Statistical Post-Processing

of

Precipitation Forecasts

Åke Johansson

January 2005

CONTENTS

1. INTRODUCTION

After half a century of existence Numerical Weather Prediction (NWP) is a mature and highly sophisticated application of basic physical science. Considerable advances have been achieved during these 50 years with the construction of ever more realistic models that describe more and more facets of the atmosphere and its surrounding media. Concomitant with these developments are a monotonic increase in forecast skill. However, despite these impressive developments only limited success has been achieved in forecasting those elements that are of greatest concern to the general public, such as the occurrence, type and amount of precipitation, cloudiness, surface winds and maximum and minimum temperature.

It should be emphasized that there continues to exist appreciable systematic errors in the NWP models, many of which have persisted through all the years of model developments. These errors limit the effectiveness of the models in almost all applications. The systematic errors are not only visible in the ensemble or time-mean fields, but are equally disturbing for the variability in the models. To understand the nature and the cause of the systematic errors continues to be a high-priority research area in numerical modeling.

NWP models have a finite spatial resolution. This implies that the direct model output (DMO) values at a specific location are derived from a smoothed field and should be regarded as a spatial mean value. In many applications, however, we are interested in forecasts at a specific point and/or in small scale features. Appropriate statistical post-processing can in many cases provide this kind of information.

As mentioned above there are many parameters of interest that the NWP models do not predict explicitly. Most notable are various forms of probabilities, e g probability of precipitation (POP). Also in this case can appropriate statistical post-processing provide the desired information.

In order to determine which method, or combination of methods, that produces the best forecasts, it is crucial to have a system that can calculate the forecast skill in a meaningful and reliable manner. It is hard to overemphasize the importance of a good verification system since it enables a determination of which method, or combination of methods, that are the best at a certain forecast lead time.

The purpose of the present study is to use well established statistical post-processing techniques to increase the forecast skill of precipitation in the small Torpshammar catchment area in central Sweden. The technique used relies heavily on the experience of the Technique Development Laboratory of the US Weather Service (Dallavalle 2002, Antolik 2000 and Glahn et al, 1991). The uniqueness of the study is the use of a forecast data set, the ERA-40 reforecasting data set, with unique properties, namely that it spans a very long time period (45 years) and that it is as homogeneous as is possible to achieve.

The question that we primary address is how much skill improvement above DMO that is possible to achieve with the proposed technique and how does it compare with alternative techniques ? A closely related question is how long should the reforecasting time series be ?

2. STATISTICAL POST-PROCESSING METHOD

a. Model Output Statistics - MOS

The post-processing method that we will use in this study is the Model Output Statistics (MOS) technique (Glahn and Lowry 1972). It has been used extensively with very good results at many operational weather services for over 30 years, most notably at the Technique Development Laboratory of the US Weather Service. It consists of determining a statistical relationship between observed weather elements (predictands) and appropriate variables (predictors). The predictors are variables from:

- NWP model forecast at some forecast lead time = Direct Model Output (DMO)
- Prior observations

The MOS technique is thus an extension of the Perfect Prog method in that it uses a larger set of predictors, most notably DMO. The predictor sample in MOS usually consists of a relatively short period of prognostic data produced by NWP model(s). In this way the bias and inaccuracy of the NWP model, as well as the local climatology, can be automatically built into the forecast system.
MOS can be looked upon as a way to extract the weather-related statistics of an NWP model. The major advantages of the MOS technique are:

- It removes some of the systematic model biases
- It accounts for certain phase errors
- It predicts explicitly the weather elements of common public concern
- Generates forecasts at specific sites
- Produces reliable probabilities
- Provides information on model predictability by producing forecasts that tend toward the sample mean with increasing forecast lead time and by selecting different predictors for different forecast lead times.

In most operational settings the principle shortcoming of the MOS method is the frequent model changes (typically 4-6 changes a year) in the dynamical model and its data assimilation system. These changes cause the forecast data sets to be inhomogeneous, which in turn decreases the reliability of the derived statistical relationships. The uniqueness of this study is that we have access to a very long forecast data set that has been produced by a model that has been held fixed. The data set is thus as homogeneous as is possible in addition to being very long (changes in the observation system is however a source of inhomogeneities in the data set).

b. Linear regression

The most common statistical method is linear regression where one variable is regressed upon another variable. The next higher step, is multiple regression, where a vector of predictor data is related to a single predictand variable. When the vector of predictor data is large any regression scheme will build a model capable of accounting for large amounts of variance. However, a large fraction of this is artificial. There is, therefore, a need for a procedure to select the most predictively important variables to explain the single predictand in such a way that the result will hold up on independent data. The most common such procedure is called stepwise multiple regression.

Canonical correlation analysis (CCA) is the generalization of all these approaches. It finds the optimum linear combination of the predictor data vector that will explain the most variance in the predictand data vector. Both predictor and predictands are thus multidimensional vectors chosen by the CCA methodology. This implies that we do not know in advance what the predictand vector is going to be. This fact makes the CCA technique not suited for the purposes discussed here.
The most common statistical method used in MOS is multiple linear regression.

c. Multiple linear regression

Multiple linear regression relates one variable, y, called the dependent variable or predictand, to k other variables, x_i, called the independent variables or predictors. The resulting equation can be used to estimate the predictand as a linear combination of the predictors:

$$\hat{y} = a_0 + a_1 x_1 + a_2 x_2 + ... + a_k x_k = a_0 + \sum_{i=1}^{k} a_i x_i$$

where the a_i's are the regression coefficients. The a_i's are usually determined such that the square of the estimation errors is a minimum on a developmental (or dependent) sample of size n, i e the minimum of the quantity

$$Q_o = \sum_{j=1}^{n} \left(y_j - \hat{y}_j \right)^2$$

is sought for. This procedure is usually referred to as the method of least squares. Then

$$Q = \sum_{j=1}^{n} \left(y_j - \bar{y} \right)^2 = Q_1 + Q_o$$

where Q_1 is the explained part of Q and Q_o the unexplained part (which is minimized). A measure of the goodness of the regression equation is the quantity

$$R^2 = \frac{Q_1}{Q}$$

which is the fractional part of the variance that is explained by the regression equation.

d. Stepwise multiple linear regression

There exist many procedures for selecting the predictors to be included in a multiple regression equation. The most widely used in meteorology and which is also used at TDL is the forward selection procedure. This algorithm selects the predictor with the best linear correlation to the predictand and computes a one-predictor equation. It then selects the predictor with the best correlation to the residual. The new equation is computed, and again the best correlated predictor is selected. Thus, intercorrelations among the predictors are accounted for and redundant predictors may be excluded from the equation. The process continues until the correlation to the residual is lower than a critical value depending on the size of the development sample and the number of potential predictors.

It is well known that all of the variables should not necessarily be used as predictors in a linear regression equation. The optimum number of predictors to include in the equation must be determined. This is not a trivial problem, as evidenced by the many different regression methods available, none having proved itself better than the rest (Carr 1988). This is because it has been found that while the skill of prediction first improves as predictors are added to the equation, at some point the skill actually deteriorate with increasing numbers of predictors. This phenomenon is referred to as overfitting. An often used technique to avoid this problem is to use an F-test. New predictors are added as long as the variance explained by this predictor exceeds what one would expect from chance.

The best results when using MOS are obtained if due regard is payed to the following points:

- Individual equations are necessary for each: Station
  Variable
  Time of day [00, 06, 12, 18 UTC]
  Time of year [Winter , Summer]
  Forecast lead time
- For each equation one has to determine the type of predictors to be used as well as the number of predictors. There are no hard rules for how this should be done and it can be considered as an art. However, it can not be overemphasized that it is very important that it is being done judiciously, with a lot of common sense and sound meteorological thinking. Otherwise there is a big danger that one end up with nonsense results.
- There are usually between 8-20 predictors in each equation. This is well in line with numbers found in the mathematical statistical literature (see e g Hjort 1996).

3. OBSERVATIONAL DATA

The observed precipitation data is taken from the SMHI (digital) climate archive, BÅK. This archive contains rain gauge observations from roughly 1000 stations in Sweden for the time period from 1961 up to present time. The data is carefully quality controlled and constitutes the highest quality data available. The observations come from different types of stations. The highest quality stations are synoptic stations which measure manually and report precipitation in real time every 12 hours, 06 and 18 UTC. However, the most common station type is so called climate stations, which only measure once a day, at 06 UTC, and most often report in delayed mode. Since the 1990-ies many stations are un-manned and the rain gauge measuring is made automatic. The numbers of rain gauge stations and their location have undergone numerous changes during the time since 1961. The present number of stations is around 700.

In order to use as many stations as possible in the area of interest we limit ourselves by only considering the accumulated precipitation during the 24 hour period from 06 UTC to 06 UTC the following day. The accumulated precipitation amount during this 24 hour period is for bookkeeping purposes referred to the first of the two days, since 75% of the time interval of accumulation is occurring during the first day. Furthermore, we limit our consideration to the 10 year period 1980-1989. This implies a total of 3653 daily precipitation observations for each station. However, for many stations there are missing data. We choose 16 stations which had as little missing data as possible. Selected information regarding the 16 stations in the Torpshammar catchment area is given in Table 1. The same information as given in Table 1 is also given in graphical form in Figs. 1-4. The location of the stations is given in Fig. 1, their serial number is given in Fig. 2, the station type is given in Fig. 3 and the number of missing data is given in Fig. 4.

Serial Number	Station Name	BÅK Number	Station Type	Latitude	Longitude	Number of Missing Observations
1	DÖDRE	12448	3	627923	146870	517
2	KÖLSILLRE	12524	3	623986	152137	0
3	HUNGE	12545	1	627503	150984	31
4	SÖSJÖ	12546	3	627667	154996	31
5	NAGGEN	12616	3	622790	159986	0
6	ULVSJÖN	12617	3	622665	164483	30
7	HÖGSVEDJAN	12634	3	625535	166992	0
8	ÖRAÅTJÄRNARNA	12643	3	627165	161775	0
9	INDAL	12734	3	625684	171253	29
10	TANDSBYN	13400	3	630032	147508	0
11	FRÖSÖN	13411	1	631974	144863	0
12	RÖSTA	13415	3	632443	145535	0
13	RISSNA	13504	3	630608	153446	90
14	GRENINGEN	13520	3	633271	154054	0
15	KRÅNGEDE	13609	1	631610	161718	62
16	STENSJÖ	13616	3	632669	164446	30

Table 1. Information regarding the 16 precipitation stations selected in the Torpshammar catchment area. Station type 1 means a synoptic station (SYNOP) and type 3 means a climate station. Latitude and longitude are multiplied by 10000 and given in decimal form.

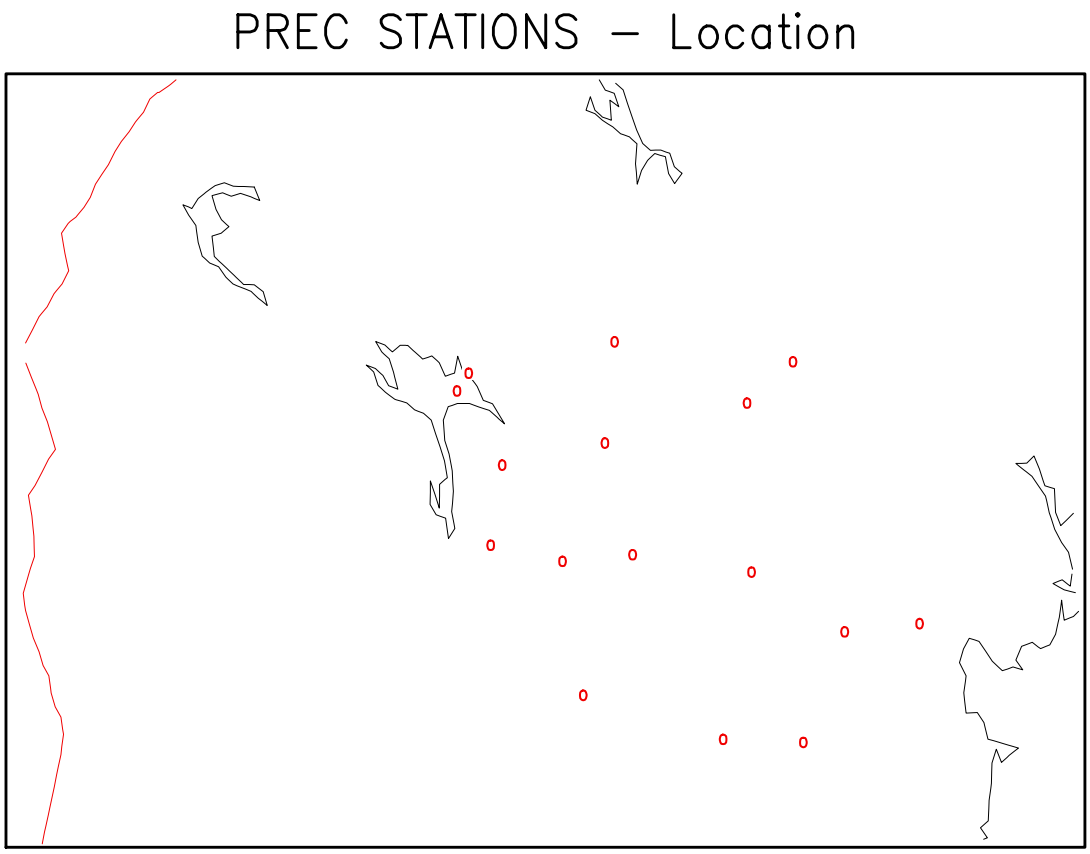

Fig. 1 Location of the 16 precipitation stations which are located within or close to the Torpshammar catchment area.

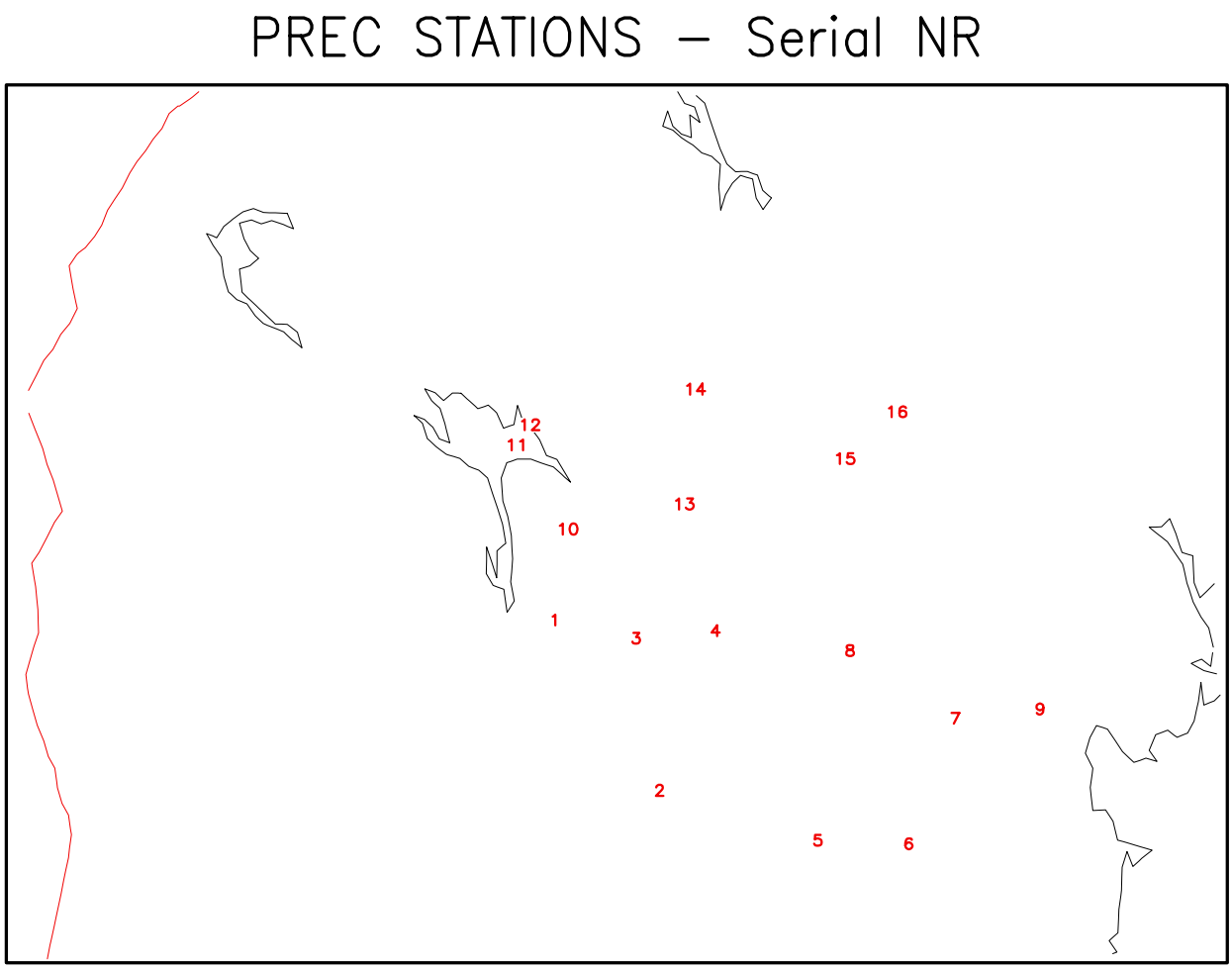

Fig. 2 Serial number of the 16 precipitation stations which are located within or close to the Torpshammar catchment area.

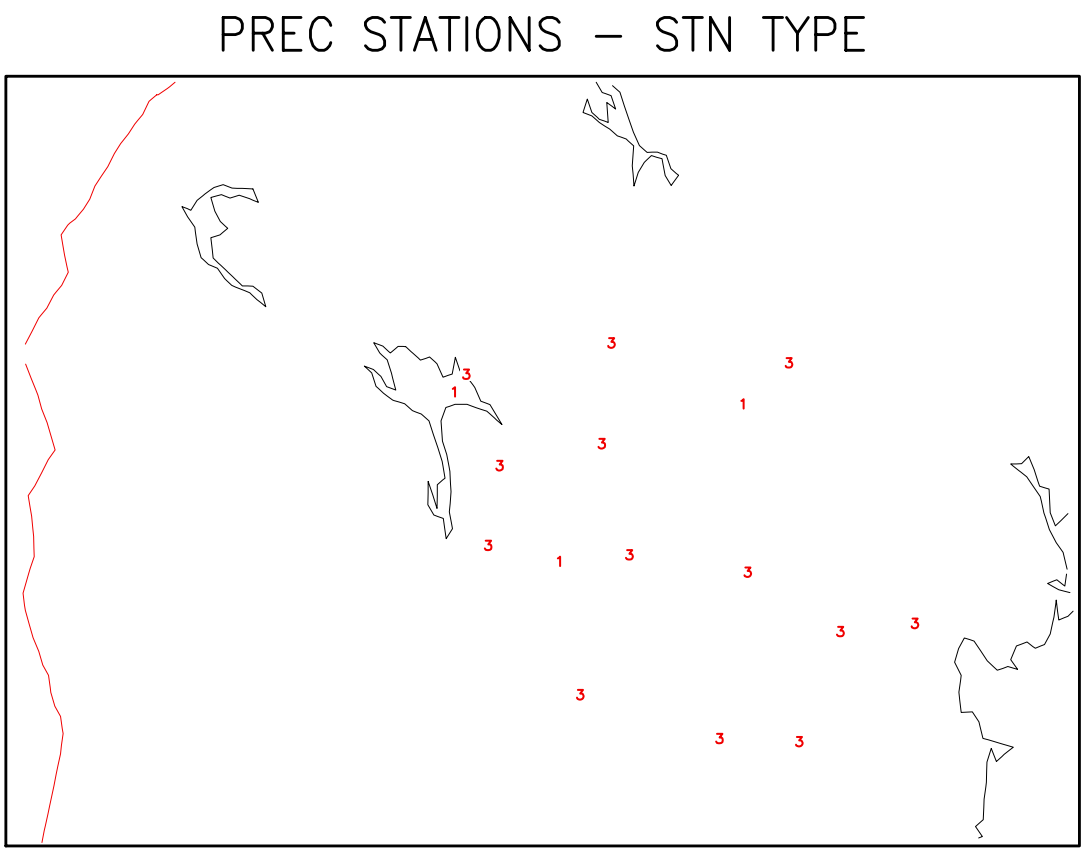

Fig. 3 The type of station for the 16 precipitation stations which are located within or close to the Torpshammar catchment area. Station type 1 means a synoptic station (SYNOP) and type 3 means a climate station.

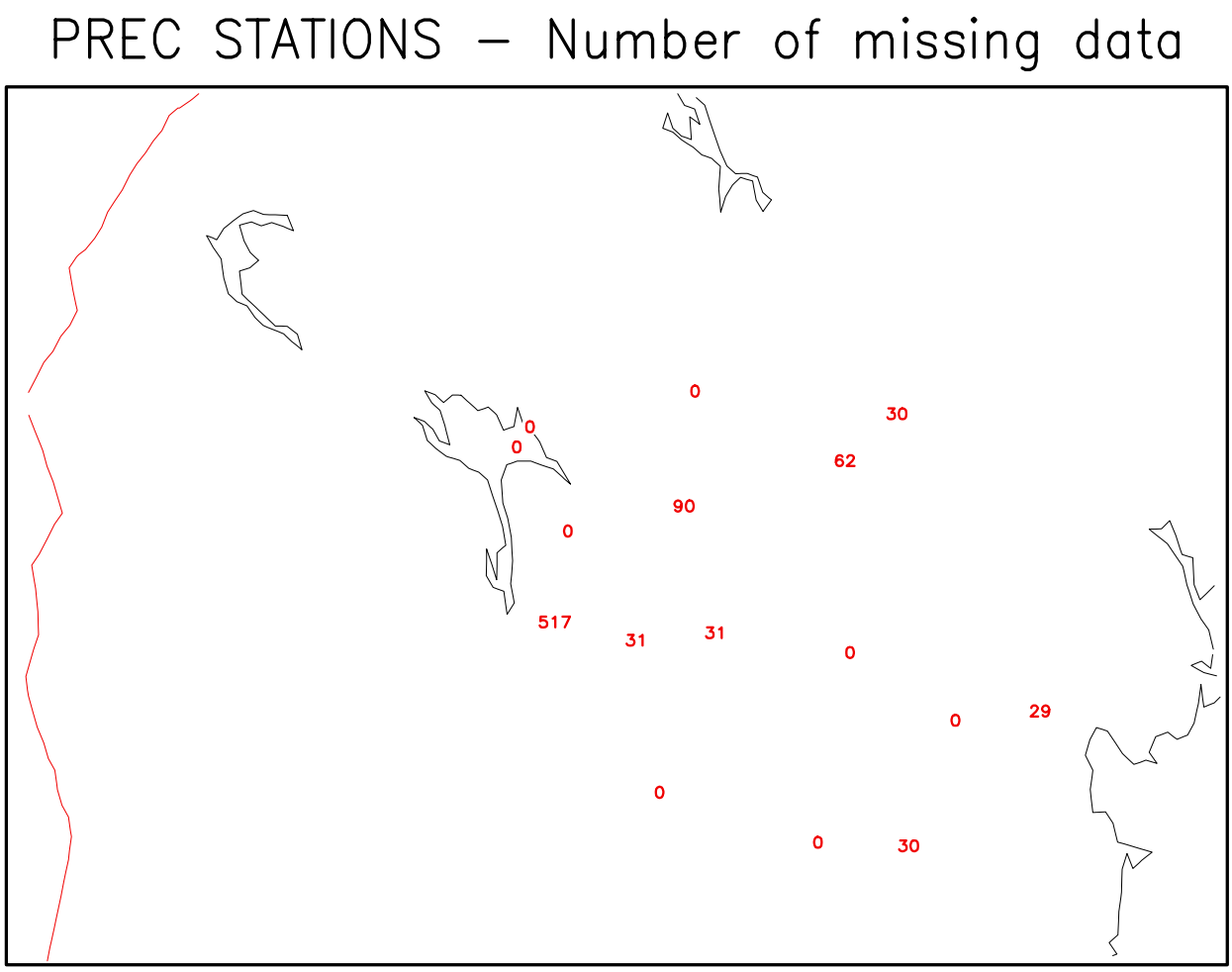

Fig. 4 Number of missing data for the 16 precipitation stations which are located within or close to the Torpshammar catchment area.

4. PREDICTAND

The predictand that will be used is in fact a set of binary precipitation amount predictands. The binary predictands represents the occurance (1) or non-occurance (0) of a set of events. Each event is defined to have occurred if the accumulation of precipitation, PREC, is equal to or surpasses a certain cutoff values, $PREC_n$ i e

$$\text{Event \#n} = \begin{cases} 0 \ , \ PREC < PREC_n \\ 1 \ , \ PREC \geq PREC_n \end{cases}$$

Based on the observed precipitation distribution in the Torpshammar catchment area we use the cutoff values given in Table 2.

Event Number n	$PREC_n$ **[mm]**
1	0.3
2	1
3	2
4	5
5	10
6	20

Table 2. Accumulated precipitation amount cutoff values used to determine the occurrence or non-occurrence of the events.

The predictands are treated cumulatively. This means that a single observation implies a simultaneous occurrence of one or more events. As an example we can consider the case of an observed daily precipitation amount of 8 mm which implies that event number 1-4 have occurred while event 5-6 have not occurred.

With the predictand defined in the above manner the linear regression used for forecasting produces a set of equations for the probability, P_n , of accumulated precipitation amount equal to or exceeding each of the above cutoff values, $PREC_n$, i e

$$P_n = \int_{PREC_n}^{\infty} p(r)\,dr$$

where p(r) is the probability density function of precipitation amount.

In order to promote consistency of the forecasts we develop the regression equations for the entire, concurrent set of predictands simultaneously. This implies that the same predictors are offered to all predictands, and if selected by any one predictand, is used for all predictands. This brand of regression that uses several binary predictands is sometimes called Regression Estimation of Event Probabilities (REEP).

5. FORECAST DATA

The forecast data that is used in this study is taken from the ECMWF 40-year ReAnalysis project (ERA-40). In addition to the core mission of the project; global reanalysis every 6th hour for the time period from September 1957 to August 2002, 36 hr global reforecasts were also carried out every 12th hour from the 00 and 12 UTC reanalyzed fields. The main reason for using this data set is that it constitutes a forecast data set that is as homogeneous as is possible to achieve. This is a quite unique situation. Normally one has to do with forecast data coming from a model that is used operationally. Such models are typically subject to frequent changes (~ 5-6 times a year) which are aimed to improve the realism and quality of the model. However, from the viewpoint of using historical forecast data for statistical postprocessing the inhomogeneities that the frequent model changes introduces can be quite detrimental.

The output from ERA-40 is very large. For the purpose of the present study we have retrieved a limited subset which is of immediate relevance. The size of this subset is of the order of 66 GB and is stored at the ECFS facility at ECMWF. A detailed description of this data set is given in Appendix A.

In Fig. 5 is shown how the model grid (in blue) is distributed in the area around the Torpshammar catchment area. Also shown is the location of the 16 precipitation stations (in red). It is thus clear that there is a substantial scale difference between what the numerical weather prediction model can resolve and the distance between the precipitation stations.

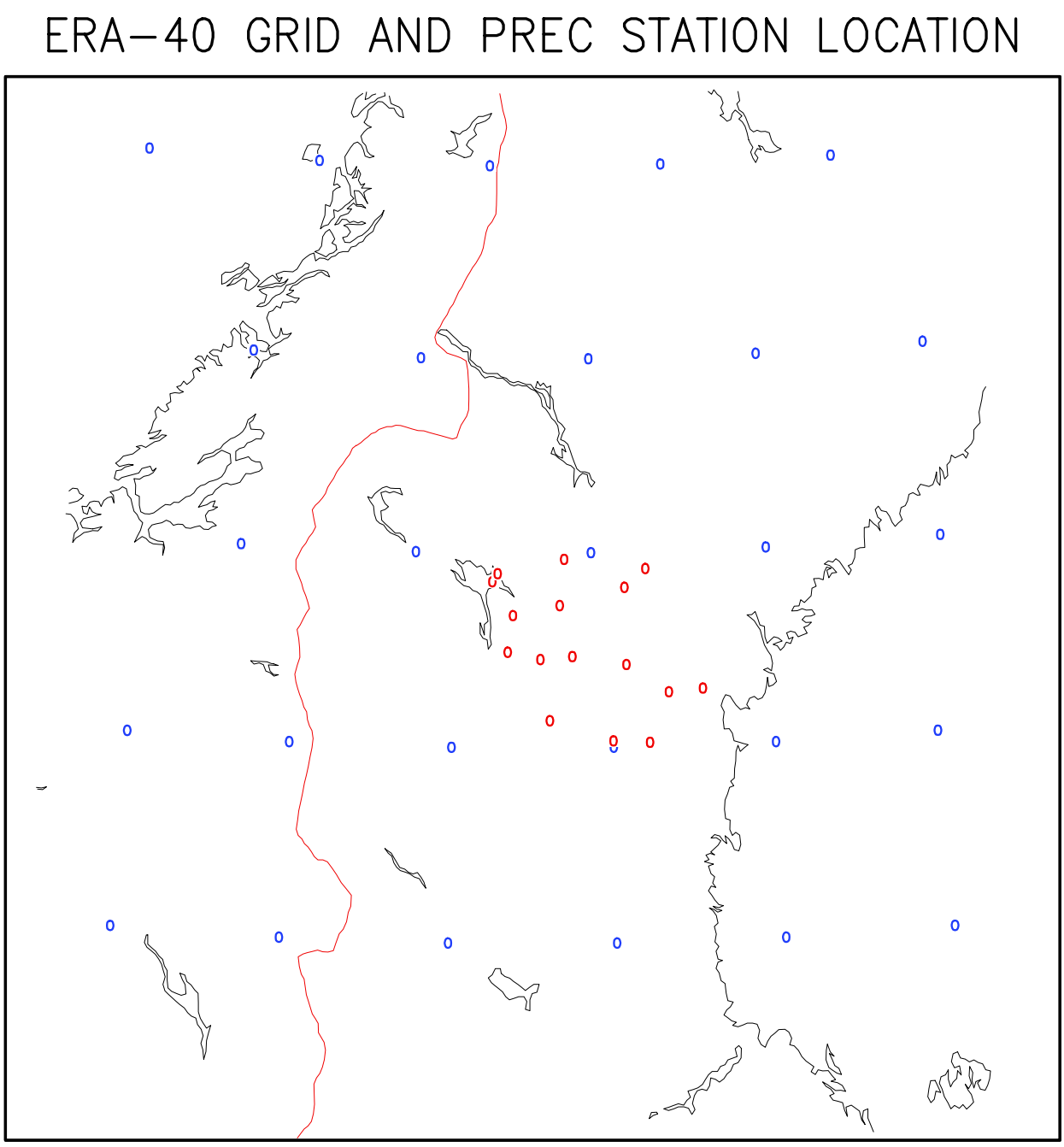

Fig. 5 The ERA-40 model grid which is a reduced Gaussian grid is shown as blue circles for the area around the Torpshammar catchment area. Also shown is the location of the 16 precipitation stations (in red circles).

6. PREDICTORS

All predictor variables are interpolated to the horizontal location of the observational site (station) where the regression equations are applied. The predictor variables are thus stacked in a column above the station. All upstream influences are therefore only taken into account in an indirect way.

The type of predictor variables is of three different kinds:

1. Raw variables. Variables taken directly from the subset of ERA-40 that is described in Appendix A.

2. Derived variables from the raw variables. These variables are defined on the basis of having physically plausible connection to the predictand variable in a meteorologically sound manner. These variables are also a convenient way to introduce nonlinearities into the linear regression procedure. A classical example in the case of predicting precipitation is to form the nonlinear predictor variable $\omega \cdot RH$, i e the product of vertical velocity and relative humidity.

3. Grid Binary (GB) variables. A transformed variable is first calculated at each grid point. Its value is either 0 or 1 dependent on whether the grid point value exceeds a cutoff value (1) or not (0). The grid binary predictor is then defined to be the interpolated value of this transformed field to the station location. The value will thus be in the interval $0 < GB < 1$.

Experiences from e g TDL have shown that if both continuous and grid binary versions of the same meteorological variable are offered to the linear regression procedure the grid binaries are almost always chosen first. The reason for this is partly the fact that grid binaries are able to capture the non-linearities in the relation between predictor and predictand. Another aspect that contributes is the fact that many meteorological variables does not vary smoothly with space but instead changes almost discontinuously in association with e g fronts. A small error in the NWP forecast of e g the exact location of a front will result in quite large errors of an interpolated value of a continuous variable. In contrast, the grid binary variables seem to smooth out such NWP errors. A typical example of a good grid binary variable in association with predicting precipitation is the mean relative humidity in the lower troposphere with a cutoff value of 70%.

7. INTERPOLATION

We will use polynomials to approximate functions along longitudes and latitudes, respectively. These polynomials will in turn be used for interpolation in first the longitudinal and then finally in the latitudinal direction. The interpolation problem of determining an m^{th}-degree polynomial which agrees with the values of a given function on a net of (m+1) points always has a unique solution, which can be expressed in the form

$$p(x) = c_o + c_1(x - x_o) + c_2(x - x_o)(x - x_1) + \ . \ . \ . \qquad (1)$$

The coefficients c_o , c_1 , c_2 . . . in (1) can be uniquely determined in such a way that p(x) takes on arbitrary predetermined values on the net $\{x_i\}_{i=o}^{m}$.

$$p(x_o) = c_o$$

$$p(x_1) = c_o + c_1(x_1 - x_o)$$

$$p(x_2) = c_o + c_1(x_2 - x_o) + c_2(x_2 - x_o)(x_2 - x_1)$$

$$p(x_3) = c_o + c_1(x_3 - x_o) + c_2(x_3 - x_o)(x_3 - x_1) + c_3(x_3 - x_o)(x_3 - x_1)(x_3 - x_2)$$

one can compute c_o , c_1 , c_2 . . . recursively from the above triangular system of equations. Thus the interpolation problem has exactly one solution.

The grid points involved in the interpolation are in the immediate surroundings of the Torpshammar catchment area and are displayed in Fig. 6. There are 5 longitude points along each of the 4 Gaussian latitudes. The longitude numbers counted from Greenwich are marked at the place of the grid points. The first step is to interpolate in the longitudinal direction. This involves calculating 5 coefficents at each Gaussian latitude and then using them to calculate the interpolated value at the longitude of the station in question. The 4 interpolated values thus obtained is then used to calculate 4 coefficients which is used to interpolate to the latitude of the station in question. An example of how this interpolation procedure works is displayed in Fig. 7.

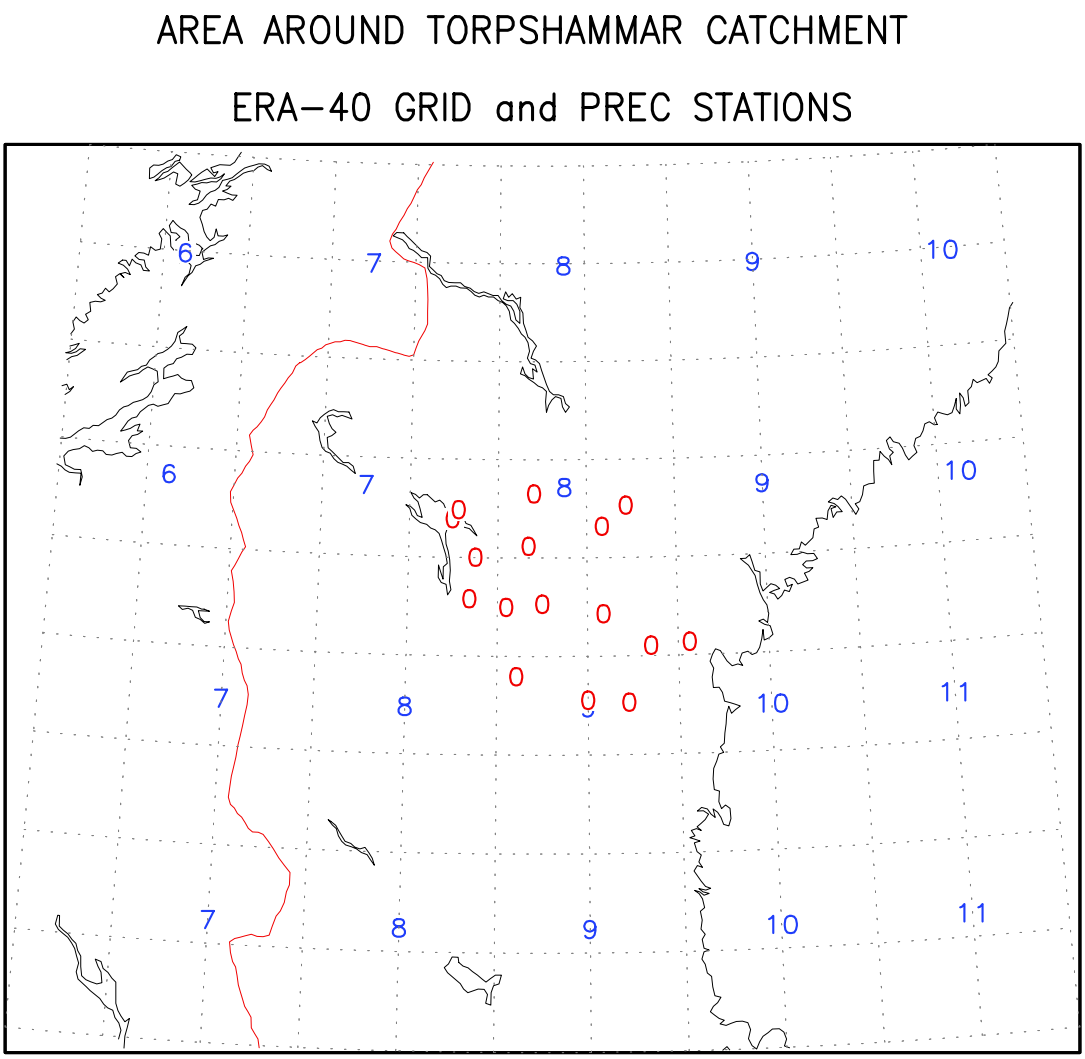

Fig 6. The grid points used for the interpolation to stations in the Torpshammar catchment area. The longitude numbers counted from Greenwich are marked with blue color and the station locations with red circles.

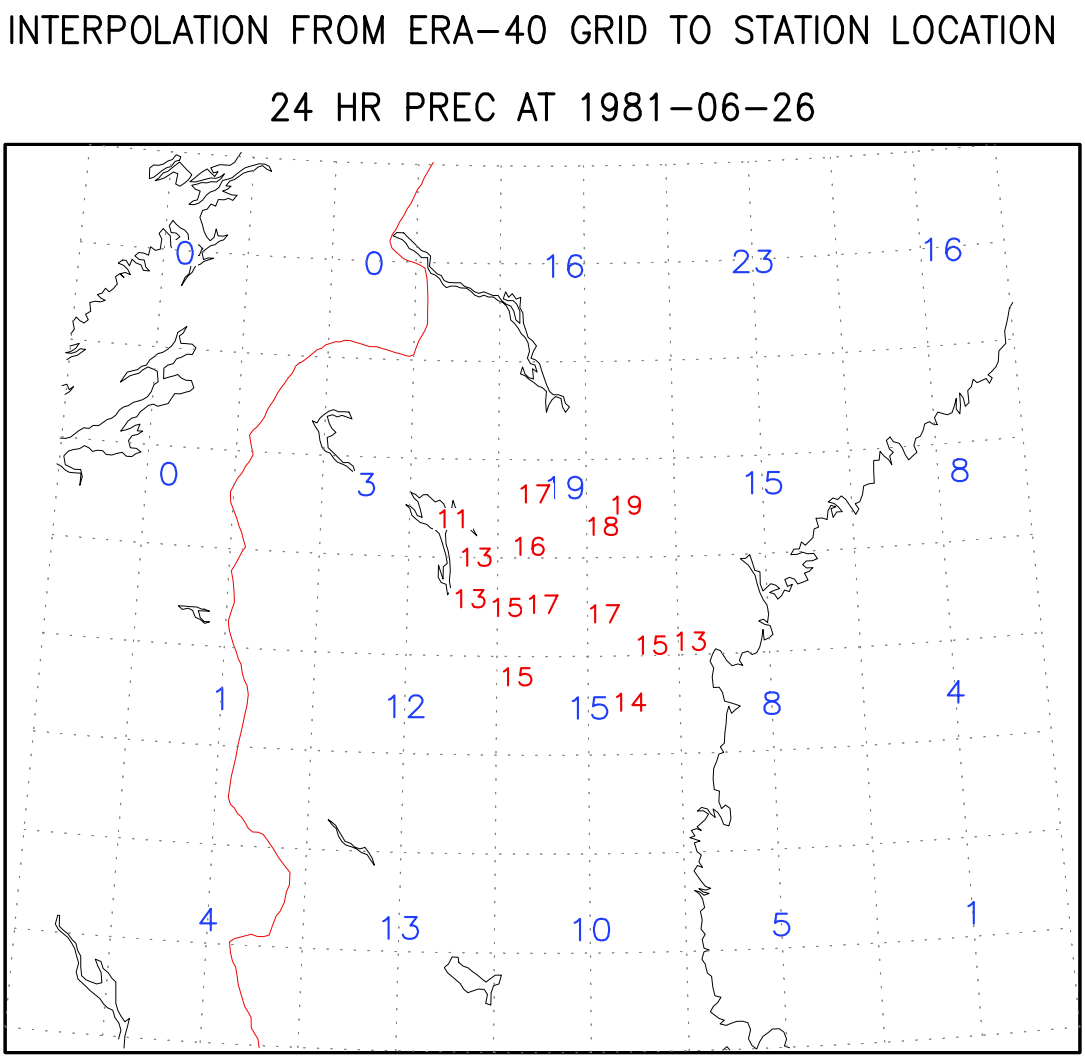

Fig. 7. An example of how the interpolation procedure works. The grid point data is the 24 hour accumulated precipitation at 26 June 1981.

8. CATEGORICAL FORECASTS

The result of the linear regression technique used, REEP, produces a set of probability forecasts of the accumulated precipitation amount exceeding certain threshold values. It is however also possible and desirable to convert the probability forecasts to categorical forecasts. The categories used in this study are given in Table 3 where one can see that the bounds are identical to the cutoff values given in Table 2.

QPF Category	Precipitation amount [mm]
0	0 – 0.3 (Trace)
1	0.3 – 1
2	1 – 2
3	2 – 5
4	5 – 10
5	10 – 20
6	20 -

Table 3. Description of the QPF categories

The conversion to categories is accomplished by using an algorithm which compares the forecast probabilities to predetermined threshold values. One threshold value per predictand is determined from the dependent data set and it is designed in such a way as to yield categorical forecasts which satisfy the following conditions:

1. The forecast skill as measured by the Critical Success Index, CSI (also known as the Threat Score) should be maximized

2. The Bias should be close to 1

3. The False Alarm Rate should not be too high

The scores and indices introduced above are closely related to the contingency table and an explanation is found in Appendix B.

When the threshold values have been determined on the dependent data set they can be used to obtain categorical forecasts. A single categorical forecast is determined by comparing each forecast cumulative probability to its threshold value. The comparison begins with the rarest event, i e largest precipitation amount, and then proceeds consecutively to the most common, i e smallest precipitation amount. The first time the forecast probability exceeds the threshold value determines the category. A graphical illustration of the conversion procedure is shown in Fig. 8.

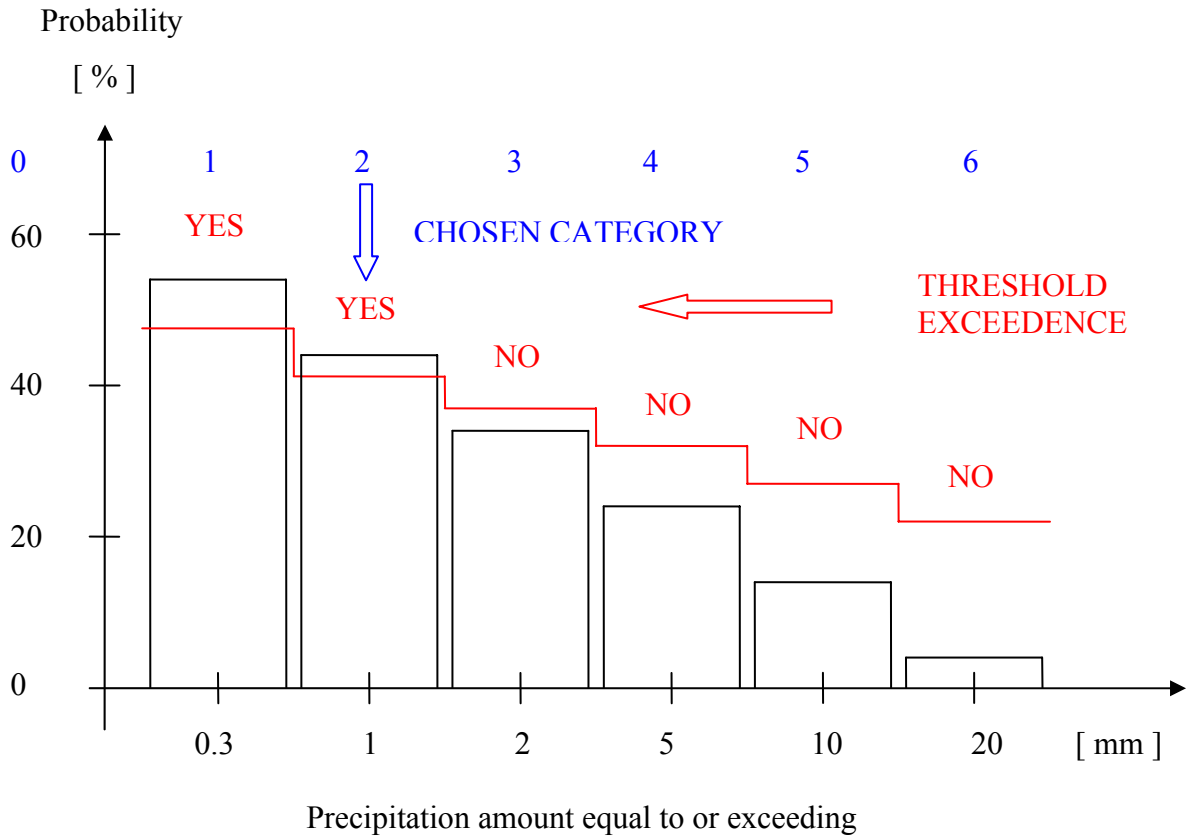

Fig. 8. Graphical illustration of how the conversion from QPF probabilities to categorical forecasts is performed. The red solid line indicates the threshold value for each predictand. The category number is written in blue above the columns. The category chosen is the first column that rises above the threshold line when starting from the highest amount going towards lower amounts (from right to left in the figure). If none is encountered then category 0 is chosen.

9. PRECIPITATION CHARACTERISTICS

Precipitation is a variable with rather specific characteristics. It is important to establish these characteristics for both the observational and forecast data. To this end we construct empirical relative frequency histograms. The construction of bins for the histograms is as follows:

a. Observed data from BÅK

Observed data from BÅK are given with an accuracy of 0.1 mm. Precipitation amount 0.01 mm is given when the rain gauge is wet but no amount is possible to measure, so called "trace" amount. We pool the observed precipitation amounts, PREC, into NCAT+3 categories according to Table 4. The value 0.01 mm is classified as belonging to the no precipitation category.

INDEX OF CATEGORY	**NAME OF CATEGORY**	**RANGE OF PRECIPITATION AMOUNT [mm]**
-1	NO PREC	$0 \leq \text{PREC} \leq 0.01$
0	SMALL PREC	$0.1 \leq \text{PREC} \leq 0.4$
1	1 mm	$0.5 \leq \text{PREC} \leq 1.4$
2	2 mm	$1.5 \leq \text{PREC} \leq 2.4$
3	3 mm	$2.5 \leq \text{PREC} \leq 3.4$
NCAT	ncat mm	$\text{NCAT}-0.5 \leq \text{PREC} \leq \text{NCAT}+0.4$
NCAT+1	LARGE PREC	$\text{NCAT}+0.5 \leq \text{PREC} < \infty$

Table 4. Division of precipitation amount, PREC, into NCAT+3 categories for observed data from BÅK.

b. Theoretical distributions

In order for a theoretical distribution to be commensurate with the categorization for observed data we divide the continuous distribution of precipitation amounts, PREC, into NCAT+3 categories according to Table 5.

INDEX OF CATEGORY	NAME OF CATEGORY	RANGE OF PRECIPITATION AMOUNT [mm]
-1	NO PREC	PREC = 0
0	SMALL PREC	$0 < PREC \leq 0.45$
1	1 mm	$0.45 < PREC \leq 1.45$
2	2 mm	$1.45 < PREC \leq 2.45$
3	3 mm	$2.45 < PREC \leq 3.45$
NCAT	ncat mm	$NCAT-0.55 < PREC \leq NCAT+0.45$
NCAT+1	LARGE PREC	$NCAT+0.45 < PREC < \infty$

Table 5. Division of precipitation amount, PREC, into NCAT+3 categories for theoretical distributions.

c. Model data from ERA-40

For model data from ERA-40 we divide the continuous distribution of precipitation amounts, PREC, into NCAT+3 categories in the same way as for the theoretical distributions according to Table 5.

In the following we will consider both observational and ERA-40 data in the time period 1980-1989. We study the 24 hour accumulated precipitation which implies data at a maximum of 3653 time points (some stations have less as can be seen in Table 1). We therefore choose NCAT=20.

Frequency of wet days

In Fig. 9 is shown the frequency of wet days at the 16 stations described in Table 1. The blue curve is for the observations and the red curve is for the ERA-40 data interpolated to the station location. As can be seen the model has considerably more wet days than the observations, a characteristic which is rather common for NWP models.

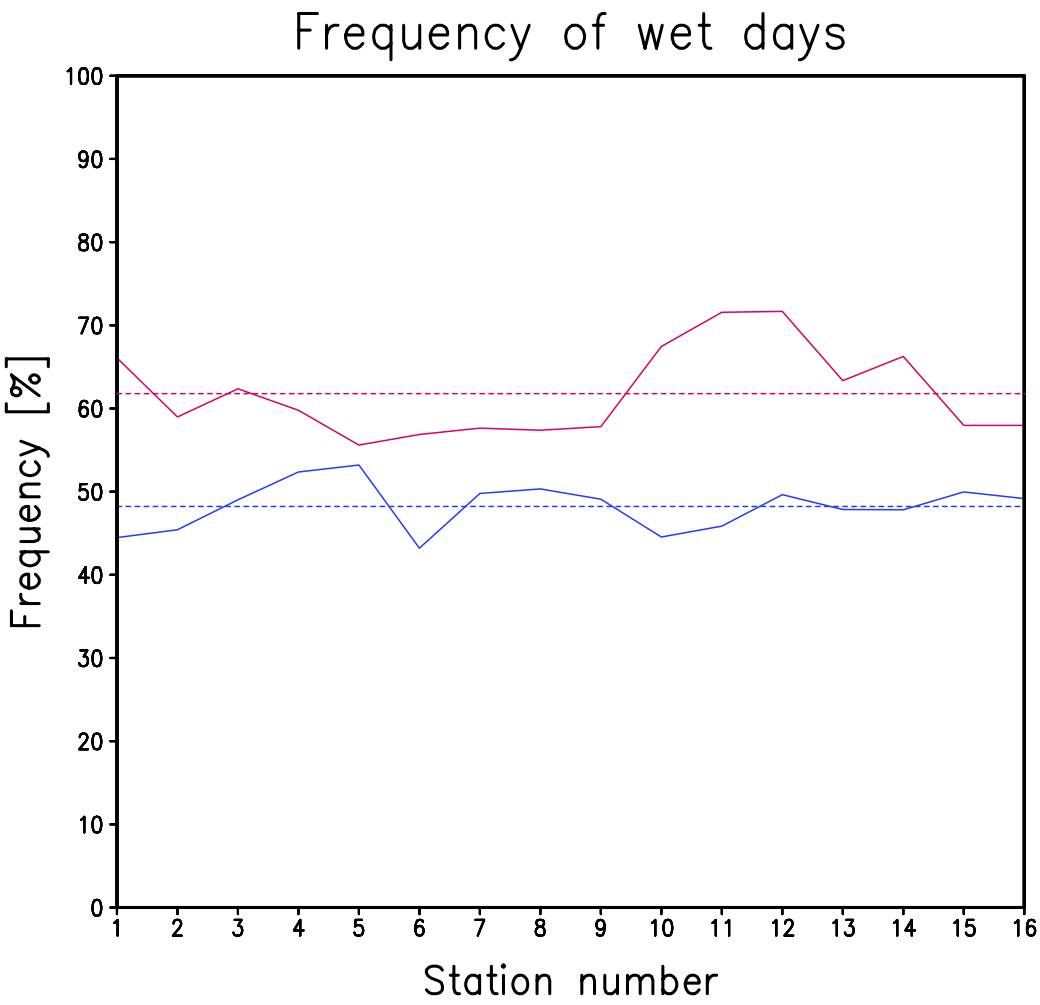

Fig. 9. The frequency of wet days at the 16 stations described in Table 1. The blue curve is for the observations and the red curve is for the ERA-40 data interpolated to the station location. The dashed curves are average values for all stations.

Mean daily precipitation

The mean daily precipitation is shown in Fig. 10 for all days. The model is seen to be quite accurate when considering the average value over all stations. However, for some stations there is a considerable overestimation while for others there is an underestimation. The variability between the stations is thus quite large within the small catchment area.

When only considering wet days, Fig. 11, the model clearly underestimates the precipitation, on the average by 20%. This is commensurate with the previous two figures since the model has more wet days but still roughly the correct total precipitation. Note that the model underestimates the station to station variability when it rains within the catchment area compared to the observations.

Variability of daily precipitation

The variability of daily precipitation amount as measured by the standard deviation is shown in Fig. 12-13 for all days and wet days, respectively. The model is underestimating the variability at almost all stations, and this underestimation is more pronounced when only considering wet days. The station to station variability is larger in the real world compared to the model world.

Maximum daily precipitation

In Fig. 14 we display how the maximum 24 hour precipitation amount recorded during the 10 years 1980-1989 vary from station to station. The ERA-40 data not only underestimate the maximum amount but also the station to station variability.

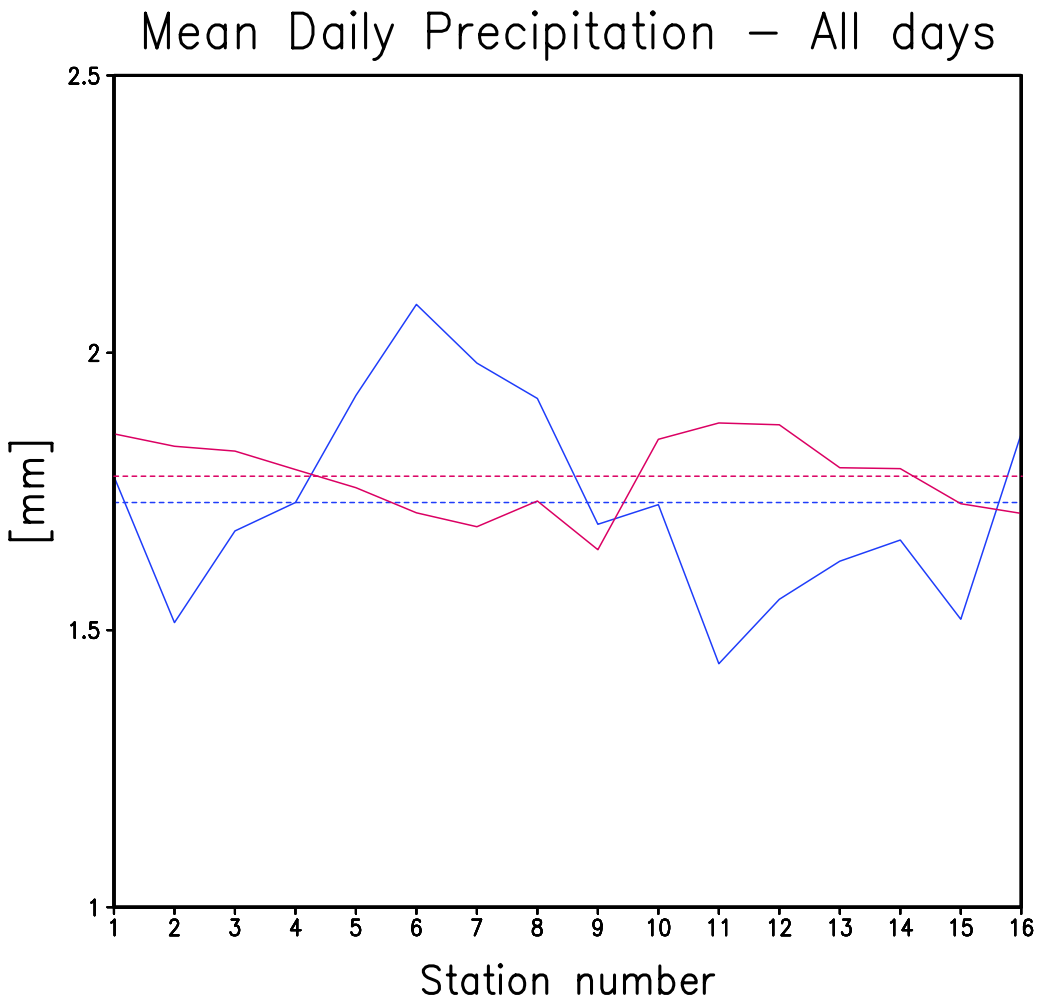

Fig. 10. The mean daily precipitation at the 16 stations described in Table 1. The blue curve is for the observations and the red curve is for the ERA-40 data interpolated to the station location. The dashed curves are average values for all stations

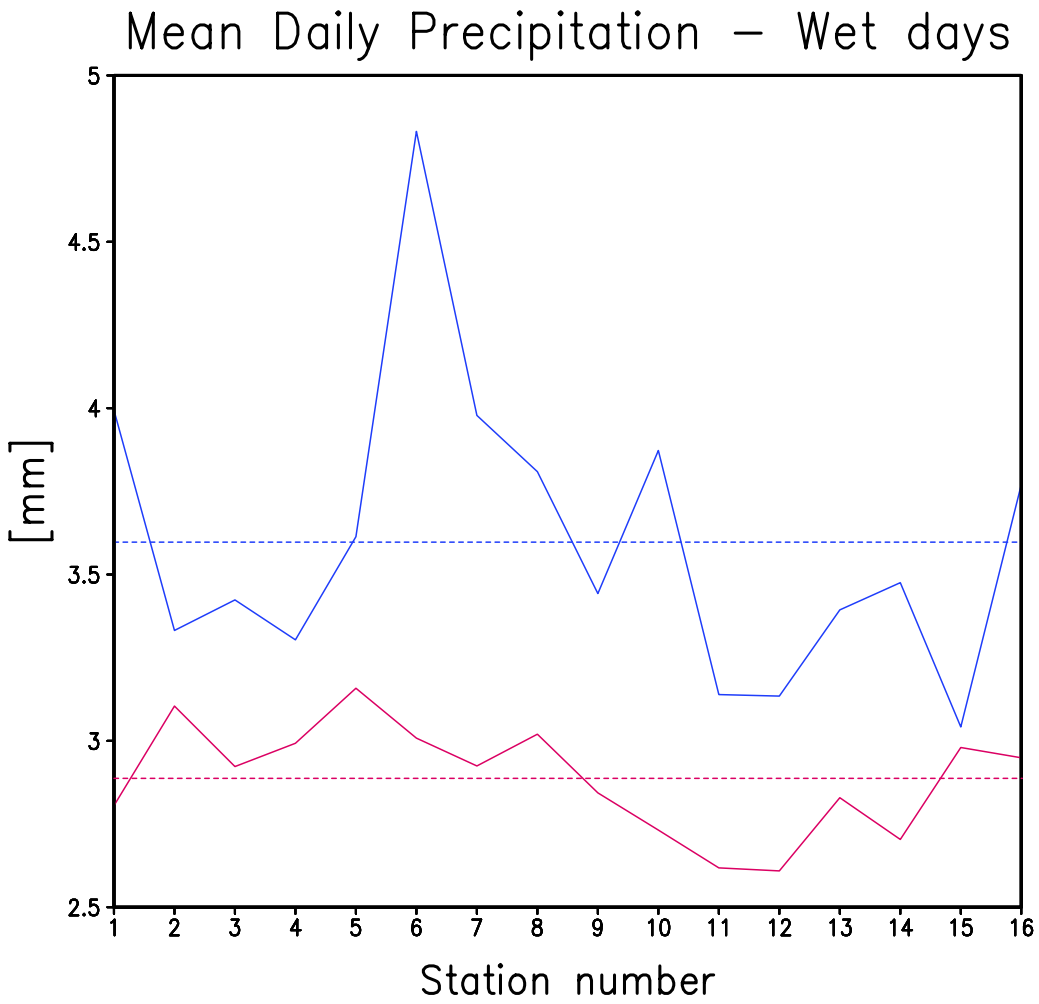

Fig. 11. The mean daily precipitation on wet days at the 16 stations described in Table 1. The blue curve is for the observations and the red curve is for the ERA-40 data interpolated to the station location. The dashed curves are average values for all stations

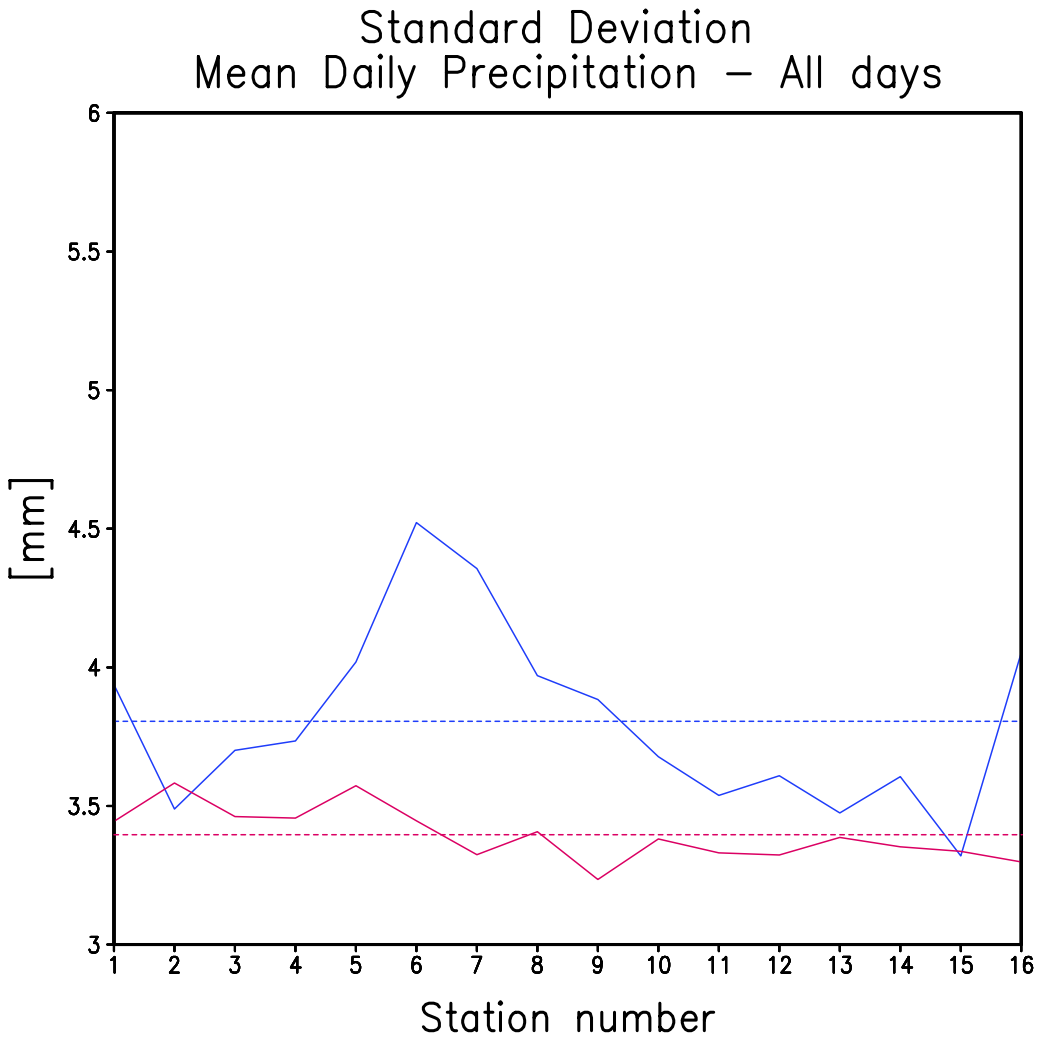

Fig. 12. The standard deviation of daily precipitation at the 16 stations described in Table 1. The blue curve is for the observations and the red curve is for the ERA-40 data interpolated to the station location. The dashed curves are average values for all stations

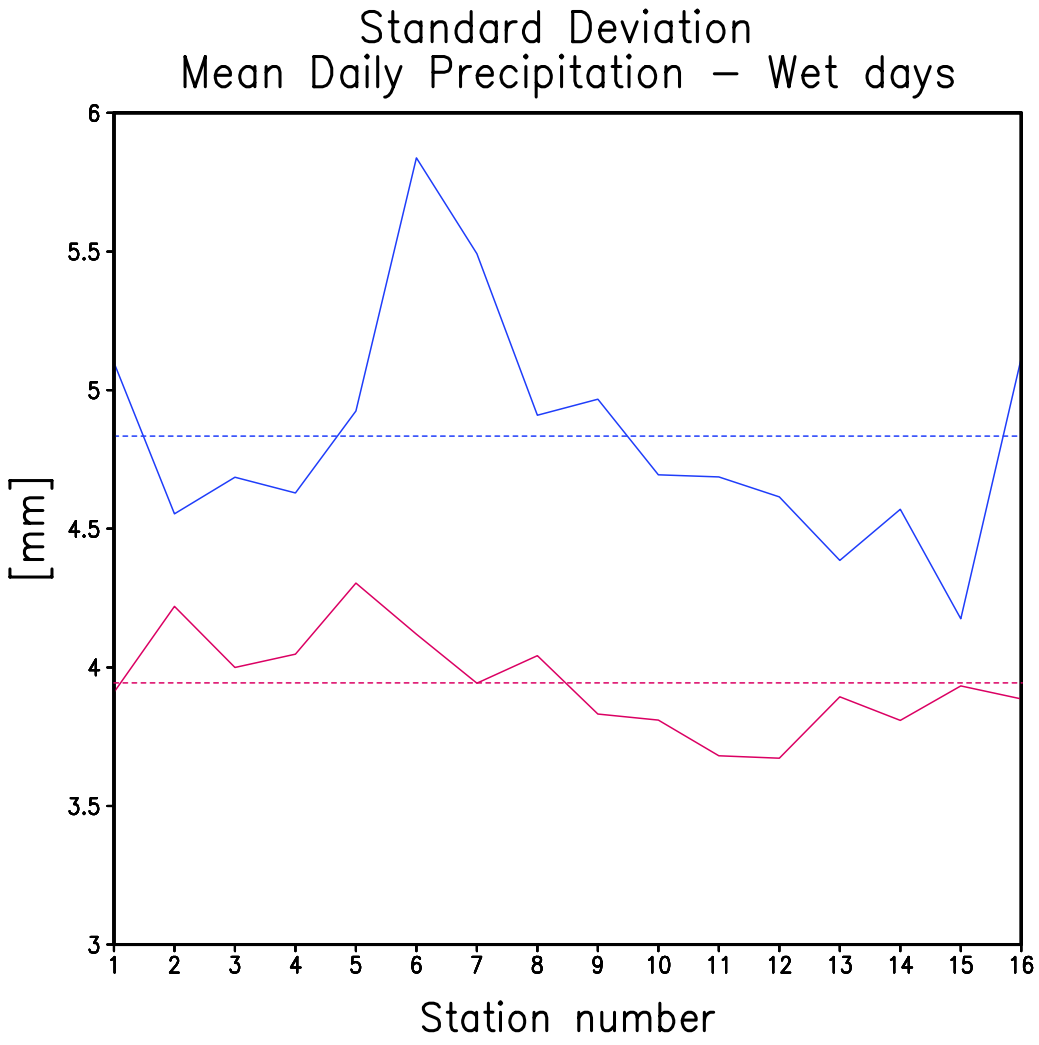

Fig. 13. The standard deviation of daily precipitation on wet days at the 16 stations described in Table 1. The blue curve is for the observations and the red curve is for the ERA-40 data interpolated to the station location. The dashed curves are average values for all stations

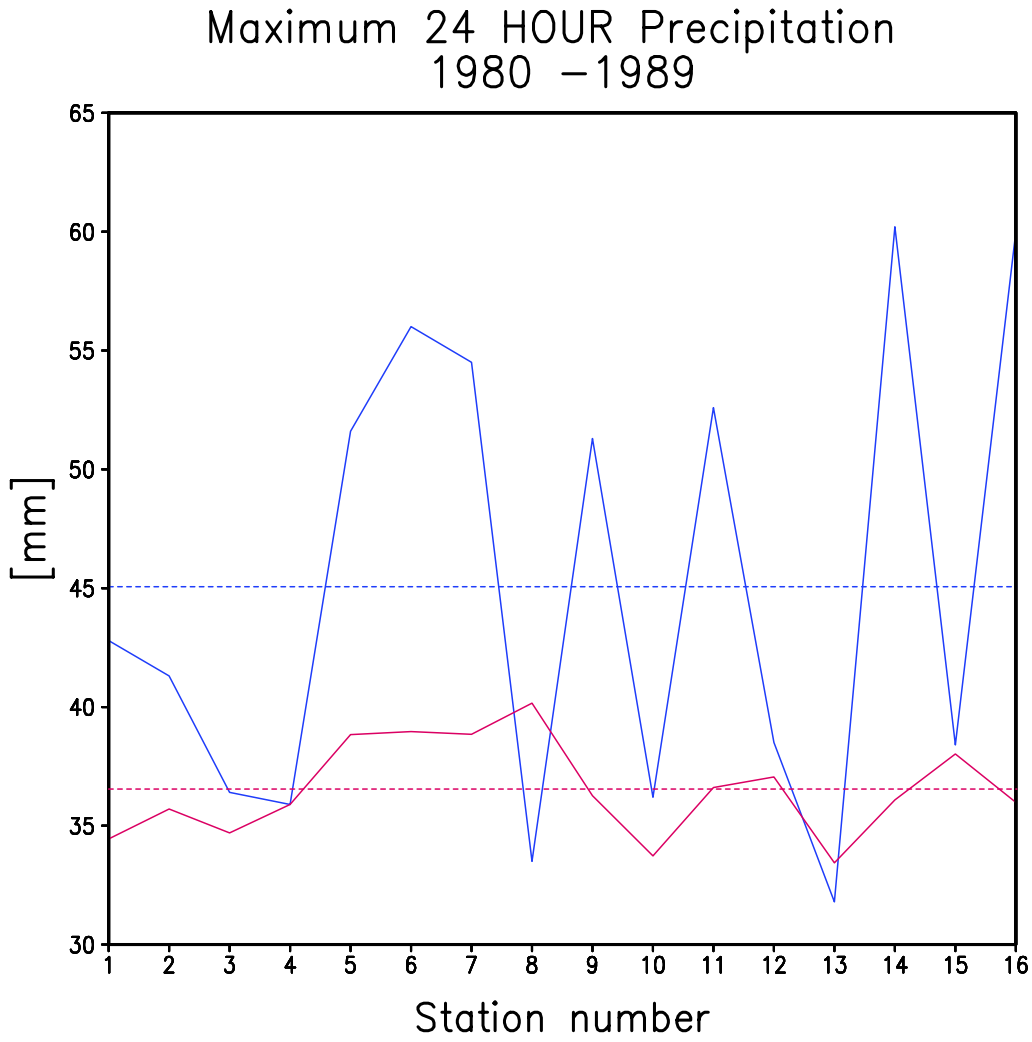

Fig. 14. The maximum 24 hour precipitation amount recorded during the 10 years 1980-1989 at the 16 stations described in Table 1. The blue curve is for the observations and the red curve is for the ERA-40 data interpolated to the station location. The dashed curves are average values for all stations

Probability density function

An important characteristic of accumulated precipitation amount is its probability density function, pdf. This pdf is known to be quite skewed and many authors, e g Krzysztofowicz and Sigrest (1997), have tried to compare the observed distributions with known theoretical distributions, such as the exponential, gamma and Weibull distributions. If such a theoretical distribution fits the data well, then the scale and shape parameters can be used to deduce many statistical properties of the distribution.

In Fig. 15 is shown the observed relative frequency of 24 hr accumulated precipitation from the station Frösön (blue solid curve). The distribution is seen to be very skewed, which is quite typical. More then half of the wet days have precipitation amounts less then 2 mm, and only 7 % of the wet days have more then 10 mm.

A fit to the gamma and exponential distributions are also displayed as solid and dashed red curves, respectively. The method of moments estimation of the shape and scale parameters are used to make the fit to the two theoretical distributions. For the exponential distribution this is identical to the maximum likelihood estimation. The gamma distribution clearly fits the observed data better then the exponential.

The corresponding distribution for ERA-40 model data interpolated to the location of the station Frösön is shown in Fig. 16 as the blue solid curve and the fit to the gamma and exponential distributions are displayed as solid and dashed red curves, respectively. Also in this case the gamma distribution fits the observed data much better then the exponential distribution and the fit is furthermore better compared to the observed data.

The models ability to capture the observed distribution is quite good as seen in Fig. 17. However, the relative frequencies are with respect to wet days, which we know differs between observations and ERA-40 as seen in Fig. 9. Therefore, if we instead consider relative frequencies with respect to all days we get Fig. 18 which shows that ERA-40 has substantially more days with small precipitation amounts.

Another indication that the station to station variability is underestimated in ERA-40 can be deduced from Figs. 19-20. They show how the shape (α) and scale (λ) parameters of the gamma distribution (see Appendix C) vary from station to station.

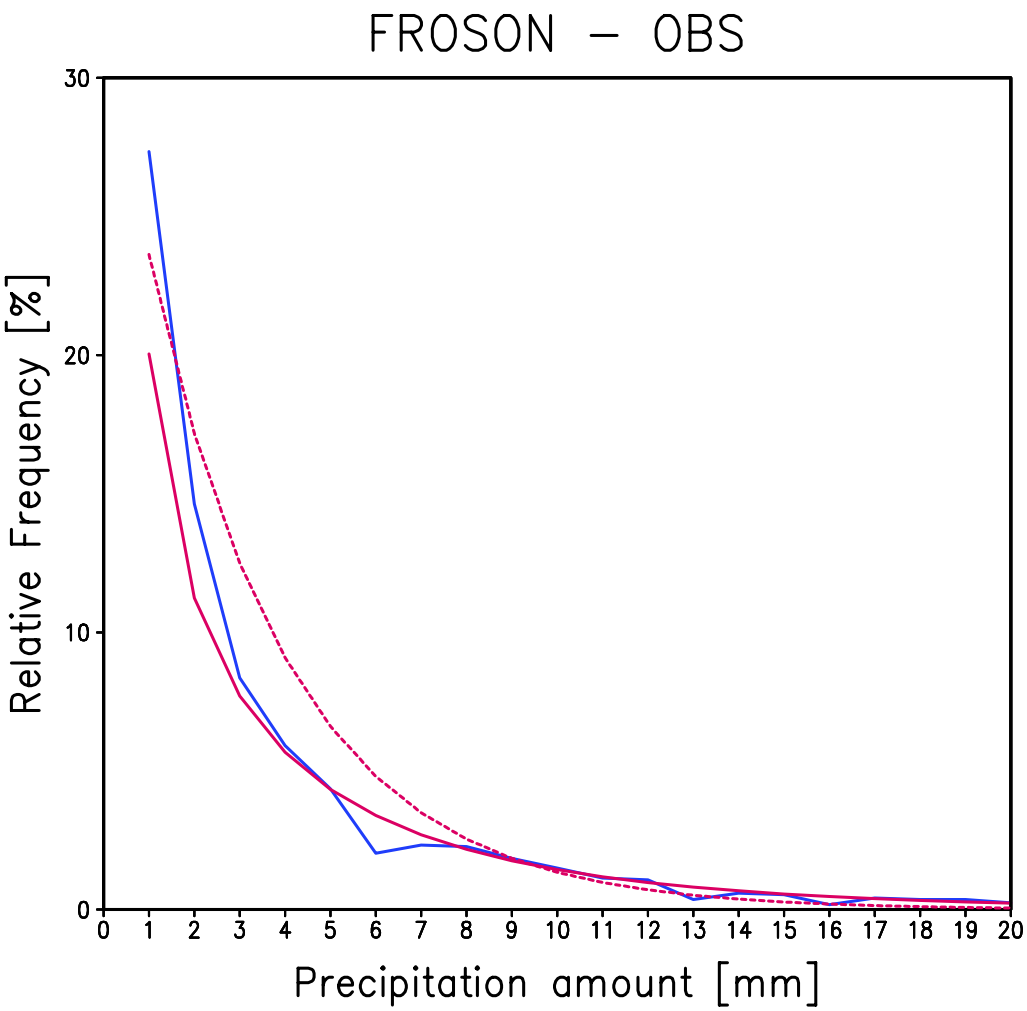

Fig. 15. The observed relative frequency of 24 hr accumulated precipitation at the station Frösön is shown in blue solid curve. A fit to the gamma and exponential distributions are also displayed as solid and dashed red curves, respectively.

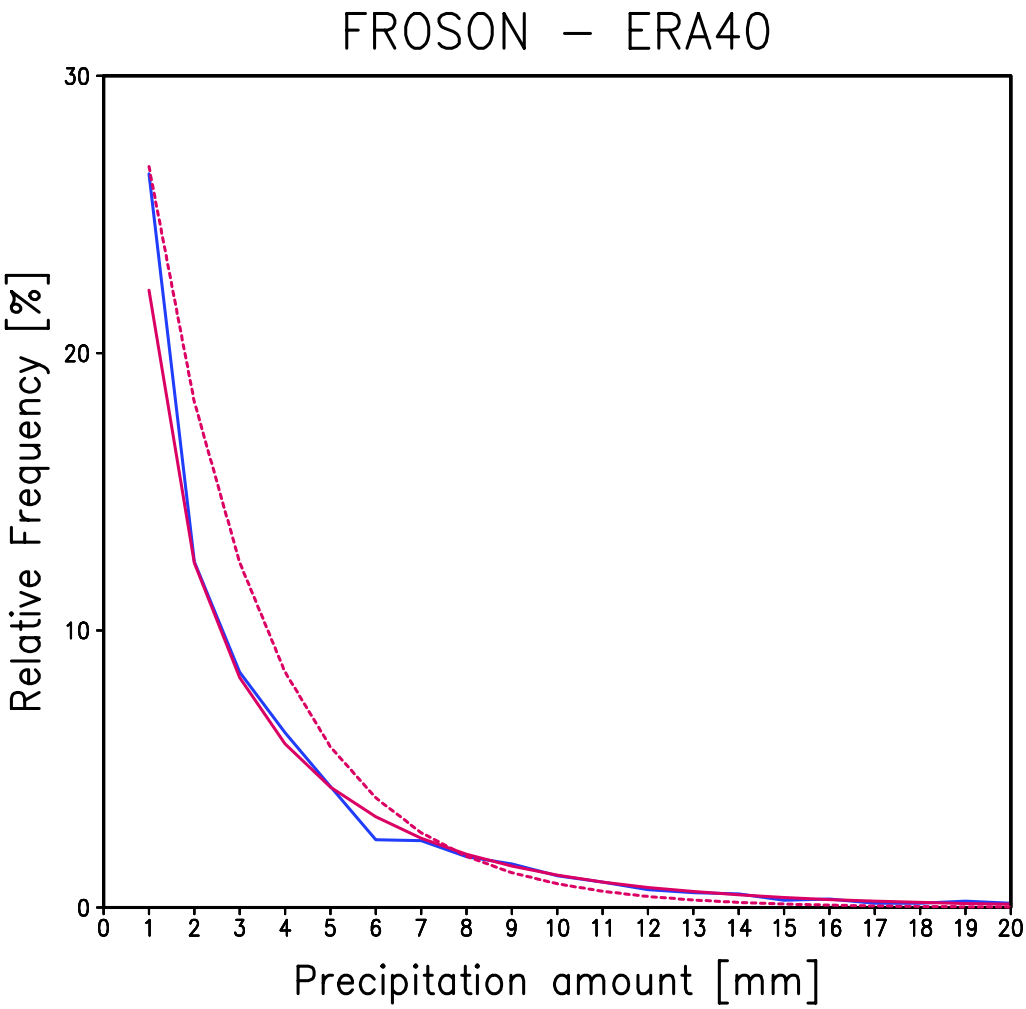

Fig. 16. The relative frequency of 24 hr accumulated precipitation as deduced from ERA-40 model data and interpolated to the location of the station Frösön is shown in blue solid curve. A fit to the gamma and exponential distributions are also displayed as solid and dashed red curves, respectively.

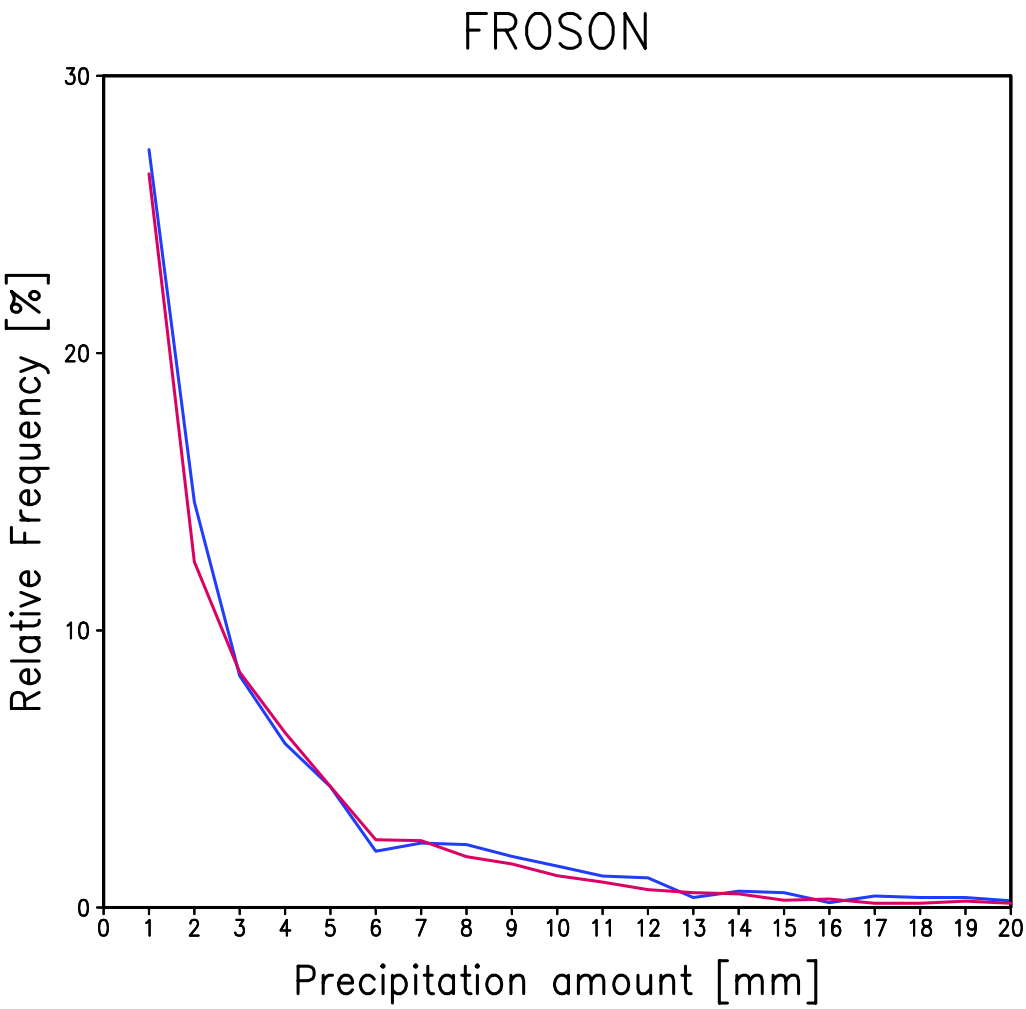

Fig. 17. The observed relative frequency of 24 hr accumulated precipitation at the station Frösön is displayed as a blue solid curve. The corresponding relative frequency as deduced from ERA-40 model data and interpolated to the location of the station Frösön is displayed as a red solid curve.

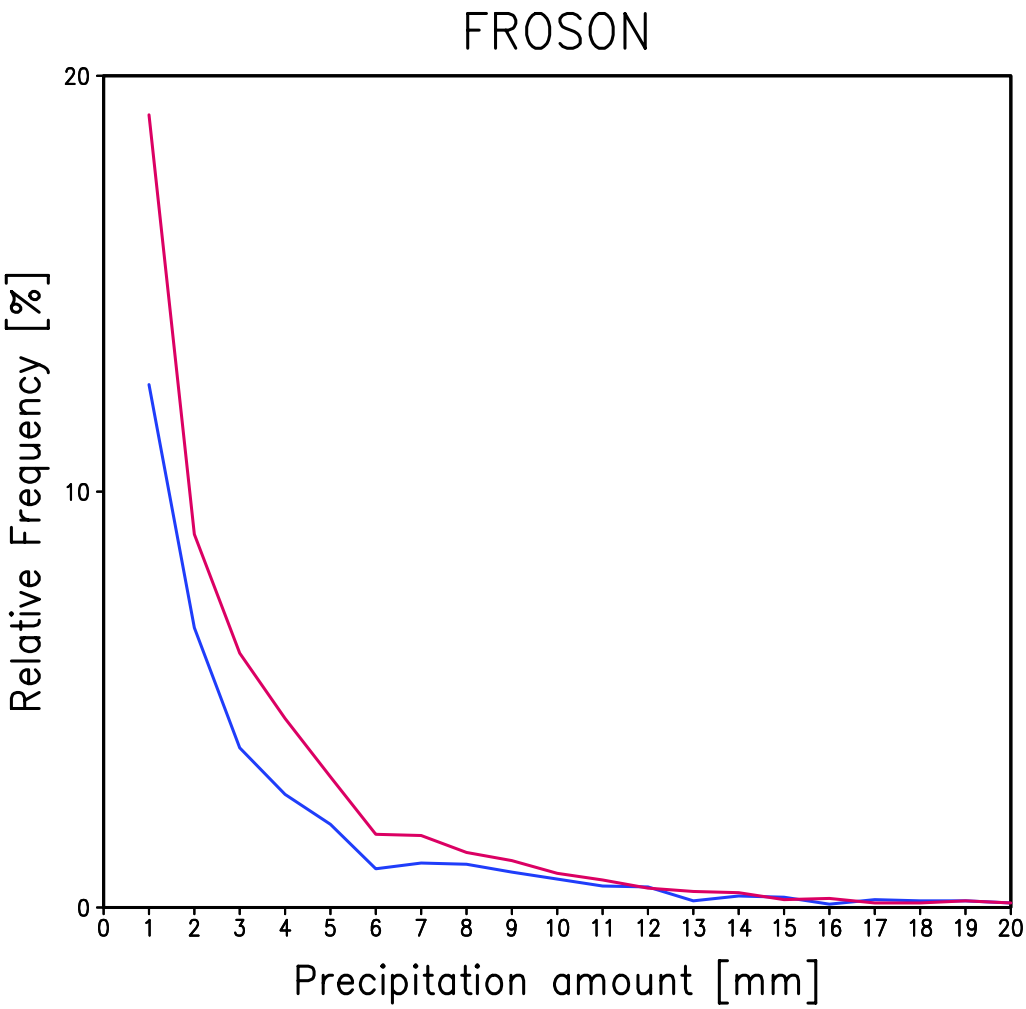

Fig. 18. Same as Fig. 16 except that the relative frequencies are with respect to all days instead of wet days.

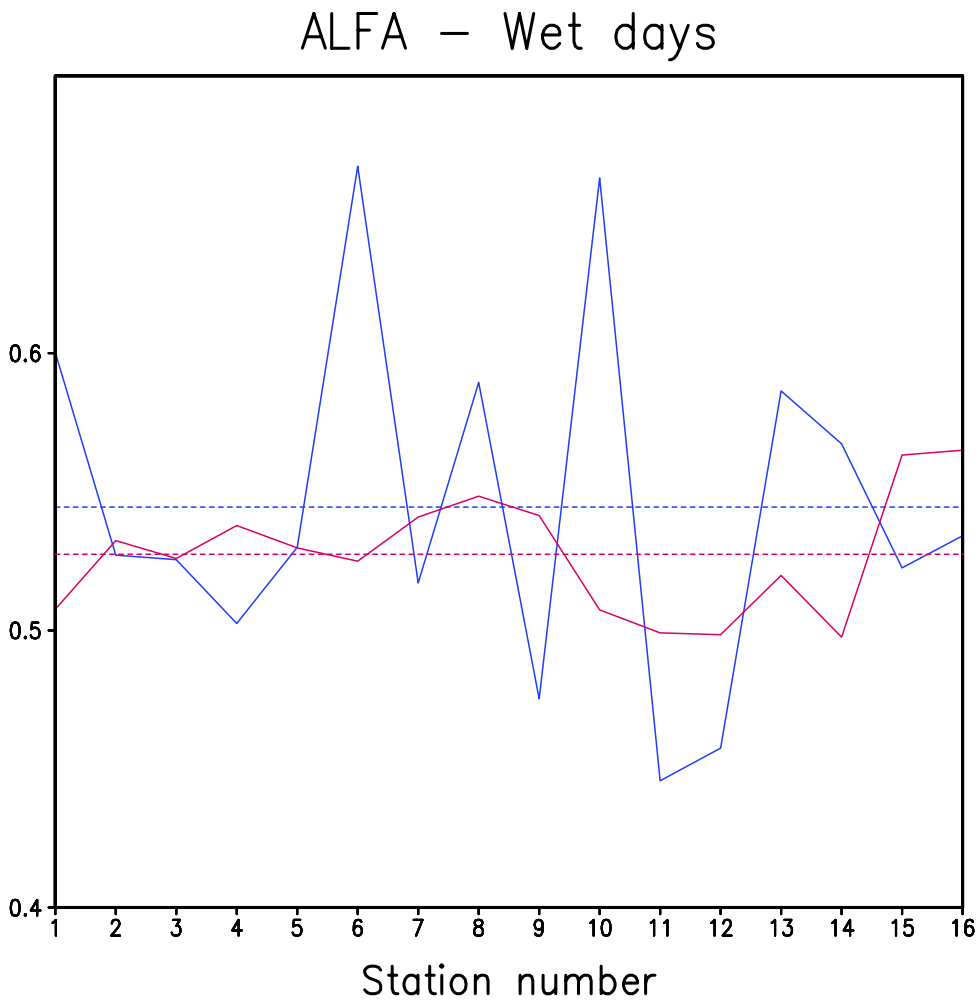

Fig. 19. The shape parameter, α, of the gamma distribution for the stations listed in Table 1.

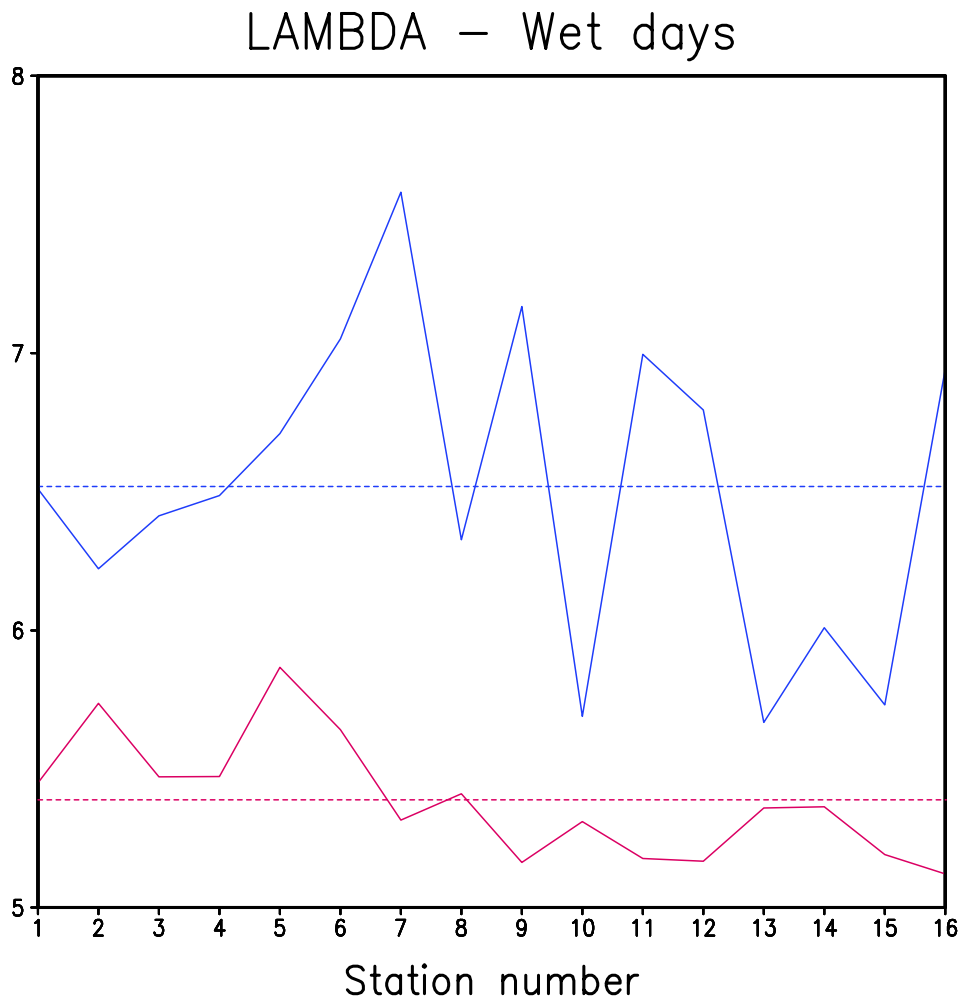

Fig. 20. The scale parameter, λ, of the gamma distribution for the stations listed in Table 1.

Spinup of precipition

A long lasting problem in numerical weather prediction is the initial transitory behavior of many variables of prime interest. One of these is the precipitation rate, which in ERA-40 is known to increase rapidly over the continental areas of the extratropical northern hemisphere (Kållberg 2004, Personal communication). In Fig. 21 is displayed how the average precipitation rate over the 16 stations in the Torpshammar catchment area develops as a function of forecast lead time. Even though there is a general spinup over the 36 hours there is also a decrease in the precipitation rate between 12 and 24 hours. The accumulated 24 hour precipitation between forecast lead time 36 and 06 which is used in this study is shown as a red solid line.

Predictability

As a measure of the predictability of precipitation in ERA-40 we display in Fig. 22 the root mean square error (RMSE) of the 24 hour accumulated precipitation for all stations. The RMSE is larger then the mean daily precipitation and has reached roughly 55 % of its saturation level.

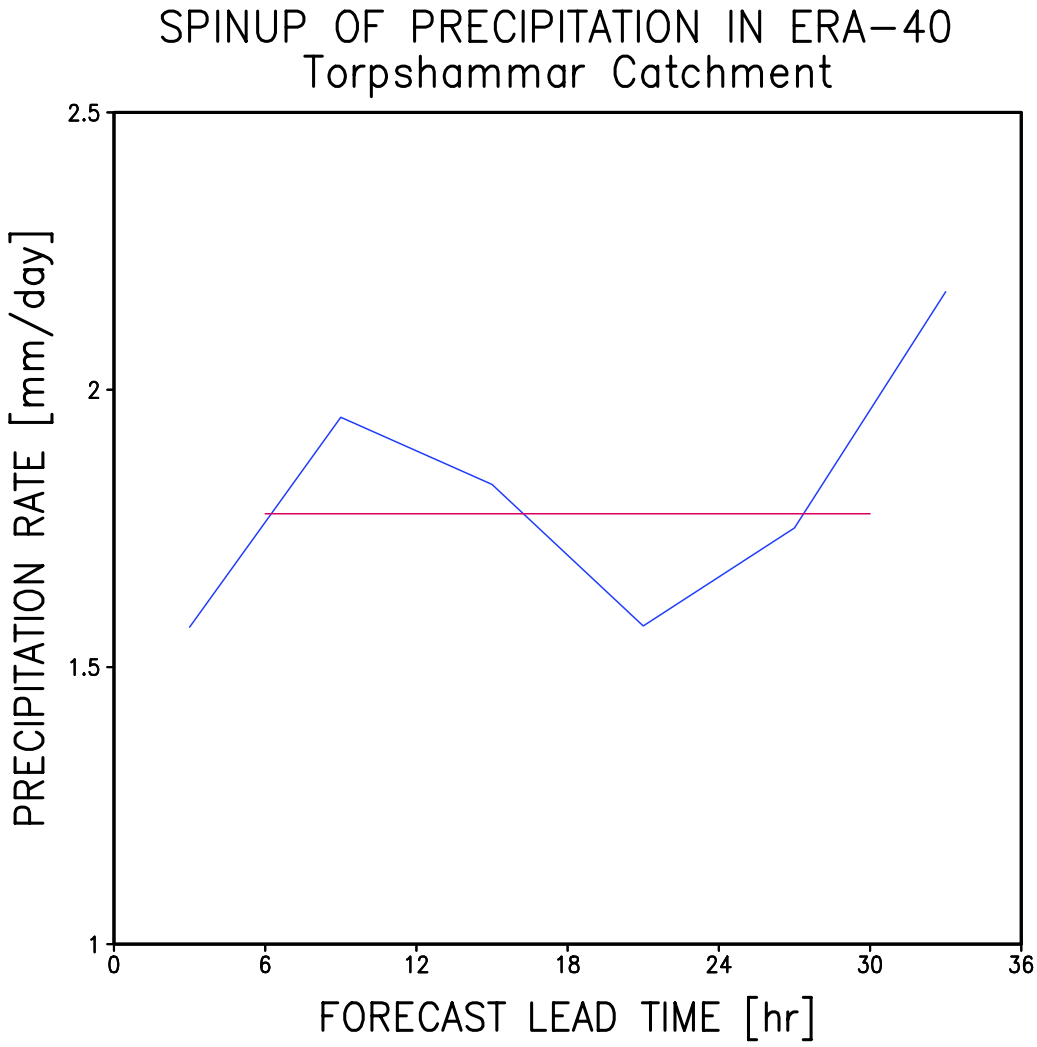

Fig. 21. The average precipitation rate over the 16 stations in the Torpshammar catchment area as a function of forecast lead time. The accumulated 24 hour precipitation between forecast lead time 36 and 06 which is used in this study is shown as a red solid line.

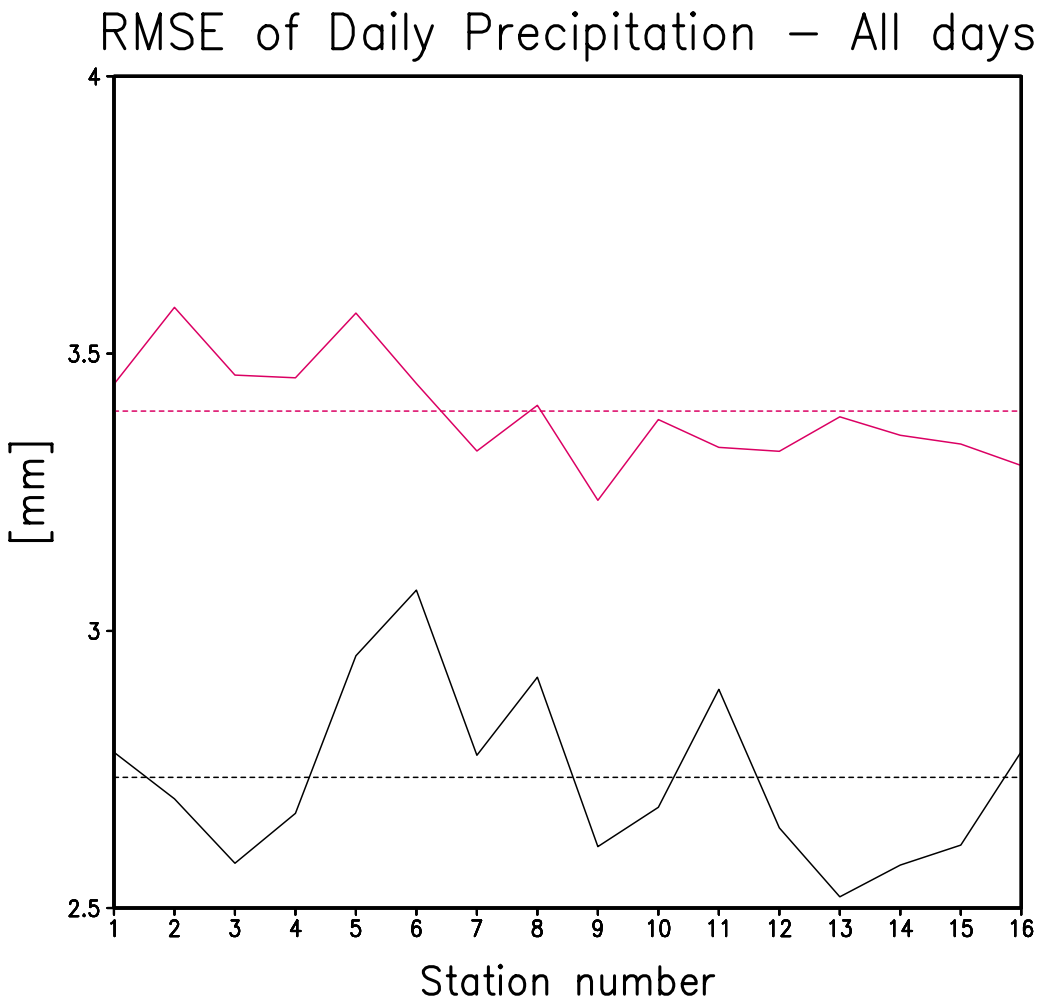

Fig. 22. The root mean square error (RMSE) of the 24 hour accumulated precipitation for the stations listed in Table 1 (black curve). For comparison is shown the standard deviation over all days of the daily precipitation amount.

10. FINAL REMARKS

This study is not complete. What is presented in this report is merely a documentation of what has been achieved up to present time. The work done is basically of a preparatory nature where the different pieces represent building blocks needed to construct the proposed statistical post-processing system. It is hoped that it will be completed in the future.

APPENDIX A

1. Introduction

ERA-40 is the third major global reanalysis effort up to date. Produced in the early 2000-s it can be said to be the second generation reanalysis following the pioneering first generation major global reanalyses, ERA-15 and NCEP-NCAR, which were produced in the middle of the 1990-ies. In Table 1 is summarized some salient features of these three major global reanalyses.

	ERA-15	**NCEP-NCAR**	**ERA-40**
Number of Years	15	56+	45
Time Period	1979JAN – 1994FEB	1948JAN -	1957SEP – 2002AUG
Horizontal Resolution	T106	T63	T_L159
Number of Vertical Levels	31	28	60
Top Level [hPa]	10	3	0.1
Number of Pressure Levels	17	17	23
Analysis Method	OI	SSI (3D-VAR)	3D-VAR
Initialization	Diabatic NNMI 5 Vertical Modes	NO	Imbedded in 3D-VAR
Time-stepping	3-D Semi-Lagrangian		Semi-Lagrangian
Model Version	IFS Cy13r4	Global Oper Model	IFS Cy23r4
Model Date	1995-04-03	1995-01-10	2001-06-12
Production Period	1994 – 1996	1994 – 1997	2001 - 2003

Table 1. Some basic facts regarding the three largest global reanalysis efforts.

The basic data for ERA-40 is quite inhomogeneous, with changes often occurring stepwise. One way to describe the inhomogenieity is to divide the time into several epochs with similarities in data sources and observation distribution. The main epochs are listed in Table 2.

Epoch	Time Period	Observational Systems and Instruments
1	1957 - 1971	No Satellite data
2	1959 – 1966	Missing SYNOP data
3	1972 – 1978	VTPR
4	1979 →	TOVS and Cloud Motion Winds
5	1987 →	TOVS and Cloud Motion Winds and SSM/I
6	1991 →	TOVS and Cloud Motion Winds and SSM/I and ERS
7	1998 →	TOVS and Cloud Motion Winds and SSM/I and ERS and ATOVS

Table 2. Main epochs based on similarities in data sources and observation distribution.

2. Time domain

The ERA-40 data set spans the 45 year period, 1 September 1957 to 31 August 2002. We will not use the entire data set, but instead restrict ourselves to the 4 decades 1961-2000. This is because the observational data from BÅK starts in 1961. The number of years considered, nyear, is thus 40 and the number of leap years, nleap, during the 40 years is 10. The total number of days, nday, in these 4 decades is then:

$$nday = nyear \cdot 365 + nleap = 14.600 + 10 = 14.610$$

Analysis data are stored every 6 hr, except for certain ground level data that are stored every 3 hr. Forecast data are also stored every 6 hr. Let t_O denote time and t_F forecast lead time. Each day data are stored at mday=18 different time pairs, (t_O, t_F), which are marked in Fig. 1 as circles and are explained in words in Table 3.

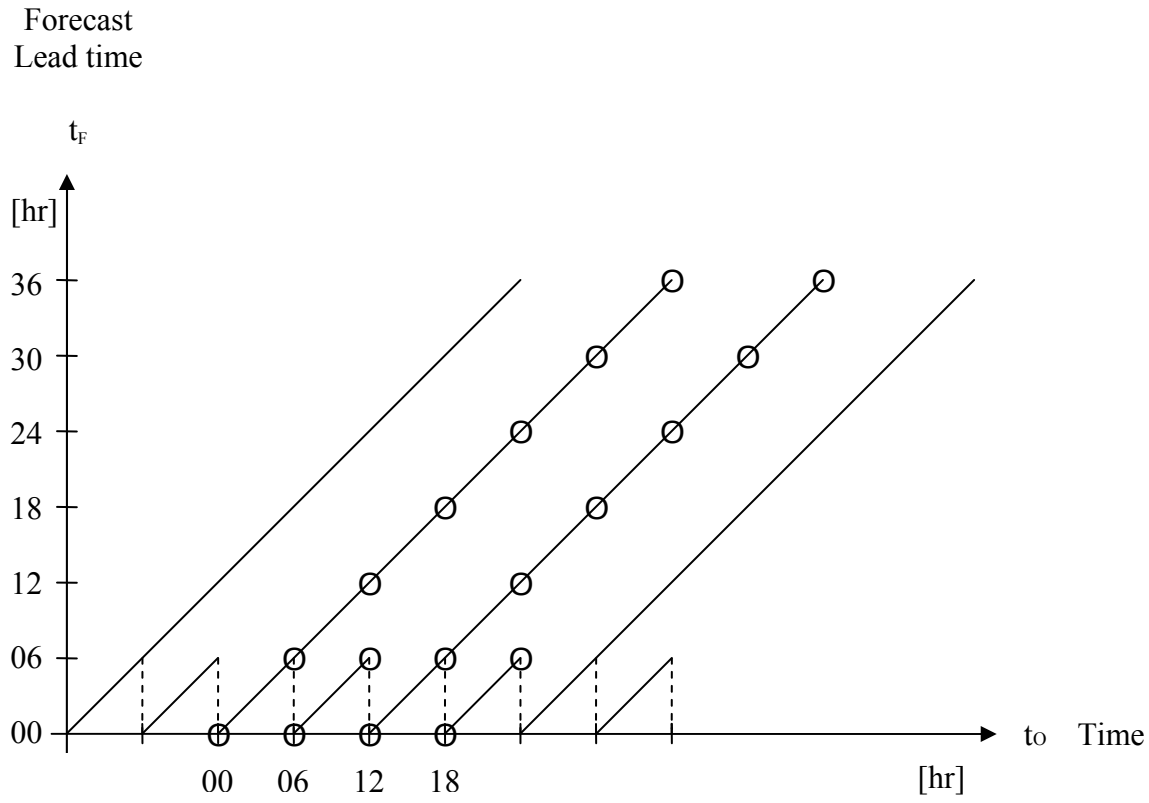

Fig. 1. Graphical illustration of time pairs.

Explanation in words of time pairs (t_O, t_F)	**Number of time pairs**
Analysis, stored every 6-hr at 00, 06, 12 and 18 UTC	4
Forecasts starting from 00 UTC, stored every 6-hr out to 36-hr forecast lead time	6
Forecasts starting from 06 UTC, stored at 6-hr forecast lead time	1
Forecasts starting from 12 UTC, stored every 6-hr out to 36-hr forecast lead time	6
Forecasts starting from 18 UTC, stored at 6-hr forecast lead time	1
Total number of time pairs per day = mday	18

Table 3. Data regarding the time pairs shown in Fig. 1.

The number of time points, ntim, is thus:

$$\text{ntim} = \text{mday} \cdot \text{nday} = 18 \cdot 14.610 = 262.980$$

This is denoted time domain T_1. We also make other partitionings into smaller time domains, decades, as given in Table 4.

Time Domain	**Time Period**	**nday**	**ntim**
T_1	1961 - 2000	14.610	262.980
T_6	1961 - 1970	3.652	65.736
T_7	1971 - 1980	3.653	65.754
T_8	1981 - 1990	3.652	65.736
T_9	1991 - 2000	3.653	65.754

Table 4. Different Time Domains.

3. Horizontal domain

The data are stored on different horizontal grids, of which some are described in Table 5 below.

Grid	Description	Number of grid-points
G_1	The **Reduced Gaussian grid** corresponding to T_L159 has almost equidistant spacing in latitude, 1.121 ^{o} , but variable spacing in longitude, ranging from 1.125 ^{o} at the equator to 20 ^{o} close to the poles.	35.718
G_2	The **linear grid** (regular lat-lon) corresponding to T_L159 has the resolution 1.125 ^{o} x 1.125 ^{o} which corresponds to 320 x 161 grid points.	51.520
G_3	The **NCEP grid** is an often used grid. It is a regular lat-lon grid with resolution 2.5 ^{o} x 2.5 ^{o} , which corresponds to 144 x 73 grid points.	10.512

Table 5. Description of different horizontal grids used to store ERA-40 data.

We will here use data which cover 2 different sub-areas of the globe. One area covers northern Europe, while the other encompasses the extratropical northern hemisphere. The data for these 2 sub-areas are given on different grids. The ensuing horizontal domains, denoted by H_1 and H_2 , are defined in Table 6.

Domain	Sub-area of the globe	Grid	nlon	nlat	Number of Grid points
H_1	[0– (28.5 – 68.4)] ^{o} E ; [48.8– 74.6] ^{o} N	G_1	20	24	480
H_2	[0 – 360] ^{o} EW ; [20 – 90] ^{o} N	G_3	144	29	4.176

Table 6. Description of different horizontal domains.

A graphical representation of the domain H_1 with all the grid points depicted is shown in Fig. 2.

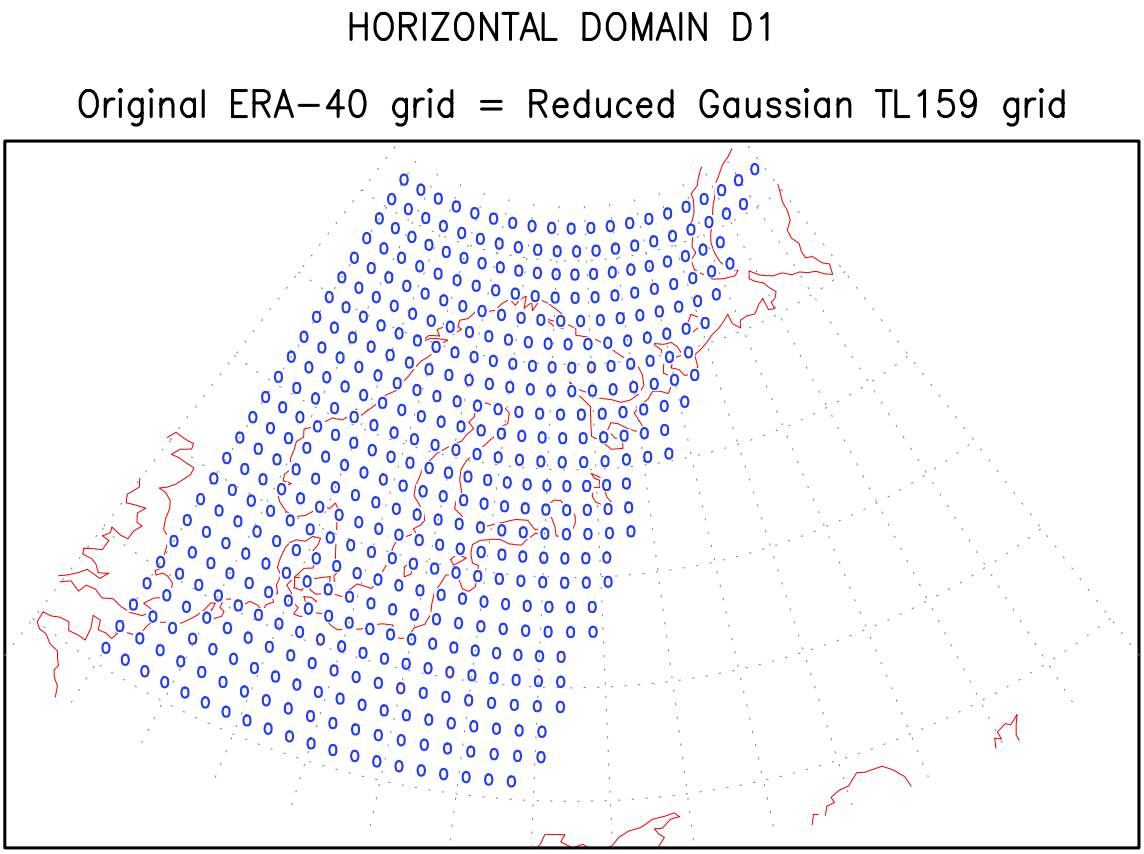

Fig. 2. Horizontal domain H_1 which covers northern Europe. All the grid points in the domain are represented by circles. They are from the model's original grid, the reduced Gaussian grid, which is very close to being equidistant.

4. Vertical domain

The ERA-40 data are stored at each of the 60 "full" model levels. The uppermost level is numbered 1 and is situated at around 65 km height, while the lowermost level is numbered 60 and is located around 10 m above the models lower boundary.

60	59	58	57	56	.	.	.	2	1
10	*30*	*60*	*100*	*160*					*64.000 m*

The model level fields have also been interpolated to the 23 pressure levels shown in Table 7. The uppermost pressure level is thus located at around 48 km height.

1000	925	850	775	700	600	500	400	300	250	200	150
100				70		50		30		20	
10				7		5		3		2	
1											

Table 7. The 23 pressure levels used in ERA-40.

We will here use 4 different sets of vertical levels, where each set has its own specific number of levels, nlev. The ensuing vertical domains, denoted by V_1 , V_2 , V_3 and V_4, are defined in Table 8.

Vertical Domain Name	**Description of levels** **Level Numbers**	**levtyp**		**nlev**
V_1	Only 1 level	001 112	surf	1
	0000			
V_2	The lowest 5 model levels + the highest model level	109	mlev	6
	60 59 58 57 56 . . . 1			
V_3	The lowest 7 pressure levels	100	plev	7
	1000 925 850 775 700 600 500			
V_4	5 pressure levels throughout the troposphere	100	plev	5
	1000 850 700 500 300			

Table 8. Definition of 4 different vertical domains

5. Variable domain

There are two sets of variables. The first set is associated with horizontal domain H_1 and consists of 3 variable domains, VAR_1 , VAR_2 and VAR_3 .The second set is associated with horizontal domain H_2 and consists of 2 variable domains, VAR_4 and VAR_5 .

The first set consists of variables that ideally should be in a column at the station location. As a substitute for a column at the station location one can use a column at the nearest grid point to the station which is not situated over a sea. The column stretches from under the ground, at the ground, at 2 m, at 10 m and further up through many levels in the boundary layer (model levels) and in the lower troposphere (pressure levels). The data is extracted from the highest resolution grid available, i e the reduced Gaussian grid of the model itself. This grid is the best available for interpolation to station location, or for using the nearest (land) grid point to the station.

The first variable domain, **VAR_1**, consists of variables at, or close to, the ground [levtyp=001], as well as vertically integrated quantities. They are described in Table 9.

Variable Name			Unit	Ana	Fcst	PAR	LEV
Short	**Long**						**TYP**
SST	sea surface temperature	(1,2)	K	A	-	34	001
T_s	soil temperature level 1	(1)	K	A	F	139	112
T2M	2 meter temperature		K	A	F	167	001
T_d2M	2 meter dewpoint temperature		K	A	F	168	001
z_o	surface roughness	(3)	m			173	001
z_{oF}	forecast surface roughness		m	-	F	244	001
z_{oHF}	log of forecast surface roughness [m] for heat			-	F	245	001
MF_x	eastward component of turbulent stress (accum)		N m^{-2} s	-	F	180	001
MF_y	northward component of turbulent stress (accum)		N m^{-2} s	-	F	181	001
HF	surface sensible heat flux (accum)		W m^{-2} s	-	F	146	001
LF	surface latent heat flux (accum)		W m^{-2} s	-	F	147	001
SLP	mean sea level pressure		Pa	A	F	151	001
PW	total column water		kg m^{-2}	A	F	136	001
Snow	snow depth		m	A	F	141	001
PP_L	large-scale precipitation (accum)	(4)	m	-	F	142	001
PP_C	convective precipitation (accum)	(4)	m	-	F	143	001
PP_S	snowfall (accum)	(4)	m	-	F	144	001
U10M	10 m eastward wind component		m s^{-1}	A	F	165	001
V10M	10 m northward wind component		m s^{-1}	A	F	166	001
Ch	Charnock parameter			A	F	148	001

Notes: (1) SST and T_s is the same over the oceans
(2) SST is a BITMAP field. SST = 0 over land
(3) Fixed field
(4) PREC = PP_L + PP_C . PP_S is included in PP_L and/or PP_C

Table 9. Description of the variables in domain VAR_1. Total number of variables, nvar=20, of which one is a fixed field.

The second variable domain, **VAR_2**, consists of variables in a column consisting of 5 model levels in the boundary layer. The highest model level is also stored, because surface pressure is stored there. They are described in Table 10.

Variable Name			Unit	Ana	Fcst	PAR	LEV
Short	**Long**						**TYP**
T	temperature		K	A	F	130	109
U	eastward wind component		m s^{-1}	A	F	131	109
V	northward wind component		m s^{-1}	A	F	132	109
ζ	vorticity		s^{-1}	A	F	138	109
ω	vertical velocity		Pa s^{-1}	A	F	135	109
q	specific humidity		kg/kg	A	F	133	109
p_s	log of surface pressure [Pa]	(1)		A	F	152	109

Notes: (1) Only at the uppermost level, mlev=1.

Table 10. Description of the variables in domain VAR_2. Total number of variables, nvar=7, of which one is stored only at the uppermost level.

The third variable domain, **VAR_3**, consists of variables in a column consisting of 7 pressure levels in the lower troposphere. They are described in Table 11.

Variable Name		Unit	Ana	Fcst	PAR	LEV
Short	**Long**					**TYP**
Φ	geopotential	m^2 s^{-2}	A	F	129	100
T	temperature	K	A	F	130	100
U	eastward wind component	m s^{-1}	A	F	131	100
V	northward wind component	m s^{-1}	A	F	132	100
ζ	vorticity	s^{-1}	A	F	138	100
ω	vertical velocity	Pa s^{-1}	A	F	135	100
q	specific humidity	kg/kg	A	F	133	100
RH	relative humidity	%	A	F	157	100

Table 11. Description of the variables in domain VAR_3. Total number of variables, nvar=8.

The second set consists of variables that describe the large scale flow in the extratropical Northern Hemisphere. Since we are interested in the large scale structure of the flow it suffices to use data from a coarse grid, and we therefore here use the regular lat-lon NCEP grid with resolution 2.5^{o} x 2.5^{o}.

The fourth variable domain, $\mathbf{VAR_4}$, consists of surface variables. They are described in Table 12.

Variable Name		Unit	Ana	Fcst	PAR	LEV
Short	**Long**					**TYP**
SST	sea surface temperature	K	A	-	34	001
SLP	mean sea level pressure	Pa	A	F	151	001

Table 12. Description of the variables in domain VAR_4. Total number of variables, nvar=2.

The fifth variable domain, $\mathbf{VAR_5}$, consists of variables at 5 pressure levels in the troposphere. They are described in Table 13.

Variable Name		Unit	Ana	Fcst	PAR	LEV
Short	**Long**					**TYP**
Φ	geopotential	$m^2\ s^{-2}$	A	F	129	100
T	temperature	K	A	F	130	100

Table 13. Description of the variables in domain VAR_5. Total number of variables, nvar=2.

6. Monthly data

The basic temporal domain for data storage is a month. The monthly data are divided into five data groups by combining different horizontal, vertical and variable domains. In Table 14 we present the five data groups that we will use.

Data group	Time domain	Horizontal domain	Vertical domain	Variable domain
D_1	Month	H_1	V_1	VAR_1
D_2	Month	H_1	V_2	VAR_2
D_3	Month	H_1	V_3	VAR_3
D_4	Month	H_2	V_1	VAR_4
D_5	Month	H_2	V_4	VAR_5

Table 14. Definitions of the five data groups used for storing monthly data in this study.

a. Number of 2-dim fields

For each data group we can calculate the number of 2-dim fields (nfield) of these monthly data by substituting the values of nlev and nvar for the various vertical and variable domains. In Table 15 is shown the values in the case of January.

Data group	Nday	mday	nlev	nvar	nfield
D_1	31	4 14 18	1	1 9 9	9.052
D_2	31	18	5 1	6 7	20.646
D_3	31	4 14	7	8	6.944 24.304
D_4	31	18	1	2	1.116
D_5	31	18	5	2	5.580
TOTAL					67.642

Table 15. Description of how the number of 2-dim fields (nfield) of the monthly data from the five different data groups are calculated.

b. Sizes of file categories

The actual sizes of the monthly files will not only depend on the number of fields (nfield) given in Table 15 since, firstly, different months have different length, secondly different data groups have different horizontal dimensions and finally, the data is written in GRIB format. In Table 16 is shown the actual sizes of the monthly files for the month of January.

Data group	File name	nfield	nlon	nlat	Size [MB]
D_1	spp_surf_an_yyyymm	9.052	20	24	10.8
D_2	spp_mlev_an_yyyymm	20.646	20	24	34.7
D_3^a	spp_plev_an_yyyymm	6.944	20	24	8.3
D_3^f	spp_plev_fc_yyyymm	24.304	20	24	29.1
D_4	nhem_surf_yyyymm	1.116	144	29	9.2
D_5	nhem_plev_yyyymm	5.580	144	29	47.5
TOTAL		67.642			139.7

Table 16. Description of the file names, the number of 2-dim fields (nfield), the dimensions of the 2-dim fields (nlon, nlat), as well as the actual sizes of the monthly GRIB files for the month of January for the six different file categories.

These GRIB files are stored at ECFS in the directory ec:/smg/spp

7. Cumulative size of decadal data

In Table 17 we summarize the total number of files and cumulative size within the total temporal domain, T_1 , as well as within each of the basic decadal temporal domains, T_6 - T_9 .

Time Domain	Time Period	Number of months	Number of file categories	Number of files	Total Size [GB]
T_1	1961 - 2000	480	6	2880	66
T_6	1961 - 1970	120	6	720	16.5
T_7	1971 - 1980	120	6	720	16.5
T_8	1981 - 1990	120	6	720	16.5
T_9	1991 - 2000	120	6	720	16.5

Table 17. Total number of files and cumulative size within each of the basic decadal temporal domains.

APPENDIX B

CONTINGENCY TABLES AND ASSOCIATED SCORES

OBS / FCST	1	2	3	TOTAL
1	a	b	c	J
2	d	e	f	K
3	g	h	i	L
TOTAL	M	N	O	T

1. Critical Success Index (CSI) or Threat Score

$$\frac{a}{M+J-a} \qquad \frac{e}{N+K-e} \qquad \frac{i}{O+L-i}$$

2. Bias

$$\frac{M}{J} \qquad \frac{N}{K} \qquad \frac{O}{L}$$

3. False Alarm Rate

$$\frac{d+g}{K+L} \qquad \frac{b+h}{J+L} \qquad \frac{c+f}{J+K}$$

APPENDIX C

The gamma distribution

The probability density function (pdf) of the gamma distribution is

$$f(x) = \frac{1}{\lambda\,\Gamma(\alpha)}\left(\frac{x}{\lambda}\right)^{\alpha-1} e^{-\frac{x}{\lambda}} \qquad , \qquad x > 0 \quad , \quad \alpha\,,\lambda > 0$$

where α is the shape parameter, λ is the scale parameter, and Γ is the gamma function.

The mean and variance of the gamma distribution are

$$E(x) = \alpha\,\lambda$$

$$VAR(x) = E(x^2) - E(x)^2 = \alpha\,\lambda^2$$

The method of moments estimation of the shape and scale parameters are

$$\hat{\alpha} = \frac{\overline{x}^2}{s^2}$$

$$\hat{\lambda} = \frac{s^2}{\overline{x}}$$

where $\overline{x}$ and s are the sample mean and standard deviation, respectively.

The maximum likelihood estimation of the shape and scale parameters are obtained from simultaneously solving the equations

$$\begin{cases} \ln\hat{\alpha} - \dfrac{d}{d\hat{\alpha}}[\ln\Gamma(\hat{\alpha})] = \ln\left[\dfrac{\overline{x}}{\left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}}}\right] \\ \hat{\lambda} = \dfrac{\overline{x}}{\hat{\alpha}} \end{cases}$$

see e g Evans et al. (2000). These equations have to be solved numerically.

REFERENCES

Antolik, M. S., 2000: An overview of the National Weather Service's centralized statistical quantitative precipitation forecasts. *J. Hydrology*. **239**, 306-337.

Carr, M. B., 1988: Determining the optimum number of predictors for a linear prediction equation. *Mon. Wea. Rev.*, **116**, 1623-1640.

Dallavalle, P., 2002: Personal communications.

Draper, N. R. and H. Smith, 1998: Applied Regression Analysis, 3rd Ed., John Wiley and Sons, 706 pp.

Evans M., N. Hastings, and B. Peacock, 2000: Statistical Distributions, 3rd Ed. John Wiley and Sons, 221 pp.

Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11**, 1203-1211.

Glahn, H. R., A. H. Murphy, L. J. Wilson, and J. S. Jensenius Jr., 1991: Lectures and papers presented at the WMO Training Workshop on the Interpretation of NWP Products in terms of Local Weather Phenomena and their Verification, Wageningen, The Netherlands, WMO, Geneva, 340 pp.

Hjort, U., 1996: Stokastiska processer i tids- och frekvensbeskrivning. Linköpings Universitet. 310 pp.

Krzysztofowicz, R. and A. A. Sigrest, 1997: Local climatic guidance for probabilistic quantitative precipitation forecasting. *Mon. Wea. Rev.*, **98**, 917-924.