Diverse profile datasets based on the CAMS atmospheric composition forecasting system

Reima Eresmaa and Anthony P. McNally

European Centre for Medium-range Weather Forecasts
Shinfield Park, Reading, RG2 9AX, United Kingdom

Overview

We have compiled a new diverse profile database using short-range forecasts from the currently operational version of the Copernicus Atmosphere Monitoring Service (CAMS) forecasting system. The profiles are given in a 60-level vertical grid extending from surface up to 0.1 hPa. The database consists of eight subsets. While one subset is produced by a fully-randomized selection process, the other seven subsets each focus on describing variations in one atmospheric variable only. A total of 40,000 vertical profiles are given to cover annual and diurnal variations in temperature, specific humidity, and mixing ratios of ozone, carbon monoxide, nitrogen dioxide, sulphur dioxide, and formaldehyde.

The new profile database is a follow-up product to the previous atmospheric composition -focussed database, released in 2012 and based on forecasts produced by the Monitoring Atmospheric Composition and Climate (MACC) project. A significant difference between the two databases is that the new database puts emphasis on distributions of reactive gases. Profiles of greenhouse gases and aerosols are not included.

1 Introduction

Copernicus Atmosphere Monitoring Service (CAMS) provides global atmospheric composition analyses and forecasts in near-real-time for a wide range of users. Building on the architecture of the Integrated Forecasting System (IFS), CAMS maintains and develops an operational numerical system with emphasis on concentrations of greenhouse and reactive gases and aerosol. Since the upgrade of June 2016, the operational CAMS system has been run in resolution T511 and five-day forecasts have been produced twice a day.

With the intention to provide a concise but comprehensive description of key parameters contained in the CAMS operational system, a diverse profile database has been compiled at the European Centre for Medium-range Weather Forecasts (ECMWF). The new database, hereafter the CAMS profile database, continues the series of ECMWF diverse profile databases. Previously, profile databases have been released using either the operational NWP system of ECMWF (Chevallier et al., 2006; Eresmaa and McNally, 2014) or a delayed-mode numerical system of the Monitoring Atmospheric Composition and Climate (MACC) project (Eresmaa et al., 2012). The NWP-focussed databases emphasized the sampling of variables with direct meteorological interest and the MACC profile database focussed on distributions of greenhouse gases and aerosol. In the CAMS profile database, the focus is put on reactive gases including O_3, NO_2, SO_2, CO, HCHO.

This document describes the production process and contents of the CAMS profile database. The generic profile selection algorithm and details of implementation to produce the CAMS profile database are described in Section 2. Statistical distributions of sampled variables in the database, as well as spatio-temporal locations of selected profiles, are discussed in Section 3. Detailed instructions on how to read the CAMS profile database are given in Section 4.

2 Selection algorithm

The CAMS profile database is produced using the selection algorithm of Chevallier et al. (2006) with modifications as in Eresmaa and McNally (2014). As the basis of the selection algorithm, the squared departure $D(s_i, s_j)$ separating profile s_i from profile s_j is defined as

$$D(s_i, s_j) \quad = \quad \sum_{k=1}^{K} \sum_{m=1}^{M} \left(\frac{\theta_{ik}(m) - \theta_{jk}(m)}{\sigma_k(m)} \right)^2 , \tag{1}$$

where k and m, respectively, are indices of variable and level, K and M are total numbers of variables and levels to be considered, $\theta_{ik}(m)$ and $\theta_{jk}(m)$ are values of variable k on level m in the two profiles, and $\sigma_k(m)$ is the standard deviation of variable k on level m. Pools consisting of input and output profiles are denoted by S_I and S_O, respectively. A candidate profile s_i, drawn from S_I, is saved in S_O if and only if the inequality

$$D(s_i, s_j) > t \ \ \forall \ \ s_j \in S_O \tag{2}$$

is true. Profiles contained in S_I are considered in random order, and threshold t is tuned empirically such that the number of profiles included in S_O in the end of the process is as desired. As an important modification on top of the basic algorithm of Chevallier et al. (2006), a pre-defined number (hereafter denoted by N) of profiles is chosen in a completely random fashion in order to make statistical properties within S_O stay reasonably close to those within S_I. This is achieved by keeping t at zero until N profiles are selected.

2.1 Implementation

Operational and pre-operational forecasts produced by the CAMS global forecasting system version 41R1_CAMS_hires are used as input. The forecasting system has been run pre-operationally since 4 November 2015 and operationally since 21 June 2016. The model forecasts are given on 60 levels (i.e., M=60 at all times) and in resolution T511 in a reduced Gaussian grid, corresponding to approximately 40 km grid spacing. The model top is at 0.1 hPa. In order to cover as much as possible of seasonal variability, the CAMS profile database contains vertical profiles retrieved from model forecasts produced during a one-year period starting on 9 November 2015. Forecast steps 12, 18, 24, and 30 hours are included in the database to allow covering diurnal variations. Due to practical constraints, the input data is thinned in time dimension such that only forecasts initialized at 00z on either 9th, 19th, or 29th day of each calendar month are considered.

Resulting from the decisions to process four forecast valid times per day and three days per month, there are 144 unique forecast valid times to be considered during the full one-year time period. As the global modelling grid consists of 348,528 grid points, the total number of profiles provided as input to the production of the CAMS profile database equals 50,188,032.

The CAMS profile database is set to consist of eight subsets, each corresponding to univariate sampling focusing on one atmospheric variable (i.e., K=1 at all times). The sampled variables are temperature (T), specific humidity (Q), and mixing ratios of ozone (OZ), carbon monoxide (CO), nitrogen dioxide (NO_2), sulphur dioxide (SO_2), and formaldehyde (HCHO). Additionally, one subset is made to consist of randomly-selected profiles only. In accordance with earlier profile databases released by ECMWF, each subset is set to contain 5,000 profiles, so that the total number of profiles included in the CAMS profile database equals 40,000. Despite that each subset is produced independently of other subsets, the final database provides the same set of atmospheric and surface parameters for all subsets.

As it is computationally not feasible to process the whole input pool of 50,188,032 profiles in one go, the selection of profiles is carried out in two stages (per subset). During the first stage, each forecast valid time is processed separately from others, and t is specified such that approximately 2,400 profiles are retained per forecast valid time (t is not changed from one forecast valid time to another). The

Subset	*First stage*		*Second stage*	
	Threshold departure	Output count	Threshold departure	Output count
temperature	0.074	352,085	0.254	5,000
specific humidity	0.31	345,034	0.626	5,000
ozone	0.34	340,175	0.67	5,000
carbon monoxide	1.16	351,593	3.09	5,000
nitrogen dioxide	6.5	342,011	4.62	5,000
sulphur dioxide	8.6	347,631	7.4	5,000
formaldehyde	2.94	344,066	4.28	5,000
random	-	345,600	-	5,000

Table 1: Applied threshold parameter values and total counts of selected profiles in the end of the first and second rounds of running the selection algorithm for each subset of the database.

target number of 2,400 is chosen for convenience such that the total number of profiles passing the first selection stage is close to the total number of grid points (i.e. the number of profiles available at each forecast valid time; 144$\times$2,400=345,600). The first 2,100 profiles are selected fully by random, i.e., N=2,100.

Output from the first selection stage, as applied separately to each forecast valid time, are put together to provide input to the second selection stage. During the second stage, N=4500 and t is specified such that exactly 5,000 profiles are selected for each subset. Due to the stochastic nature of the process, the number of output profiles varies even if t is fixed, and therefore this stage is repeated as many times as necessary to get exactly the target number of output profiles.

Table 1 shows the applied threshold parameters and number of profiles selected during the first and second stages of the process for each subset of the CAMS profile database.

3 Sampled distributions

3.1 Distribution in space and time

Figure 1 is an overview of the distribution of selected profiles in space and time dimensions. Each dot represents one profile location. Majority of profiles are chosen by random selection and these are shown in gray dots. Black dots are for the 10% of profiles that have passed the selection algorithm even with a non-zero t specification and thus represent circumstances that are atypical in one or another way. Each block of 5,000 profiles corresponds to one subset of the database as indicated at the top. Within each block, profile index increases primarily with forecast valid time and secondarily with grid point index. In essence, therefore, profiles collected during Northern (Southern) Hemispheric winter season are in the left-hand-side (right-hand-side) in each subset block. The y-axis shows the grid point index such that grid points near the North (South) pole are at the bottom (top) and those at the Equator are in the middle.

As expected, randomly-selected profiles profiles spread out homogeneously within the range of possible grid point indices. The atypical profiles tend to concentrate in areas where variability of the sampled variable is comparatively high. For temperature (profiles 1–5000), high variability results in many profiles selected from both Northern and Southern high latitudes during local Winter. For humidity, black dots concentrate on midlatitudes in the Summer, but more profiles are selected in the Northern hemisphere than in the South. In the case of ozone mixing ratio, the variability is relatively high in the Northern extratropics and in particular over the Arctic during Winter and Spring seasons. Sampling distributions of other reactive gases are concentrated around certain episodes and are most easily interpreted with reference to Fig. 2 that shows locations of selected profiles on map. In case of carbon monoxide, the most relevant episodes include the widespread forest fires in Indonesia, Canada and central Siberia.

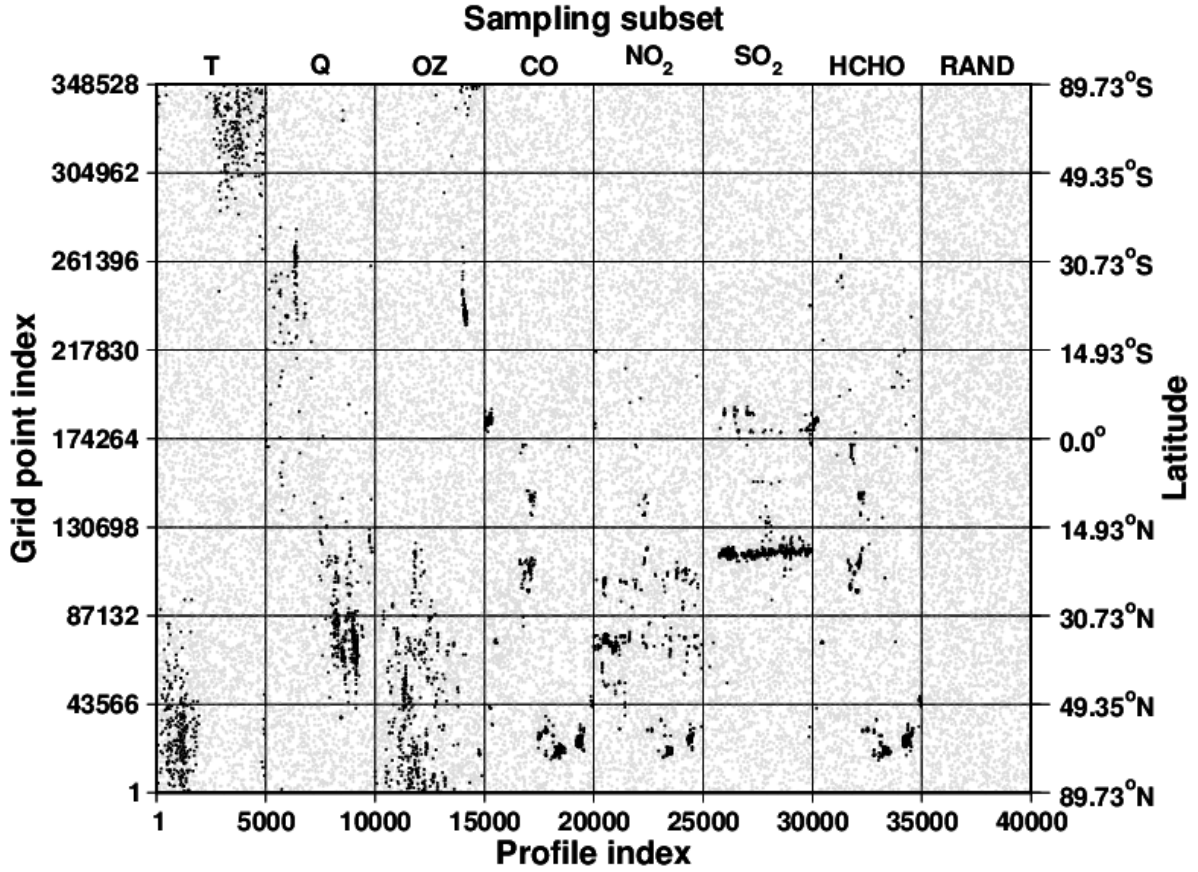

Figure 1: Grid point indices of selected profiles organised as a function of profile index. Gray dots indicate profiles selected by random, and black dots are for profiles passing the selection algorithm. Each block of 5,000 profiles corresponds to one subset of the database as indicated at the top. Within each subset, profiles are ordered first by forecast valid time and then by grid point index.

Variability in nitrogen dioxide shows maxima near emission sources in China, central Siberia and Northern Iran. For sulphur dioxide, sampling is enhanced in the context of volcanic eruptions particularly in Mexico and Hawaii. Collocated with the region concentrated sampling in carbon monoxide and nitrogen dioxide subsets, sampling distribution of formaldehyde shows a prominent maximum in central Siberia.

Figure 3 compares locations of selected profiles in T-, Q-, and OZ-sampled subsets with those included in the previous NWP-focussed diverse profile database (IFS-137 profile database; Eresmaa and McNally, 2014). Both the CAMS profile database (left column) and the IFS-137 profile database contain 5,000 profiles for each subset, and the percentage of profiles chosen by random is the same (90%). In the temperature subset, more profiles are selected from southern parts of South America and Antarctic Peninsula in the IFS-137 profile database than in the CAMS profile database. In the specific humidity subset, the CAMS profile database appears to contain fewer profiles in tropical continents and more profiles in high and mid-latitudes as compared with the IFS-137 profile database, but these differences are small. With regard to the ozone subset, the CAMS profile database contains more profiles selected from North America and Greenland. As the sampling methods used in the two profile databases are practically identical, the differences seen in coverage maps are believed to originate from system differences in horizontal grid resolution (T511 vs. T1279) and vertical discretization (60 levels with top at 0.1 hPa vs. 137 levels with top at 0.01 hPa). The differences in the sampling of ozone mixing ratio are potentially associated with more realistic model description in the CAMS forecasting system.

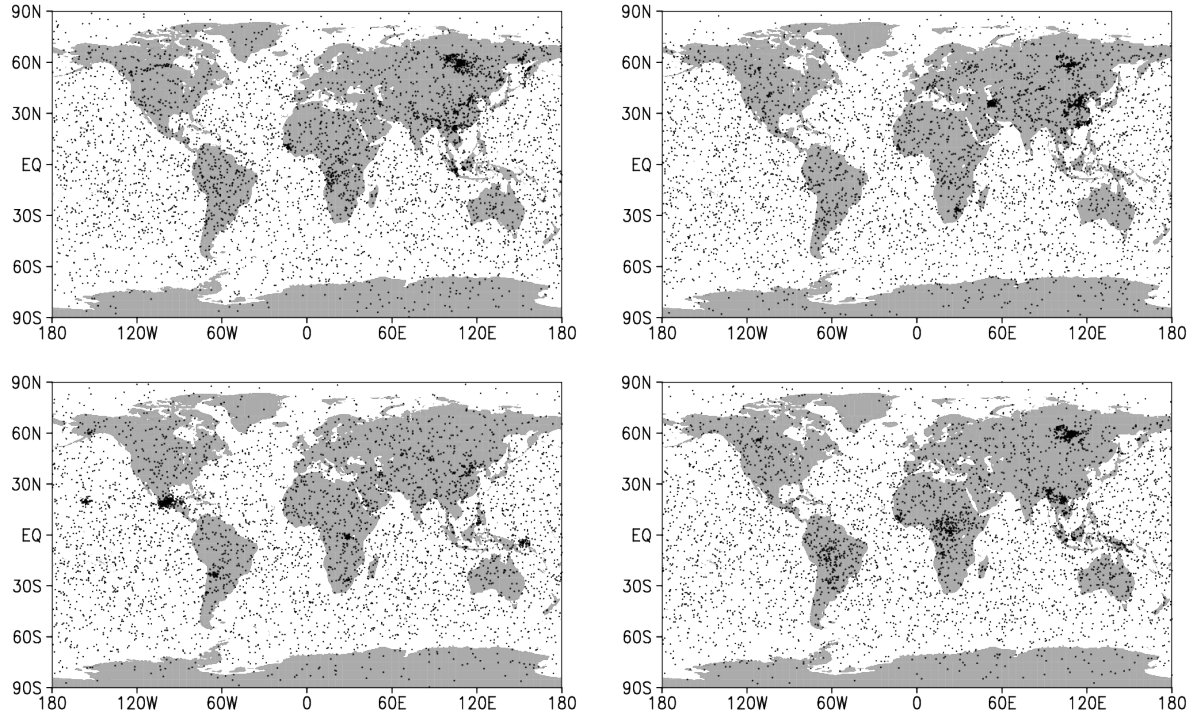

Figure 2: Locations of selected profiles in CO (top left), NO_2 (top right), SO_2 (bottom left), and HCHO (bottom right) -sampled subsets of the CAMS profile database.

3.2 Sampled variable distributions

Distributions of the sampled variables in each subset of the CAMS profile database are illustrated in Figs. 4–10. In each figure, the panel on the left shows the sampled distribution in terms of vertical profiles corresponding to the median (black line), 25th and 75th percentile (i.e., lower and upper quartile; range constrained by red shading), 10th and 90th percentile (orange shading), 5th and 95th percentile (green shading), 1st and 99th percentile (cyan shading), and minimum and maximum (blue shading). Panels in the middle and on the right show the difference (Fig. 4) or ratio (Figs. 5–10) of percentile curves as compared with the median. Dashed lines show the difference (or ratio) curves for the relevant variable in the random subset and they indicate the extent to which sampled variables spread out in typical conditions. The effect of applying the sampling algorithm to produce other subsets shows up as the difference between solid and dashed lines.

Considering the temperature subset (Fig. 4), including 500 atypical profiles in the total sample of 5,000 profiles widens the distribution especially by making the 5th and 10th percentile profiles colder in stratosphere and the 25th percentile profile colder in troposphere. There is little effect on upper quantiles (i.e., 75th, 90th, 95th, and 99th percentiles). Indicated also by the median difference in the random subset (black dashed line), the median profile in the T-sampled subset is notably colder than in the random subset.

The specific humidity subset (Fig. 5) shows the effect of the sampling algorithm by slightly increased separation between lower and upper quantile profiles. There is a general shift towards higher values throughout the troposphere. The largest impact is in the shift of the 99th percentile curve in upper troposphere, indicating that the sampling algorithm is successful in finding and selecting a handful of extremely moist profiles.

In the case of sampling the ozone mixing ratio (Fig. 6), the algorithm prefers selecting profiles with relatively high lower-stratospheric and tropospheric ozone concentration. This results in the median

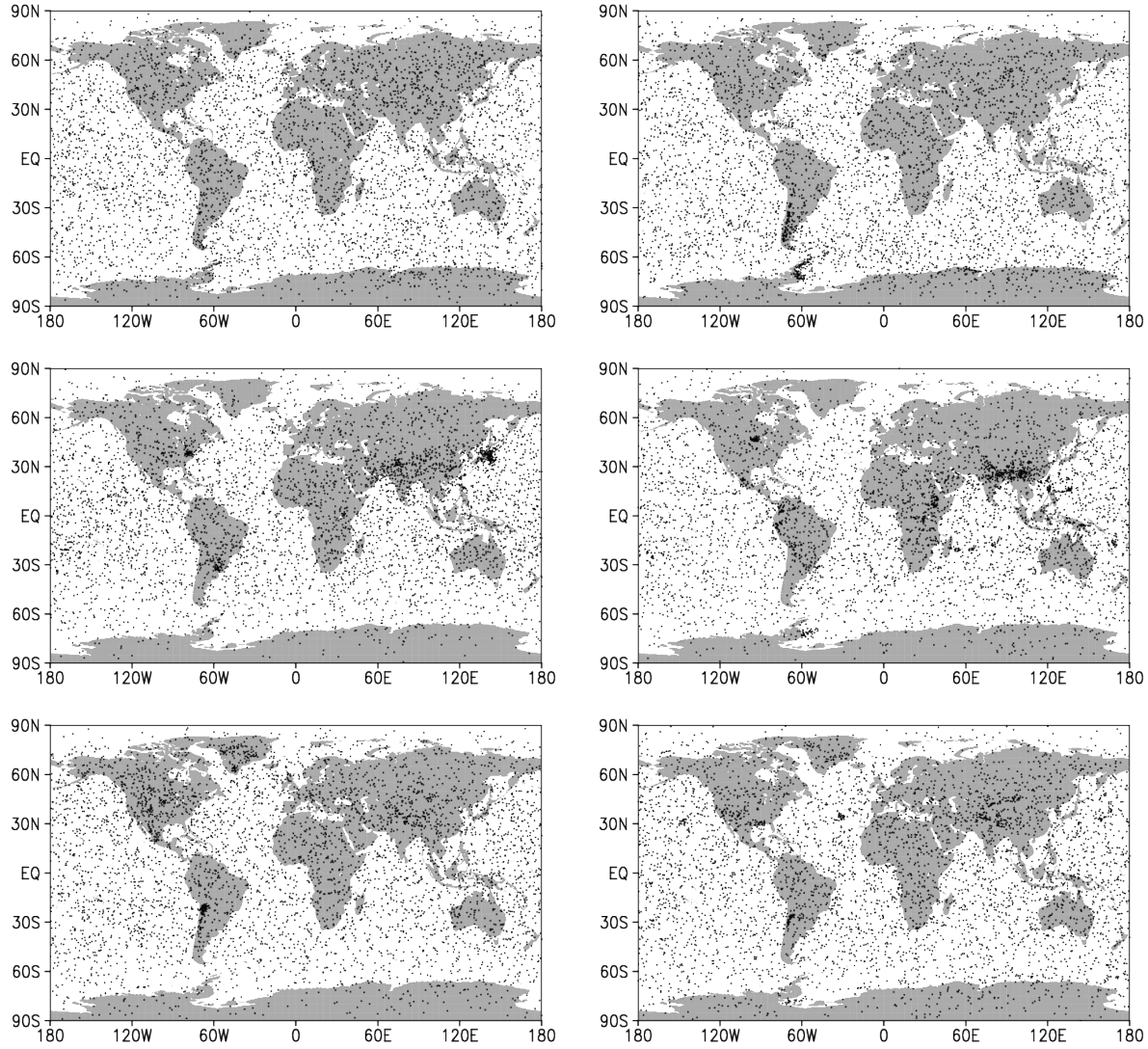

Figure 3: Locations of selected profiles in T (top), Q (middle), and OZ (bottom) -sampled subsets of the CAMS (left) and IFS-137 (right) profile databases.

profile shifting towards higher values between 50–300 hPa. All upper quantile profiles are shifted towards higher values in troposphere, but not above tropopause. The most notable shift in lower quantile profiles is towards higher values at the upper troposphere in the 25th percentile profile.

Fig. 7 shows the sampling distribution for carbon monoxide. The distribution shows a strong maximum captured by the 90th, 95th, and 99th percentile profiles in the lower troposphere, while the median shows a mixing ratio profile that decreases only slowly with increasing height in the troposphere. With regard to upper quantiles and their relation to the median, the use of the sampling algorithm makes a large impact on 90th percentile profile in the lower troposphere, and 95th and 99th percentile profiles show the large impact extending up to the tropopause. Again, very little impact from the selection algorithm is seen in the lower quantile profiles.

The vertical distribution of nitrogen dioxide (Fig. 8) shows a strong maximum near 10 hPa. At the level of the maximum, the global distribution is very homogeneous as percentile profiles are close to each other and the only visible effect from using the sampling algorithm is the shift in the 1st percentile profile towards lower values. Elsewhere, a big difference is made in the troposphere. Especially the 90th, 95th,

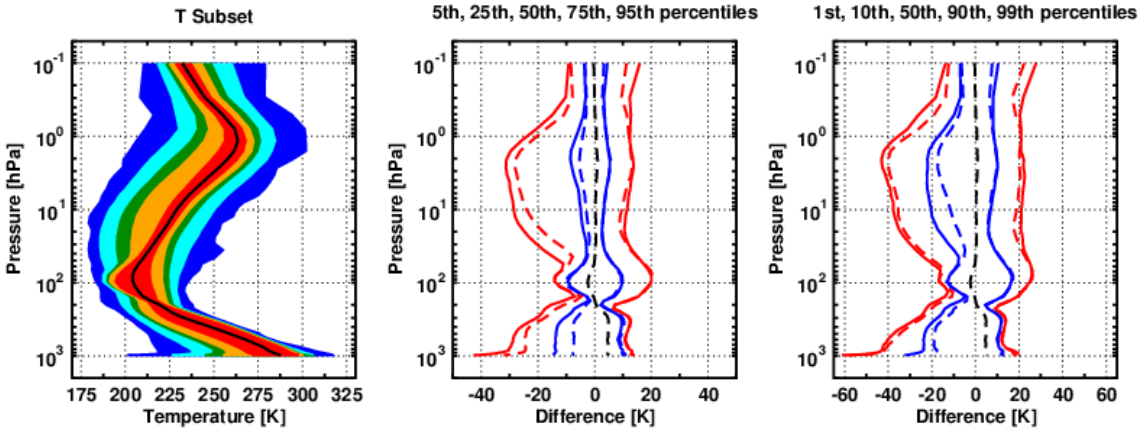

Figure 4: (left) Temperature distribution in the T-sampled subset. Blue, cyan, green, orange, and red shading indicate the range constrained by minimum and maximum, 1st and 99th percentile, 5th and 95th percentile, 10th and 90th percentile, and 25th and 75th percentile, respectively. Black line shows the median profile. (middle) Temperature difference between selected percentile profiles and the T-sampled median profile. Dashed (solid) lines are for the random (T-sampled) subset. Red lines show the 5th and 95th percentiles, blue lines show the 25th and 75th percentiles, and the black line shows the random-sampled median. (right) As middle, except that red lines show the 1st and 99th percentile, and blue lines show the 10th and 90th percentile.

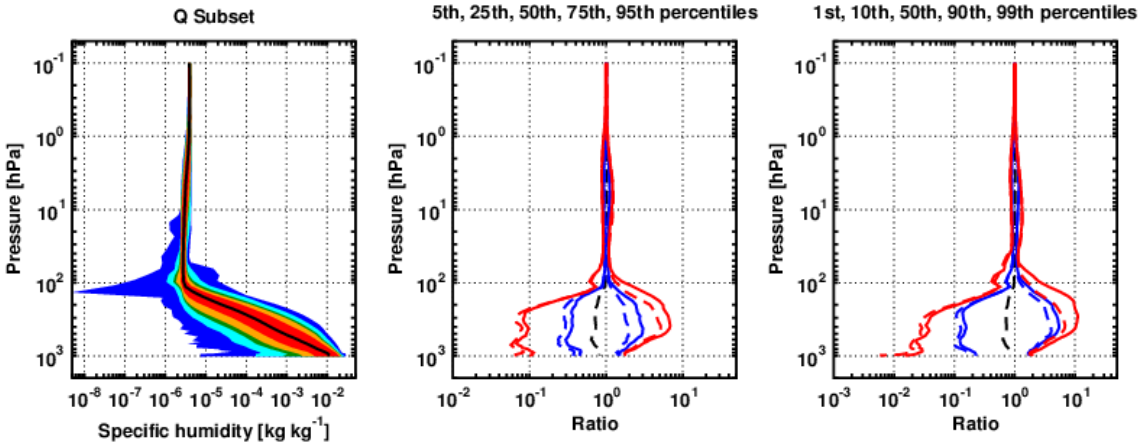

Figure 5: As Fig. 4, but for specific humidity distribution in the Q-sampled subsets. Percentile profiles in middle and right-hand-side panels are normalized by the Q-sampled median profile.

and 99th percentile profiles are shifted to the right to represent considerably higher values, as compared with the random subset.

Sulphur dioxide distribution (Fig. 9) is vertically rather homogeneous in troposphere but the mixing ratio reduces towards a constant small number in stratosphere. The sampling algorithm produces a slight shift towards higher values in the 25th percentile and median curves. A more prominent impact from the sampling algorithm is seen in upper quantiles, particularly in 90th, 95th and 99th percentile curves.

With other reactive gases consistently showing a shift towards higher values in upper percentile profiles in troposphere, and little impact from the sampling algorithm in lower quantiles or in stratosphere, the sampling of formaldehyde (Fig. 10) makes no exception. The median profile shows a steadily decreasing mixing ratio with increasing height in troposphere.

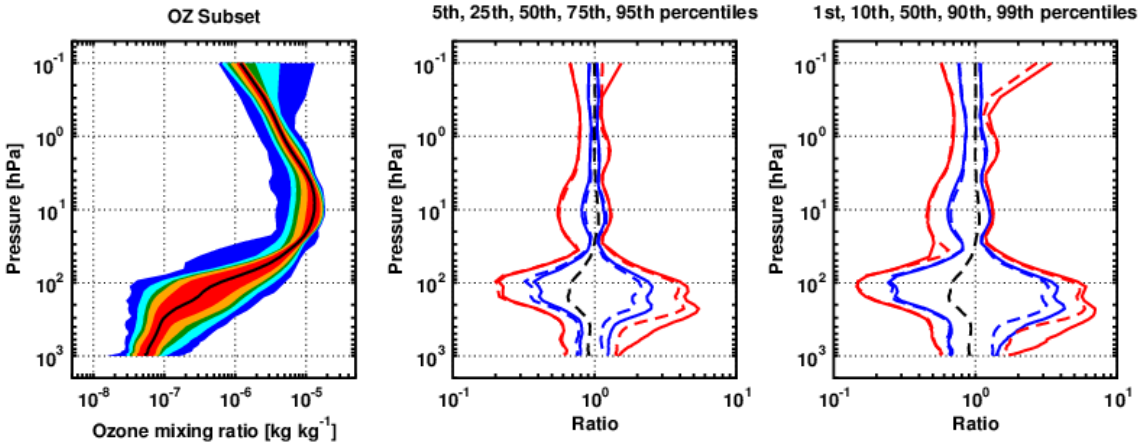

Figure 6: As Fig. 5, but for ozone mixing ratio distribution in the OZ-sampled subset.

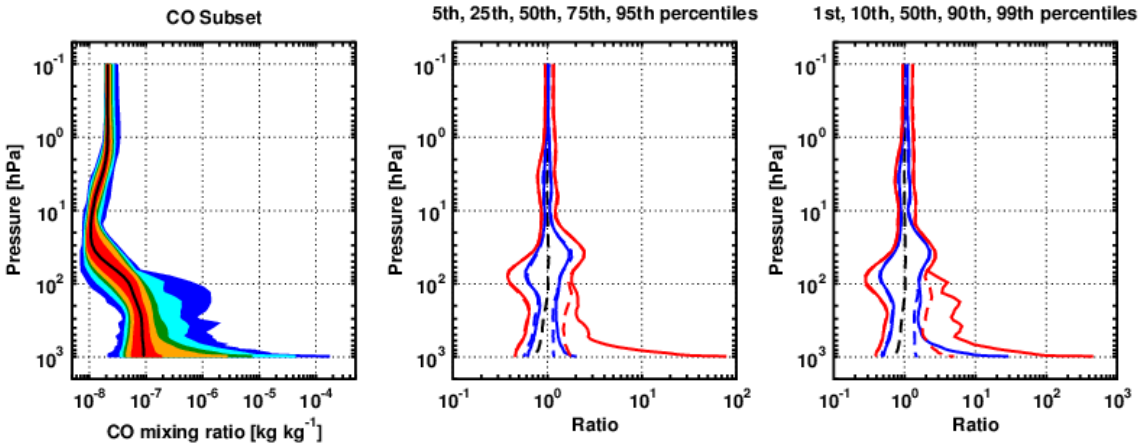

Figure 7: As Fig. 5, but for carbon monoxide mixing ratio distribution in the CO-sampled subset.

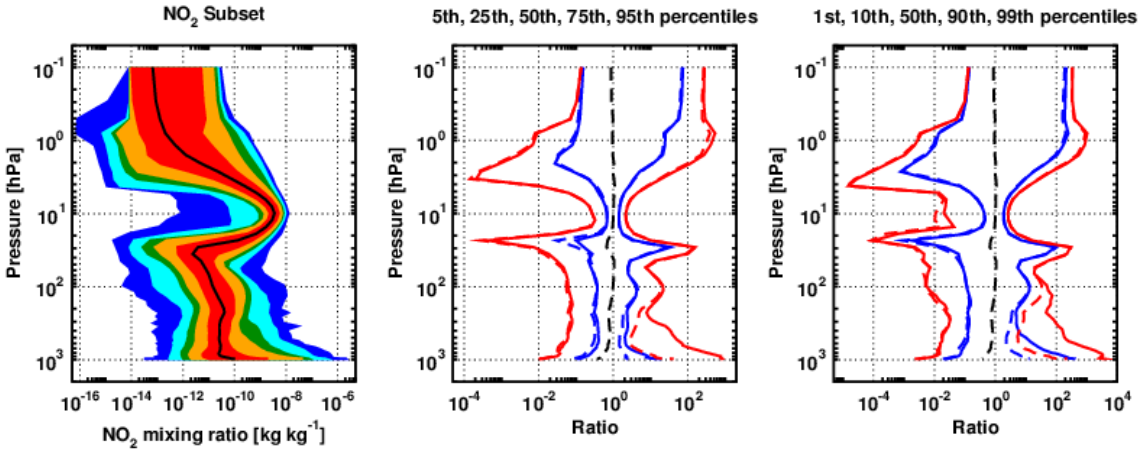

Figure 8: As Fig. 5, but for nitrogen dioxide mixing ratio distribution in the NO_2-sampled subset.

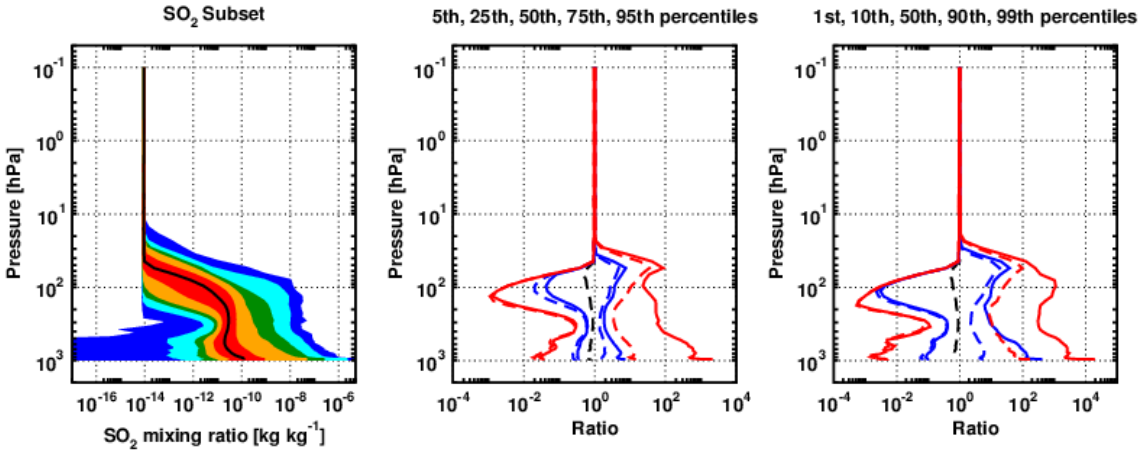

Figure 9: As Fig. 5, but for sulphur dioxide mixing ratio distribution in the SO_2-sampled subset.

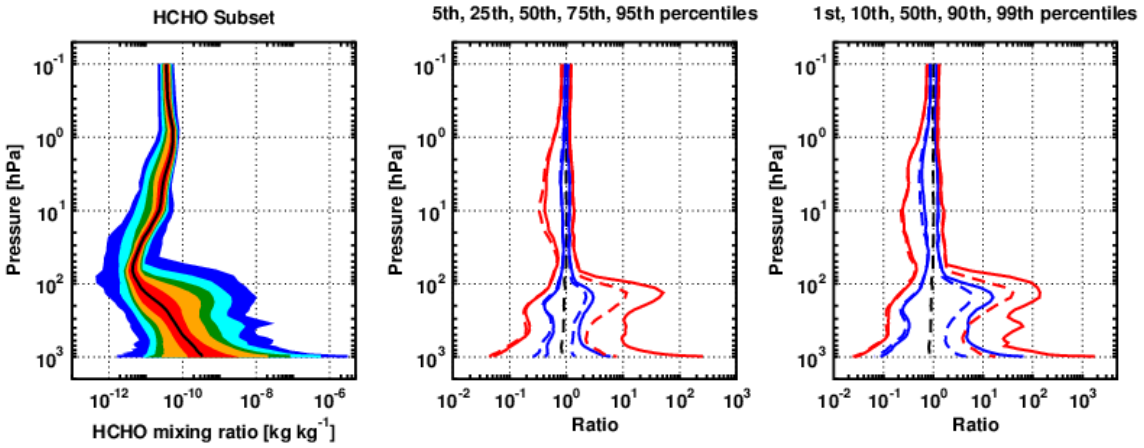

Figure 10: As Fig. 5, but for formaldehyde mixing ratio distribution in the HCHO-sampled subset.

4 Reading the database

The CAMS profile database is distributed as a compressed and tarred data file `nwpsaf_cams.tar.gz`. Assuming that the user has successfully copied the file from the web site of the NWP SAF project, the database is extracted by entering commands

```
--> gunzip nwpsaf_cams.tar.gz
--> tar -xvf nwpsaf_cams.tar
```

in a unix/linux shell.

The profile data is given in a set of eight ASCII files. The data files are named according to the generic pattern

`nwpsaf_`{subset}`_sampled.dat`,

where {subset} identifies the subset of the database (`t` for temperature, `q` for humidity, `oz` for ozone mixing ratio, `co` for carbon monoxide mixing ratio, `no2` for nitrogen dioxide mixing ratio, `so2` for sulphur dioxide mixing ratio, `hcho` for formaldehyde mixing ratio, and `rand` for random).

The package also contains an example FORTRAN program `readsaf60.f90` intended to assist building

an interface for the use of the database. The example program is compiled using a Fortran 90 compiler (gfortran in this example), and the resulting executable is run from the command line

```
--> gfortran readsaf60.f90
--> ./a.out
```

During the run, the user is asked to enter identification for the subset that is to be read:

```
Enter the identification of the sampled variable:
- t (for temperature)
- q (for humidity)
- oz (for ozone)
- co (for carbon monoxide)
- so2 (for sulphur dioxide)
- no2 (for nitrogen dioxide)
- hcho (for formaldehyde)
- rand (for random)
```

In the end of a successful run, a confirmation is prompted on the screen:

```
 Number of profiles found in the file:  5000
```

As the example program does not produce any output files, users are encouraged to modify the program code according to their specific requirements.

Atmospheric and surface-related variables included in the CAMS profile database are listed in Table 2.

Acknowledgements

The authors have received funding for this work through the EUMETSAT NWP SAF Programme.

References

Chevallier, F., S. Di Michele, and A. McNally, 2006: Diverse profile datasets from the ECMWF 91-level short-range forecasts. *NWP SAF Report No. NWPSAF-EC-TR-010, 14 p.*

Eresmaa, R., A. Benedetti, and A. McNally, 2012: Diverse profile database of aerosol and trace gas concentrations from the Monitoring Atmospheric Composition and Climate short-range forecasts. *NWP SAF Report No. NWPSAF-EC-TR-015, 12 p.*

Eresmaa, R., and A. McNally, 2014: Diverse profile datasets from the ECMWF 137-level short-range forecasts. *NWP SAF Report No. NWPSAF-EC-TR-017, 12 p.*

Atmospheric variables (given on model levels)	
Variable name	Unit
Temperature	K
Specific humidity	kg kg^{-1}
Ozone mixing ratio	kg kg^{-1}
Carbon Monoxide mixing ratio	kg kg^{-1}
Sulphur Dioxide mixing ratio	kg kg^{-1}
Nitrogen Dioxide mixing ratio	kg kg^{-1}
Formaldehyde mixing ratio	kg kg^{-1}
Fractional cloud cover	
Cloud liquid water content	kg kg^{-1}
Cloud ice water content	kg kg^{-1}
Rain rate	kg m^{-2} s^{-1}
Snow rate	kg m^{-2} s^{-1}
Vertical velocity	Pa s^{-1}

Surface variables	
Variable name	Unit
Logarithm of surface pressure in Pa	
Surface geopotential	m2 s^{-2}
Surface skin temperature	K
2-meter temperature	K
2-meter dew point temperature	K
10-meter wind speed U component	m s^{-1}
10-meter wind speed V component	m s^{-1}
Surface albedo	
Roughness length	m
Snow temperature	K
Snow depth	m
Fractional cover of land (land/sea mask)	

Additional information	
Variable name	Unit
Latitude	$^{\circ}$
Longitude	$^{\circ}$
Forecast base year	
Forecast base month	
Forecast base day	
Forecast step	h
Grid point index	
Profile index	
Random selection flag	

Table 2: Variables included in the CAMS profile database.