METEOROLOGICAL APPLICATIONS
Meteorol. Appl. **15**: 113–123 (2008)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/met.62

RMetS
ROYAL METEOROLOGICAL SOCIETY

Terminal aerodrome forecast verification in Austro Control using time windows and ranges of forecast conditions

Guenter Mahringer*
Guenter Mahringer, Austro Control, MET Office Linz, Flughafenstrasse 1, A-4063 Hoersching, Austria

ABSTRACT: Terminal aerodrome forecasts (TAFs) are widely used meteorological forecasts for flight planning. Therefore, there is considerable interest in assessing their accuracy, skill and value. TAFs give information about the expected conditions of wind, visibility, significant weather and clouds at airports. Using different types of change groups, the forecaster gives a range of possible values valid for a time interval, the shortest interval being 1 h. A TAF thus contains a range of forecast conditions for each hour.

Point verification has proved to be difficult for TAFs. To ease these difficulties, time and meteorological state constraints are relaxed in the method described in this paper. This is done by verifying two conditions for each hour of the TAF. The highest (or most favourable) observed value is used to score the highest forecast value, and the lowest (or most adverse) observed value is used to score the lowest forecast value. Entries are made accordingly into two contingency tables. The contingency tables are specific for weather element and lead time.

Verification results should give feedback to forecasters. Contingency tables show the strengths and weaknesses of TAF, and displays for individual TAFs are available in the sense of 'eyeball verification'. For management information, common verification measures for categorical events (such as the Gerrity Score and the Heidke Skill Score) are calculated from the contingency tables. For answering specific customer questions, verification results in respect to certain values of weather elements are available. Copyright © 2008 Royal Meteorological Society

KEY WORDS forecast quality; aeronautical meteorology; verification method; contingency tables

Received 14 September 2007; Revised 2 January 2008; Accepted 18 January 2008

1. Introduction

Terminal aerodrome forecasts (TAFs) are widely used meteorological forecasts for flight planning. Therefore, there is considerable interest in assessing their accuracy, skill and value. TAF verification methods and systems have been established by many aviation weather services mainly since the early 1980s. However, until now there have been many different approaches, and all attempts to establish a commonly accepted method have so far been successful only for limited regions.

The main areas of differences in approaches concern observational data used for verification, the treatment of change groups, the verification of meteorological elements and the scores used to display the results.

For a long time, Meteorological Aviation Routine Weather Reports (METARs) were the only data source available and used for verification (Balzer, 1995; Harris, 1998). However, several systems nowadays also include SPECIs (Fuller, 2003; Kluepfel, 2005). Continuously measured sensor data are until now hardly used for TAF verification.

The verification of weather elements is somehow confused by contradicting requirements of International Civil Aviation Organization (ICAO) Annex 3, 'Meteorological Services for International Air Navigation' (ICAO, 2004). TAF amendment criteria given in ICAO Annex 3 are thresholds, e.g. for visibility and ceiling ('ceiling' refers to a layer cloud cover $>4/8$ or obscured sky with a vertical visibility observed or forecast). On the other hand, Appendix B of Annex 3 contains quality criteria ('operationally desirable accuracy of forecasts') based on absolute and/or relative deviations between observed and forecast values.

As flight planning is essentially based on thresholds, categorical verification schemes are more commonly used. Balzer (1995) uses a mixed system verifying wind speed and wind vector as continuous variables, whereas for wind gusts, visibility, ceiling and present weather a categorical verification is made. Kluepfel (2005) uses thresholds for all elements, even for wind direction, for which Annex 3 (ICAO, 2004) gives amendment criteria based on the deviation between forecast and observation. For present weather, there are different approaches of grouping the phenomena, and thereafter verifying the groups by using 2-category contingency tables (e.g. Kluepfel, 2005). Combined criteria for visibility and ceiling are described in Kluepfel (2005) and Fuller (2003). However, as TAF forecasts do not contain forecasts of

* Correspondence to: Guenter Mahringer, Austro Control, MET Office Linz, Flughafenstrasse 1, A-4063 Hoersching, Austria.
E-mail: guenter.mahringer@austrocontrol.at

Copyright © 2008 Royal Meteorological Society

runway visual range (RVR), a verification of instrument flight approach categories is not exactly possible.

Verification schemes evaluating the percentage of correct forecasts based on ICAO Annex 3 Attachment B are less widespread.

Most controversial is the verification scheme for TAF change groups. These are used to forecast transitions or temporary changes in values or states of weather elements. Transitions at a given time are forecast using a 'from' (FM) statement (followed by the time of the change). For forecasting transitions within a time interval, 'becoming' (BECMG) is used, followed by the time interval in which the change is forecast to occur. Temporary changes are forecast by 'temporarily' (TEMPO), followed by the time interval within which these changes are forecast to occur. 'Probably' (PROB) is used to forecast alternative conditions that will occur with a certain probability during a specified time period, followed by the forecast probability (only 30 and 40% are allowed). PROB may also be used in combination with TEMPO when forecasting temporary changes with a certain probability.

Gordon (1993) states that one cannot directly compare observed conditions at a single time with what the forecast from the TAF was (remark: because there is more than one forecast state valid for many points of time in a TAF). A more complex approach is required. Gordon suggests checking, for blocks of time (such as three hours), the worst (minimum) forecast conditions against the worst observed conditions. This approach is followed by the NORTAF verification scheme (Hilden *et al*., 1996, 1998). Many other schemes assign probabilities for the conditions forecast by PROB, PROB TEMPO and TEMPO groups (Balzer, 1995; Harris, 1998; Fuller, 2003). Kluepfel (2005) defines an Operational Impact Forecast as the forecast in effect that is most likely to have the largest impact on flight operations, and uses a very complex approach to verify TEMPO groups by investigating the variability within a time interval of ± 90 min from the observation. However, such ideas tend to be difficult to understand for forecast users, if not even for forecasters.

For BECMG groups, a transition in probabilities from the beginning towards the end of the period has been used by the UK Met Office (Harris, 1998). Balzer (1995) and Fuller (2003) regard a forecast correct as long as the observed value lies within the range opened by the BECMG group. However, this encourages forecasters to use BECMG groups excessively to improve scores.

All the verification systems discussed compare one forecast state with one observed state. The idea of this paper is to understand a TAF as a forecast of a range of possible conditions within defined time intervals. As will be shown in the following section, such an approach is able to overcome many of the problems coming along with the idea of verification based on single observations.

2. TAF verification method

2.1. Principles

Three different types of applications are seen for TAF verification:

- Management information: how good are the forecasts in an overall view? How is forecast quality developing over the years? This approach is also used for deriving the figures necessary in the framework of a quality management system.
- Forecaster feedback: where are the strengths and weaknesses of our TAFs? Where should we most urgently try to improve? How did yesterday's TAF perform (individual verification)?
- Customer information: how can customers make the best use of the TAFs in their planning procedures?

There are two basic principles for the Austro Control TAF verification. These arise from considerations of what a TAF really is, how it is produced by the forecaster and how it is used by the customer.

Firstly, a TAF is considered a forecast for *time periods* rather than for points of time. This is due to the fact that, by using the TAF change groups BECMG, TEMPO, PROB and PROB TEMPO, changes within periods are forecast, the shortest meaningful interval being 1 h. Only the FM statement allows forecasting changes at a given point of time.

Secondly, a TAF is considered to contain a *range of forecast conditions* rather than a single state. All change groups except FM give alternative conditions for a certain time interval. Even if there is no change group valid for a certain hour, the condition stated is typical for a range (which is assumed to have a uniform effect on flight operations) delimited by thresholds like the Amendment criteria contained in ICAO Annex 3.

To evaluate the correctness of a forecast range, the highest (or most favourable) and lowest (or most adverse) conditions valid for each hour of the TAF are taken for verification. For this purpose, all observations within the respective hour are used (METAR and SPECI), which span a range of observed conditions. So, for each hour, two comparisons are made: the highest observed value is used to score the highest forecast value and the lowest observed value is used to score the lowest forecast value.

The '*range of forecast conditions*' approach avoids the need of assumptions about probabilities for conditions forecast by TEMPO and PROB TEMPO, or ambiguous conditions during a BECMG period. The '*time period*' approach allows a range of observed conditions rather than a single observation to be compared with a range of forecast conditions. For customer-oriented verification, it is possible to exclude the change groups PROB, TEMPO and/or PROB TEMPO from verification. This accounts for operational procedures applied by many TAF users. Furthermore, the effect of certain change groups on TAF quality can be studied by using this option.

Copyright © 2008 Royal Meteorological Society

DOI: 10.1002/met

The TAF verification is based on half-hourly METARs. Additionally, SPECIs are used. Variations between METAR observations are additionally verified by taking RE (recent) and VC (vicinity) groups into account. It should be noted that verification without SPECIs is not adequate for a continuous forecast because short-term variations are systematically underrepresented in the observation set. For every hour, at least two representative observations are required. The first METAR to be used for an hour is the one at or shortly before the start of the hour. All METARS observed within the hour are also used, as well as all SPECIs within the period from the first METAR to the end of the hour.

As airline planning procedures require a TAF, there is no obvious purpose in investigating the quality of the TAF in comparison with any reference forecast such as climate or persistency. However, the verification method can be used for comparisons between TAFs, AUTOTAFs and other forecast guidances.

TAFs are also checked for syntax errors and for the occurrence of more than one change group of the same type with overlapping validity. Such incorrectly coded TAFs (WMO, 1995) are counted and their percentage is one of the results, but they are excluded from the verification of TAF quality.

2.2. The verification of weather elements

The four weather elements wind, visibility, significant weather and ceiling (defined as a layer cloud cover $>4/8$ or obscured sky with vertical visibility (VV) stated in METAR or TAF) are verified separately. They are verified according to the Amendment criteria given by Annex 3 (ICAO, 2004), with the possibility of additionally taking local regulations into account. The recommendations concerning the 'operationally desirable accuracy of forecasts' given in Attachment B of Annex 3 are in general not used for verification. It should be noted in this context that any known operational planning procedure uses TAFs in respect to threshold values. Therefore, the needs of customers are better reflected by a method based on threshold values.

With the exception of wind direction, for each element and hour, entries into two contingency tables are calculated (Jolliffe and Stephenson, 2003). For this purpose, the values of the elements (observed and forecast) are transformed into classes according to predefined criteria. The first table corresponds to the highest observed/forecast value of the weather element within an hour; the second table corresponds to the lowest observed/forecast value of the weather element within an hour. The contingency tables are specific for the considered weather element and the lead time (this is the time difference between the time when the TAF is issued and the hour considered). A summarizing table for all lead times is set up for each weather element.

For visibility, ICAO Annex 3 contains the thresholds 150, 350, 600, 800, 1500, 3000 and 5000 m. If more than one visibility value appears in a METAR, the first value is used.

The verification of clouds is carried out in respect to the ceiling. ICAO Annex 3 contains ceiling thresholds of 100, 200, 500, 1000 and 1500 feet above ground. If more than one ceiling value appears in a METAR/TAF, the lowest one is taken into account. If no height is given after the cloud amount statement, no verification is possible. If there is no ceiling observed/forecast, the class of a ceiling higher than the highest threshold value applies. The existence of towering cumulus (TCU) or cumulonimbus (CB) is not verified. This is due to the fact that these clouds often appear in METAR observations even when they are far from the airport, especially in situations with very good visibility.

Present weather is observed/forecast according to the code tables in the Manual on Codes (WMO, 1995). For TAFs, ICAO Annex 3 indicates the following weather phenomena to be relevant:

- Freezing fog
- Freezing precipitation (intensity is relevant)
- Moderate or heavy (showers of) precipitation (drizzle and rain, snow, hail)
- Low drifting dust, sand or snow
- Blowing dust, sand or snow (intensity is not relevant)
- Dust storm or sandstorm (intensity is relevant)
- Thunderstorm (with or without precipitation)
- Squall
- Funnel cloud

To ensure proper statistical scores, the phenomena may be grouped in classes containing "similar" events (e.g. by not verifying different intensities). For Austrian airports, the classes used are shown in Table I. The verification

Table I. Present weather classes used for Austrian airports.

Class no.	Class name	Weather phenomena
0	NSW	No significant weather, all phenomena not appearing below
1	FZFG	Freezing fog (patches)
2	RA	Moderate or heavy drizzle and rain (showers) including combinations with other types of precipitation
3	BLSN	Drifting or blowing snow
4	SN	Moderate or heavy (showers of) snow and hail including combinations with other types of precipitation
5	FZRA	Freezing rain, freezing drizzle
6	TS	Thunderstorm, squall line, funnel cloud

Copyright © 2008 Royal Meteorological Society

DOI: 10.1002/met

procedure itself is similar to the visibility and ceiling verification. If more than one group of present weather phenomena is reported in one METAR/TAF, the highest relevant class number is taken for verification (e.g. if a METAR contains rain and snow (RASN) and mist (BR), only RASN is verified).

When verifying surface wind forecasts, one has to bear in mind that wind is a very variable element. METAR observations containing 10 min mean wind direction and speed measurements often do not reflect the complete behaviour of wind in a time interval. For this purpose, an analysis of continuous sensor data would be necessary, which requires data that are usually not available.

For wind direction verification, significant deviations between observations and forecasts are investigated. Forecast errors are only regarded significant when the wind speed reaches a certain value (in Austro Control, a threshold of 7 kt is used) and the deviation is greater than a given limit (in Austro Control, a threshold of 30° is used). If the observed mean wind speed is greater or equal than a pre-defined speed threshold, all differences between the observed wind direction and the forecast wind directions valid for the respective hour are calculated. A VRB ('variable') forecast is assigned a difference of 180°. The smallest difference is taken for verification. If it is smaller than the direction threshold, the forecast direction is considered correct. If the observed mean wind speed is smaller than the speed threshold, the forecast wind direction is considered correct because any direction deviation is operationally insignificant, and the forecaster was not requested to give a change group in this case.

Wind speed is verified using contingency tables. In accordance with ICAO Annex 3, Appendix 5, Austro Control MET and Air Traffic Management (ATM) agreed on operationally significant wind speed thresholds: 7, 15, 25, 35, 45 and 55 kt. Wind speed is then verified as described for visibility. As the METAR wind observations only reflect the conditions in one-third of the time, they will in many cases not include the extreme values having appeared within any period of time. Therefore, one must expect the range of forecast values to be larger than the variation in observed values.

The criterion for forecasting wind gusts in TAFs is the expectation that the gustiness (short term variability of wind speed) will be $\geq$10 kt. Without continuous data or suitable SPECIs, the verification of wind gusts is not meaningful. In Austro Control, hourly maximum wind speed values are used. The verification of wind gusts is done using classes delimited by thresholds 30 and 45 kt. Only the maximum observed/forecast value is investigated.

All thresholds and criteria mentioned in this section can easily be adapted to local requirements where necessary.

2.3. Treatment of change groups

The Austro Control TAF verification method takes the best and worst value forecast in a TAF for a certain hour for verification, regardless of the kind of change group but with the possibility of disregarding certain change groups.

As long as no change groups appear, the highest and lowest observed values are verified against the basic forecast, and two entries in contingency tables (one for 'best observed condition', one for 'worst observed condition') are calculated.

The statement FM refers to a sudden change in the weather element from state 1 to state 2 at the point of time GGgg (GG stands for the hour, gg for the minute of the change). The values valid before the FM statement are taken to verify the hour(s) before the hour GG. Both the values before and after the FM statement are used to verify the hour GG to GG + 1 in the sense of a range. The values valid after the FM statement are taken to verify the hour(s) after GG + 1. The minute of the change gg is disregarded in verification.

Together with the basic forecast, the change groups BECMG, TEMPO, PROB and PROB TEMPO give at least two different forecast conditions. For each hour, the highest/lowest forecast values are verified against the highest/lowest observed values for each weather element. If more than one change group is valid within an hour, the highest/lowest forecast value of all valid groups is determined and used for verification. For BECMG groups, no assumption about the change mode (regular or irregular) within the period is necessary. However, BECMG groups over long time intervals and/or a wide range of values are less likely to get good results.

For TEMPO groups, the method ensures that too long TEMPO groups and TEMPO groups forecasting changes that in fact last for much longer than an hour get worse scores.

Note that the method fails to distinguish between TEMPO, PROB30, PROB40, PROB30 TEMPO and PROB40 TEMPO, because the values forecast by all these groups are used to determine the highest/lowest forecast value. However, an airline operator cannot do much more than to account for or disregard certain groups in TAFs. In practice, the use of these groups often tells more about the forecast phenomenon than about the kind of expectation. For example, convective phenomena are likely to be forecast with TEMPO or PROB TEMPO; fog would rather be forecast with a PROB statement if not with BECMG or FM. By excluding TEMPO, PROB and/or PROB TEMPO groups from verification, one can determine the consequences of disregarding these groups in aircraft operation planning procedures.

2.4. Example

An example explaining the method described in the previous sections is shown in Table II. Only visibility is displayed, SPECI observations are not shown. Note that observations at H + 50 are also representative for the (beginning of) the following hour. Shaded cells in Table II indicate visibility categories that were forecast and/or observed according to the legend. This display

Copyright © 2008 Royal Meteorological Society

DOI: 10.1002/met

Table II. TAF verification example for visibility.

TAF 0615	0700 TEMPO 0609 0200 BECMG 0911 4000 FM1200 9999

Time (UTC)	06–07	07–08	08–09	09–10	10–11	11–12	12–13	13–14	14–15
OBS H+20	1800	0100	0500	0300	1700	3500	8000	9999	6000
OBS H+50	0300	0400	0400	1000	2300	6000	9999	9999	0300
Forecast and observed VIS classes									
TIME (UTC) Visibility (m)	06–07	07–08	08–09	09–10	10–11	11–12	12–13	13–14	14–15
5000 – 9999									
3000 – <5000									
1500 – <3000									
0800 – <1500									
0600 – <0800									
0350 – <0600									
0150 – <0350									
0000 – <0150									

Not forecast and not observed
Not forecast but observed
Forecast and observed
Forecast but not observed

form is also used for individual TAF evaluation, which is available for the forecaster in the sense of 'eyeball verification'.

For scoring, maximum and minimum forecast and observed values are taken. For instance, looking at the hour 0700–0800 UTC, visibility is forecast to be within 700 m (prevailing condition) and 200 m (TEMPO condition). The observed values are 100 m and 400 m. Therefore, the maximum forecast value is 700 m, the maximum observed value is 400 m; an entry is inserted into the contingency table for maximum values (Table III(a)) accordingly. The minimum forecast value is 200 m, the minimum observed value is 100 m; an entry is inserted into the contingency table for minimum values (Table III(b)) accordingly. The same is repeated for each hour.

3. Verification measures

When evaluating a time series of TAFs, the verification system generates the following:

- Two contingency tables for the highest and the lowest forecast/observed value
- For the elements visibility, ceiling, weather and wind speed
- For each hour of TAF validity, plus a summarizing table.

Depending on the scope of the investigation, many different conclusions and scores can be derived from this detailed set of information.

3.1. Overall scores

3.1.1. Scores for 2-category contingency tables

From 2-category contingency tables (Table IV), measures are calculated according to Table V. These scores all contain relevant information, but all have deficiencies regarding their ability to contain all aspects of forecast quality in a single number. Jolliffe and Stephenson (2003) show that none of these and other known scores show all the desired properties of a verification score. However, the odds ratio skill score (ORSS) is the only one of them that is not depending on the probability of event and therefore preferable to the other scores. This property makes comparisons of forecasts in different climates easier, and therefore it is used in the Austro Control TAF verification system. Furthermore, this measure has the property to be very sensitive to the quality of rare event forecasts, which is desirable for aviation forecasts. Problems only arise with small sample sizes. They are met by determining confidence intervals for the ORSS.

For making comparisons with other systems easier, the more widely used HSS is used as a second reference.

In Austro Control TAF verification, ORSS and HSS are used for present weather, where each class of phenomenon is scored separately. For this, the 7-category table is reduced to six 2-category contingency tables containing the information in respect to each class.

For wind direction, the percentage of correct forecasts is calculated. As additional information, the fraction of

Copyright © 2008 Royal Meteorological Society

DOI: 10.1002/met

Table III. Contingency tables for the TAF example of Table II. The diagonal is shaded.

(a) Maximum values forecast-/observed for each hour of TAF validity

VIS class FCST/OBS	0000– <0150	0150– <0350	0350– <0600	0600– <0800	0800– <1500	1500– <3000	3000– <5000	5000– 9999
0000–<0150								
0150–<0350								
0350–<0600								
0600–<0800			2			1		
0800–<1500								
1500–<3000								
3000–<5000					1	1		1
5000–9999								3

(b) Minimum values forecast/observed for each hour of TAF validity

VIS class FCST/OBS	0000– <0150	0150– <0350	0350– <0600	0600– <0800	0800– <1500	1500– <3000	3000– <5000	5000– 9999
0000–<0150								
0150–<0350	1	1	1					
0350–<0600								
0600–<0800		1			1			
0800–<1500								
1500–<3000								
3000–<5000						1		1
5000–9999		1						1

Table IV. Schematic 2-category contingency table (from Jolliffe and Stephenson, 2003).

Forecast	Observed		
	Yes	No	Total
Yes	a	b	$a+b$
No	c	d	$c+d$
Total	$a+c$	$b+d$	$n=a+b+c+d$

Cell meanings:
Hit a
False alarm b
Miss c
Correct rejection d.

cases with mean wind speeds above the threshold for the investigation is provided.

3.1.2. Scores for n-category contingency tables

Jolliffe and Stephenson (2003) describe that the generalized forms of the Heidke Skill Score (HSS) and the Peirce's Skill Score (PSS) can be used for multi-category forecasts. However, they have some severe deficiencies: they are dependent on the forecast distribution; they fail to draw adequate attention to correct forecasts of rare events; and they do not use the off-diagonal information of the n-category contingency tables.

These deficiencies can be overcome by a type of scores called Gandin and Murphy Equitable Scores. They involve equitable scoring matrices, which consist of rewarding or penalizing factors for every cell of the n-category contingency table. The factors are calculated according to a defined set of rules.

A subset of these scores is the Gerrity Score (GS), for which the scoring matrix is derived based on the observed frequencies of classes. This ensures that correct forecasts of rare events get high rewards compared to correct forecasts of very common events. This score can alternatively be calculated by averaging the PSS values for all 2-category contingency tables which are created from the n-category table using the threshold-related events.

The GS is therefore used for producing overall measures of forecast quality for the TAF elements visibility, ceiling and wind speed.

Examples of results are shown in Section 4.

3.2. Scores for gaining forecaster feedback

Forecaster feedback is most efficient when a forecaster has the possibility to look at his/her TAFs on an individual basis shortly after the event. For this, the verification can be done for single TAFs. Forecast and observed conditions are displayed like in Table II. This display shows the times and values of agreement as well as disagreement of forecast and observed conditions.

For longer periods, internal investigations of strengths and weaknesses of TAFs can be done on the basis of

Copyright © 2008 Royal Meteorological Society

DOI: 10.1002/met

Table V. Verification measures used in Austro Control TAF verification (after Jolliffe and Stephenson, 2003).

Name of measure	Definition
Probability of event (base rate)	$p(E) = (a + c)/n$
Bias	$\text{Bias} = (a + b)/(a + c)$
Hit rate (probability of detection)	$\text{H} = \text{POD} = a/(a + c)$
Proportion correct	$\text{PC} = (a + d)/n$
False alarm ratio	$\text{FAR} = b/(a + b)$
False alarm rate	$\text{F} = b/(b + d)$
Conditional probability of an event, given:	
The event was forecast	$p(E)$ when $\text{FCST} = a/(a + b)$
The event was not forecast	$p(\text{E})$ when not $\text{FCST} = c/(c + d)$
Heidke Skill Score (HSS) with E = PC for random forecasts	$\text{HSS} = (a + d - E)/(1 - E)$
	$\text{E} = ((a + b) \times (a + c) + (b + d) \times (c + d))/n$
Peirce's Skill Score (PSS)	$\text{PSS} = \text{H} - \text{F} = (a \times d - b \times c)/(a + c) \times (b + d)$
Critical success index (CSI) = Threat score (THS)	$\text{CSI} = \text{THS} = a/(a + b + c)$
Odds ratio skill score ORSS (Yule's Q)	$\text{ORSS} = (a \times d - b \times c)/(a \times d + b \times c)$

the contingency tables themselves. Valuable insight can especially be gained from looking at different lead times. From the table entries, one can see for which thresholds or phenomena the forecast quality is good or where it should be improved.

3.3. Verification results for customers

For individual forecast users, usually not all the threshold values will be of the same importance. Therefore, for investigating the TAF quality in respect to their needs, the n-category contingency table is reduced to 2-category tables in respect to certain thresholds.

For users with little experience with statistical scores, figures that can easily be understood and explained are selected. For example, the importance of events can be explained using p(E). The ability of the forecasts to discriminate between occurrence and non-occurrence of the event can be shown using the conditional probabilities *p(E) when forecast* and *p(E) when not forecast*. Another possibility is to show a combination of hit rate/false alarm ratio (H = probability of detection (POD)/FAR). However, it must be kept in mind that FAR values tend to be rather high with time intervals of only 1 h.

4. Results

TAF quality is measured separately for the weather elements visibility, ceiling, present weather, wind direction and wind speed. Contingency tables, overall scores and event-related scores are shown. The investigation period covers 3 months (September to November, 2006) and all 9-h TAFs are issued 3-hourly by the local MET Offices. Amendments are not considered.

Since this article concentrates on demonstrating the verification method, the presentation of results focuses on one airport (Graz, LOWG) and one parameter (visibility) only. For the other weather elements and all other Austrian airports, similar tables and figures have been compiled, which follow the same method and use the scores as presented in Section 3.

4.1. Contingency tables

Many conclusions about forecast quality can be drawn from looking at the contingency tables resulting from the verification algorithm. Owing to the use of the *forecast range* principle, we get a pair of n-category contingency tables for each investigation, one for the upper limit of the range and the other for the lower limit.

As an example, Table VI shows the contingency tables for visibility at LOWG (Graz Airport).

The contingency tables show observed and forecast values clustered at the bottom and the top of the range. Low visibilities are somewhat underforecast in respect to maximum values (Table VI(a)), but slightly overforecast in respect to minimum values (Table VI(b)), which means that forecast ranges usually tend to be too large, and fog events are forecast for slightly longer time periods than observed. The tables do not show if fog formation is forecast too early or too frequently, or if fog dissolution is forecast too late. However, apparently there is very good skill in catching dense fog events as there are relatively few missed events (Table VI(b)).

Slight visibility reductions appear to be overforecast in the 'minimum visibility' table (Table VI(b)). The reason for this is probably that temporary visibility reductions associated with precipitation events happen less frequently or for shorter periods than forecast.

Low maximum visibility often fails to be correctly forecast (Table VI(a)). This may be due to either long BECMG groups or the use of TEMPO, PROB and PROB TEMPO together with good visibility in the basic forecast, where the visibility reduction lasts for longer than one hour.

4.2. Overall scores from contingency tables

Table VII shows scores derived from the contingency tables discussed above. Looking at the diagonal, the forecasts of maximum visibility look better as the relative frequencies sum up to 0.812 compared to 0.695 for minimum visibility. Nevertheless, all scores are higher

Copyright © 2008 Royal Meteorological Society

DOI: 10.1002/met

Table VI. Contingency tables for TAF visibility forecasts at Graz (LOWG) for 1 September 2006–30 November 2006, all lead times.

(a) Maximum values forecast/observed for each hour of TAF validity

FCST/OBS	0–149 m	150–349 m	350–599 m	600–799 m	800–1499 m	1500–3499 m	3500–4999 m	5000–9999 m	Sum
0–149 m	13	4	3	0	7	4	5	6	42
150–349 m	29	22	2	0	8	5	1	10	77
350–599 m	8	16	1	0	1	3	6	4	39
600–799 m	0	0	1	0	1	1	0	1	4
800–1499 m	2	2	5	1	6	7	5	18	46
1500–3499 m	9	17	13	2	3	19	13	30	106
3500–4999 m	26	39	4	0	23	51	81	123	347
5000–9999 m	15	29	6	0	22	68	155	3374	3669
Sum	102	129	35	3	71	158	266	3566	4330

(b) Minimum values forecast/observed for each hour of TAF validity

FCS/OBS	0–149 m	150–349 m	350–599 m	600–799 m	800–1499 m	1500–3499 m	3500–4999 m	5000–9999 m	Sum
0–149 m	139	93	20	4	24	38	20	44	382
150–349 m	43	62	8	2	17	47	16	127	322
350–599 m	3	4	3	0	7	9	3	21	50
600–799 m	0	1	0	0	0	0	1	5	7
800–1499 m	2	6	1	2	6	24	24	26	91
1500–3499 m	1	11	6	2	12	56	78	113	279
3500–4999 m	7	18	9	0	6	38	74	253	405
5000–9999 m	9	8	4	0	12	32	59	2670	2794
Sum	204	203	51	10	84	244	275	3259	4330

for the forecasts of minimum visibility, the difference being smallest for the HSS and largest for the GS. As mentioned in Section 3.1.2, the GS rewards correct or almost correct forecasts of rare events by high weights in the scoring matrix (not shown). On the other hand, hits in the most frequent class of visibilities of $\geq$5000 m are rewarded much less. The HSS and the Peirce Skill Score (PSS) are putting some extra weight on rare events by the subtraction of hits 'by chance', but much less than the GS. As they are not using the off-diagonal information, they, for example, account much less for the relatively frequent cases in the lower left corner of Table VI(a). Therefore, the use of the GS is preferred.

A bias towards forecasting high maximum and low minimum values can be seen from the fact that observed values higher than maximum forecast and lower than minimum forecast values are rare as compared to the opposite. TAFs more often tend to include possible scenarios that fail to appear then, than to missing certain developments. This is especially valid for minimum values, which are frequently forecast lower than observed.

Figure 1 shows the dependence of the GS values on lead time. There is a moderate overall drop in forecast quality between the first and last hour of the 9-hour TAF, irregularities may be due to the sample size (eight TAFs per day over 3 months for one airport).

Table VII. Scores derived from the contingency tables in Table VI.

(a) For maximum values forecast/observed	
Gerrity Score (GS)	0.349
Heidke Skill Score (HSS)	0.363
Peirce Skill Score (PSS)	0.341
FC max < OBS max	0.062
FC max = OBS max	0.812
FC max > OBS max	0.126
(b) For minimum values forecast/observed	
Gerrity Score (GS)	0.698
Heidke Skill Score (HSS)	0.386
Peirce Skill Score (PSS)	0.455
FC min < OBS min	0.236
FC min = OBS min	0.695
FC min > OBS min	0.068

4.3. Scores for events derived from the contingency tables

Events are defined in relation to certain thresholds related to flight operations. For example, if an operator uses a visibility of 600 m as his/her planning threshold, it is meaningful to investigate how well the events of visibility lower than 600 m were forecast, no matter how much

Copyright © 2008 Royal Meteorological Society

DOI: 10.1002/met

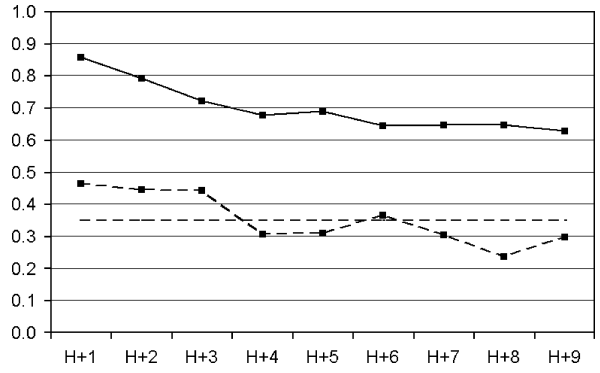

Figure 1. Gerrity scores (GS) as a function of lead time for TAF visibility at LOWG, September to November 2006. Bold lines: GS for minimum visibility forecast/observed for each hour. Dashed lines: GS for maximum visibility forecast/observed for each hour. Horizontal lines represent mean values, respectively.

lower or higher the actual and forecast values really were.

The n-category contingency table is reduced to a 2-category contingency table in respect to one threshold by summing up all cell entries of forecasts/observations below/above the threshold. Table VIII shows scores calculated from all 2-category tables derived from Table VI, again for maximum and minimum values for each hour, respectively.

In Table VIII, $p(E)$ indicates the relevance of the event considered, showing that low visibilities are relevant for flight operations in LOWG.

The hit rate H (or probability of detection POD) shows rather low ability to forecast low maximum values, but much better scores for minimum values. The false alarm ratio (FAR) is the number of 'yes' forecasts – 'no' observations relative to the number of all 'yes' forecasts, whereas the false alarm rate F relates the former to the number of all 'no' observations. For both maximum and minimum visibility, H values are generally lower and FAR values generally higher for lower thresholds, indicating that low visibility forecasting is a considerable challenge for forecasters.

The bias values confirm a moderate tendency to forecast larger ranges of values than observed as low maximum values are forecast too seldom, low minimum values too often.

Conditional probabilities show that TAF forecasts clearly have the ability to discriminate between events

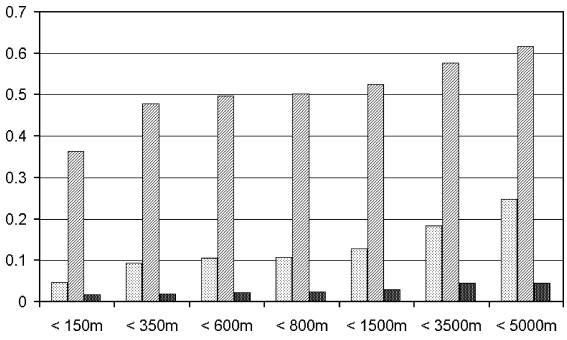

Figure 2. Probabilities and conditional probabilities for visibility events (minimum hourly visibility forecast/observed below threshold) at LOWG, September to November 2006: Stippled (left column): probability of event. Hatched (central column): probability of event when being forecast. Dark stippled (right column): probability of event when being not forecast.

Table VIII. TAF scores for 2-category tables related to thresholds derived from the contingency tables in Table VI.

(a) For maximum visibility values forecast/observed

Event: max vis	$p(E)$	H = POD	F	FAR	Bias	$p(E)$ when FCST	$p(E)$ when not FCST	PSS	ORSS	HSS
<150 m	0.024	0.127	0.007	0.690	0.412	0.310	0.021	0.121	0.910	0.169
<350 m	0.053	0.294	0.012	0.429	0.515	0.571	0.039	0.282	0.941	0.366
<600 m	0.061	0.368	0.015	0.380	0.594	0.620	0.040	0.354	0.950	0.436
<800 m	0.062	0.368	0.016	0.389	0.602	0.611	0.041	0.353	0.947	0.433
<1500 m	0.079	0.388	0.019	0.365	0.612	0.635	0.050	0.369	0.941	0.449
<3500 m	0.115	0.432	0.026	0.315	0.631	0.685	0.070	0.406	0.933	0.484
<5000 m	0.176	0.614	0.054	0.290	0.865	0.710	0.080	0.560	0.931	0.591
Mean								0.349	0.936	0.418

(b) For minimum visibility values forecast/observed

Event: min vis	$p(E)$	H = POD	F	FAR	Bias	$p(E)$ when FCST	$p(E)$ when not FCST	PSS	ORSS	HSS
<150 m	0.047	0.681	0.059	0.636	1.873	0.364	0.016	0.622	0.943	0.440
<350 m	0.094	0.828	0.094	0.521	1.730	0.479	0.019	0.734	0.958	0.553
<600 m	0.106	0.819	0.098	0.503	1.646	0.497	0.023	0.721	0.953	0.561
<800 m	0.108	0.816	0.098	0.498	1.626	0.502	0.024	0.718	0.952	0.563
<1500 m	0.127	0.810	0.107	0.475	1.543	0.525	0.030	0.703	0.945	0.570
<3500 m	0.184	0.820	0.135	0.423	1.421	0.577	0.045	0.685	0.934	0.589
<5000 m	0.247	0.884	0.181	0.383	1.434	0.617	0.044	0.703	0.944	0.614
Mean								0.698	0.947	0.556

Copyright © 2008 Royal Meteorological Society

DOI: 10.1002/met

and non-events (see also Figure 2). The expectation of an event is always much higher than p(E) when it is forecast and much lower when it is not forecast. The expectation of experiencing unexpected visibility conditions 'below minimum' can thus be reduced by a factor of 3–6 (depending on threshold) by observing the TAF forecasts of minimum visibility.

PSS, HSS and ORSS all show the smallest values for event 1 (<150 m), for the other thresholds, the behaviour is different. Scores are generally higher for minimum than for maximum visibility. Mean PSS values may be used as a simple workaround for calculating GS values (see Jolliffe and Stephenson, 2003, p. 91).

Confidence intervals for ORSS have been calculated according to Stephenson (2000). As the number of cases is relatively large in all classes, they are generally narrow, the widest 95% confidence interval has been found for maximum visibility <150 m ranging from 0.813 to 0.958 (where ORSS = 0.910).

The same type of presentation is used for ceiling, wind speed and present weather classes.

5. Conclusions

In Austro Control TAF verification, a TAF is considered a forecast for *time periods* rather than for points of time as changes within time periods are forecast by using change groups. Furthermore, a TAF is considered to contain a *range of forecast conditions* rather than a single state. All change groups (except FM) give alternative conditions for a certain time interval.

To evaluate the correctness of a forecast range, the highest (or most favourable) and lowest (or most adverse) conditions valid for each hour of the TAF are taken for verification. For this purpose, all observations within the respective hour are used (METAR and SPECI), which span a range of observed conditions. So, for each hour, two comparisons are made: the highest observed value is used to score the highest forecast value, and the lowest observed value is used to score the lowest forecast value.

The '*range of forecast conditions*' approach avoids the need of assumptions about probabilities for conditions forecast by TEMPO and PROB TEMPO, or ambiguous conditions during a BECMG period. The '*time period*' approach allows a range of observed conditions to be compared with a range of forecast conditions. To study the effect of certain types of change groups, TEMPO, PROB and PROB TEMPO groups can be excluded from verification.

As airlines usually use the lowest (most adverse) condition for flight planning, the verification results for the lower margin can be used as user-oriented results.

Scores used are mostly taken from standard literature (e.g. Jolliffe and Stephenson, 2003). Additionally, conditional probabilities are used to investigate the ability to discriminate between events and non-events. These measures can easily be understood by customers.

Results are available to the forecaster for individual TAFs ('eyeball verification') and in the form of contingency tables. For management, overall scores are available for each airport and weather element. For customers, results are presented specifically for important flight planning and operations thresholds. The question of forecast value will be investigated.

Appendix

ATM = Air Traffic Management
AUTOTAF = Automatically produced TAF
BECMG = Becoming (used in TAF code)
BKN = Broken (5–7/8 cloud cover; used in METAR and TAF code)
BLSN = Blowing snow (used in METAR and TAF code)
BR = Mist (used in METAR and TAF code)
CB = Cumulonimbus (used in METAR and TAF code)
Ceiling = Layer cloud cover >4/8 or obscured sky with vertical visibility
CSI = Critical Success Index
E = Event
F = False alarm rate
FAR = False alarm ratio
FCST = Forecast
FM = From (used in TAF code)
FZFG = Freezing fog (used in METAR and TAF code)
GGgg = Time statement (hours and minutes; used in METAR and TAF code)
GS = Gerrity score
H = Hit rate (probability of detection POD)
HSS = Heidke Skill Score
ICAO = International Civil Aviation Organization
MET = Meteorological
METAR = Meteorological aviation routine weather report
NORTAF verification scheme = TAF verification scheme developed by the Northern European Meteorological Institutes
NSW = No significant weather (used in METAR and TAF code)
OBS = Observed
ORSS = Odds ratio skill score
OVC = Overcast (8/8 cloud cover; used in METAR and TAF code)
p(E) = Base rate; probability of event E
p(E) when forecast = Probability of event E when being forecast
p(E) when not forecast = Probability of event E when being not forecast
PC = Percent correct
POD = Probability of detection (hit rate H)
PROB.. = Probability forecast (used in TAF code) followed by percentage value
PSS = Peirce's Skill Score
RA = Rain (used in METAR and TAF code)
RASN = Rain and snow (used in METAR and TAF code)
RE.. = Recent, followed by weather phenomenon (used in METAR code)

Copyright © 2008 Royal Meteorological Society

DOI: 10.1002/met

RVR = Runway visual range
SN = Snow (used in METAR and TAF code)
SPECI = Special weather report in METAR code, issued when an operationally significant deterioration or improvement in airport weather conditions is observed
TAF = Terminal aerodrome forecast
TCU = Towering Cumulus (used in METAR and TAF code)
TEMPO = Temporarily (used in TAF code)
THS = Threat Score
TIPS = TAF Interactive Production System (EUMETNET project)
TS = Thunderstorm (used in METAR and TAF code)
UTC = Universal time coordinated
VC.. = Vicinity, followed by weather phenomenon (used in METAR code)
VIS = Visibility
VRB = Variable wind direction (used in METAR and TAF code)
VV = Vertical visibility (used in METAR and TAF code)
WMO = World Meteorological Organization

References

Balzer K. 1995. TAF Verifikation. Eine Dokumentation der Methodik. Deutscher Wetterdienst.

Fuller S. 2003. Verification of Terminal Aerodrome Forecasts. Met Office NWP Gazette October 2003.

Gordon ND. 1993. Verification of terminal forecasts. In *Proceedings Fifth International Conference on the Aviation Weather System 2–6 August 1993*, Vienna, Virginia.

Harris G. 1998. The UKMO TAF Verification Scheme. Report presented at the TIPS Task 6 Project Workshop, October 1998, Stockholm.

Hilden A, Kjaer O, Feldberg R. 1998. *Forecast Verification Report January 1998*. DMI: Copenhagen.

Hilden A, Heidegård A, Midtbø KH, Santakari J, Vävargård T. 1996. *The NORTAF Verification System, Status Report May 1996*. DMI: Copenhagen.

International Civil Aviation Organization (ICAO). 2004. Meteorological services for international air navigation. *Annex 3 to the Convention on International Civil Aviation*, 15th edn ICAO: Montreal.

Jolliffe IT, Stephenson DB (eds). 2003. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. John Wiley and Sons: Chichester.

Kluepfel CK. 2005. TAF verification in the U.S. National Weather Service. NWS Instruction 10–1601.

Stephenson DB. 2000. Use of the, odds ratio' for diagnosing forecast skill. *Weather and Forecasting* **15**: 221–323.

World Meteorological Organization WMO. 1995. International codes – Volume I.1 Part A: Alphanumeric codes.

Copyright © 2008 Royal Meteorological Society

DOI: 10.1002/met