NWP SAF

Satellite Application Facility
for Numerical Weather Prediction

Diverse profile datasets from the ECMWF 91-level short-range forecasts

Frédéric Chevallier[1], *Sabatino Di Michele*[2] *and Anthony P. McNally*[2]

[1] *Laboratoire des Sciences du Climat et de l'Environnement, France*
[2] *European Centre for Medium–Range Weather Forecasts, UK*

Document No. NWPSAF–EC–TR–010

Version 1.0

December 2006

Diverse profile datasets from the ECMWF 91-level short-range forecasts

Frédéric Chevallier[1], Sabatino Di Michele[2] and Anthony P. McNally[2]

[1]Laboratoire des Sciences du Climat et de l'Environnement, France

[2]European Centre for Medium–Range Weather Forecasts, UK

This documentation was developed within the context of the EUMETSAT satellite Application Facility on Numerical Weather Prediction (NWP SAF), under the Cooperation Agreement dated 25 November 1998, between EUMETSAT and the Met Office, UK, by one or more partners within the NWP SAF. The partners in the NWP SAF are the Met Office, ECMWF, KNMI and Météo France.

Change record			
Version	Date	Author / changed by	Remarks

Abstract

This report summarises the characteristics of five databases that respectively sample the atmospheric temperature, water vapour, ozone, cloud condensate and precipitation simulated by the European Centre for Medium-Range Weather Forecasts system. Each database contains 5000 profiles described on 91 pressure levels. Their potential applications include statistical regressions, the provision of first-guesses for inversion algorithms and the validation of various models, in particular in the field of radiation.

1 Introduction

Building on the experience from the *Thermodynamic Initial Guess Retrieval* databases (TIGR: Chédin *et al.*, 1985; Escobar-Nunoz, 1993; Chevallier *et al.*, 1998), a series of diverse profile datasets from atmospheric simulations has been set up at ECMWF. Each one of them aims at providing a collection of representative cases, small enough to apply computationally expensive algorithms, like line-by-line radiation models. Obviously, each collection bears some of the qualities and weaknesses of the ECMWF forecasting system that produced them. Therefore, effort has been made to update the dataset so that it follows the continuous improvement in the modelling and the analysis of the atmosphere at ECMWF. Starting in 1998 and a version of the model that used 31 vertical pressure levels (Chevallier *et al.*, 2000), the dataset was renewed in 1999 and 2002 with respectively the 50-level and the 60-level versions of the system (Chevallier, 1999, 2002). The ECMWF operational system has been upgraded to 91 levels in February 2006 and a new release has been consequently made, which is described here. The sampling approach has been revised and is detailed in section 3, starting from a description of the previous sampling method in section 2. The new data are described in section 4.

2 Previous sampling strategy

The sampling strategy for the previous 60-level dataset was made of two parts. The first one consisted in filtering the infinity of possible profiles in the atmosphere, by gathering a much reduced sample of them. This initial database S was composed of 3D descriptions of the global atmosphere from the ECMWF 40-year re-analysis and included about 7 million profiles. The sampling of S with a topological approach was the second part of the method. It was iterative and relied on a distance D, that measured the dissimilarity between two atmospheric situations. At step one, a first atmospheric situation from S, s_1, was randomly drawn and archived in a new set E. At step i, an i^{th} atmospheric situation, s_i, was randomly drawn and archived in E if it was different enough from the already selected situations (i.e, if the distance D between the current profile and each one of the already-selected situations was larger than a predefined threshold d). The distance was defined as:

$$D(s_i, E) = \sum_{k=1}^{3} \mu_k D_k(s_i, E) \tag{1}$$

with:

$$D_k(s_i, E) \quad = \quad Min_{s_j \in E} \sqrt{\sum_{m=1}^{N} \left(\frac{\theta_{ik}(m) - \theta_{jk}(m)}{\sigma_{\theta k}(m)} \right)^2} \tag{2}$$

The μ_ks are predefined weights. N is the number of atmospheric pressure levels. k indicates one of the atmospheric variables among temperature, specific humidity and specific ozone. $\theta_{jk}(m)$ represents variable k at pressure level m for profile j. $\sigma_{\theta k}(m)$ is the standard deviation of $\theta_{jk}(m)$ in S.

This approach tends to cover the space of possible profiles with regularly spread samples. The size of the mesh is controlled by the sampling threshold d. The fact that extreme variabilities are as much selected as frequent ones reinforces the robustness of the regressions computed on the dataset.

Eventhough the sampling distance for the 60-level dataset only took temperature, humidity and ozone information into account, most of the variables archived from the original ECMWF simulations were provided as well in the delivered dataset.

3 Evolution of the sampling strategy

3.1 The new approach

Some recent work at ECMWF focused on cloud and precipitation, that motivated the development of another dataset with a somewhat different sampling methodology (Di Michele and Bauer, 2006). In order to homogenize the various datasets, the new release of the SAF diverse profile dataset had to take such variables into account in the sampling.

In principle, the method described above allows one to sample any selection of variables together by introducing the corresponding terms in Equation (1). In practice, the sampling results from a compromise between the sampling of the different variables. Adding more terms obviously degrades the distribution of each individual variable, without any benefit for users not interested in the representation of all the variables together.

As a consequence, it was decided to create as many datasets as there are types of variables to sample, with the same generic approach. To do that, we chose to apply the above-described one to the temperature profile, the humidity profile and the ozone profile separately. For these three datasets, Eq. (1) and (2) reduce to:

$$D(s_i, E) \quad = \quad Min_{s_j \in E} \sqrt{\sum_{m=1}^{N} \left(\frac{\theta_i(m) - \theta_j(m)}{\sigma_\theta(m)} \right)^2} \tag{3}$$

For cloud condensate and precipitation, the high variability of the vertical distribution of such variables does not seem to be that interesting to sample in comparison with the vertical columns per water phase. Therefore, we created two datasets for these two types of variables using:

$$D(s_i, E) \quad = \quad Min_{s_j \in E} \sqrt{\sum_{m=1}^{2} \left(\frac{\theta_i(m) - \theta_j(m)}{\sigma_\theta(m)} \right)^2} \tag{4}$$

with $\theta_i(m)$ the cloud condensate (respectively the precipitation) total column for liquid ($m = 1$) and solid water ($m = 2$).

3.2 Implementation

An initial database S was gathered using data from cycle 30R2 of the ECMWF forecasting system. The spectral model is truncated at wavenumber 799, which makes the horizontal resolution close to 25km. 91 pressure levels are used between 0.02hPa and the surface. The 3D description of the atmosphere was extracted from the 36-, 42-, 48- and 54-hour ranges of the forecasts that start at day 1, 10 and 20 of every month between July 2005 and June 2006. The data before February 2006 correspond to pre-operational experiments of the forecasting system. Such a set-up includes a total of 144 global snapshots of the atmosphere. Each snapshot is made of 843,490 profiles. Altogether, S contains 121,462,560 profiles. Testing the sampling set-up (i.e., mostly testing different d values) of this large dataset is tedious and the 144 individual snapshots were pre-sampled with *ad hoc* d values for each dataset. The characteristics of this preliminary phase are reported in Table 1.

Variable	T	q	oz	condensate	precipitation
Threshold	0.06	0.30	0.30	0.10	0.08
Selected	191,746	122,684	202,123	135,681	131,814

Table 1: Main characteristics of the preliminary sampling. For each dataset (temperature T, specific humidity q, specific ozone oz, cloud condensate and precipitation), the distance used and the number of selected profiles are indicated. The sampling operates on 144 files of 843,490 profiles.

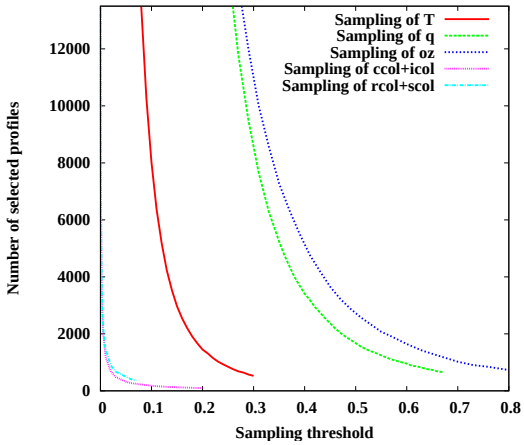

Figure 1: Number of selected profiles as a function of the sampling threshold d in the final sampling. The curve is shown for each one of the five datasets: temperature T, specific humidity q, specific ozone oz, cloud condensate $(ccol + icol)$ and precipitation $(rcol + scol)$.

The final sampling of the pre-sampled databases relied on the same algorithms. It seemed important to us to make the various datasets of the same size so that one can merge two or more of them with the same weight. **Figure 1** shows how the sampling threshold determines the number of selected profiles for each one of the datasets. To help the decision about this number, we investigated the variations of the standard deviation of the sampled variables. When sampling a single variable that is Gaussian-distributed, an increase of the standard deviation is expected with increasing thresholds because the sampling thins the population close to the mean. In our case, the variation is more complicated due to the interaction between the values at different altitudes in the case of profiles, and between the two water phases for the column values. A positive correlation between standard deviation and threshold is observed for the four cloud and precipitation variables (**Figure 3**). The opposite behaviour is seen for the profile variables with the smallest thresholds (**Figure 2**). In the case of specific humidity, some irregular variations are observed for the large thresholds (that select less than about 5000 profiles). However, it seems difficult to find any reliable objective criterion to choose d and practical considerations were favoured. We decided to set d so that exactly 5000 profiles were selected for each dataset. The characteristics of the final sampling are given in Table 2. **Figure 4** illustrate the sampling procedure by showing the closest profiles from the composite modal profile and the farthest ones from it, for the three of the final datasets.

4 Five new datasets

4.1 Available variables

Each situation in the five 91-level sampled database, is indexed by its space-time location:

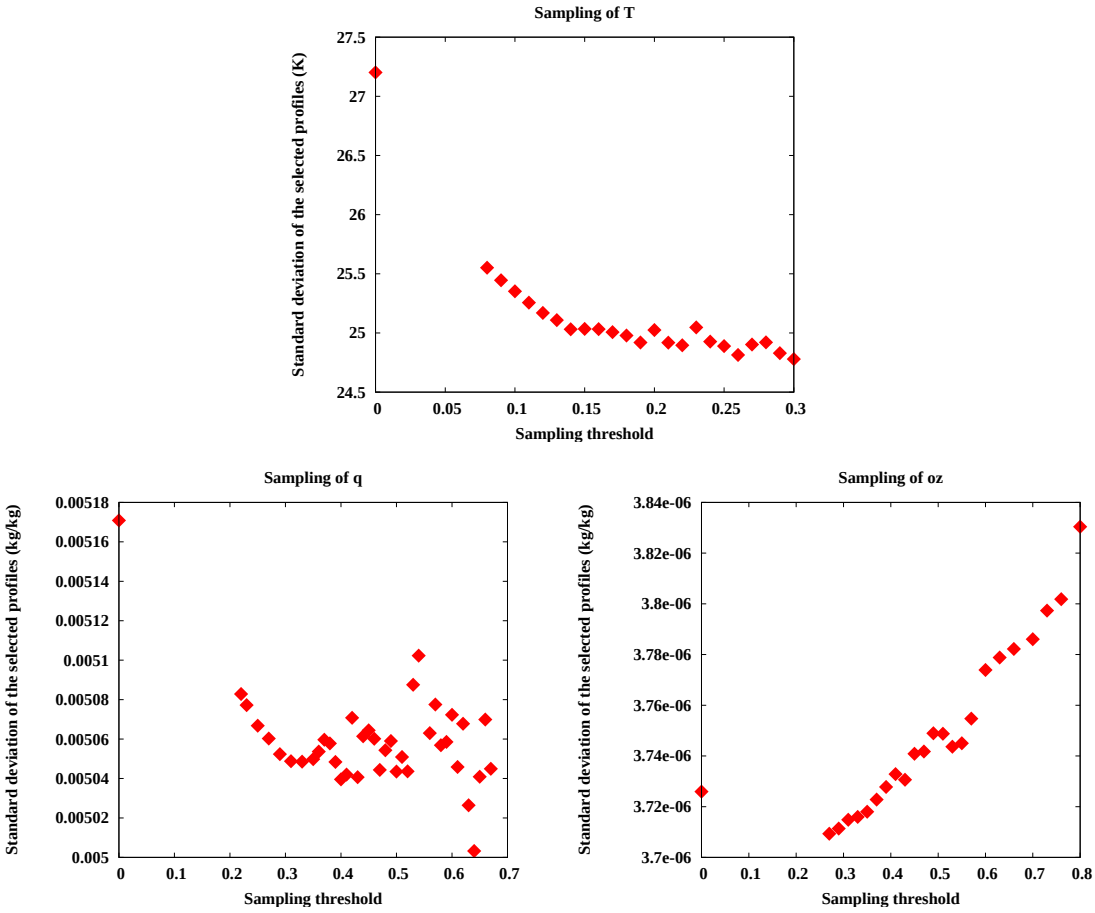

Figure 2: Standard deviation of the temperature T, specific humidity q and specific ozone oz, all pressure levels compined, as a function of the sampling threshold d in the corresponding datasets.

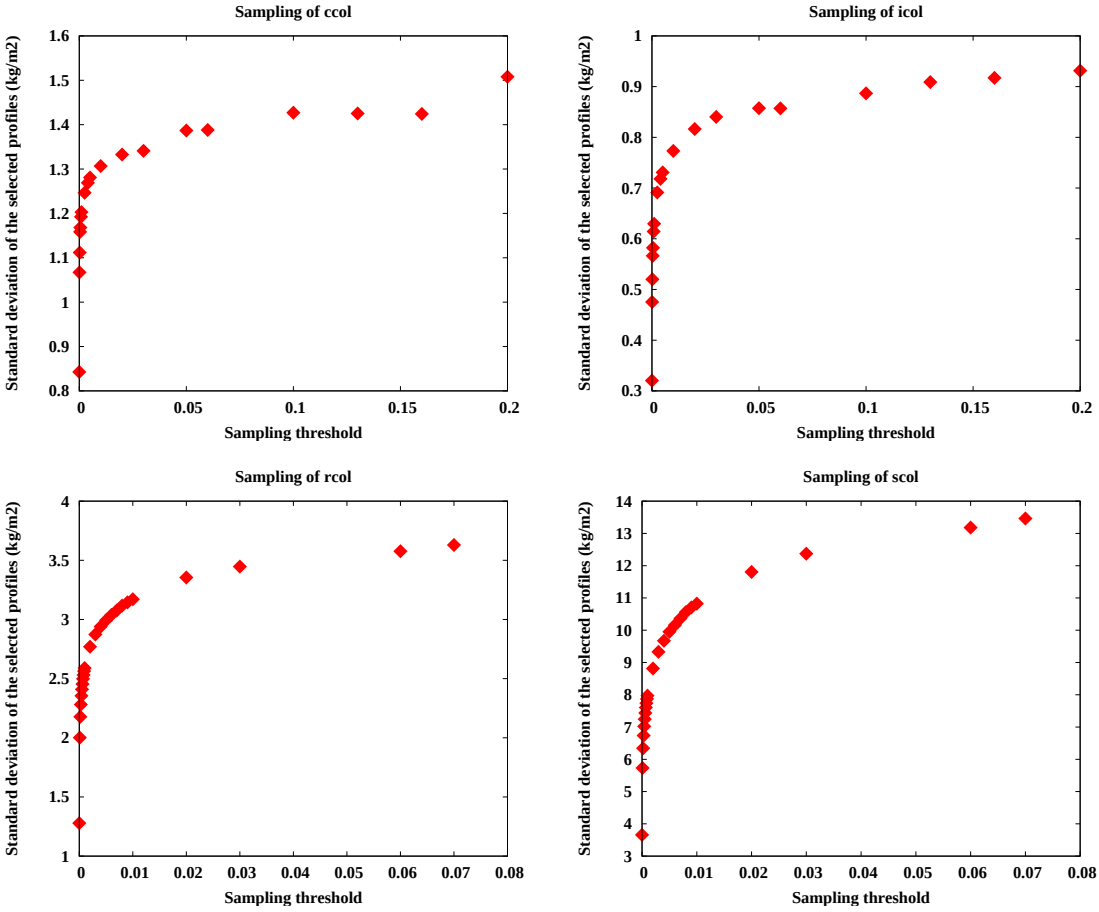

Figure 3: Standard deviation of the cloud liquid water total column *rcol*, of the cloud ice water total column *icol*, of the rain total column *rcol* and of the snow total column *scol* as a function of the sampling threshold d in the corresponding datasets. Note that *ccol* and *icol* are sampled together (see text). So are *rcol* and *scol*.

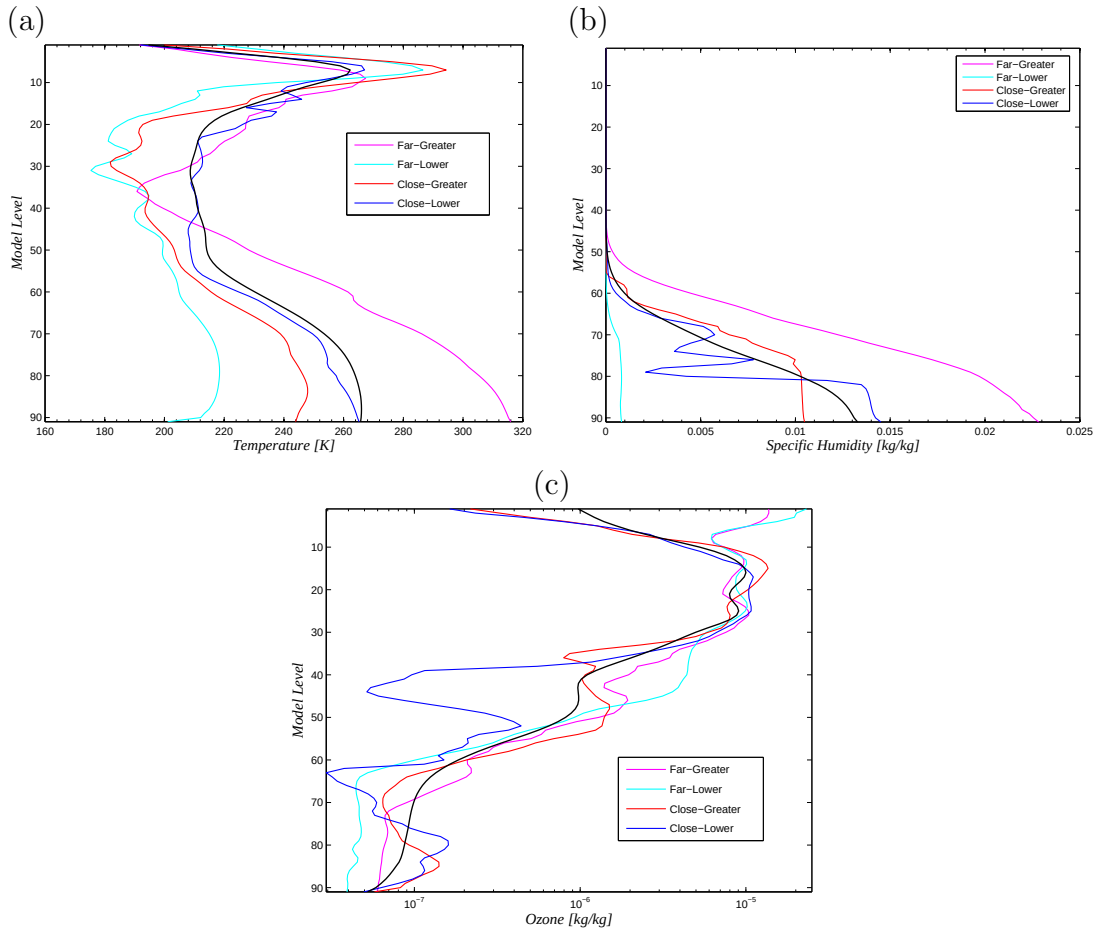

Figure 4: Composite profile of the modal values (thick black line), two closest and two farthest profiles from it, on the sides of the distribution. The distances are defined from Equation (3).

- the longitude, between -180° and 180°, eastward counted
- the latitude, between -90° and 90°
- the date (year, month, day, and synoptic hour) of the profile, actually defined from the date of the forecast start and from the time step of the forecast

As said before, the new datasets focus on:

- the atmospheric temperature, in K, on the 91-level grid
- the atmospheric specific humidity, in kg/kg, on the 91-level grid
- the atmospheric specific ozone, in kg/kg, on the 91-level grid
- the cloud liquid water, in kg/kg, on the 91-level grid
- the cloud ice water, in kg/kg, on the 91-level grid
- the rain, in kg/(m^2.s), on the 91-level grid
- the snow, in kg/(m^2.s), on the 91-level grid

The vertical pressure grid is a linear function of the surface pressure P_s. Indeed for each level l, the pressure $P(l)$ is expressed as: $P(l) = a_l + b_l P_s$. The pressure grid is illustrated at http://www.ecmwf.int/products/changes/high_resolution_2005/#model_levels_L91 . The minimum pressure is 2 Pa.

Other variables of the sampled situations have been extracted from the ECMWF archive and complete the database:

- the Neperian logarithm of the surface pressure (Pa)
- the surface geopotential (m^2.s^{-2})
- the surface skin temperature (K)
- the 2-meter temperature (K)

Variable	T	q	oz	condensate	precipitation
Threshold	0.122197	0.350505	0.400305	0.00157535	0.0015734
Selected	5000	5000	5000	5000	5000

Table 2: Main characteristics of the final sampling. For each dataset (temperature T, specific humidity q, specific ozone oz, cloud condensate and precipitation), the distance used and the number of selected profiles is indicated. The sampling operates on the selected profiles of Table 1.

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

Table 3: Definition of the vegetation types in the ECMWF forecasting system, from http://www.ecmwf.int/research/ifsdocs/CY28r1/Physics/Physics-08-03.html .

- the 2-meter dewpoint temperature (K)
- the 2-meter specific humidity (kg/kg)
- the 10-meter u and v components of the wind (m/s)
- the land fraction (0 corresponds to sea-only points)
- the stratiform rain at the surface (kg/(m^2.s))
- the convective rain at the surface (kg/(m^2.s))
- the snow at the surface (kg/(m^2.s))
- the cloud cover, on the 60-level grid
- the vertical velocity, in Pa/s, on the 60-level grid
- the type (see Table 3) and cover of low vegetation
- the type (see Table 3) and cover of high vegetation
- the temperature (K) and volumetric water (m^3/m^3) in four soil layers. Downward from the surface, the depth of the layers is successively: 7, 21, 72 and 189 cm.
- the ice cover and its temperature (K) in four layers. Downward from the surface, the depth of the layers is successively: 7, 21, 72 and 50 cm.
- the snow temperature (K), depth (m), density ($kg.m^{-3}$) and albedo (0-1)
- the surface albedo (0-1)
- the surface roughness (m)
- the index of the gridpoint on the ECMWF Gaussian grid
- The distance from the mean profile from Equation (3) or (4)

The samplings were performed on the ECMWF model vertical layers and not on fixed pressure layers. As a consequence, the sampled databases gather profiles corresponding to various ocean conditions as well as to land conditions, including high elevated grounds. The lowest surface pressure in the databases are 479 hPa and the highest 1052 hPa.

4.2 Statistical distribution of the variables

The histograms and some statistical characteristics (mean, mode, minimum and maximum values per model level) of the databases are shown in **Figure 5**.

The profile characteristics share some similarities with the previous 60-level dataset (see Figure 7 in Chevallier, 2002), but some differences can be noted. The first obvious feature is the expected increased vertical variability, in particular for ozone. This will be quantified in the next paragraph. Further, compared to the 60-level dataset, the distribution of the 91-level temperature database appears to be shifted toward colder values in the troposphere and the 91-level humidity database is shifted toward wetter profiles. This change was caused by the separation of the temperature sampling and of the humidity sampling, that avoided an artificial compromise between temperature variability (largest for cold temperature) and humidity variability (largest for humid profiles).

Principal Component Analyses have been performed on the temperature, humidity and ozone fields of the corresponding databases, in order to compare the vertical resolution with that of the previous ECMWF diverse profile datasets. The cumulated variance as a function of the number of leading eigenvectors is presented in **Figure 6**. The temperature and ozone plots (Figure 6a and 6c) illustrate the increasing resolution obtained when increasing the number of levels from 50 to 60 and then to 91. The humidity plot also shows improvement when going from 50 to 91 levels, but the 60-level version has less variability than the 50-level one. This feature was discussed in Chevallier (2002) and attributed to the relatively low horizontal resolution of the 60-level simulations used (125 km) compared to the 50-level ones (60 km).

4.3 Availability

The five datasets are available from the NWP-SAF[1]. All comments or questions can be sent to A. P. McNally[2].

They are provided in the form of ten ASCII files :

- nwp_saf_ccol_sampled.atm : part one of the dataset sampled for cloud condensate
- nwp_saf_ccol_sampled.sfc : part two of the dataset sampled for cloud condensate
- nwp_saf_oz_sampled.atm : part one of the dataset sampled for ozone
- nwp_saf_oz_sampled.sfc : part two of the dataset sampled for ozone
- nwp_saf_q_sampled.atm : part one of the dataset sampled for humidity
- nwp_saf_q_sampled.sfc : part two of the dataset sampled for humidity

[1] http://www.metoffice.com/research/interproj/nwpsaf/rtm
[2] Tony.McNally@ecmwf.int

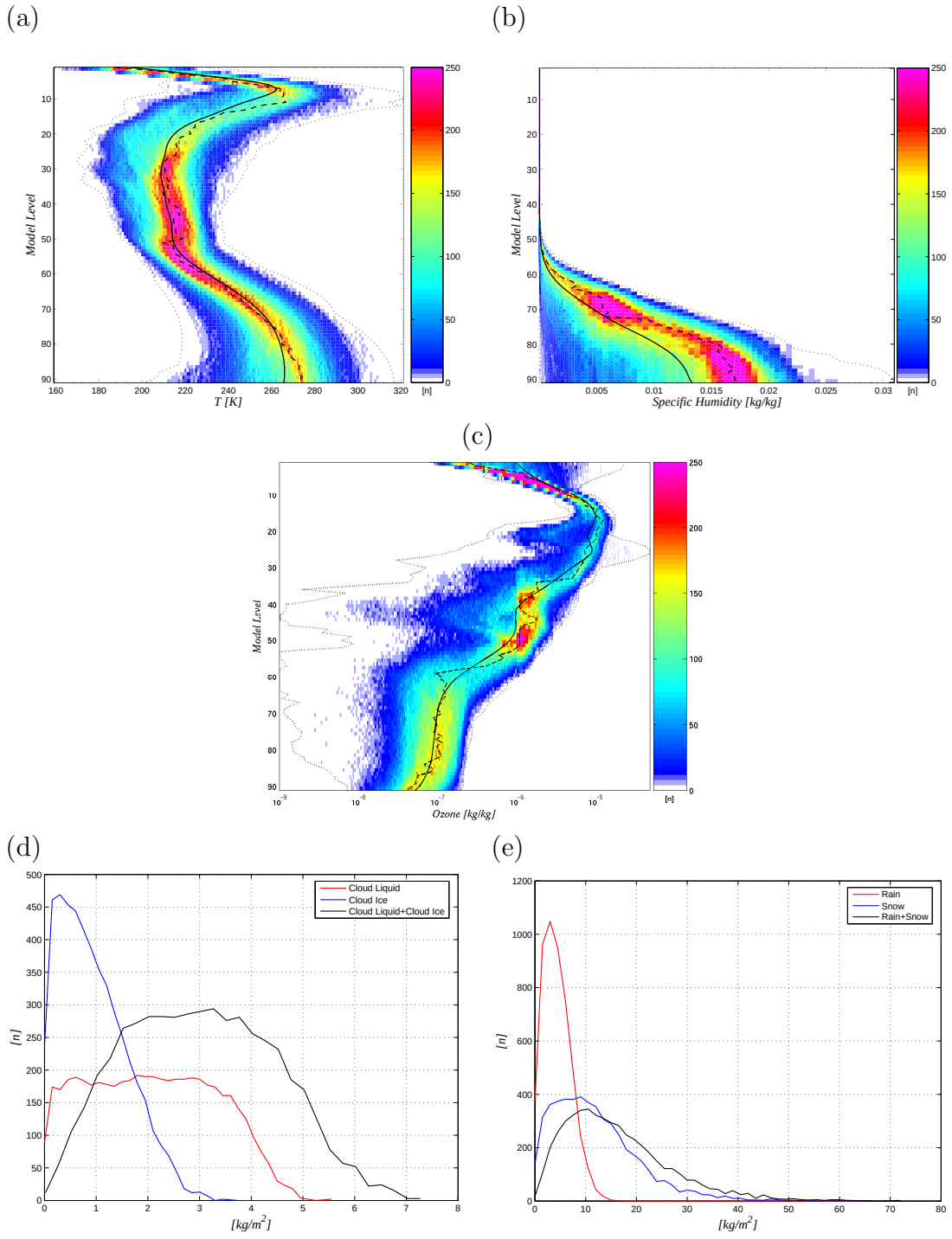

Figure 5: Histograms of the sampled databases for temperature (a), specific humidity (b), ozone (c), cloud condensate (d) and precipitation (e). Figures (a)-(c) also show the composite profiles of the minimum (left dotted line), the maximum (right dotted line), the mean (thick black line) and the mode (dashed line). The profiles are displayed on model levels. The surface pressure bounds vary from one dataset to another. For temperature, the pressure of the surface varies between 479 hPa and 1052 hPa.

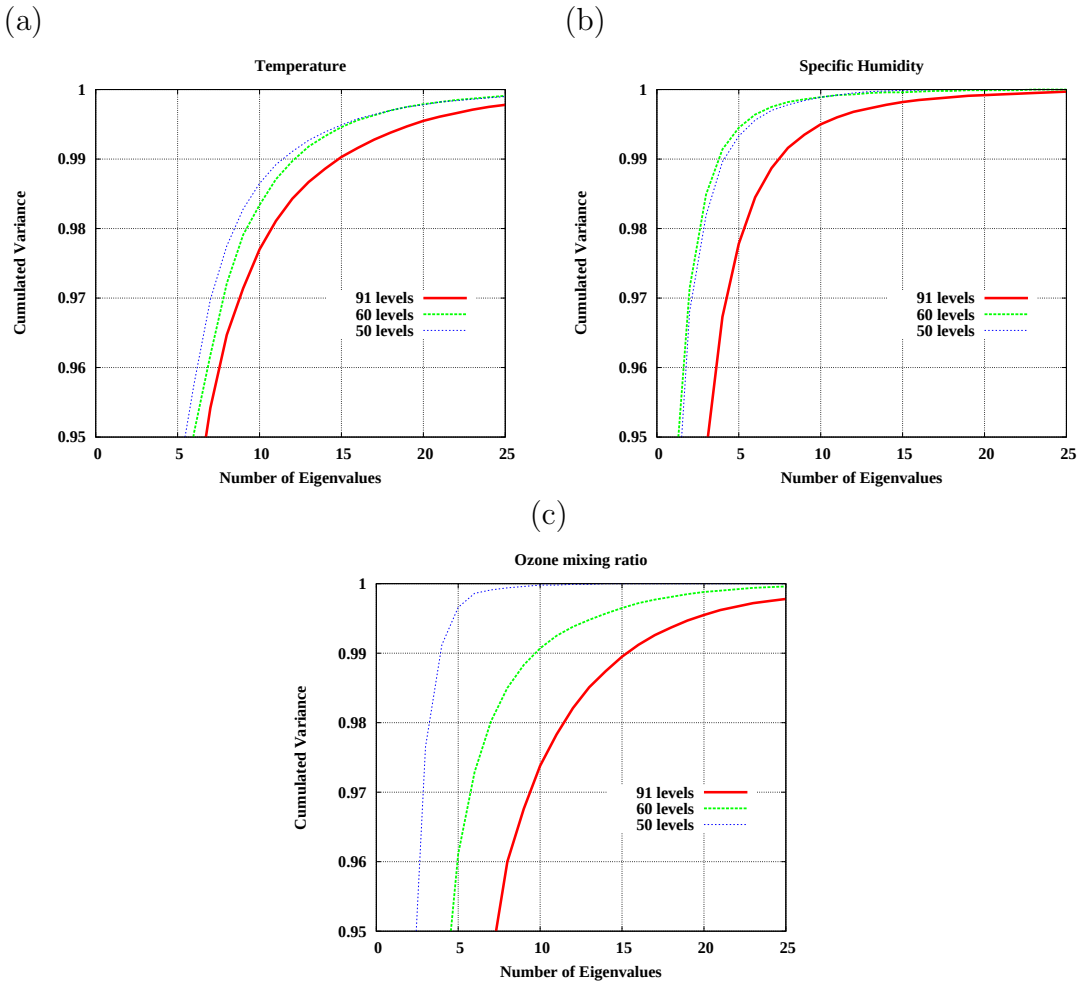

Figure 6: Cumulated variance as a function of the number of leading eigenvalues in the Principal Component Analysis of the temperature (a), specific humidity (b) and specific ozone (c) fields for the 50-level dataset, the 60-level dataset and the corresponding 91-levels datasets.

- nwp_saf_rcol_sampled.atm : part one of the dataset sampled for precipitation
- nwp_saf_rcol_sampled.sfc : part two of the dataset sampled for precipitation
- nwp_saf_t_sampled.atm : part one of the dataset sampled for temperature
- nwp_saf_t_sampled.sfc : part two of the dataset sampled for temperature

A FORTRAN program, *readsaf91.f90*, that demonstrates how to read them is provided.

These sampled databases presented here should not be considered as final ones. They carry both qualities and weaknesses from the ECMWF assimilation-forecast system. Further improvements of the system will enable further improvements of the databases.

Acknowledgements

This work was initiated during a stay of the first author at ECMWF as a NWP-SAF visiting scientist in June 2006. The authors wish to acknowledge the support of and the fruitful interaction with Peter Bauer and Jean-Noël Thépaut.

References

Chédin, A., N. A. Scott, C. Wahiche and P. Moulinier, 1985: The Improved Initialization Inversion method : a high resolution physical method for temperature retrievals from satellites of the TIROS-N series. *J. Climate Appl. Meteor.*, 24, 128-143.

Chevallier, F., 1999: TIGR-like sampled databases of atmospheric profiles from the ECMWF 50-level forecast model. *NWP SAF Report No. NWPSAF-EC-TR-001, 18 p.*

Chevallier, F., 2002: Sampled databases of 60-level atmospheric profiles from the ECMWF analyses. *NWP SAF Report No. NWPSAF-EC-TR-004, 27 p.*

Chevallier, F., F. Chéruy, N. A. Scott, and A. Chédin, 1998b: A neural network approach for a fast and accurate computation of longwave radiative budget. *J. Appl. Meteor.*, 37, 1385-1397.

Chevallier, F., A. Chédin, F. Chéruy, J.-J. Morcrette, 2000: TIGR-like atmospheric profile databases for accurate radiative flux computation. *Q. J. R. Meteor. Soc., 126, 777-785.*

Di Michele, S., and P. Bauer, 2006: Passive microwave radiometer channel selection based on cloud and precipitation information content estimation. *Quart. J. Roy. Meteor. Soc.*, 132, 1299-1324.

Escobar-Munoz, J., 1993: Base de données pour la restitution de variables atmosphériques à l'échelle globale. Étude sur l'inversion par réseaux de neurones des données des sondeurs verticaux atmosphériques satellitaires présents et à venir. PhD thesis, Univ. Paris VII, 190 pp. [Available from LMD, Ecole Polytechnique, 91128 Palaiseau cedex, France].