Numerical Weather Prediction

Satellite Application Facility

SAF Programme

Research Report No. 1

TIGR-like sampled databases of atmospheric profiles from the ECMWF 50-level forecast model

by

F. Chevallier

October 1999

Abstract

This report summarizes the characteristics of two databases, that sample the temperature and water vapour profiles simulated by the European Centre for Medium-Range Weather Forecasts system. The first one contains 13766 atmospheric situations and has been successfully used for non-linear regressions. The second one contains less than 200 situations in order to suit the applications that require a higher degree of sampling of the database. Both databases are available through the EUMETSAT Satellite Application Facility.

1 Introduction

Collections of atmospheric profiles are needed for many different applications, for example the modelling of atmospheric processes, like radiation, or the initialization of retrieval schemes. In practice, the sampling of associated variables (temperature, moisture, ...) is made difficult by the high dimension of the model space. A major attempt to sample this kind of a priori information on global scales has been the constitution of successive versions of the *Thermodynamic Initial Guess Retrieval* database (TIGR: Chédin *et al.*, 1985 ; Achard, 1991 ; Escobar-Nunoz, 1993 ; Chevallier *et al.*, 1998b) from Laboratoire de Météorologie Dynamique (LMD). Each version groups together hundreds of soundings sampled from larger databanks of observations of the atmosphere: radiosonde reports and, for the latest version (TIGR-3) only, satellite-retrieved atmospheric profiles. Recent applications of the neural network-based techniques, like the computation of longwave flux profiles in General Circulation Models (Chevallier *et al.*, 1998a) have led to further improvements of the sampling strategy initiated in the TIGR databases. Automatic procedures, as described by Chevallier *et al.* (1999a), have been developed and applied to the sampling of profiles simulated by the ECMWF forecast system.

This report summarizes the characteristics of two sampled databases that correspond to the ECMWF model in its 50-level vertical resolution. The general sampling method is described in section 2. Its application for the sampling of atmospheric situations from the ECMWF short-range forecasts is shown in section 3. Section 4 presents a further sampling of the database, so as to reduce it from 13,766 situations to less than 200. Section 5 provides an overall summary.

2 The sampling technique

2.1 General description

As in the TIGR databases, the sampling strategy is a two-step method. The first step consists in filtering the infinity of possible profiles in the atmosphere, by gathering a much reduced but representative sample of them. Let us call S this initial database. The sampling of S with a topological approach is the second step of the method. It relies on an index I, that measures the dissimilarity between two atmospheric situations. The process is iterative. At step one, a first atmospheric situation from S is randomly drawn and archived in a new set E. At step n, a n^{th} atmospheric situation is randomly drawn and archived in E if it is different enough from the already selected situations, relatively to index I. With that approach, the distribution of E over the space of the various atmospheric variables is smoother than that of

S. In practice, restriction to some variables has to be made. In the following, temperature and water vapour are taken into account. The reader is referred to Chevallier *et al.* (1999a) for the details of the method.

2.2 The sampling of the ECMWF model profiles

This sampling strategy is used for the sampling of profiles generated from the ECMWF atmospheric model. Two databases are gathered in this way. The first one corresponds to the 31-layer vertical grid (from the top of the atmosphere to the surface) that has been used at ECMWF between 1991 and 1998. The second one uses the 50-layer vertical grid that has been introduced into operation in March 1999. Only the second one will be referred to in the following.

In this application, S mainly results from the aggregation of eleven days of profiles from the ECMWF short-range (6-hour) forecasts. Each day is selected from a different month of data. It includes a complete description of the atmosphere on the corresponding vertical grid (50 levels) and an horizontal $1.125° \times 1.125°$ grid representation every six hours. S consists of about 1,500,000 profiles. It is divided into seven subgroups differing by the total precipitable water vapour content of the profiles: the first group ranges from 0 to 0.5 cm, the second from 0.5 to 1.5 cm, the third from 1.5 to 2.5 cm, and so on, until the seventh one that goes from 5.5 cm up to the highest values.

The sampling approach described above is used for the extraction of about the same number (N) of samples from each class, except for the first one, where twice as many ($2 \times N$) profiles are extracted, in consideration of the higher temperature variability: this class includes all types of situations from polar to tropical. N determines the density of the sampled database. For the 50-level database, $N \simeq 1700$ was chosen. The whole sampled database includes 13766 profiles.

3 Characteristics of the 50-level sampled database

3.1 Variables

Each situation in the 50-level sampled database, hereafter referred to as 50L-SD, is indexed by its space-time location:

- the longitude, between 0° and 360°, eastward counted
- the latitude, between -90° and 90°
- the date, as *yyyymmddhh*, where *yyyy* is the year, *mm* the month, *dd* the day, and *hh* the synoptic hour

As said before, the sampled variables are:

- the atmospheric temperature, in K, on the ECMWF 50-level grid
- the atmospheric specific humidity, in kg/kg, on the ECMWF 50-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 in table 1.

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

- the surface pressure (hPa)
- the surface temperature (K)
- the 2-meter temperature (K)
- the 2-meter specific humidity (kg/kg)
- the land/see mask, where 1 corresponds to land and 0 corresponds to sea

In addition, an ozone profile has been added from the Fortuin and Langematz (1994) climatology dependent on season and latitude:

- the specific ozone, in kg/kg, on the ECMWF 50-level grid

The sampling is performed on the ECMWF model vertical layers and not on fixed pressure layers. As a consequence, the sampled database gathers profiles corresponding to various ocean conditions as well as to land conditions, including high elevated grouds. The lowest surface pressure in the database is 508 hPa and the highest 1049 hPa. This is one the main differences between the 50L-SD and the TIGR-3 database from LMD, that used the same sampling method. The other one is the origin of the profiles.

3.2 Characteristics of the profiles

The histograms of the 50L-SD are presented on figures 1 to 3 for each geotype (sea and land) as a function of the following variables: the total water vapour content, the skin temperature, the date (month and local time), the location (longitude and latitude), the surface pressure, the temperature and the specific humidity in model layer 44. Layer 44 corresponds to a pressure level of 800 hPa when the surface pressure is 1000 hPa (see table 1) and has been chosen as an example of the layer histograms.

An ideal sampling would lead to a regular distribution of the variable values, but is impossible because of the constraints imposed by the physical laws. From the various histogram shapes it is clear that the 50L-SD results from compromises. If the distribution of the situations as a function of month, local time and longitude is regular for each geotype, the other histograms are more irregular due to physical constraints. As an example, the wing in the temperature (respectively specific humidity) in layer 44 histograms[1] between 240 and 280 K (respectively 0.002 and 0.01 kg/m^2) illustrates the weak variability of specific humidity (respectively temperature) in this temperature (respectively specific humidity) range in this layer and in the initial set S. The difference between the specific humidity histogram for sea and that one for land also stems

[1] see figures 3c and 3d for the temperature and figures 3e and 3f for the specific humidity

from the different occurence of each type of situation, even if the natural distribution has been strongly smoothed. Since the representation of high water vapour contents has been forced in the sampling (see section 2.2), the wing in the water vapour histogram is more regular than that of the temperature histogram.

3.3 Extreme values

The 50L-SD extrema have been compared to those of two observation databases. The first database is the 43-profile set that is used in the RTTOV fast radiative transfer model (RTTOV-5: Eyre *et al.*, 1993; Saunders *et al.*, 1999). It is a modified subset of the TIGR-2 database (Achard, 1991; Escobar-Munoz, 1993). The second one is the TIGR-3 database (Chevallier *et al.*, 1998b). Quality controls have ensured that the highest measured temperature in the radiosonde data that the RTTOV set and TIGR-3 use, reaches 30 hPa at minimum (Escobar-Munoz, 1993). Above 30 hPa, the temperature profiles have been extrapolated above the highest measured pressure level using a statistical procedure (Moulinier, 1983). Also the specific humidity profiles reach at least 300 hPa in each set, before any extrapolation.

The extreme values of the atmospheric temperature and of the specific humidity are shown on figure 4 for each set.

The temperature extrema (figures 4a and 4b) of the RTTOV set and those of TIGR-3 are very similar. Compared to them, the 50L-SD has colder minima in the troposphere, similar maxima below 500 hPa and colder maxima above 500 hPa, in particular in the stratosphere. The colder tropospheric minima in the 50L-SD are due to the presence of profiles from the Antarctic plateau, that is sampled neither in the RTTOV set nor in TIGR-3. The colder maxima in the low stratosphere tend to show that the 50L-SD does not sample the hottest stratospheric profiles. As said before, in the middle and in the high stratosphere, both the RTTOV set and TIGR-3 may suffer from extrapolation artefacts. It should be noted that the 50L-SD corresponds to a version of the ECMWF forecast system that did not assimilate the Advanced Microwave Sounder Unit-A (AMSU-A) data. While this may be a weakness, the operational monitoring of the ECMWF system suggests that, at least below 30 hPa, the biases of the system are within 1.5 K, with or without AMSU-A.

The specific humidity maxima (figures 4c and 4d) are very different from one dataset to the next. The RTTOV set has the driest values in the troposphere. This is due to the sampling method used in TIGR-2, that took only temperature into account in the choice of the profiles. This has been improved in the TIGR-3 dataset, as show on the figure. The 50L-SD specific humidity maxima are rather similar to those of TIGR-3 between 200 hPa and 750 hPa. They lie in between the two other datasets below 750 hPa, and are dryer above 200 hPa. The fact that the 50L-SD maxima are mostly dryer than those of TIGR-3 may indicate that the ECMWF forecast system is not able to provide the extremely humid situations. In the case of the middle- and upper-tropospheric humidity, it should be noted that a misinterpretation of the radiosonde humidity data, which has existed for some time, has been found in the ECMWF assimilation system (Andersson and Viterbo, personal communication, 1999). This problem results in an artificial reduction of about 10% of the analysed relative humidities between 400 and 100 hPa, in the regions covered by the radiosondes. However, the 50L-SD only contains 6-hour forecast data that are less affected. As an illustration, the histograms for 200 hPa specific humidity of the 50L-SD are compared to those of TIGR-3 on figure 5. The main range

of variability are very similar between the two databases. Only the extreme values significantly differ. Now, as said before, the TIGR-3 water vapour profiles at 200 hPa may be extrapolated values, and therefore less reliable.

4 Strategy for a further reduction of the 50-level database

4.1 Description of the 150 profile database

For some computationally expensive applications, such as radiative "line-by-line" computations, the number of profiles, 13766, may still be too high.

A further reduction of the size of the database could be achieved by re-sampling the initial 1,500,000 profile initial set (the set S of section 2.2) with a lower value for N. However, the high computational burden of this approach makes it not very flexible and impractical. A simpler approach, that makes use of the 50L-SD, is used here. Under the restrictions examined in section 3, the sampled 50-level database is a regular mesh of the 1,500,000 profile initial set. A random sampling of it enlarges the mesh without modifying its distribution. This approach is used to select a reduced set of 150 profiles out of the 13766 profile database. This subset is referred to 50L-SDs in the following. Figures 6 to 8 present the histograms of the 50L-SDs. As expected, the shapes are similar to those for the 50L-SD. The main effect of the random sampling is the reduction of the extrema (see also figure 9).

4.2 The extrema

The reduction of the extrema in the 50L-SDs mainly concerns the application of the database as a validation set, when one wants to know how an algorithm performs on extreme cases. The temperature extrema and the specific humidity maxima in the 50L-SD correspond to 150 situations. 26 of them are selected by random and added to the 50L-SDs. This alternate version of the reduced database will be referred to as 50L-SDse. The effect of this addition on the extrema of the reduced dataset is shown on figure 10. The extension clearly makes the reduced dataset closer to the 50L-SD as far as the extrema are concerned, but also makes it heterogeneous. However, because of the relatively small number of profiles introduced (only about 16% more) the histograms of the 50L-SDse are not significantly modified (not shown).

5 Summary and future developments

Two datasets have been sampled from the 50-level ECMWF model outputs, that may be used for a wide range of applications, depending on the computational expense. The sampling method used allows for a regular distribution of physically consistent atmospheric temperature and water vapour profiles in each set. As illustrated by Chevallier *et al.* (1999b), these databases are suitable for regression applications. They can also serve as independent validation sets for various algorithms.

The sampled databases presented here should not be considered as final ones. They carry both qualities and weaknesses from the ECMWF assimilation-forecast system. Further im-

provements of that system will enable further improvements of the databases. The forthcoming 60-level ECMWF model (with an increased resolution in the boundary layer, changes to water vapour and stratospheric data assimilation) will lead to improved databases. The on-going assimilation of ozone at ECMWF will also enable to replace the current climatology.

Acknowledgments

The 50-level databases presented here are available through the EUMETSAT Satellite Application Facility. This work has been initiated at Laboratoire de Météorologie Dynamique (France). Thanks are due to R. Armante, A. Chédin, F. Chéruy and N. A. Scott. The interaction with M. Matricardi at ECMWF was also fruitful. T. McNally carefully reviewed and commented the original manuscript.

References

Achard, V., 1991: Trois problèmes clés de l'analyse 3D de la structure thermodynamique de l'atmosphère par satellite : mesure du contenu en ozone ; classification des masses d'air ; modélisation hyper rapide du transfert radiatif. PhD thesis, University Paris VI, 168 pp. [Available from LMD, Ecole Polytechnique, 91128 Palaiseau cedex, France].

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., F. Chéruy, Z. X. Li, and Scott, N. A., 1998a: A fast and accurate neural network-based computation of longwave radiative budget application in a GCM. In *Proceedings of the Am. Meteor. Soc. Conference*, Paris, France.

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, 1999a: TIGR-like atmospheric profile databases for accurate radiative flux computation. Accepted in *Quart. J. Roy. Meteor. Soc.*.

Chevallier, F., J.-J Morcrette, F. Chéruy, and N. A. Scott, 1999b: Use of a neural network-based LW radiative transfer model in the ECMWF atmospheric model. Accepted in *Quart. J. Roy. Meteor. Soc.*.

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].

Eyre, J. R., 1991: A fast radiative transfer model for satellite sounding systems. ECMWF Technical Memorandum No. 186 [Available from the librarian at ECMWF].

Fortuin, J. P. F. and Langematz, U., 1994: An update on the global ozone climatology and on concurrent ozone and temperature trends. *Proceedings SPIE*, 2311, 207-216.

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].

Saunders, R., M. Matricardi, and P. Brunel, 1999: An improved fast radiative transfer model for assimilation of satellite radiance observations. *Quart. J. Roy. Meteor. Soc.*, 125:556, 1407-1425.

level	pressure (hPa)	level	pressure (hPa)	level	pressure (hPa)	level	pressure (hPa)
1	0.10	14	11.49	27	155.70	40	626.02
2	0.32	15	14.24	28	180.77	41	671.31
3	0.59	16	17.64	29	208.01	42	716.89
4	0.95	17	21.85	30	237.35	43	762.33
5	1.38	18	27.08	31	268.76	44	807.04
6	1.89	19	33.55	32	302.17	45	850.23
7	2.47	20	41.56	33	337.46	46	890.88
8	3.14	21	51.49	34	374.50	47	927.68
9	3.93	22	63.49	35	413.17	48	958.96
10	4.88	23	77.58	36	453.32	49	982.63
11	6.04	24	93.83	37	494.84	50	996.14
12	7.49	25	112.24	38	537.57		
13	9.27	26	132.86	39	581.36		

Table 1: 50-level vertical grid of the ECMWF model, when the surface pressure equals 1000 hPa. The general formulation depends on the surface pressure.

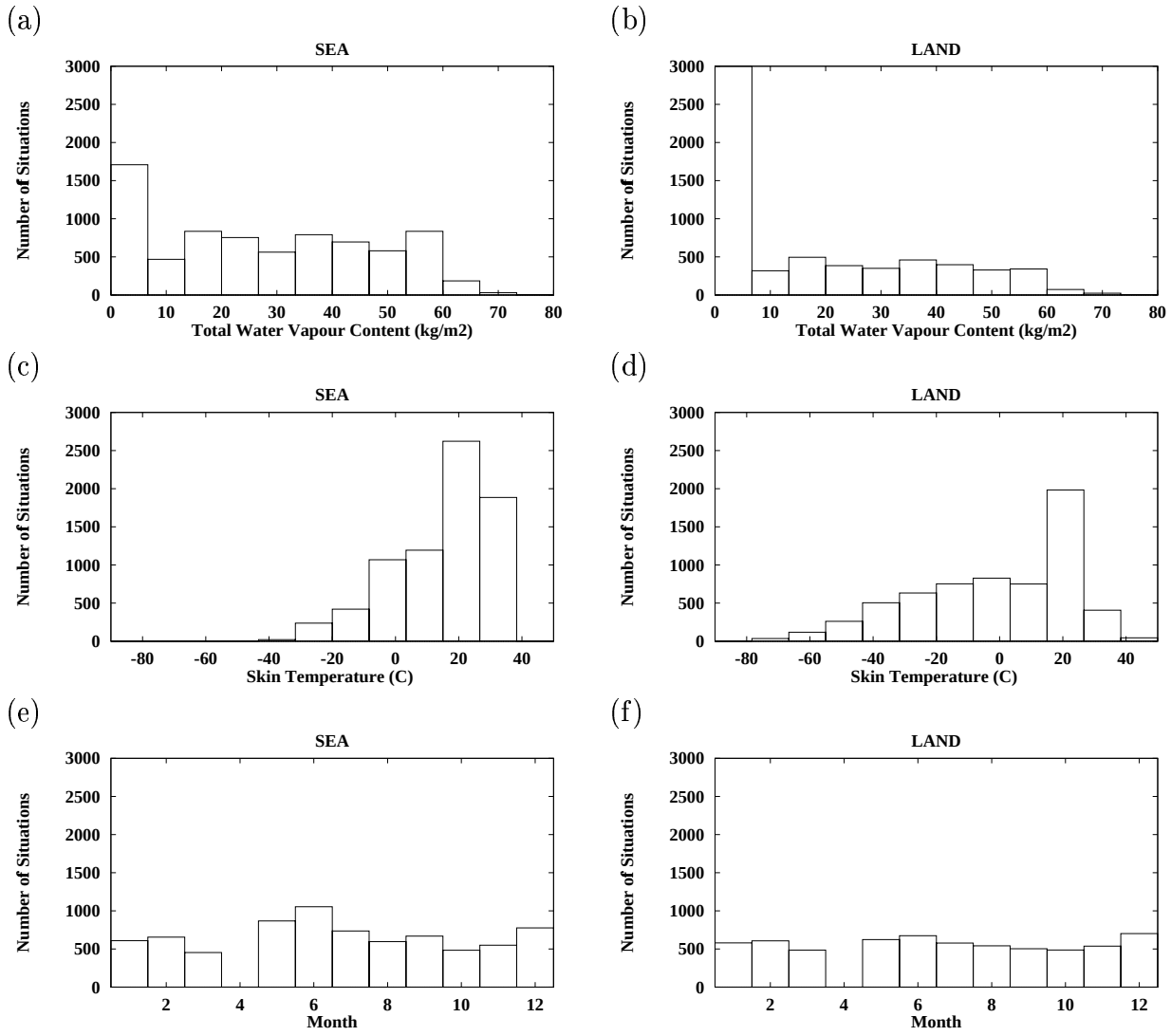

Figure 1: Distribution of the situations in the 50-level sampled database (50L-SD, 13766 situations) as a function of some variables and for each geotype.

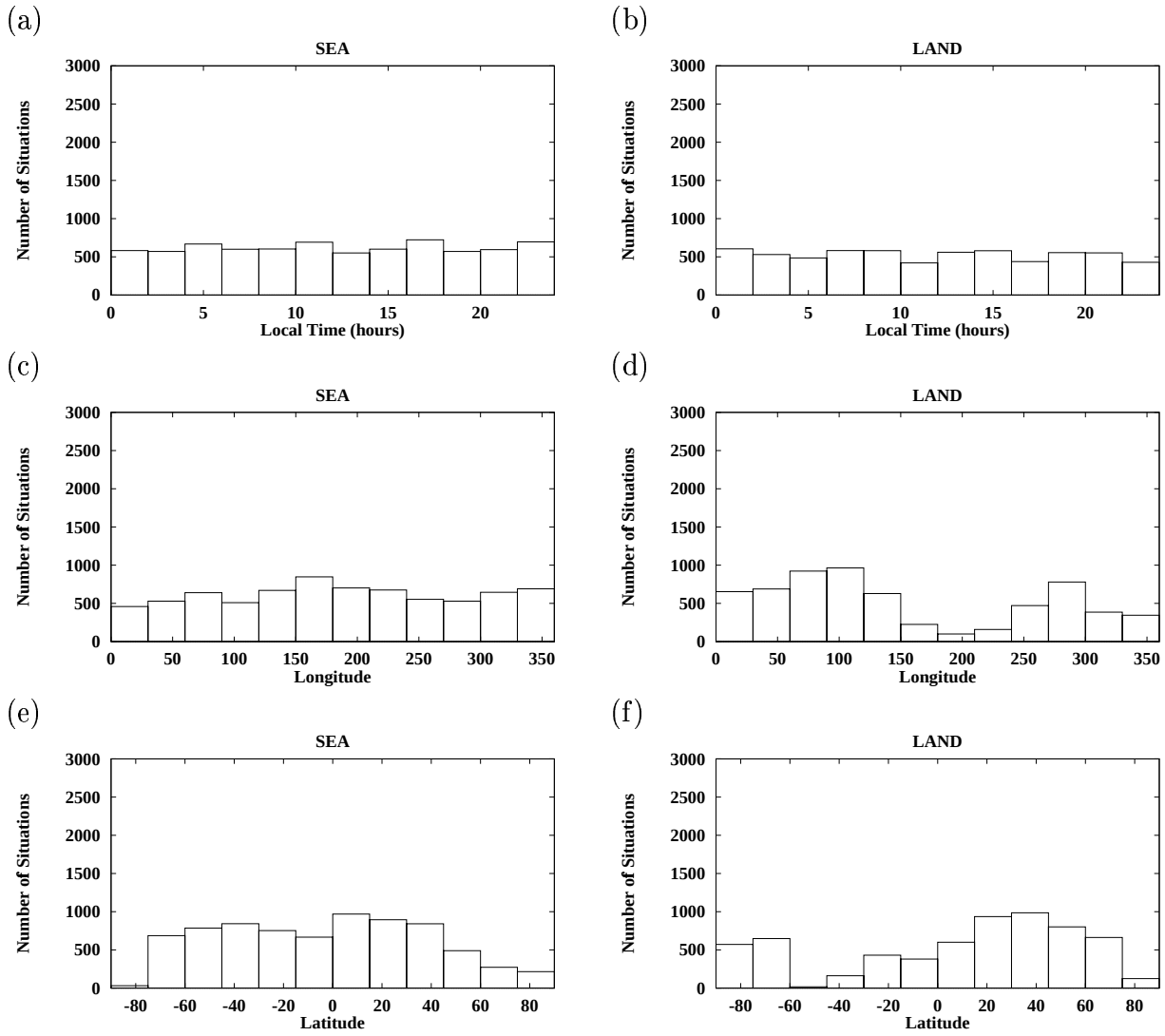

Figure 2: Same as previous.

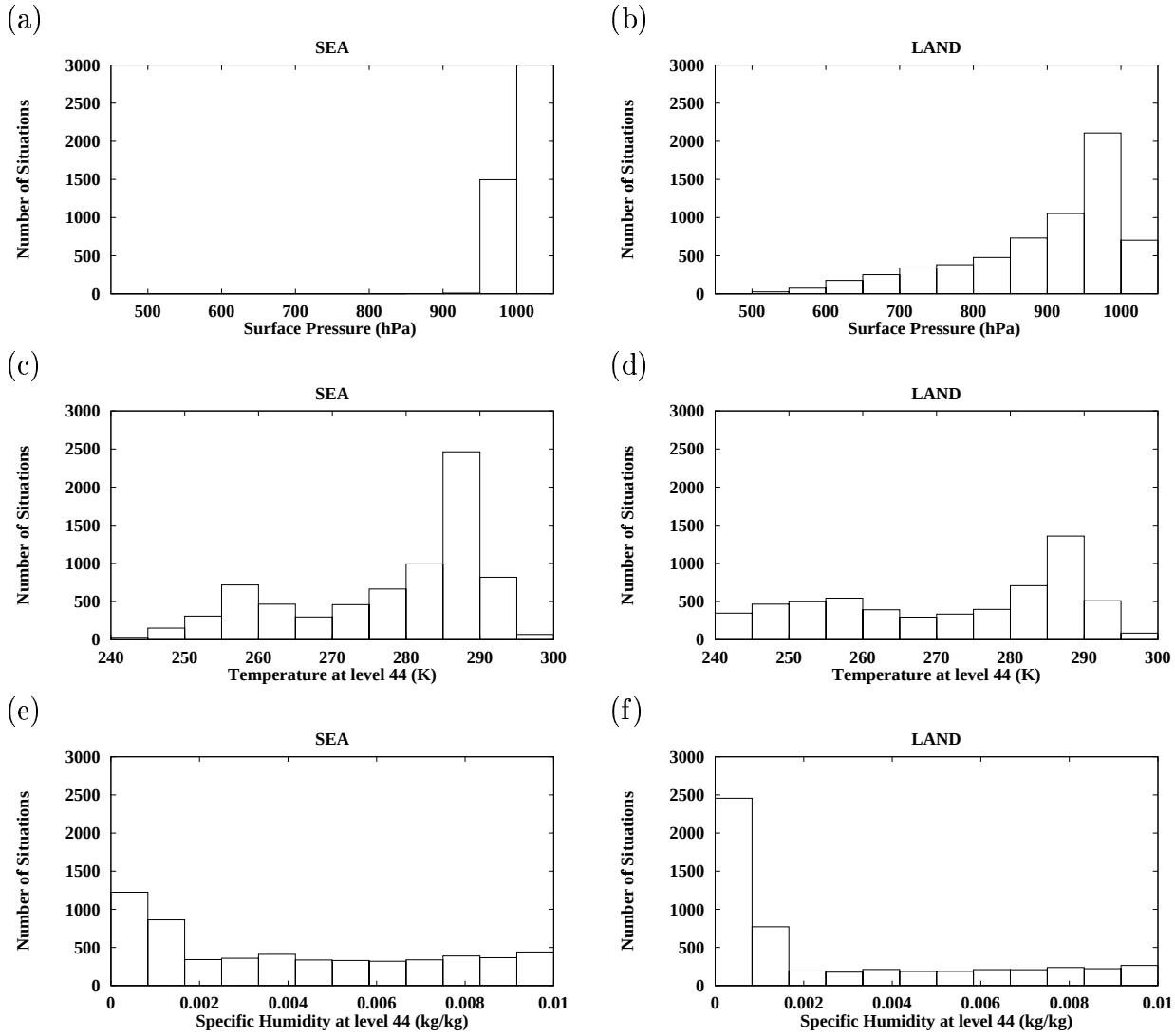

Figure 3: Same as previous. Layer 44 corresponds to a pressure level of 800 hPa when the surface pressure is 1000 hPa (see table 1) and has been chosen as an example of the layer histograms.

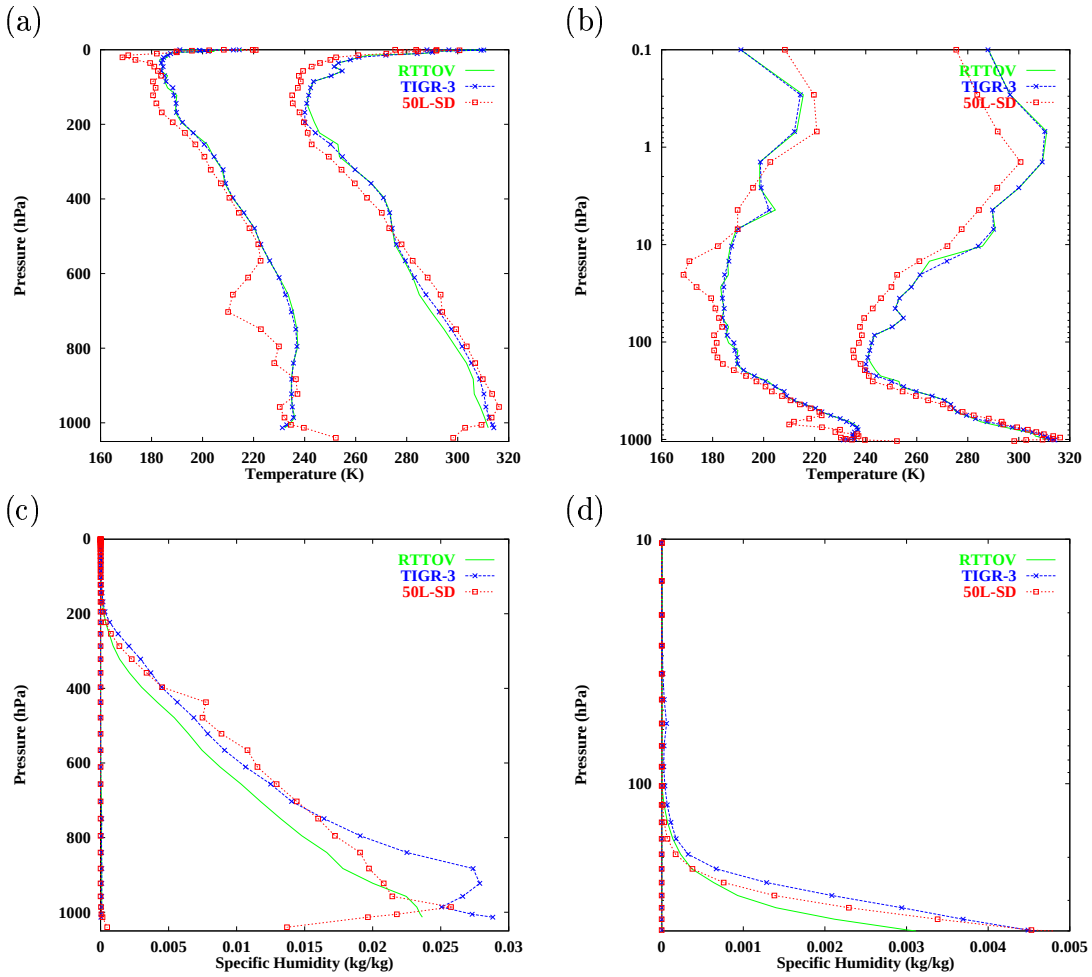

Figure 4: Extreme values of temperature ((a) and (b)) and of specific humidity ((c) and (d)) in the 50-level sampled database (noted 50L-SD), in TIGR-3, and in the 43-profile set used in RTTOV. All datasets are interpolated on a fixed pressure level grid: the 43-level RTTOV grid to which a 1040 hPa level has been added. Only the 50-level sampled dataset has values at 1040 hPa. The pressure values on figure (d) range from 400 hPa to 10 hPa only.

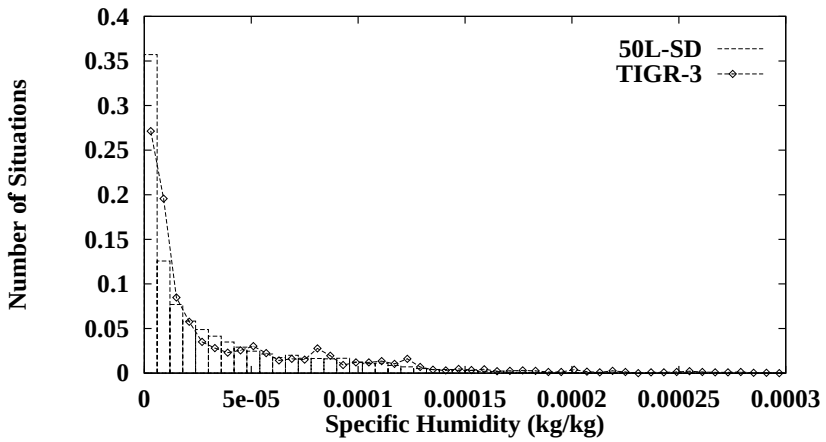

Figure 5: Histograms for specific humidity at 200 hPa of the 50-level sampled dataset and of TIGR-3, normalized by their respective number of profiles.

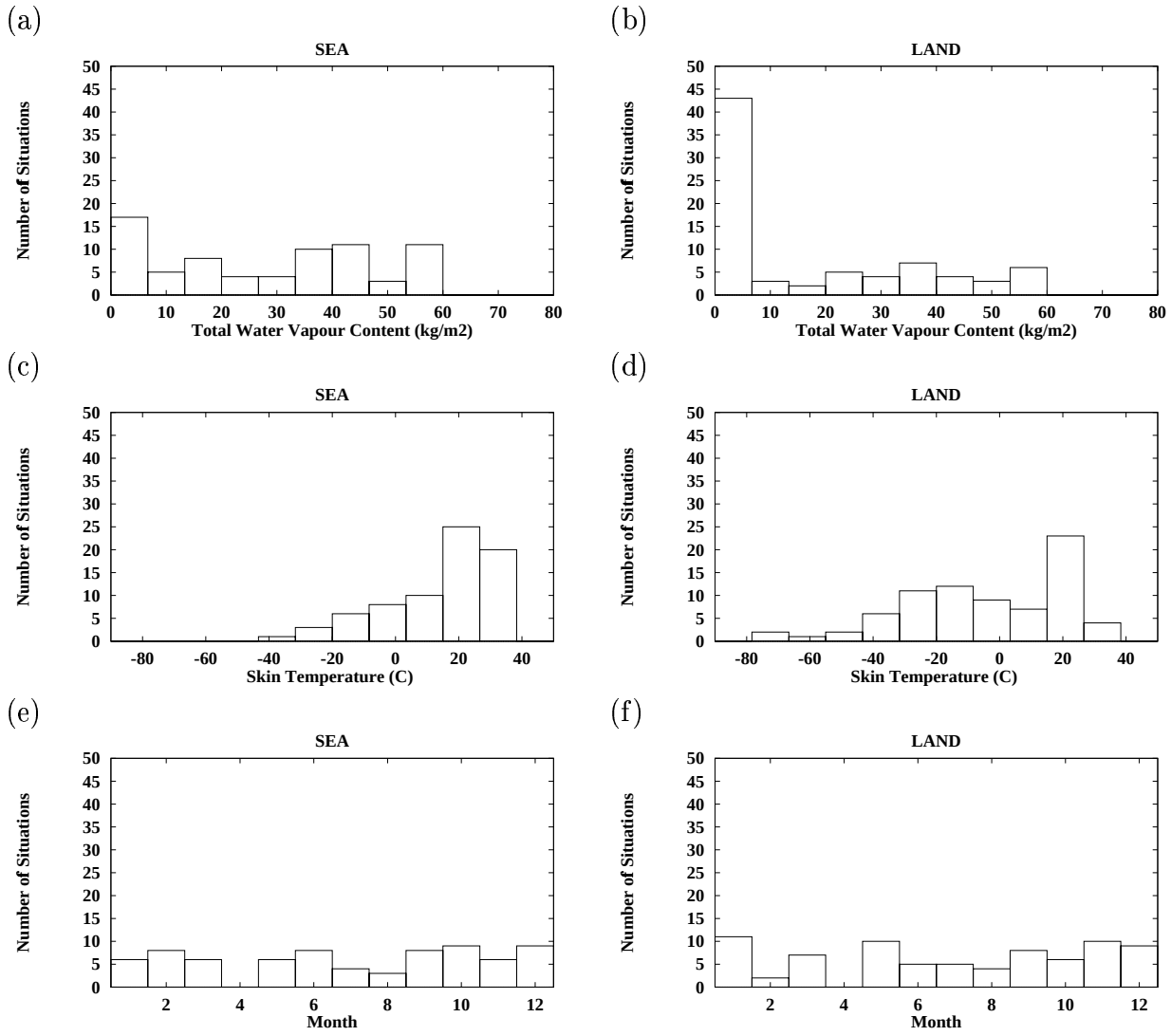

Figure 6: Distribution of the situations in the reduced 50-level sampled database (noted 50L-SDs) as a function of some variables and for each geotype.

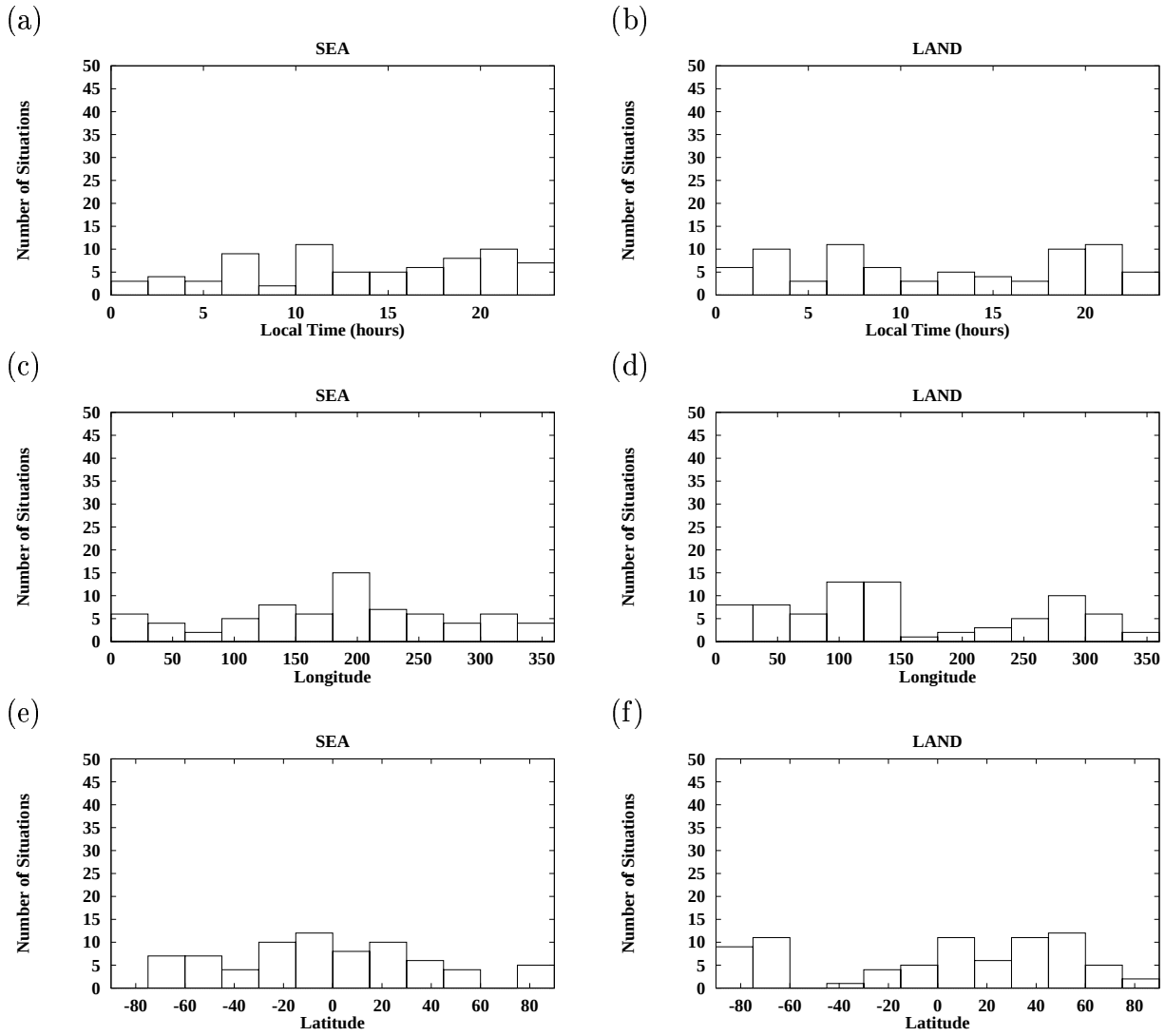

Figure 7: Same as previous.

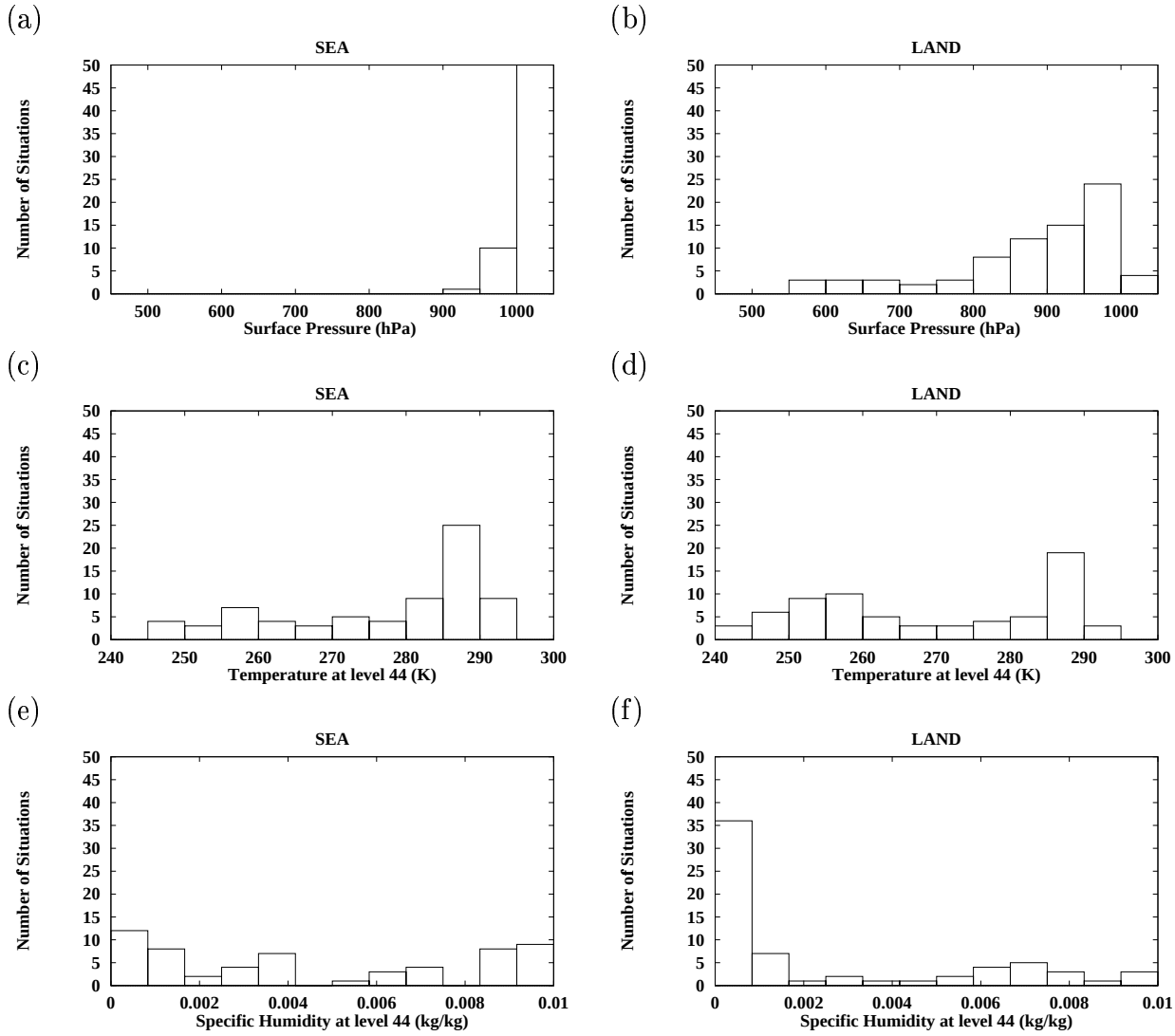

Figure 8: Same as previous. Layer 44 corresponds to a pressure level of 800 hPa when the surface pressure is 1000 hPa (see table 1) and has been chosen as an example of the layer histograms.

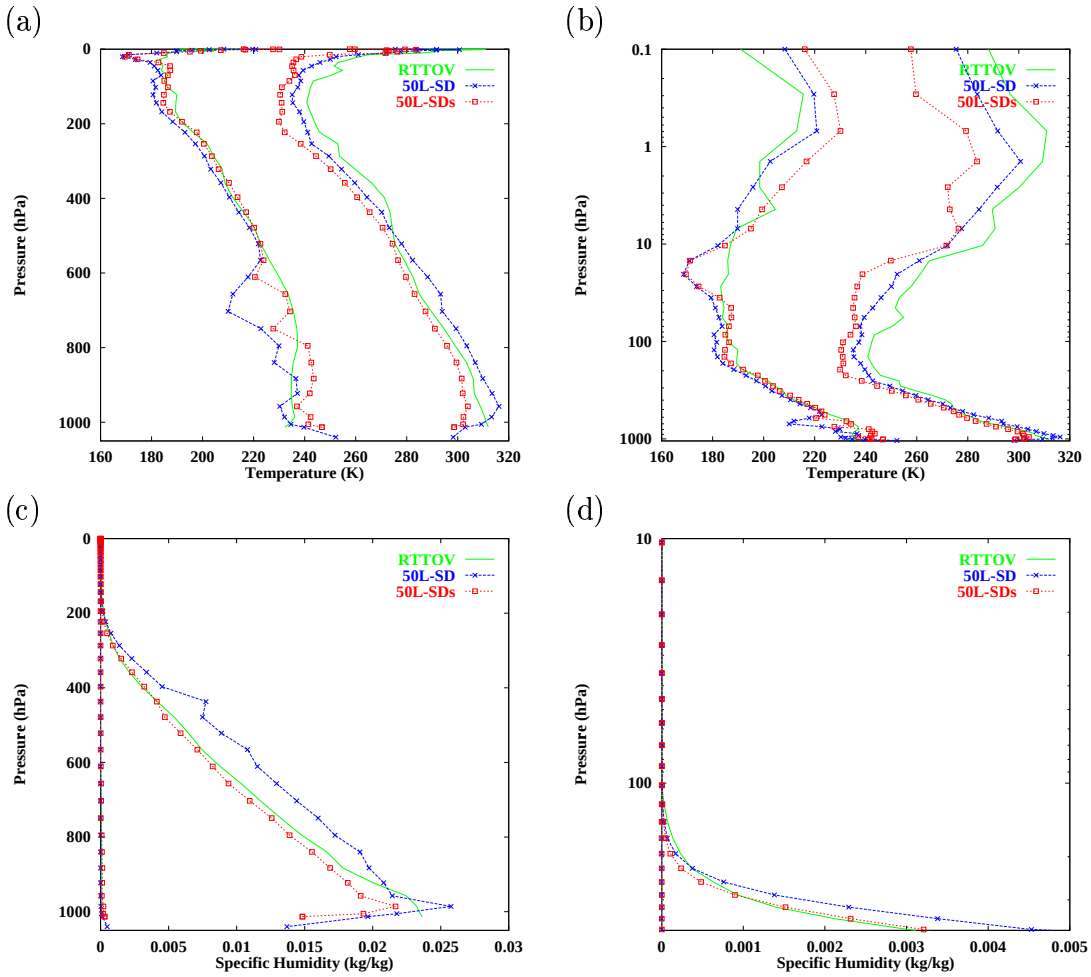

Figure 9: Extreme values of temperature ((a) and (b)) and of specific humidity ((c) and (d)) in the 50-level sampled dataset (noted 50L-SD), in its subset (noted 50L-SDs), and in the 43-profile set used in RTTOV. All datasets are interpolated on a fixed pressure level grid: the 43-level RTTOV grid. The highest atmospheric pressure in the 50L-SDs is 1029 hPa. Therefore the 1040 hPa level of figure 5 has been suppressed. The pressure values on figure (d) range from 400 hPa to 10 hPa only.

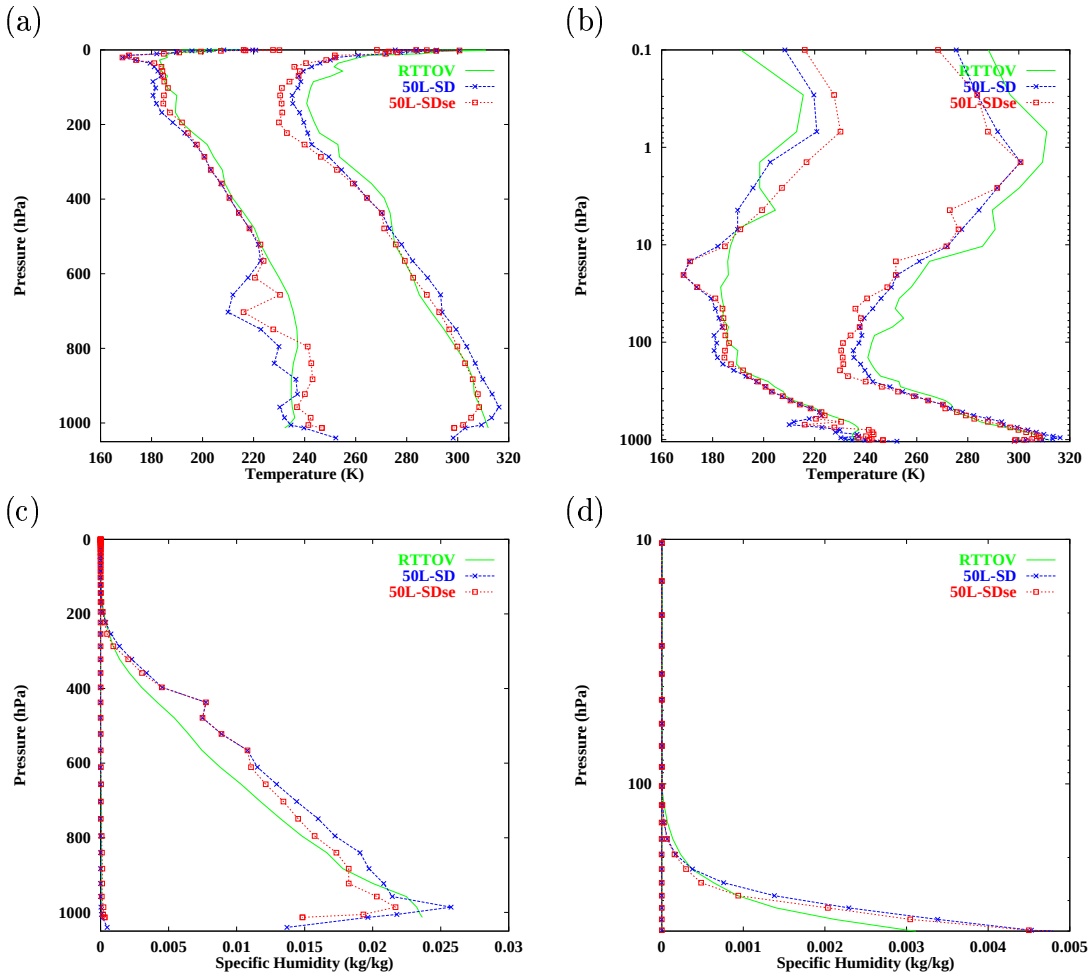

Figure 10: Extreme values of temperature ((a) and (b)) and of specific humidity ((c) and (d)) in the 50-level sampled dataset (noted 50L-SD), in the subset with some extrema (noted 50L-SDse), and in the 43-profile set used in RTTOV. The pressure values on figure (d) range from 400 hPa to 10 hPa only.