson, C., Polreich, E., Prodanova, M., Robertson, L., Sørensen, J., Syrakov, D., 2008. Multi-model ensemble analysis of the ETEX-2 experiment. Atmos. Environ. 42, 7250–7265.
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.
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.
Shemyakin, V., Haario, H., 2018. Online identification of large-scale chaotic system. Nonlinear Dynam. 93, 961–975.
Solazzo, E., Galmarini, S., 2015. The Fukushima Cs-137 deposition case study: properties of the multi-model ensemble. J. Environ. Radioact. 139, 226–233. https://doi.org/10.1016/j.jenvrad.2014.02.013.
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.
Stohl, A., Seibert, P., Wotawa, G., Arnold, D., Burkhart, J.F., Eckhardt, S., Tapia, C., Vargas, A., Yasunari, T.J., 2012. Xenon-133 and caesium-137 releases into the atmosphere from the Fukushima Dai-ichi nuclear power plant: determination of the source term, atmospheric dispersion, and deposition. Atmos. Chem. Phys. 12, 2313–2343. https://doi.org/10.5194/acp-12-2313-2012.
Straume, A.G., Koffi, E.N., Nodop, K., 1998. Dispersion modeling using ensemble forecasts compared to ETEX measurements. J. Appl. Meteorol. 37, 1444–1456.
Sun, Y., Carrigan, C.R., 2012. Modeling noble gas transport and detection for the Comprehensive Nuclear-Test-Ban Treaty. Pure Appl. Geophys. 171, 735–750. https://doi.org/10.1007/s00024-012-0514-4.
Tipka, A., Haimberger, L., Seibert, P., 2020. Flex_extract v7.1 – a software to retrieve and prepare ECMWF data for use in FLEXPART. Geosci. Model Dev. (GMD) 13, 5277–5310. https://doi.org/10.5194/gmd-13-5277-2020.
UNSCEAR, 2013. Sources, Effects and Risks of Ionizing Radiation – United Nations Scientific Committee on the Effects of Atomic Radiation, vol. I. Scientific Annex A. Report. United Nations.
Vitart, F., Balsamo, G., Bidlot, J.R., Lang, S., Tsonevsky, I., Richardson, D., Alson-Balmaseda, M., 2019. Use of ERA5 to initialize ensemble re-forecasts. Technical report. https://doi.org/10.21957/w8i57wuz6. https://www.ecmwf.int/node/18872.
Wilks, D., 2011. Statistical Methods in the Atmospheric Sciences, 100. Academic Press.
WMO, 2019. Manual on the Global Data-Processing and Forecasting System: Annex IV to the WMO Technical Regulations, 485. WMO.
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.
ZAMG, 2018. FLEXible PARTicle dispersion model - official FLEXPART website. https://www.flexpart.eu/. (Accessed 27 June 2020).