Diverse profile datasets from the ECMWF 137-level short-range forecasts

Reima Eresmaa and Anthony P. McNally

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

Overview

Resolution upgrades and improvements in representation of physical processes in the forecasting system of the European Centre for Medium-range Weather Forecasts (ECMWF) justify production of a new version of the diverse profile database. We have compiled a new database including a representative collection of 25,000 atmospheric profiles from global operational short-range forecasts. The profiles are given in a 137-level vertical grid extending from surface up to 0.01 hPa. The database is divided in five subsets focussing on diverse sampling of temperature, specific humidity, ozone mixing ratio, cloud condensates, and precipitation. In contrast to earlier releases of the ECMWF diverse profile database, the 137-level database puts an increased emphasis on preserving statistical properties of sampled distributions (i.e., including a realistic amount of frequently-occurring atmospheric states). This is achieved by applying randomized selection to provide majority (90%) of profiles in the output database. Because of the modification in the sampling process, most parameters in the new database show reduced amount of variability, as compared with the previous 91-level database. The effect is partially compensated by the increased horizontal and vertical resolutions in the operational forecasting system.

1 Introduction

ECMWF has produced several diverse profile databases in the past. Recently, Eresmaa et al. (2012) released a database consisting of 60-level vertical profiles of aerosol and trace gas concentrations retrieved from numerical forecasts produced as part of the Monitoring Atmospheric Composition and Climate (MACC) project. Since the release of the previous diverse profile database for NWP applications (hereafter IFS-91 database; Chevallier et al., 2006), the operational forecasting system of ECMWF has undergone continuous development including several major modifications and improvements. Amongst other scientific and technical modifications, the operational system was upgraded in 2009 from T799 to the present-day resolution T1279 in horizontal (corresponds to transitioning from 25 to 16 km in terms of grid spacing) and, in 2013, from 91-level to the present-day 137-level discretization in vertical. In order to ensure practical relevance of applications based on the diverse profile databases, it has become necessary to introduce an update to reflect the recent operational changes.

The new IFS-137 database described in this document is produced using a method very similar to the one presented in Chevallier (2006), although increased emphasis is put on preserving statistical distributions of the sampled meteorological parameters. Consequently, the new database contains fewer extreme situations than the IFS-91 database, and distribution of each sampled variable is less homogeneous across its range of values. The applied sampling method is explained in Section 2, while Section 3 discusses the practical implementation to produce the IFS-137 database. Section 4 focuses on statistical characterization of profiles included in the IFS-137 database. Emphasis in this characterization is put on the comparison with the IFS-91 database.

2 Selection algorithm

The IFS-137 database is produced using a sampling method that is, in essence, very similar to the one described in Chevallier et al. (2006). We denote pools of input and output profiles as S_I and S_O, respectively. The selection process is started by creating an array that contains all input profiles in random order. The profile ranked first in the random array is saved in S_O. The process is then continued by repeatedly comparing the next candidate input profile with all those profiles already saved in S_O during preceding selection rounds. Formally, the comparison is based on squared inter-profile departure, that between profiles s_i and 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 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. Profile s_i is saved in S_O if the inequality

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

is true. The threshold value t is manually tuned such that number of profiles saved in S_O after considering all input profiles is as desired.

The method applied in the production of the IFS-137 database benefits from two extensions to the earlier method. Firstly, a quality control step has been introduced to prevent unphysical profiles from entering the output database. The quality control was found necessary because numerical representations of cloud and precipitation fields are often found to contain unrealistically small-scaled local features (so-called grid-point storms; see Fig. 1 for a typical example), that are very likely to be saved in the output pool because of their large departure from the vast majority of realistic profiles. The quality control identifies and rejects those grid points where the value of the sampled variable (on any level) differs from the field mean value by more than 25 standard deviations. Typically, while sampling cloud condensates or precipitation, this means rejecting some hundreds of grid points in a global field consisting of 2,140,702 grid points, whereas there are virtually no such rejections while sampling temperature, specific humidity or ozone mixing ratio.

Secondly, an option has been added to select a large number of profiles by random without the requirement on sufficient departure with respect to all other selected profiles. This is implemented simply by

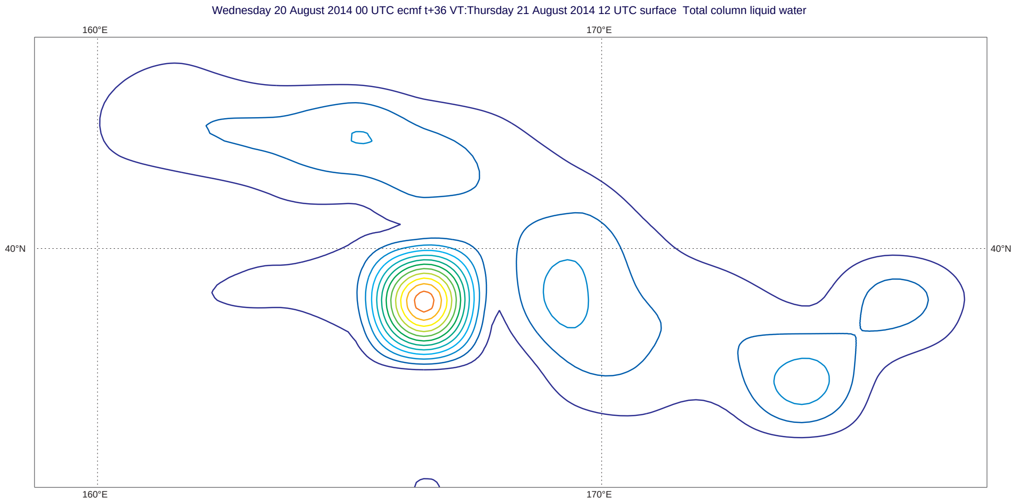
Wednesday 20 August 2014 00 UTC ecmf t+36 VT:Thursday 21 August 2014 12 UTC surface Total column liquid water
160°E
170°E
40°N
40°N
160°E
170°E

Figure 1: A grid-point storm in a 36-hour forecast of total column cloud liquid water. Contour interval is 0.4 kg m^{-2}.

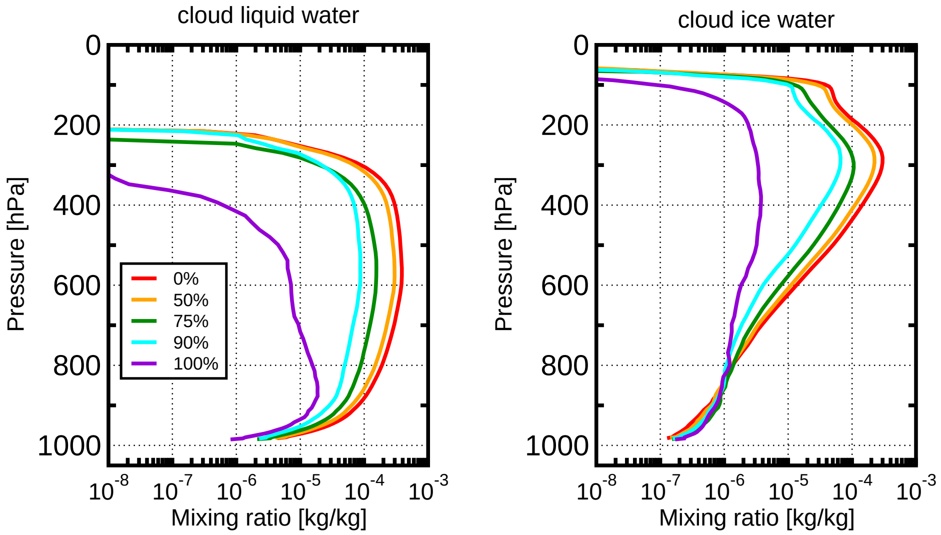

Figure 2: The effect of randomized selection on sample-mean profiles of cloud liquid water (left) and ice (right) content. Red, orange, green, blue, and violet curves refer to including 0%, 50%, 75%, 90%, and 100% of randomly-selected profiles in the output.

selecting and saving a fixed number of profiles from the top of the randomly-ordered input array. The purpose of the randomized sampling is to ensure that statistical properties of profiles in the output pool do not differ too much from those in the input pool. The effect of the randomized selection is illustrated in Fig. 2 for the case of selecting 5,000 profiles using column-integrated cloud condensates (cloud liquid water and cloud ice) as sampling variables. The most extreme strategies would be to either select all profiles by random (fully randomized selection without applying the selection algorithm at all; violet lines) or to rely only on the selection algorithm and not include any profiles by random (as was done during the production of the IFS-91 database; red lines). Given that the output sample is sufficiently large, the fully randomized selection accurately preserves statistical properties of the input pool. In contrast, relying only on the selection algorithm maximises homogeneity across the range of occurring values. Different compromises between the two extremes are shown in orange, green, and blue lines. Even if fraction of randomly-selected profiles is increased to 90% (i.e., only 10% of output profiles are allowed to pass the selection algorithm), sample-mean profiles are relatively unrepresentative of the input population.

3 Implementation

Similarly to the previous diverse profile databases, the IFS-137 database attempts to represent global and seasonal atmospheric variability as described by the operational forecasting system of ECMWF. The database is targeted to containing 25,000 profiles and it is divided in five equally-sized subsets that correspond to univariate (i.e., K=1 in Eq. (1)) sampling for (1) temperature, (2) specific humidity and (3) ozone mixing ratio and to bi-variate (K=2) sampling using column-integrated values of (4) cloud condensate and (5) precipitation.

The profile database is compiled from the short-range forecasts spanning the time period of 1 September 2013 – 31 August 2014. The forecasts are produced by the version Cy40r1 of the Integrated Forecasting System (IFS), that became operational at ECMWF on 19 November 2013 (earlier forecasts used in this work are from pre-operational testing period). The major challenge is to keep the final concise database representative of the initially huge input pool that contains billions of profiles produced by the operational forecasting system in a global grid during one year. There are two operational analyses each day (at 00z and 12z), and the modelling grid contains 2,140,702 grid points. Even if only four forecast steps are considered from each analysis time, number of available input profiles is around 6.2 $\times$ 10^9, while

	First round		*Second round*	
Subset	Threshold departure	Output count	Threshold departure	Output count
temperature	0.18	229,078	0.274	5,000
specific humidity	0.66	237,144	0.99	5,000
ozone	1.00	209,460	0.85	5,000
cloud condensate	1.80	218,223	0.0147	5,000
precipitation	2.50	215,423	0.220	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.

the target number for profiles to be included in the final database is only 25,000. As a first step towards reducing the data volume, the available input data is thinned in time dimension such that only forecasts originating from analyses valid at 00z on either 1st, 10th or 20th day of each month are considered. With regard to forecast lead times, only forecast steps 36, 42, 48, and 54 hours are considered. The temporal thinning is necessary primarily because of practical constraints, as it is time-consuming to retrieve a large number of full-resolution forecast files from the operational archive. As a result of the temporal thinning, number of distinct forecast valid times during one year is reduced to 144, while the total number of input profiles is reduced to 3.1×10^8. This is the total number of profiles processed during the production of the IFS-137 database.

Even after the temporal data thinning, the number of input profiles is far too large for a simple one-time execution of the selection algorithm. As a next production step, therefore, the selection algorithm is applied separately to each forecast valid time. During the first round of running the selection algorithm, the threshold parameter of Eq. (2) is specified such that approximately 1,500 profiles become selected from at forecast valid time. The quality control is applied as described in Section 2. Also, the option for randomized selection is applied to produce 90% of output profiles. The process is repeated separately for the five subsets of the database. Applied inter-profile departure threshold parameter values and counts of profiles remaining after the first selection step are shown in Table 1.

The final step of the process combines the output profiles from all forecast valid times into one input pool and applies the selection algorithm on this. During the final run, the quality control option is switched off, but the randomized selection is again applied to provide 90% of output profiles. Threshold parameters (shown in Table 1) are specified such that the output pool contains exactly 5,000 profiles for each subset.

4 Sampled distributions

4.1 Distribution of selected profiles in space and time

Locations of selected profiles in temperature, specific humidity, and cloud condensate subsets of the IFS-91 and IFS-137 databases are plotted on map in Fig. 3. In the IFS-91 database, the sampling is fully determined by the selection algorithm, which makes the geographical distributions very inhomogeneous. Selected profiles represent those regions where gradients of the sampled variable are the strongest: in the case of temperature, mid- and high latitudes dominate, while humidity and cloud condensate subsets concentrate to low latitudes. The IFS-137 database shows a much more homogeneous spatial distribution in all sampling subsets, which is a consequence of the randomized selection.

Temporal distribution of the selected profiles is illustrated in Fig. 4. Again, the lack of randomized selection results in large variations from one month to another in the case of the IFS-91 database (left panel). These are most strongly contributed by variations in the ozone subset (green parts of each column). Dominance of randomly-selected profiles in the IFS-137 database leaves only little room for monthly variations in the data count (right panel).

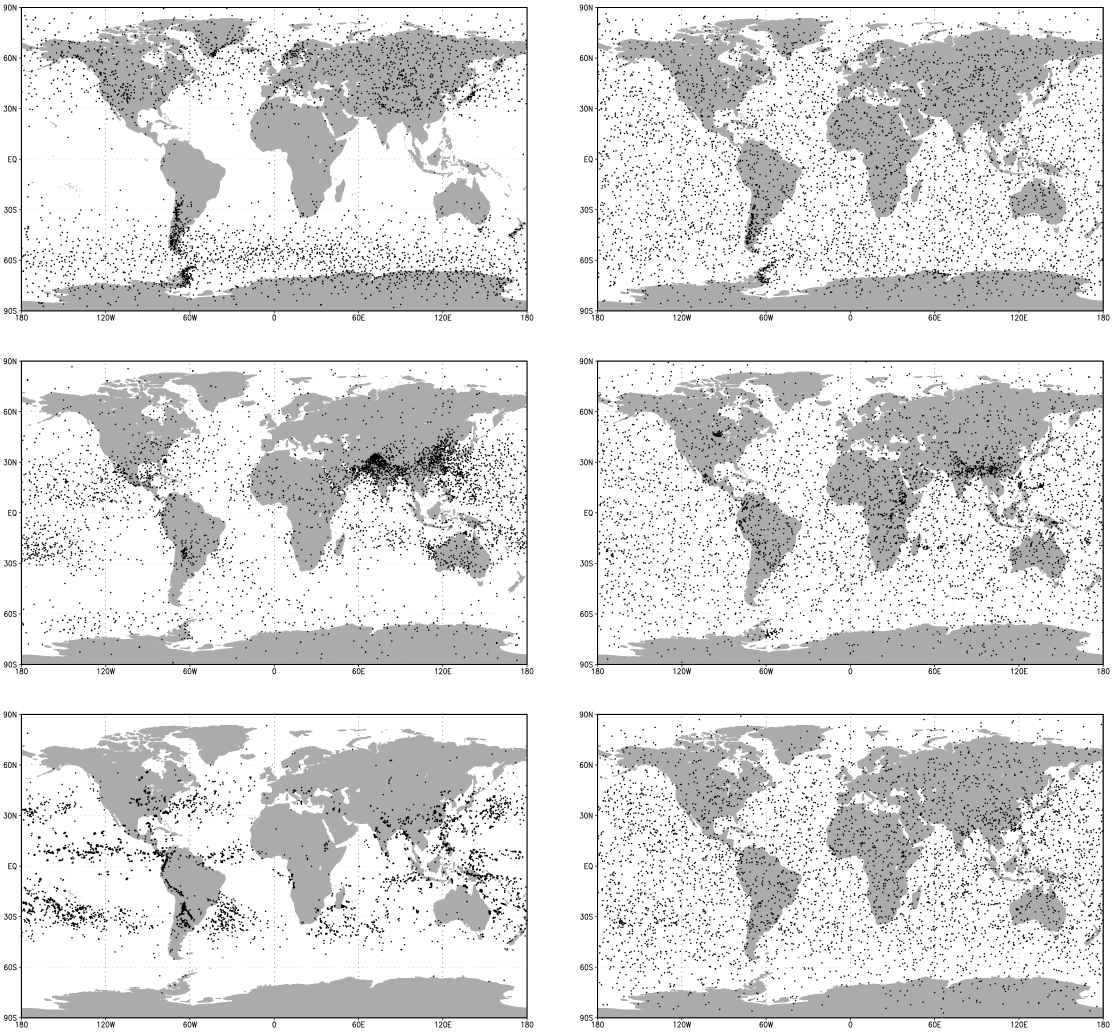

Figure 3: Locations of selected profiles in the temperature (top), specific humidity (middle), and cloud condensate (bottom) -sampled subsets of the IFS-91 (left) and IFS-137 (right) databases.

4.2 Distributions of sampled variables

Figures 5–8 show distributions of the sampling variables in each subset of the IFS-91 and IFS-137 databases. In each case, mean profile is shown by black solid line, while shaded regions show range determined by minimum and maximum values (gray), 10th and 90th percentiles (orange), and 25th and 75th percentiles (i.e., lower and upper quartiles, red). Because of the increased emphasis on preserving statistical properties in the output database, one would expect to see narrower ranges of values and distributions that are more confined around mean profiles in the new IFS-137 database (right panels) than in the old IFS-91 database (left panels), although this should be (at least partially) compensated by the effect of increased horizontal and vertical resolution in the input operational forecast model. In the case of sampling for temperature, the IFS-137 database shows slightly more variability in troposphere (top panels of Fig. 5), but considerably less variability in stratosphere (top panels of Fig. 6). The large number of tropical profiles contained in the new database (see Fig. 3) results in clearly defined tropopause in the mean profile and in bulk of the distribution being shifted towards warmer tropospheric temperatures.

Differences between the two databases are small in the case of specific humidity (bottom panels of Fig.

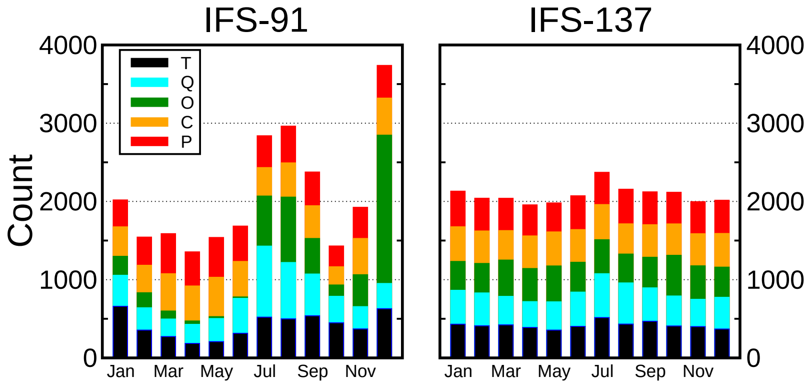

Figure 4: Distribution of profiles within calendar months in IFS-91 (left) and IFS-137 (right) databases. Different subsets are shown in different colours.

5). Because of the large percentage of randomly-selected profiles, very moist profiles are less common in the new database, resulting in a shift of mean towards drier profiles. Variable range appears to be slightly wider in the new database, but this is an artefact attributed to the logarithmic scale of the x-axis; in absolute sense, there is less variability in the new database than in the IFS-91 database.

The randomized selection has a major effect on the sampling for ozone mixing ratio. Lower panels of Fig. 6 show that the variable range in the IFS-137 database is only a fraction of that in the IFS-91 database, and also the mean profile appears smoother than before. The new database contains only a few profiles from polar late winter conditions, where the ozone concentration typically reaches its minimum.

Cloud condensate distributions shown in Fig. 7 highlight not only differences in the sampling method and grid resolution of the operational forecasting system, but also changes in physical parameterizations of cloud liquid water (top panels) and cloud ice (bottom panels). Including many randomly-selected profiles (some of which have no cloud liquid or ice water at all) shifts distributions consistently towards low mixing ratios, although extreme values are higher in IFS-137. The distribution of cloud condensates is highly non-gaussian. Because of the high sensitivity of mean to outliers, the mean profile consistently exceeds the upper quartile of the distribution. The most prominent effect from the modifications in physical parameterizations is that the new database contains some cloud liquid water above 300 hPa.

Non-gaussianity also affects shape of precipitation distributions shown in Fig. 8. Again, mean profiles are poor representations of the population as a whole. Extreme of rain rate (top panels) is roughly the same in the two databases, but that of snow rate (bottom panels) is lower in IFS-137 than in IFS-91. New physical parameterizations limit presence of rain below 500 hPa only.

4.3 Constituents of the database

The profiles included in the IFS-137 database are divided in five subsets of 5,000 profiles each, corresponding to sampling for temperature, specific humidity, ozone mixing ratio, cloud condensates, and precipitation rates. Meteorological and surface parameters included in the database are listed in Table 2. Additionally, each profile is given information to allow identification by geographical coordinates, forecast base and lead times, grid point index, and selection ranking number (1 for the first and 5,000 for the last selected profile in each subset). Vegetation type indices are explained in Table 3, and soil and ice layer depths in Table 4.

The constituents of the IFS-137 database are made as far as possible similar to those in the IFS-91 database. However, changes in archiving practice at ECMWF have forced us to make the following adjustments:

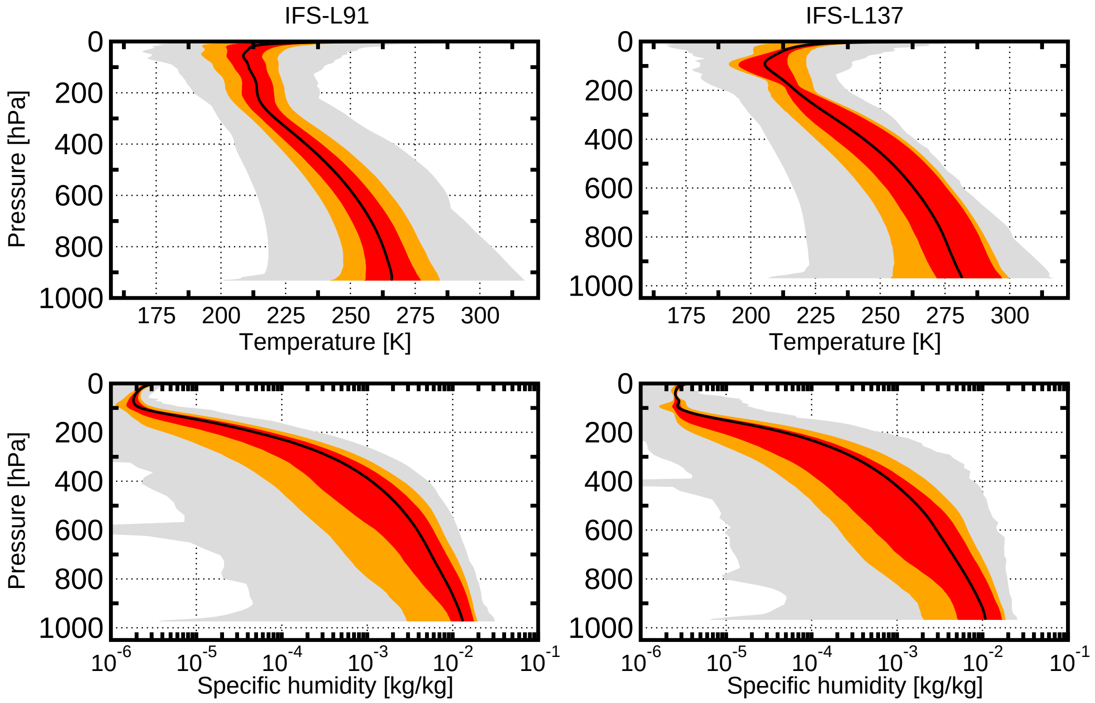

Figure 5: Distribution of temperature (top) and specific humidity (bottom) in the respective subsets of the IFS-91 (left) and IFS-137 (right) databases. Gray shading indicates the range constrained by minimum and maximum values, orange shading that constrained by 10th and 90th percentiles, and red shading that constrained by lower and upper quartiles (25th and 75th percentiles). Black solid line shows the mean profile.

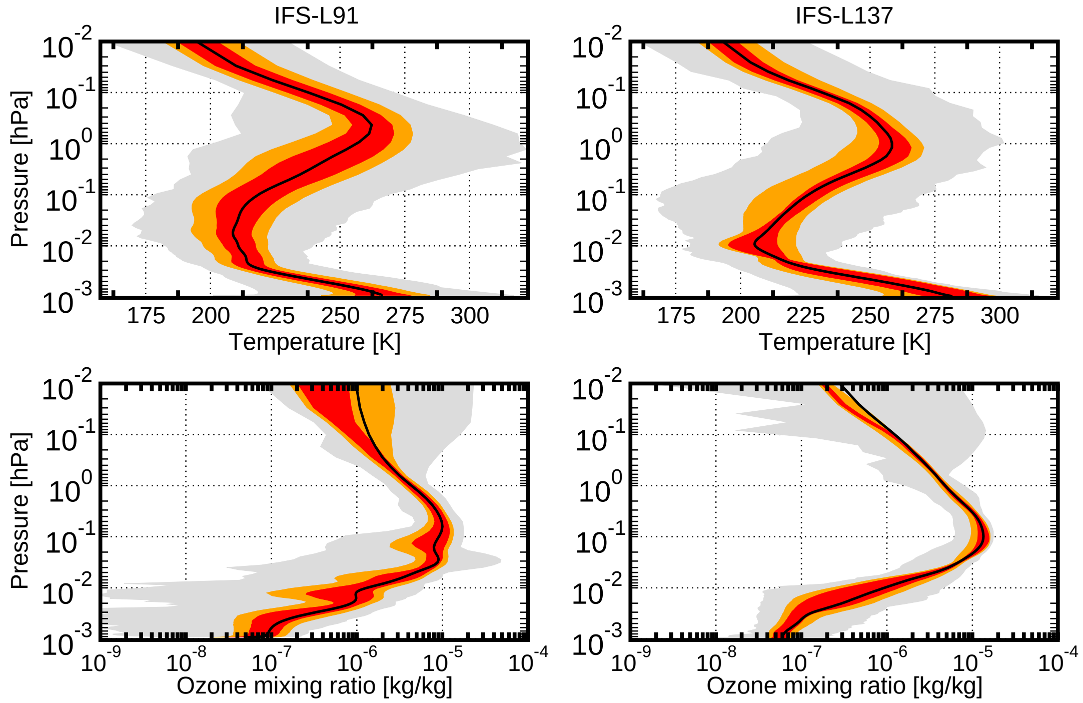

Figure 6: As Fig. 5, but for temperature (top) and ozone mixing ratio (bottom) in the respective subsets. Note the logarithmic y-axis.

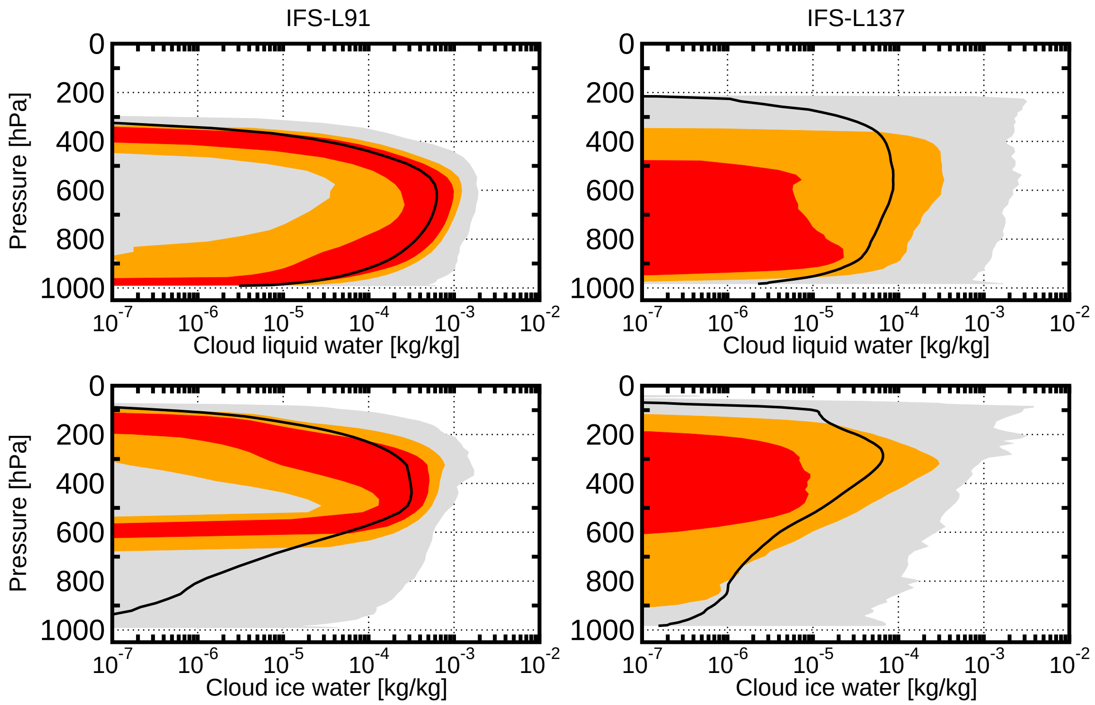

Figure 7: As Fig. 5, but for cloud liquid water (top) and cloud ice water (bottom) mixing ratios in the cloud condensate -sampled subsets.

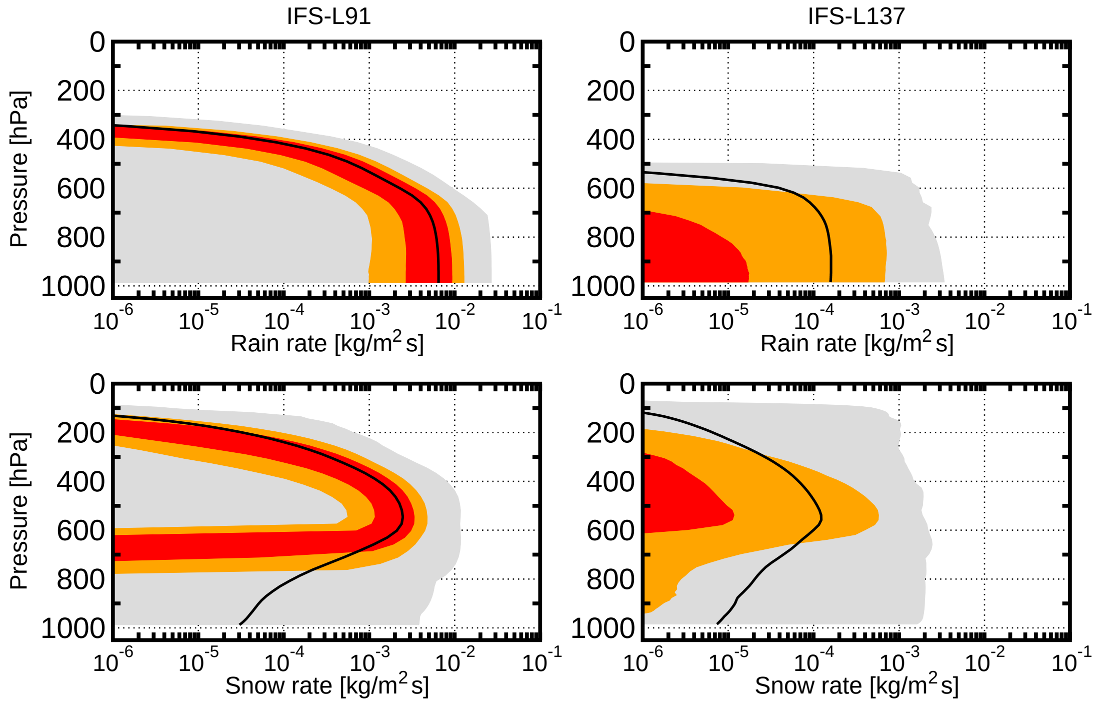

Figure 8: As Fig. 5, but for rain (top) and snow (bottom) rates in the precipitation-sampled subsets.

Atmospheric variables (given on model levels)	
Variable name	Unit
Temperature	K
Specific humidity	kg kg^{-1}
Ozone 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	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}
Stratiform precipitation at surface	m
Convective precipitation at surface	m
Snowfall (water equivalent) at surface	m
Fractional cover of land (land/sea mask)	
Surface albedo	
Roughness length	m
Type of low vegetation	
Fractional cover of low vegetation	
Type of high vegetation	
Fractional cover of high vegetation	
Fractional cover of sea-ice	
Snow albedo	
Snow density	kg m^{-3}
Snow temperature	K
Snow depth	m
Soil temperature (in four layers)	K
Volumetric soil water content (in four layers)	m^{3} m^{-3}
Ice temperature (in four layers)	K

Table 2: Atmospheric and surface parameters provided for each profile in the IFS-137 database.

Index	Vegetation type
1	Crops, Mixed Farming
2	Short Grass
3	Evergreen Needleleaf Trees
4	Deciduous Needleleaf Trees
5	Evergreen Broadleaf Trees
6	Deciduous Broadleaf Trees
7	Tall Grass
8	Desert
9	Tundra
10	Irrigated Crops
11	Semidesert
12	Ice caps and glaciers
13	Bogs and Marshes
14	Inland water
15	Ocean
16	Evergreen Shrubs
17	Deciduous Shrubs
18	Mixed Forest/woodland
19	Interrupted Forest
20	Water and Land Mixtures

Table 3: Vegetation type definitions.

Index	Soil	Ice
1	0–7	0–7
2	7–28	7–28
3	28–100	29–100
4	100–289	100-150

Table 4: Soil and ice layer (in cm) depths.

1. 2-meter specific humidity is no longer archived and is not included in the database. If needed, this parameter can be computed from 2-meter temperature and 2-meter dewpoint temperature.
2. Instead of providing instantaneous large-scale rain, convective rain and snow rates at the surface, we provide large-scale precipitation, convective precipitation, and snowfall amounts at the surface. These values are accumulated during the forecast and they are given in meters (of water equivalent in the case of snowfall).

5 Reading the database

The IFS-137 database is distributed as a compressed and tarred data file `profiles137.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 profiles137.tar.gz
--> tar -xvf profiles137.tar
```

in a unix/linux shell.

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

`nwp_saf_`{subset}`_sampled.`{part}`.gz`,

where {subset} identifies the subset of the database (`t` for temperature, `q` for humidity, `oz` for ozone mixing ratio, `ccol` for cloud condensate, and `rcol` for precipitation), and {part} is `atm` for files containing atmospheric profiles and `sfc` for files containing surface parameters only. Prior to reading the data, the compressed files need to be uncompressed by command

```
--> gunzip nwp_saf_*_sampled.*.gz
```

The package also contains an example FORTRAN program `readsaf137.f90` intended to assist building an interface for the use of the database. The example program is compiled using a Fortran 90 compiler (pgf90 in this example), and the resulting executable is run from the command line

```
--> pgf90 readsaf137.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)
- ccol (for cloud condensate)
- rcol (for precipitation)
```

A confirmation of a successful run will be printed on the screen as the reading of data files is finished:

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

As the example program does not make any output files, users are requested to modify the code according to their specific needs.

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.*