TECHNICAL MEMORANDUM

200

Recommendations on the verification of local weather forecasts

P. Nurmi

Operations Department

May 1994

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

Recommendations on the verification of local weather forecasts

Pertti Nurmi

1 Introduction

The Technical Advisory Committee (TAC) of ECMWF has repeatedly emphasised the value of the annual report, (Application and) Verification of ECMWF Products in Member States. This report includes, attached to it as an Appendix, a proposal for the verification of local weather forecasts, which was prepared by Harald Daan of KNMI in 1985. Daan's recommendations were mainly based on two WMO publications (Daan, 1984 and Murphy, 1985) and represented the state-of-the-art in forecast verification of the time. Much of the information is still valid today. Unfortunately this valuable undertaking, with an aim to exchanging verification results based on uniform methodology, has to a large extent been overlooked by the Member States. (Only a single country has largely adopted the proposal, and only a few countries have seen it appropriate to include verification results of end products delivered to the users, whereas most have ignored this area entirely.) Realizing the situation, the 18th session of the TAC (1993) requested that a revised proposal for the verification of local weather forecasts be issued as new guidance to Member States. The TAC further recommended that Member States would undertake more verification of end products to be included in their contributions to the verification report.

This proposal provides some simple and basic, as well as somewhat more extensive recommendations covering the weather elements to be evaluated and the measures for their verification. The basic elements and measures should be quite easy to adopt but hopefully a most comprehensive set of recommendations will be employed. Even these more extensive specifications can provide only a rather general insight of the overall forecast quality, but the design of a uniform scheme which can be generally accepted and implemented requires compromises and simplifications. This should, however, not prevent anybody from utilizing additional, more sophisticated, forecasting and verification measures.

It is hoped that the present recommendations would serve as a solid basis for operational verification practices within the Member States. It is also considered beneficial to adjust any existing verification practices of local weather forecasts (e.g. among groups running high resolution limited area models like EWGLAM) in line with these recommendations.

2 General guidelines

Harald Daan's proposal provided useful guidance for the present recommendations. Some of the earlier predictands and verification measures are kept more or less similar. However, some amendments and additions, but also simplifications, are proposed.

Perhaps the most comprehensive change to the approach of the previous proposal is the reduction of probabilistic forecasts, because it has been observed that they are quite rarely used in the Member States. The undeniable merits of probability forecasts are acknowledged and the use of them is by no means rejected. Therefore methods for their verification are still provided. In the foreseeable future, when the output from ECMWF Ensemble Prediction System (EPS) is expected to become more widely used, the probabilistic approach will eventually become more obvious and the verification guidelines will be readjusted accordingly.

Resulting from the advent of computerization in the weather services, the parts in the earlier proposal (Annexes 1 and 2), which included specific forms for data input and manipulation are now considered outdated. Due to this, and because of the abundant and diverse computer systems of the weather services, no explicit recommendations concerning the methods for data input are provided here.

Another extensive modification to the previous proposal is the redefinition of the required forecast ranges (lead times). Although the role of ECMWF is to produce medium-range forecasts, verification results at all ranges, from short to late-medium range, are relevant and therefore included in the present recommendations. This is rational as the same evaluation procedures can be applied to cover all existing local weather forecasts being produced within a weather service. This will also facilitate evaluation of the output from short-range limited-area models, emphasising the use of same verification measures for their verification.

The Member State contributions to the annual verification report have to a large extent dealt with general evaluation of ECMWF model performance in the free atmosphere, i.e. the field which is extensively covered by the Centre itself. However, verification results of the local area weather forecasts within individual countries, as evaluated by the respective countries, are primarily required. A loose definition of "local area weather forecasts" in this context is appropriate:

(1) **Direct Model Output (DMO)** of near surface weather parameters;
Forecasts of the corresponding parameters produced directly by the model.
The following predictands are considered here:

- two metre temperature
- precipitation
- ten metre wind speed

(2) **Post-Processed Products (PPP)** of the corresponding parameters;
These are usually statistically adapted model output by perfect prog, MOS or Kalman filtering techniques.

(3) **End Products (EP)** of the corresponding parameters;
Operational forecasts produced and delivered either to the general public or to various special users.

DMO and PPP have become an increasingly important guidance in the production of final forecasts (EP) to the consumers. In some cases DMO or PPP are known to be delivered untouched by human hand directly to the end user. The production chain leading to a final local area forecast can thus be viewed (with severe simplification) as:

DMO --> PPP --> EP

Each of these components should naturally be subject to a comprehensive, comparative, quality controlling process.

DMO and PPP are usually available in digital form enabling straightforward verification (archiving of the data is advisable). EP will apparently have to be interpreted in most cases from their original form (e.g. worded forecasts by duty forecasters) to facilitate similar treatment, presumably requiring some labour.
Such efforts will, nevertheless, be eventually very rewarding in providing valuable feedback of the various components of the production chain to the operational forecasting environment as well as to product and model developers.

Source of forecasts and observations

It is obvious that the forecasts to be verified should include, whenever available, all following products:

- **DMO**
- **PPP**
- **EP**

It is then possible to see, e.g. how much the PPP schemes can improve over DMO in different countries and, further, what is the improvement (if any) gained by EP.
PPP are typically generated for a number of selected synop stations. For DMO, the closest gridpoint to the relevant station(s), rather than interpolated values, should be applied. The verifying values are observations at the stations.

Reference forecasts

Climatology and/or **persistence** are needed for computing the skill of forecasts. Persistence provides usually better short-range reference forecasts. Climatological mean (or median) values should be defined to be compatible with the predictands they are to be used with. For the verification of probabilistic forecasts, climatological frequencies of the events are needed.

Reference stations

Work is in progress within the European Working Group on Limited Area Modelling (EWGLAM) for the definition of a common list of synoptic stations for verification. Pending the completion of this work, the selection of station(s) is left to each country to decide. The minimum requirement is one **representative** station (for a small country). For larger countries, and for countries with diverse distinct meteorological or climatological areas, as many stations as seen adequate may be used. The number of stations, as such, is not a crucial point for verification. To facilitate wind verification, oceanic and/or coastal station(s) must be chosen (all weather elements need not be verified at same stations).

Forecast ranges

All forecast ranges should be included in the same verification process. Ranges can be defined either with respect to the initial analysis time, or with respect to the day of issue of the local weather forecast. The first definition corresponds to forecast model verification, it should be mentioned first in all results to avoid confusion. The second definition matches end-product verification. It takes account of the common time lag between ECMWF model output (resulting from the 12UTC analysis time) and the issued forecasts (e.g. a D+1 accumulated precipitation forecast may correspond to ECMWF +36 to +60 hour forecast range). In verification results, product range may be indicated in addition to model output range.

It is recommended that **daily forecasts** are verified separately for all available ranges. The upper forecast limit is not defined, because practices (especially concerning EP) vary between Member States. In many cases DMO and PPP are either available, or easily obtainable, over the whole ECMWF forecast range up to +240 hours. Forecast ranges may be different for different predictands.

In addition to providing forecasts for individual days, **mean values** (temperature) and **accumulations** (precipitation) over specified time intervals are proposed: average forecast **over the early, middle and late part of the forecast**, e.g. D+1 through D+3, D+4 through D+6 and D+7 through D+10. These, again, may be defined according to preferable or existing practices.

Forecast types

The forecasts should in the first place be **point estimate values. Probability forecasts** are supported as additional predictands.

Besides point estimate values, **alternative forecasts** are proposed for certain weather events (e.g. rain vs. no rain, gale warnings). They can be often interpreted as "by-products" from point estimate values. It is good to notice that, by using alternative forecasts, one is restricted to verification methods of nominal level predictands. More details are given in the respective sections covering the different weather elements.

Verification measures

Forecast quality should be addressed by using both absolute measures, **reliability** and **accuracy**, as well as relative measures, **skill**. The explicit verification measures are provided separately for each predictand in the next section.

3. Predictands and measures for their verification

This section provides recommendations for the verification of the three weather elements, temperature, precipitation and wind speed, respectively.

3.1 Temperature

Predictands

T_{min} (18UTC - 18UTC)
minimum temperature valid at D+1, D+2, D+3, D+4, D+5, ...

T_{max} (18UTC - 18UTC)
maximum temperature valid at D+1, D+2, D+3, D+4, D+5, ...

Time averages of temperature should also be verified, for example:

T_{1-3} **mean temperature** averaged over day 1 through day 3

T_{4-6} **mean temperature** averaged over day 4 through day 6

T_{7-10} **mean temperature** averaged over day 7 through day 10

Notes:

- The first period of T_{min} / T_{max} starts on the day of issue of the forecast
- Alternatively, nighttime T_{min} (18UTC-06UTC) and daytime T_{max} (06-18UTC) may be used, because T_{min} / T_{max} are measured usually during night/day. However, especially in northern latitudes in winter this is often not the case.
- At the moment T_{min} / T_{max} DMO are not included in the product dissemination of ECMWF, but they could be made available if required.
- If no T_{min} / T_{max} forecasts are produced either as PPP or EP then 12- (or 6-) hourly temperature forecasts valid at synop observation times can be used instead. For DMO, they should be used.
- The quality of daily two metre temperature forecasts is known to deteriorate rapidly after about D+3, but daily verification results beyond that are still considered useful in addressing possible changes or trends in forecast quality in the medium range .
- The averaging periods may be defined according to relevant applications and needs in the Member States.

- Additionally, T_{min} / $T_{max\ 1\text{-}3,\ 4\text{-}6}$ and $_{7\text{-}10}$ may be provided, i.e mean minimum/maximum temperatures averaged over the given periods.

- If alternative forecasts like, "$T_{4\text{-}6}$ is two degrees below/above normal", are preferred, they can easily be translated from point value forecasts. Vice versa is not possible

Verification Measures

Measure of reliability, Mean Error (or bias):

$$ME = \frac{1}{N}\sum_{i=1}^{N}(f_i - o_i)$$

where f = forecast, o = observation

Measure of accuracy, Mean Absolute Error:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|f_i - o_i|$$

MAE for a reference forecast:

$$MAE_{ref} = \frac{1}{N}\sum_{i=1}^{N}|ref_i - o_i|$$

where ref = reference forecast

Measure of skill: $SS_{MAE} = 1 - MAE / MAE_{ref}$

In the shorter, one-to-two day, forecast range persistence usually provides a better guidance than climatology (i.e. smaller MAE). Therefore, in order to avoid unrealistically "good" short-range forecasts, it is recommended to compute the mean absolute error for both persistence and climatology and then to select the better reference forecast, MAE_{ref} for computing the skill.

In addition, a simple informative and straightforward verification means is:

Error distribution chart (e.g. as a bar chart)

It does not provide a single quality measure but provides in an easy-to-interpret form information on how the forecast errors are distributed, giving feedback of the general forecast quality as well as of the possible biases. Temperature errors of two (five) degrees are quite commonly used as thresholds for "correct" ("false") forecasts but the thresholds should, naturally, be somehow related with the climatological and/or day-to-day variability of the observed temperature in the location/area being analysed.

Guidelines for the production and presentation of temperature verification results are shown in the examples of Figures 3.1 to 3.3 (the data are fictitious).

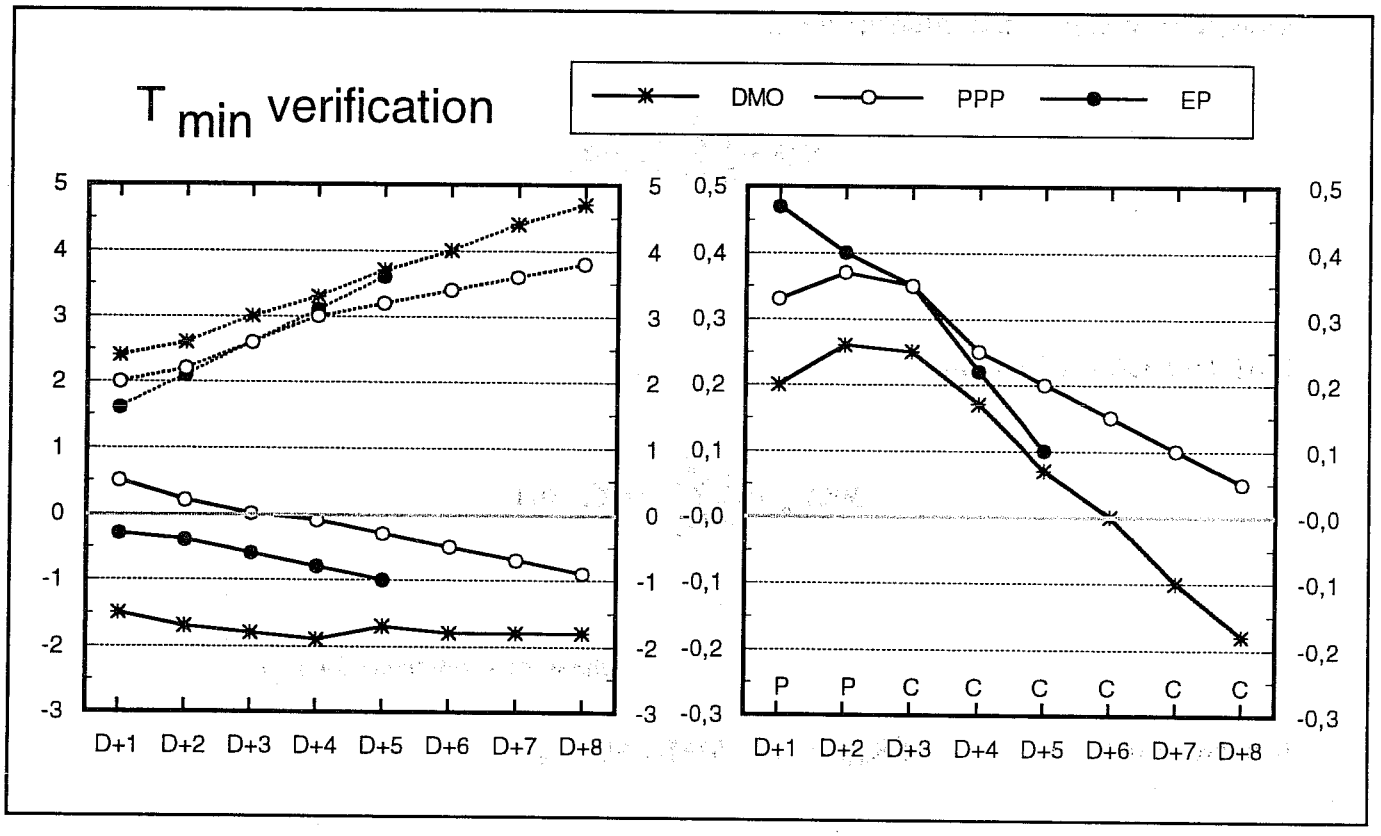

Figure 3.1 Mean Error (left, solid lines), Mean Absolute Error (left, dotted lines) and Skill (right) of one-to-eight day T (min) forecasts (DMO, PPP, EP) at station "xyz" averaged over months " abc" in 1993. DMO are temperature forecasts valid at 06UTC rather than T (min). EP extend only to D+5. The letters above lead times in the skill figure denote the reference forecast, persistence (P) or climatology (C)

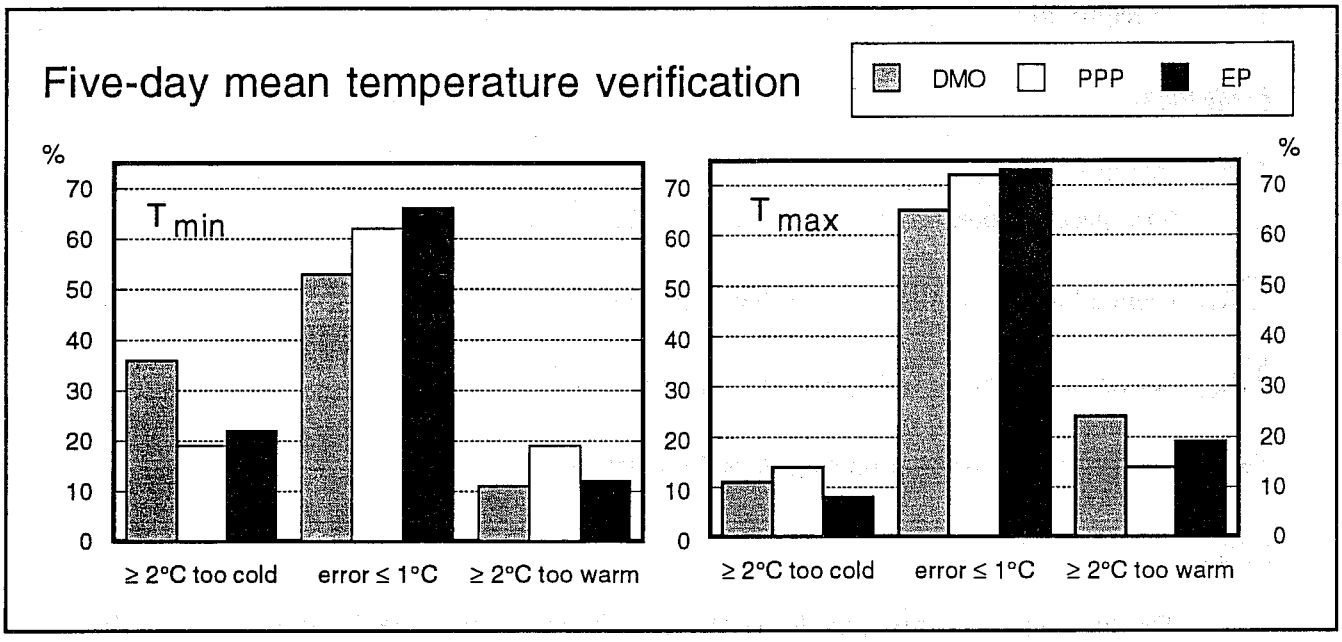

Figure 3.2 Error distributions of five-day mean minimum (left) and maximum (right) temperature forecasts (DMO, PPP, EP) at station xyz averaged over months abc in 1993.
DMO are temperature forecasts valid at 06UTC (for T_{min}) and at 18UTC (for T_{max})

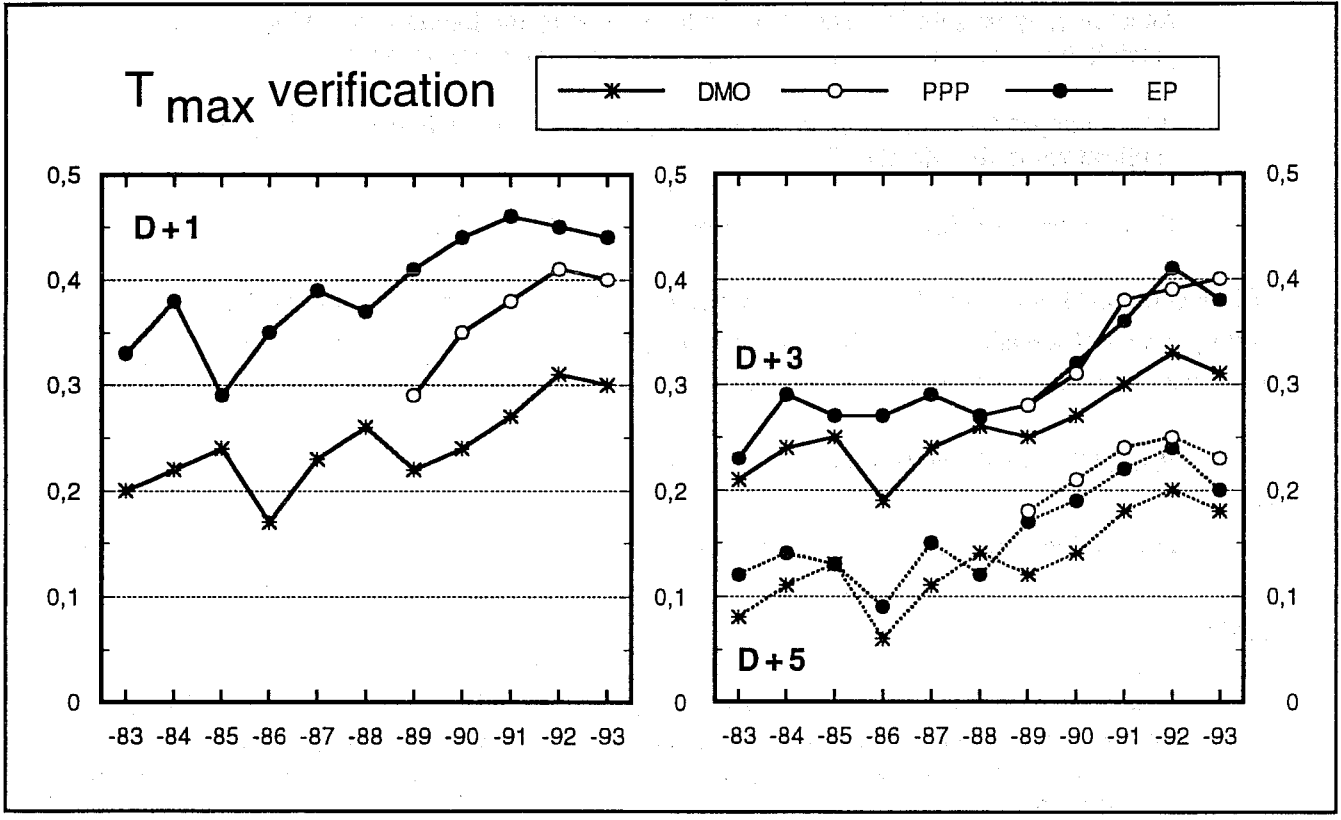

Figure 3.3 Skill of D+1 (left) and D+3/D+5 (right) T (max) forecasts (DMO, PPP, EP) averaged over stations "xyz", "yzx" and "zxy", 1983 through 1993. DMO are temperature forecasts valid at 18UTC rather than T (max). The PPP scheme was introduced in 1989. Persistence was used as reference at D+1, climatology at other lead times.

3.2 Precipitation

Predictands

$\sum R_{24}$ (00UTC - 24UTC)
total precipitation during D+1, D+2, D+3, D+4, D+5, ...

$\sum R_{1-3}$ **accumulated precipitation** over days 1 through 3

$\sum R_{4-6}$ **accumulated precipitation** over days 4 through 6

ΣR_{7-10} **accumulated precipitation** over days 7 through 10

Notes:

- The accumulation period can also be 06UTC - 06UTC or whatever is most convenient, but the lead time to the beginning of the first period (D+1) should be at least 12 hours
- Forecasts are easily available as DMO
- Daytime precipitation may be more appropriate from the users' point-of-view. The forecasts may be split into shorter periods, but due to the intermittent and local nature of precipitation, deterioration of results due to timing errors are apparent
- As for temperature, different accumulation periods may be defined according to applications in the Member States
- ΣR forecasts can simply be summed from individual daily ΣR_{24}

If it is considered unacceptable to verify point estimate forecasts of precipitation against corresponding observations, and to consider precipitation rather as an event, the following additional (or optional) forecasts are proposed:

PoP **Probability of Precipitation** (given in tenths),
or

alternative forecasts of, rain vs. no rain

Notes:

- It is up to individual Member States to define the threshold for the rain/no rain event. A threshold of 0.3 mm/24 hrs is recommended. Using 0.1 mm/24 hrs in forecasting a rain event will most likely lead to systematic errors since so small rainfall amounts can even be caused by fog.
- Alternative forecasts (rain vs. no rain) can be easily translated from absolute point estimate values, whereas vice versa is not possible.

Verification measures

Because of the sporadic and local character of precipitation, especially during the warm season, direct comparison between forecasted and observed total precipitation amounts is not sufficient to provide a full picture of the quality of precipitation forecasts. By extending the accumulation period over 24 hours, these effects are at least somewhat reduced.

The same reliability, accuracy and skill measures, **ME, MAE,** $\mathbf{SS_{MAE}}$, defined in the previous paragraph for temperature forecasts can be applied for the verification of point estimate forecasts of precipitation. **Error distribution diagrams** are useful and applicable.

If precipitation forecasts are formulated in probabilistic terms, corresponding measures for accuracy and skill can be defined as:

Measure of accuracy, (half) Brier Score:

$$BS = \frac{1}{N}\sum_{i=1} (p_i - o_i)^2$$

where p = forecast probability of a rain event
o = 1, if rain occurred, otherwise o = 0

BS for a reference forecast:

$$BS_{ref} = \frac{1}{N}\sum_{i=1}^{N} (rp_i - o_i)^2$$

where rp = climatological probability of a rain event
(also persistence, having explicit values = 1 or = 0, can be used as rp)

BS is the equivalent of the mean square error for forecasts given in probabilistic terms.

Measure of skill for PoP forecasts: $\mathbf{SS_{BS}} = 1 - BS / BS_{ref}$

As with temperature, the forecast to be used as reference in the early forecast ranges should be persistence or climatology, depending on which of them scores best.
The point estimate forecasts can be translated both into alternative forecasts, rain vs. no rain, and for defining the occurrence of more obscure events like heavy rain where the threshold may be set, e.g. at 10 mm/24 hrs. Verification should be done by accumulating the (N) events in a 2 * 2 contingency table, which can be shown in a simplified symbolic form:

		OBS		
		Event	no event	
FC	Event	A	B	A + B
	no event	C	D	C + D
		A + C	B + D	N

and by using the following verification measures:

Hit Rate (actually, probability of detection): **HIR** = A/(A+C)

False Alarm Rate: **FAR** = B/(A +B) (= 1 - A/(A+B))

The hit rate is, by definition, an overall measure of the accuracy of forecasts over all N events (i.e. including "hits" of the complement event, D/(B+D), but it is more informative just to evaluate the event of interest (here, occurrence of rain or heavy rain).

Especially in the verification of rare and extreme events these measures provide valuable feedback which can be computed very simply. Together they provide information on the (partial) accuracy and reliability of the alternative forecasts:

	>	systematic overforecasting of the event
HIR + FAR	= 1	no bias
	<	systematic underforecasting of the event

Guidelines for the production and presentation of precipitation verification results are shown in the examples of Figures 3.4 (the data are again fictitious).

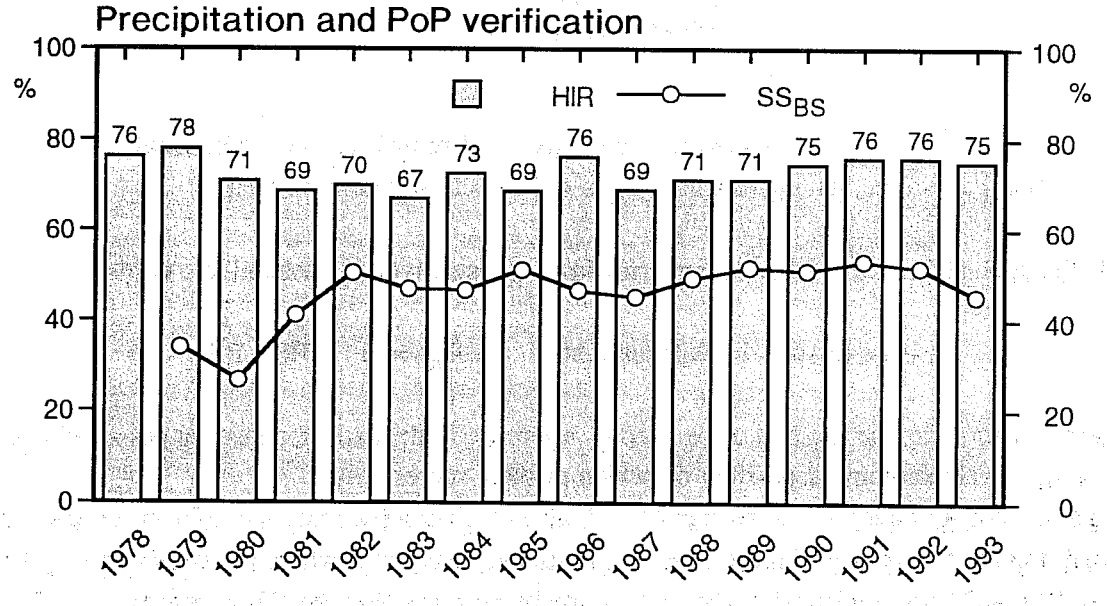

Figure 3.4 Hit rates (HIR, shaded bars) of precipitation event (≥ 0.3 mm/24 hrs) EP forecasts and the Brier Skill score (SS, solid line) of probability of precipitation (PoP) EP forecasts at station "xyz", 1978 through 1993. Lead time is D+1.

3.3 Wind speed

Predictands

$\mathbf{FF_{24}}$ (18UTC - 18UTC)

or

$\mathbf{FF_{12}}$ (06UTC - 18UTC)

maximum wind speed during D+1, D+2, D+3, D+4, D+5, ...

Notes:

- The verifying observation is the highest measured (10 minute integrated) wind speed during the 24 (or 12) hour period.
- The verifying period can be other than those defined, provided that the lead time to the beginning of the first period (D+1) is at least 12 hours.
- Due to the difficulties in forecasting wind speed, the longest forecast ranges for EP are presumably seldom over D+3. However, DMO is easily available for longer lead times.
- The selection of the verifying observation(s) should be done carefully, e.g. noting the anemometer elevation at the station(s) etc.
- If warnings against winds exceeding certain defined thresholds (gale, storm winds) are being issued for specified ocean areas, they should be verified against a representative set of stations in the area and by selecting the highest of all observations as the verifying value.

Verification measures

The same reliability, accuracy and skill measures, **ME, MAE,** $\mathbf{SS_{MAE}}$, defined in Section 3.1 can be applied for the verification of point estimate forecasts of wind speed. **Error distribution diagrams** are applicable.

Wind warnings should be verified with the alternative forecast approach by using the measures presented in the context of alternative precipitation forecasts, i.e. the hit rate, **HIR**, and false alarm rate, **FAR**.

If probabilistic methods are utilized in wind speed forecasting, or for the production of wind warnings, then the methods presented for the verification of PoP forecasts can be applied, i.e. **BS** and $\mathbf{SS_{BS}}$.

Figure 3.5 provides an example (now, based on real data) of the verification of warnings against gale winds.

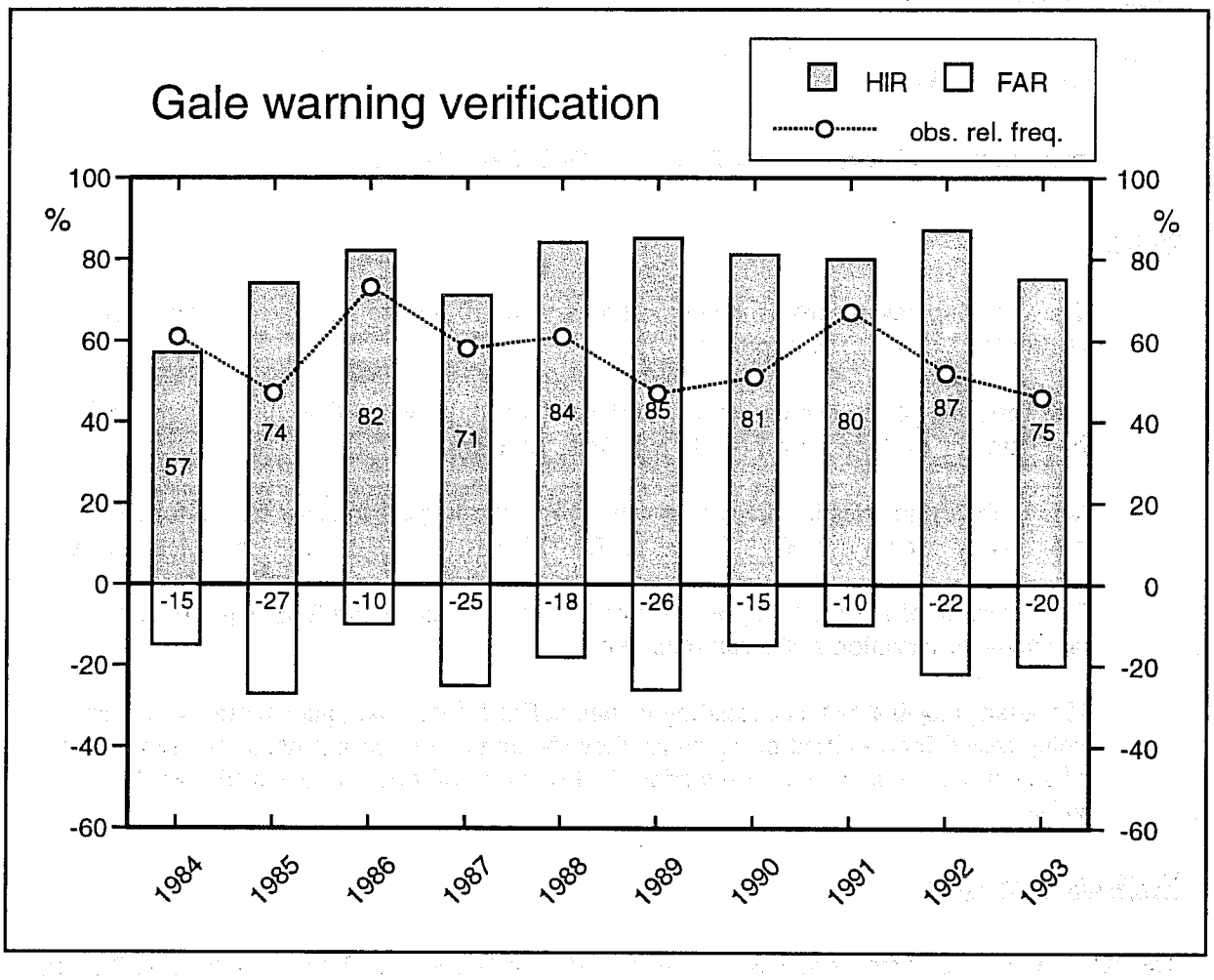

Figure 3.5 Hit rates (HIR, positive shaded bars) and false alarm rates (FAR, negative open bars) of gale wind warnings averaged over months September-December, 1984 through 1993, in the Archipelago Sea area. Lead time is D+1. The dotted curve shows the observed relative frequency of gale winds in the area. The maximum observed wind speed from three stations in the area during a 24 hour period was selected as the verifying observation.

References

Daan, H., 1984. Scoring Rules in Forecast Verification. WMO PSMP Report No. 4

Murphy, A.H., 1985. Proposed Standard Procedures for Verification of Local Weather Forecasts. WMO PSMP Report No. 15