

# NWP SAF

## *Satellite Application Facility for Numerical Weather Prediction*

Document NWPSAF-EC-VS-013

Version 1.0

30 June 2006

### 91-level ECMWF diverse profile dataset: assessment

Frédéric Chevallier,

LSCE, France



|                |   |  |
|----------------|---|--|
| <b>NWP SAF</b> | <b>91-level ECMWF<br/>diverse profile dataset</b> | Doc ID : NWPSAF-EC-VS-013<br>Version : 1.0<br>Date : 30.6.06 |
|----------------|---|--|

## 91-level ECMWF diverse profile dataset: assessment

Frédéric Chevallier,  
LSCE, France

This documentation was developed within the context of the EUMETSAT Satellite Application Facility on Numerical Weather Prediction (NWP SAF), under the Cooperation Agreement dated 16 December, 2003, 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.

**Copyright 2006, EUMETSAT, All Rights Reserved.**

# NWP SAF: Visiting Scientist Report

## 91-level ECMWF diverse profile dataset

Frédéric Chevallier, LSCE, France

Host Institute: ECMWF (UK)  
Contact Point: A.P. McNally  
Duration: 26-30 June 2006

### 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 is consequently planned.

The preparation of the new release motivated this mission in June 2006. It consisted in revising the sampling approach and implementing it. This work is reported in section 3, starting from a description of the previous sampling method in section 2.

### 2 Previous sampling strategy

The sampling strategy for the 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:

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

with:

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

The  $\mu_k$ s 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 the regressions computed on the dataset.

Note that 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

Some recent work at ECMWF focused on cloud and precipitation, that motivated the development of an other 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 should 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 is suggested to create as many datasets as there are types of variables to sample, with the same generic approach. To do that, we choose to apply the above-described separately to the temperature profile, the humidity profile and the ozone profiles. For these three datasets, Eq. (1) and (2) reduce to:

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

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 to the vertical columns per water phase. Therefore, we create two datasets for these two types of variables using:

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

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

## 4 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 by S. Di Michele 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-operationnal 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. A FORTRAN program and a kornshell script were developed to perform this task. The characteristics of this preliminary phase are reported in Table 1.

The final sampling of the pre-sampled databases was prepared with similar softwares, then tested, but has not been finalized yet because no decision has been made about the number of profiles to select. It seems important to make the various datasets of the same size so that one can merge two or more of them with the same weight, but the question remains open about the appropriate number.

| 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 is indicated. The sampling operates on 144 files of 843,490 profiles.

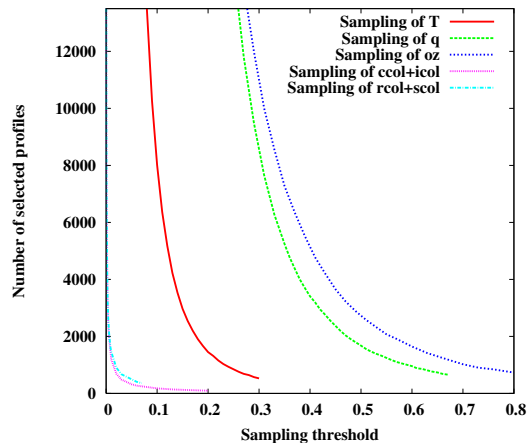


Figure 1: Number of selected profiles as a function of the sampling threshold  $d$  in the final sampling. The curve is shown for each one of the five datasets: temperature  $T$ , specific humidity  $q$ , specific ozone  $oz$ , cloud condensate ( $ccol + icol$ ) and precipitation ( $rcol + scol$ ).

Figure 1 shows how the sampling threshold determines the number of selected profiles for each one of the datasets. These curves will be used to choose the thresholds given the desired number of profiles. 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 (Figures 2). In the case of specific humidity, some irregular variations are observed for the large thresholds (that select less than about 5000 profiles).

So far, it seems difficult to find any reliable objective criterion to choose  $d$  and practical considerations from the database users should be favoured.

## Acknowledgements

This project results from the fruitful interaction between several people at ECMWF and me: P. Bauer, S. Di Michele, A.P. McNally and J.-N. Thépaut. The technical implementation was shared between S. Di Michele and me.

## 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

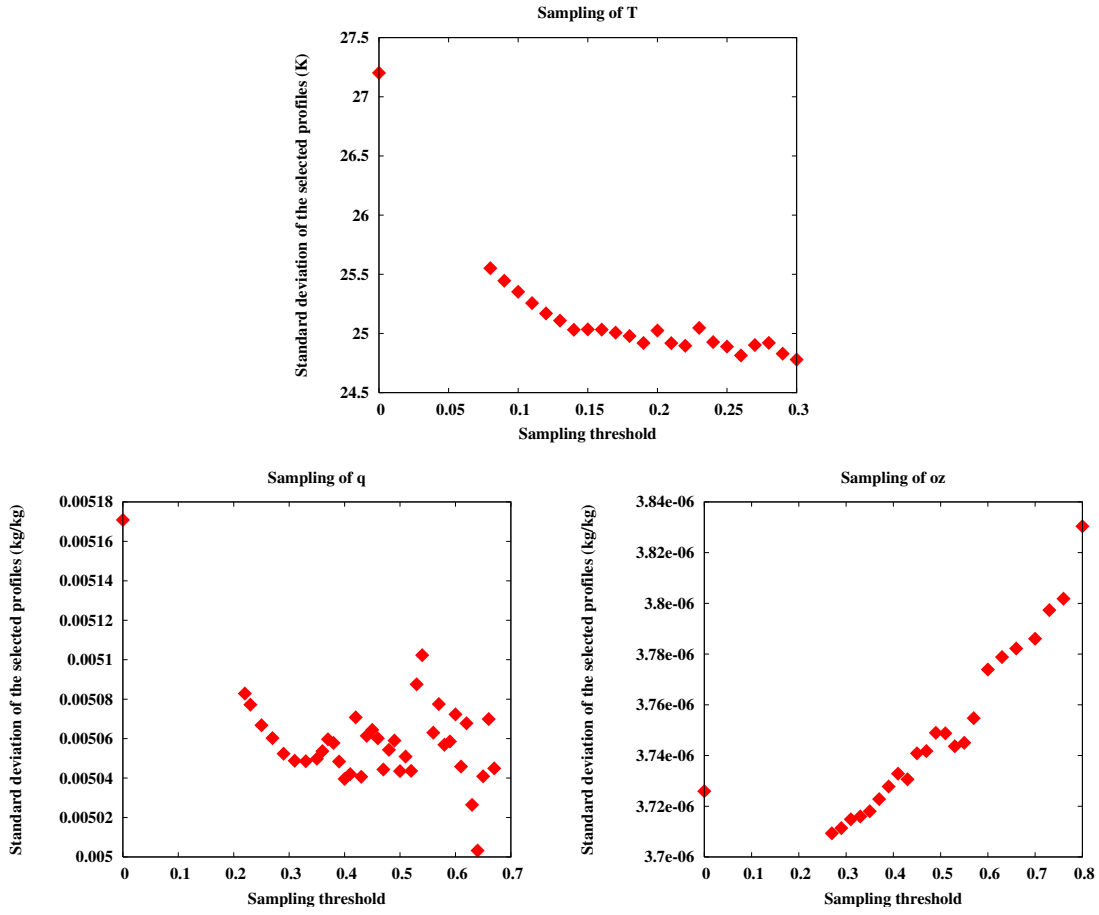


Figure 2: Standard deviation of the temperature  $T$ , specific humidity  $q$  and specific ozone  $oz$ , all pressure levels compined, as a function of the sampling threshold  $d$  in the corresponding datasets.

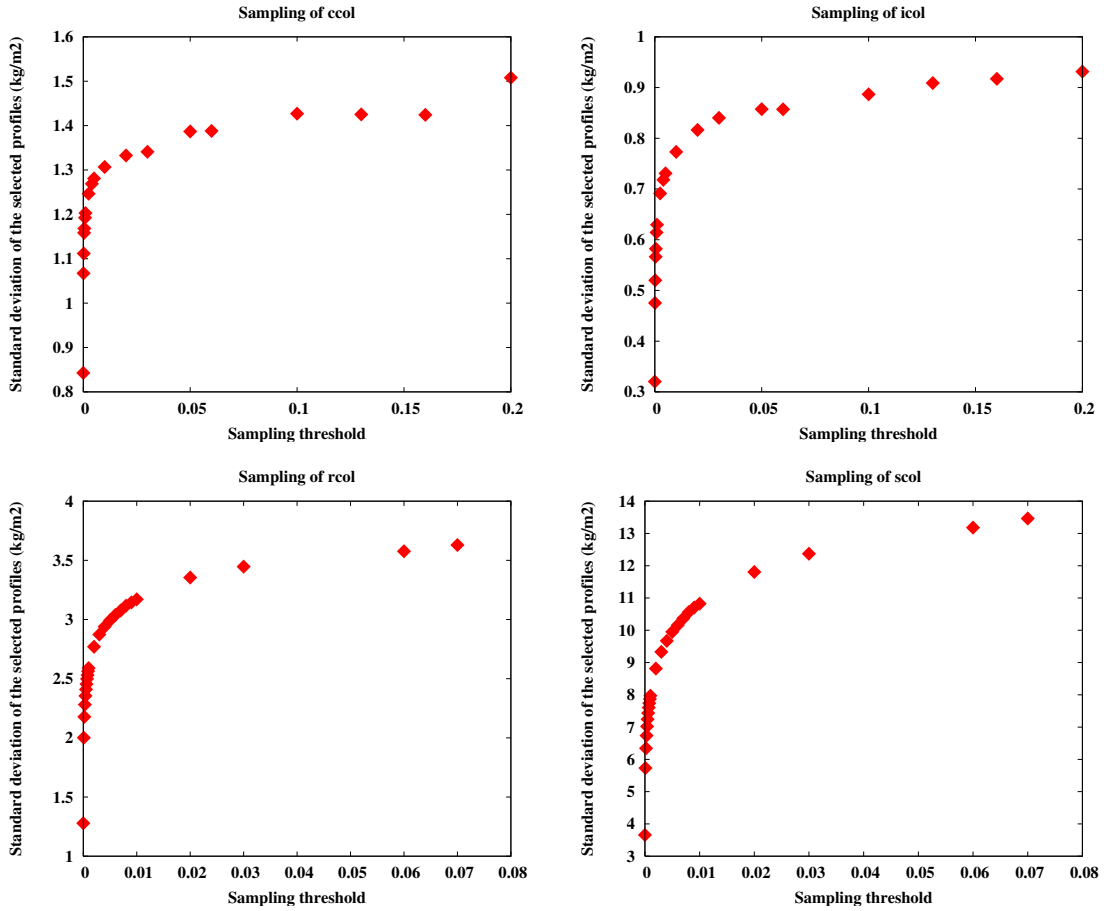


Figure 3: Standard deviation of the cloud liquid water total column  $rcol$ , of the cloud ice water total column  $icol$ , of the rain total column  $rcol$  and of the snow total column  $scol$  as a function of the sampling threshold  $d$  in the corresponding datasets. Note that  $ccol$  and  $icol$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



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. Chérut, 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. Chérut, 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].
- Moulinier, P., 1983: Analyse statistique d'un vaste échantillonnage de situations atmosphériques sur l'ensemble du globe. *LMD Internal note 123*, 30 pp., in French [Available from LMD, Ecole Polytechnique, 91128 Palaiseau cedex, France].