A gratis Two-Model-Ensemble versus EPS

P.Emmrich and K.Balzer
Deutscher Wetterdienst, Zentralamt Offenbach

Introduction

The daily problem of the forecaster involved in medium-range forecasting is at least twofold: How to judge

- **The large day to day variation of the skill and**
- **The dissent between available operational models [1].**

Namely, because of the large day to day variation of the model's skill, EP was required by the meteorological community in order to get an a-priori-information of the confidence of a deterministic prediction.

Considering simple ensemble statistics [1]:
Due to the fact that all members of an ensemble - including the (unperturbed) control forecast - are equally likely, we have to concentrate on the ensemble mean and variance.
With the ensemble size i=1,...,m the ensemble average is $\overline{E} = \frac{1}{m}\sum_{i=1}^{m} e_i$.
Introducing a reference (verification) X, the mean quadratic distance becomes

$$m^{-1}\sum (e_i - X)^2 = (\overline{E} - X)^2 + m^{-1}\sum (e_i - \overline{E})^2,$$

where the first term on the right is the error of the ensemble mean and the second term expresses the ensemble variance around the average.
This means that the error of $\overline{E}$ compared with the mean error of the individual members (left term) is reduced by the ensemble dispersion.
Therefore, **basically we expect from an ensemble prediction system**

1) an improvement of the skill by means of the ensemble mean,
2) a forecast of forecast skill by means of the ensemble variance.

[1] The day to day variation in model's skill is nearly of the same magnitude as the variation between two models at the same forecast time. Comparing the models GM and ECMWF, the latter varies naturally stronger because of the difference in model resolution (GM is a T106-version of ECMWF-T213).
Noticable but not surprising: The ratio of dissent can be reduced significantly by applying the arithmetic mean of both models.

Furthermore, it is possible to derive

3) a most probable solution from the ensemble by means of the ensemble mode $\frac{1}{n}\sum_{j=1}^{n} e_{j_{\mathrm{mode}}}, n \leq m$ and

4) a probability forecast of weather elements by means of their ensemble distribution.

The daily practical approach

The most important requirement for the daily practical work is, to have available proper local (regional) weather parameters derived from the basic models available. In Germany this is realized by means of a statistical interpretation scheme (AFREG) which is applied to both operational models GM and ECMWF and, in an experiment so far, to the ensemble mean. Considering these interpretation results, indeed large daily differences are frequently observed. This is a common serious problem. The simplest approach to partially overcome this problem is to average both routine forecasts. In doing so, a gratis two-member-ensemble is created representing two different models and initial analyses as well. Perhaps, this can be considered as one of the simplest EP approaches. The clear disadvantage is the limitation to only 2 members. It may be an advantage that both models are highly developed in contrast to the EPS model version T63. Making consequential use of the AFREG procedure for deriving real weather parameters from all the basic guidance available, 3 strategies have to be tested:

- the gratis 2-model-mixture,
- the best deterministic solution from the ensemble ($\overline{E}$),
- the most probable solution from the ensemble (EMode).

As far as EPS is considered at least the following basic approaches or combinations are to be examined:

Basic information	**derive weather parameter by means of**
[A] Ensemble Mean	[a] statistical interpretation
[B] Ensemble Mode	[b] direct model output
[C] Whole Ensemble	[c] forecaster's subjective interpretation

Versions [Aa] and [Ba] have been examined by means of the AFREG interpretation scheme. Versions [Ab] and [Bb] or [Cc] are not yet examined/verified.

It is first the intention to present latest EPS verification results with respect to the two very essential questions from the practical point of view:

- **Is the spread of EP an useful predictor of the skill of the EMean and EMode forecast?**
- **How does EP perform compared to the skill of the operational model?**

Finally, these results are compared with those obtained from the simple two-model-average based an the operational models GM and ECMWF.

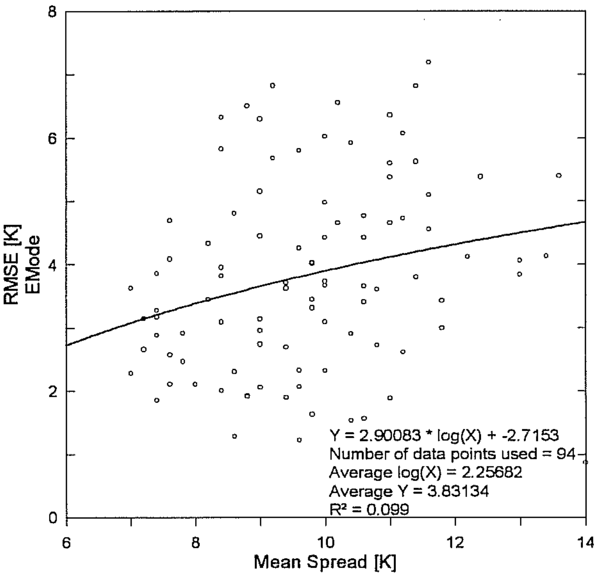
EPS verification

- **Is the spread of EP an useful predictor of the skill of EMean and EMode forecast?**

Various verifications show that the relationship between the mean spread per run and the mean skill (RMSE) derived from plume diagrams is too weak for a successful prognostical usage and in addition, it seems that there is a remarkable decrease in EPS skill in 1994/95 compared to 1993, which affects the skill-spread-relationship negatively. In 1993 the verified common explained variance between both parameters was 17%. Verification from 1994 as well as latest results show a decrease to about unexceptable 8-10% (figure 1). Furthermore, this relationship is to some extent significantly smaller considering single forecast days.

Figure 1:

Plume Diagram Aachen, Temperature Forecast 850 hPa
January - June 1994
Mean RMSE of EMode as Function of Mean Spread
(days 1-9 averaged)



The same decrease in skill is found according to cluster verification compared to the skill of the operational forecast T213 (figure 2).

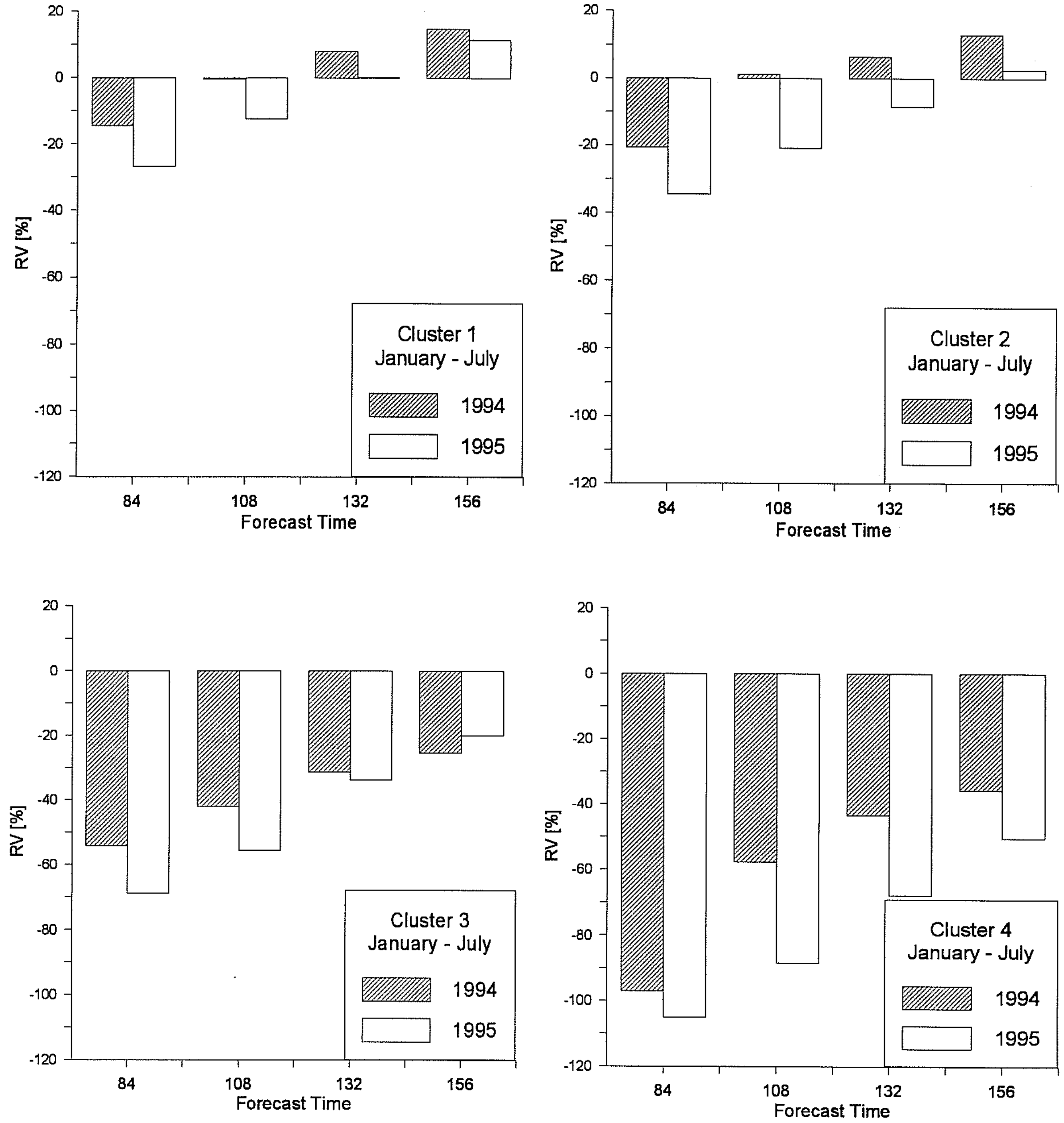


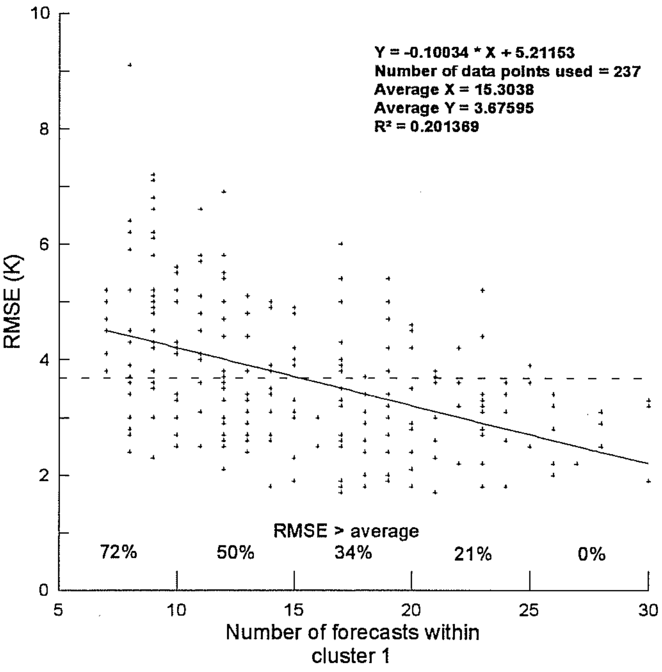
Figure 2: Verification of clusters
Geopotential height forecast 500 hPa, Europe
January - July 1994/95

Reduction of error variance RV in relation to T213-forecast

➢ **Conclusion:** The first expectation of EP as an useful aid for predicting the forecast skill has mainly failed so far.

But, it seems that the choice of the proper skill predictor is important. The number of clusters determined and/or the number of members within a cluster express the ensemble spread as well. We found that the cluster-1-size works better as a predictor (coefficient of determination up to $R^2 = 0.25$) than the mean spread of plumes as in figure 3. Indeed, this relationship is still not strong enough compared to what is expected by the forecaster. Furthermore, this relationship expresses fundamental behaviour of RMSE: The larger the sample size (cluster size) the better the forecast skill in terms of RMSE.

Figure 3:
Relationship between RMSE of cluster-1-mean and cluster-1-size
Temperatur forecast 500 hPa, H+156, Europe
January - August 1995



➢How does EP perform compared to the skill of the operational model?

We consider 4 strategies forecasting T850 by means of plume diagrams

(Average over days 0 to 10)

1993

strategy taken

	T213	T63	EMode
RMSE [K]	3.83	3.64	3.39
RV relative to T213		**10%**	**22%**

1994

strategy taken

	T213	T63	EMode	E(Max+Min)/2
RMSE [K]	3.97	3.96	3.61	3.45
RV relative to T213		1%	17%	**24%**

RV: Reduction of error variance

Conclusion:

1) The former superiority of T63 (RV = 10% in 1993) over the much more sensible T213 disappears in 1994/95. Nevertheless, T213 and T63 show the same level of skill, which may reflect the fact that T63 is the less sensitive model (the smoother the field verified the better scores RMSE, relatively).

2) The EMode is more successful than T213 already from day 2 on (whereby the relative break from day 4 to day 5 is not understood). Indeed, this may be in general an effect of averaging over the part of the ensemble representing the mode forecast and so far it coincides with our general expection.

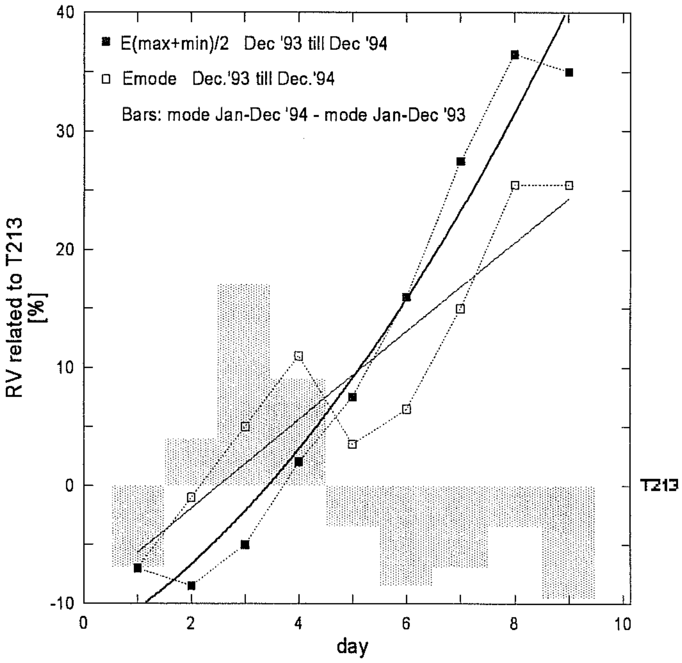
3) The most successful strategy is clearly E(max+min)/2 (RV = 24%), in particular for days 5 to 9, as illustrated in figure 4. This reflects exactly what we expect from averaging an ensemble.

4) Hypothetically, the cluster-1-mean should be equivalent to the EMode. And no doubt, from the theoretical point of view it speaks much in favour of the mode- and cluster-1-mean-philosophy. But, in practice the simple EMean is clearly more skillful than the EMode. So, the question arises whether this philosophy is really appropriate and useful. Following the general ensemble statistics it may be the more useful approach to concentrate strictly on the ensemble mean and dispersion.

Figure 4:

Plume Diagram Rostock, Temperature 850 hPa
December 1993 - December 1994

Reduction of Error Variance RV related to T213

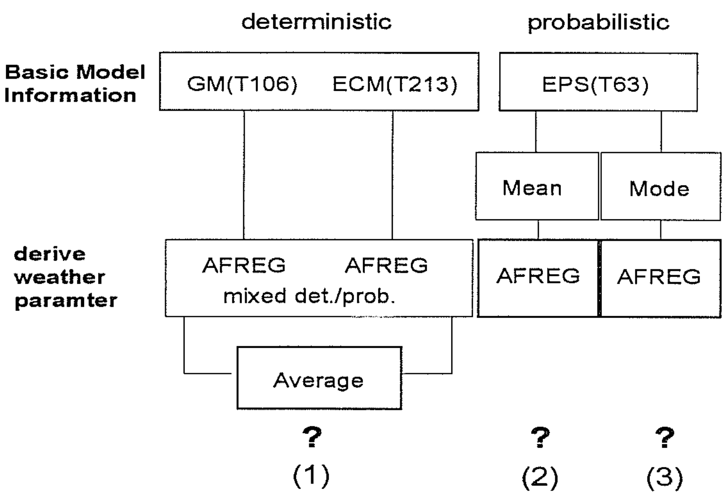


The winner: A simple two-model-ensemble

As known and mentioned previously, in Germany the two operational models GM(T106) and ECMWF(T213) are interpreted in terms of local weather by means of the statistical interpretation scheme AFREG. The simplest approach to overcome the problem of large daily differences between both model interpretations is, to average both forecasts.

On this background the question arises

Which strategy performs best ?



A first verification (sample: Dec.'94 - April '95, N = 1015) concerning forecast days 4 and 5 showed that strategy (1) was clearly the best followed by (2).

For the second period of verification (June to August 1995) all days up to 10 were included. Figure 5 summarizes this verification (average over 6 german locations and over the parameters Tmax, Tmin, dd, ff, PoP and SD).

The result is highly surprising and sensational:

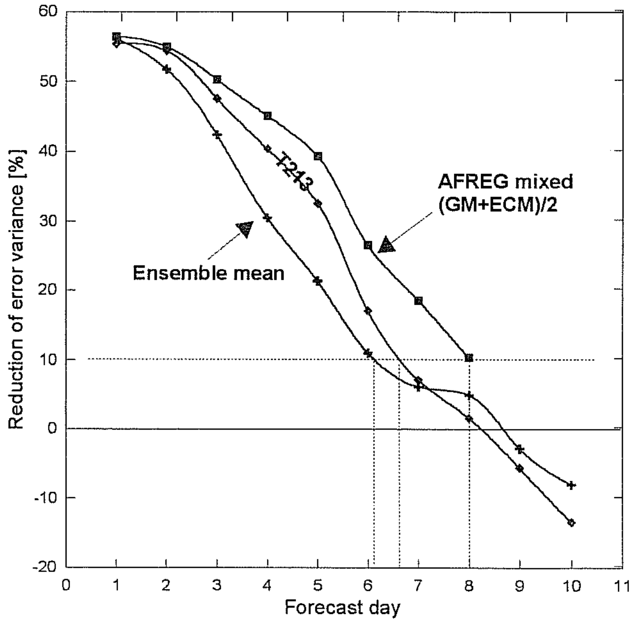
- The gratis two-model-mixture is clearly the winner of the 3 competing approaches. Its difference to T213 and EMean is significant in the medium-range from day 3 on.
- Most surprising: The T213-interpretation performs better than the EMean-interpretation up to day 7. In earlier verifications this was the case up to day 4, only.

Important to notice: The integrity of this verification lacks due to the fact that only one summer season is considered. So, it may be a simple seasonal effect that T213 performs so well. Before presenting serious

Figure 5:

SKILL COMPARISON (preliminary result)
based on 6 PPM-interpreted weather parameters*):
Tmin, Tmax, dd, ff, PoP, SD

6 stations averaged, test period: June - August 1995



*) PPM procedure AFREG applied to EMean, T213 and (GM+ECM)/2

conclusions such seasonal effects have to be eliminated by enlarging the sample size significantly. Indeed, in the event that this result would stabilize, it should be considered as indicator for making use of higher sophisticated models in ensemble prediction: A signal toward using model version T106 instead of version T63.

[1] K. Fraedrich, C. Ziehmann, 1995
Praktische Vorhersagbarkeit: Persistenz in rotem Rauschen
Meteorol. Z., N.F. 4, 139-149