[redacted]

Arbeitsberichte der Schweizerischen Meteorologischen Zentralanstalt
Rapports de travail de l'Institut Suisse de Météorologie
Rapporti di lavoro dell'Istituto Svizzero di Meteorologia
Working Reports of the Swiss Meteorological Institute

Zürich

52705

No. 139

PRINCIPAL COMPONENT ANALYSIS OF ECMWF ANALYSIS FIELDS TAKEN AS PREDICTORS FOR A REGRESSION MODEL OF PRECIPITATION FORECASTS

Francis Schubiger

October 1986

Multivariate statistics 519.2

Summary

Principal component analysis, a very powerful method to compress data in an "optimum" way, is applied on different ECMWF analysis fields such as geopotentials (direct fields) or geostrophic relative vorticity (derived fields). An overview of the principal component analysis is presented and the commonly used truncation criteria are reviewed. Examples of the winter 1981/1982 are worked out and special attention is payed to a synoptic-oriented truncation criterion. The main principal components of the different fields will be the predictors for a regression model with, as predictands, the precipitations at the stations of the swiss automatic network (ANETZ).

Résumé

L'analyse en composantes principales est une méthode très efficace pour comprimer un grand nombre de données d'une façon "optimale". Cette méthode est appliquée à différents champs d'analyse du CEPMMT, tels que le géopotentiel (champ direct) ou la vorticité relative géostrophique (champ dérivé). L'analyse en composantes principales est revue rapidement et les différents critères de troncation sont exposés. Quelques exemples de l'hiver 1981/1982 sont apportés et une attention spéciale est donnée à un critère de troncation basé sur une approche synoptique. Les composantes principales les plus importantes seront ensuite les prédicteurs pour un modèle de régression avec comme prédictands les précipitations aux stations du résau automatique suisse (ANETZ).

Zusammenfassung

Die Hauptkomponentenanalyse ist eine sehr effiziente Methode für eine "optimale" Datenkomprimierung. Sie wird auf verschiedene EZMW-Analysenfelder angewandt, wie z.Bp. Geopotentiale (als direkte Felder) oder geostrophische relative Vorticity (als abgeleitete Felder). Die Hauptkomponentenanalyse wird kurz erläutert und die verschiedenen Trunkations-Kriterien dargelegt. Beispiele aus dem Winter 1981/1982 werden aufgezeigt. Besondere Aufmerksamkeit wird einem synoptisch-orientierten Trunkations-Kriterium gegeben. Die wichtigsten Hauptkomponenten werden als Prädiktoren für ein Regressionsmodell verwendet werden, das die Niederschlagsdaten an den Stationen des schweizerischen automatischen Netzes (ANETZ) bestimmt.

Riassunto

Vengono presentati un riassunto dell'analisi delle componenti principali e una revisione dei criteri di troncamento più comunemente usati. Sono inoltre esposti alcuni esempi dell'inverno 1981/82. Particolare attenzione viene prestata ad un criterio di troncamento, basato su un approccio sinottico. Le componenti principali più importanti dei diversi campi saranno in seguito usate quali predittori per un modello di regressione che ha come predittando le precipitazioni alle stazioni della rete automatica svizzera (ANETZ).

Table of contents Page

1. Introduction 1

2. Principal component analysis 3

3. The aims of the projects DIAGNO1 and DIAGNO2 7

4. The predictor fields for DIAGNO2 12

5. Results of the PCA for the winter 1981/1982 14

5.1 Interpretation of the results 16

5.2 Truncation criterion 26

References 30

1. Introduction

The forecast fields of the numerical weather models build today the basis for the weather forecast (the short-range and especially the medium-range). But these forecast fields cannot be directly applied as local forecasts. High-resolution forecasts in space and time with large-scale models, such as that of the European Centre for Medium Range Weather Forecasts (ECMWF) 1) are only possible with supplementary computations.

There are two main ways to do this (Fig. 1).

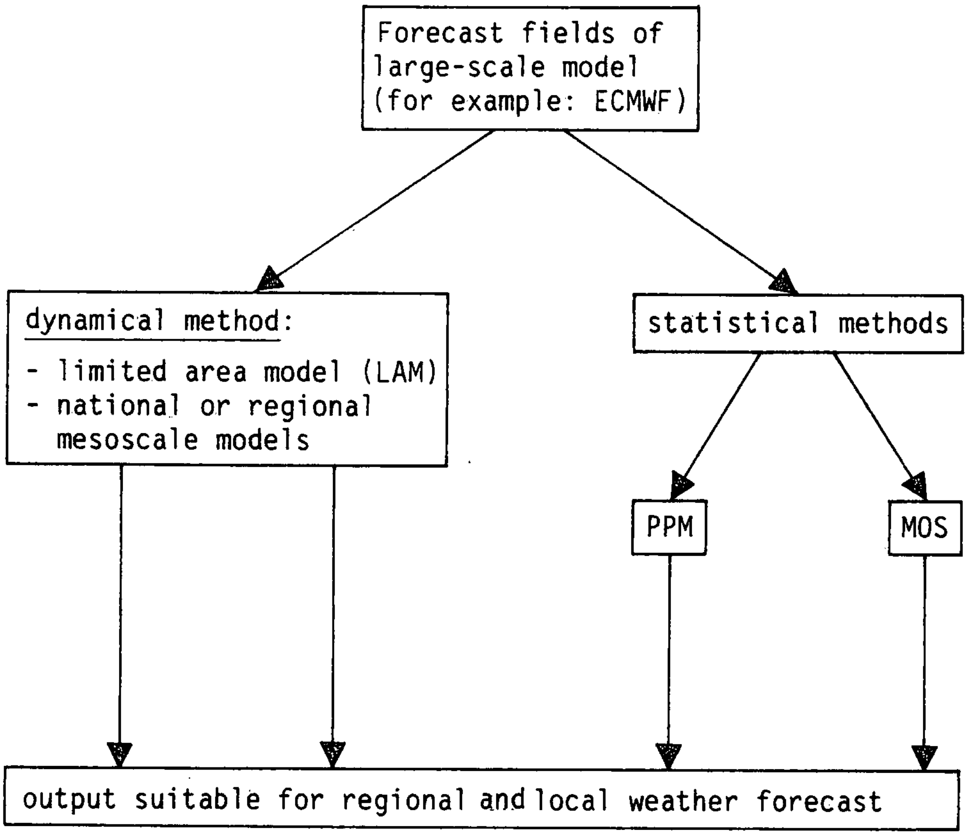

Forecast fields of large-scale model (for example: ECMWF)
dynamical method:
- limited area model (LAM)
- national or regional mesoscale models
statistical methods
PPM
MOS
output suitable for regional and local weather forecast

Fig. 1 Applications of large-scale numerical models for local weather forecast

1) spectral model in the horizontal with triangular truncation at wavenumber 106 and physical processes in a gaussian grid with a resolution of 1.125°, i.e. resolution of horizontal motions with a half wavelength of approximately 190 km.

The first is to run a mesoscale model (grid mesh of less than 50 km) that is driven at the boundaries by a large-scale model. Such a model is in development in Switzerland (grid mesh of less than 15 km).

The second way is a dynamic-statistical method. The forecast fields of numerical models give the input data, i.e. are the predictors, for a statistical interpretation model. The results, i.e. the predictands, are local weather elements, as for example 6- or 12-hour accumulated rainfall, minimal or maximal temperature, thunderstorms and so on.
To build up such a statistical method there are two possibilities. The first one is to use analysis fields as predictors for the development, but then, of course to use forecast fields to apply the method. This is the so-called Perfect Prognosis Method (PPM), where we assume that the forecast fields are always correct and do not have a systematic bias.
The second possibility is to use forecast fields as predictors and then to apply the method with forecast fields of the same range. This is the so-called Model Output Statistics (MOS) method. It has the advantage that systematic errors of the numerical model are filtered out. But each time the numerical model will be changed, the statistical model should be recomputed. Therefore it is difficult to have a sample large enough between two (major) changes in the numerical model.

At the Swiss Meteorological Institute (SMI) there is a project called DIAGNO2. It is a statistical interpretation model for the rainfall forecast based on the Perfect Prognosis Method (PPM). The predictands are the measured rainfall data at the 60 stations of the swiss automatic observing network (ANETZ). The predictors are on one side 3 fields of vertical velocities at 850, 700 and 500 hPa computed by the omega-equation at the SMI (mesoscale predictors) and on the other side 27 direct or derived fields from the ECMWF (large-scale predictors). The method is based on the PPM (development with analysis fields, application with forecast fields) because the ECMWF model is still updated quite often (higher resolution, new analysis scheme or new parameterization of convection,...).

The data of the 27 direct and derived fields from the ECMWF in a grid of 72 points chosen for this project give a very large number of predictors. So, in a first step, this large number of predictors has to be compressed in an "optimum" way, i.e. with a minimum loss of information. The method chosen to this end is the principal component analysis (PCA). Each predictor field is compressed with a PCA. The working reports of the SMI of Quiby (1984) and Ambühl (1984) treat the more theoretical part of the project.

This working report deals with the choice of the different predictors and with some impressive results of the PCA for the winter 1981/1982. Another important problem treated concerns the question of how many principal components must be retained for an optimal separation between significant information and "noise". It will be shown that for our project retaining the principal components responsible for a total variance of 99% will fulfil this condition. The next step of the project DIAGNO2, that will be treated in further working reports, is a forward selective multiple linear regression between the principal components responsible for a total variance of 99% as predictors, and the 6-hourly rainfall data for each station of the automatic network (ANETZ) as predictands.

2. Principal component analysis

The principal component analysis (PCA) is a very efficient method of the multivariate statistics for the compression of the amount of data. Since its introduction several decades ago, the PCA, also known as empirical orthogonal functions (EOF), even though they are not identical (Richman, 1981), has recently become a popular tool in atmospheric sciences (e.g.: Craddock and Flood, 1969; Rinne and Karhila, 1979; LeDrew, 1980; Horel, 1981; Cohen, 1983). PCA of rainfall data has been applied by Molteni et.al. (1983), Melice and Wendler (1984) and Goossens (1985). Use of EOF for a goal similar to ours (EOF followed by a regression analysis) can be found in White et.al. (1958), Paegle and Haslam (1982) and a very similar project is described in Mori (1984). A good overview of properties and applications of the PCA is given in ECMWF (1977) and Sneyers and Goossens (1985).

The principal component analysis (PCA) amounts to a rotation of the coordinate axes in n-dimensional variable space to a new reference frame. The first principal component extracted accounts for the maximum possible variance in the data set. Each succeeding principal component accounts for the maximum remaining variance. Although the number of principal components equals the number of original variables, in general only the first few principal components are required to explain a large fraction of the variance. In other words PCA is used to filter the data by separating the factors which account for the signal from those which account for the noise.

PCA is not only a method of compressing data sets in an optimum way as it is mostly used in atmospheric sciences. It is also possible to identify some of the individual principal axes and, at least in a tentative way, associate them with a particular physical process (Savijärvi, 1978; North, 1984) as will be seen later in this report. But special care is of rigour for small sample sizes, for which Storch and Hannoschök (1985) recommend renouncing a physical interpretation.

PCA can also be used for data classification (Christenson and Bryson, 1966) or to find out extreme or erroneous data (Flury and Riedwyl, 1983).

Let us see the computations of the principal components (PCs) for our project.

Let f(x,t) be a value of the field f at location x and time t, where x, t are integers in the range $1 \leqslant x \leqslant P$, $1 \leqslant t \leqslant N$. So, the sample of data consists of N analyses of P grid-point-values of the field f. (We have M different fields, and for each field we compute a PCA).

We compute first the standardized field F:

$$(1)\quad F(x,t) = \frac{f(x,t) - \bar{f}(x)}{s(f(x))} \quad \text{, where: } \bar{f}(x) = \frac{1}{N} \sum_t f(x,t)$$

$$s(f(x)) = \sqrt{\frac{1}{N-1} \left(\Sigma_t f(x,t) - N \bar{f}(x) \right)}$$

Why to standardize the data ? For a PCA of only one field as here, i.e. with data of the same unit, it is not necessary to standardize the data. On the contrary, the physical interpretation of the PCs is easier with non-standardized fields. But for a PCA of data of different units a standardization is necessary. For our project we left it open to do a second PCA between the first PCs of each field. For that reason we standardized the data.

One forms the P x P - symmetric variance-covariance matrix:

$$(2)\quad VCV(x,x') = \frac{1}{N-1} \sum_t F(x,t) \cdot F(x',t)$$

and we compute its P eigenvectors $E_i = (E_i(1),\dots, E_i(P))$ and P non-negative eigenvalues λ_i, $1 \leq i \leq P$.

The total variance is: $\mathrm{Tr}(VCV) = \sum_i \lambda_i$

The eigenvalues λ_i and their corresponding eigenvectors E_i are ordered, such that: $\lambda_1 \geqslant \lambda_2 \geqslant \dots \geqslant \lambda_P$. Now the first eigenvalue, with the highest variance, corresponds to the eigenvector with the highest amount of information and to the first, main, principal component, as will be seen next.

We compute the principal components for each time t:

$$(3)\quad Z(i,t) = \sum_x E_i(x) \cdot F(x,t)$$

We can reconstitute the standardized and initial field:

$$(4)\quad F(x,t) = \sum_i Z(i,t) \cdot E_i(x)$$

$$(5)\quad f(x,t) = \left(F(x,t) + \overline{f}(x)\right) \cdot s(f(x))$$

Useful properties of PCA are:

$$\sum_x E_i(x) \cdot E_j(x) = \delta_{ij}\ , \text{ for all } i,j.$$

$$\frac{1}{N} \sum_i Z(i,t) = 0$$

$$\frac{1}{N-1} \sum_t Z(i,t) \cdot Z(j,t) = \lambda_i \delta_{ij}\ , \text{ for all } i,j.$$

$$\frac{1}{N-1} \sum_t \sum_x F^2(x,t) = \sum_i \lambda_i$$

The demonstration of these properties is given in Quiby (1984).

Our next concern is the problem of the truncation criterion after a PCA. The following question must be answered: How many PCs must be retained to keep the significant meteorological information ? Several different truncation criteria have been defined in the past to separate the eigenvalues due to random noise from those that may contain an interesting geophysical signal. For a recent review of the various criteria available, see the paper of Preisendorfer et.al. (1981).

We will briefly review three main kinds of truncation criteria:

1) The first criteria defined in the past recommended to retain only those eigenvalues whose component variance are greater than a fixed limit, normally 1%. These criteria are known as, for example, Guttmann lower bound criterion or are referenced to Kaiser or also others. They are still widely used (Rinne and Järvenoja, 1979; Horel, 1981).

2) Another criterion results from the observation that the eigenvalues corresponding to the noise components are approximately in a geometric progression. Craddock and Flood (1969) developped the LEV-graph, that is, a graph in which the eigenvalues are plotted in logarithmic scale against their ordinal numbers. In fact the eigenvalues corresponding to the noise components give on the graph a straight line. So, one can consider significant the PCs whose eigenvalues are above this straight line. Examples of using this criterion can be found in Rinne and Järvenoja (1979), Molteni et.al. (1983) and Cohen (1983). A similar criterion, the "scree test", was developped earlier, in 1966, by Cattell.

3) More recently, Preisendorfer and Barnett (1977) suggest a method based on a Monte Carlo technique. For a description of that procedure, also called rule N, see Overland and Preisendorfer (1982). The authors of recent articles more and more apply this method as truncation criterion (Ashbaugh et.al. 1984; Goossens, 1985).

For our project we will follow a synoptic approach to define the point of truncation. We reconstitute the initial field with a variable number q of PCs:

$$(6)\ f(x,t)=\left(\sum_{i=1}^{q} z(i,t)\cdot E_i(x)+\bar{f}(x)\right)\cdot s(f(x))$$, where $1 \leq q \leq I$

The fraction of total variance explained by these PCs is: $\dfrac{\sum_{i=1}^{q}\lambda_i}{Tr(VCV)}$

We compare subjectively, in a synoptic way, the reconstituted fields with the original field. We will retain as much PCs as necessary to get the same synoptic pattern for both fields. For our set of data we will find that 99% of total variance depicts the real atmosphere (see Section 5). Thus the truncation criteria will be reached when the PCs included in the expansion account for 99% of the total variance, i.e.:

$$\frac{\sum_{i=1}^{q}\lambda_i}{Tr(VCV)} = 99\%$$

A rotation of the PCs is often used for a better representation of the input map types. For exploratory analyses, or for the best reproduction of the "correct" input map, a rotation is very useful (Richman, 1981; Horel, 1981). Richman (1986) gives a good review article of that topic and explains some shortcomings of unrotated PCs, such as domain shape dependence, subdomain instability or sampling errors. As for our project the only concern for the use of the PCA is the reduction of the amount of data, we will not do any rotation of our PCs. So we will use the principal component analysis in the way as it was created and used initially, i.e. as a technique to reduce dimensionality.

3. The aims of the projects DIAGNO1 and DIAGNO2

In the mid-seventies a project for forecasting precipitations by dynamical-statistical methods (based on the omega-equation) has been developed at the SMI (Kuhn and Quiby, 1976). The model relies on large-scale grid point data and computes then on a fine grid mesh the field of vertical velocities at 850, 700 and 500 hPa. A regression technique, developed for 20 cases with the model based on a grid from the Deutscher Wetterdienst (DWD), gave satisfactory results for the precipitations.

The project DIAGNO1

In early eighties the project DIAGNO1 had the aim to make the above model compatible to run operationally with the ECMWF analysis and forecast fields on a polar-stereographic grid with a resolution of 300 km at 60°N. The results of the vertical velocities were obtained on a grid of 24 x 25 points with a resolution of 1/10 of the initial ECMWF fields, i.e. about 28 km (Fig. 2). An orography of that grid had to be developed. Unfortunately the US-Navy Orography with a resolution of 10' x 10' (available from ECMWF) has some big errors in mountaineous regions of our country, especially in the Valais (Fig. 3), so an orography for our grid was developed at the SMI by Piaget (Fig. 4). The regression method for the precipitations, developed on the DWD grid, no longer gave useful results, so only the vertical velocities are computed daily. But for the forecasters these diagnostically derived fields of vertical velocities are only of limited usefulness. The main problem is still to interpret these fields and to do a qualitative-quantitative estimate of the amount of precipitation that should occur.

The project DIAGNO2

This new project has for aim to determine the precipitations at the locations of the swiss automatic network (ANETZ) by more elaborated statistical methods (principal component analysis followed by a forward selective multiple linear regression) and with more predictor fields than only the vertical velocity fields of DIAGNO1 (see second part of Section 1, p. 2-3, for more details).

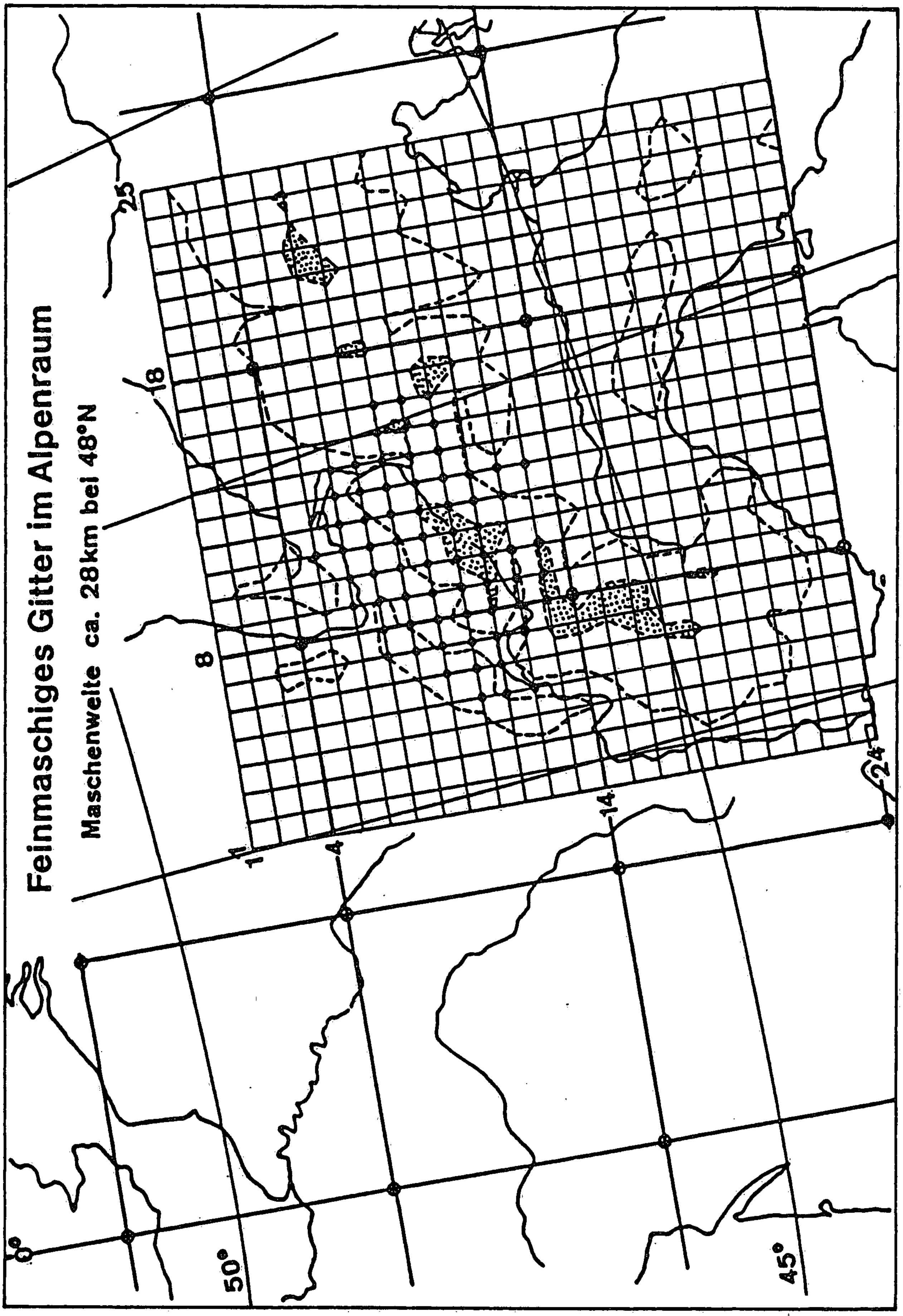

Feinmaschiges Gitter im Alpenraum
Maschenweite ca. 28km bei 48°N
25
18
8
1
4
14
24
50°
45°
6°

Fig. 2 Representation of the two grids of DIAGNO 1: the coarse one has a resolution of about 280 km, the nested one is ten times finer

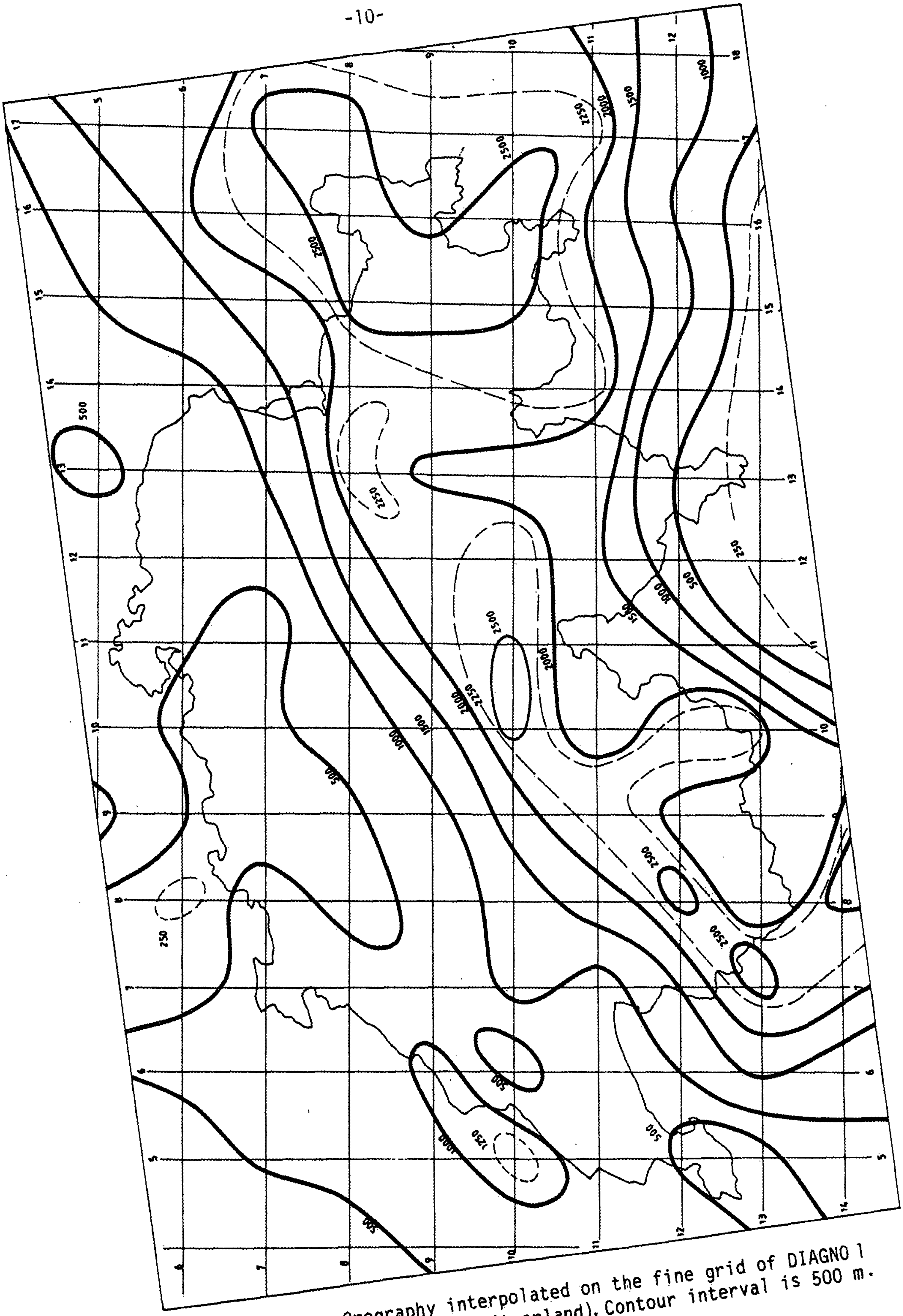


Fig. 3 US-Navy Orography interpolated on the fine grid of DIAGNO 1 (extract of area of Switzerland). Contour interval is 500 m.

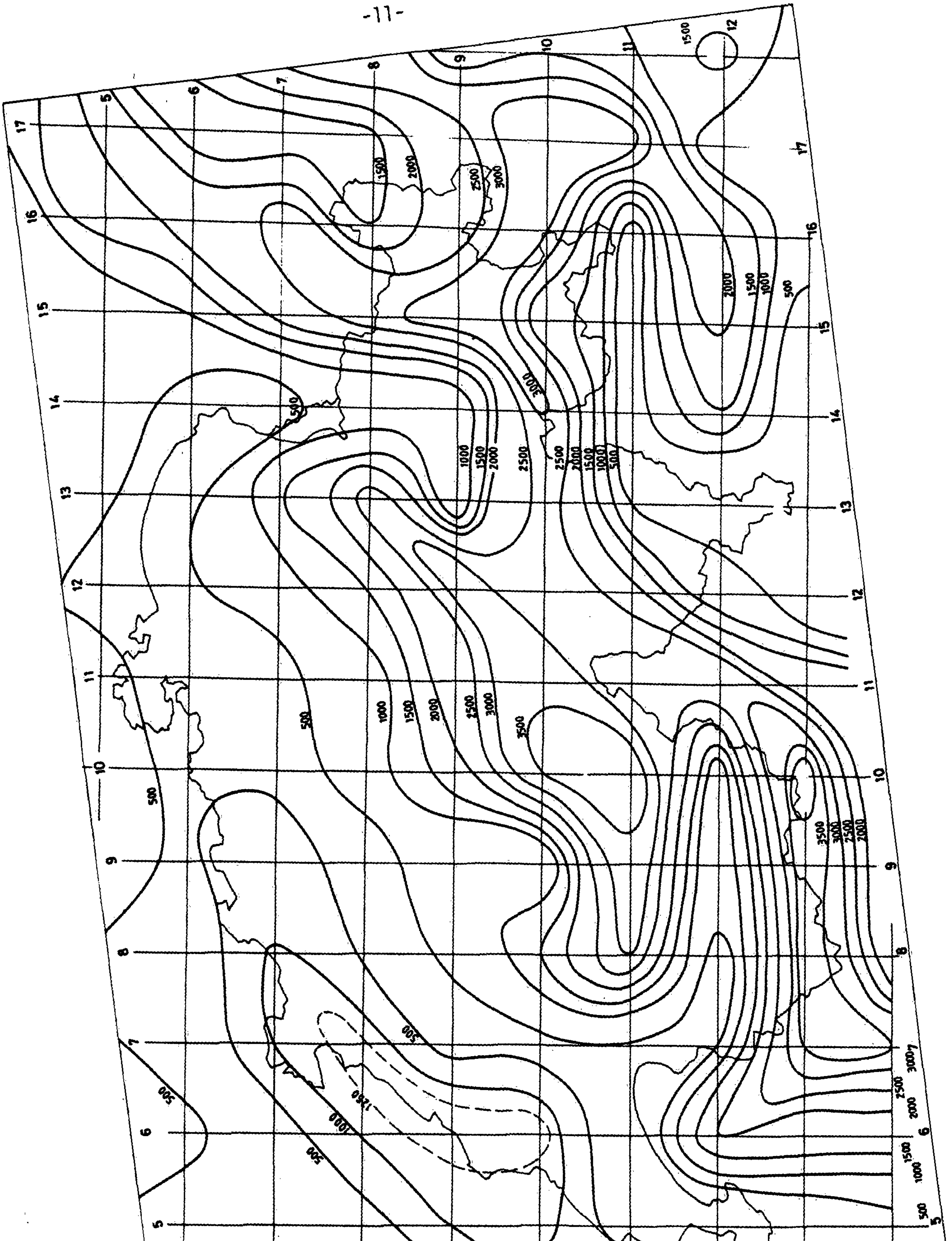


Fig. 4 Orography for the fine grid of DIAGNO 1 developped at the SMI by piaget (extract of the area of Switzerland). Contour interval is 500 m.

4. The predictor fields for DIAGNO2

The predictor fields are on one side mesoscale predictors, the three vertical velocities of DIAGNO1 (see Section 3), and on the other side large-scale predictors, 27 direct or derived fields from ECMWF in a grid 42N-52.5N and 1.5E-13.5E with a mesh of 1.5° in latitude and longitude, i.e. 72 grid-points (Fig. 5).

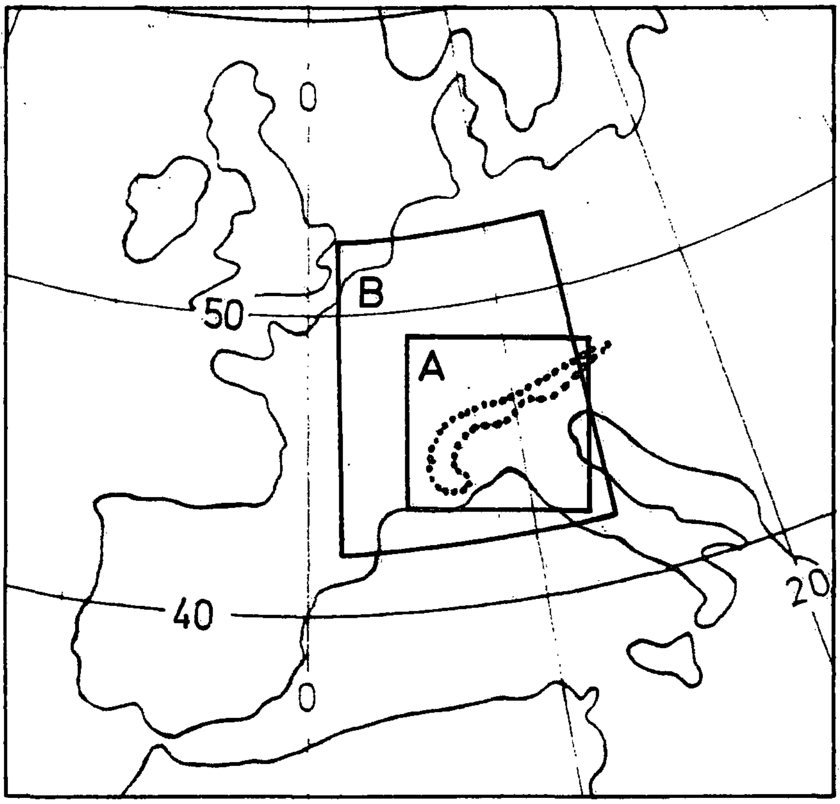


Fig. 5 The grid size of the 3 mesoscale predictors (A) and of the 27 large-scale predictors (B). Area B is used for the figures of Section 5.

The fields from ECMWF are chosen in a way to account for all physical processes responsible for precipitations. There are the following fields:

1)- 4): the geopotentials z at 850, 700, 500 and 300 hPa (direct fields)

5)-12): the geostrophic wind components u_g and v_g at 850, 700, 500 and 300 hPa (computed from the geopotentials)

13)-16): the relative vorticity ζ_g at 850, 700, 500 and 300 hPa (computed from the geopotentials)

17)-20): the vertical velocity ω at 850, 700, 500 and 300 hPa (direct fields)

21)-23): the thickness Δz of the layers 850-700, 700-500 and 500-300 hPa (direct from the geopotentials)

24)-26): the static stability σ of the layers 850-700, 700-500 and 500-300 hPa (computed from the temperature and geopotentials)

27): the water vapor content Q between 850 and 300 hPa (computed from the temperature, relative humidity and geopotentials).

The formula for the computations of the derived fields are the following:

geostrophic wind u_g, v_g and relative vorticity ζ_g:

$$u_g = -\frac{g}{f}\left(\frac{\partial z}{\partial y}\right)_p \quad ; \quad v_g = \frac{g}{f}\left(\frac{\partial z}{\partial x}\right)_p$$

$$\zeta_g = \frac{g}{f}\nabla^2 z + \frac{\beta}{f} u_g$$

in spherical coordinates:

$$\frac{\partial z}{\partial x} = \frac{1}{r\cos\varphi}\,\frac{\partial z}{\partial \lambda}$$

$$\frac{\partial z}{\partial y} = \frac{1}{r}\,\frac{\partial z}{\partial \varphi}$$

$$\nabla^2 z = \frac{1}{r^2\cos^2\varphi}\,\frac{\partial^2 z}{\partial \lambda^2} + \frac{1}{r^2}\,\frac{\partial^2 z}{\partial \varphi^2} - \frac{\operatorname{tg}\varphi}{r^2}\,\frac{\partial z}{\partial \varphi}$$

where:

g: acceleration due to gravity (= 9.80665 m s^{-2})

f: Coriolis-parameter; f = $2\Omega \sin\varphi$

Ω: angular velocity of the earth (= $7.292\cdot 10^{-5}$ s^{-1})

φ: latitude

λ: longitude

r: earth radius (= $6371\cdot 10^{3}$m)

β: beta plane coefficient; $\beta = \left(\frac{df}{dy}\right)_\varphi = \frac{2\Omega}{r}\cos\varphi$

Static stability σ:

$$\sigma = \frac{g}{\theta}\,\frac{\partial \theta}{\partial z}$$, where $\sigma = N^2$; N: Brunt-Vaisala frequency

where θ : potential temperature.

Water vapor content Q:

We compute first the vapor pressure e:

$$e = \frac{RH}{100} \cdot e_s(T)$$

where:

RH: relative humidity

e_s (T): saturation vapor pressure, calculated from 'WMO-tables météorologiques internationales' (chap. 4.6).

The density of water vapor ϱ_w is then:

$$\varrho_w = \frac{e[hPa] \cdot 100}{R_w \cdot T[K]} \cdot \frac{1}{C_v} \qquad [kg \cdot m^{-3}]$$

where:

Rw: gas constant for water vapor (= 461.51 $J \cdot kg^{-1} \cdot K^{-1}$)

C_v: factor of compressibility for non-ideal gases (= 1, assumed).

Then Q is the sum of $\overline{\varrho_w} \, \Delta z$ of the three layers (units (kg water/m^2)).

5. Results of the PCA for the winter 1981/1982

The PCA is developped separately for each season to take the seasonal effects into account. The seasons overlap for 15 days over the adjacent ones; this has the advantage to smooth the transition from one season to the other, and also to enlarge a little bit the sample of data:

Spring:	February 16	- June 15
Summer:	May 16	- September 15
Automn:	August 16	- December 15
Winter:	November 16	- March 15

Results for the winter 1981/1982 will be shown. It is the period of 120 days from 16.11.-30.11.82 and 1.12.81-15.3.82. (The two weeks of November 1982 instead of November 1981 have been taken, because only one year of data from 1.12.81-30.11.82 were available at SMI. But this does not influence the sample of data or the interpretation of the results). The analyses of each day for 12 UTC will be considered. Table 1 gives the percentage of total variance with only 1 to 15 PCs, the number of PCs necessary to reach 99% of the total variance, and the index of the highest PC which had at least once in the season the largest value. The first column of the table gives the part of the standard deviation to the mean field.

Table 1: Percentage of total variance with i PCs for winter 1981/1982

f(x)		$\frac{s(f(x))}{f(x)}$	% of total variance with i PCs						Number of PCs to reach 99% of total variance	highest PC with largest value
Par.	level (hPa)		i = 1	i = 2	i = 4	i = 8	i = 12	i = 15		
geopotential	850	6%	86.4	94.4	99.1	99.91	99.98	99.993	4	5
	700	~2-3.5%	84.8	93.8	99.1	99.91	99.98	99.995	4	4
	500		79.1	91.1	98.8	99.88	99.97	99.993	5	4
	300		74.0	89.0	98.7	99.88	99.98	99.993	5	4
geostr. Wind u	850	~1	59.7	84.2	95.0	98.7	99.76	99.90	8	4
	700		64.5	88.0	95.9	99.3	99.82	99.93	7	4
	500		65.3	87.9	95.4	99.3	99.83	99.94	8	5
	300		67.8	88.8	95.6	99.4	99.84	99.93	8	5
geostr. wind v	850	~2	67.5	82.6	93.9	98.5	99.61	99.84	10	7
	700		68.7	84.5	95.0	98.9	99.73	99.88	9	5
	500		63.2	82.9	94.6	99.0	99.76	99.90	9	5
	300		64.4	84.0	95.8	99.1	99.75	99.91	8	5
geostr. relative vorticity	850	~1.5	27.5	48.6	70.7	89.8	96.0	97.7	20	8
	700		30.5	49.8	71.5	90.4	96.6	98.2	19	8
	500		31.1	48.3	72.8	91.7	97.3	98.6	17	8
	300	6	33.7	52.1	76.0	91.8	97.2	98.7	16	8
vertical vorticity	850	~2	30.9	48.3	70.0	87.9	95.1	97.4	20	8
	700		27.6	47.3	70.5	88.1	95.4	97.9	19	9
	500		31.9	51.8	73.5	89.7	96.2	98.0	18	9
	300		29.0	46.7	67.9	87.7	95.4	97.8	19	15
thickness	850-700	~2-3%	61.6	82.6	96.3	99.5	99.82	99.93	7	4
	700-500		63.5	82.6	96.6	99.5	99.89	99.96	7	5
	500-300		64.0	83.6	97.0	99.6	99.90	99.96	6	5
static stability	850-700	~0.25-0.5	38.1	62.7	84.6	93.7	98.5	99.3	14	7
	700-500		41.5	62.7	84.3	95.9	98.8	99.4	13	6
	500-300		58.3	74.5	91.2	98.2	99.6	99.81	10	5
water vapor content	850-300	0.3	50.6	70.0	87.5	97.2	99.1	99.6	12	6
									302	

5.1 Interpretation of the results

For the fields of geopotential 850 hPa and geostrophic relative vorticity 850 hPa we shall look more in detail to the results of the PCA.

Geopotential field 850 hPa

Fig. 6 shows the mean field and the standard deviation. A well-known fact is that standard deviation is only a few percent of the mean field. Fig. 7 shows the first four eigenvectors (EVs) (figures are multiplied by 10^3). The first one accounts for already 86% of total variance. The values at all grid-points are about equal. As will be seen later, this first EV gives mainly the information of the mean field. The second EV gives 8% of total variance. It gives the zonal contribution to the field. The sign of the values must not be interpreted; this EV gives the information for the west-circulation (high pressure in the south) as well as for the east-circulation (low pressure in the south). The third EV gives with 4% of total variance the meridional contribution to the field. From the fourth EV (only 0.5% of total variance) a physical interpretation is almost impossible.

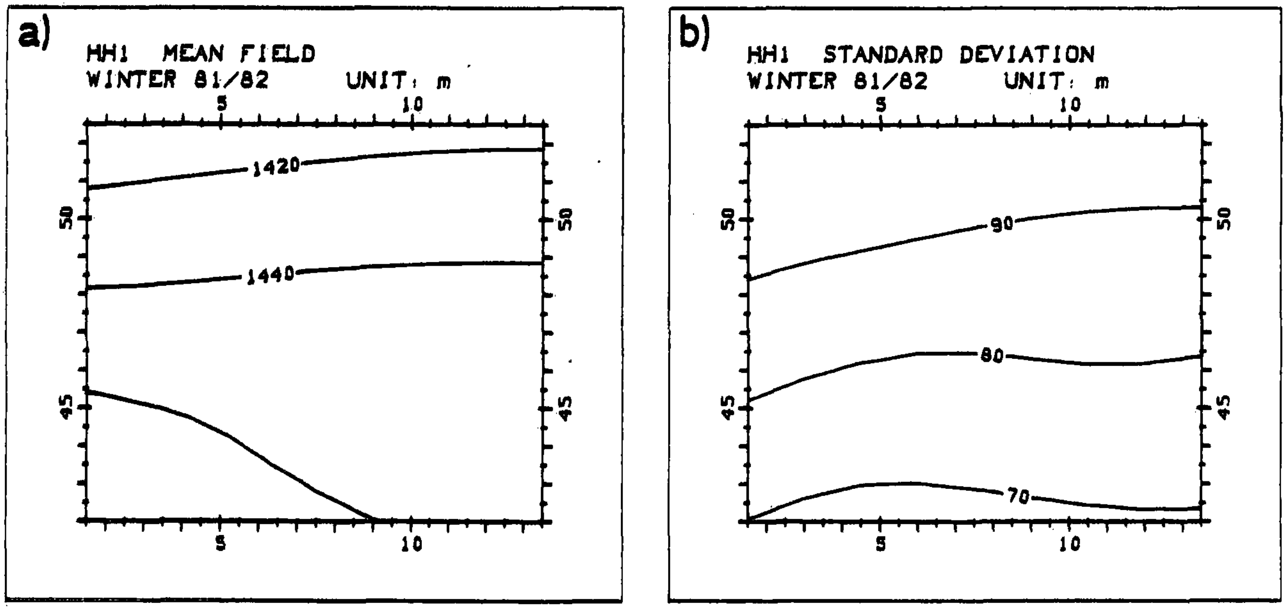


Fig. 6 Mean field (a) and standard deviation (b) of geopotential 850 hPa in the grid 42 - 52.5 N and 1.5 - 13.5 E for winter 1981/1982.

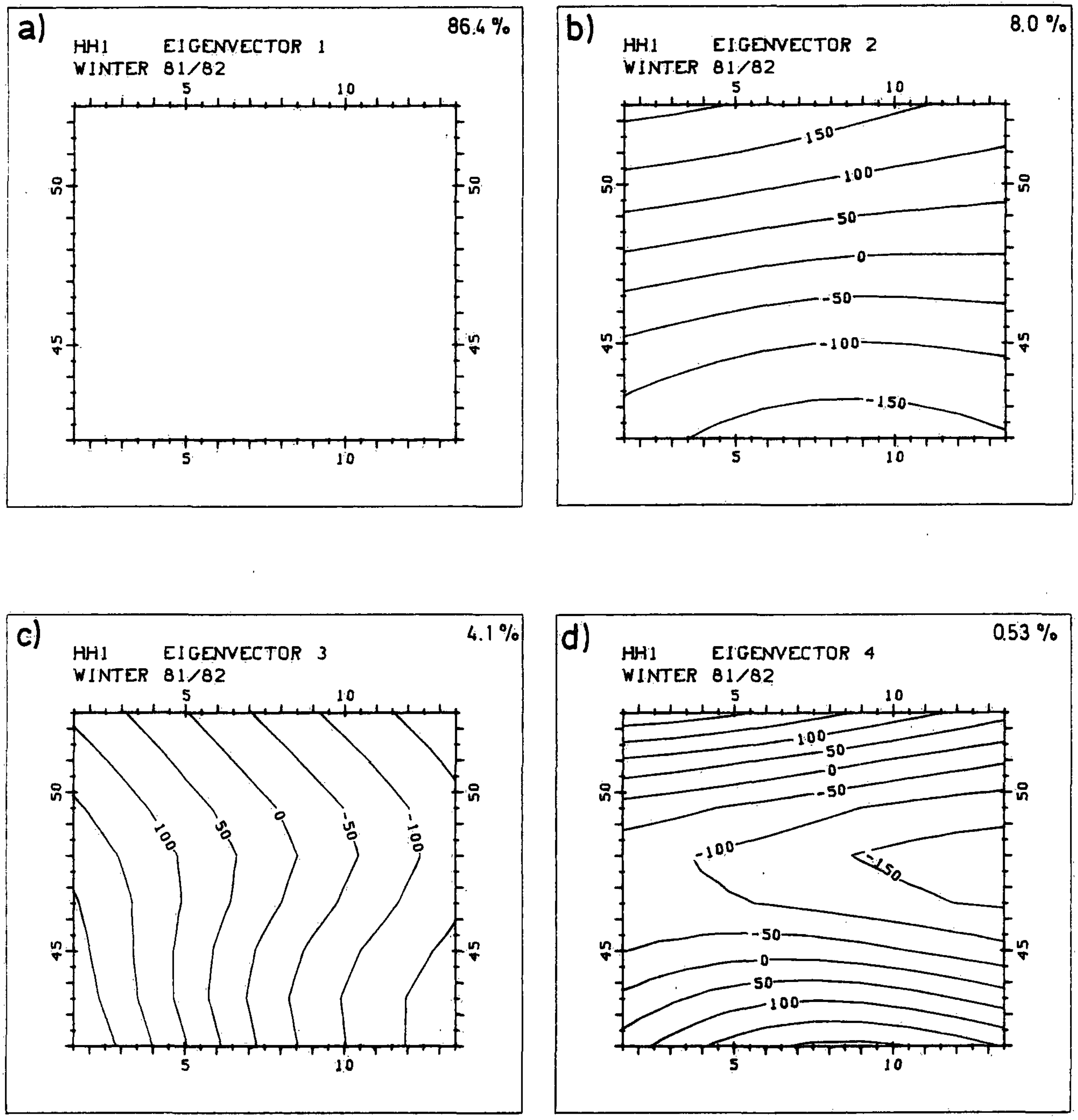


Fig. 7 The first four eigenvectors of geopotential 850 hPa in the grid 42 - 52.5 N and 1.5 - 13.5 E for winter 1981/1982. At the right upper corner are the percentage of total variance of the eigenvector. Figures are multiplied by 10^3.

Fig. 8a shows the analysis field for December 1, 1981, 12 UTC. With Eq. (6) we shall reconstruct the initial field with only the few first PCs (Fig. 8b-g). With one PC the field is very different from that of the analysis. It looks like the mean field for the whole winter (Fig. 6a). So it can be said the first eigenvector gives about the mean field (this is not the case for fields like vorticity (see later) or vertical velocity, because for these fields the mean standard deviation for the whole winter has the same order of magnitude as the mean field itself). Horel (1981, p. 1082) points out that "the first PC often represents a useful objectively-determined weighted average of the original data". With 2 PCs the field is still very different from that of the analysis, in spite of the already 94.5% of total variance. So, we can agree with Richman (1981, p.1151) that "the percent of variance extracted may not be a very useful criterion for determining how well the map types depict the data". Only with the third PC the field looks like the analysis (errors of maximum 20 meters). It is so because on December 1 (Fig. 8a) we had a meridional N-flow over the Alps, and the meridional component of a field comes only with 4% of total variance in the third eigenvector (Fig. 7c). With 4 PCs (99.1% of total variance) the errors are now of maximum 10-15 meters. With 8 PCs the errors are of only a few meters.

So, for that analysis the third PC had the largest value (of 3.5, compared to 2.4 for PC1 and 2.7 for PC2). It happened even once in the 120 days of the sample that the fifth PC (with only 0.5% of total variance) had the largest value (see Table 1). It was the situation of January 8, 1982, 12 UTC where all five first PCs had small values (-1.3 for PC1, -0.6 for PC2, -0.5 for PC3, + 1.4 for PC4 and + 1.5 for PC5). Fig. 9 gives the projection of the 120 analyses in the plane formed by the second and third PC-axis. 12 different symbols identify each a period of 10 days. The days from December 1, 1981 - December 6, 1981 are marked on the figure. The analyses which are dominated by meridional circulation will be recognized as nearer to the third PC as to the second (N-flow in the upper part, S-flow in the lower part). This kind of figure is therefore also an interesting tentative method for classifying different weather types. An example is the point on the axis of the second PC at value +5: it is the situation of November 29, 1982, 12 UTC, with a low pressure in the Mediterranean, and so a zonally east circulation in the grid domain.

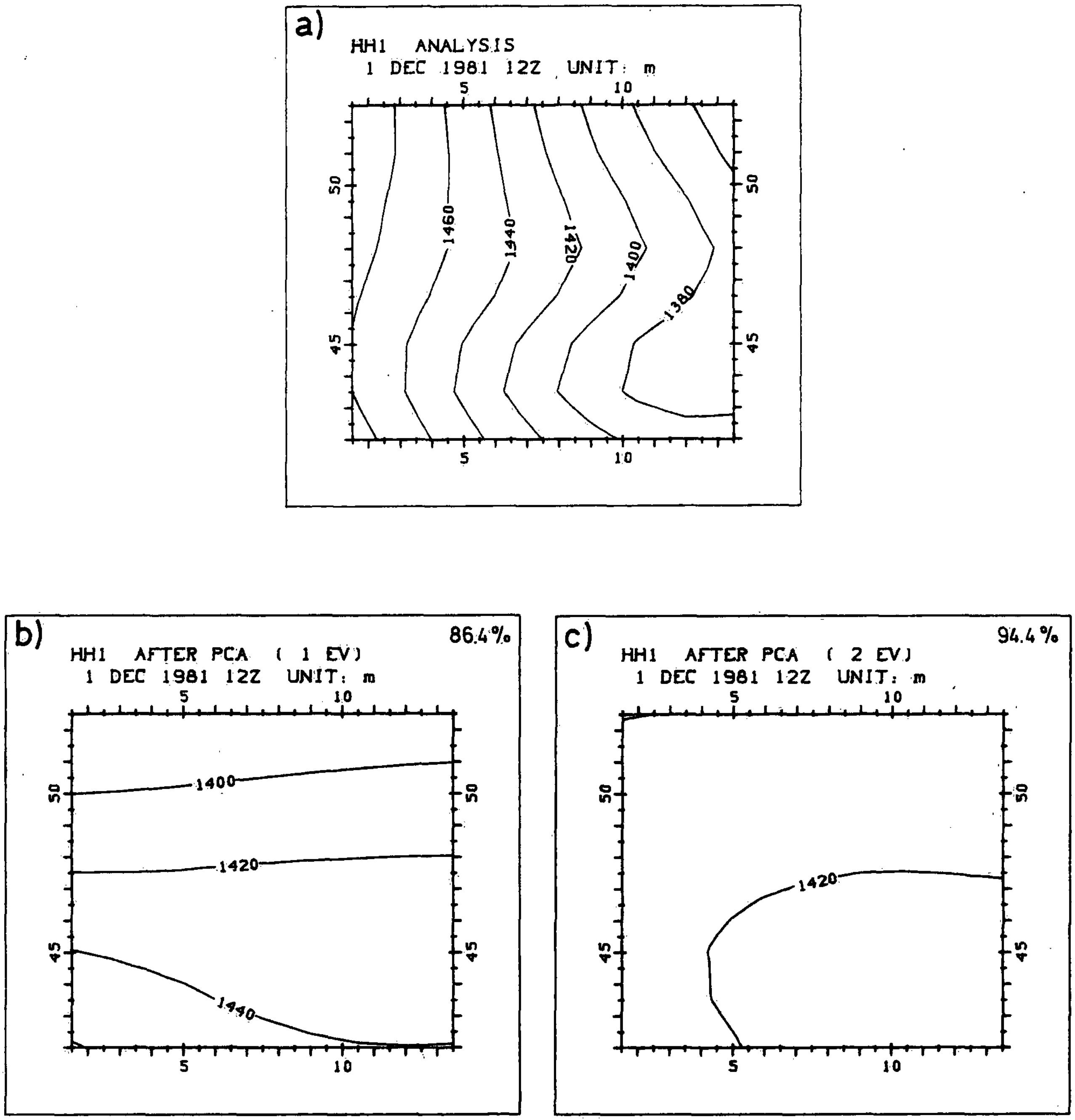


Fig. 8 Geopotential 850 hPa for December 1, 1981, 12 UTC in the grid 42 - 52.5 N and 1.5 - 13.5 E. Analysis from ECMWF (a) and reconstructed field with 1 (b), 2 (c), 3 (d), 4 (e), 8 (f) and 15 (g) PCs. At the right upper corner are the percentage of total variance of the PCs used for the reconstructed field.

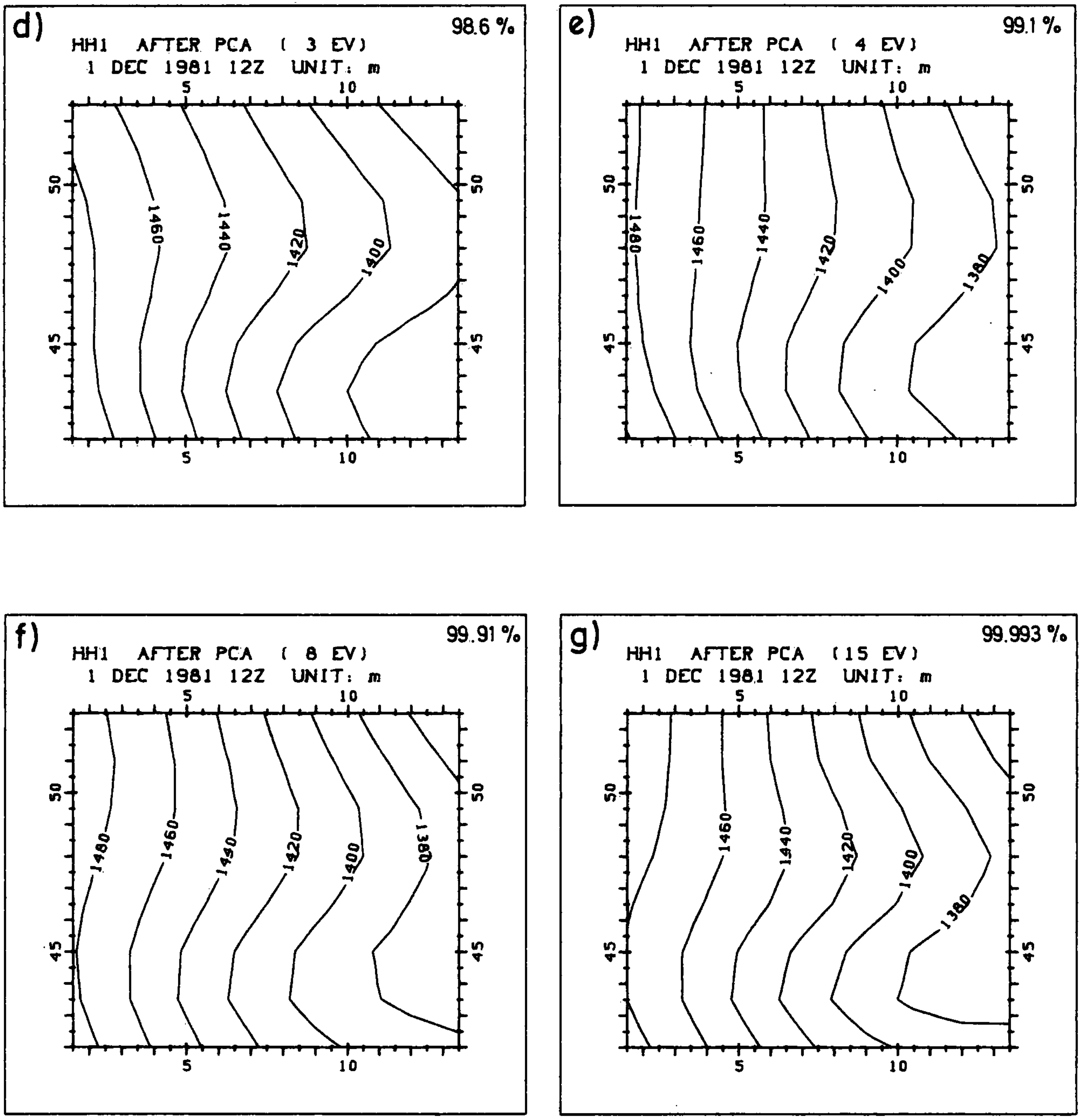


Fig. 8 d - g: see previous page.

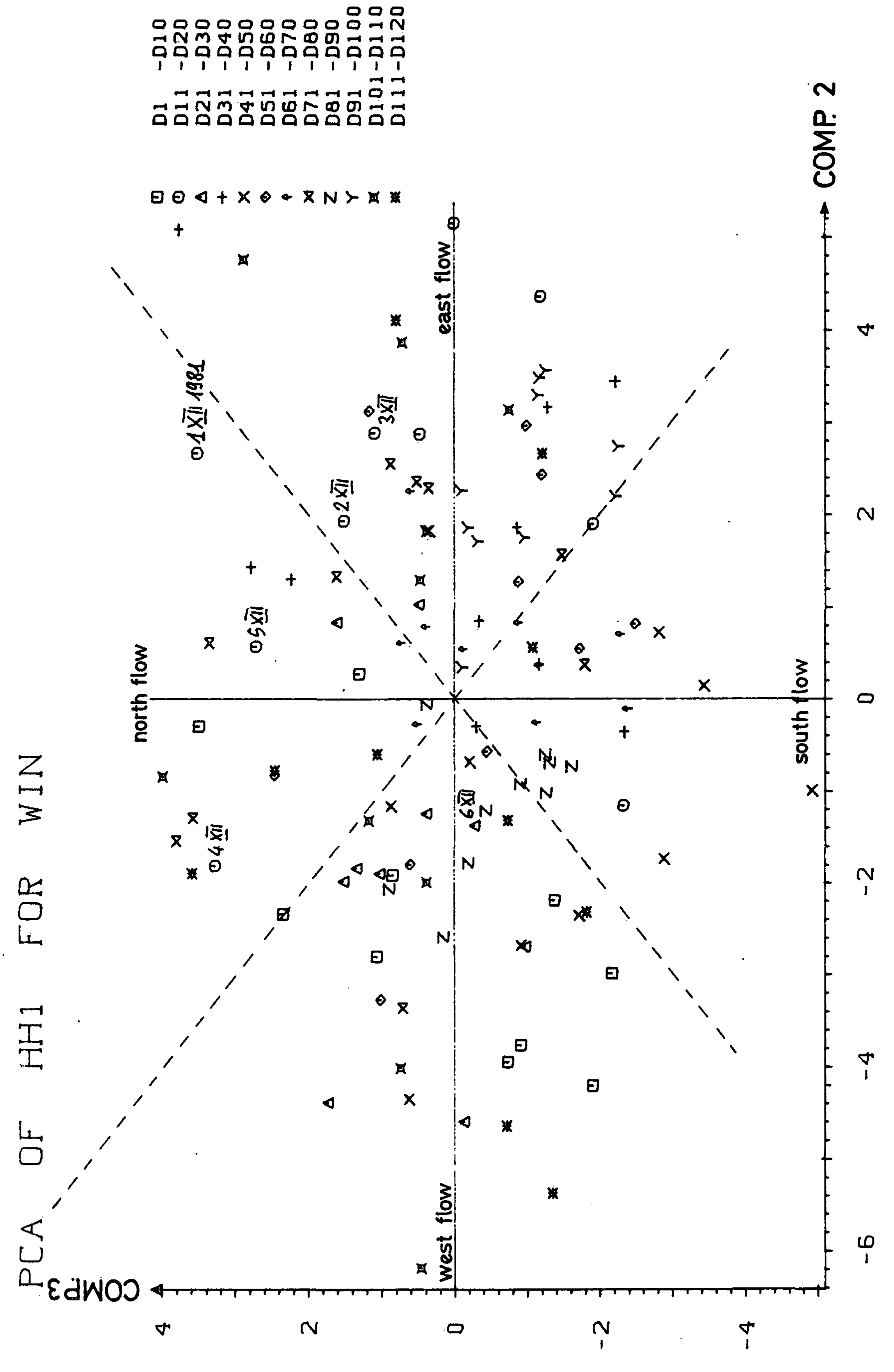


Fig. 9 Projection of the 120 analyses of geopotential 850 hPa in the grid 42 - 52.5 N and 1.5 - 13.5 E for winter 1981/1982 on the second and third PC-axis. 12 different symbols identify each a period of 10 days. The days from December 1, 1981 - December 6, 1981 are labeled.

Field of geostrophic relative vorticity 850 hPa

Fig. 10 shows the mean field and the standard deviation. The standard deviation has the same order of magnitude as the mean field itself. Fig. 11 depicts the first four EVs. The first EV accounts for only 27% of total variance, and the second for 21%. Already for these two first EVs a physical interpretation is difficult. It must not be forgotten that the EVs give information relative to the standardized fields. It looks as the first EV gives a zonal contribution to the deviations from the mean field.

As for geopotential we will reconstruct the analysis field of geostrophic relative vorticity of December 1, 1981, 12 UTC, with the first PCs. Fig. 12 gives these results. With two PCs, and a total variance of 48%, we already have a field with the same pattern as the initial one but the center of maximum positive vorticity is 3° of longitude located far to the west. With 4 PCs (total variance of

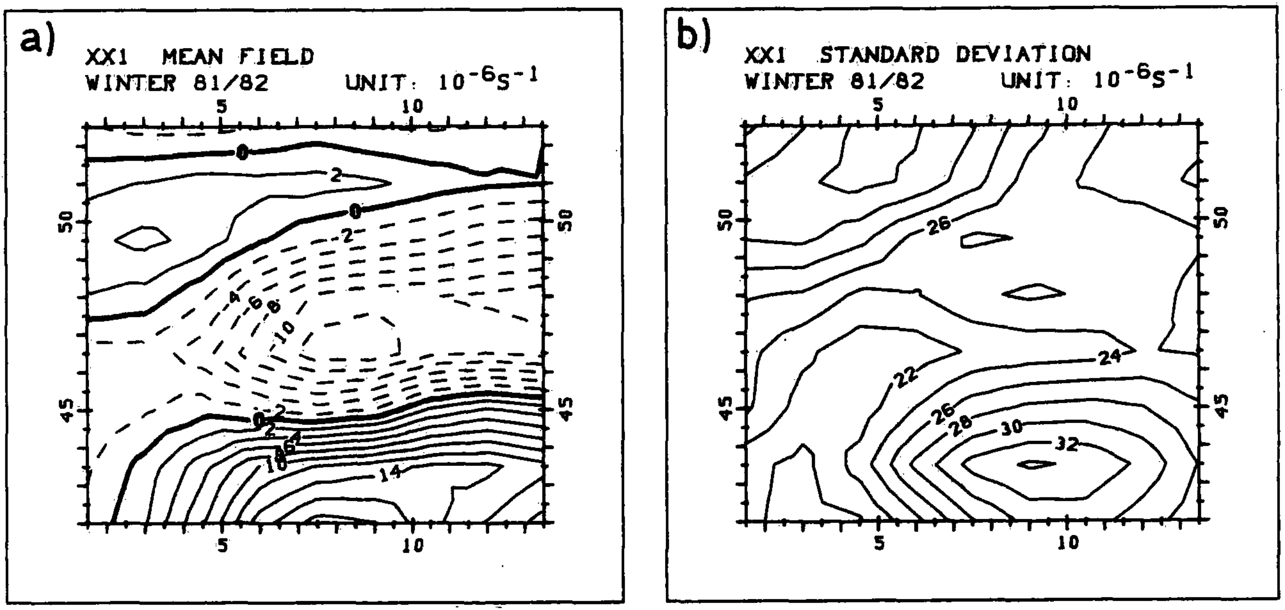


Fig. 10 Mean field (a) and standard deviation (b) of geostrophic relative vorticity 850 hPa in the grid 42 - 52.5 N and 1.5 - 13.5 E for winter 1981/1982.

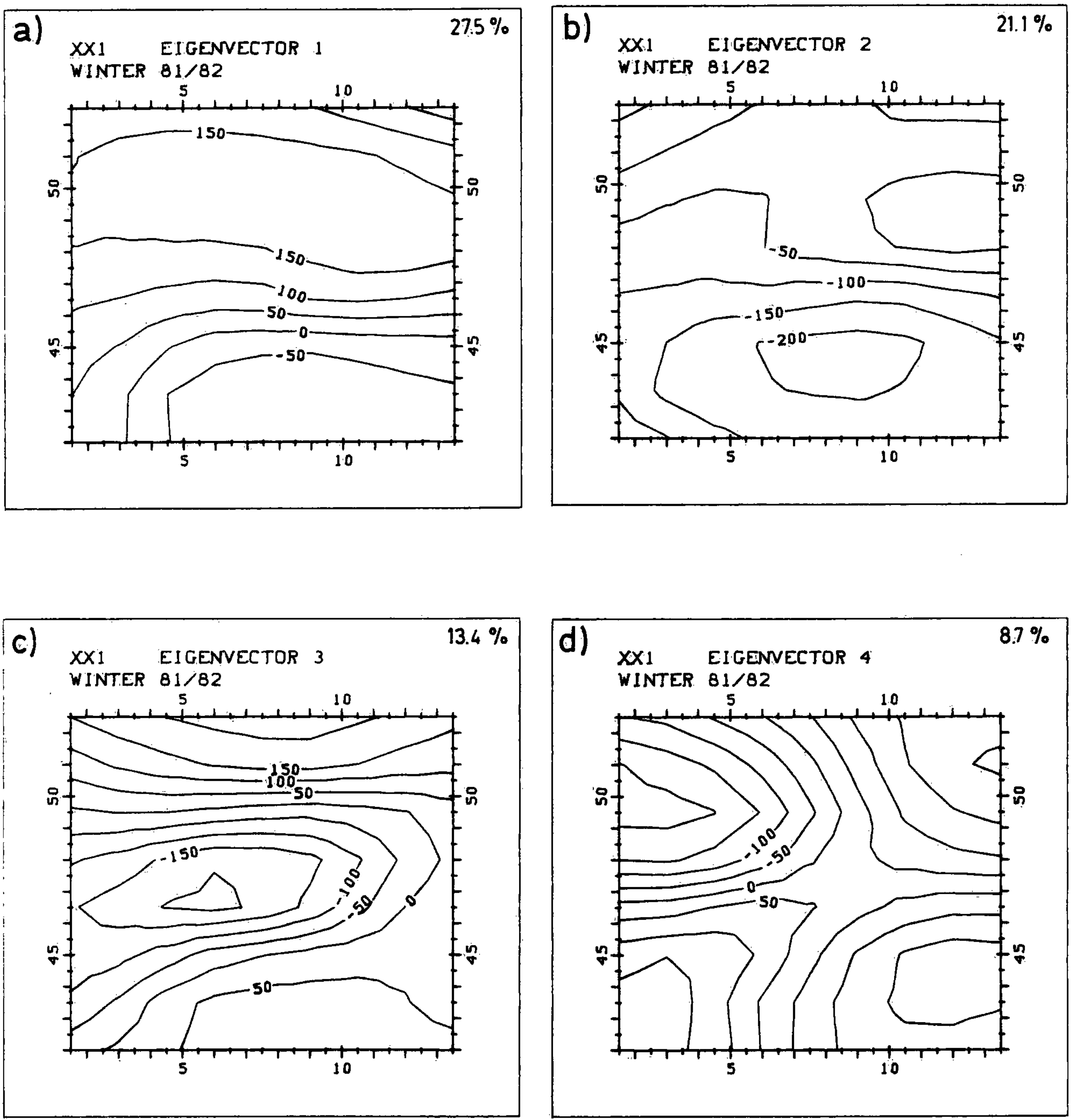


Fig. 11 The first four eigenvectors of geostrophic relative vorticity 850 hPa in the grid 42 - 52.5 N and 1.5 - 13.5 E for winter 1981/1982. Figures are multiplied by 10^3. At the right upper corner are the percentage of total variance of the eigenvector.

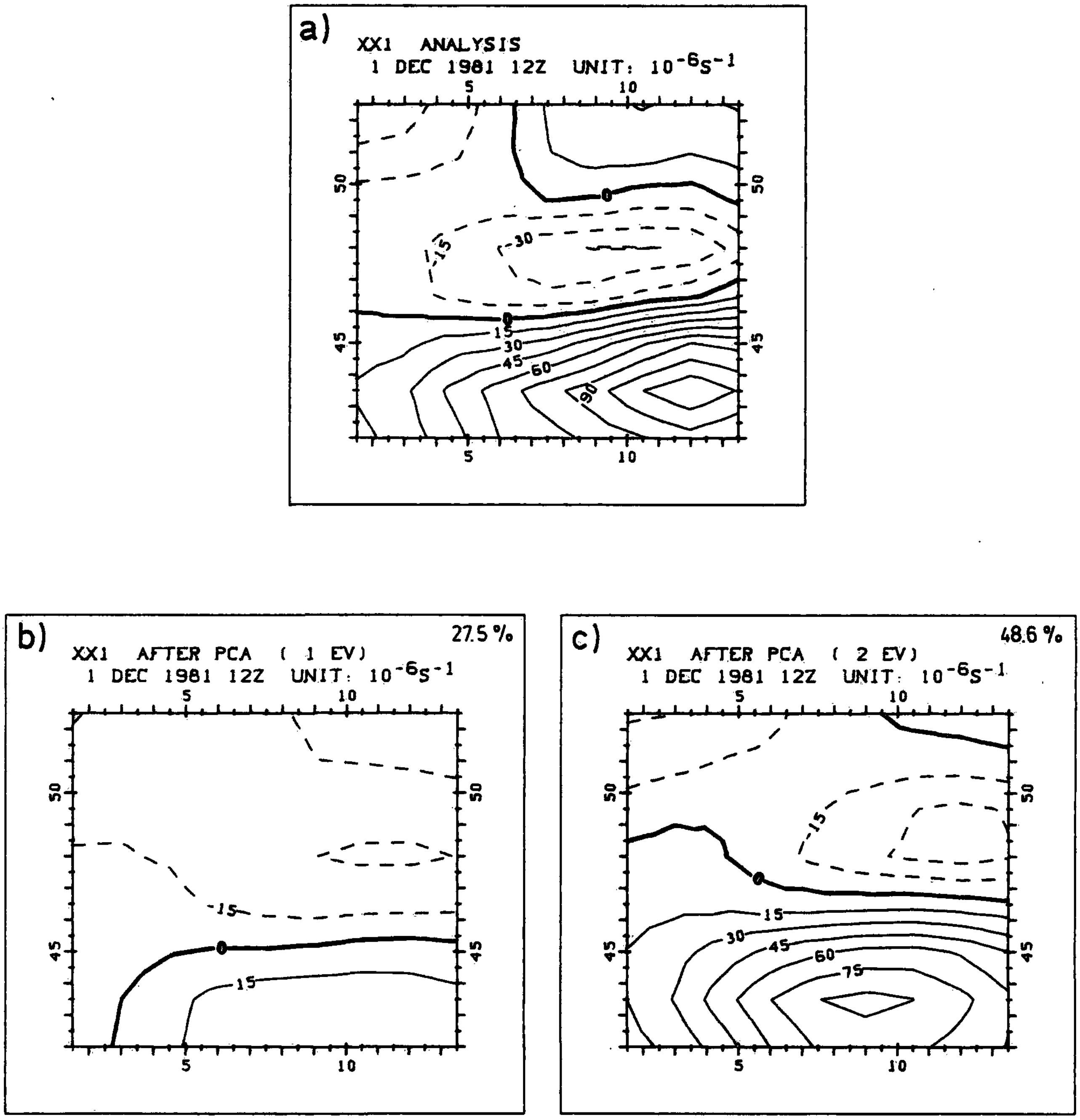
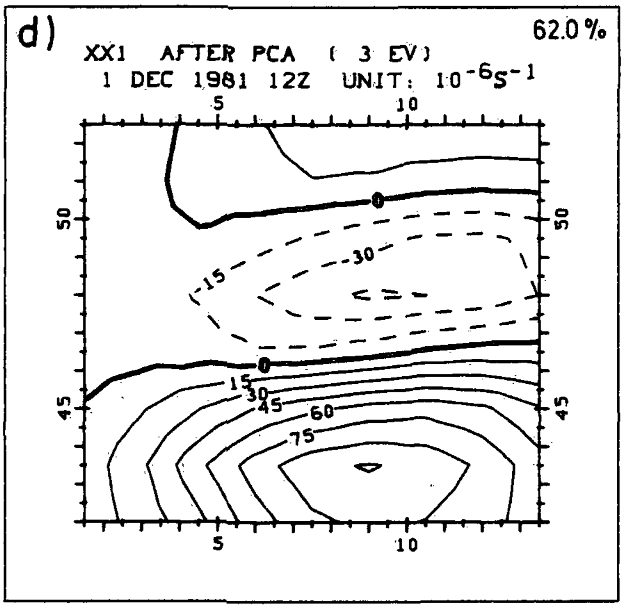
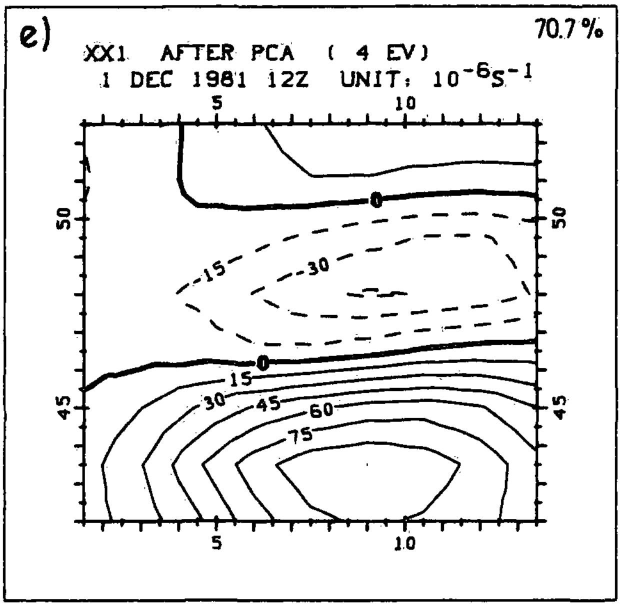
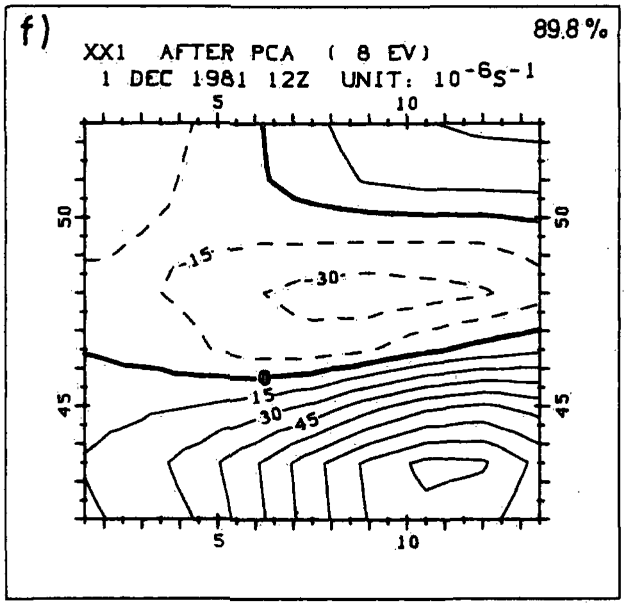
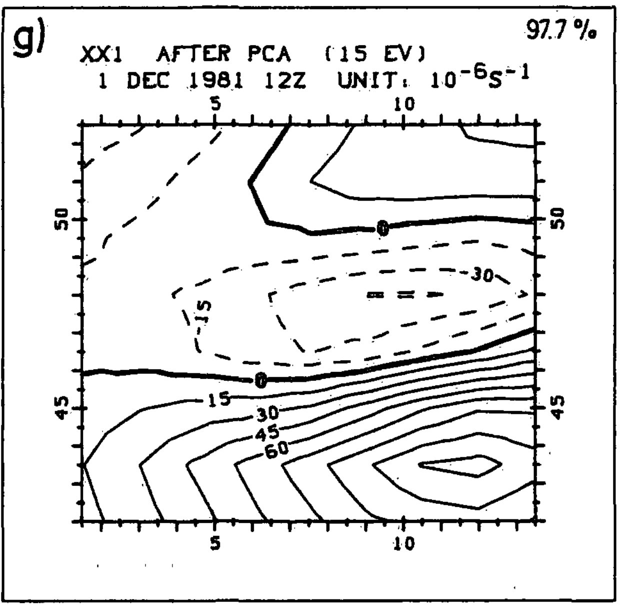


Fig. 12 Geostrophic relative vorticity 850 hPa for December 1, 1981, 12 UTC in the grid 42 - 52.5 N and 1.5 - 13.5 E. Analysis from ECMWF (a) and reconstructed field with 1 (b), 2 (c), 3 (d), 4 (e), 8 (f) and 15 (g) PCs. At the right upper corner are the percentage of total variance of the PCs used for the reconstructed field.


d)
62.0 %
XX1 AFTER PCA (3 EV)
1 DEC 1981 12Z UNIT: 10⁻⁶s⁻¹
5
10
50
45
0
-15
-30
15
30
45
60
75


e)
70.7 %
XX1 AFTER PCA (4 EV)
1 DEC 1981 12Z UNIT: 10⁻⁶s⁻¹
5
10
50
45
0
-15
-30
15
30
45
60
75


f)
89.8 %
XX1 AFTER PCA (8 EV)
1 DEC 1981 12Z UNIT: 10⁻⁶s⁻¹
5
10
50
45
0
-15
-30
15
30
45


g)
97.7 %
XX1 AFTER PCA (15 EV)
1 DEC 1981 12Z UNIT: 10⁻⁶s⁻¹
5
10
50
45
0
-15
-30
15
30
45
60

70%) the errors are still of the same order (up to $6.5 \cdot 10^{-5}$ s^{-1}). With 8 PCs (90% of total variance) the errors still reach $2 \cdot 10^{-5}$ s^{-1}, and then with 15 PCs (97.7% of total variance) the differences between the reconstructed and initial field are only of maximum 10^{-5} s^{-1}.

5.2 Truncation criterion

Fig. 13 shows the percentage of total variance of the first PCs of geopotential 850 hPa and geostrophic relative vorticity 850 hPa. It is already evident that for the geopotential a large part of the information is contained in very few PCs, and that on the other hand for the vorticity the compression of the data is less efficient.

The different truncation criteria defined in the literature have been exposed in Section 2. For our project we will follow a synoptic approach to define the criterion. Our objective is to retain at least the whole meteorological significant information for the precipitations. For geopotential 850 hPa and geostrophic relative vorticity 850 hPa of December 1, 1981, 12 UTC, we have seen that if we retain all PCs responsible for 99% of total variance we can reconstruct the initial field with enough precision, i.e. the whole meteorological information is contained in these first PCs. Other such analyses for other dates and also for the other predictor fields show that for our project a total variance of 99% is a good criterion for separating "noise" from meteorological information. Table 1 shows how many PCs are necessary to fulfil this condition. For the geopotentials 4-5 PCs (instead of the 72 grid points) are already sufficient; but for the fields of geostrophic relative vorticity and vertical velocity 18-20 PCs are necessary.

To do a comparison with the literature we have also computed the LEV-graphs (see Section 2). The test for this criterion is not as evident as in many papers. For the geopotential (Fig. 14) the quasi-straight line from eigenvalue 7 to 40 is (or seems to be) already due to noise components. For the geostrophic relative vorticity (Fig. 15) the LEV-graph would give a truncation after approximately 30 PCs (total variance of 99.7%).

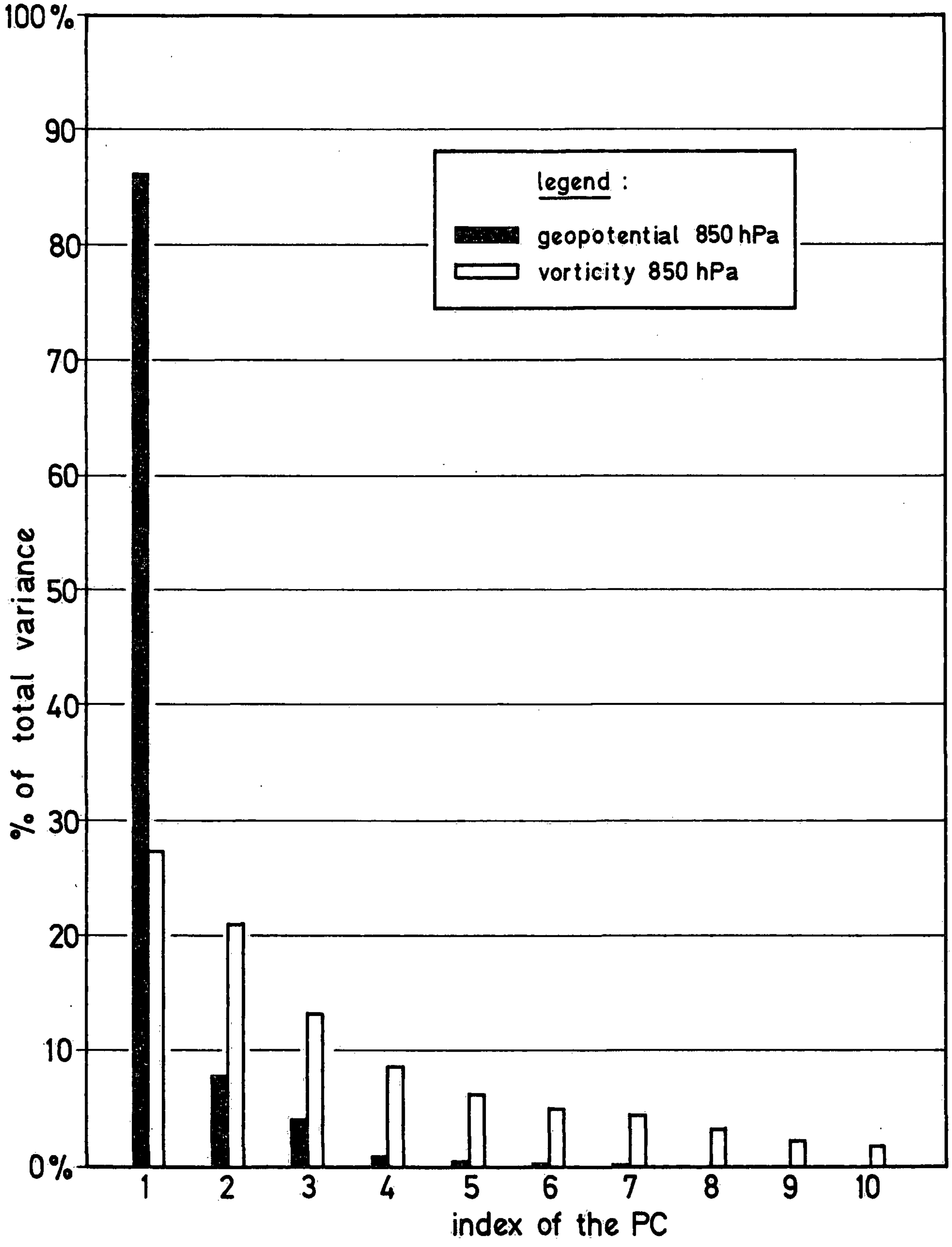


Fig. 13 Percentage of total variance of the first PCs of geopotential 850 hPa and geostrophic relative vorticity 850 hPa for the winter 1981/1982 in the grid 42 - 52.5 N and 1.5 - 13.5 E.

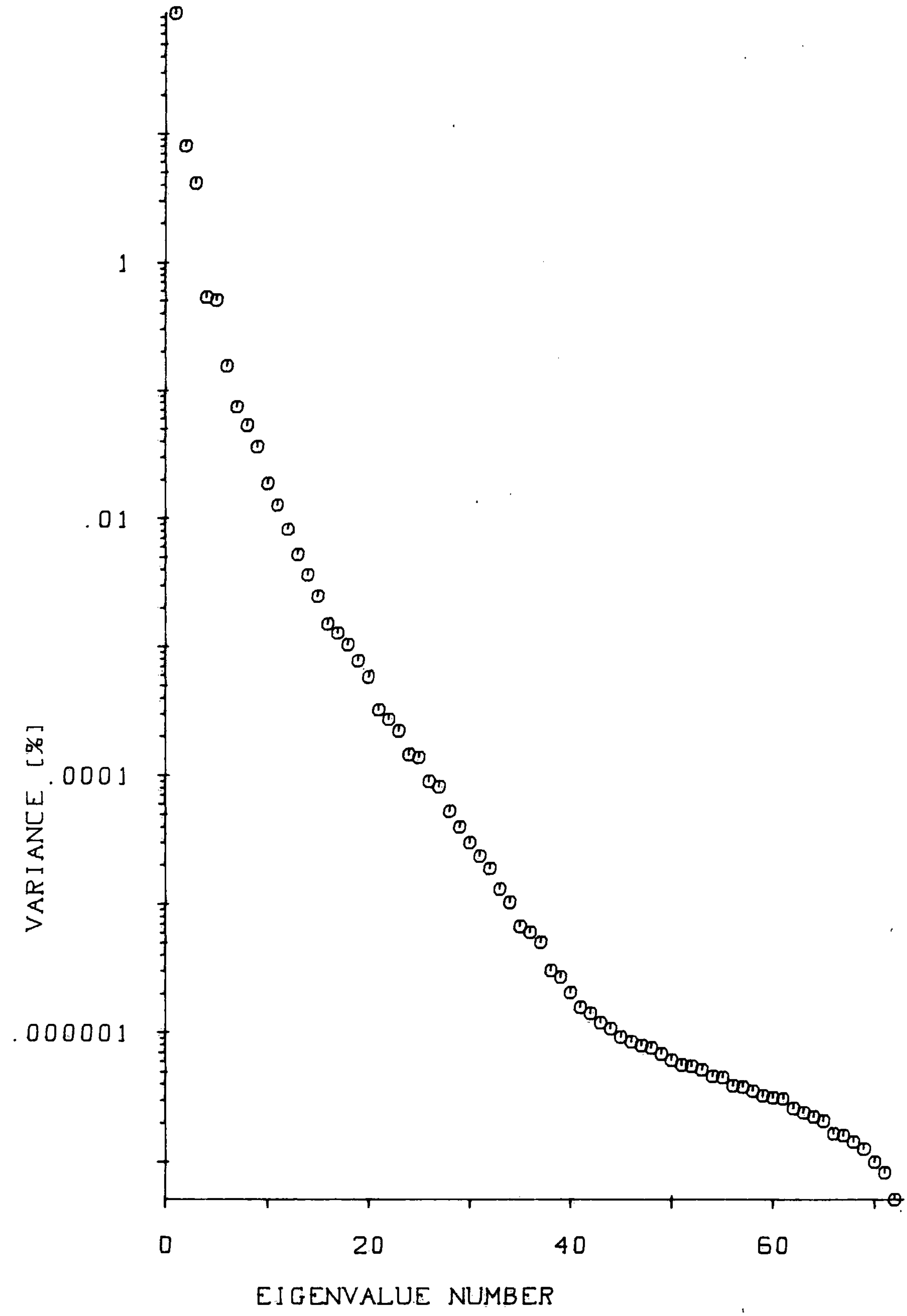


Fig. 14 LEV-graph for geopotential 850 hPa for winter 1981/1982 in the grid 42 - 52.5 N and 1.5 - 13.5 E.

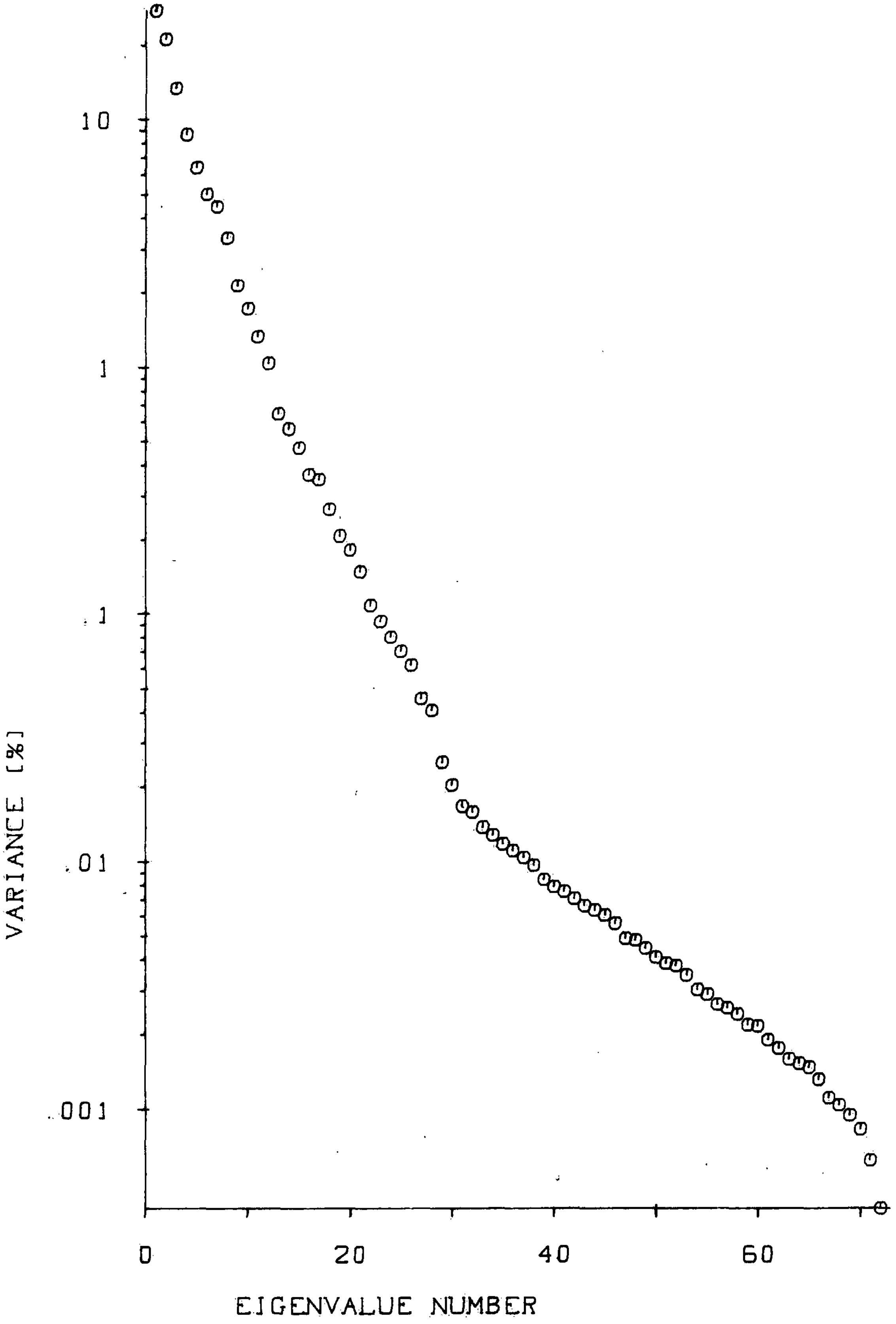


Fig. 15 LEV-graph for geostrophic relative vorticity 850 hPa for winter 1981/1982 in the grid 42 - 52.5 N and 1.5 - 13.5 E.

References

Ambühl, J., 1984: Interprétation statistique des prévisions numériques. Working Report of the Swiss Meteorological Institute No. 123. 52 pp.

Ashbaugh, L.L., Myrup, L.O., Flocchini, R.G., 1984: A principal component analysis of sulfur concentrations in the western United States. Atmospheric Environment, 18, 783-791.

Christenson, W.I., Bryson, R.A., 1966: An investigation of the potential of component analysis for weather classification. Mon.Wea.Rev., 94, 697-709.

Cohen, S.J., 1983: Classification of 500 mb height anomalies using obliquely rotated principal components. J. Climate Appl. Meteor., 22, 1975-1988.

Craddock, J.M., Flood, C.R., 1969: Eigenvectors for representing the 500 mb geopotential surface over the northern hemisphere. Quart. J.R. Met. Soc., 95, 576-593

ECMWF, 1977: Workshop on the use of empirical orthogonal functions in meteorology. European Centre for Medium Range Weather Forecasts. 155 pp.

Flury, B., Riedwyl, H., 1983: Angewandte multivariate Statistik. Gustav Fischer Verlag. Stuttgart, New York. 187 pp.

Goossens, C., 1985: Principal component analysis of mediterranean rainfall. J. Climat., 5, 379-388.

Horel, J.D., 1981: A rotated principal component analysis of the interannual variability of the northern hemisphere 500 mb height field. Mon.Wea.Rev., 109, 2080-2092.

Kuhn, W., Quiby, J., 1976: Dynamical-statistical methods of meso-scale precipitation forecasting over mountaineous terrain. Pure Appl. Geophys., 114, 945-964.

LeDrew, E.F., 1980: Eigenvector analysis of the vertical velocity field over the eastern canadian arctic. Mon.Wea.Rev., 108, 1992-2005.

Melice, J.L., Wendler, G., 1984: Precipitation statistics in Southern Tunisia: a contribution to the desertification problems in the Sahel Zone. Arch.Met.Geoph.Biocl., Ser. B33, 331-348.

Molteni, F., Bonelli, P., Bacci, P., 1983: Precipitation over Northern Italy: A description by means of principal component analysis. J. Climate Appl. Meteor., 22, 1738-1752.

Mori, N., 1984: 192-hour prognosis of precipitations by using numerical prediction products: application of EOF to multiple-regression formula. J.Meteor.Soc.Japan, Ser. II, 62, 183-189.

North, G.R., 1984: Empirical orthogonal functions and normal modes. J. Atmos.Sci., 41, 879-887.

Overland, J.E., Preisendorfer, R.W., 1982: A significance test for principal components applied to a cyclone climatology. Mon.Wea.Rev. 110, 1-4.

Paegle, J.N., Haslam, R.B., 1982: Statistical prediction of 500 mb height field using eigenvectors. J.Appl.Meteor., 21, 127-138.

Preisendorfer, R.W., Barnett, T.P., 1977: Significance tests for empirical orthogonal functions. Preprints Fifth Conf. Probability and Statistics in Atmospheric Sciences, Las Vegas, Amer. Meteor. Soc., 169-172.

Preisendorfer, R.W., Zwiers, F.W., Barnett, T.P., 1981: Foundations of Principal Component Selection Rules, SIO Reference Series 81-4, Scripps Institution of Oceanography, La Jolla, CA.

Quiby, J., 1984: L'analyse en composantes principales et son application en météorologie. Working Report of the Swiss Meteorological Institute no. 122, 25 pp.

Richman, M.B., 1981: Obliquely rotated principal components: an improved meteorological map typing technique ? J. Appl. Meteor., 20, 1145-1159.

Richman, M.B., 1986: Rotation of principal components. J.Climat., 6, 293-335.

Rinne, J., Järvenoja, S., 1979: Truncation of the EOF series representing 500 mb heights. Quart. J.R. Met.Soc., 105, 885-897.

Rinne, J., Karhila, V., 1979: Empirical orthogonal functions of 500 mb height in the northern hemisphere determined from a large data sample. Quart. J.R. Met. Soc., 105, 873-884.

Savijärvi, H.: Verification and storing with empirical orthogonal functions. ECMWF Internal Rep. 18, 36 pp.

Sneyers, R., Goossens, C., 1985: L'analyse par la méthode des composantes principales: Application à la climatologie et à la météorologie. Organisation météorologique mondiale. 9e session de la Commission de climatologie, Appendice au Doc. 5, 24 pp. + 26 pp.

Storch, H. von, Hannoschök, G., 1985: Statistical aspects of estimated principal vectors (EOFs) based on small sample sizes. J. Climate Appl. Meteor., 24, 716-724.

White, R.M., Cooley, D.S., Derby, R.C., Seaver, F.A., 1958: The development of efficient linear statistical operators for the prediction of sea-level pressure. J.Meteor., 15, 426-434.

Author's address:

Francis Schubiger
Swiss Meteorological Institute
Krähbülstrasse 58

CH-8044 Zurich