TECHNICAL MEMORANDUM

114

# Statistical assessment of an observing system experiment based on frequency distributions of 500 hPa differences

P. Emmrich and P. Kållberg

Research Department

December 1985

European Centre for Medium-Range Weather Forecasts
Europäisches Zentrum für mittelfristige Wettervorhersage
Centre européen pour les prévisions météorologiques à moyen

# ABSTRACT

In a set of analysis-forecast experiments, using the ECMWF data assimilation and forecast system, different subsets of the FGGE observational data set have been evaluated。 The present work examines the statistical significance of differences between forecasts resulting from the different analyses. Four observational subsets are compared with the complete set (the controls): a 'ground-based' set, a simulated 'space-based' set, a set excluding cloud drift winds and aircraft data, and a set excluding satellite temperature profiles and aircraft data. The results show that the 'space-based' system is most different to the control in its forecast impact, closely followed by the 'ground-based' system. Both these subsets show statistically significant impacts. The other two experiments show little or no impact in the standard statistical meansures applied in this study.

CONTENTS

## 1. INTRODUCTORY REMARKS

One of the objectives of the 'First GARP Global Experiment' (FGGE) was to evaluate the performance and impact of several new, automated observing systems. In particular, the FGGE observational database contains cloud drift wind data (SATOBS) from five geostationary satellites, giving a complete data coverage at all longitudes. Temperature profiles from polar orbiting satellites (SATEMS) are also available with a reasonably high resolution of the order 250 km. In a special programme, large amounts of high quality wind observations (ASDAR) were collected from wide bodied jet aircraft equipped with inertia navigation systems.

The impact of meteorological observing systems on the quality of analyses and forecasts may be tested in Observing System Experiments (OSEs). In a typical OSE two or more data assimilations are run for a period of time, usually a week or more. In each assimilation different subsets of observations from the complete set are used. In a 'control' experiment all available data are used. From each assimilation, forecasts are run from selected times. The impact of the different data sets is evaluated by comparing the analyses and forecasts, both between themselves and with observed and analysed verifications. For a short review of OSEs, see Gilchrist (1985).

At ECMWF two major OSEs have been carried out based on the FGGE data. For the experiments a slightly modified version of the operational ECMWF analysis-forecast system was used. In one experiment, OSE I, different combinations of satellite and aircraft data were tested during an 11-day period in November 1979. In the other, OSE II, an 8-day period in February - March was selected. The two experiments are described in Uppala et al. (1985) and Kållberg (1985).

One valuable diagnostic in evaluating OSEs is the time separation of the forecasts. It is a common experience in OSE work that after a few days all forecasts from a common initial time are more similar to each other, than to the verifying analysis. It has been shown by Arpe et al. (1985) that the relative contribution of the model error to the total forecast error dominates over the contribution of the analysis error between days 2 and about 7, i.e. in the so-called medium range. Thus forecast evaluations by comparisons with the reality are difficult since they are to a high degree 'contaminated' by forecast model errors. Comparisons between the different forecasts, on the other hand, are more revealing, since all differences are due only to differences in the initial analyses.

In the present study, a statistical evaluation of the relative differences between the forecasts from the five assimilations in OSE II is carried out. The symbols used to describe the various statistics are given in Table 1.

The control and the four experiments of OSE II are as follows:

M0 control experiment: all observations (FGGE) used, except SATOBS over land and SATEMS over land below 100 hPa

M1 observations used as in M0, but excluding all SATEMS, SATOBS and AIREPS, i.e. a 'ground based' system

M2 observations used as in M0, but excluding all SATOBS and AIREPS

M3 observations used as in M0, but excluding all SATEMS and AIREPS

M6 observations used as in M0, but excluding TEMPS, PILOTS and winds from SYNOPS and SHIPS; simulating a totally space-based observing system.

From each of the five analysis sets M0 - M6, nine forecasts were run from consecutive days.

Statistical estimates like the RMS error or the correlation coefficient provide measures of the separation of the forecasts (experiments). In order to assess the experiments we compare the statistical measures derived from them. However, usually there is no a priori information about the significance of such an error difference which may be random in many cases. Therefore the object of this statistical study is:

- to describe the separation of the four experiments of OSE II on the basis of frequency distributions of geopotential differences between the individual experiment and the control experiment M0.
- to estimate the statistical significance of the separation among the experiments.

| | |
|---|---|
| n | number of data of one run |
| N | number of data of one experiment |
| nc | number of cells occupied within a histogram |
| cw | cell width |
| $x_a$ | arithmetic mean of geopotential differences |
| $\hat{x}_a$ | arithmetic mean of differences of the whole experiment |
| $SD^2$ | variance (second moment $m_2$) of differences |
| SD | standard deviation of differences |
| $CI_k$ | confidence interval at the level k |
| $CHI^2$ | Chi-square value |
| $Q(CHI^2)$ | cumulative distribution function $CHI^2$ (probability of making a type I error by rejecting $H_o$) |
| F | Fisher's F-statistic |
| Q(F) | cumulative probability density function F |
| t | Student's t-statistic |
| P(t) | cumulative probability density function t |
| u | Mann-Whitney test statistics |
| Q(u) | right tail cumulative probability density function of the (0,1) normal distribution |
| df | degrees of freedom |
| R | correlation coefficient |
| $m_3$ | third moment |
| $m_4$ | fourth moment |
| SK | skewness $(m_3/m_2)^{2/3}$ |
| KU | kurtosis $m_4/SD^4$ |

Table 1: Explanation of symbols

In this study the geopotential differences between each experiment and the control at 500 hPa have been evaluated grid-point by grid-point in a lat-lon grid with a resolution of 1.875° between 30°N and 70° ; the number of grid points in this geographical belt is 4246. The evaluation was made in every 12th hour through each forecast run from H+0 to H+240. Therefore the number of differences per run is n = 89166. Frequency distributions have been computed with a cell width of either cw = 25 or cw = 100 metres. Unless declared otherwise, all statistical parameters discussed are calculated on the basis of distributions with cw = 25. Consequently the means and modes calculated are only approximations of the corresponding real values. However, it should be noted that in this report all the frequency distributions shown in the tables and figures have a cell width of cw = 100 metres. It should also be noted that the first and the sixth forecast of experiment M6 are not available and one case from M1 - M0 is questionable due to a minor programming error.

## 2. SOME FUNDAMENTAL REMARKS ON STATISTICAL TESTING

Statistical testing is based on the logical attack of a concise, well-defined statement or null-hypothesis $H_o$. The hypothesis $H_o$ has always a negative character (there is no difference between ... or ... has no effect). Usually it is impossible to prove that $H_o$ is true; it can only be proved false. The alternative hypothesis $H_a$, the converse of $H_o$, is assumed to be true if the $H_o$ is proved to be false. Hypothesis $H_o$ is either rejected or not rejected, but never accepted.

In order to fail or to reject $H_o$ for a particular statistical test, a rejection criterion has first to be defined. This criterion is based on the probability of making an error in rejecting $H_o$. This probability is called the significance level $\alpha$ and the value of the test statistic corresponding to the probability is the critical value for the statistic. It may be with this type of testing that a true $H_o$ may occasionally be rejected. This error will be committed with a frequency $\alpha$ (that is, for example, 1% of the rejections of $H_o$ made at the 1% significance level will be incorrect). This is the so-called type I error. Now, the error of not rejecting $H_O$, when it is false, is a type II error with a frequency of say B. The power of a test is defined as 1-B, which is the probability of rejecting $H_o$ when it is false.

## 3. RESULTS

### 3.1 Summarized frequency distributions

The frequency distributions (FD) of differences to the control experiment of all runs belonging to one experiment have been summarized according to predefined classes. The result may be called the summarized experiment FD (SEFD). The SEFD of the four experiments are shown in Table 2a. Figure 1 shows the corresponding relative frequencies. The statistical parameters belonging to these distributions are listed in Table 2b.

By means of these parameters it is possible to characterise the SEFD. We can see from SK $\triangleq$ 0 that all distributions are nearly symmetric, like a normal distribution. But since a normal distribution is characterised by SK = 0 and KU = 3, we conclude that the SEFD are significantly different from normal, since KU >> 3. The SEFD are very leptokurtic.

### 3.2 Domain of geopotential differences

First consider, run by run, the number of cells occupied within the histograms of the geopotential differences between the experiments and the control. Table 2a contains the resulting frequency distributions. It provides a simplified picture of the variation of the atmospheric conditions predominant during the time period chosen. Clearly the atmospheric conditions did vary to some extent during this 9-day period. The last row of Table 2a shows that there is a minimum of variation of differences towards the middle of the period, and there is an indication of an approximately 4-day wave-like behaviour of the atmospheric conditions. In fact these differences shown in the horizontal row SUM do not turn out to be of significance. Only the comparison R2/R4 leads, by means of the U-test, to a clearly significant difference (Q(u) = .0104).

Table 2a: Frequency distribution of geopotential differences, summarized for all runs available.

| class<br>(meters) | M6-MO<br>7 runs | M1-MO<br>9 runs | M2-MO<br>9 runs | M3-MO<br>9 runs |
|---|---|---|---|---|
| -550 | 21 | | | |
| -450 | 194 | 70 | | |
| -350 | 2251 | 678 | 239 | 41 |
| -250 | 10301 | 5154 | 3402 | 1406 |
| -150 | 44038 | 31687 | 23036 | 11630 |
| -50 | 254308 | 356509 | 372656 | 380430 |
| 50 | 243069 | 370400 | 376568 | 398233 |
| 150 | 53554 | 32137 | 23365 | 9722 |
| 250 | 13073 | 5120 | 3060 | 894 |
| 350 | 2962 | 714 | 168 | 128 |
| 450 | 391 | 25 | | 10 |

Table 2b: Statistical parameters of the distributions of table 2a

| | | | | |
|---|---|---|---|---|
| $\bar{x}_a$ | -3.81 | -.39 | -.12 | .78 |
| $CI_{.05}$ | .25 | .15 | .14 | .10 |
| $-CI_{.05}$ | 74.46 | 50.11 | 42.75 | 31.36 |
| SD | 89.71 | 60.37 | 51.51 | 37.78 |
| $+CI_{.05}$ | 113.93 | 76.67 | 65.42 | 47.98 |
| N | 624162 | 802494 | 802494 | 802494 |
| nc | 42 | 38 | 31 | 30 |
| KU | 5.54 | 8.17 | 8.49 | 12.13 |
| SK | -.13 | .02 | .07 | .18 |

Figure 1: Relative frequencies of differences M(X)-M(O) summarized over all runs belonging to an experiment

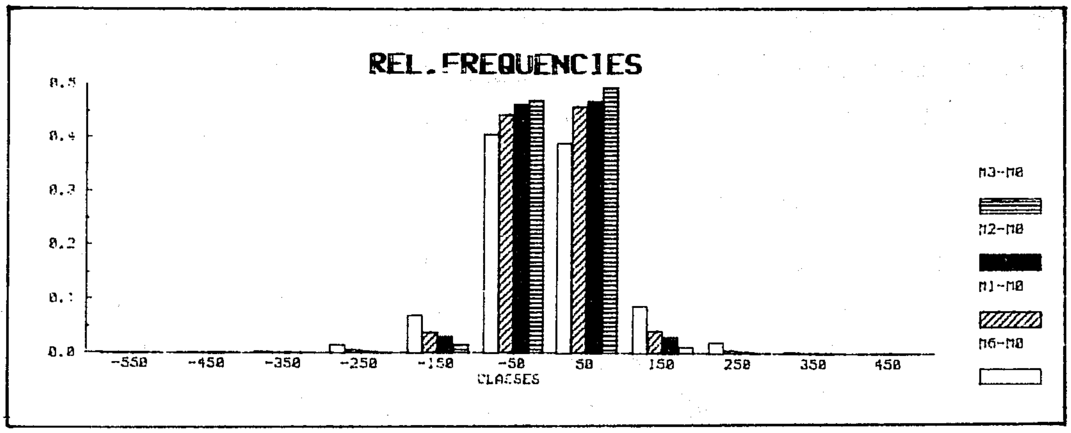

Table 3a: Number of cells occupied within the histograms (cw = 25)

| | R1 | R2 | R3 | R4 | R5 | R6 | R7 | R8 | R9 | SUM |
|---|---|---|---|---|---|---|---|---|---|---|
| M6 | | 43 | 30 | 24 | 29 | | 38 | 35 | 31 | (312) |
| M1 | 38 | 30 | 27 | 21 | 18 | 34 | 27 | 27 | 28 | 250 |
| M2 | 25 | 25 | 21 | 23 | 29 | 25 | 29 | 28 | 25 | 230 |
| M3 | 29 | 27 | 21 | 18 | 24 | 17 | 16 | 25 | 23 | 200 |
| SUM | (139) | 125 | 99 | 86 | 100 | (111) | 110 | 115 | 107 | 992 |

Table 3b: U-test referring to the last column SUM in table 3a

| | M6/M1 | M6/M2 | M6/M3 | M1/M2 | M1/M3 | M2/M3 |
|---|---|---|---|---|---|---|
| u | 1.64 | 2.49 | 2.91 | 1.02 | 2.03 | 1.63 |
| Q(u) | .0505 | .0064 | .0018 | .1539 | .0212 | .0516 |
| P(u) | .9495 | .9936 | .9982 | .8461 | .9788 | .9484 |

From the last column SUM we get an overall view of the domain of differences between the experiments. It shows a considerable decrease from M6 to M3. Experiment M6 is therefore the one which is most dissimilar to the control with the largest variation of geopotential differences. The domains occupied by both M1 and M2 are roughly equal.

Considering the results shown in Table 3a we ask the question whether the frequency distribution of the number of cells occupied varies between the experiments. By means of the U-test the hypothesis $H_o$ that the distributions given by the experiments are independent from each other may be tested. The result of this test is shown in Table 3b. The value Q(u) represents the probability of making a type I error by rejecting $H_o(SUM_i = SUM_j)$. As the Q(u) shows, $H_o$ has to be rejected for M6/M2 and for M6/M3. At the 95% significance level the difference M1/M3 also becomes significant.

**Conclusion 1:**

Experiment M6 differs very significantly from experiments M2 and M3, but it does not differ significantly from experiment M1.

## 3.3 Run means $x_a$ and standard deviations SD

### (a) Standard deviation SD

For each experiment and each run the SD of the differences between the experiments and control was calculated every 12 hours through the forecasts; Table 4 contains this information. It is very clear that the physical meaning of the standard deviation of the differences SD is much stronger than that of the number of cells occupied. The largest $\overline{SD}$ is connected to M6 and the smallest one refers to M3, whilst experiments M1 and M2 have a similar $\overline{SD}$. Qualitatively the same is true for the variation, SSD, of the SD during the time period. Using the SD we test the hypothesis $H_o$ that the resulting $\overline{SD}$ are equal. We use the U-test again, and the calculated Q(u) and P(u) are given in Table 5. These show that at the significance level of 99% the hypothesis $H_o$ cannot be rejected for M1/M2. All the other differences in $\overline{SD}$ are of high significance. That means that as far as the SD are considered, only experiments M1 and M2 react similarly during the time period.

Table **3** : Standard deviation of differences SD

| | R1 | R2 | R3 | R4 | R5 | R6 | R7 | R8 | R9 | SUM | $\overline{SD}$ | SSD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| M6 | | 114.73 | 78.82 | 68.86 | 75.37 | | 108.12 | 83.65 | 86.26 | 615.81 | 87.97 | 17.08 |
| M1 | 73.64 | 64.24 | 55.01 | 52.68 | 42.22 | 66.59 | 58.34 | 53.62 | 69.77 | 536.11 | 59.57 | 9.87 |
| M2 | 54.55 | 50.53 | 47.80 | 46.24 | 38.95 | 46.12 | 61.37 | 60.59 | 52.88 | 459.03 | 51.00 | 7.23 |
| M3 | 45.46 | 49.94 | 38.43 | 33.81 | 33.89 | 32.09 | 28.00 | 35.14 | 38.14 | 334.90 | 37.21 | 6.80 |
| SUM | 173.65 | 279.44 | 220.06 | 201.59 | 190.43 | 144.80 | 255.83 | 233.00 | 247.05 | 1945.85 | | |
| $\overline{SD}$ | 57.88 | 69.86 | 55.02 | 50.40 | 47.61 | 48.27 | 63.96 | 58.25 | 61.76 | | | |

Table 5: U-test referring to the $\overline{SD}$ caused by the experiments (see table 4, column $\overline{SD}$)

| | M6/M1 | M6/M2 | M6/M3 | M1/M2 | M1/M3 | M2/M3 |
|---|---|---|---|---|---|---|
| u | 3.12 | 3.33 | 3.33 | 1.90 | 3.40 | 3.13 |
| Q(u) | $\underline{.0009}$ | $\underline{.0004}$ | $\underline{.0004}$ | .0287 | $\underline{.0003}$ | $\underline{.0009}$ |
| P(u) | $\underline{.9991}$ | $\underline{.9996}$ | $\underline{.9996}$ | .9713 | $\underline{.9997}$ | $\underline{.9991}$ |

Table 6 : Means $x_a$

| | R1 | R2 | R3 | R4 | R5 | R6 | R7 | R8 | R9 | $\hat{x}_a$ | $S(x_a)$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| M6 | | -16.72 | 5.64 | 1.94 | 3.22 | | -1.66 | -11.82 | -7.25 | -3.81 | 8.36 |
| M1 | 1.89 | 2.19 | 2.74 | -1.14 | 2.69 | 1.27 | -1.17 | -5.05 | -6.91 | -0.39 | 3.52 |
| M2 | 3.15 | .55 | 3.03 | -1.30 | 1.01 | - .61 | .50 | -3.55 | -3.91 | -0.13 | 2.51 |
| M3 | - .26 | - .69 | 2.57 | 1.17 | 2.32 | .68 | 2.42 | -2.19 | .97 | 0.78 | 1.60 |

(b) Means $x_a$

In the same way as for SD, the mean $x_a$ was calculated (see Table 6). On average (see column $\hat{x}_a$) there is a considerable change from M6 to M3, namely from a quite large negative mean bias to a positive one. We subject these $\hat{x}_a$ to the U-test in order to test $H_o(\hat{x}_{a1} = \hat{x}_{a2})$ against $H_a(\hat{x}_{a1} > \hat{x}_{a2})$ it leads to the conclusion that the differences are all non-significant. That is the bias independent of the experiments.

Conclusion 2:

Because of the large SD, experiment M6 is to be considered as the most unstable one. But it seems that this large difference is mainly caused by the contribution of the two runs R2 and R7, which show a very large SD of differences from the middle to the end of each of them. In contrast, experiment M3 is the most stable. It gives the smallest SD, SSD and $S(x_a)$. Experiments M1 and M2 show, compared to the others, a more or less similar result. No statistically significant difference is found between these two experiments. The results suggest that the experiments have a strong effect on the frequency distribution of differences. In other words, the starting conditions, i.e. the different observing subsystems defined for each experiment are giving frequency distributions that are significantly different, except for experiments M1 and M2.

## 3.4 Initial state

It is most important to take account of the initial conditions of each experiment. In doing so we consider Table 7 which contains the frequency distributions of the geopotential differences at the starting time H = 0. According to Table 7 the bias $x_a(H = 0)$ does not vary significantly between the experiments, though the $\overline{SD}(H = 0)$ very clearly do.

In order to test the significance of the differences of the $\overline{SD}(H = 0)$, we make use of the F-test, even though the F-test requires in its strict sense the basic assumption that the distribution under consideration is Gaussian. Table 7 shows that this is not really true; nevertheless we apply the F-test as a first approximation. We test $H_o(SD_1^2 = SD_2^2)$ against $H_a(SD_1^2 > SD_2^2)$ at the significance level of 95% ($\alpha = 0.05$). With $Q(F) > \alpha$, $H_o$ cannot be rejected. As we can see from Table 8, $H_o$ has to be rejected in the three

Table 7: Frequency distribution of geopotential differences at H = 0 (starting conditions)

| class (metres) | M6-MO | M1-MO | M2-MO | M3-MO |
|---|---|---|---|---|
| -187.5 | 33 | | | |
| -162.5 | 89 | | | |
| -137.5 | 190 | | | |
| -112.5 | 443 | 10 | 4 | |
| -87.5 | 860 | 66 | 40 | |
| -62.5 | 1269 | 303 | 189 | |
| -37.5 | 2825 | 897 | 909 | 8 |
| -12.5 | 8459 | 21706 | 21194 | 26817 |
| 12.5 | 11132 | 14707 | 15779 | 10995 |
| 37.5 | 3411 | 477 | 99 | 363 |
| 62.5 | 793 | 48 | | 31 |
| 87.5 | 195 | | | |
| 112.5 | 23 | | | |
| | 29722 | 38214 | 38214 | 38214 |
| $x_a$ | -4.26 | -3.30 | -2.98 | -4.78 |
| $-CI_{.05}$ | 28.03 | 11.02 | 9.95 | 7.92 |
| SD | 36.88 | 15.52 | 14.42 | 12.19 |
| $+CI_{.05}$ | 56.06 | 27.94 | 27.69 | 28.89 |

Table 8: F-test referring to SD(H=0) as shown in table 7

| | F | Q(F) |
|---|---|---|
| M6/M1 | 5.64 | 0.0148 |
| M6/M2 | 6.54 | 0.0153 |
| M6/M3 | 7.15 | 0.0231 |
| M1/M2 | 1.16 | 0.4368 |
| M1/M3 | 1.62 | 0.3352 |
| M2/M3 | 1.40 | 0.3884 |

Table 9: Paired sample t-test according to FD(H=0) as shown in table 7 but excluding all frequencies smaller than |D| = 25

| | t | df | P(t) |
|---|---|---|---|
| M6/M1 | 3.36 | 5 | 0.9899 |
| M6/M2 | 2.96 | 4 | 0.9793 |
| M6/M3 | 3.04 | 2 | 0.9534 |
| M1/M2 | 1.42 | 4 | 0.8854 |
| M1/M3 | 1.23 | 2 | 0.8285 |
| M2/M3 | 0.51 | 1 | 0.6493 |

cases where M6 is involved. That means the $\overline{SD}(H = 0)$ of M6 is significantly larger than the $\overline{SD}(H = 0)$ of the other three experiments. Referring to M1, M2 and M3, no significant differences in $\overline{SD}(H = 0)$ exist. A very lucid picture of this result can be obtained by constructing the 95%-confidence interval of the $\overline{SD}(H = 0)$ as shown in Table 7.

An alternative approach is to use the paired sample t-test. This does not require the normality and equality of variances assumptions, but instead assumes that the differences $d_i = x_i - y_i$ come from a normally distributed population. If there is pairwise association of the data of the two samples x and y, the paired t-test is usually more appropriate, since it takes into account the pair-to-pair variability amongst the data. If we make use of the whole sample size, as shown in Table 7, it follows when $N_x = N_y$ that $t = \bar{d}/s_d = 0$ because $\bar{d} = n^{-1}*SUM(x_i - y_i) = 0$. This is true for M1, M2 and M3. Therefore we consider geopotential differences $|D| < 25$ m (the central domain of the geopotential differences) to represent the very trivial analysis noise or, say, the very ordinary inaccuracy of the analysis. This domain will be neglected. Doing so results in the t-statistic given in Table 9. The test-hypothesis $H_o$ is that the two sample populations under consideration have statistically equivalent means. The test is performed using n paired events from the sample population with the number of degrees of freedom df = n-1. The significance level is taken to be 95%. The results show that $H_O$ cannot be rejected as far as the comparisons M1/M2, M1/M3 and M2/M3 are concerned because of $P(t) < 1-\alpha$. For the same reason $H_o$ has to be rejected referring to M6/M1, M6/M2 and M6/M3. This result is very close to that obtained from the F-test referring to the differences in the variances.

**Conclusion 3:**

Considering the complete set of runs, the initial conditions of experiment M6 are different from those of the other three experiments at the significance level of 95%. Statistically the initial conditions of experiments M1, M2 and M3 are similar. This result shows that the contribution of the surface-based subsystem (TEMP, PILOT, SYNOP) is of the highest importance for the accuracy of the initial analysis. For the same reason the step by step exclusion of the components of the space-based subsystem, as done in experiments M1, M2 and M3, seems to be less important.

## 3.5 The mean growth rate of the geopotential differences

Next we consider how the standard deviation of differences grow in twelve hour steps from $\overline{SD}_{12}(H = 0)$ to $\overline{SD}_{12}(H = 240)$; Table 10 shows the increase of these $\overline{SD}_{12}$ (the average standard deviation of differences over all runs belonging to one experiment in steps of 12 hours). Figure 2 shows the calculated 99%-confidence interval of these $\overline{SD}_{12}$. Clearly the $\overline{SD}_{12}$ grow in a more or less similar manner in all experiments. It is also evident that this growth is caused only by the model. If the initial SD is large (see M6), then the growth rate is also large.

In order to perform a simple significance test we divide the $\overline{SD}_{12}$ into the two classes with $\overline{SD}_{12} \lesseqgtr 50$ and test the resulting contingency table by means of the $CHI^2$-test (see Table 11). The chosen hypothesis $H_o$ is that no significant differences in frequencies exist. The probability of being correct in rejecting $H_o$, when it is false, is $P(CHI^2) = 1-Q(CHI^2)$. The $P(CHI^2)$ results in Table 10 show that the frequencies obtained from M6 are different from those obtained from M2 and M3, at least at the significance level of 98%. However, even the difference between the frequencies of M6 and M1 is non-significant (the α-risk of making a type I error by rejecting $H_o$ is greater than 5%). Again the differences in frequencies between M1 and M2, M1 and M3, and M2 and M3 are also not significant. This result is necessarily in line with Conclusion 1.

Conclusion 4:

On the basis of these results we have to reconsider Conclusion 3 in respect of M1. Considering the change of $\overline{SD}_{12}$ during the averaged run it follows that the significant difference in initial $\overline{SD}(H = 0)$ between M6 and M1 (see Table 6 and 7) is probably not very meaningful. It cannot be verified by referring to the differences in $\overline{SD}_{12}$.

It is further of interest to consider the mean growth rate of the $\overline{SD}_{12}$ called $\Delta\overline{SD}_{12}$ (see Table 12), from which some important features are seen. The $\Delta\overline{SD}_{12}$ of M6 grows very rapidly, reaching its maximum rate at H+66. It appears that even short-range forecasts are unfavourably affected under M6 conditions. The maximum growth rates of M1 and M2 are reached much later at H+174. This indicates that medium-range forecasts are

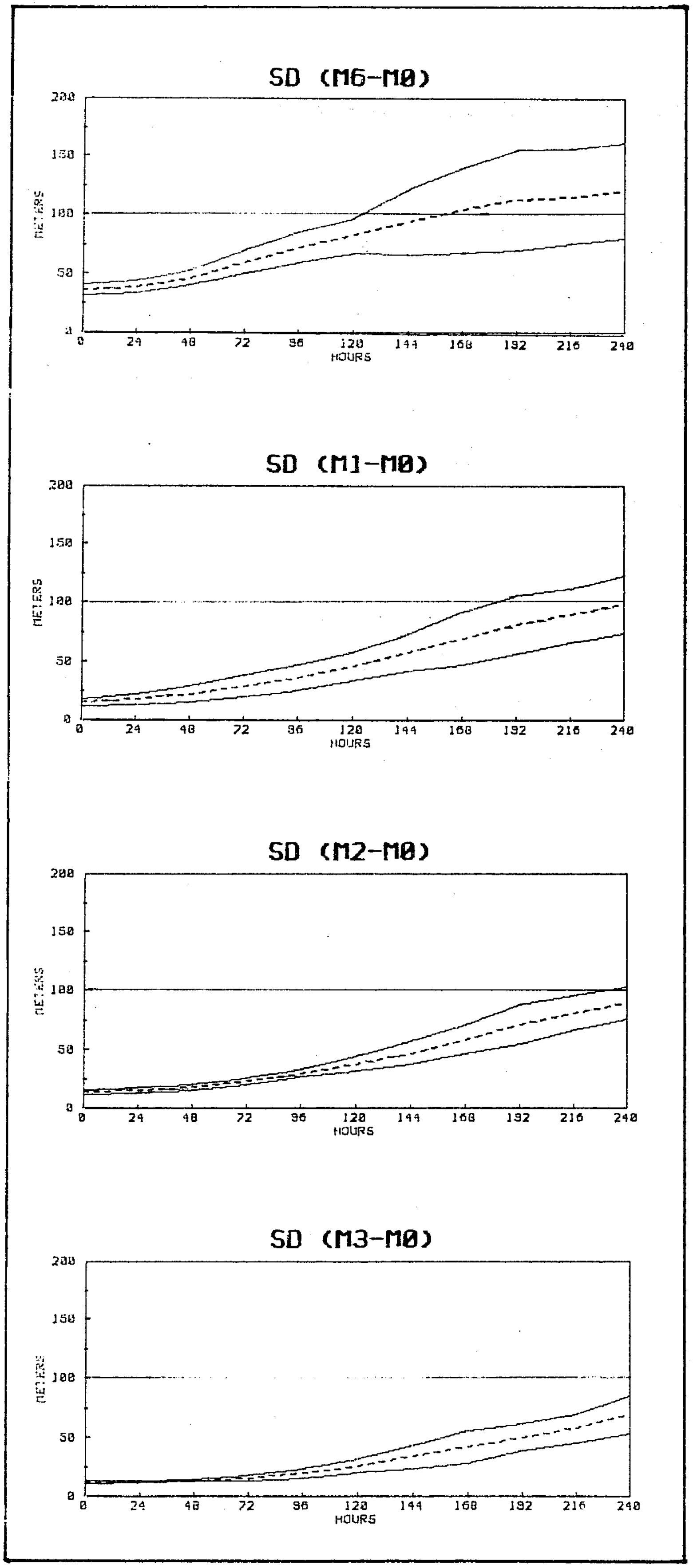

Figure 2: 99%-confidence interval of the mean standard deviation SD of differences M(x)-M(0)

Table 10: Mean SD of differences in steps of 12 hours from H=0 to H=240

| H | M6 | M1 | M2 | M3 |
|---|---|---|---|---|
| 0 | 36.7 | 15.0 | 14.1 | 12.1 |
| 12 | 37.1 | 15.8 | 15.3 | 12.7 |
| 24 | 39.3 | 17.6 | 15.8 | 12.9 |
| 36 | 43.4 | 19.5 | 16.6 | 13.3 |
| 48 | 47.2 | 22.3 | 18.3 | 13.6 |
| 60 | 53.5 | 25.7 | 20.5 | 14.3 |
| 72 | 61.6 | 29.6 | 23.4 | 15.8 |
| 84 | 67.9 | 33.0 | 26.4 | 17.7 |
| 96 | 73.5 | 37.0 | 30.2 | 20.1 |
| 108 | 78.6 | 41.4 | 34.2 | 22.9 |
| 120 | 84.0 | 46.6 | 38.2 | 26.1 |
| 132 | 90.0 | 52.6 | 42.5 | 30.0 |
| 144 | 96.1 | 58.7 | 47.6 | 34.0 |
| 156 | 101.7 | 64.7 | 53.5 | 38.3 |
| 168 | 106.3 | 70.3 | 59.2 | 42.8 |
| 180 | 111.5 | 76.8 | 65.5 | 46.7 |
| 192 | 115.3 | 82.3 | 71.9 | 50.5 |
| 204 | 117.3 | 86.1 | 77.2 | 54.4 |
| 216 | 118.6 | 89.9 | 81.2 | 58.2 |
| 228 | 120.3 | 94.1 | 85.0 | 63.3 |
| 240 | 124.0 | 99.7 | 89.3 | 70.1 |

Table 11: Contingency table for $SD_{12} \gtrless 50$

| | SD < 50 | SD > 50 | |
|---|---|---|---|
| M6 | 5 | 16 | 21 |
| M1 | 11 | 10 | 21 |
| M2 | 13 | 8 | 21 |
| M3 | 16 | 5 | 21 |
| | 45 | 39 | 84 |

$CHI^2$-test

| | $CHI^2$ | df | $P(CHI^2)$ |
|---|---|---|---|
| total | 12.40 | 3 | 0.9938 |
| M1/M2/M3 | 2.60 | 2 | 0.7278 |
| M6/M1 | 3.63 | 1 | 0.9434 |
| M6/M2 | 6.22 | 1 | 0.9874 |
| M6/M3 | 11.52 | 1 | 0.9993 |
| M1/M2 | 0.39 | 1 | 0.4671 |
| M1/M3 | 2.59 | 1 | 0.8926 |
| M2/M3 | 1.00 | 1 | 0.6833 |

Table 12: Growing rate of $\overline{SD}_{12}$

| H | M6 | M1 | M2 | M3 |
|---|---|---|---|---|
| 6 | 0.4 | 0.8 | 1.2 | 0.6 |
| 18 | 2.2 | 1.8 | 0.5 | 0.2 |
| 30 | 4.1 | 1.9 | 0.8 | 0.4 |
| 42 | 3.8 | 2.8 | 1.7 | 0.3 |
| 54 | 6.3 | 3.4 | 2.2 | 0.7 |
| 66 | 8.1 | 3.9 | 2.9 | 1.5 |
| 78 | 6.3 | 3.4 | 3.0 | 1.9 |
| 90 | 5.6 | 4.0 | 3.8 | 2.4 |
| 102 | 5.1 | 4.4 | 4.0 | 2.8 |
| 114 | 5.4 | 5.2 | 4.0 | 3.2 |
| 126 | 6.0 | 6.0 | 4.3 | 3.9 |
| 138 | 6.1 | 6.1 | 5.1 | 4.0 |
| 150 | 5.6 | 6.0 | 5.9 | 4.3 |
| 162 | 4.6 | 5.6 | 5.7 | 4.5 |
| 174 | 5.2 | 6.5 | 6.3 | 3.9 |
| 186 | 3.8 | 5.5 | 6.4 | 3.8 |
| 198 | 2.0 | 3.8 | 5.3 | 3.9 |
| 210 | 1.3 | 3.8 | 4.0 | 3.8 |
| 222 | 1.7 | 4.2 | 3.8 | 5.1 |
| 234 | 3.7 | 5.6 | 4.3 | 6.8 |

Table 13: U-test and unpaired sample t-test according to growing rate of $\overline{SD}_{12}$ as shown in table 12

|  | df | t | P(t) | u | Q(U) | P(U) |
|---|---|---|---|---|---|---|
| M6/M1 | 38 | 0.23 | 0.5892 | -0.30 | 0.6170 | 0.3830 |
| M6/M2 | 38 | 1.01 | 0.8402 | -0.93 | 0.8246 | 0.1754 |
| M6/M3 | 38 | 2.41 | 0.9896 | -2.27 | 0.9885 | 0.0115 |
| M1/M2 | 38 | 0.89 | 0.8097 | -0.73 | 0.7674 | 0.2326 |
| M1/M3 | 38 | 2.45 | 0.9906 | -2.22 | 0.9867 | 0.0133 |
| M2/M3 | 38 | 1.50 | 0.9292 | -1.60 | 0.9448 | 0.0552 |

Table 14: Experiment by experiment correlation of the mean 12 hourly growing rate of SD

|  | $R^2$ | F | df1 | df2 | Q(F) |
|---|---|---|---|---|---|
| M6/M1 | 0.21 | 4.88 | 1 | 18 | 0.04 |
| M6/M2 | 0.04 | 0.80 | 1 | 18 | 0.38 |
| M6/M3 | 0.00 | 0.02 | 1 | 18 | 0.89 |
| M1/M2 | 0.62 | 29.51 | 1 | 18 | 0.00 |
| M1/M3 | 0.79 | 66.62 | 1 | 18 | 0.00 |
| M2/M3 | 0.58 | 24.76 | 1 | 18 | 0.00 |

unfavourably affected with M1 or M2 conditions. In contrast, the maximum growth rate of M3 is obtained at the end of the run. A nearly constant and very small growth rate is found up to H+54; after H+54 a slow increase takes place. Therefore M3 shows the smallest mean growth rate.

The significance of differences in the mean growth rates, as shown in Table 12, is easy to test by means of the U-test and the unpaired sample t-test. The latter can be used because the $\Delta\overline{SD}_{12}$ are distributed normally and the differences in variances are non-significant. The hypothesis $H_o$ is that no significant differences in mean growth rates do exist. From Table 13 is seen that the mean growth rate of M3 differs from those of M6 and M1 at least at the significance level of 98% (see P(t) or Q(u)). The difference between M3 and M2 is non-significant since the α-risk of making a type I error by rejecting $H_o$ is greater than 5%. The exclusion of SATEM and AIREP, in M3 is therefore probably of least impact by reference to the error growth. Experiment M3 has the best starting conditions and the resulting growth rate of SD is smallest. Therefore experiment M3 represents the best observation set of the four experiments under consideration.

Finally we correlate the change in 12 hourly mean growth rate of $\overline{SD}$ experiment by experiment. The resulting $R^2$ (see Table 14) is tested by means of the F-distribution with $F = R^2 (N-2)/(1-R^2)$, df1 = 1 and df2 = N-2 = 18. The test hypothesis $H_o$ used is that a significant correlation does exist; the probability that $H_o$ can be rejected is Q(F). It is found that only a weak correlation exists betwen the growth rates of M6 and M1, but there is no significant correlation between M6 and M2, or between M6 and M3. This is again in line with Conclusions 1 and 4. The growth rates of M1, M2 and M3 are similar because they do correlate very strongly with one another.

## 4. FINAL CONCLUSION

Analysed and forecast 500 hPa geopotentials have been compared, gridpoint by gridpoint, from nine forecasts from each of five data assimilations using different combinations of the FGGE data. It is found that, when compared with the complete FGGE set (experiment M0), exclusion of SATEMS and AIREPS (experiment M3) shows the least impact (expressed as differences) in a statistical sense, closely followed by experiment M2 (where SATOBS and AIREPS were excluded). Subsets M6 (excluding TEMPS, PILOTS and surface winds, i.e. a space-based subset) and M1 (excluding all space-based data, i.e. the observing system before the advent of satellites) both show statistically significant differences to the control set M0. M6 is more different from the control than any other set. The relative difference between M1 and M2, i.e. the net impact of SATEM data is in most cases not statistically significant.

The somewhat puzzling result that removal of SATEM and AIREP data from the full FGGE data set does not show a statistically significant impact, is discussed further in Uppala et al. (1984), where other measures of the impact are applied. During OSE II there was little meteorological activity over oceans where these data ought to be most important. Furthermore there was a persistent gap in the potentially available SATEM data over the eastern Pacific, off the coast of North America that could well have affected the results. Uppala et al. concluded that "the much reduced activity over the oceans, coupled with the absence of SATEM data led to negligible impact of the SATEM and SATOB data on the forecast skill in this second period".

Our result that experiments M3 (and M2) did not show any statistically significant data impact when compared with the full FGGE set in M0, is thus in good agreement with the results of Uppala et al. It may be said in hindsight that the choice of the period for OSE II was unfortunate for the purpose of proving the usefulness of satellite data. On the other hand, the results demonstrate the very great difficulties,and costs, encountered when trying to judge different observing systems against each other in an Observing System Experiment.

## Acknowledgement

Thanks have to be addressed to the European Centre for Medium Range Weather Forecasts for making my two week visit possible. During this stay the idea for this statistical judgement took shape. I am very obliged to Dr. Cubasch and Mr. Uppala for their friendly assistance and also to Dr. Riddaway for editing and tidying-up the manuscript.

## REFERENCES

Arpe, K., A. Hollingsworth, A.P. Lorenc, M.S. Tracton, G. Cats and P. Kållberg, 1985: The response of numerical weather prediction systems to FGGE level IIb data. Part II: Forecast verification and implications for predictability. Q.J.Roy.Met.Soc., 111, 67-101.

Gilchrist, A., 1985: Observing system experiments - review and outlook. ECMWF Seminar on Data Assimilation System and Observing System Experiments with particular Emphasis on FGGE, 3-7 September 1984, 145-164.

Kållberg, P., 1985: Performance of some different FGGE observation subjects for a period in November 1979. ECMWF Seminar on Data Assimilation Systems and Observing System Experiments with Particular Emphasis on FGGE, 3-7 September 1984, 203-228.

Uppala, S., A. Hollingsworth and S. Tibaldi, 1985: Results from two recent observing system experiments at ECMWF. ECMWF Seminar on Data Assimilation Systems and Observing System Experiments with Particular Emphasis on FGGE, 3-7 September 1984, 165-202.