Verification of ECMWF products at the Deutscher Wetterdienst (DWD)

DWD - Martin Göber, Cristina Primo

1 Summary of major highlights

The usage of a combined GME-MOS and ECMWF-MOS continues to form the best available guidance for the production of local short and medium range forecasts. ECMWF high resolution forecasts in conjunction with GME forecasts are now being used for the production of a probabilistic warning guidance based on the MOS technology.

2 Use and application of products

2.1 Post-processing of model output

2.1.1 Statistical adaptation

The high resolution ECMWF model (12 UTC run) and DWD's model GME (00 UTC) are statistically interpreted up to 7 days in terms of near surface weather elements by means of a perfect prog scheme (AFREG) as well as by MOS and subsequent weighted averaging of the two interpretations to form „AFREG/MIX“ and “MOS/MIX”. Since 2008 ECMWF high resolution forecasts in conjunction with GME forecasts have been used for the production of a probabilistic warning guidance based on the MOS technology.
Usage of the 00 UTC run of the ECMWF model within the MOS-scheme has been prepared and will become operational in 2010.
Kalman filtering of some EPS surface variables has been terminated.

2.1.2 Physical adaptation

2.1.3 Derived fields

2.2 Use of products

The high resolution ECMWF model forms together with DWD's model GME the general operational data base. ECMWF‘s high resolution model is always used together with other models in short- and medium-range forecasting. For medium range forecasting the EPS is used additionally; in the short range the LEPS (Local model nested into EPS clusters) provides ensemble information. EPS products are used intensively in order to create a daily simple confidence number and describe alternative solutions. Furthermore, they are used to estimate the prospect for extreme weather events. Here, extensive use of the Extreme Forecast Index (EFI) is made.

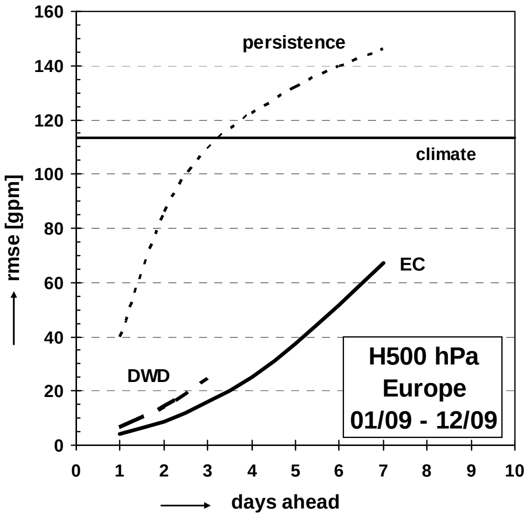
3 Verification of products

3.1. Objective verification

3.1.1 Direct ECMWF model output (both deterministic and EPS)

3.1.2 ECMWF model output compared to other NWP models

Upper air forecasts from ECMWF continued to exhibit smaller errors than DWD-GME forecasts (Fig. 1). The RMSE of the ECMWF model for 500hPa geopotential height has hardly changed from 2008 to 2009 and it decreased by about 5 % (0,5 gpm) in the short range for the GME. ECMWF MSLP error growth with forecast range is about one day better than for DWD-GME in the short range (fig. 2). The RMSE of the GME model for MSLP has hardly improved from 2008 to 2009 and only slightly for ECMWF (about 0,1 hPa in the short range and by 0,3 hPa in the medium range).



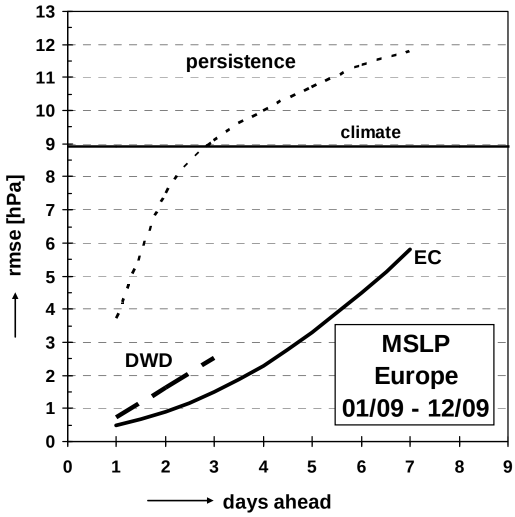


Fig. 1 RMSE 500hPa geopotential over Europe. DWD (Numerical Weather Prediction model GME), EC (high resolution ECMWF model), persistence (analysis from the initial state is used as a forecast for all following days), climate (long term mean of the predictand (H500, MSLP) serves as a constant forecast).

Fig. 2 Same as fig. 1, but for RMSE of mean sea level pressure.

3.1.3 Post-processed products

Here, various statistically post-processed model forecasts are compared for the following:

Predictands

MIN = daily minimum temperature (°C)
MAX = daily maximum temperature (°C)
SD = daily relative sunshine duration (%)
dd = surface wind direction (°) 12 UTC. Only verified, if ff(obs) ≥ 3 m/s
ff = surface wind speed (m/s) 12 UTC
PoP = Probability of Precipitation > 0 mm/d
PET = potential evapotranspiration (mm/d)
RR = a binary predictand: precipitation amount > 0 mm/d: Yes/No;

Forecast Types

AFREG/MIX = Perfect prog product AFREG(MIX) = AFREG(EC)+AFREG(DWD)/2
EC = high res. ECMWF model, DWD = operational DWD Global Model "GME" (initial time: 00 UTC). AFREG is generated for several *areas* of the whole Germany, but verified against *point* observations at 6 stations.

MOS/MIX = post processed product, a weighted average of Model Output Statistics of MOS/GME and MOS/EC

and **Verification measures**

rmse is used for both categorical and probabilistic forecasts (equals square root of the Brier Score)
RV = Reduction of Variance against reference, $1-(rmse/rmse^*)^2$, here: mean value for day 2 ... 7
$rmse^*$ = smoothed climate as the best reference forecast to evaluate forecast skill
HSS = Heidke Skill Score, only for binary predictands
$\overline{\mathbf{HSS}}$ = mean value for day 2 ... 7

rmse		day							rmse*	
		+2	+3	+4	+5	+6	+7	+8	(climate)	RV [%]
MIN	AFREG/MIX	2,43	2,48	2,63	2,83	2,97	3,24	3,55	3,97	55
	MOS/MIX	1,72	1,98	2,27	2,59	2,82				66
MAX	AFREG/MIX	2,34	2,43	2,59	2,94	3,32	3,80	4,21	4,36	60
	MOS/MIX	1,77	2,03	2,38	2,77	3,11				68
SD	AFREG/MIX	24,8	25,5	26,2	27,4	28,5	30,0	30,9	29,5	15
dd[1)]	AFREG/MIX	44,6	47,5	52,9	59,8	66,7	77,1	81,8	90,2	64
	MOS/MIX	33,7	40,4	49,1	59,8	67,1				69
ff	AFREG/MIX	1,64	1,70	1,80	1,95	2,09	2,20	2,22	2,10	23
	MOS/MIX	1,50	1,62	1,76	1,92	2,00				29
PoP	AFREG/MIX	39,5	40,2	41,2	43,9	45,5	47,6	49,2	46,8	22
	MOS/MIX	35,5	38,6	40,7	43,6					28
PET	AFREG/MIX	0,756	0,779	0,790	0,816	0,857	0,910	0,946	0,896	16
HSS%										HSS
RR	AFREG/MIX	50	45	41	30	25	18	8	0	42
	MOS/MIX	60	52	42	29					46

Table 1 Verification of operational medium range forecasts for 6 stations in Germany (Hamburg, Potsdam, Düsseldorf, Leipzig, Frankfurt/M., München); 01/2009 - 12/2009; rmse and HSS, respectively. Day of issue = day +0 = today at noon. [1)] Here, persistence is used as a 'reference forecast'.

The skill (RV) of minimum temperature and wind (speed and direction) forecasts was significantly lower in 2009 than in the preceding years. However, these were the variables with the largest improvements in the years before. MOS/MIX forecasts have substantially smaller errors than AFREG/MIX, which is only partly due to the lower (and thus less realistic) variability of MOS forecasts. The lower variability of MOS, especially in the medium range, is an obstacle for the use of it for forecasts of more severe weather. Here, the more variable solutions of the EPS serve as an important additional guidance.

The application of post-processing lead to largely reliable probability of YES/NO precipitation (PoP) forecasts (fig. 3), yet with a slight underconfidence in situations of low PoP forecasts.

Figs. 4-5a,b show two things: i) the MOS technology performs better than a perfect prog technology (AFREG) ; ii) mixing post-processed products from both models leads to a very moderate improvement of the forecast. However, in the medium range the gain in skill is about half a day.

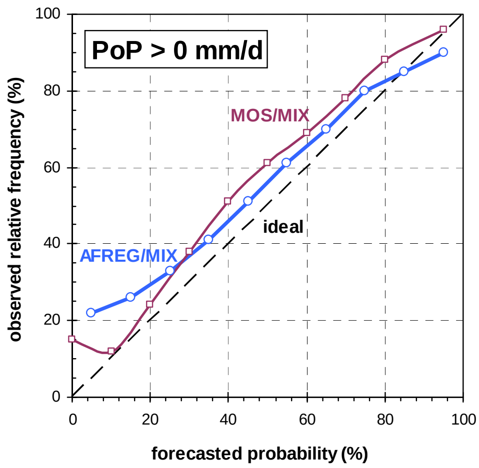


Fig. 3 Reliability diagram (6 stations, 01/09 – 12/09, day+2 ... day+7; only up to day+5 for MOS(MIX))

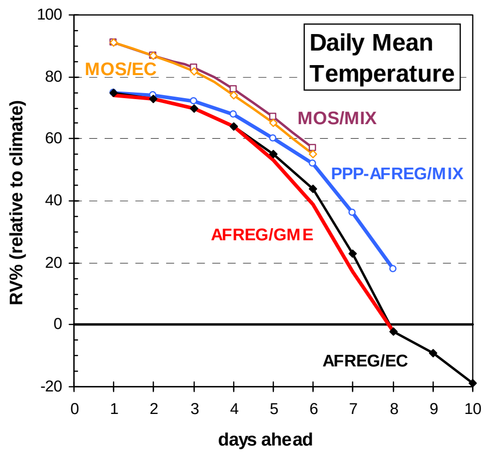
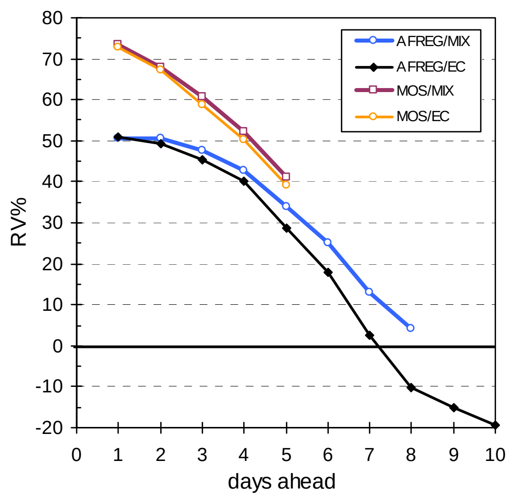


Fig. 4 Forecast skill RV for Daily Mean Temperature (DWD, 6 stations, 01/09 – 12/09



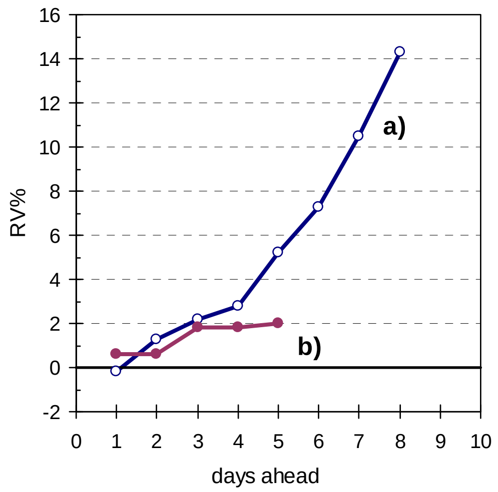
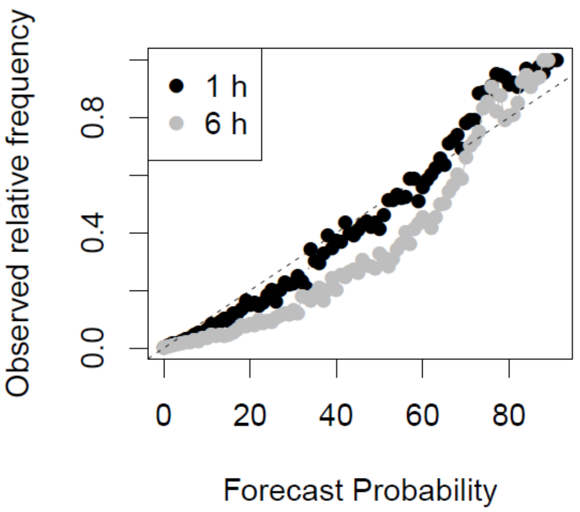


Fig. 5a Forecast skill RV as a function of range, averaged for all predictands taken in table 1 (without PET and RR)

Fig. 5b follows from fig. 5a:
a) Blue line:
RV(AFREG/MIX) - RV(AFREG/EC)
b) Claret red line:
RV(MOS/MIX) - RV(MOS/EC)

Finally, first results are presented of the verification of a probabilistic guidance for warning purposes (WarnMOS, a MOS for extreme events based on ECMWF and GME models). One hour forecast are very reliable (Fig. 6a), yet the larger the lead time, the larger the tendency to be overconfident in cases of low probability forecasts. Unbiased forecasts (Fig. 6b) would be achieved at a decision probability of 44%, leading to a rather low hit rate of 65% and a false alarm ratio of 50%. Strong overforecasting is necessary to reach a hit rate of 80%, which is one target for warnings. This could only be accomplished by accepting false alarm ration well above 50%. Notice, that the optimum HSS is reached for an almost unbiased forecast.



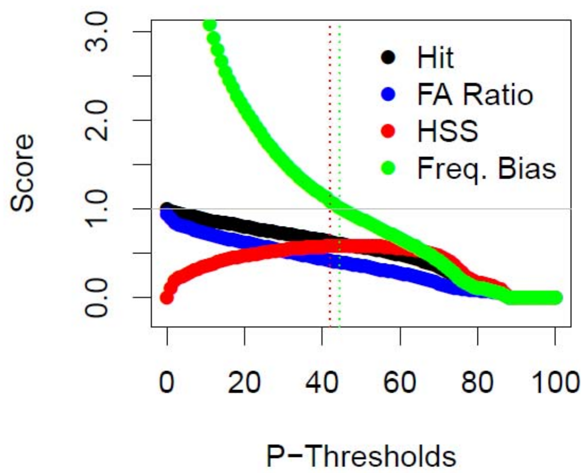


Fig. 6a Reliability of one and six hourly forecast of gusts above 14 m/s, summer 2009.

Fig. 6b Different scores as a function of the probability threshold for the “Yes” forecast of gusts above 14 m/s with a lead time of 6 hours, summer 2009. The green vertical line highlights the scores achievable for an unbiased forecast and the decision threshold of 44 %; the red vertical line cuts through the maximum HSS at a decision threshold of 42 %.