2019

Adding value to numerical weather predictions for the aviation industry in South Africa.

Quinton Jacobs Mr.
South African Weather Service, Quinton.Jacobs@weathersa.co.za
Willem A. Landman Prof.
University of Pretoria, willem.landman@up.ac.za

Introduction

The aviation industry forms part of the backbone of modern society. Air travel enables people and goods to be transported over vast distances in short periods of time. The aviation industry, therefore, enables the populace of the world to interact in a way no other form of transport could do before. This heavy reliance on air travel brings with it a great responsibility to the safety of passengers and cargo. One of the greatest safety hazards the aviation industry faces is the weather, since an aircraft, on the ground or in the air, is subjected to weather-related hazards. Such hazards include reduced visibility due to fog or precipitation, windy conditions (crosswinds or gust fronts), ice accumulation on the wings of the aircraft reducing lift, turbulence, and the numerous hazards associated with thunderstorms.

Aeronautical Meteorological Services are allowed by the World Meteorological Organization (WMO) to charge for services rendered to the aviation industry since aviation-related forecasts are deemed specialized forecasts and not public good forecasts (WMO, 2007). The South African Weather Service (SAWS) charges the aviation industry for the services supplied as the designated Aeronautical Meteorological Service for South Africa. These changes to the aviation industry amount to about 24–31% of the annual income of SAWS, whereas other commercial income is less than 10% (SAWS, 2008, 2010, 2013). As a significant source of income, aviation forecasts are important to SAWS and therefore the quality of these products need to be evaluated. Establishing the quality of aviation products from SAWS is also to the benefit of the aviation industry which has a right to know the quality of the forecast products provided. Moreover, knowing which areas of the forecast need more attention can lead to research being conducted to improve the shortcomings and further enhancement of the aviation forecast products, leading to an improved product which benefits the end-user.

Another factor to take into account is the availability of weather forecasts over the internet. These online weather forecasts and products are usually directly derived from raw numerical weather prediction (NWP) model outputs, sourced from the various model output generating agencies around the globe. The purpose of this reserach is to determine whether the SAWS human forecasters can outperform a state-of-the-art NWP model with regards to Terminal Aerodrome Forecasts (TAF). The NWP model used in this study is the Unified Model (UM) from the Met Office of the United Kingdom, as administrated by SAWS. This version of the UM has a 12 km resolution, with hindcasts available for 1-hour intervals over a 48 hour period. (Landman, Engelbrecht, Engelbrecht, Dyson, & Landman, 2012). This UM configuration can, therefore, be verified at the same time resolution as the operational TAF messages and its output can generate the model-based TAF data required for the verification presented in this paper.

For this study, the airports considered are the five main commercial and international airports in South Africa. Figure 1 indicates the location of these airports. The five airports are:

1. Johannesburg O.R. Tambo International Airport (FAOR)
2. Cape Town International Airport (FACT)
3. Durban King Shaka International Airport (FALE)
4. Bloemfontein Bram Fischer Airport (FABL)
5. Port Elizabeth International Airport (FAPE)

These five airports are all serviced by different regional forecasting teams of SAWS, and thus represent a well-rounded overview of the forecasting capabilities of SAWS forecasters in the aviation industry. TAF bulletins are issued for other airports also; these airports are selected because they are the airports with the most air traffic in the country. For instance, in the period March 2011 to February 2012 the number of movements (take-offs and landings) at these airports were (Airports Company South Africa, 2013):

- FAOR-212580
- FACT-97935
- FALE-55194
- FABL-20088
- FAPE-68893

These five airports are also in different climatic zones (Landman et al., 2001). FACT is situated in the south-western Cape (winter rainfall region dominated by frontal passages); FAPE is situated on the south coast (all-year-round rainfall region, affected by both summer tropical troughs and frontal passages); FALE is located near the KwaZulu-Natal coast (region dominated by summer tropical troughs, but also affected by the ridging of the Atlantic ocean high); FABL is located in the central interior (region dominated by summer tropical troughs); and FAOR is situated in the north-eastern interior (region dominated by summer tropical troughs, located on a ridge (the Witwatersrand Ridge). Since the airports are geographically separated, the results will not be compared against each other as each set of forecasters have different challenges. Comparing the forecasting teams is not the goal of this study-comparing the forecasting systems is the main objective.

Human-compiled TAF messages and observations are available from 2002 in the electronic archives of SAWS, but the version of the Unified Model used in the study only became operational in January 2011. Therefore, the study is conducted on a two-year data set from 1 February 2011 to 31 January 2013. The 12 km UM analyses here has since been discontinued and has been replaced with higher resolution models (4 km and 1.5 km). The higher resolution models will be used in future studies but are not used in this study.

Data and Methodology

To conduct the evaluation of the forecasting systems, three hindcast datasets were created, one for each forecast system – the human forecasters; the NWP model forecasts; and a persistence forecast as control. The data for the three datasets have three origins: the actual observed data at the airports (in aviation coded form METAR and SPECI, where METAR messages are the routine reports and SPECI messages are special reports); the human-generated TAF messages; and raw NWP model output sourced from the Unified model from which model-based TAF messages are created.

Data

Observed weather observations at the five airports serve as the verification data. This data is also used to construct persistence forecasts. The persistence forecasts assume that the forecast for tomorrow is exactly the same as the weather observed today. Therefore, the observations are used as the forecast 24 hours later in the construction of the persistence forecast TAF messages. Thus, for the persistence forecast TAF messages, the observational data is moved forward by one day and then used to construct TAF messages.

The human-generated TAF messages are obtained from the archives of SAWS and are those which were routinely issued by aviation forecasters at the various airports. Amended TAF messages are disregarded in this study (i.e., TAF messages that were issued at non-standard times, due to unexpected changes in the conditions at the aerodromes). The routine operational TAF messages for the given airports are issued every six hours at 04Z, 10Z, 16Z, and 22Z. These TAF messages, therefore, have a lead time of two hours as they all start on the next main synoptic hour. Bloemfontein Bran Fisher and Port Elizabeth are 24 hour TAF messages, therefore valid for 06/06, 12/12, 18/18, and 00/24. Cape Town, Johannesburg OR Tambo, and Durban King Shaka are 30 hour TAF messages, therefore valid 06/12, 12/18, 18/24, and 00/06. This configuration of the TAF messages are duplicated for the persistence and UM generated TAF messages. This was primarily done to ensure the data sets are equal and that the forecasts being evaluated are of equal temporal range. This eliminated bias towards any of the systems and ensures the data compared are the same.

For the UM data, the model grid-point closest to the airport is used. The UM model used here has a resolution of 12 km (Landman et al., 2012), but ICAO stipulates that weather given in an observation or TAF is valid for a 15 km radius from the airport (ICAO, 2012). Thus the selected grid-point is within the area allowed by ICAO.

The UM model provides easy extraction of most atmospheric variables required to build a TAF message. Only the present weather and cloud height had to be determined independently from the model output. Rain and snow are determined

using rain and snow categorical fields of the model. Rain is additionally identified as convective or non-convective and considered convective only if the convective precipitation is greater or equal to half of the total rainfall. When rainfall is found to be convective the total totals instability index is subsequently calculated in order to determine if the rainfall is from cumulonimbus cloud (Revering, n.d.). To determine between thundershowers and showers, the Total Totals Index values where considered. A value of 45 or above is used to determine thundershowers. If the value is lower than 45, it is deemed to be only showers. A value of 45 for the Total Totals Index is considered where thunderstorms will become scattered and more likely to occur (Miller, 1972). Values of 44 can also lead to thundershowers but are only a few and isolated.

Determining the cloud base height requires careful consideration. With the atmospheric variables available from the UM, no ideal system to determine the cloud base height could initially be found. The best solution is found when calculating the lifting condensation level pressure and then calculating the height of this pressure above sea level. Calculating the pressure of the lifting condensation level is done by iteration of the wet-bulb temperature and the dry adiabatic lapse temperature as pressure decreases, as calculated on a skew-T plot. The process algorithms used are those presented by Schlatter and Baker (1991). Once the pressure value is determined, the hypsometric equation is used to determine the height above the airports, as the airport height above sea-level is known (Holton, 1992).

Methodology

To evaluate the various hindcasts, a verification system was developed by using the criteria found in ICAO Annex 3 (ICAO, 2012). The TAF message is evaluated against the aviation observation message (METAR or SPECI) valid during the validity period of the TAF. For this study, only the TAF messages issued on the standard-issue times (04Z, 10Z, 16Z, and 22Z) are considered. This is done to eliminate the advantage human-forecasters may have by amending their TAF messages.

The verification process divides the aviation messages (observation (METAR and SPECI) and TAF) into six components and into hourly intervals:

1.   Wind direction-direction is only considered significant if the wind speed ≥ 10 kt. Variable winds are ignored. Wind direction is evaluated as a hit or miss. If the difference between the forecasted and actual wind direction exceeds 60º it is considered a miss, otherwise a hit (ICAO, 2012).

2.   Wind Speed and gusts-Winds ≥ 10 kt is considered significant, therefore if both the actual and forecasted winds are < 10 kt, the event is considered a correct rejection. The forecasted wind speed is subtracted from the actual wind speed. If the absolute difference is ≤ 10 kt, the event is considered a hit. If the result

is > 10 kt, the actual wind is stronger than the forecast and therefore a miss. If the result is < -10 kt, the actual wind speed did not reach the forecasted value and is a false alarm (ICAO, 2012). The same procedure applies when significant gusts are reported/forecasted. The 10 kt change in wind speed is the only stipulation made within the criteria set out in Annex 3 and therefore used as the only consideration (ICAO, 2012). Wind speed 10-minute average exceeding 40 kt is a rare occurrence in South Africa, and should therefore not present a major impact in operations if used as the verification guideline.

3.   Visibility-Visibility is handled by setting the visibility into a category, based on the limits for visibility as defined in ICAO Annex 3. The CAVOK (Ceiling And Visibility OKay) code is defined as visibility of greater than 10 km and is therefore assigned a visibility value of 9999. The categories are ordered with higher visibilities allocated higher values. All visibilities above 5000 m are given a category value of 9. The visibility categories start at 5 for visibilities between 3000 m and 5000 m and drop to 1 for visibilities below 100 m. The VV (Vertical Visibility) Code is interpreted as zero visibility, as this code implies that horizontal visibility is almost zero and only the vertical visibility can be observed. The categories are compared-if both are 9, visibility is considered a correct rejection. Otherwise, if both are the same the forecast is correct is considered a hit; if the actual visibility category > forecasted visibility category, a false alarm is recorded; and actual visibility category < forecasted visibility category, a miss is recorded (ICAO, 2012).

4.   Present weather is only considered when visibility is ≤ 3000 m. The text in the forecasted present weather is checked to determine whether the actual contains the same text-if it does it is considered a hit; if not it is a miss. If both present weather texts are empty it is considered a correct rejection (ICAO, 2012).

5.   Cloud amount-Cloud amounts are only considered if the cloud base is ≤ 1500 ft. Cloud amount is also considered significant if the change is between broken (BKN) and scattered (SCT) cloud. Therefore, if both the actual and forecasted cloud amount is broken or overcast it is considered a hit. If both the actual and forecasted cloud amounts are scattered, few or nil, it is considered a correct rejection. If the actual cloud amount is overcast or broken, but the forecast cloud amount is not, it is considered a miss. If the forecasted cloud amount is overcast or broken and the actual cloud amount not, then it is a false alarm. CAVOK and NSC codes are taken as NIL999, implying no cloud (as it is not significant). The VV Code is interpreted as OVC000, implying overcast condition on the ground (ICAO, 2012).

6.   Cloud Base-Cloud base is only significant when the cloud amount is broken or overcast. Therefore, if both actual and forecasted cloud amounts are not broken or overcast, the event is considered a correct rejection. Cloud base is

also assigned ranges in ICAO Annex 3 and is therefore treated similarly to the visibility, by assigning categories to the ranges (ICAO, 2012).

BECMG (BECMG = becoming) groups are used to show weather parameters that are expected to change from the previous prevailing conditions. Omitted weather parameters are considered to stay the same. These BECMG groups are handled by assigning the average value between the component value before the group and the component value after the group to the hour value between the times of the BECMG group. Therefore, the verification system interprets the change as a linear change of the two hour period of the BECMG period. This is done to ensure consistency and simplicity in the results.

TEMPO (TEMPO = Temporal) groups are considered to be of secondary importance in the forecasts. TEMPO suggest alternative weather parameters which are expected to only occur less than half of the time of the validity period. This is usually interpreted as temporal changes in the predominant weather conditions, resulting from the weather parameters given in this group, for example when showers are expected the weather conditions will deteriorate to the values given in the TEMPO group, only when the showers are active over the aerodrome. Once the showers have passed the prevailing conditions should return. Therefore, the TEMPO value will only be evaluated if the original value did not yield a hit or correct rejection. The TEMPO values are evaluated the same as the prevailing values of the forecast. Since the values are considered ranges of correct values the PROB (PROB = probability) code is ignored. Therefore, a PROB TEMPO group will be handled as a TEMPO group. The TAF messages are evaluated for the entire duration of its validity period, using the corresponding observation in hourly intervals.

The SPECI (SPECI = Special Aviation Meteorological Report) criteria apply to observations and TAF messages. The SPECI criteria are the limits of significant changes and are determined by ICAO and stipulated in Annex 3 (ICAO, 2012). Since the TAF message is considered a concise forecast of the most likely meteorological conditions at the aerodrome, the SPECI criteria also apply and are considered as the significant changes in the conditions. There are also some other criteria given such as a change of runway, but this differs for the aerodrome to aerodrome and is therefore not considered. Since the SPECI criteria consist of ranges of values, the range of values within SPECI criteria can be considered correct for each of the six verified variables and thus the probabilistic nature of TAF messages can be simplified into a binary system (Mahringer, 2008). The verification system handles TAFs as binary forecasts, therefore generating 2x2 contingency tables (Table 1) and one can subsequently calculate the verification indices derived from the resulting contingency tables (Doswell et al., 1990; Jolliffe, et al., 2012; Joint Working Group on Forecast Verification Research (JWGFVR), 2014). The verification indices used in the study are limited to the following:

- Proportion Correct (PC) – This index is evaluated because it is used by both SAWS and ICAO as the index for the desired level of accuracy (ICAO, 2012; SAWS, 2008).
- Hit Rate (H) – This index indicates the ability of the forecast system to determine the onset of significant events, which is important to both the Meteorological Authority and the end-users.
- False Alarm Ratio (FAR) – The false alarm ratio is the percentage of false alarms, with relation to the forecasts.
- Critical Success Index (CSI) – This index determines the number of correct forecasts out of the sum of hits, misses and false alarms. This index, therefore, takes both misses and false alarms into consideration. It is important to note that correct rejections are not considered at all within the CSI calculation.
- Heidke Skill Score (HSS) – The Heidke Skill Score is a skill score designed to make the proportion correct measure linear. It favours under-predicting systems and can be harsh to an over-predicting system (Jolliffe et al., 2012).
- Pierce Skill Score (PSS) – This skill score is, in essence, the difference between the hit rate and false alarm rate. Since the PSS only considers the hit rate and false alarm rate, bias does not affect the outcome. Unlike the Heidke Skill Score, it favours over-predicting systems and is harsh to under-predicting systems (Jolliffe et al., 2012).

A skill score can be used to determine which forecast system is best, but is meaningful only if the skill score is equitable. An equitable skill score treats random forecasts and a constant forecast in an equal manner. Equitability also ensures that the forecasts do not distort due to the occurrence of common events. Correct forecasts of rare events would carry more weight than more common events. The Heidke and Pierce Skill Scores are both equitable (Wilks, 2011).

A Monte Carlo or Bootstrap Method is employed to determine whether the values calculated from the three forecast systems are significant or not, with the significance level set to the 95th percentile. The bootstrap method is analogous to writing the TAF messages on strips of paper and throwing it into a hat. The hat is shaken and for every day in the period, a strip is drawn randomly from the hat. The drawn TAF is written down as the TAF for the day and returned to the hat and shaken again (Wilks, 2011). Thus the original data sets of TAF messages are read to memory and evaluated. The original data sets of TAF messages in memory are then used to rebuild the TAF data randomly from the original data. These random values are also verified using the same verification system as the original data. 5000 random sequences are generated for each TAF dataset. The results are sorted and the significance cut-off value is set to 95%.

The forecaster versus unified model skill score (FvUMSS).

To determine which forecast system is best, between the human-forecasters or the Unified Model Forecasts, an additional skill score is developed to compare the two systems. From the generic skill score formula (equation 1) (Wilks, 2011):

$$FvUMSS = \frac{F_{ev} - F_{Ref}}{F_{Perf} - F_{Ref}} \quad (1)$$

Where F_{ev} is the verification score for the forecast to be evaluated; F_{Ref} is the verification score for the Reference forecasts and F_{Perf} is the verification score for the value of a perfect forecast:

$$FvUMSS = \frac{F_{Human} - F_{UM}}{F_{Perf} - F_{UM}} \rightarrow F_{Perf} = \begin{cases} 1 \\ 0 \; for \; FAR \end{cases} \quad (2)$$

The FvUMSS value (equation 2) is calculated for all the resulting indices used in the study and the results are interpreted as follows (equation 3):

$$if \; FvUMSS = \begin{cases} |FvUMSS| \leq 10\% \rightarrow Neutral \; result \\ 10\% < |FvUMSS| \leq 25\% \rightarrow Minor \; in \; favour \; of \; Result \\ 25\% < |FvUMSS| \leq 50\% \rightarrow Major \; in \; favour \; of \; Result \\ |FvUMSS| > 50\% \rightarrow Comprehensively \; in \; favour \; of \; Result \end{cases} \quad (3)$$

A positive FvUMSS value implies the human forecaster is performing better than the UM. A negative FvUMSS value implies the UM is performing better than the human forecasters. The thresholds for the categories are chosen as such to identify the degree by which the one system outperforms the other – minor if the result is between 10 and 25%, major if the result is between 25 and 50% and comprehensive if the result is greater than 50%.

Results

The three data sets are created. The evaluation and bootstrap methods are run on each of these data sets. Upon investigating the results, the persistence data set results yielded little useful results if one considers the contingency tables of the variables as separate entities. Therefore, as the results will be discussed in terms of the individual variables, only the human forecasters and the UM data sets will be discussed. These are the two data sets that the study is investigating.

Wind Direction

Table 2 shows the results of the wind direction for all five airports. Results found to be above the 95th percentile are underlined and considered significant. The human forecasters managed proportion correct values between 76–85%, of which all the results are found to be significant. The UM proportion correct values fared better with values between 81–91%, but none are found to be significant. Thus with the values of the UM being higher, chance still outperformed it. Hit rate, CSI, and PSS produced results that are very close for both data sets across all the airports. The HSS score is in favour of the UM with values consistently in the 70's,

except for FAPE, whereas the HSS for the human forecasters is consistently lower than the UM values.

Table 3 shows the FvsUMSS for all the airports regarding wind direction. Since the Proportion Correct values are higher than the human forecaster's values the results are all in favour of the UM. FABL and FACT exceeded 50% and are therefore considered comprehensive results; FAOR a major result; and FAPE and FALE minor results in favour of the UM. The hit rate recorded two minor results in favour of the human forecasters; the CSI also recorded two minor results in favour of the human forecasters and the same goes for the PSS, also two minor results in favour of the human forecasters.

Wind Speed

Table 4 shows the results of wind speed for all the airports for both the human forecaster and the UM. Wind speed is a variable that both forecast systems handle quite well. Proportion Correct values for the human forecasters are between 92–100% and the UM values between 85–97%. False alarm ratio values are low too, with the highest computed value of 12.8% for the UM at FALE and 10.6% for the human forecasters at FAPE. The hit rate for the human forecasters are all above 97%, but the UM model struggled with values dropping to just below 70 for FABL and FACT. The CSI values for the human forecasters are between 87 and 99%, whereas the UM values are between 68 and 91%. The HSS and PSS show similar trends similar to the CSI. Of interest is that all the results of the human forecasters are found to be significant, whereas only false alarm ratio values are significant for the UM.

Table 5 shows the FvsUMSS results for wind speed variable at all the airports. The UM only secured the false alarm ratio results for FABL. FACT, FAOR and FAPE and the proportion correct, HSS and PSS results for FABL in its favour. It is noteworthy to mention that these results are all comprehensive for the false alarm ratios and had major results for the proportion correct, HSS and PSS. All the other results are in favour of the human forecasters, as well as both major and comprehensive results.

Present Weather

From Table 6 it can be seen that the human forecasters handle present weather quite well. This can mainly be contributed to the specific criteria enforced by only considering present weather as significant when visibility drops to 3000 m or less. This ensures that the present weather is correctly predicted more often than not. It is also noteworthy that all human forecast values are significant.

The UM has a problem with detecting the onset of present weather as seen by the hit rate values dropping to the 60's and 70's. This, in turn, affects the other indices negatively as well. It is also noteworthy to mention that none of the results

is found to be significant. Since present weather is considered as a hit, miss or as insignificant, as no false alarms are recorded.

With both systems scoring near-perfect proportion correct values, the FvsUMSS values are in favour of the UM as indicated in Table 7. Since the UM is almost perfect all these results are comprehensive. However, since the hit rate of the UM is significantly lower than the human forecasters, hit rate, CSI and Pierce skill score FvsUMSS values are all comprehensively in favour of the human forecasters because CSI and Pierce's score is heavily dependent on hit rate. The Heidke skill scores had one result leaning to the UM and the rest in favour of the human forecasters.

Visibility

Visibility is the variable the human forecasters struggle the most with. From Table 8 proportion correct values are high, but hit rates are in the 50s and 60s, except for FALE, which managed value of 81%. False alarm ratio values are also very high – from 59% at FABL to 82% at FALE and FAPE. These values also had negative effects on the rest of the indices. This appears to indicate severe over forecasting of visibility. It is noteworthy that all values are significant.

The UM also struggled with the visibility, showing signs of extreme over forecasting of visibility. Because of this trend of over forecasting, the hit rates are all 100%, therefore, no misses, but the false alarm ratio values are also all above 98%, where values close to 0% are desired. These high false alarm values dropped the CSI and Heidke skill scores to almost 0%. All the values, except the false alarm ratio values, are found to be insignificant.

The FvsUMSS results, given in Table 9, show a strong tendency towards the human forecasters, but the Pierce skill scores of the UM are significantly higher than the human forecasters and are therefore comprehensively in the UM's favour.

Cloud Amount

Table 10 shows the calculated results for the cloud amount variable for all five airports and both the human forecasters and UM forecasting system. Forecasters believe the cloud amount of the Numerical Weather Prediction (NWP) model output to be quite reliable. Though the NWP model used by the forecaster might not have been the UM, the results for cloud amount are almost the same across the board. The only value that did not work out the same is the hit rate of FABL, where the UM beat the forecaster by 1.7%, although both are still above 97%. One interesting thing to note is that only the results of FALE are found to be significant, above the 95 percentile.

Table 11 shows the FvsUMSS results for the cloud amount variable. Since almost all the values in Table 2 are the same, low skill is recorded. The 1.7%

difference in the hit rate of FABL's cloud amount resulted in a conclusive result in favour of the UM.

Cloud Base

Cloud base is the other variable both forecasting systems struggle with as can be seen in Table 12. Here the UM has a disadvantage as there is no consolidated method of determining cloud base for both convective and non-convective cloud. The values for both systems are lower than desired, with the exception of proportion correct at FABL. From these results, the hit rates are low and the false alarm ratios are high. This indicates that there is a timing issue with regards to the onset and cessation of significant broken and overcast cloud layers in both forecasting systems. From Table 13, the FvsUMSS values are varied, with the majority showing small differences between the two systems.

Totals of the Variables

Table 14 shows the results when considering the TAF forecast as a whole. When considering the forecasts both forecast systems still deliver within the desired ranges as are set forth by ICAO (ICAO, 2012). It is noteworthy that the UM has lower hit rates and higher false alarm ratio values than the human forecasters, which consequently lowers the CSI, Heidke and Pierce skill scores. It is also evident that all the results of the human forecasters are significant, whereas some of the UM results are not.

Table 15 shows that with the exception of some of the proportion correct FvsUMSS scores, all the results are in favour of the human forecasters.

Discussion and Conclusion

Considering the results of the various components of the TAF forecast between the human forecasters and the UM the following can be deduced:

1.   For the human forecaster to produce quality forecasts, quality guidance is required. Therefore, it is paramount that the guidance given to the human forecaster to aid in decision making should be a quality product. Highly skilled NWP models such as the UM are required for quality guidance to the human forecaster, enabling the production of quality products.

2.   A synergy exists between quality guidance and properly trained forecasters to produce quality products.

3.   The human forecaster's ability to detect the onset of significant changes in the criteria of the various variables are superior to the UM. This can mainly be contributed to the human forecaster's ability to adapt the forecast to prevailing conditions, whereas the model is only run once a day.

4.   Both forecasting systems seem to forecast wind direction, wind speed, present weather and cloud amounts adequately and seem to be on equal footing, with the exception of the hit rates.

5.   Proportion correct, whilst useful, should not be considered in isolation as underlying features of the forecast system can be masked. This shortcoming is best displayed by the visibility variable, which is excessively over forecasted.

6.   Visibility is over forecast by both systems excessively with the UM resulting in enormously high false alarms.

7.   Cloud base seems to be the most difficult variable to forecast for. Timing of the onset and cessation of significant cloud bases of broken and overcast cloud seem to be difficult for both systems.

Both forecast systems present a strong case for themselves. NWP models have come a long way in reliability and accuracy. It is clear from the result above that the UM has a problem with hit rates and therefore gives the human forecaster the edge. The onset and cessation of significant weather at an aerodrome have significant impacts on operations at the aerodrome. Therefore the higher hit rates do indicate that the human forecaster is indeed adding value to the raw output of the UM. It is also noteworthy that the two variables causing the greatest problems at an aerodrome – visibility and cloud base, are the two variables both the human forecaster and the UM have the most problems with.

Since the human forecaster uses the NWP model as a guidance forecast, further work on the calibration of the UM should result in an improved guidance forecast. If the guidance forecast becomes more reliable, the end product should therefore also improve. There are numerous online websites providing free NWP forecasts, but as to the needs of the aviator, speaking to the human forecaster and using the aviation products they generate, would provide a more useful set of data to plan and execute flight plans.

References

Airports Company South Africa. (2013, 9 11). *ACSA airport passengers and movements*. Retrieved from http://www.acsa.co.za/home.asp?pid=117

Doswell, C. A., Davies-Jones, R., & Keller, D. L. (1990). On summary measures of skill in rare event forecasting based on contingency tables. *Weather and Forecasting, 5*, 576-586.

Holton, J. R. (1992). *An Introduction to dynamic meteorology* (3rd ed.). London: Academic Press.

International Civil Avaition Organization. (2012). *Standards and recommended practices annex 3 to the convention on international civil aviation meteorological service for international air navigation annex 3* (17th ed.). Montréal: Author.

Joint Working Group on Forecast Verification Research. (2014). *Forecast verification: Issues, methods and FAQ*. Retrieved April 22, 2015, from http://www.cawcr.gov.au/projects/verification/

Jolliffe, I. T., & Stephenson, D. B. (2012). *Forecast verification: A practitioner's guide in atmospheric science* (2nd ed.). West Sussex: John Wiley & Sons.

Landman, S. (2012). *A multi-model ensemble system for short-range weather prediction in South Africa.* Pretoria: University of Pretoria.

Landman, S., Engelbrecht, F. A., Engelbrecht, C. J., Dyson, L. L., & Landman, W. A. (2012). A short-range weather prediction system for South Africa based on a multi-model approach. *Water SA, 38*(5), 765-774.

Landman, W. A., Mason, S. J., Tyson, P. D., & Tennant, W. J. (2001). Retro-active skill of multi-tiered forecasts of summer rainfall over southern Africa. *International Journal of Climatology, 21*, 1-19.

Mahringer, G. (2008). Terminal aerodrome forecast verification in austro control using time windows and ranges of forecast conditions. *Meteorological Applications, 1*(15), 113-123.

Miller, R. C. (1972). *Technical report 200: Notes on analysis and severe-storm forecasting procedures of the air force global weather central.* Air Weather Service (MAC) United States Air Force.

Revering, A. (n.d.). *Andrew Revering's list of meteorological formulas*. Retrieved July 11, 2014, from http://www.aprweather.com/pages/calc.htm

South African Weather Service. (2008). *Annual report 2007/2008.* Pretoria: South African Weather Service.

South African Weather Service. (2010). *Annual report 2009/2010.* Pretoria: South African Weather Service.

South African Weather Service. (2013). *Annual report 2012/2013.* Pretoria: South African Weather Service.

Schlatter, T. W., & Baker, D. V. (1991, June 12). *Algorithms, comparisons and source references by Schlatter and Baker*. (NOAA Environmental

Research Laboratories). Retrieved July 11, 2014, from
http://wahiduddin.net/calc/density_algorithms.htm

Wilks, D. S. (2011). *Statistical methods in atmospheric sciences* (3rd ed.).
Oxford: Academic Press.

World Meteorological Organization. (2007). *Guide to aeronautical
meteorological services and cost recovery WMO-No. 904* (2nd ed.).
Geneva: Author.
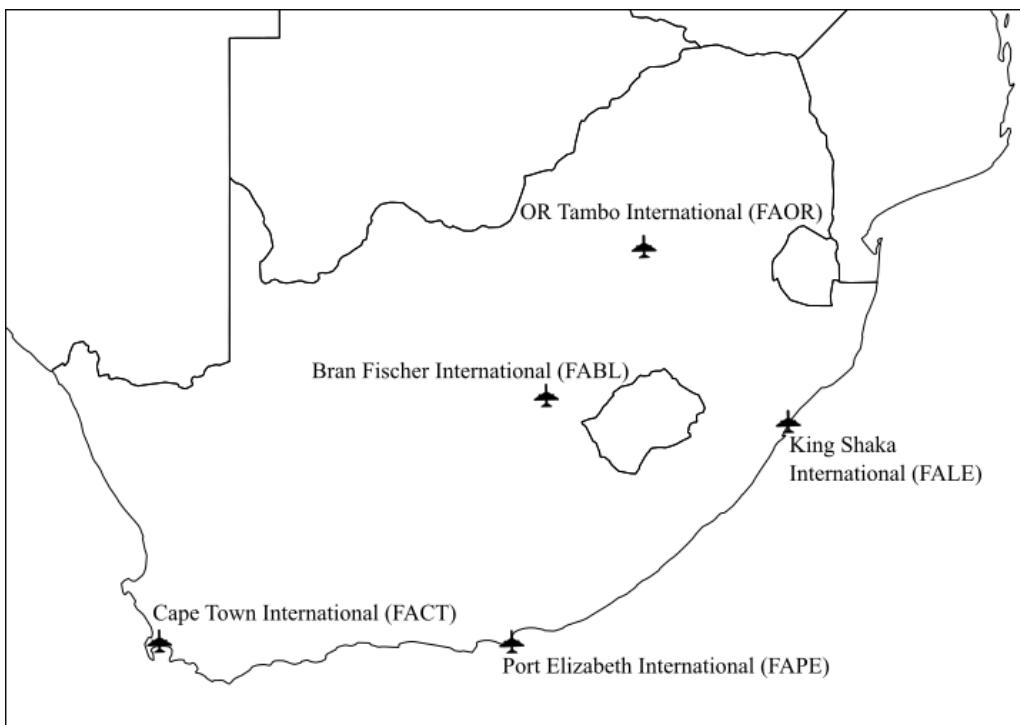
Figure 1. Locations of main international airports in South Africa.

Table 1
2x2 Contingency Table (Jolliffe & Stephenson, 2012)

		Event Observed		
		Yes	No	Total
Event Forecasted	Yes	A	b	a + b
	No	C	d	c + d
	Total	a + c	b + d	n = a + b + c + d

Table 2

Calculated Results as a Percentage for Wind Direction Variable for the Human Forecaster (F) and the Unified Model (U) for all Five Airports

	Wind Direction									
	FABL		FACT		FALE		FAOR		FAPE	
	F	U	F	U	F	U	F	U	F	U
Proportion Correct	**79.8**	90.3	**76.7**	85.0	**83.0**	85.4	**84.8**	89.7	**78.3**	81.6
Hit Rate	**74.3**	74.6	**73.0**	74.6	**74.8**	72.5	**77.1**	73.9	**74.1**	70.2
False Alarm Ratio	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**
Critical Success Index	**74.3**	73.2	**73.0**	74.6	**74.8**	72.5	**77.1**	73.9	**74.1**	70.2
Heidke Skill Score	**55.2**	77.7	**42.6**	70.6	**66.0**	71.1	**69.2**	77.4	**48.3**	64.3
Pierce Skill Score	**74.3**	73.2	**73.0**	74.6	**74.8**	72.5	**77.1**	73.9	**74.1**	70.2

Note. Values are given for the six verification indices considered. Bold values are found to be significant, above the 95th percentile.

Table 3

Calculated FvUMSS Values as a Percentage for Wind Direction Variable. Positive Values are in Favour of the Human Forecaster and Negative Values in Favour of the Unified Model

	Wind Direction				
	FABL	FACT	FALE	FAOR	FAPE
Proportion Correct	-108.9	-54.8	-15.9	-47.8	-17.7
Hit Rate	-1.2	-6.1	8.5	12.4	13.1
False Alarm Ratio	-	-	-	-	-
Critical Success Index	4.2	-6.1	8.5	12.4	13.1
Heidke Skill Score	-100.9	-94.9	-17.9	-36.1	-44.8
Pierce Skill Score	4.2	-6.1	8.5	12.4	13.1

Table 4

Calculated Results as a Percentage for Wind Speed Variable for the Human Forecaster (F) and the Unified Model (U) for all Five Airports

	Wind Speed									
	FABL		FACT		FALE		FAOR		FAPE	
	F	U	F	U	F	U	F	U	F	U
Proportion Correct	**95.8**	97.0	**93.8**	85.3	**94.5**	90.0	**99.3**	96.0	**92.1**	86.0
Hit Rate	**99.6**	69.3	**97.8**	69.3	**98.5**	80.1	**99.6**	89.4	**97.3**	72.4
False Alarm Ratio	**5.4**	**1.3**	**5.4**	**2.1**	**8.2**	**12.8**	**0.7**	**0.3**	**10.6**	**2.0**
Critical Success Index	**94.2**	90.3	**92.6**	68.3	**90.6**	71.7	**98.9**	89.2	**87.3**	71.4
Heidke Skill Score	**89.7**	92.8	**80.0**	69.6	**88.9**	76.3	**98.5**	91.2	**83.8**	71.8
Pierce Skill Score	**86.8**	90.8	**76.1**	68.1	**88.4**	74.7	**98.4**	89.2	**82.9**	71.1

Note. Values are given for the six verification indices considered. Bold values are found to be significant, above the 95th percentile.

Table 5

Calculated FvUMSS values for Wind Speed Variable

	Wind Speed				
	FABL	FACT	FALE	FAOR	FAPE
Proportion Correct	-42.5	58.0	45.2	82.9	43.5
Hit Rate	98.6	92.9	92.5	96.3	90.4
False Alarm Ratio	-333.1	-162.6	36.0	-161.3	-432.8
Critical Success Index	40.2	76.7	66.6	89.8	55.7
Heidke Skill Score	-42.5	34.1	53.0	83.3	42.8
Pierce Skill Score	-43.9	25.0	54.2	85.3	40.8

Note. Positive values are in favour of the Human Forecaster and Negative values in favour of the Unified Model.

Table 6

Calculated Results as a Percentage for Present Weather Variable for the Human Forecaster (F) and the Unified Model (U) for all Five Airports

Present Weather										
	FABL		FACT		FALE		FAOR		FAPE	
	F	U	F	U	F	U	F	U	F	U
Proportion Correct	**98.0**	99.4	**96.6**	97.8	**96.6**	98.5	**97.2**	98.9	**96.0**	98.2
Hit Rate	**97.7**	74.7	**95.3**	74.7	**94.3**	63.7	**95.8**	79.6	**95.4**	65.4
False Alarm Ratio	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**
Critical Success Index	**97.7**	86.5	**95.3**	74.7	**94.3**	63.7	**95.8**	79.6	**95.4**	65.4
Heidke Skill Score	**90.9**	92.4	**91.5**	84.3	**93.0**	77.1	**93.9**	88.1	**84.8**	78.2
Pierce Skill Score	**97.7**	86.5	**95.3**	74.7	**94.3**	63.7	**95.8**	79.6	**95.4**	65.4

Note. Values are given for the six verification indices considered. Bold values are found to be significant, above the 95th percentile.

Table 7

Calculated FvUMSS values for Present Weather Variable

Present Weather					
	FABL	FACT	FALE	FAOR	FAPE
Proportion Correct	-248.1	-54.6	-128.1	-153.4	-123.9
Hit Rate	90.9	81.4	84.2	79.6	86.8
False Alarm Ratio	-	-	-	-	-
Critical Success Index	83.1	81.4	84.2	79.6	86.8
Heidke Skill Score	-20.3	45.8	69.5	48.8	30.3
Pierce Skill Score	83.1	81.4	84.2	79.6	86.8

Note. Positive values are in favour of the Human Forecaster and Negative values in favour of the Unified Model.

Table 8

Calculated results as a percentage for Visibility Variable for the Human Forecaster (F) and the Unified Model (U) for all Five Airports

	Visibility									
	FABL		**FACT**		**FALE**		**FAOR**		**FAPE**	
	F	**U**	**F**	**U**	**F**	**U**	**F**	**U**	**F**	**U**
Proportion Correct	**99.6**	98.7	**95.8**	91.0	**94.6**	96.5	**98.2**	94.5	**95.9**	95.7
Hit Rate	**68.8**	100.0	**59.2**	100.0	**81.1**	100.0	**68.6**	100.0	**55.9**	100.0
False Alarm Ratio	**59.3**	**99.7**	**78.4**	**98.7**	**81.1**	**98.8**	**63.2**	**99.0**	**81.7**	**99.0**
Critical Success Index	**34.4**	0.3	**18.8**	1.3	**18.1**	1.2	**31.5**	1.0	**16.0**	1.0
Heidke Skill Score	**51.0**	0.6	**30.0**	2.3	**29.0**	2.2	**47.1**	1.9	**26.1**	1.8
Pierce Skill Score	**68.4**	98.7	**55.6**	90.9	**75.9**	96.5	**67.2**	94.5	**52.4**	95.7

Note. Values are given for the six verification indices considered. Bold values are found to be significant, above the 95th percentile.

Table 9

Calculated FvUMSS values for Visibility Variable

	Visibility				
	FABL	**FACT**	**FALE**	**FAOR**	**FAPE**
Proportion Correct	66.3	53.4	-53.5	67.6	4.3
Hit Rate	-	-	-	-	-
False Alarm Ratio	40.6	20.6	18.0	36.2	17.6
Critical Success Index	34.2	17.8	17.2	30.8	15.2
Heidke Skill Score	50.7	28.3	27.4	46.1	24.7
Pierce Skill Score	-2418.9	-390.7	-588.1	-492.6	-1005.5

Note. Positive values are in favour of the Human Forecaster and Negative values in favour of the Unified Model.

Table 10

Calculated results as a percentage for Cloud Amount Variable for the Human Forecaster (F) and the Unified Model (U) for all Five Airports

Cloud Amount										
	FABL		**FACT**		**FALE**		**FAOR**		**FAPE**	
	F	**U**	**F**	**U**	**F**	**U**	**F**	**U**	**F**	**U**
Proportion Correct	97.9	97.9	99.5	99.5	**99.1**	**99.1**	98.9	98.9	98.6	98.6
Hit Rate	97.7	99.4	99.4	99.4	**99.1**	**99.1**	98.8	98.8	98.4	98.4
False Alarm Ratio	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	0.0
Critical Success Index	97.7	97.7	99.4	99.4	**99.1**	99.1	98.8	98.8	98.4	98.4
Heidke Skill Score	84.9	84.9	97.6	97.6	**92.7**	**92.7**	95.4	95.4	**91.9**	**91.9**
Pierce Skill Score	97.7	97.7	99.4	99.4	**99.1**	**99.1**	98.8	98.8	98.4	98.4

Note. Values are given for the six verification indices considered. Bold values are found to be significant, above the 95th percentile.

Table 11

Calculated FvUMSS values for Cloud Amount Variable

Cloud Amount					
	FABL	**FACT**	**FALE**	**FAOR**	**FAPE**
Proportion Correct	0.0	0.0	0.0	0.0	0.0
Hit Rate	-294.8	0.0	0.0	0.0	0.0
False Alarm Ratio	-	-	-	-	-
Critical Success Index	0.0	0.0	0.0	0.0	0.0
Heidke Skill Score	0.0	0.0	0.0	0.0	0.0
Pierce Skill Score	0.0	0.0	0.0	0.0	0.0

Note. Positive values are in favour of the Human Forecaster and Negative values in favour of the Unified Model.

Table 12

Calculated results as a percentage for Cloud Base Variable for the Human Forecaster (F) and the Unified Model (U) for all Five Airports

	Cloud Base									
	FABL		FACT		FALE		FAOR		FAPE	
	F	U	F	U	F	U	F	U	F	U
Proportion Correct	81.5	62.5	67.3	53.7	53.6	59.8	**64.3**	60.8	53.5	**56.3**
Hit Rate	6.3	48.9	18.8	48.9	**49.5**	**44.6**	**58.5**	37.4	47.5	**44.0**
False Alarm Ratio	70.0	**64.6**	**69.4**	75.6	81.0	**79.3**	**54.3**	56.3	64.5	**59.7**
Critical Success Index	5.5	17.1	**13.2**	19.4	**15.9**	16.5	34.5	25.2	25.5	**26.6**
Heidke Skill Score	4.8	4.8	4.0	3.0	2.4	**5.3**	23.8	11.4	3.7	**7.2**
Pierce Skill Score	3.3	4.4	3.5	4.1	4.0	**7.7**	**25.5**	**11.0**	4.0	**7.3**

Note. Values are given for the six verification indices considered. Bold values are found to be significant, above the 95th percentile.

Table 13

Calculated FvUMSS Values for Cloud Base Variable. Positive Values are in Favour of the Human Forecaster and Negative Values in Favour of the Unified Model

	Cloud Base				
	FABL	FACT	FALE	FAOR	FAPE
Proportion Correct	50.6	29.2	-15.5	8.9	-6.5
Hit Rate	-83.2	-58.7	8.7	33.7	6.3
False Alarm Ratio	-8.4	8.2	-2.1	3.6	-8.1
Critical Success Index	-14.0	-7.7	-0.6	12.4	-1.6
Heidke Skill Score	0.0	1.0	-3.0	14.0	-3.7
Pierce Skill Score	-1.2	-0.6	-4.1	16.3	-3.5

Table 14

Calculated Results as a Percentage for the Totals of Variables for the Human Forecaster (F) and the Unified Model (U) for all Five Airports

	Totals									
	FABL		**FACT**		**FALE**		**FAOR**		**FAPE**	
	F	**U**	**F**	**U**	**F**	**U**	**F**	**U**	**F**	**U**
Proportion Correct	**94.7**	95.7	**90.8**	87.3	**89.4**	89.9	**93.0**	93.6	**88.7**	88.0
Hit Rate	**93.8**	77.9	**89.2**	77.9	**89.7**	81.7	**91.1**	81.2	**89.4**	75.2
False Alarm Ratio	**0.9**	**4.8**	**3.4**	**17.2**	**9.9**	**17.2**	**3.4**	**9.7**	**5.4**	**11.2**
Critical Success Index	**93.0**	78.2	**86.5**	67.0	**81.6**	69.9	**88.3**	74.7	**85.1**	68.7
Heidke Skill Score	**86.5**	85.2	**80.3**	71.0	**78.8**	75.2	**85.8**	81.4	**73.2**	72.7
Pierce Skill Score	**91.2**	80.5	**83.2**	69.9	**78.8**	74.9	**86.7**	78.5	**76.4**	70.1

Note. Values are given for the six verification indices considered. Bold values are found to be significant, above the 95 percentile.

Table 15

Calculated FvUMSS Values for the Totals of the Variables

	Totals				
	FABL	**FACT**	**FALE**	**FAOR**	**FAPE**
Proportion Correct	-24.6	27.6	-5.0	-8.5	5.9
Hit Rate	72.2	51.4	43.7	52.9	57.4
False Alarm Ratio	80.7	80.3	42.2	65.5	51.9
Critical Success Index	68.0	59.1	39.0	54.0	52.4
Heidke Skill Score	8.5	32.2	14.4	23.7	2.0
Pierce Skill Score	54.8	44.2	15.5	38.2	20.9

Note. Positive values are in favour of the human forecaster and negative values in favour of the Unified Model.