Verification and performance measures of Meteorological Services to Air Traffic Management (MSTA)

Background

Information on the accuracy, reliability and relevance of products is provided in terms of verification methods, metrics and/or performance measurements. For "Scientific" verification and assessment of accuracy, the MET community currently uses several methods that were developed from a MET perspective, addressing primarily the need to ensure that system changes lead to improved results. Nevertheless this community understands the need to move from a pure MET assessment of accuracy towards an ATM -impact based or scenario based evaluation. Several initiatives seeking to quantify impacts and express success in user-oriented terms currently exist in several States. This document aims at giving an overview of these developments.

MSTA Forecast

MSTA forecasts are designed to facilitate decision support for aviation stakeholders. Hence, weather forecast elements would be mapped to impacts on air traffic operations. These weather elements and the quantitative thresholds and criteria to be used for this mapping are to be determined by agreement between MET authority, the service provider and user communities concerned. Moreover, information on the level of confidence (measures of reliability) that the users can put on the forecasts is required as an essential element of a decision support system. This confidence level needs to be derived from verification and validation processes, and would typically be scenario-dependent.

Verification and performance measures of MSTA forecast

MSTA forecasts aim to facilitate decision support. Hence, it is crucial to forecast whether the weather elements concerned (e.g. ceiling height, visibility, crosswind, thunderstorms hitting certain holding area) will reach certain criteria or thresholds which impact on ATM (reduced capacity of airspace, operating categories and procedures).

To facilitate end user's decision making process, verification can be made in various ways to express in terms that are closely related to the user decision processes. In particular:

(A) Time-success diagram
This can be constructed based on past performance of the system, to provide a "tolerance limit" that users define what can be considered acceptable (e.g. the location of thunderstorm hitting certain holding area, say 20 nm radius region, crosswind reaching certain thresholds, the onset or cessation of a phenomenon such a freezing rain has to be predicted within 10, 20 or 30 minutes), and based on these definitions of a "hit", a diagram can be constructed that allows to determine the likely lead time (warning time before onset) for any desired threshold of success. As different user communities (TWR controller, APCH, Supervisor, ATFM) work to very different time horizons in their decision making process, they can be provided with the expected success rate at the required time horizon, or, if a minimum success rate is defined (say 75% of all cases of below IFR conditions need to be

predicted), the available lead time (say 3 hours) can be read from the diagram.

For thunderstorms/significant convection, identification of a "hit" or a "false alarm" could be made by comparison with ground truth (radar/lightning) data with reference to pre-defined categorized criteria (e.g. intensity, spatial and temporal coverage). The thresholds are chosen that correspond to operating regulations/impact that are directly relevant to the customer (i.e. user-relevant measures).

The performance measures or ability of the forecasting system to correctly predict the parameter in the given categories can be measured by performance indicators such as Probability of Detection (POD) (or hit rate), False Alarm Ratio[1] (FAR) or other measures of success (e.g. Critical Success Index[2] (CSI)) composed from these two. The presentation could be time-success diagram (Figure 1) or a contingency table from which the POD, FAR and CSI could be computed (Figure 2). Annual performance can also be presented to provide a trend of performance (Figure 3).
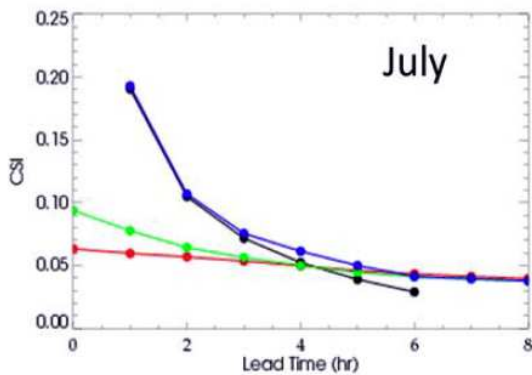

Figure 1: Monthly mean CSI of convection forecasts (different colours represent different methods) as a function of lead time

		OBSERVATIONS								
		00-09	10-19	15-19	20-24	25-34	35-44	45-54	55-99	*
F	00-09	31643	3690	296	33	5	0	0	0	35667
O	10-19	3243	1784	342	32	0	0	0	0	5401
R	15-19	431	434	192	14	2	0	0	0	1073
E	20-24	133	223	188	78	23	0	0	0	645
C	25-34	1	34	54	39	20	0	0	0	148
A	35-44	0	0	0	0	0	0	0	0	0
S	45-54	0	0	0	0	0	0	0	0	0
T	55-99	0	0	0	0	0	0	0	0	0
S	*	35451	6165	1072	196	50	0	0	0	42934

Figure 2: Contingency table of crosswind forecasts versus crosswind observations

---

[1] FAR = (false alarms) / (number of forecasts of the event)
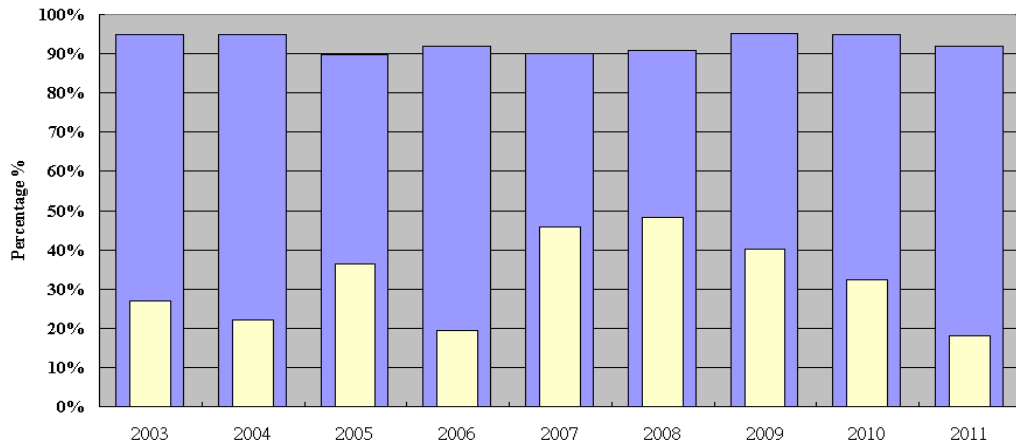[2] CSI = (hits) / (hits + false alarms + misses)

Figure 3: Year to year performance statistics (POD (blue colour) and FAR (cream colour)) of windshear alert service

In some real-life verification examples, capability of forecast thunderstorms or significant convection hitting certain specific area which are of interest to ATM (typical radius of 20 nm) drops rather rapidly with lead time, especially after the first couple of hours (Figure 1). In addition, verification results indicate that forecasters' contribution could generally improve the forecast capability as seen in some cases. This demonstrates a need for manual input to tackle limitation of automatic forecasting systems.

(B) Onset/end time verification
As different ATM users have different lead time requirements for severe weather, verification could be made on onset time and end time of these events. Figure 4 illustrates an example of verification of onset time and end time of thunderstorms. Generally speaking, performance of onset time is better than end time forecast.
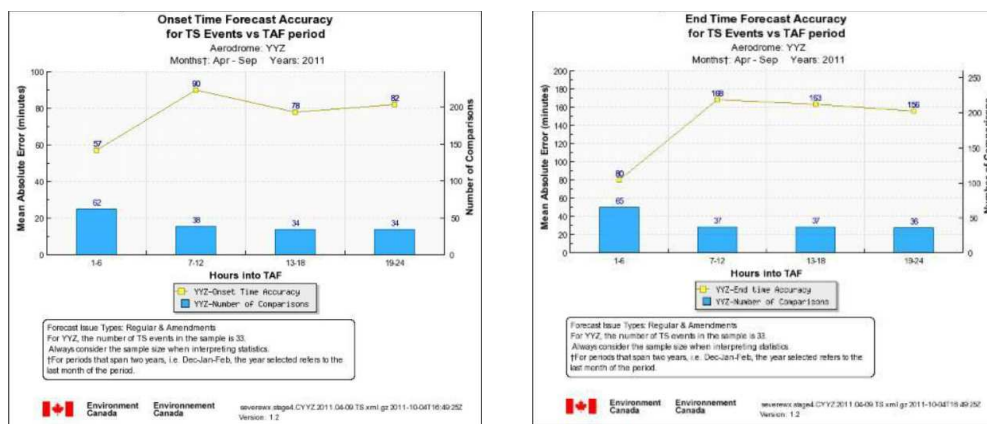

Figure 4: Verification of onset time and end time of thunderstorm forecasts

(C) Object-oriented verification
Forecasts of significant weather given in graphical form could be expressed in object form (e.g. severe icing polygon). Where thresholds have been determined for gridded forecasts (e.g. radar reflectivity exceeding 35 dBZ), similar significant weather objects could be identified. Location errors of the forecast objects could be computed by comparing the location of the forecast object versus that of the actual object, e.g. by referring to the centroid of the identified object. The overlapping area as well as the

difference area between the forecast and actual objects could also be computed and evaluated (Figure 5).
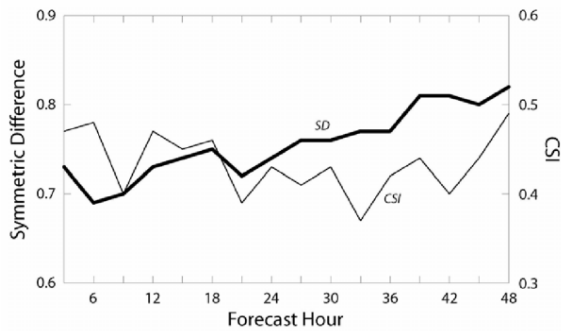

Figure 5: Difference between the forecast and actual objects (symmetric difference) and CSI of a precipitation forecast as a function of the lead time (forecast hour)

(D) Impact-based verification

More sophisticated verification methods include verification of weather impact to ATM, e.g. aircraft route deviations and weather avoidance fields, which attempt to quantify whether a pilot will fly near weather hazards; evaluation of flight delay or reduction of airspace capacity. Figure 6 illustrates the "mincut" approach in evaluating flow reduction of a jetway within a sector due to a convection area.
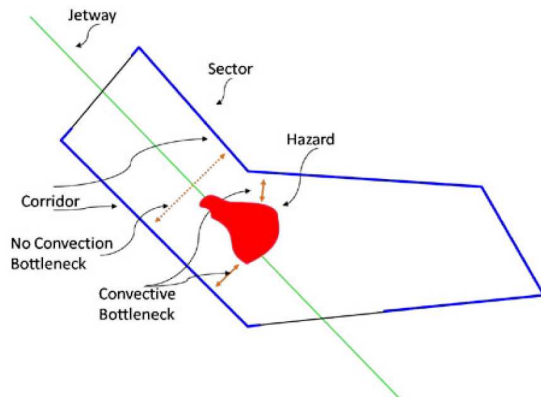

$$\text{Flow Reduction} = 1 - \frac{Mincut_{convection}}{Mincut_{corridor}}$$

Figure 6: Schematic diagram indicating the concept of "mincut" in evaluating flow reduction

Predictability of weather phenomena

Predictability and performance are location- and weather-type dependent and also depend on season, climate and local environment, e.g. complex topography. For different weather event types, performance could vary quite a lot (Figure 7). POD and forecast skill score (CSI) are higher for synoptic scale weather phenomena, e.g. approach/depart of cold/warm front. However, for mesoscale convective system (MCS), i.e. severe organized thunderstorm system, POD and forecast skill score could be significantly lower.
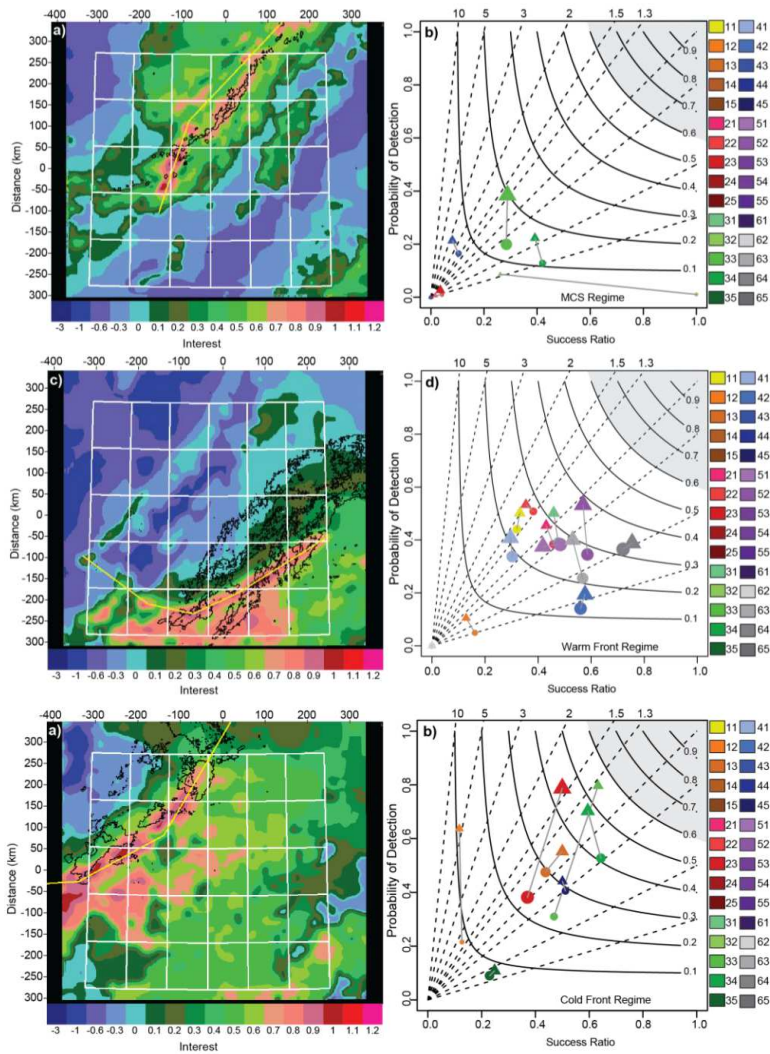
Figure 7: Performance scores (POD and CSI) of convection forecasts for three different weather regimes: MSC, warm front and cold front