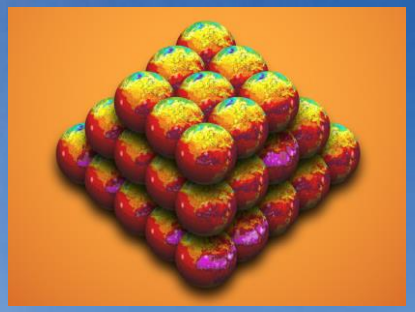


UEF 2019

Quantifying uncertainties and confidence level in ATM simulations – scientific aspects*

C. Maurer[1], D. Arnold[1], M. Haselsteiner[1], J. Brioude[2], L. Haimberger[3], and F. Weidle[1]

[1] Central Institute for Meteorology and Geodynamics (ZAMG), Vienna, Austria

[2] Atmosphere and Cyclone Lab (LACy), University of La Réunion Island, France

[3] Institute for Meteorology and Geophysics (IMGW), University of Vienna, Austria

ZAMG
Zentralanstalt für Meteorologie und Geodynamik

*work 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.

1. Comprehensive Nuclear Test-Ban Treaty (CTBT)

ARTICLE I - BASIC OBLIGATIONS

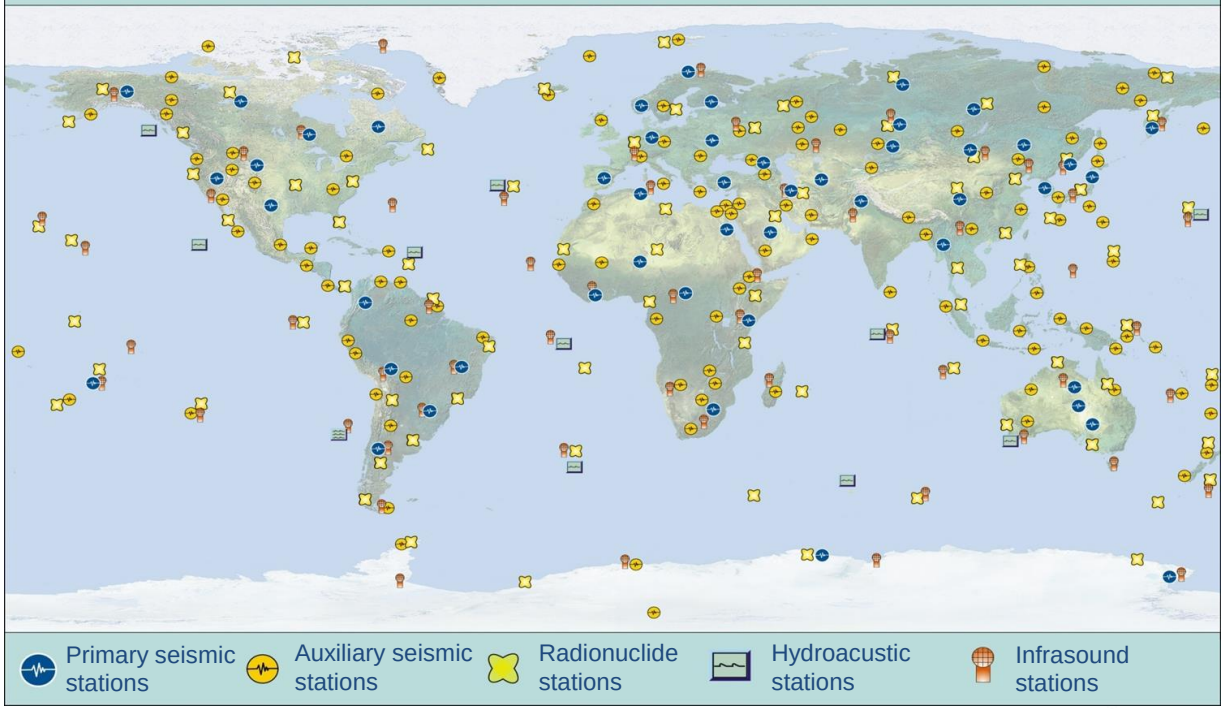
Each State Party undertakes ***not to carry out any nuclear weapon test explosion or any other nuclear explosion****, and to prohibit and prevent any such nuclear explosion at any place under its jurisdiction or control.*

Entry into force: As soon as 44 countries which operate nuclear reactors have ratified the Treaty (not just signed).

The **International Monitoring System (IMS)**

80 radionuclide stations foreseen, whereof **40** will be equipped with **noble gas** measuring devices

>80 % of the network currently installed



2. Aim

"**CTBTO wants to develop tools to estimate uncertainties and confidence levels in (operational) ATM simulations using meteorological data from an Ensemble Prediction System (EPS).** New tools will become part of CTBTO ATM pipeline in the second phase of this project."

Tasks:

1. Perform a scientific literature review and define requirements.
2. Prepare a plan in consultation with the ATM team.
3. Define a set of products to display uncertainties and confidence levels in ATM simulations.
4. Prepare data to estimate uncertainties.
5. Run a prototype to generate products to display uncertainties and confidence levels.
6. Produce a short report including recommendations for the second phase of the project.

ZAMG
Zentralanstalt für Meteorologie und Geodynamik

3.1. Literature review for EPS-based ATM

Main messages:

- **If computationally affordable, ensemble** dispersion **modelling** is the **method of choice** to **take uncertainties of models and measurement data into account** and create probabilistic forecasts.
- Authors tend to **use the full ensemble (51 members - 50 plus control run - in case of ECMWF) or, to reduce resources, draw a (mostly randomly) subset.** However, **accuracy of forecast probabilities** in certain scenarios clearly **increases with** the **ensemble size**.
- **Clustering/finding representative members** of EPS input **for the purpose of ATM** remains a **research topic with no final solution.** So far only special cases have been considered (e.g., high altitude transport of volcanic ash with strong horizontal forcing). Moreover, **clustering is always confined to a certain region** (e.g., Europe).
- **No redundancy has to be expected** due to the design of the NWP ensemble system **– which may not be true for multi-model ensembles.**

ZAMG
Zentralanstalt für Meteorologie und Geodynamik

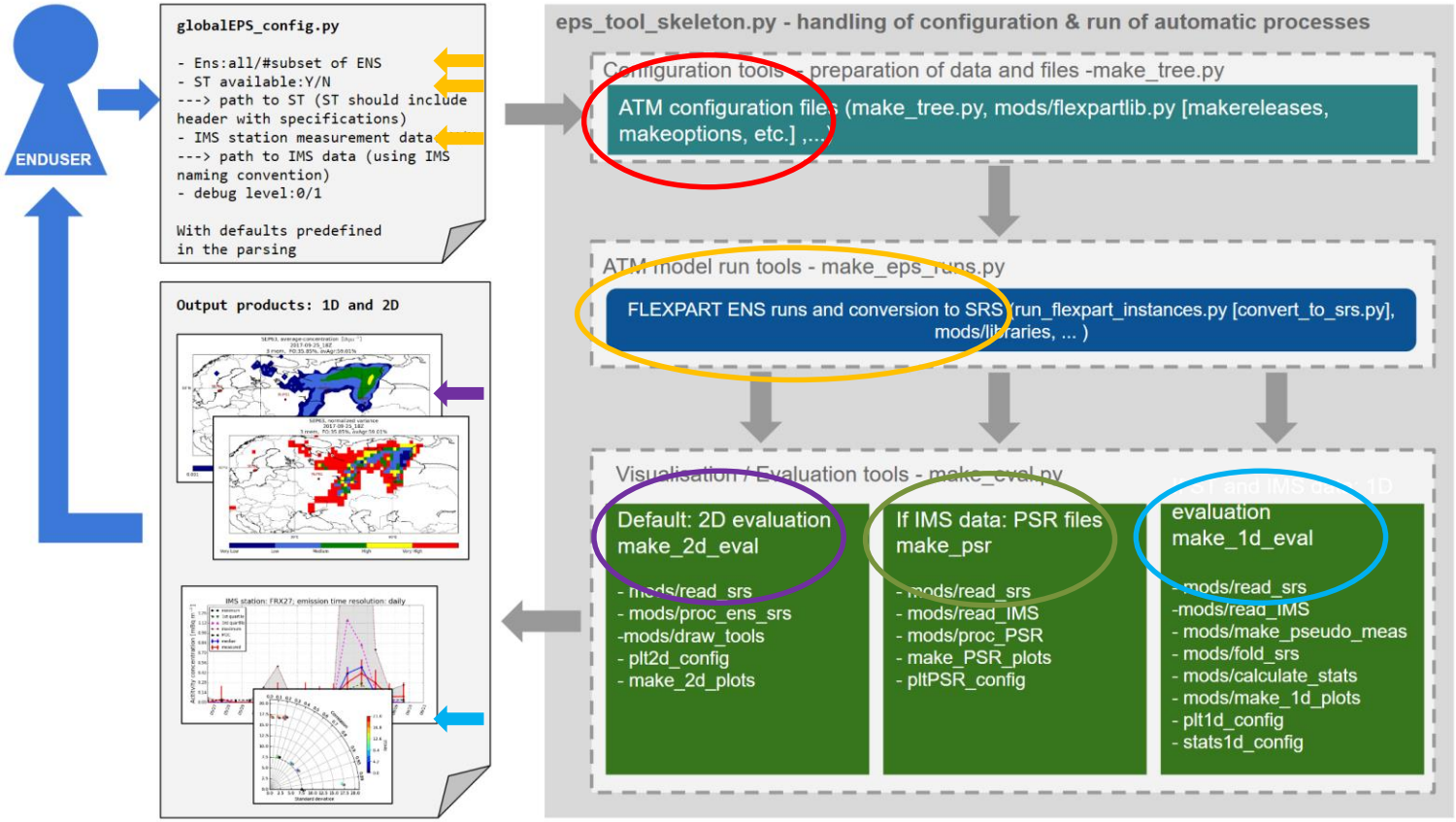
3.2 Literatur review for EPS-based ATM

Display and evaluation:

- ***Agreement in Threshold Level***, ***Agreement in Percentile Level***, ***Overlap Plots***, ***Maximum Concentration Plots*** and ***Box Plots*** are most **common.**
- Many of the **usual statistical scores suffer from** important **drawbacks.** The **main judgment** of the model performance should **not be based on the correlation coefficient.** Also, **statistics** can be **heavily influenced if** some **modelled values are near zero while nuisance sources cause the measured values to be at or just above a detection limit.** To mitigate the deficiencies of independent statistical measures**, a combined Rank measure should be considered.**

ZAMG
Zentralanstalt für Meteorologie und Geodynamik

4. System Prototype design

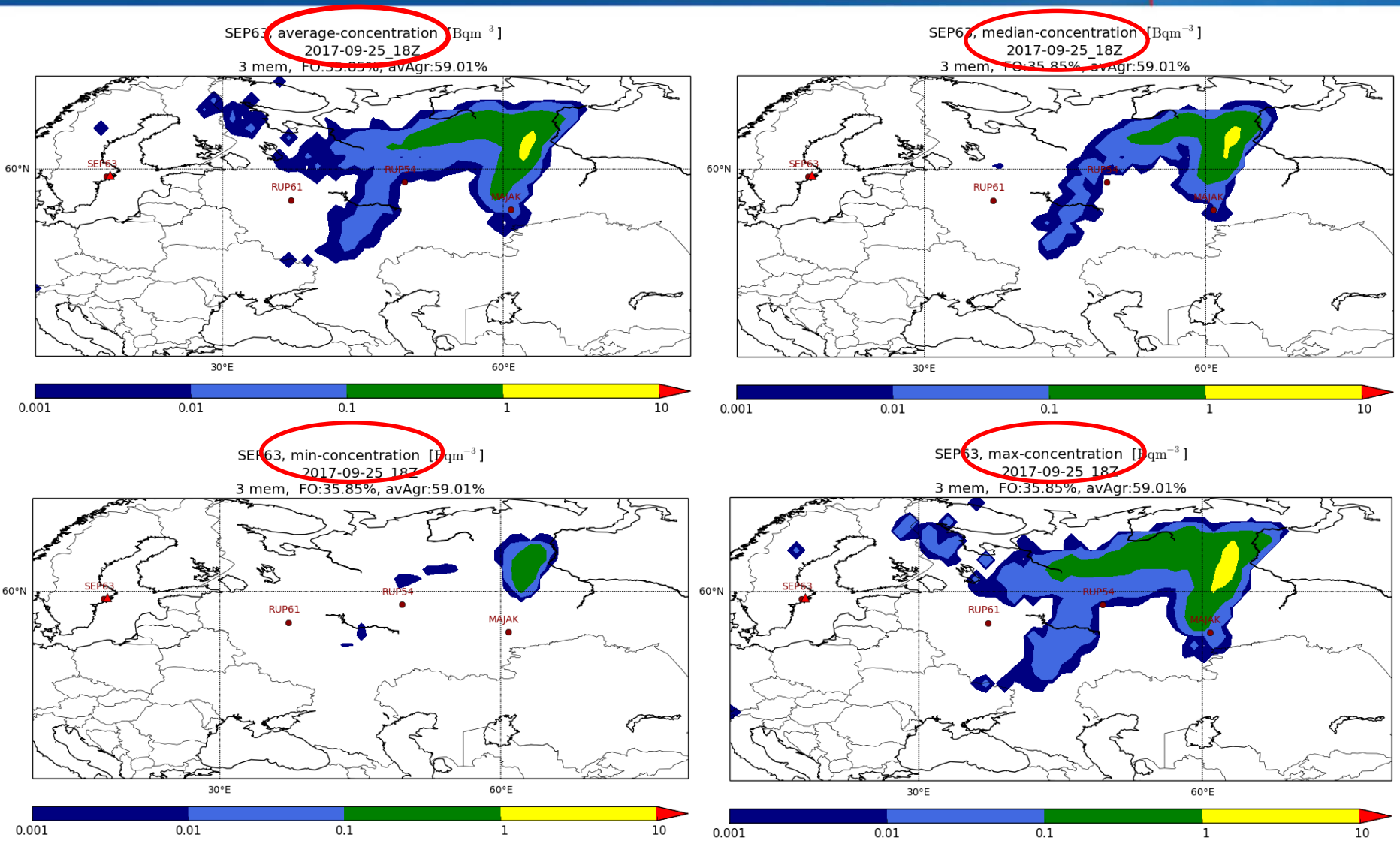


- Namelist file for configuration
- Launch atmospheric transport model FLEXPART runs based on (randomly selected or full) ECMWF-EPS input at selected locations, either the source location (fwd) or any (or all) of the IMS stations (bwd).
- Postprocess FLEXPART results.

- Plot (ensemble) Source Receptor Sensitivity (SRS) fields/Fields of Regard (FOR).
- Calculate and plot the (ensemble) Probable Source Region (PSR)-fields if measurements are available and can be correlated with SRS values for each grid point and time step.
- Fold SRS fields with a source term if both the source term and measurements are available, plot resulting time series and perform statistical evaluation.

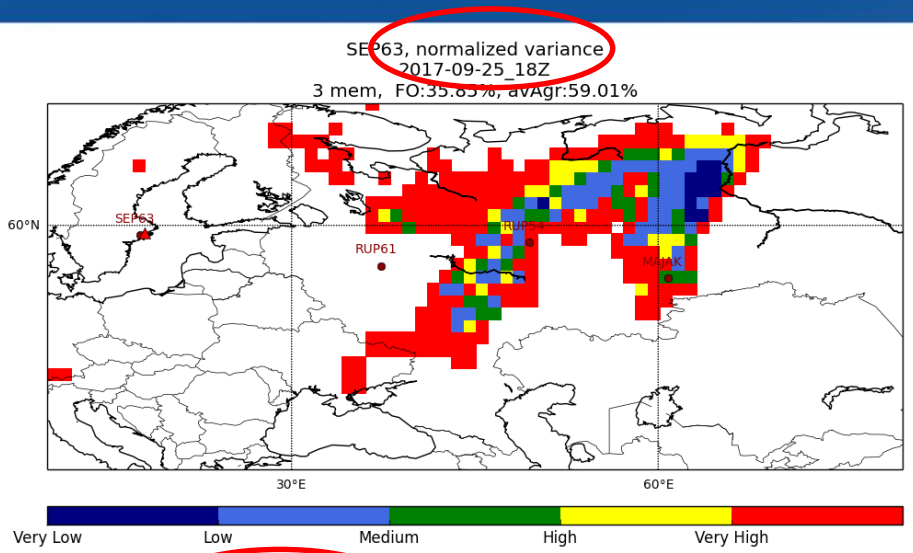
ZAMG
Zentralanstalt für Meteorologie und Geodynamik

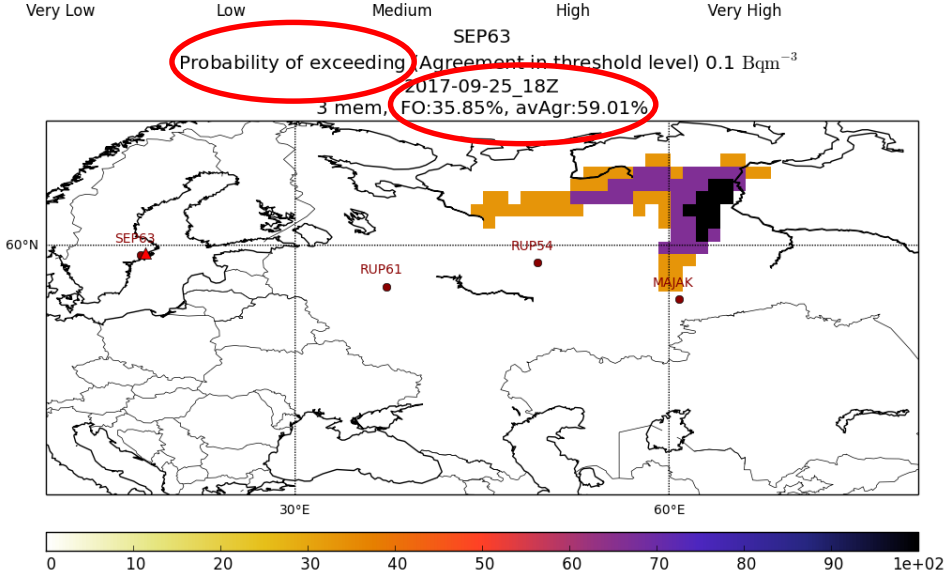
5.1. 2D-Products (SRS)



Prototype design based on the „Ruthenium event in Sept./Oct. 2017

5.2. 2D-Products (SRS)



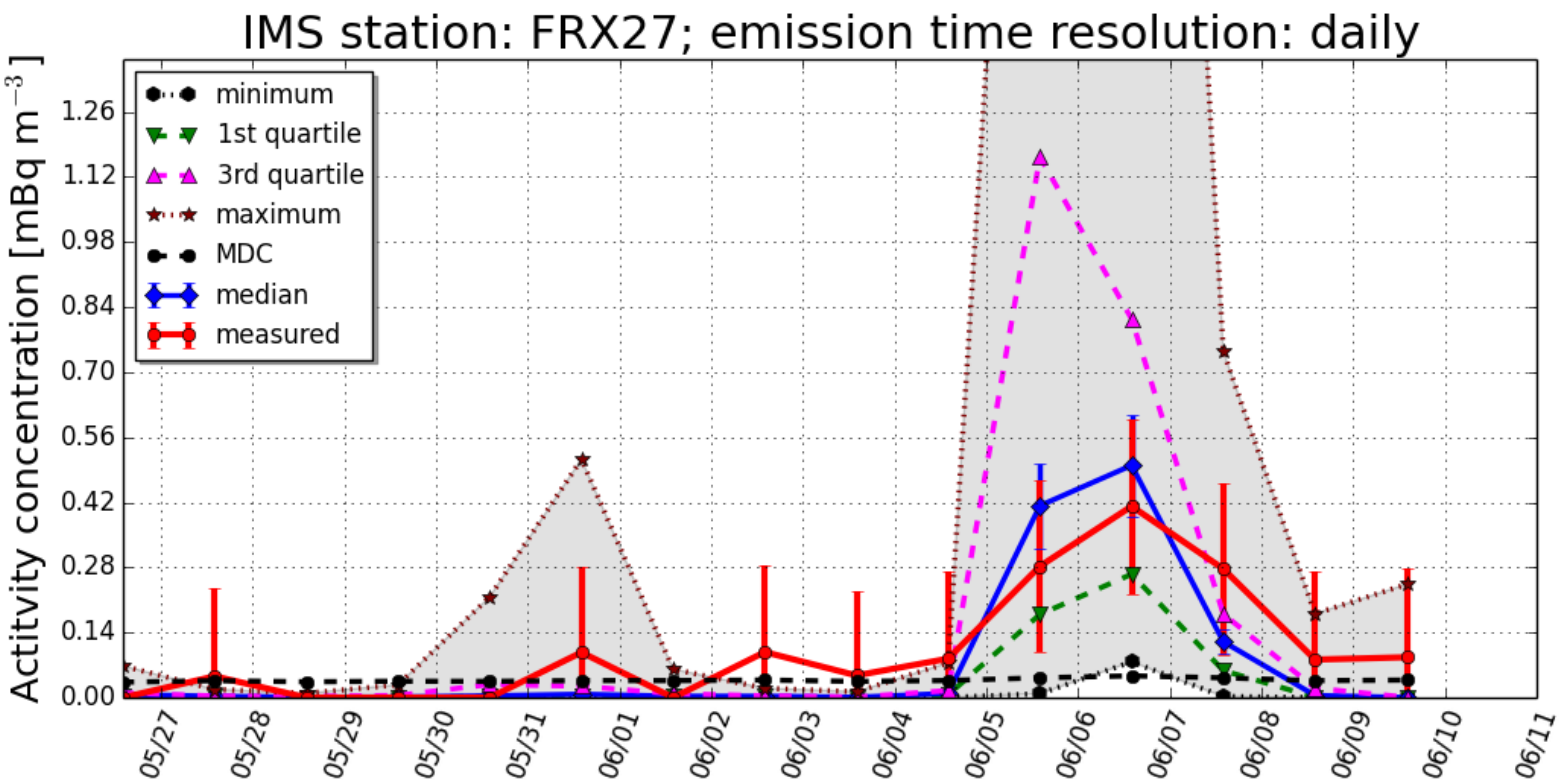


„*Figure of Overlap*“ and „*Average Agreement*“ as indicator of consistency between ensemble members

- Both **median** and **mean** are used according to literature in the context of ensemble dispersion modelling. However, **median is preferred.**
- **Minimum** and **maximum indicate extreme events**, which **may not be visible using** the **median** display.
- The **normalized variance** (normalized with the squared mean value, maximum value N-1) provides a **rough estimate of uncertainty.**
- The **probability of exceedance (Agreement in Threshold Level)** includes a threshold definition in order to **focus on values above a certain level.**

ZAMG
Zentralanstalt für Meteorologie und Geodynamik

6. 1D-Products (concentration time series) & numerical statistics

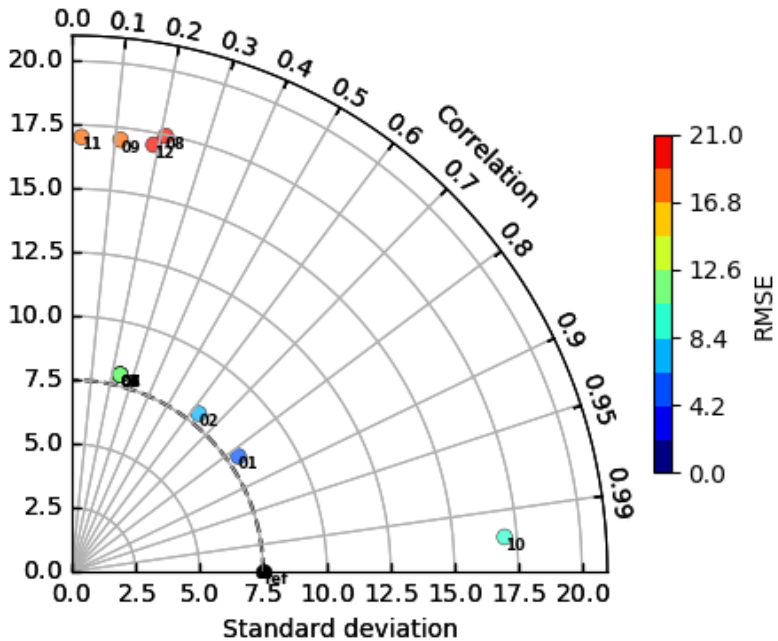


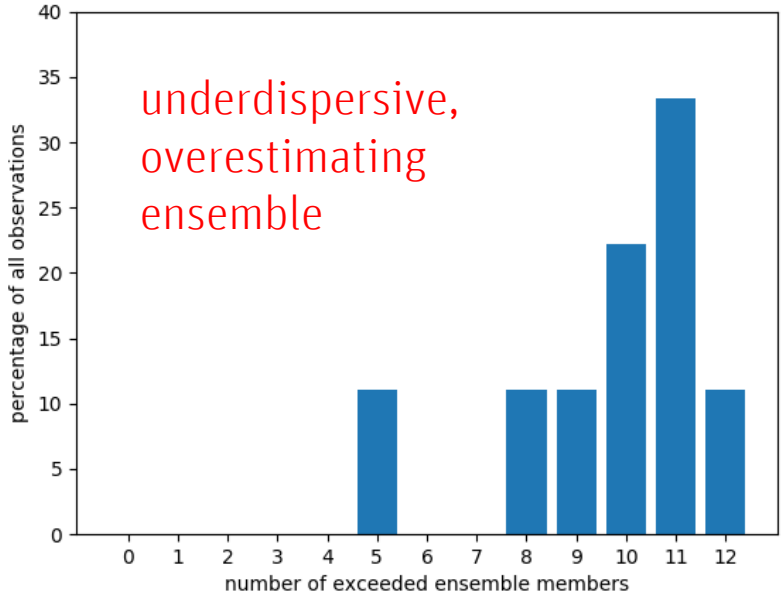
Statistics:

- **R^2**, the **Fractional Bias**, the **Kolmogorov-Smirnov Parameter and** the **Accuracy** (based on MDC) will be given individually and **combined into a Rank measure.**
- **Brier Score** based on MDC will be calculated.
- An alternative **skill score S** (personal communication with P. Seibert) will be tested.

ZAMG
Zentralanstalt für Meteorologie und Geodynamik

7. Graphical statistical evaluation





Taylor Diagram: Data points represent the time series of a member or an ensemble statistic at a measurement station.

Talagrand Diagram: Evaluating the structure of the ensemble. Ideally one wants a flat diagram.

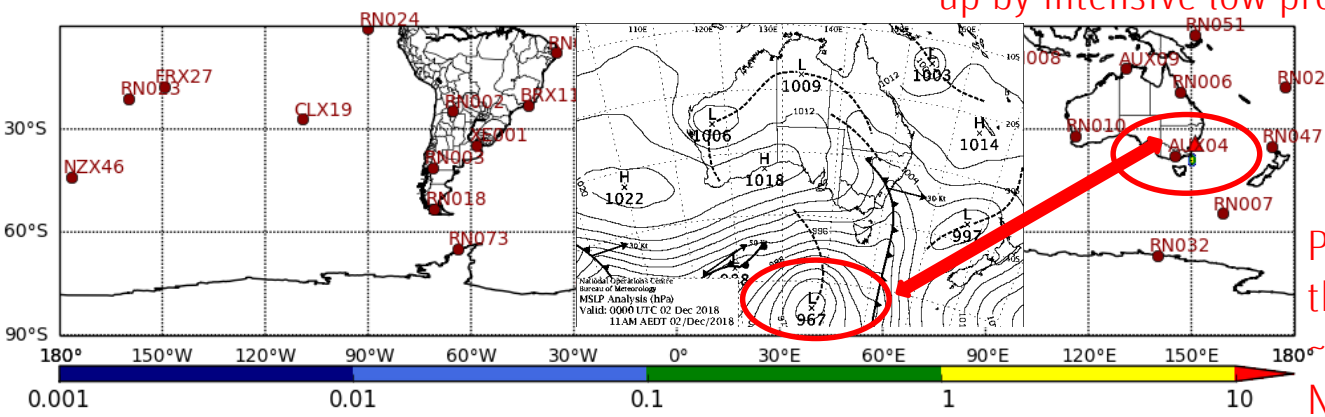
ZAMG
Zentralanstalt für Meteorologie und Geodynamik

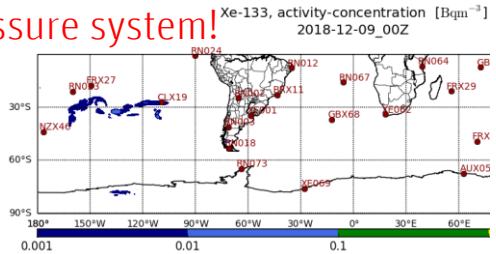
8. Test Cases (real & synthetic)

1. **Real case** (11.3.-25.3.2011) - **Fukushima NPP accident**: Real IMS Cs-137 and Xe-133 measurements and source term from *Stohl et al.* (2012).
2. **Synthetic cases** (1.12.2018-15.12.2018), producing pseudo-measurements based on FLEXPART and NCEP-GFS data:
   1. **Hypothetical ANSTO puff release** (1E15 Bq Xe-133, Dec., 1st, 00:00-01:00 UTC, summer time conditions)
   2. **Hypothetical DPRK-test site puff release** (1E15 Bq Xe-133, Dec., 1st, 00:00-01:00 UTC, winter time conditions)

Xe-133, activity-concentration [$\mathrm{Bqm^{-3}}$]
2018-12-01_18Z

Hypothetical ANSTO plume gets sucked up by intensive low pressure system!





Plume reappears in the ground layer only ~week later based on NCEP data!

ZAMG
Zentralanstalt für Meteorologie und Geodynamik

9. ECMWF EPS data

- Update of flex_extract_v7.0.3 (www.flexpart.eu) for the current purpose.
- **Re-caluculation of the full ensemble** (12-hourly short-term forecasts) for the real test case (from 2011). **Perturbed forecasts** on model level fields are available only 30 days after creation!
- Using the **Ensemble Analysis** product (available at 00/06/12/18 UTC, *stream=ELDA*,) for the synthetic test cases: 25 members, **observations and model tendencies perturbed with stochastic noise.** Another 25 symmetric members are added by calculating the differences between each member and the mean.
- Combining analysis (6- or 12-hourly) and short term-forecasts (up to 9 hours).
- Perform **test cases both with full ensemble and a 10 member subset.**
- ERA5-10-members ensemble is no alternative, due to not being real-time data.

ZAMG
Zentralanstalt für Meteorologie und Geodynamik

10. Selected samples for the synthetic test cases

Favourable conditions: Lots of pseudo-measurements

STAT	COLL START	COLL STOP	CONC NCEP	CONC ECMWF	CONC NCEP V10
CNX22	20181211 150000	20181212 150000	0.161E-02	0.0	0.165E-02
CNX22	20181212 150000	20181213 150000	0.527E-02	0.0	0.520E-02
CNX22	20181213 150000	20181214 150000	0.149E-02	0.0	0.139E-02
JPX38	20181204 180000	20181205 060000	0.144E-02	0.0	0.145E-02
JPX38	20181209 180000	20181210 060000	0.815E-02	0.0	0.822E-02
JPX38	20181210 060000	20181210 180000	0.527E-02	0.320E-04	0.552E-02
JPX38	20181210 180000	20181211 060000	0.360E-02	0.594E-05	0.408E-02
JPX38	20181213 060000	20181213 180000	0.136E-02	0.577E-06	0.139E-02
XE058	20181201 120000	20181202 000000	0.718E-02	0.0	0.718E-02
XE058	20181202 000000	20181202 120000	0.144E-01	0.0	0.145E-01
XE058	20181203 000000	20181203 120000	0.385E-02	0.0	0.391E-02
XE058	20181203 120000	20181204 000000	0.848E-02	0.0	0.855E-02
XE058	20181212 000000	20181212 120000	0.136E-02	0.732E-05	0.154E-02
XE058	20181212 120000	20181213 000000	0.166E-02	0.191E-04	0.196E-02
XE058	20181213 000000	20181213 120000	0.228E-02	0.147E-05	0.246E-02
XE060	20181206 000000	20181206 120000	0.132E-01	0.550E-06	0.131E-01
XE060	20181206 120000	20181207 000000	0.345E-02	0.663E-05	0.357E-02
XE060	20181209 120000	20181210 000000	0.184E-02	0.276E-04	0.177E-02
XE060	20181210 000000	20181210 120000	0.142E-02	0.474E-06	0.12
USX74	20181213 000000	20181213 120000	0.256E-02	0.814E-05	0.265E-02
USX74	20181213 120000	20181214 000000	0.547E-02	0.694E-05	0.528E-02
USX74	20181214 000000	20181214 120000	0.230E-02	0.0	0.222E-02
USX74	20181214 120000	20181215 000000	0.254E-02	0.602E-07	0.253E-02
USX75	20181211 120000	20181212 000000	0.126E-02	0.0	0.123E-02
USX75	20181212 000000	20181212 120000	0.350E-02	0.125E-03	0.320E-02
USX75	20181212 120000	20181213 000000	0.146E-02	0.317E-04	0.147E-02
USX79	20181213 030000	20181213 150000	0.149E-02	0.300E-05	0.152E-02
USX79	20181213 150000	20181214 030000	0.140E-02	0.644E-04	0.143E-02

Table 4: List of selected samples for the DPRK test case. Columns from left to right: IMS station name, collection start [YYYYMMDD HHMMSS], collection stop [YYYYMMDD HHMMSS], FLEXPART V8 average concentration [Bq/m^3] based on NCEP input, FLEXPART V8 average concentration [Bq/m^3] based ECMWF control run input and FLEXPART V10 average concentration [Bq/m^3] based on NCEP input.

At least one order of magnitude difference between NCEP and ECMWF driven run!

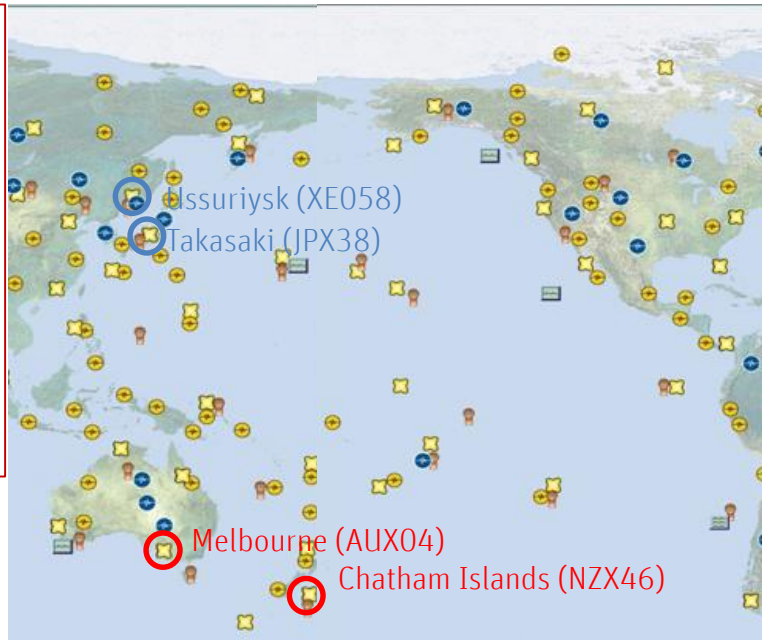
Stations (JPX38, XE058) typically expected to be affected by DPRK event.

Unfavourable conditons: Few pseudo-measurements

STAT	COLL START	COLL STOP	CONC NCEP	CONC ECMWF	CONC NCEP V10
XE001	20181213 000000	20181213 120000	0.114E-02	0.171E-04	0.115E-02
XE001	20181213 120000	20181214 000000	0.311E-02	0.656E-04	0.287E-02
XE001	20181214 000000	20181214 120000	0.220E-02	0.902E-06	0.204E-02
CLX19	20181208 090000	20181209 090000	0.171E-02	0.201E-03	0.182E-02
CLX19	20181209 090000	20181210 090000	0.743E-02	0.135E-04	0.733E-02
CLX19	20181210 090000	20181211 090000	0.845E-02	0.156E-04	0.836E-02
CLX19	20181213 090000	20181214 090000	0.161E-02	0.403E-04	0.159E-02

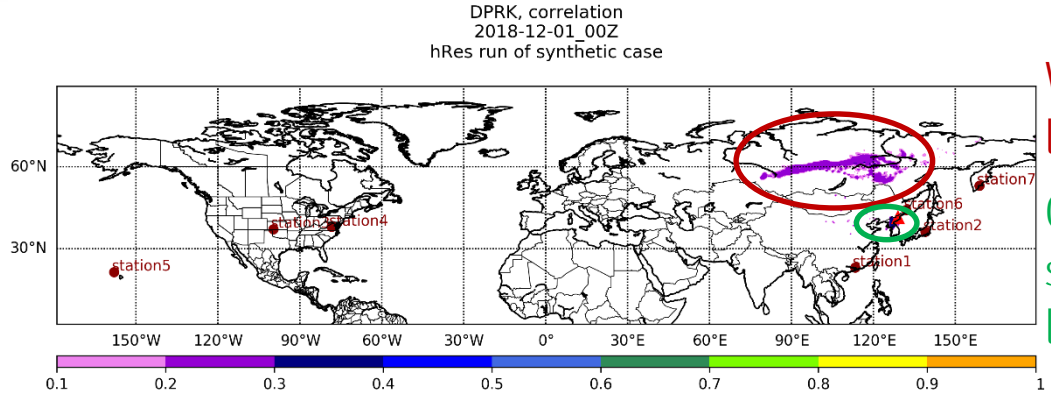
Table 5: List of selected samples for the ANSTO test case. Columns from left to right: IMS station name, collection start [YYYYMMDD HHMMSS], collection stop [YYYYMMDD HHMMSS], FLEXPART V8 average concentration [Bq/m^3] based on NCEP input, FLEXPART V8 average concentration [Bq/m^3] based ECMWF control run input and FLEXPART V10 average concentration [Bq/m^3] based on NCEP input.

Stations (AUX04 or NZX46) typically would be expected to be affected by ANSTO release.



11.1. PSR fields for the synthetic DPRK test case

Can we confine the PSR at known release time better than with the operational run only based on the metrics as introduced above?

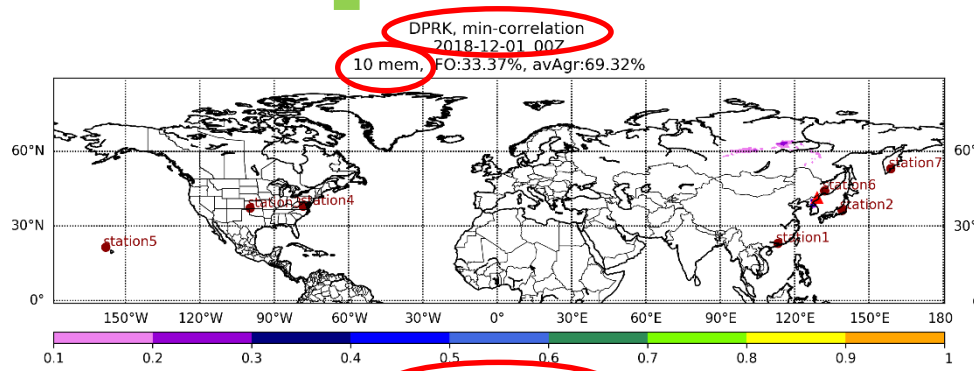


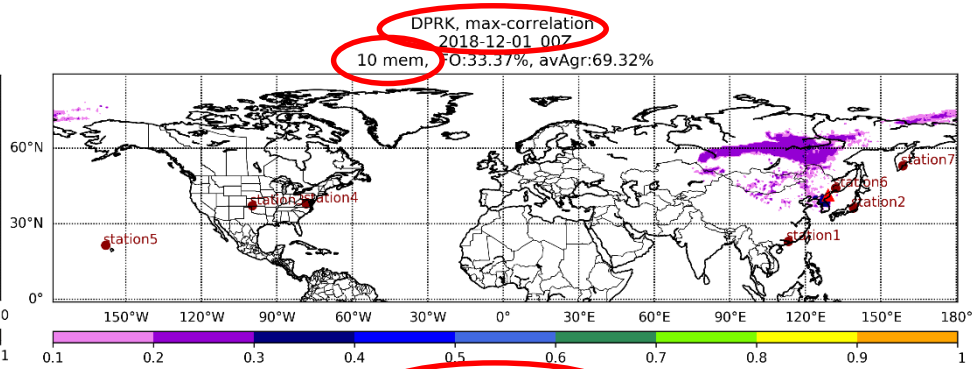
Wrong source location

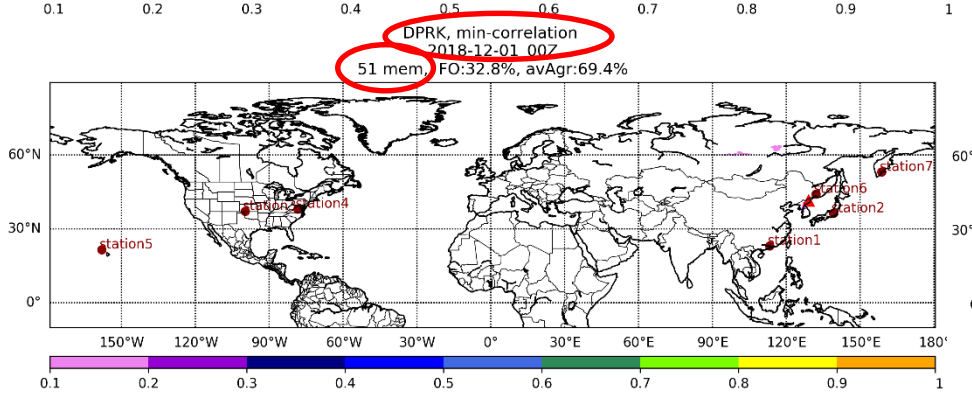
Correct source location

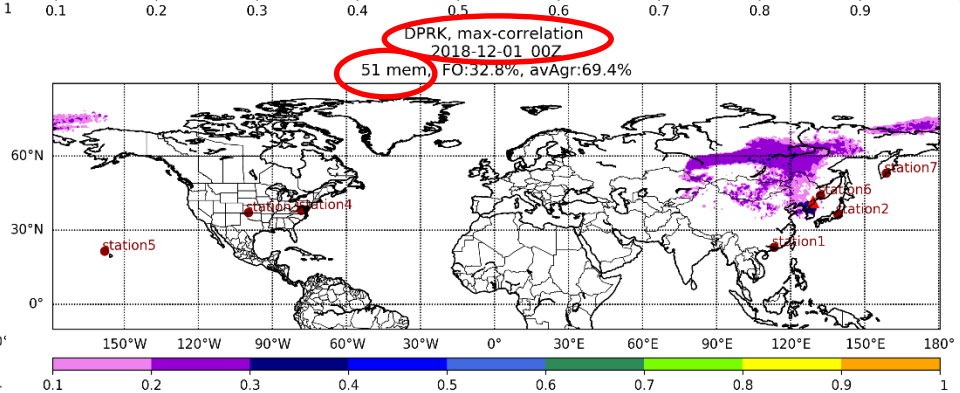
Improvement

Deterioration

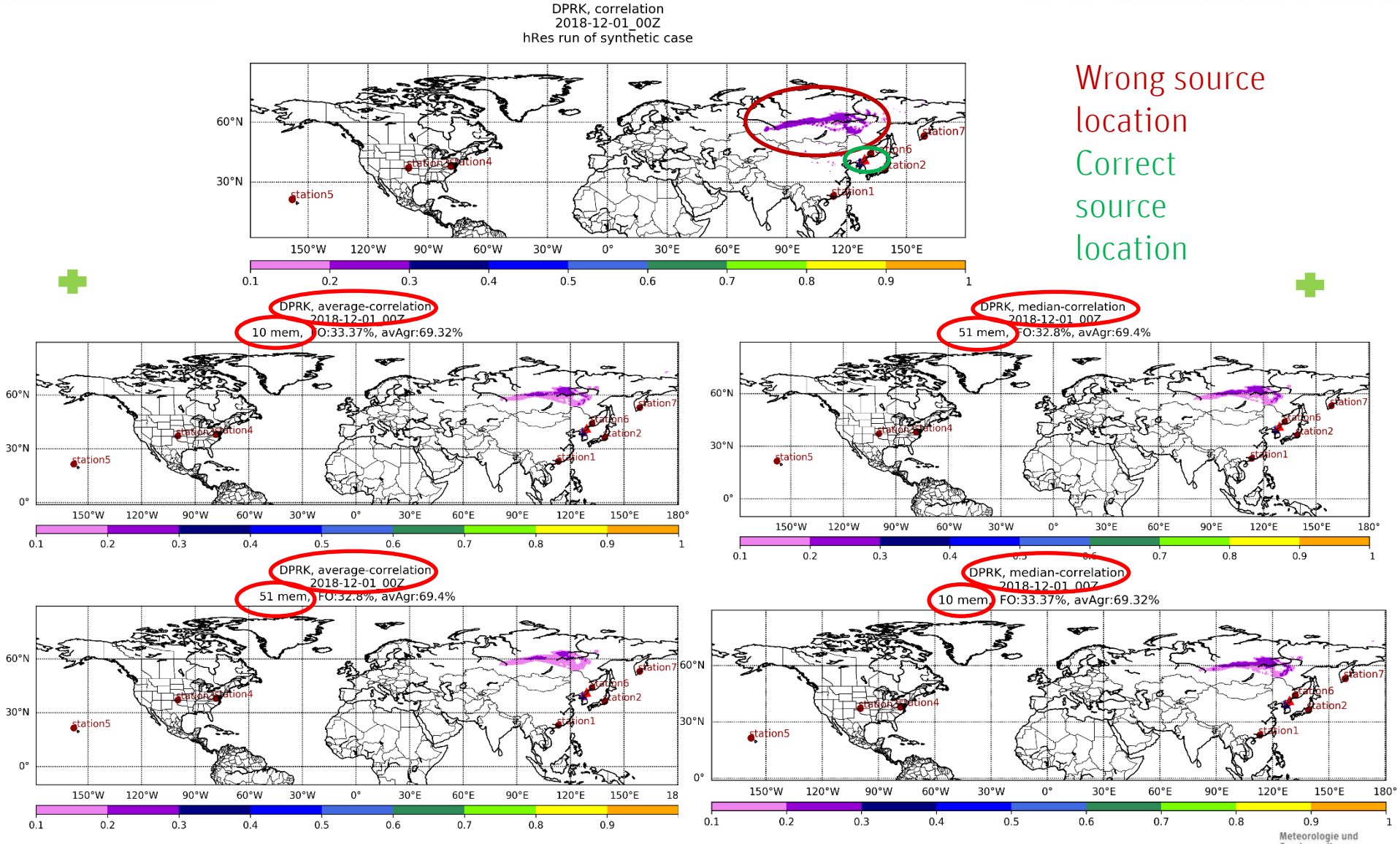








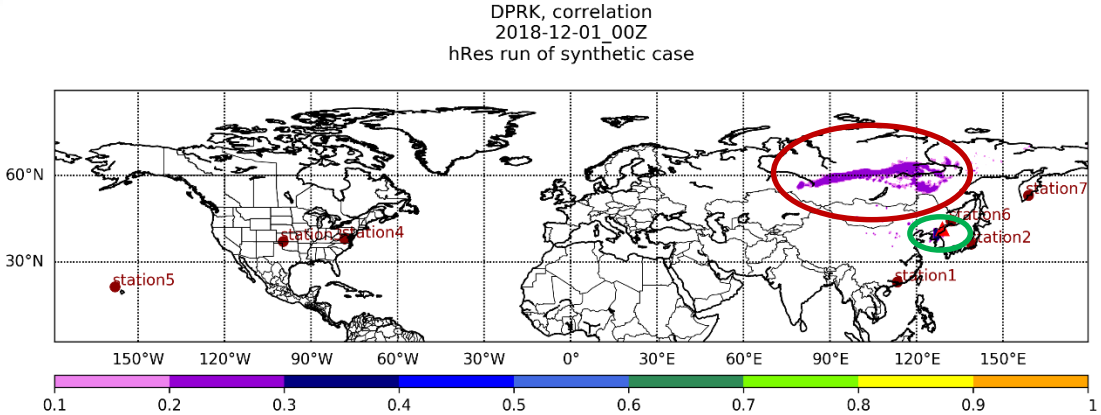
11.2. PSR fields for the synthetic DPRK test case



Wrong source location

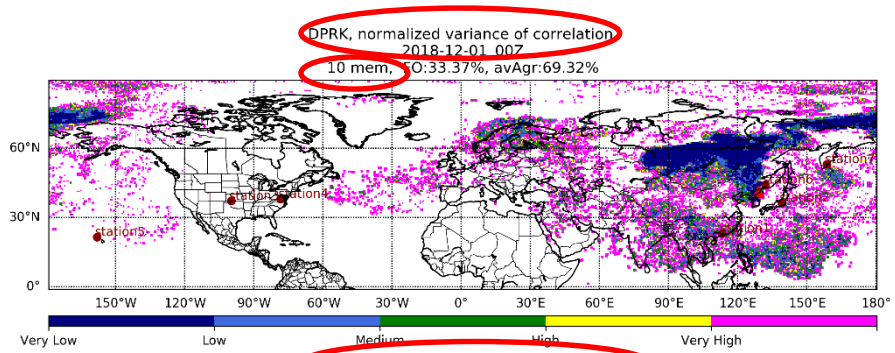
Correct source location

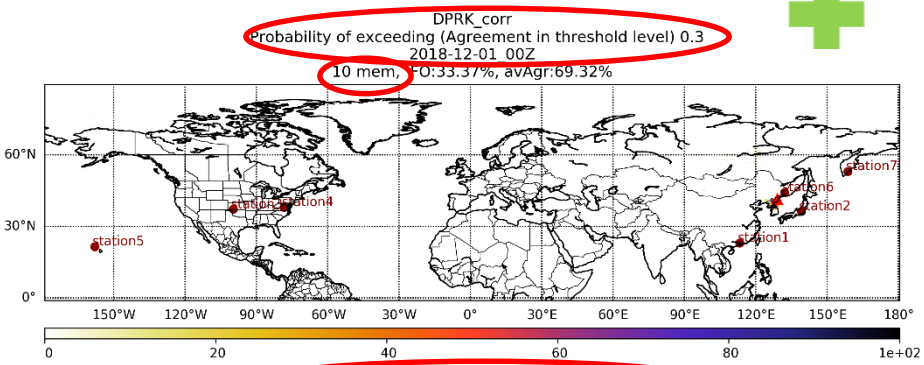
11.3. PSR fields for the synthetic DPRK test case

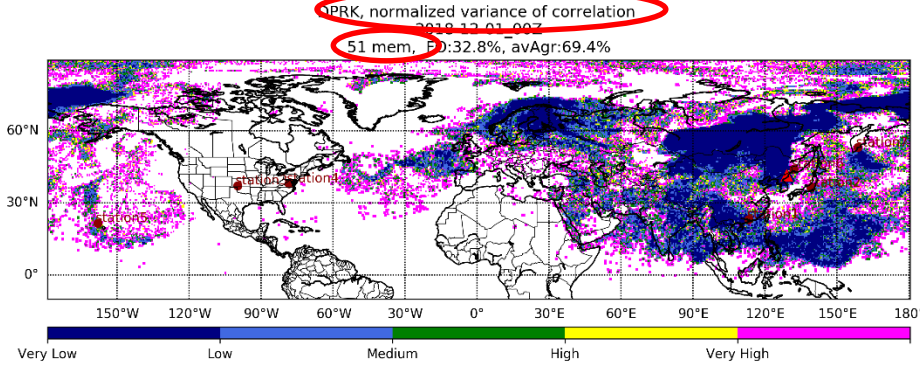


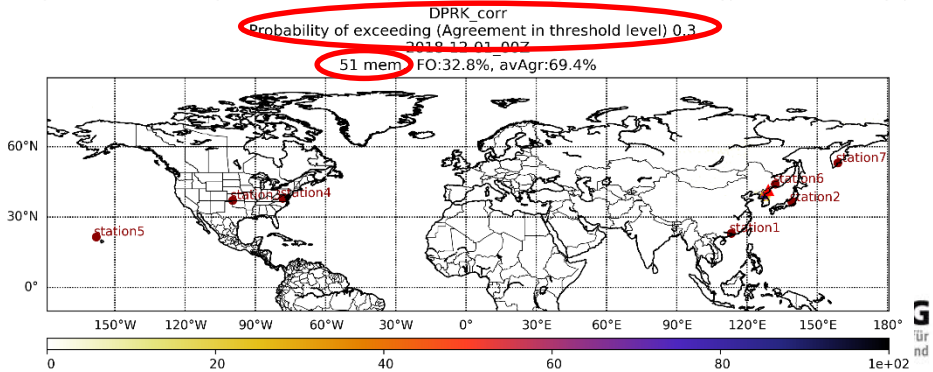
Wrong source location

Correct source location

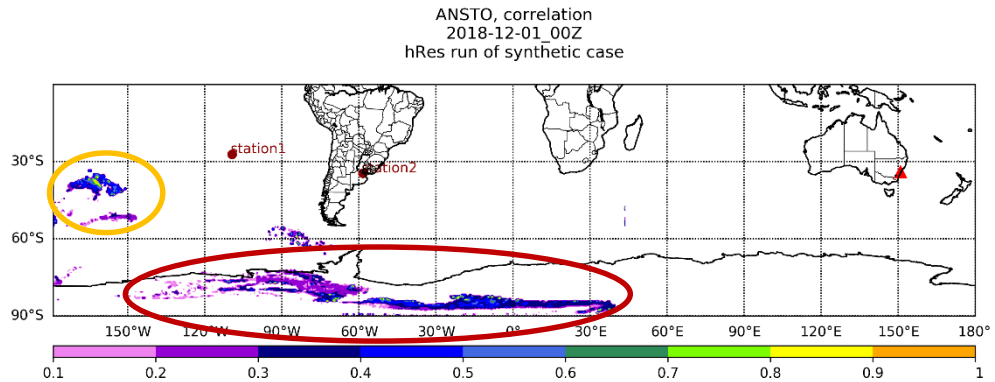






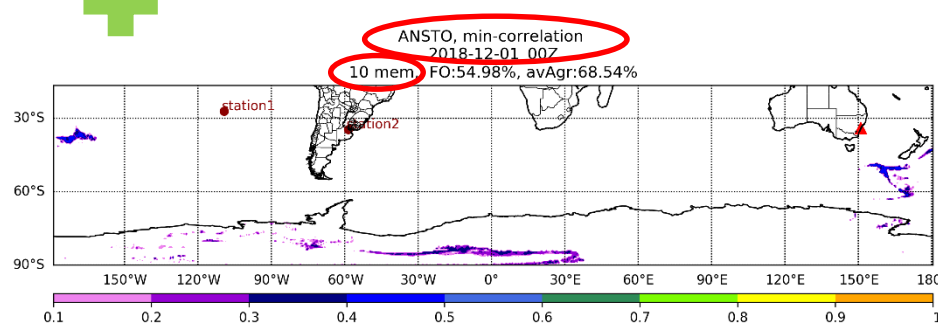


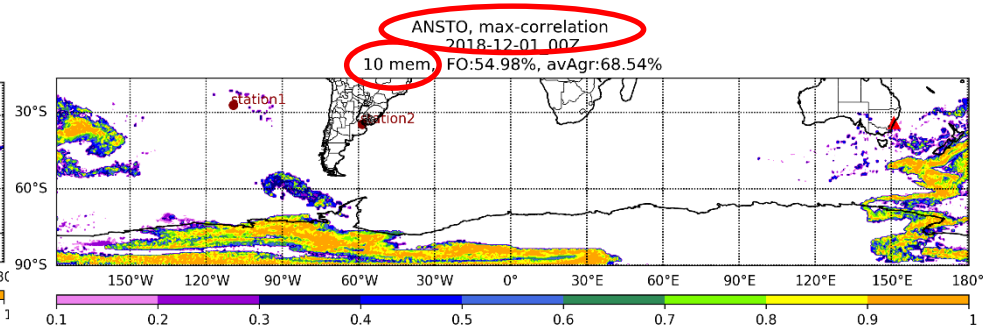
12.1. PSR fields for the synthetic ANSTO test case

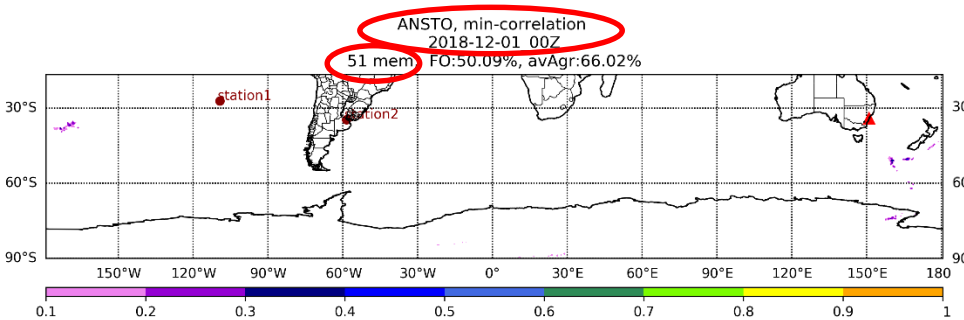


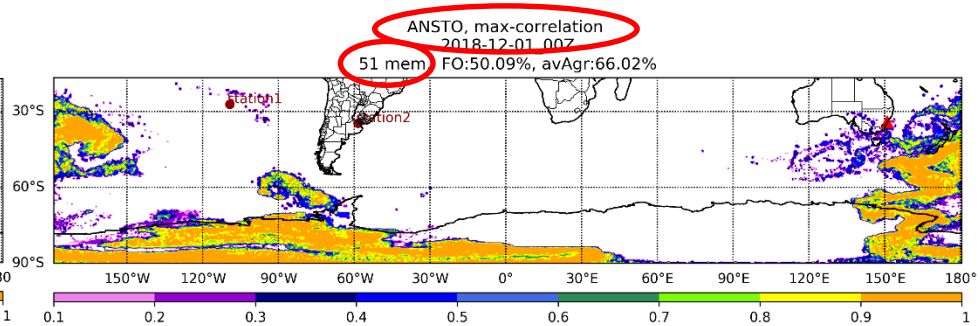
Wrong source location

Approximate source location

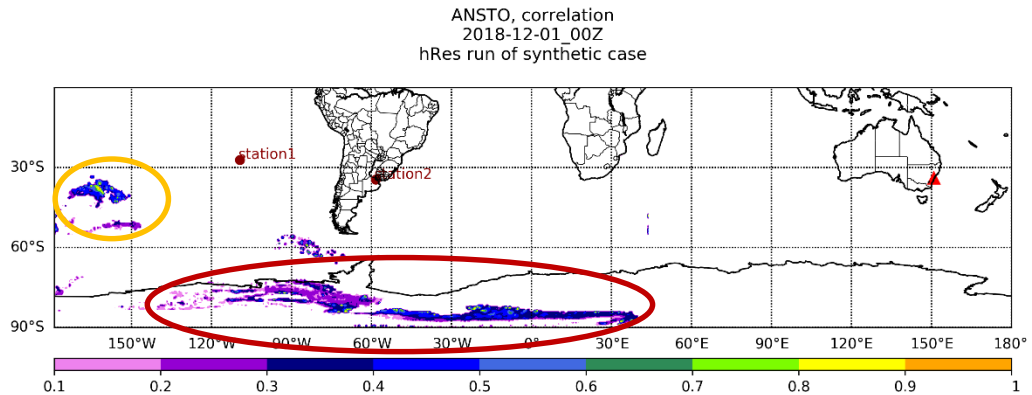






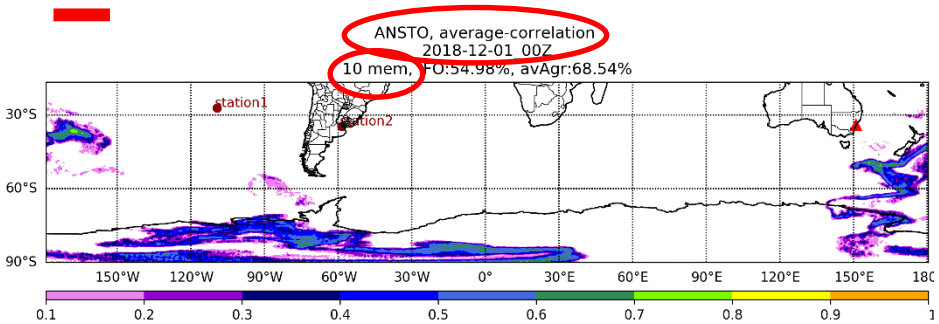


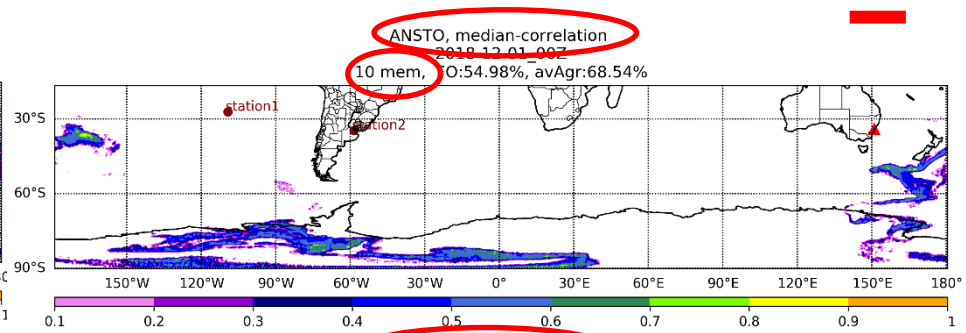
12.2. PSR fields for the synthetic ANSTO test case

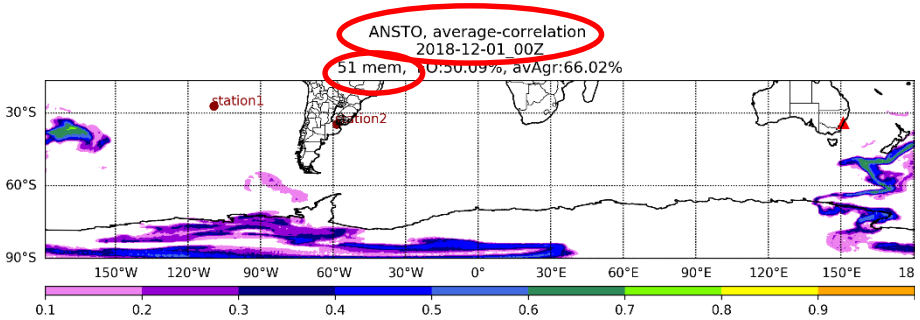


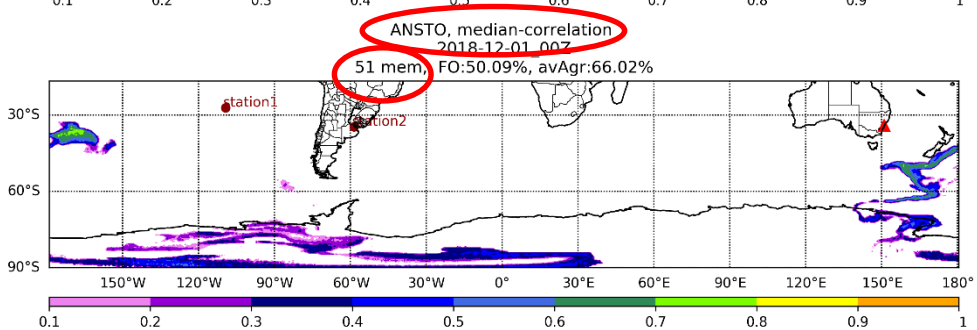
Wrong source location

Approximate source location

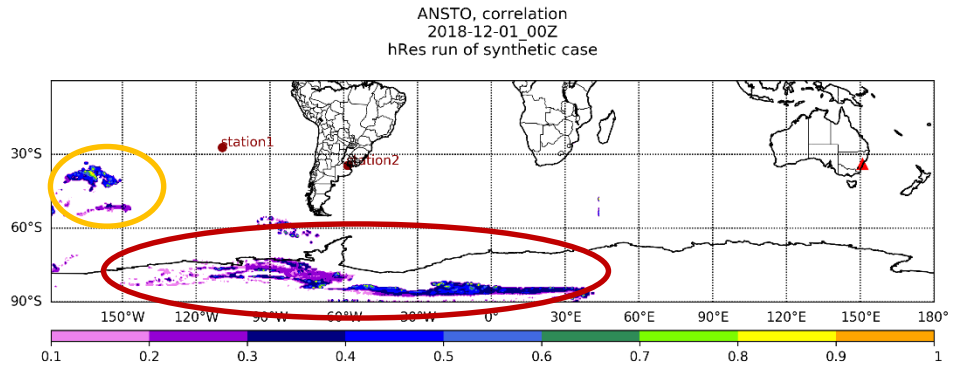






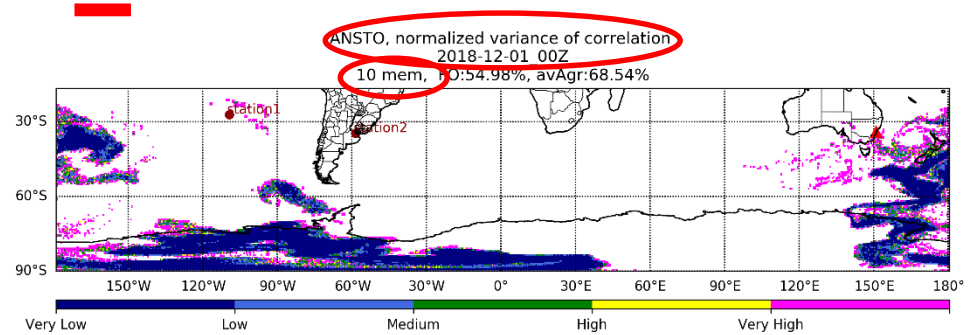


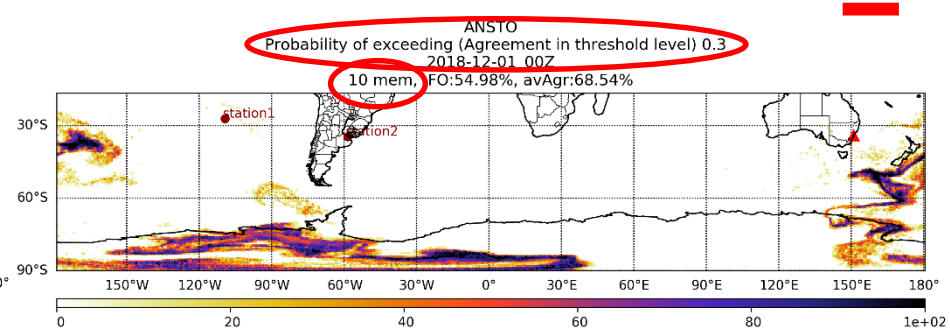
12.3. PSR fields for the synthetic ANSTO test case

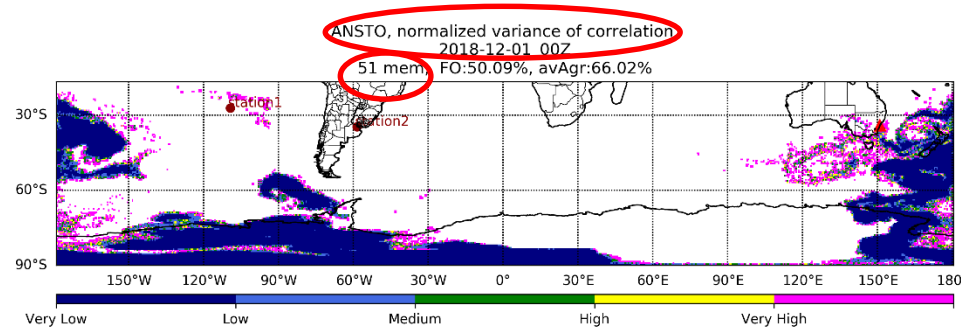


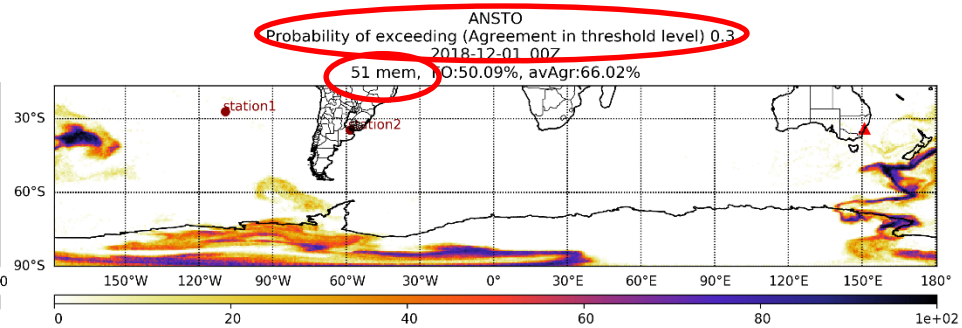
Wrong source location

Approximate source location









13.1. Preliminary conclusions

- The **best PSR ensemble metric** turns out to be the **ensemble minimum** as only those **parts of the PSR fields are kept** which are **at** (DPRK test case) **or near** (ANSTO case) to the **actual source** and **no threshold** needs **to be defined**. The **performance increases** when **switching from** a **10 member to** the **full ensemble**. Whereas when dealing **with activity concentrations** the **"worst case" (ensemble maximum)** is extremely **relevant**, it is **not so for PSR fields**.
- **Probability of exceedance** plots applied to PSR fields **suffer from using a fixed threshold**. This is **unlike the situation with activity concentration values**, where a generally useful threshold is easier to define. In essence **thresholds would need to be redefined for every case** based on those values **deduced from the operational run**. For the DPRK test case this prerequisite was met by chance.

ZAMG
Zentralanstalt für Meteorologie und Geodynamik

13.2. Preliminary conclusions

- **Average and median results** are **very similar**. They show a **small improvement** in terms of being able to locate the source region **for the DPRK test case**.
- The concept and the **usability of the "Normalized Variance"** when applied to PSR fields **needs to be re-visited**. Probably a **threshold for** confining the region of interest to **PSR values** above a certain value would be **necessary**. Variances for very tiny PSR values are not of interest.

THANK YOU FOR YOUR ATTENTION!

ZAMG
Zentralanstalt für Meteorologie und Geodynamik

I. Additional material

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, $cRNK_p$, these authors preferred to give the percentage of the maximum $cRNK$ value (3.0) while excluding the trivial auto-correlation result as follows:

$$cAgreement_p = \frac{100}{3(N-1)} \sum_{i=1}^{N} \epsilon_{ip} cRNK_p \tag{4}$$

with N, total number of experiments' participants and ϵ_{ip} =1 for $i \neq p$ and 0 for $i = p$.

The belonging case and model specific anomaly, $cAnomaly_p$, to the case specific across participants' average agreement (cAV) is calculated as:

$$cAnomaly_p = cAgreement_p - cAV \tag{5}$$

with

$$cAV = \frac{1}{N} \sum_{p=1}^{N} cAgreement_p \tag{6}$$

ZAMG
Zentralanstalt für Meteorologie und Geodynamik

II. Additional material

In order to assess probabilistic capabilities of the ensembles the *Brier Score* as function of time has been calculated for different thresholds in Galmarini et al. [16]. It is defined as the mean square error of a probability forecast:

$$BS = 1/n \sum_{i=1}^{n} (F_i - O_i)^2 \tag{13}$$

where n is the number of forecasts, F_i is the forecast probability on occasion i, O_i is the observation (0 or 1) on occasion i. The score weights larger errors more than smaller ones. De Meutter et al. [5] calculate the simulated probability as

$$F_i = \frac{t + 2/3}{n + 4/3} \tag{14}$$

where t is the number of members that simulate a value above a threshold and n is the total number of ensemble members n.

ZAMG
Zentralanstalt für Meteorologie und Geodynamik

III. Additional material

the division by $M = 0$ is kept the $NMSE$ adopts infinity. Further, Seibert [34] states that a high value of the bias implies a high value of the RMSE, though the two data sets might just be shifted by some offset and otherwise agree quite well. Therefore, the definition of a third quantity, the bias-corrected RMSE, BC_RMSE, makes sense:

$$BC\_RMSE = \frac{1}{n}\sqrt{\sum_{i}^{N}[(M_i - \bar{M}) - (O_i - \bar{O})]^2} \tag{10}$$

According to Sachs [33] (p. 128), this can also be calculated as:

$$BC\_RMSE = \sqrt{RMSE^2 - B^2} \tag{11}$$

is then the BC_RMSE. According to Taylor [42] model standard deviation, BC_RMSE and R can be combined into a single skill score S_r.

$$S_r = 2(1+R)(\frac{\sigma_m}{\sigma_o} + \frac{\sigma_o}{\sigma_m})^2 \tag{20}$$

with σ_m and σ_o being the standard deviations of predictions and observations.

The correlation contribution becomes important for large values of BC_RMSE. If one wants to include also the relative bias FB into the skill score, Seibert [34] suggests:

$$S_b = \frac{1}{1 + \alpha\, \mathsf{F}\, B^2} \tag{21}$$

A value of $\alpha = 10$ appears to give a relationship fulfilling Seibert's [34] subjective idea about such a skill score. Finally, both skill scores can be combined into a total skill score S:

$$S = \alpha S_r + (1 - \alpha S_b) \tag{22}$$

The value of α is rather arbitrary and would depend on the application. An additive and not multiplicative combination of the two scores is suggested because a model that has skill either in reproducing the mean or in reproducing the patterns should be attributed some total skill; the product of the two scores would be zero with one of the factors being zero. In any case, data sets with strongly non-normal distributions might better be transformed before applying any of the measures.

ZAMG
Zentralanstalt für Meteorologie und Geodynamik