Forecast Verification for the African Severe Weather Forecasting Demonstration Projects

World Meteorological Organization

Weather • Climate • Water

WMO-No. 1132

Forecast Verification for the African Severe Weather Forecasting Demonstration Projects

World Meteorological Organization

Weather · Climate · Water

WMO-No. 1132

2014

EDITORIAL NOTE

METEOTERM, the WMO terminology database, may be consulted at http://www.wmo.int/pages/prog/lsp/meteoterm_wmo_en.html. Acronyms may also be found at http://www.wmo.int/pages/themes/acronyms/index_en.html.

WMO-No. 1132

Chairperson, Publications Board
World Meteorological Organization (WMO)
7 bis, avenue de la Paix
P.O. Box 2300
CH-1211 Geneva 2, Switzerland

Tel.: +41 (0) 22 730 84 03
Fax: +41 (0) 22 730 80 40
E-mail: publications@wmo.int

ISBN 978-92-63-11132-6

CONTENTS

Page

ACKNOWLEDGEMENTS

After starting his career as an operational weather forecaster, Mr Laurence Wilson has worked in the research branch of the Meteorological Service of Canada for over 30 years, specializing in the methods of forecast verification and statistical post-processing of numerical model forecasts. This work was often accompanied by training activities, both inside the Canadian Weather Service and internationally through the World Meteorological Organization (WMO). Mr Wilson led the preparation of the first general practical guide to verification, entitled *A Survey of Common Verification Methods in Meteorology* (WWW-8, WMO/TD-No. 358), which was published in 1989, and is the author of numerous research papers in verification and statistical interpretation of numerical model forecasts. Mr Wilson was engaged in the WMO Severe Weather Forecasting Demonstration Project to introduce forecast verification as a key activity in the project.

FORECAST VERIFICATION IN THE AFRICAN SEVERE WEATHER FORECASTING DEMONSTRATION PROJECTS

1. INTRODUCTION – PRINCIPLES AND IMPORTANCE OF VERIFICATION

Allan Murphy, who built his scientific career on the science of verification, has said: "Verification activity has value only if the information generated leads to a decision about the forecast or system being verified." This immediately suggests that there must be a user for the verification output, someone who wants to know something specific about the quality of a forecast product, and who is in a position to make a decision based on verification results. The user could be, for example, a forecaster who is provided with the output from several models on a daily basis and wishes to know which of the models can be relied on most for forecast guidance. Or, the user could be the manager of a project such as the WMO Severe Weather Forecasting Demonstration Project (SWFDP), who wishes to know whether the increased access to model guidance products is leading to a measurable improvement in forecasts issued by a National Meteorological and Hydrological Service (NMHS).

1.1 Purposes of verification

In general, different users of verification results will have quite different needs, which means that the target user or users must be known before the verification system is designed, and also that the verification system design may need to be varied or broadened to ensure that the needs of all the users can be met. To summarize briefly, the first principle of verification is: Verification activity has value only if the information generated leads to a decision about the forecast or system being verified. Thus, the user and the purpose of the verification must be known in advance.

Purposes of verification can be classified as either administrative or scientific or, rarely, a combination of both. Administrative verification goals include justifying the cost of a weather service or the cost of new equipment, or monitoring the quality of forecasts over long periods of time. Administrative verification usually means summarizing the verification information into as few numbers as possible, using scoring rules. Scientific verification, on the other hand, means identifying the strengths and weaknesses of a forecast in enough detail to be able to make decisions about how to improve the product, that is, to direct research and development activity. Scientific verification therefore means more detail in the verification methodology, and less summarizing of the verification information. The term diagnostic verification is often applied to verification with specific scientific goals; for example, "Does the European Centre for Medium-Range Weather Forecasts (ECMWF) model forecast extreme precipitation more accurately than the National Centers for Environmental Prediction (NCEP) model, and under what conditions?"

For the SWFDP, it is fair to say that verification needs to be done for both main purposes, administrative and scientific. At the administrative level, the need is to demonstrate the impact of the project in terms of improved operational forecasting services. It might also be to demonstrate improvements in forecast quality, although this implies that there exists some objective measure of forecast quality. At the scientific level, the main need is to establish the level of accuracy of severe weather forecasts and to determine the accuracy of the various guidance products.

1.2 Three main principles of verification

The above discussion of purposes of verification can be summarized into a first principle of verification: The user and purpose of the verification must be known in advance. Preferably, users and purposes should be defined in great detail, as specifically as possible. It is useful to actually state the purpose beforehand, for example: "To determine whether the NMHS forecasts are increasing in accuracy with the introduction of Regional Specialized Meteorological Centres (RSMC) guidance forecasts of extreme precipitation."

A second principle of verification is that no single verification measure provides complete information about the quality of a forecast product. Scores that are commonly used in verification are limited in the sense that they measure only a specific aspect or attribute of the forecast quality. Use of a single score by itself can lead to misleading information; the forecast can be improved according to the score, but at the same time the performance degraded in other ways not measured by the score. Thus it is advisable to use two or more complementary scores to obtain a more complete picture of the forecast accuracy.

A third principle of verification is that the forecast must be stated in such a way that it is verifiable, which implies a completely clear statement about the exact valid time or valid period of the forecast, and the location or area for which the forecast is valid, along with the nature of the predicted event. For example, "Rain accumulations of more than 50 mm are expected in the southern half of Madagascar tomorrow." is a verifiable forecast if both forecasters and users know to what southern half refers and exactly what are the hours of tomorrow (Is it 00 UTC to 00 UTC, 06 UTC to 06 UTC, or defined as the 24-hour day in local time?).

1.3 Verification as a component of quality assurance of forecast services

Verification is actually only one aspect of the overall goodness of a forecast. Verification is usually understood to mean the evaluation of the quality of the forecast, by objectively measuring how well the forecast corresponds with the actual weather, as revealed by observations. Another aspect of forecast goodness, no less important, is its value. Value is defined as the increase or decrease in economic or other benefit to the user, resulting from using the forecast. The assessment of value requires specific quantitative information on the consequences to the user of taking action on the basis of the forecast, in addition to verification information. Value is most often objectively assessed using methods of decision theory such as cost–loss analysis. In the context of the SWFDP, forecast value accrues mostly in the form of reduction of risk to life and limb arising from severe weather events, which could be assessed subjectively in consultation with disaster management organizations in the SWFDP countries. The present discussion is limited to the verification aspects of forecast goodness.

Along with the evaluation of forecast goodness, verification is an integral part of the quality assurance of a forecast and warning production system. A complete evaluation system might also include efforts to answer questions such as: "Are the forecasts issued in time to be useful?" (timeliness); "Are the forecasts delivered to the users in a form they can understand and use?" (relevance); and "Are the forecasts always delivered on time?" (robustness). Efforts to answer such questions imply continuing dialogue with user communities such as disaster preparedness agencies in the case of severe weather forecasts.

1.4 The importance of verification

Verification, as an activity, has always been recognized as important, an essential ingredient in the forecasting process; however, in reality it has been poorly understood and not well implemented, and often not maintained as a continuing activity.

Over the past 10 years, there has been a proliferation on the Internet of daily weather forecasts for hundreds of cities, produced by national and private forecasting centres. In many cases, they are not accompanied by information on their quality. The majority of these forecasts are interpolated automatically from the raw output of the surface weather parameters of the global models, which have not been verified or even validated (during product development), except perhaps within the region of responsibility of the issuing centre. This is very poor practice.

In the context of the SWFDP, this means that all the direct model output products made available to the project had not been verified at all for any country, a situation which has recently been changing due to SWFDP activities. Given that it is also generally known that models have systematic weaknesses in the tropics, it becomes even more risky to use these products without

verifying them. At the very least, verification results should quickly indicate which model performs most reliably as forecasting guidance.

Comprehensive verification of forecast products for the global models is probably best done at the source of the model output, since it is easiest to transfer relatively small datasets of observations to the global centre rather than to transfer much larger archives of gridded model output to the individual NMHSs for verification. Unfortunately, there are often impediments to the free transfer of observational data across national boundaries, which further hinder the ability to bring observation and global or even regional model forecast data together with corresponding observations for verification purposes. That being said, the methods presented in this publication can be applied to the output from the global deterministic models quite easily as well as to verification of the local severe weather forecasts, and forecasts from the RSMCs.

While this publication describes procedures for objective verification of SWFDP forecasts, there is a role for subjective verification, and in fact it may be difficult to completely eliminate all subjectivity from the process even in objective verification efforts. For the SWFDP, subjective verification or evaluation may be needed in data-sparse areas, and is useful for the evaluation of guidance for specific case studies of events. If subjective judgments are used in any part of the verification process, this must be stated clearly.

And lastly, this publication is about objective verification procedures for severe weather forecasts, which derive extra significance because of the need for rapid protective action. The emphasis is on assessment of the meteorological content of the forecasts, and not on the perceived or real value of these forecasts to users, or the effectiveness of the delivery of these forecasts to users, both of which require additional information to evaluate. Severe weather warnings are considered to embody the advance public alert of potentially hazardous weather, and for the purposes of the verification measures described herein, are taken to be the most complete description of the severe conditions expected, including location, start and end times and the type of severe weather expected. If a warning is not issued, it is assumed that no severe weather is expected to occur.

2. VERIFICATION PROCEDURE FOR THE SWFDP SEVERE WEATHER FORECASTS

The best procedure to follow for verification depends not only on the purpose of the verification and the users, but also on the nature of the variable being verified. For the African SWFDPs, the main forecast variables are extreme precipitation and strong winds, with extreme defined by thresholds of 30 or 50 mm in 6 hours, 30, 50 or 100 mm in 24 hours, and strong winds being defined by thresholds of 20 and 30 kt (SWFDP implementation plans specify these thresholds). These are therefore categorical variables, and verification measures designed for categorical variables should be applied. In each case, there are two categories, referring to occurrence or non-occurrence of weather conditions exceeding each specific threshold.

The following subsections describe the suggested procedures for building contingency tables and calculating scores.

2.1 Defining the event

Categorical and probabilistic forecasts always refer to the occurrence or non-occurrence of a specific meteorological event. The exact nature of the event being predicted must be clearly stated, so that the user can clearly understand what is being predicted and can choose whether to take action based on the forecast. The event must also be clearly defined for verification purposes, specifically as follows:

– The location or area of the predicted event must be stated;

– The time range over which the forecast is valid must be stated;

– The exact definition of the event must be clearly stated.

Sometimes these aspects will be defined at the beginning of a season or the beginning of the provision of the service and will remain constant, for example, the establishment of fixed forecast areas covering the country. As long as this information is communicated to the forecast user community, then it would not be necessary to redefine the area to which a forecast applies unless the intent is to subdivide the standard area for a specific forecast.

The time range of forecast validity has been established as part of the project definition, for example, 6-h and 24-h total precipitation, and wind maxima over 24 hours. The 24-h period also needs to be stated (the UTC day, 00 to 24, the climatological day, for example, 06 to 06 UTC, or the local time day, 00 to 24. The definition which corresponds to the observation validity period needs to be used for verification.

For the SWFDP, it would be best if the larger countries were to be divided geographically into fixed (constant) areas of roughly the same size, areas which are climatologically homogeneous. Each region should have at least one reporting station. The smaller the area size, the more the forecast is potentially useful. However, the predictability is lower for smaller areas, giving rise to a lower hit rate and higher numbers of false alarms and missed events (terminology is defined in section 2.2 below), that is, more difficult to make a good prediction. The sparseness of observational data also imposes constraints on the subdivision of areas. A forecast cannot be verified without relevant observations. On the other hand, larger areas make the forecasts potentially less useful, for example, to disaster management groups or other users who need detailed enough location information associated with the predicted severe weather to effectively deploy their emergency resources, or to implement effective protective or emergency actions.

To summarize, in choosing the size and location of fixed domains for severe weather warnings, several criteria should be taken into account:

(a) The location and readiness of disaster relief agencies: The domains should be small enough that disaster relief agencies can respond effectively to warnings within the lead time that is normally provided.

(b) The availability of observation data: Each domain should have at least one representative and reliable observation site for forecast verification purposes.

(c) Climatology/terrain type: It is most useful to define regions so that they are as climatologically homogeneous as possible. If there are parts of the domain that are much more likely to experience severe weather than others, these could be kept in separate regions.

(d) Severe weather impacts: The domain locations and sizes should take into account factors affecting potential impacts such as population density and disaster-prone areas.

Within these guidelines, it is also useful if the warning areas are roughly equal in size, as that will help ensure consistent verification statistics. Also, within each country, the warning criteria should be constant for all domains. Finally, for the purposes of the African SWFDPs, and for possible comparisons with the results of verification of the global model forecasts over multiple countries, it would be useful if the subdomains in all countries would be roughly similar in size.

2.2 Preparing the contingency table

The first step in almost all verification activity is to collect a matched set of forecasts and observations. The process of matching the forecast with the corresponding observation is not always simple, but a few general guidelines can be stated. If the forecast event and the forecast

are clearly stated, then it is much easier to match with observations. For the SWFDP, the forecast event is the expected occurrence of severe weather conditions somewhere in the forecast area, sometime during the valid time period of the forecast. Then:

– A "hit" is defined by the occurrence of at least one observation of severe weather conditions, as defined by the thresholds anywhere in the forecast area, any time during the forecast valid time. Note that by this definition, more than one report of severe weather within the forecast valid area and time period does not add another event; only one hit is recorded.

– A "false alarm" is recorded when severe weather is forecast, but there is no severe weather observed anywhere in the for which the forecast is valid during the valid period.

– A "missed event" is recorded when severe weather is reported outside the area and/or the time period for which the warning is valid, or whenever severe weather is reported and no warning is issued. Only one missed event is recorded on each day, for each region where severe weather has occurred that is not covered by a warning.

– A "correct negative" or "correct non-event" is recorded for each day and each fixed forecast region for which no warning is issued and no severe weather is reported.

If observational data are sparse, it may be difficult to determine whether severe weather occurred, as there is much space between stations for smaller scale convective storms which characterize much of the severe weather occurrences. It is permissible to use proxy data such as reports of flooding to infer the occurrence of severe weather in the absence of observations, but full justification of these subjective decisions must be included with verification reports.

It is possible to incur missed events, false alarms and hits all at once. Consider the following example, represented schematically in Figure 1.

In Figure 1, the yellow regions represent forecast severe weather areas and the stars represent observations of severe weather; O represents observations of non-severe weather. This case contains one hit (because there are observations of severe weather within the forecast severe weather area), one miss (because there are one or more observations of severe weather that do not lie in a forecast severe weather area) and one false alarm (because there is no severe weather reported in a severe weather forecast area). Note that a false alarm is recorded only because there is a separate forecast area with no report of severe weather. The fact that not all the stations in the larger area reported severe weather does not matter; only one severe weather report is needed to score a hit. If there are no reporting stations in a forecast severe weather area, then forecasts for that area cannot be verified.

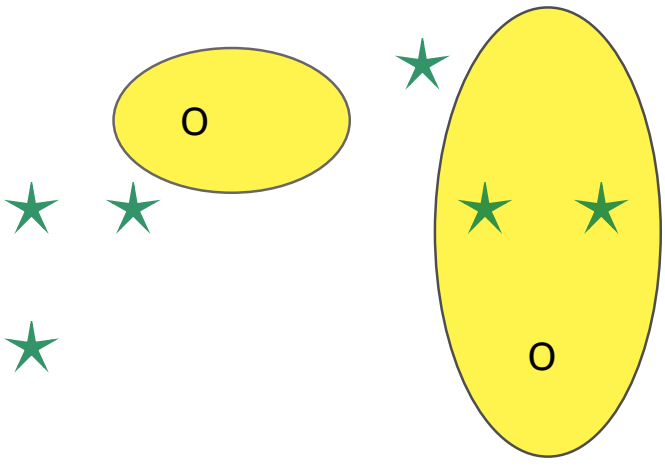

Figure 1. Schematic showing the matching of forecast severe weather threat areas with point precipitation observations

In this system, the number of hits cannot be increased by increasing the size of the forecast area. However, increasing the size of the forecast area might reduce the chance of a missed event. This should be kept in mind. If the size of the forecast severe weather area is increased merely to reduce the chance of a missed event, the forecast also becomes less useful, because disaster mitigation authorities may not know where to deploy their resources to assist the needy. Each NMHS must seek to achieve its own balance between scale (size) of forecast areas and risk of false alarms and missed events.

A contingency table illustrated in Figure 2 is then produced by totalling up the number of hits, misses, false alarms and correct negatives for a sufficiently large number of daily cases. Since the nominal verification period is one day, it makes sense to record a single case for each day and each fixed geographical region of each country. If more than one result is recorded for a particular day's forecast, for example, both a hit and a false alarm, then the result for that day should be divided by the number of different outcomes, 2 or 3. The result is the addition of 1 case to the totals of a, b, c and/or d for each day, though the 1 case may be partitioned over 2 or 3 boxes of the table. The sum total of the table in the bottom right corner will then equal the number of days times the number of separate geographical parts of the country for which observation data were available.

It might be most convenient to make two columns of ones and zeros, one each for the forecast and the observation. Then the logic functions of Excel, for example, can be used to automatically produce the totals of a, b, c and d over a sample of cases. A table which is built this way could include several columns for forecasts from different sources, for example, the RSMC guidance, and the model output from each of the global centres. Each forecast, when combined with the observations, would lead to a different table. The different tables could be scored to give comparative results. Some examples of Excel spreadsheets are available with the electronic version of this publication:

Event forecast	Event observed		
	Yes	No	Marginal total
Yes	Hit	False alarm	Fc Yes
No	Miss	Correct non-event	Fc No
Marginal total	Obs Yes	Obs No	Sum total

Event forecast	Event observed		
	Yes	No	Marginal total
Yes	a	b	a + b
No	c	d	c + d
Marginal total	a + c	b + d	a + b + c + d = n

Figure 2. The contingency table for dichotomous (yes–no) events

(a) ECMWF CT calculator – deterministic model forecasts for eastern African locations, matched to observations from eastern African countries that were available on the Global Telecommunication Network (GTS) from September 2010 to May 2011.

(b) NCEP CT calculator – deterministic model forecasts for eastern African locations, matched to observations from eastern African countries that were available on the GTS from September 2010 to May 2011.

(c) CT calculator, containing a sample of NMHS severe weather forecasts and observations from Botswana.

A description of how to use these Excel files to carry out verification of forecasts for specific locations and forecast projection times may also be found with the electronic version of this publication.

2.3 Calculating scores using the contingency table

Scores that can be computed from the contingency table entries are listed in this section, along with their characteristics, strengths and weaknesses. This is not an exhaustive list of scores that can be computed from the contingency table, but those listed here are considered to be the most useful for verification of severe weather forecasts. These scores are all functions of the entries of the contingency table as shown in Figure 2 and are easily computed. The formulae shown below are incorporated into the sample Excel spreadsheet available with the electronic version of this publication.

Computation of these scores should be considered part of analysis and diagnosis functions that are routinely performed by forecasters. These scores all have specific interpretations, discussed below, which help the forecaster perform these diagnosis tasks. The scores provide the most meaningful information if they are computed from large enough samples of cases, say 100 or so. However, severe weather occurrences are rare events, thus the number of forecasts and observations of severe weather may be small (fortunately), which makes the task of verification not only more important but also more challenging.

2.3.1 *Probability of detection (PoD) (hit rate (HR) or prefigurance)*

$$PoD = HR = \frac{a}{a+c}$$

The hit rate (HR) has a range of 0 to 1 with 1 representing a perfect forecast. As it uses only the observed events a and c in the contingency table, it is sensitive only to missed events and not false alarms. Therefore the HR can generally be improved by systematically overforecasting the occurrence of the event. The HR is incomplete by itself and should be used in conjunction with either the false alarm ratio or the false alarm rate both explained below.

2.3.2 *False alarm ratio (FAR)*

$$FAR = \frac{b}{(a+b)}$$

The false alarm ratio (FAR) is the ratio of the total false alarms (b) to the total events forecast ($a + b$). Its range is 0 to 1 and a perfect score is 0. It does not include c and therefore is not sensitive to missed events. The FAR can be improved by systematically underforecasting rare events. It also is an incomplete score and should be used in connection with the HR.

2.3.3 *Frequency bias (B)*

$$Frequency\ bias = \frac{a+b}{a+c}$$

The frequency bias (B), hereinafter referred to as bias, uses only the marginal sums of the contingency table, and so is not a true verification measure, as it does not imply matching individual forecasts and observations. Rather, it compares the forecast and observed frequencies of occurrence of the event in the sample. The forecast is said to be unbiased if the event is forecast with exactly the same frequency with which it is observed, so that the frequency bias of 1 represents the best score. Values higher than one indicate overforecasting (too frequently) and values less than 1 indicate underforecasting (not frequent enough). When used in connection with the HR or the FAR, the bias can be used to explain the forecasting strategy with respect to the frequencies of false alarms or misses. Note that the bias also can be computed for the non-events, as $(c + d)/(b + d)$. If the frequency bias is computed for all the categories of the variable, then it gives an indication of the differences between the forecast and observed distributions of the variable.

2.3.4 *Threat score (TS) (critical success index, CSI)*

$$CSI = \frac{a}{a+b+c}$$

The threat score (TS), or critical success index (CSI), is frequently used as a standard verification measure, for example in the United States of America. It has a range of 0 to 1 with a value of 1 indicating a perfect score. The CSI is more complete than the HR and FAR because it is sensitive to both missed events and false alarms. Thus it is harder to adopt a systematic forecasting strategy that is guaranteed to improve the score. It does, however, share one drawback with many other scores: it tends to go to 0 as the event becomes rarer. This score is affected by the climatological frequency of the event; if forecasts need to be compared (for example, same forecasts from different sources) using this score, but based on different verification samples, it might be wiser to use the equitable threat score (ETS), which adjusts for the effects of differences in the climatological frequencies of the event between samples. For evaluation of a forecast or for comparison of the accuracy of forecasts based on the same dataset, the CSI is a good general score. The ETS is given by:

$$ETS = \frac{a-a_r}{a+b+c-a_r}$$

$$a_r = \frac{(a+b)(a+c)}{T}$$

where T is the sample size. The quantity a_r is the number of forecasts expected to be correct by chance, by just guessing the category to forecast.

2.3.5 *The Heidke skill score (HSS)*

$$HSS = \frac{(a+d) - \dfrac{(a+b)(a+c)+(c+d)(b+d)}{T}}{T - \dfrac{(a+b)(a+c)+(c+d)(b+d)}{T}}$$

In verification, the term skill has a very specific meaning: Skill is the accuracy of a forecast compared with the accuracy of a standard forecast. The standard forecast is usually chosen to be a forecast which is simple to produce, and may already be available to users. The idea of a skill

score is to demonstrate whether the forecast offers an improvement over the choice of an unskilled standard forecast.

The Heidke skill score (HSS) uses the number correct for both categories to measure accuracy, and the standard forecast is a simple random guess which of the two categories will occur. The score is in the format:

$$HSS = \frac{Number_{correct} - Number_{chance}}{Total - Number_{chance}}$$

where the number correct by chance is the total number of forecasts, both severe and non-severe, that would be expected to be right just by random guessing. When forecasting severe weather, a guess could be made of which of the two categories would occur, like tossing a coin. Anyone can do this; there is no need to be a good forecaster. Yet, some of these guesses would by chance be correct. The idea of the Heidke skill score is to adjust for the number of forecasts that would be correct just by guessing.

The number correct by chance is defined in the same way as for the ETS, but both categories are used. The number of forecasts correct is simply the sum of the diagonal elements of the contingency table, (a + d).

The HSS ranges from negative values to +1. Negative values indicate that the standard forecast is more accurate than the forecast; skill is negative and a better score would have been obtained by just guessing what the forecast should be. The HSS represents the fraction by which the forecast improves on the standard forecast. A perfect forecast gives an HSS of 1, no matter how good the standard forecast is.

The HSS defined this way is the easiest to apply and use. All the information needed is contained in the contingency table. It turns out that pure chance offers a pretty low standard of accuracy. It is quite easy to improve on a chance forecast. Other standards of comparison are persistence ("no change from the observed weather at the time the forecast was issued" or "what you see is what you get") or climatology, which for a categorical forecast is defined as the most likely of the two categories. That is, a climatological forecast is a forecast of no severe weather all the time. This would not be a very useful forecast, but it would score well on most scores since (fortunately) no severe weather occurs much more often than severe weather. In the contingency table, d is much larger than a, b or c. A climatological forecast of no severe weather may therefore be difficult to beat. In practice, though, persistence and climatology are not often used in the HSS, because a separate contingency table for the reference forecast must be compiled.

2.3.6 *The false alarm rate (FA)*

$$FA = \frac{b}{(b+d)}$$

The false alarm rate (RA) is unfortunately often confused with the false alarm ratio. The false alarm rate is simply the fraction of observed non-events that are false alarms. By contrast, the false alarm ratio is referenced to the total number of forecasts; it is the fraction of forecasts that were false alarms. The best score for the FA is 0, that is, the wish is to have as few false alarms as possible. The FA is not often used by itself but rather is used in connection with the HR in a comparative sense. The HR is also referenced to the observations, specifically, the total number of observed events.

2.3.7 *The Hanssen–Kuipers score (KSS) (Pierce score) (true skill statistic (TSS))*

$$KSS = TSS = (HR - FA) = \frac{(ad - bc)}{(a + c)(b + d)}$$

The Hanssen–Kuipers score (KSS), also knows as the true skill statistic (TSS), is easiest to remember as the difference between the hit rate and the false alarm rate, as defined in 2.3.1 and 2.3.6, respectively. This score measures the ability of the forecast to distinguish between occurrences and non-occurrences of the event. The best possible score value is 1, which is obtained when the HR is 1 and the FA is 0. If *HR* = *FA*, then the score goes to 0, which is the worst value possible.

This score is used to indicate whether the forecast is able to discriminate situations that lead to the occurrence of the event from those that do not. If, for example, the forecaster attempts to improve the hits by forecasting the event more often, this score will indicate whether too many false alarms are incurred by doing so. The idea is to increase the HR without increasing the FA too much.

One disadvantage of this score for rare events is that it tends to converge to the HR because the value of d becomes very large.

2.3.8 *The extreme dependency family of scores (SEDS, EDI and SEDI)*

$$SEDS = \frac{\left[\log\left(\frac{(a+b)}{T}\right) - \log\left(\frac{a}{a+c}\right)\right]}{\left[\log\left(\frac{(a+c)}{T}\right) + \log\left(\frac{a}{a+c}\right)\right]}$$

$$EDI = \frac{\left[\log\left(\frac{b}{b+d}\right) - \log\left(\frac{a}{a+c}\right)\right]}{\left[\log\left(\frac{b}{b+d}\right) + \log\left(\frac{a}{a+c}\right)\right]}$$

$$SEDI = \frac{\left[\log\left(\frac{b}{b+d}\right) - \log\left(\frac{a}{a+c}\right) - \log\left(1-\frac{b}{b+d}\right) + \log\left(1-\frac{a}{a+c}\right)\right]}{\left[\log\left(\frac{b}{b+d}\right) + \log\left(\frac{a}{a+c}\right) + \log\left(1-\frac{b}{b+d}\right) + \log\left(1-\frac{a}{a+c}\right)\right]}$$

These are quite new scores, all described and analysed in a paper published in 2011. They are successors to a score called the extreme dependency score (EDS), which was published earlier but has since been shown to have some less-desirable properties compared with these newer scores. All of the EDS score family is designed to apply to the verification of rare (infrequent) events, exactly the type of extreme weather of concern in the SWFDP. Several of the other scores shown above have a tendency to go to small values near 0 when the event in question occurs infrequently. This can make it difficult to determine differences in performance, or to track improvements in the performance of a forecast system. The three new scores described in this section generally do not have this property, and therefore are more sensitive to real changes in accuracy of the forecast for rare events.

As these scores are new and are just beginning to be used in verification activities, the meaning of the values obtained is still being explored. Computation of these scores in the SWFDP provides an opportunity to contribute to their understanding in the meteorological community. Some aspects of their behaviour can be discerned from the equations. First, they are all ratios of logarithms,

which may seem, but is not really, complicated. As ratios, it does not matter whether natural logarithms or logarithms to base 10 are used; the results will be the same. Secondly, they all use the more familiar contingency table quantities illustrated in Figure 2.

The extremal dependency index (EDI) is the difference of the log of the false alarm rate and the log of the hit rate, divided by the sum of the logs of the false alarm rate and hit rate. As this score depends only on the hit rate and false alarm rate, it is related to the Hanssen–Kuipers score, which is merely the difference between the HR and the FA. This also means that it relates to the same forecast attribute as the Hanssen–Kuipers score, discrimination. The EDI is therefore of use when the aim is to assess the quality of the forecast for discriminating the antecedent conditions leading to the occurrence of extreme weather from those which do not.

The symmetric extremal dependency index (SEDI) is similar to the EDI; the added terms make this score "symmetrical" in the sense that relabelling the forecasts of the events as non-events and vice versa leads to the same value of the score, but negative. This is a rather theoretical property which is not often important, and unlikely to be important in the practice of the SWFDP, so the computation of the SEDI is not necessary; the score is included only for completeness.

The stable extreme dependency score (SEDS) is different from the others in that it uses something called the forecast frequency, that is, the number of times the event is forecast divided by the total number of cases in the verification sample. This particular score should be of interest to forecasters, because they can control the forecast frequency. A strategy of forecasting the event more often, for example, will increase this score only if the hit rate is increased proportionately more than the false alarms.

2.4 Interpreting the scores

This raises the question of whether it is worth the effort to compute all these scores, or even most of them. Once the table is prepared, then the scores are easily computed in any case, each requiring only one equation (on a spreadsheet, for example) to compute from the entries of the contingency table. More importantly, however, the different scores measure different aspects of forecast quality, and the use of several scores permits these different aspects or attributes to be assessed. This section discusses aspects of the interpretation of the different scores.

2.4.1 *Attributes of the forecast measured by the scores – accuracy, skill and discrimination*

The scores defined above can be grouped according to which attributes of the forecast they measure. The HR, FAR, TS, ETS and SEDS measure accuracy. As an attribute, the accuracy of the forecast is simply the level of agreement between forecasts and observations. These scores all measure accuracy in slightly different ways, and are especially useful in different situations. For example, it is best to use the ETS when the purpose is to compare results on different samples, since differences in the observed frequency of the event are taken into account. Both the HR and the FAR can be improved by altering the forecasting strategy, so should not be used alone. The SEDS may be more useful than other scores if the observed frequency of the event in the sample is small.

The frequency bias measures the characteristics of the distribution of forecasts, compared with the distribution of observations, as mentioned above. It is more of a diagnostic tool for the forecast rather than a true verification measure.

The HSS measures the attribute skill, as defined above.

The HR and FA, when used together, and the KSS measure the attribute discrimination; in fact, the KSS is sometimes called the Hanssen–Kuipers discriminant. While the accuracy and skill attributes are of particular importance to forecasters in deciding their forecasting strategy,

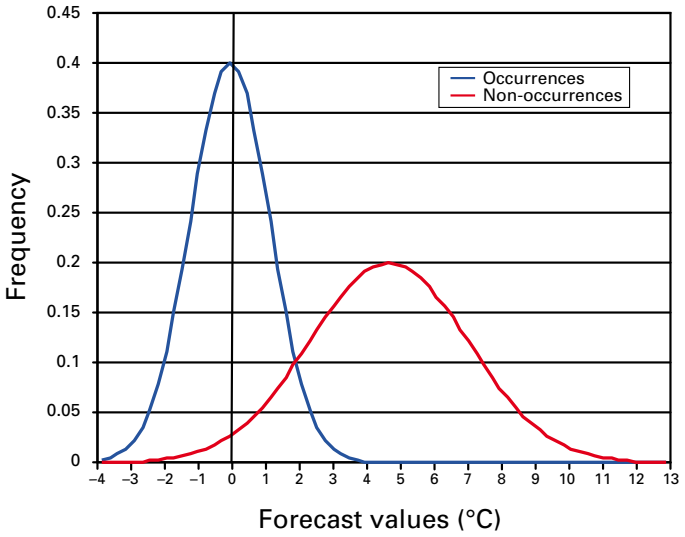

Figure 3. Concept of discrimination; plot of the frequency of forecasts of temperatures when the event "temperature lower than 0" occurred (blue curve) and did not occur (red curve)

discrimination is an attribute that relates more to the needs of the user of the forecast. Measures of discrimination tell users whether they can rely on the forecast to identify hazardous situations. The following example illustrates the concept.

Suppose a user is interested in knowing whether the minimum temperature will be below freezing (<0°C). Figure 3 shows a set of temperature forecasts, divided into two groups. The red curve shows the frequency of forecast temperatures when the observed minimum temperature was above freezing (non-occurrences of the event) and the blue curve shows the frequency of forecast temperatures when the observed minimum temperature was below freezing (occurrences). It can be seen from Figure 3 that when the observed minimum was above freezing, most of the forecasts were above freezing. There is just a small tail of the "red" distribution where forecasts are below freezing (false alarms). On the other hand, about half of the time when below freezing temperatures occurred, the forecast was for above freezing (missed events). If a user receives a temperature forecast of +1°C for example, Figure 3 shows that more often than not the actual minimum temperature was below 0°C (the blue curve is higher than the red curve at temperatures around 1°C). On the other hand, forecasts above about +4°C always verified (blue curve near 0°C), and forecasts of minimum temperatures colder than –0.5°C also nearly always verified (red curve near 0°C). It is the area of overlap of the two curves which is of concern to the user. The larger this area, the harder it becomes for the users to be confident in their use of the forecast.

The amount of separation of the two curves is in fact a measure of the ability of the forecast system to discriminate between the two categories. However, the important overlap region, where the user would be unsure of the forecast, should be as small as possible. For a particular separation of the two categories, the overlap is minimized if the variation (variance) of the forecast distributions is small. In summary, the usefulness of the forecast for decision-making by a user depends on the ability of the forecast system to discriminate events from non-events. This is measured by comparing the hit rate and the false alarm rate (not ratio).

2.4.2 *An example of discrimination/decision-making ability when risk information is included in the forecast.*

For Madagascar, the risk forecasts from RSMC Pretoria were verified for the period November 2008–June 2009 from a user perspective. The guidance forecasts include an estimate of low, medium or high risk. Using each of these risk estimates as a threshold for forecasting the occurrence of severe weather (more than 50 mm rain in 24 hours), three contingency tables can be obtained (see Table 1). There are 211 cases in total.

Table 1. Three contingency tables for forecasts of the occurrence of 24-h precipitation greater than 50 mm for Madagascar: issue a warning if the RSMC forecast indicates at least low risk (top) and at least medium risk (middle), and only issue a warning if the RSMC forecast indicates high risk (bottom)

Low	*Obs yes*	*Obs no*	*Totals*
Fcst yes	35	34	69
Fcst no	15	127	142
Totals	50	161	211
Med	*Obs yes*	*Obs no*	*Totals*
Fcst yes	31	18	49
Fcst no	19	143	162
Totals	50	161	211
High	*Obs yes*	*Obs no*	*Totals*
Fcst yes	13	4	17
Fcst no	37	157	194
Totals	50	161	211

Table 1 shows that the use of a more restrictive threshold for forecasting the event (high risk only) reduces the number of hits but also reduces the number of false alarms, while the number of misses increases significantly. To determine whether the false alarms are being reduced enough to be useful, the hit rate can be plotted against the false alarm rate, as shown in Figure 4.

This is called a relative operating characteristic (ROC) plot. The ROC diagrams have been widely used in many fields, for example, determining the ability of an X-ray picture to show a pathology clearly enough that a doctor can see it in the presence of a noisy background. In the present application, the noisy background is the imperfect model guidance forecast and the pathology is the severe weather event attempted to be forecast. Brought into meteorology in 1982, the ROC diagram is now widely used to verify ensemble probability forecasts.

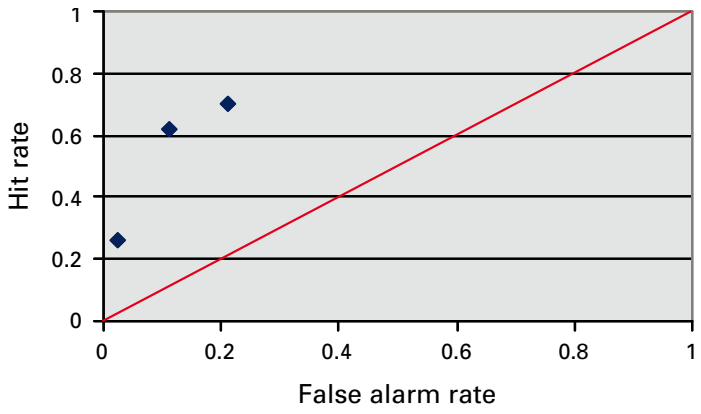

Figure 4. A ROC plot for the Madagascar forecasts showing discrimination. No matter which threshold is used, the forecasts show at least some ability to separate active days from non-active days, and show no discrimination if the hit rate = false alarm rate, along the diagonal line.

Shown in Figure 4, for the Madagascar/RSMC Pretoria data, are the three points obtained by plotting the HR versus the FA for each of the three contingency tables in Table 1. The fact that the three points remain above the diagonal line is most important here. This means that, whatever the threshold chosen, the HR is always greater than the FA, and the forecast is able to distinguish situations leading to severe weather from those that do not lead to severe weather, with some skill. If the points lay on the red diagonal line, the user would not be able to distinguish occurrences from non-occurrences on the basis of the forecast, and the forecast would be completely useless for decision-making. The closer the points are to the upper left corner (HR=1 and FA=0), the better the discriminating ability of the forecast.

2.4.3 *The troublesome "d"*

The number of correct negatives, d, is often hard to define in the contingency table. The most common problem is the sparseness of observations for determining severe weather occurrences. As severe weather often happens over a relatively small area, it is often not known whether "no report" of severe weather is a non-occurrence, or an occurrence that is missed by the observations. The effect on the contingency table is possibly to cause hits to be reported as false alarms (forecast but not seen), and to cause missed events to be reported as correct negatives (not forecast and not seen, but occurred).

It is also difficult to define the spatial and temporal boundaries of the non-event. The option proposed in this publication is to allow one correct negative for each specific forecast region per day, since the predicted variable is accumulated over 24 hours.

Some of the scores defined above do not use d from the contingency table. These can be emphasized when there are doubts about the accuracy of the table entries because of missed observations. These scores are the HR, FAR, TS and ETS.

3. EXAMPLE – APPLICATION AND INTERPRETATION OF CONTINGENCY TABLE RESULTS

Table 2 shows data for Botswana. The tables and the scores were computed automatically from the dataset of matched observations and forecasts of the event (observation = 0 or 1 and forecast = 0 or 1, respectively), as entered into a version of the CT calculator spreadsheet available with the electronic version of this publication.

Table 2 shows two contingency tables for forecasts of >50 mm precipitation for Botswana, for the period November 2008 to March 2009. The top table was created from the list of events provided by the Botswana Meteorological Services. In this case, each observation of severe weather was defined as a severe event, while inactive days were assigned one event each. For the middle table, days with multiple observations of severe conditions were weighted to match the weight for inactive days, so that each day totals to one event. This reduced the total to 131 events, which is the total number of days covered by the sample. Entries in the table have been rounded to the nearest whole number for simplicity. Of the 131 events, 43 were severe weather occurrences and 88 were inactive days. The scores were computed for both versions of the table; the differences in the results were not very large in general. For the interpretation, the middle table is emphasized though most comments also apply to the top table.

First, note the frequency bias. The severe weather event was predicted only a little more than half as often as it occurred (24 forecasts versus 43 occurrences). The hit rate (0.46) probably could be increased by forecasting the event more often, but the low false alarm ratio (.15) might also rise. If false alarms are to be avoided (so that users will be sure to always heed the forecast, for example), then it may be desirable to keep the false alarms low even at the expense of higher missed events (23).

Table 2. Contingency tables for a set of severe precipitation forecasts for Botswana, along with the scores[a] for these tables. The lower table and the right-hand column show results when forecasts are weighted so that one day produces one event.

Contingency table – Botswana original			
	OBS YES	OBS NO	
FCST YES	26	5	31
FCST NO	27	84	111
	53	89	142

Contingency table – Botswana weighted			
	OBS YES	OBS NO	
FCST YES	20	4	24
FCST NO	23	84	107
	43	88	131

Scores	*Unweighted*	*Weighted*
Per cent correct	0.77	0.79
Hit rate	0.49	0.46
False alarm rate	0.06	0.04
Frequency bias	0.58	0.55
False alarm ratio	0.16	0.15
Threat score	0.45	0.42
Equitable threat score	0.31	0.31
No. correct by chance	81	80
Fraction correct by chance	0.57	0.61
Heidke skill	0.47	0.47
Hanssen–Kuipers score	0.43	0.42
Extreme dependency score	0.16	0.18
Stable extreme dependency score	0.36	0.38
Extremal dependency index	0.60	0.61
Symmetric extremal dependency index	0.64	0.64

[a] The CT calculator, an Excel file available with the electronic version of this publication, contains the equations to compute the scores.

Next, consider the hit rate, the FAR and the TS together. The FAR is quite low in this case, the hit rate (.46) is in the medium range, and the TS is also in the medium range, but a little lower than the HR because of the false alarms.

The ETS is much lower than the TS in this case because of the number expected correct by chance. When the event happens fairly often (43/131 = 0.33 or 33% of the cases), then the number correct by random guessing would be large enough to matter, so the ETS is lower than the TS. When the event becomes rare, the difference between the TS and ETS would be smaller, and both would normally be lower because of the difficulty of forecasting rare events. The total forecasts of occurrences and non-occurrences that would be correct by guessing is 61%. This compares with a total fraction correct (both categories) of $(a + d)/T$ = 104/131 = 0.79 = 79%. Thus, the Heidke skill score (0.47) shows improvement over pure guessing. In this publication, the fraction correct $(a + d)/T$ is not emphasized; it is a less useful score for rare events, because it becomes dominated by the number of correct negatives (d), which may be very large and which obscures the accuracy of forecasts of the event. The hit rate is the preferred accuracy measure.

For the Botswana forecasts, the Hanssen–Kuipers score is also reasonably large, indicating an important positive difference between hit rate and false alarm rate and, correspondingly, a good discrimination between severe weather days and non-severe weather days.

The extreme dependency family of scores, EDS, SEDS, EDI and SEDI, have been included, because they are new scores designed specifically for extreme (or rare) events. As with any new score, it will take time and experience to get a feeling for the meaning of these score values. In this case, the EDS has a low positive value. This is likely due to the fact that a rare event is being predicted. More recent research on this score has shown that it is in fact still dependent on the rarity of the event, and therefore may not be any more useful than other established scores such as the ETS. More significantly, it is also known that the EDS can be improved by forecasting more false alarms, which is definitely an undesirable property of that score. For these main reasons, its use is no longer recommended.

The SEDS, EDI and SEDI all show modest positive values in Table 2. Since these scores are new, further experience will be needed to fully assess the meaning of specific values of these scores. It can be seen that the EDI and SEDI are larger than the SEDS, which is consistent with other experience. All three of these scores improve slightly with the weighting, contrary to other scores. For the EDI and SEDI, this is because the drop in the false alarm rate from .056 to .041, while not as large numerically as the drop in the hit rate from .49 to .46, is nevertheless a more important decrease proportionally than the decrease in the HR.

Finally, given that the weighting of the data to one event per day had a relatively small impact, it is probably not worth the effort to do this unless there are many duplicated event-days.

4. CONTINGENCY TABLE VERIFICATION OF SPATIALLY-DEFINED FORECASTS – THE RSMC SEVERE WEATHER CHARTS

It has been agreed by all concerned that verification of the RSMC daily severe weather forecasting guidance forecasts is a good idea. The question arises: How should this be done? There are several new techniques available that are specifically designed for spatially-defined forecasts. The method described here is consistent with the contingency table method discussed above, and should give results that can be compared with the contingency table results calculated at the NMHSs.

Consider a spatial definition of the four quantities in the contingency table – hits, false alarms, misses and correct negatives – as shown in Figure 5. Shown in Figure 5 are false alarms (areas where severe weather is forecast but is not observed), hits (areas where severe weather is both observed and forecast) and misses (areas where severe weather is observed but not forecast).

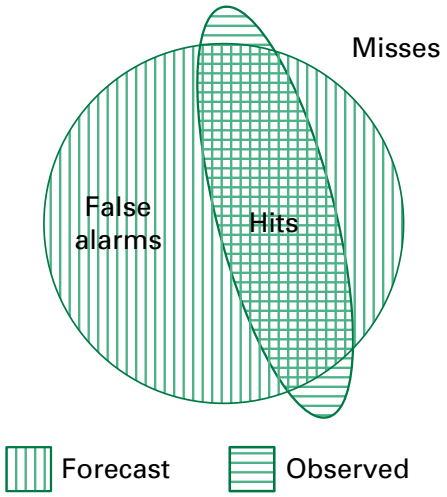

Figure 5. Schematic of contingency table verification for spatial forecasts

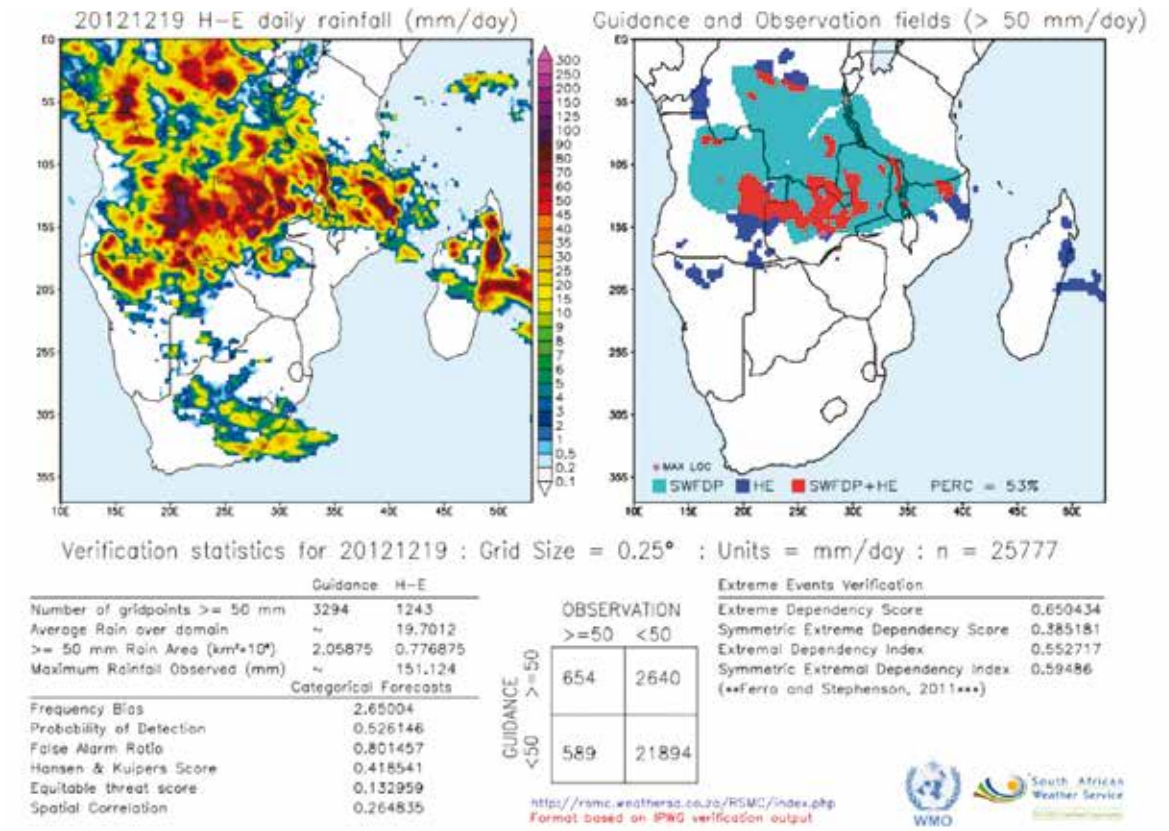

Figure 6. Example of verification of the severe weather forecast charts from RSMC Pretoria: (left) hydro-estimator data used as the observation and (right) the forecast in turquoise, the hits in red and missed events in blue

Definition of the correct negatives is more difficult. In general, this would be the whole area that is not covered by any of the other three, which would usually be most of the domain of the forecast.

It is clear that computation of these areas requires at least quasi-continuous observations and forecasts. The RSMC forecasts are shown as continuous areas, but standard observations are far from sufficient to verify spatially continuous forecasts. However, the data from the European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT) satellite-based hydro-estimator program (using the Met Office (UK) numerical weather prediction model) are quasi-continuous and would be of interest to use for this verification. A word of caution is needed here. The hydro-estimator data use a model to assist with the satellite estimates of precipitation (the Met Office model). Therefore, if these data are used to verify the Met Office model precipitation forecasts, the results will be artificially inflated; the forecasts will appear better than they should, because there is a statistical dependence between the model and the observations used to verify that model. It should also be noted that the observations themselves are remotely-sensed, and will have lower spatial resolution than surface-station observations.

The problem of verification of the RSMC severe weather charts against hydro-estimator data has recently been tackled for the southern African SWFDP by the South African Weather Service. More information will be available on this initiative as research papers are published, but Figure 6 shows an example of the output of this project. The key to the work is the development of a method to decode jpg images to isolate the regions where the severe weather event (> 25 mm and > 50 mm precipitation in 24 hours) has been observed according to the hydro-estimator, and to apply the same method to decode the forecast images. The advantage of this method is that it is the pictures themselves, as shown in Figure 7, which are automatically decoded. The forecast image and observation (hydro-estimator) image are then compared using grid boxes of 0.25 degrees to determine the hits, misses, false alarms and correct negatives as shown in the contingency table at the bottom centre of Figure 6.

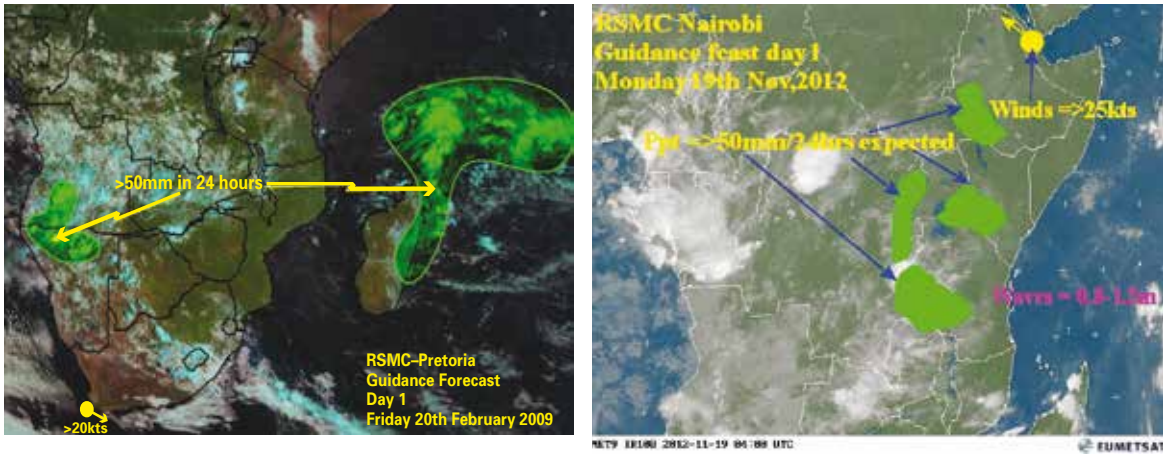

Figure 7. Examples of guidance forecasts from RSMCs Pretoria (left) and Nairobi (right)

The decoding tool will be extremely valuable in verification applications of many types. In addition to the basic contingency table scores, the data can be used for some of the newer diagnostic tools that are available for spatial forecasts, which will help determine whether there are systematic spatial errors in the model forecasts compared with the hydro-estimator data.

5. A FEW WORDS ABOUT VERIFICATION OF ENSEMBLE PROBABILITY FORECASTS

All the verification discussion up to this point relates to the deterministic, categorical forecasts of severe weather events that are produced by the RSMCs and the NMHSs. The set of forecast guidance products available from the global centres and from the RSMCs also includes forecasts of a probabilistic nature, mainly from ensemble forecasts, but also including the risk tables produced at the RSMC. Ensemble forecasts have generally not been saved for the purpose of verification, which is unfortunate because it is clear that some of the products are very popular as forecast guidance, for example, the metgrams that are prepared for a predetermined set of locations for each NMHS.

To date, very little verification of the ensemble forecasts available to the SWFDPs has been done. This is partly because the data management problems are larger; it is not really feasible to set up Excel spreadsheets to tackle this verification. Rather, it will be necessary to use a more powerful freeware package such as "R" to do the verification. Fortunately, ensemble forecasts for ECMWF, the Met Office (UK) and NCEP are available for the same period as was used for the verification of global model deterministic forecasts (see section 6 below). These data have been only briefly explored so far, as part of a verification study assignment at the Fifth International Verification Methods Workshop, held in Melbourne, Australia, in December 2011. A few results from this experiment are briefly described in this section, along with a summary of verification methods for probability forecasts.

Verification of probability forecasts requires different methods than verification of categorical forecasts. The observation, which is nearly always categorical – the event occurs or it does not – must be compared to a probability forecast of the category. With a categorical observation and a probability forecast, meaningful verification results cannot be obtained from a single forecast; it is necessary to collect a large sample of forecasts and corresponding observations to obtain useful verification results for probability forecasts.

Accuracy is commonly measured by scores such as the Brier score (PS) and the rank probability score. To measure skill, the Brier skill score (BSS) is often used. This score is in the form:

$$BSS = \frac{PS_{average} - PS_{forecast}}{PS_{average}}$$

Here, $PS_{average}$ is usually the average Brier score for the verification sample ("climatology", or the average frequency of occurrence of the event), and $PS_{forecast}$ is the Brier score obtained for the forecasts. The interpretation of this score is exactly the same as for the Heidke skill score presented above; it is the per cent of improvement of the forecast compared with the standard unskilled forecast (climatology in this case). The BSS is negative if $PS_{average}$ is < $PS_{forecast}$, meaning that climatology is better than the forecast, since the Brier score is negatively oriented – smaller is better. A perfect BSS is 1, which occurs when $PS_{forecast}$ = 0, a perfect Brier score.

An important and frequently assessed forecast attribute of probability forecasts is reliability. Reliability is the degree to which the forecast probability matches the actual frequency of occurrence of the event. If, on all occasions when a 30% probability of the event is forecast the event occurs 30% of the time, then the forecast is reliable.

Reliability is a little like the average error. If the event occurs on average on 40% of the occasions when 30% probability is forecast, then this is an underforecast (the probability forecast is too low). Conversely, if only 20% of the forecasts of 30% are associated with occurrences of the event, then this is an overforecast (the probability forecast is too high). Reliability cannot be measured on a single forecast, because the observation is categorical. Since numerous forecasts of each forecast probability (30%, 40%, 50%, etc.) must be collected to obtain a good estimate of the observed frequency of occurrence of the event, quite a large sample of forecasts and observations is needed to calculate the reliability.

Another property of probability forecasts that is often measured is discrimination, as described in section 2.4.1. The tool used is the ROC, which is obtained in the same way as described in section 2.4.2. When the forecasts are expressed as probabilities, thresholds are established usually for each decile probability value, 0.1, 0.2...0.9, contingency tables are set up based on these thresholds, and the HR and FA are computed for all these tables and plotted on a graph, similar to that shown in Figure 4. The points usually suggest a curve which should lie above the diagonal.

Figure 8 shows an example of the Brier skill score results obtained for the ECMWF and Met Office (UK) Global and Regional Ensemble Prediction System (MOGREPS) ensemble forecasts for all the GTS stations in eastern Africa that were available at ECMWF during the rainy season from September 2010 to May 2011. The data on which Figure 8 is based are described in more detail in section 6 below.

The results in Figure 8 indicate that the ECMWF forecasts show approximately 0 skill for all thresholds, for all forecast ranges to six days. Skill is slightly positive for the shortest range ECMWF forecasts. The MOGREPS forecasts show negative skill, which is strongest for the lowest thresholds and decreases towards 0 for the highest thresholds. This means that a climatological precipitation forecast would be more accurate than the MOGREPS forecast. More study would be needed to determine the reasons for the difference, but one possibility is that the MOGREPS forecasts show more spatial and temporal variations, which does not match the observations.

Figure 9 shows the ROC curve for the 24-h ECMWF precipitation forecasts for all available eastern African stations. Points for probability values 0.1, 0.2...0.9 are plotted. Each of these thresholds is associated with a hit rate and false alarm rate, which are also plotted.

As in Figure 4, all the points lie above the 45-degree line, indicating that the probability forecasts can be used to discriminate between situations which support the occurrence of severe weather and those which do not. Error bars are shown on Figure 4, indicating that the points are well above the diagonal with a high degree of confidence. It should be emphasized here that the ROC is a completely independent measure from the Brier skill score and reliability diagrams; there is no particular reason to expect that these different measures will provide the same result. The ROC involves setting a threshold, and stating: "Yes, we will assume severe precipitation will occur if the

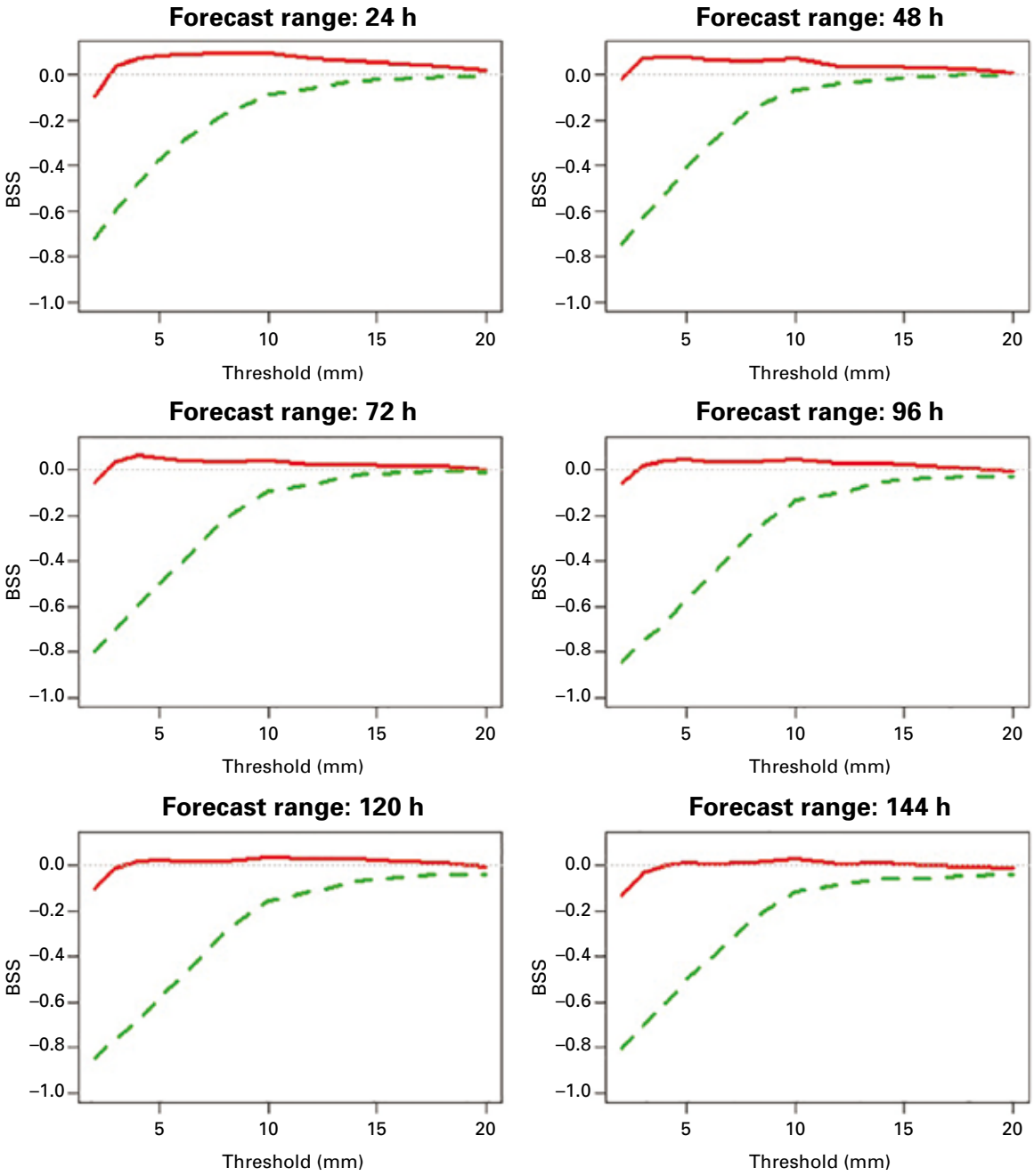

Figure 8. Brier skill scores (BSS) obtained for ECMWF (red, solid) and MOGREPS (green, dashed) probability of precipitation forecasts as a function of threshold in mm/24h

forecast probability is more than 0.4." The user would then take action based on that decision (issue a warning, alert the authorities, etc.). The ROC computation thus implies conversion of a probability forecast into a categorical one, and therefore is related to decision-making in response to the forecast. That is why it is considered a measure that helps assess the utility of the forecast for decision-making processes.

It should also be emphasized that these first results from the verification of ensemble forecasts were obtained as part of a student project, and therefore might bear more checking and further work before complete confidence can be placed in their validity. At the same time, however, it is impressive that such results could be obtained with less than a week of work using "R". That is an indication that the R package is indeed useful, and should be explored further to assist in verification activity for the SWFDPs.

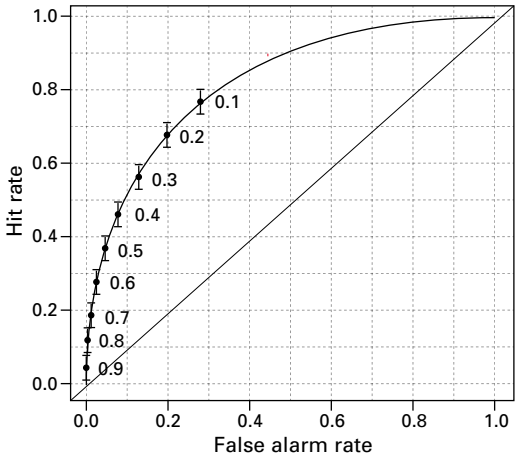

Figure 9. The ROC curve for ECMWF 24-h precipitation forecasts for all available eastern African stations, September 2010 to May 2011

6. EXAMPLE: SOME VERIFICATION RESULTS FOR THE EASTERN AFRICAN SWFDP

The direct model output products made available from ECMWF, NCEP and the Met Office (UK) to the SWFDP had not been verified at all for any country in Africa. In this section, some first results of verification of global models with respect to GTS observations from eastern African countries are shown.

6.1 ECMWF and NCEP global models verified with respect to GTS observations for the 2010–2011 rainy season (September 2010 to May 2011)

6.1.1 *Data*

Both ECMWF and NCEP were kind enough to supply matched observation and forecast data for their respective global models for all stations in the six countries for which they received observations on the GTS between September 2010 and May 2011, that is, one rainy season. These data are in a consistent format, as determined by ECMWF and followed by NCEP. The data format and suggestions for manipulation of the data are described in more detail with the electronic version of this publication; the Excel spreadsheet has been set up to facilitate further verification and exploration of this dataset.

Figure 10 shows the locations of the stations represented in the dataset. Unfortunately, there are no data available from Ethiopia or Burundi, and very little from Rwanda. Data are available mostly from Kenya and the United Republic of Tanzania. Up to a point, it can be assumed that the general conclusions about the accuracy of model forecasts apply throughout the region, but it would definitely be useful to have more detailed datasets available in future, and especially important to have datasets available if statistical post-processing of model output were to be undertaken.

6.1.2 *Scatter plots – looking at the data*

The first step in forecast verification is to look at the data, which is easily accomplished by preparing a forecast-observed scatter plot. Figure 11 shows such a plot for 24-h ECMWF and NCEP forecasts for all stations.

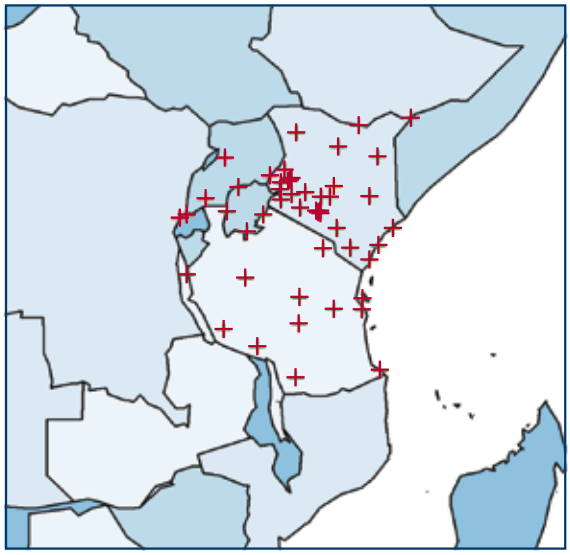

Figure 10. Stations available in the 2010–2011 verification dataset

The first point to note about these two scatter plots is that the forecasts do not match the observations very well. That is to say, the forecasts are not good quality in terms of the forecast precipitation amounts. Convective precipitation is notoriously difficult to predict by a model, because of its small scale, especially in the tropics.

The second point to note is that there are quite a few severe observed events included in the dataset (observed precipitation over 50 mm). All of these, however, are missed events. Neither model forecasts more than 50 mm when it is observed. ECMWF never predicts more than 50 mm at any of the stations at any time, the maximum forecast being about 46 mm. NCEP does attempt to predict more than 50 mm numerous times, but all except perhaps one case are false alarms.

On both scatter plots, the straight line is an attempt to determine a predictive equation for rainfall at the station, given the model forecast rainfall. In both cases, since the quality of the model prediction is poor, neither predictive equation can be used. The lack of quality is evidenced by the R-squared value, which is near 0 for both. This means that the predictive equation explains practically none of the variation in the observations; the error in the forecast will be as large with the equation as without it. However, reading from the straight line provides some idea of the

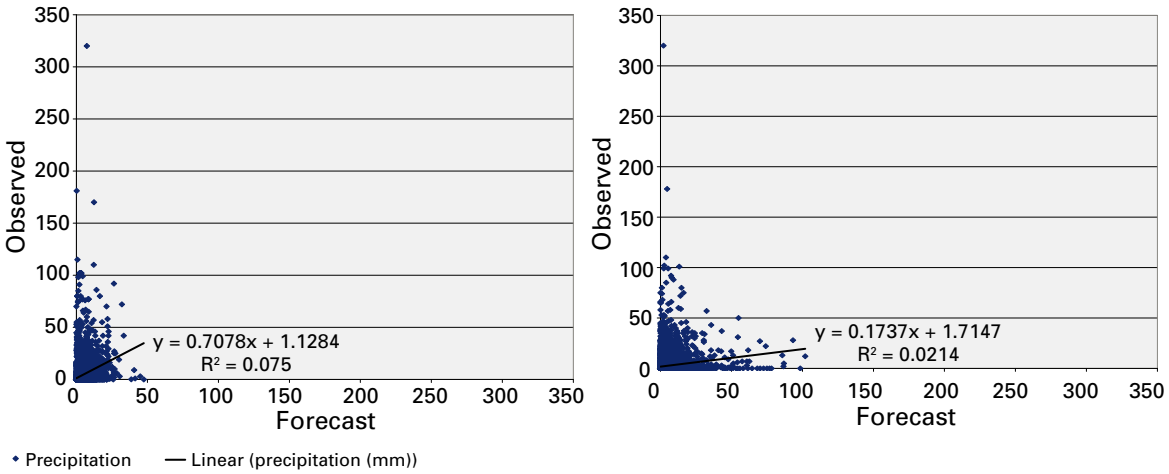

Figure 11. Scatter plot of 24-h ECMWF (left) and NCEP (right) forecast precipitation amounts versus observations for the 2010–2011 rainy season in eastern Africa. The straight line is the best-fit line to the data.

average errors in the forecasts. For example, when 100 mm is predicted by NCEP, the expected observed precipitation is about 20 mm, indicated by the location of the straight line for a forecast value of 100 mm. This suggests that there is a tendency for the NCEP forecasts to overforecast precipitation. For ECMWF, a forecast of 50 mm corresponds to an expected observed value of about 40 mm, according to the straight line, and so ECMWF is overforecasting only slightly on average.

6.1.3 *Contingency table scores*

This section refers to the contingency table scores, as defined above. As there are relatively few cases of extreme rainfall, and because the models appear not to predict the extreme rainfall events well at all, it will be simpler to understand the performance of the models by using lower thresholds and computing the scores for these. In the following, thresholds of 1, 5, 10, 20 and 30 mm have been used.

Figure 12 shows the frequency bias for the 2010–2011 rainy season, for all available stations, for the six thresholds and for both the ECMWF and NCEP models. The frequency bias indicates whether the model is forecasting the event as often as it occurs (value of 1.0), more often than it occurs (>1.0), that is, overforecasting, or less often than it occurs (<1.0), that is, underforecasting.

The first point to note is that both models overforecast significantly when the event is defined as >1.0 mm / 24 h. This is an indication of the known tendency towards overforecasting of drizzle in many models, the tendency to "leak" small amounts of precipitation. For thresholds of 5 mm and higher, there is a difference in the performance of the two models. For 5 mm, ECMWF forecasts are nearly unbiased, while NCEP continues to overpredict. For 10 mm and higher, ECMWF underpredicts the frequency of occurrence; in fact, the underforecasting is extreme for thresholds over 20 mm. On the other hand, NCEP shows a bias of near 1 for the 10-mm threshold but also underforecasts for higher thresholds. The NCEP bias, however, never drops much below 0.50, even for 30 mm. This is consistent with the scatter plot results. The attempts by NCEP to predict higher amounts of precipitation are not accurate. The frequency bias does not consider accuracy.

The second point to note is that the bias is usually higher for both models for day 1 forecasts. For day 2 and beyond, the bias drops for the NCEP forecasts, but stays approximately the same for the ECMWF forecasts.

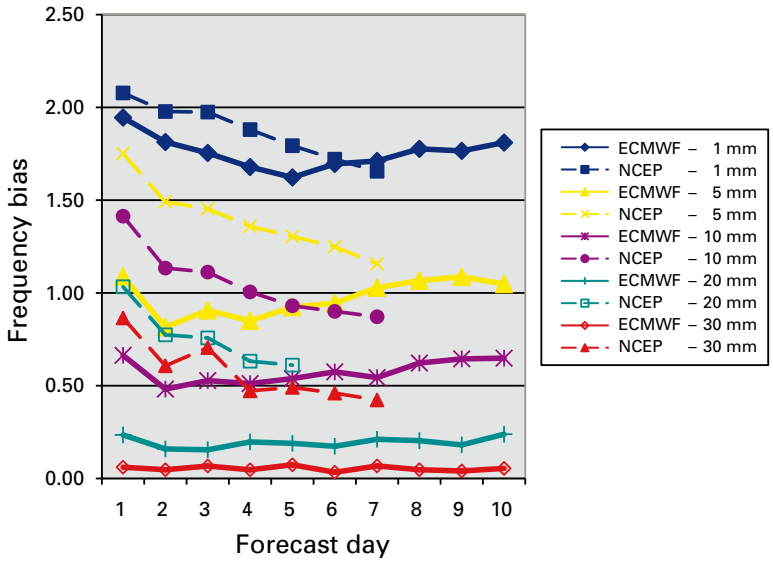

Figure 12. Frequency bias (B) for ECMWF (solid lines) and NCEP (dashed llnes) forecasts of precipitation categories greater than 1-, 5-, 10-, 20- and 30-mm precipitation in 24 hours

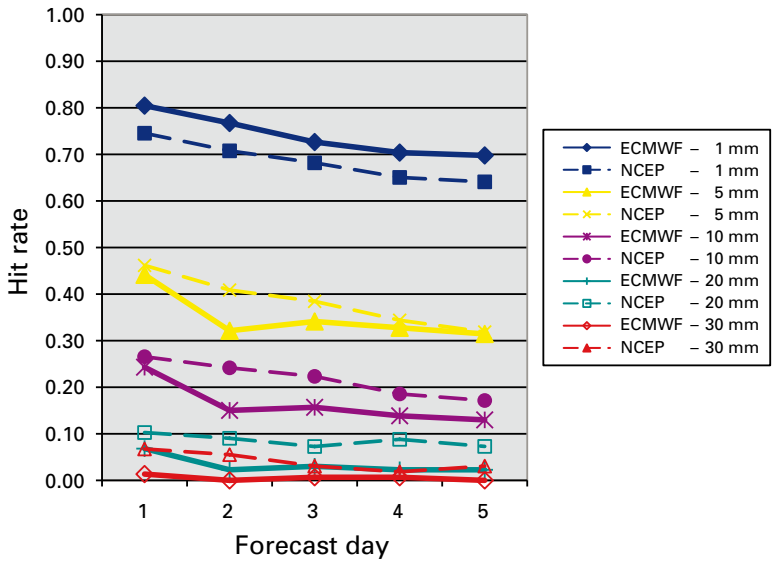

Figure 13. Hit rate (HR) for ECMWF (solid lines) and NCEP (dashed lines) forecasts of greater than 1, 5, 10, 20 and 30 mm in 24 hours for the 2010–2011 rainy season

The hit rate (Figure 13), shows the accuracy of the forecasts, specifically the percentage of the observed events that were correctly forecast. The hit rate is highest for the lowest threshold, and decreases steadily with increasing threshold. It should be noted that, according to the hit rate alone, the NCEP forecasts are slightly more accurate than the ECMWF forecasts for all thresholds except the lowest. The hit rate decreases with increasing forecast projection, as expected. Results are shown up to day 5.

The false alarm ratio (Figure 14) shows the percentage of the forecasts of the event that were false alarms. False alarms are generally undesirable, so this should be as low as possible (0 is best). Like the hit rate, the false alarm ratio should never be used alone; it is important to consider it in connection with the hit rate. In Figure 14, the false alarm ratios are all high for both models, but it can be seen that the NCEP false alarm ratios are higher than the ECMWF ones at all thresholds, and in some cases considerably higher. This result sheds additional light on the quality of the

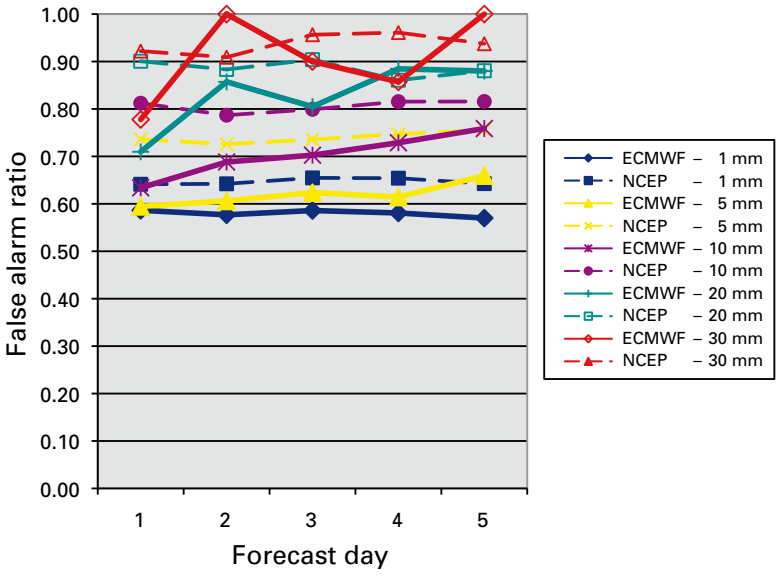

Figure 14. False alarm ratio (FAR) for ECMWF (solid lines) and NCEP (dashed lines) forecasts of greater than 1, 5, 10, 20 and 30 mm in 24 hours for the 2010–2011 rainy season

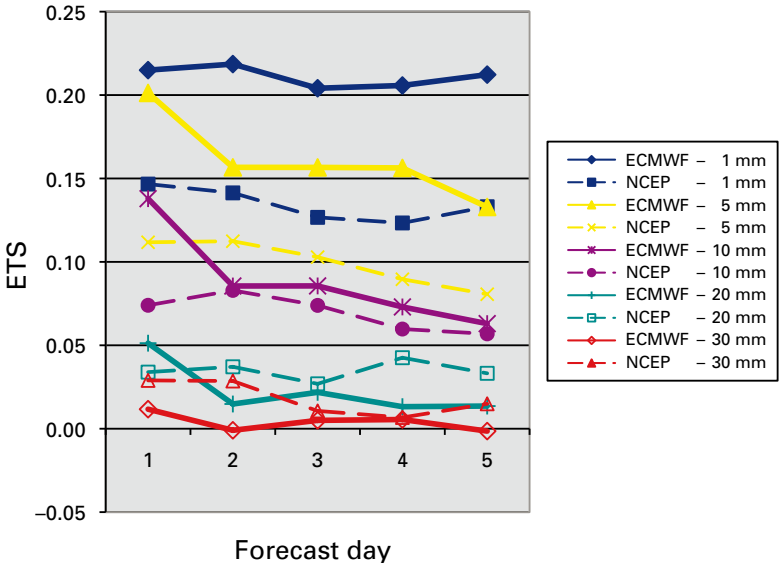

Figure 15. Equitable threat score (ETS) as a function of lead time for ECMWF forecasts (solid lines) and NCEP forecasts (dashed lines) of 24-h precipitation amounts greater than 1, 5, 10, 20 and 30 mm

NCEP forecasts relative to ECMWF. The higher hit rates at NCEP are achieved at a cost of higher false alarms. This means that the NCEP forecasts might be preferred if higher false alarm ratios are acceptable to users; otherwise, the ECMWF forecasts, achieving hit rates almost as high (higher at the lowest threshold) but with much lower false alarm ratios, might be preferable.

The ETS is a commonly used score to measure the accuracy of categorical forecasts. Unlike the HR and the FAR, it takes into account both types of errors, misses and false alarms, and therefore it can be used alone as a general measure of accuracy. This score does tend towards 0 for rarer events, as is clearly shown in Figure 15. This effect is undesirable for verification of rarer events (higher thresholds), because the score becomes insensitive to differences in accuracy of forecasts. This is the main reason new scores such as the EDI were developed. The ETS, however, is a good choice for general use, and the equitability property. It takes into account differences in the underlying climatology of the event so that results from samples with different frequencies of occurrence of the event can be compared.

Figure 15 shows a "truer" story than either the HR or the FAR alone. The ECMWF forecasts are clearly more accurate for the lowest two thresholds (light rain), while there is little difference in the forecast quality for thresholds of 10 mm and above. NCEP seems to have a slight edge overall, but at these low values, that difference is not likely significant. It should be noted that the score values are low for all thresholds for both models, indicating again that the precipitation amount forecasts cannot be used to indicate 24-h total precipitation total at specific locations.

The Pierce skill score (Hanssen–Kuipers score, true skill statistic) measures the accuracy of the forecasts in a different way. This score is an indicator of how well the forecast can distinguish situations leading to the occurrence of the event from those leading to the non-occurrence of the event. As mentioned above, the KSS is the difference between the hit rate and the false alarm rate. Positive values indicate the hit rate is greater than the false alarm rate, indicating that the forecast is able to correctly distinguish occurrences from non-occurrences. The greater the difference, the higher the KSS and the clearer the distinction made by the forecast.

For the model forecasts tested, Figure 16 indicates that ECMWF scores better than NCEP, but only at the lowest threshold. For all thresholds above 1 mm, there is no significant difference between the ECMWF and NCEP model forecasts, and neither does a good job of discriminating occurrences from non-occurrences.

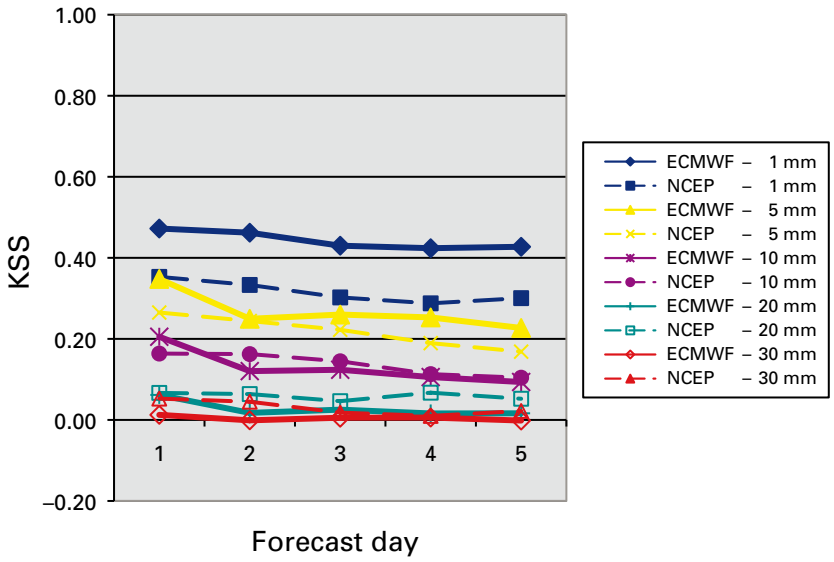

Figure 16. Pierce skill score (KSS, TSS) for ECMWF (solid lines) and NCEP (dashed lines) forecasts of 24-h precipitation amounts greater than 1, 5, 10, 20 and 30 mm, for lead times of 1 to 5 days; all available GTS stations in the eastern African region, September 2010 to May 2011

The stable extreme dependency score was designed to avoid the problem of small values for events that are not common, and limitation which applies to other common scores such as the ETS and hit rate. Studies of this score since it was developed indicate that it does not fully achieve this goal, but nevertheless the score is expected to be useful for forecaster-oriented evaluation of forecasts. Since it uses the forecaster frequency (the total number of forecasts of the event divided by the total sample size), it is sensitive to changes in the forecaster's forecast strategy. If the event is forecast too often, then the score may be lower because of excessive false alarms.

In Figure 17, best (highest) values are obtained for both lowest thresholds (1 and 5 mm), and it can be seen that ECMWF scores better than NCEP, presumably because of the NCEP tendency to overforecast. For the 20-mm threshold, the SEDS values are about the same for both models,

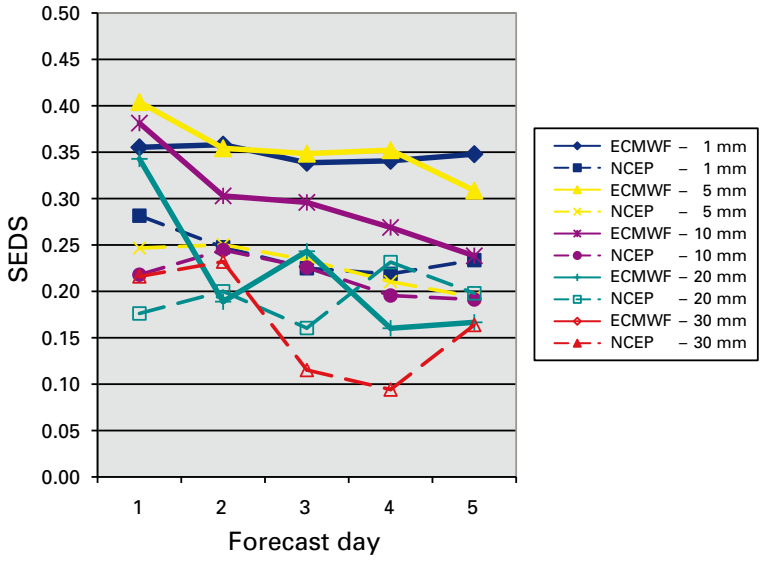

Figure 17. Stable extreme dependency score (SEDS) for forecasts of 24-h precipitation accumulation for African stations, from ECMWF (solid lines) and NCEP (dashed lines); thresholds of 1, 5, 10, 20 and 30 mm

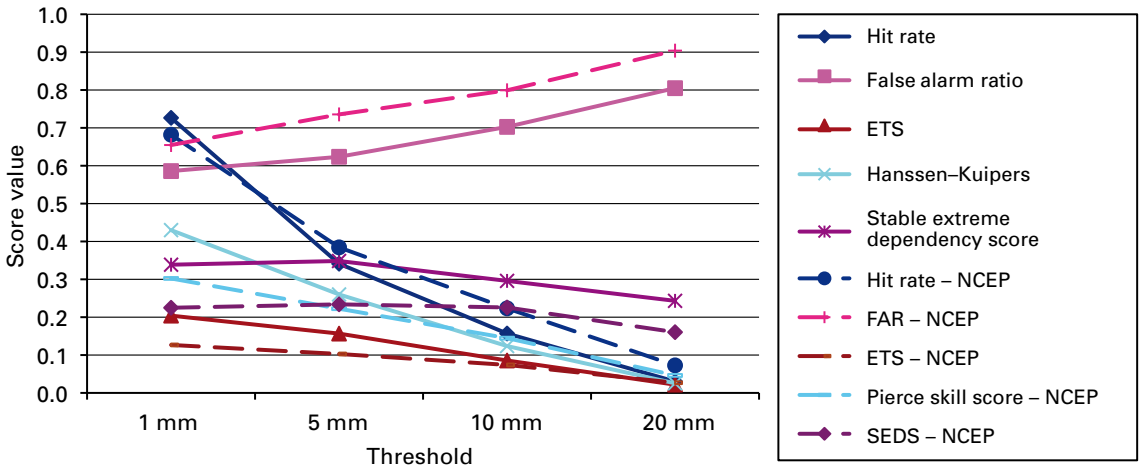

Figure 18. Scores for eastern Africa 24-h precipitation forecasts as a function of threshold, for day 3 ECMWF results (solid lines) and NCEP results (dashed lines)

while for the 30-mm threshold, only NCEP scores could be computed, because ECMWF did not score any hits for this threshold, and did not forecast it often. The score for NCEP is quite low, indicating that, although the event was forecast, the accuracy is too low to be of use in practice. Finally, it should be noted that there is a decrease with increasing threshold, partly because of the low frequency effect mentioned above, but also because of decreasing accuracy.

The same results are shown in a different way in Figure 18. Here the scores are plotted for a single forecast projection time (3 days) for different thresholds. It is clear that most of the scores decrease for increasing threshold, which illustrates the effect of lower frequencies of occurrence of the event (the base rate) on the score values. Coupled with this is a genuine degradation of forecast accuracy for the higher thresholds. The tendency of the FAR to increase towards 1.0 for lower base rates is also evident, as is the fact that the NCEP forecasts overforecast more than the ECMWF forecasts at all thresholds, incurring higher false alarms. The SEDS is less sensitive to the base rate than the other scores, which is consistent with the design of that score, and indicates that it would indeed be useful for higher thresholds (rare events).

6.2 Verification of 6-hour precipitation forecasts by the Met Office (UK) global model

It had been originally intended that all three global centres involved in the project would supply matched forecasts and observations for 24-h precipitation accumulation for all the eastern African stations for which they had GTS data. However, for technical reasons, the Met Office (UK) was unable to compute 24-h precipitation amounts from the SYNOP observations and so it supplied forecasts and corresponding observations of 6-h amounts for a shorter period, from September 2010 to March 2011. Forecasts are for the range 0 to 48 hours; some sample results are shown below for 12-, 24-, 36- and 48-h forecast projections, that is, 6-h accumulations 6–12 h ahead, 18–24 h ahead, 30–36 h ahead and 42–48 h ahead. Results are also for thresholds of 1, 5, 10 and 20 mm. Higher thresholds were rarely exceeded in the model predictions.

6.2.1 *Frequency bias*

Frequency bias is shown in Figure 19 for the whole six-month period. As with the ECMWF and NCEP forecasts, there is a tendency to forecast precipitation too often when smaller amounts are included, and to underforecast the higher amounts. This tendency is more extreme for these 6-h forecasts, with biases higher than 3. This means the precipitation of more than 1 mm in 6 hours is forecast more than three times as often as it occurs.

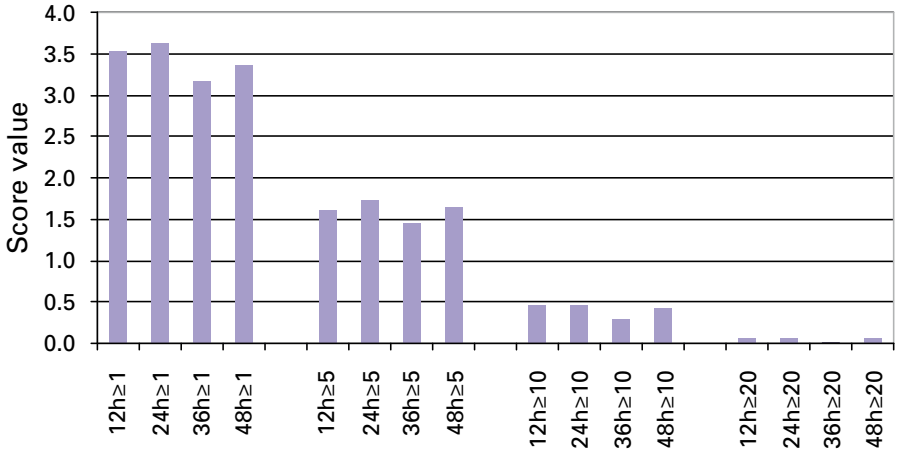

Figure 19. Frequency bias for Met Office (UK) forecasts of 6-h total precipitation, each group of four bars representing the results for a particular precipitation threshold, 1, 5, 10 and 20 mm, from left to right

On the other hand, for the 10- and 20-mm thresholds, the event is seriously underforecast, with bias of less than 0.5 even at 10 mm.

6.2.2 *Hit rate and false alarm ratio*

The frequency bias does not indicate forecast accuracy; it only shows whether the event is forecast as often as it is observed. The hit rate and false alarm ratio are shown together in Figure 20.

As mentioned above, these two scores should be used together, since either can be systematically improved by adopting a less-than-ideal forecasting strategy. The HR indicates that only the forecasts for the 1-mm threshold show some accuracy; hit rates are below 0.2 for the other thresholds. It should also be noted that the higher hit rates at the 1-mm threshold are achieved only with very high overforecasting biases (Figure 19). The false alarm ratios are high (undesirable) for all thresholds, generally higher than 0.8. That means that over 80% of the forecasts of the event were false alarms.

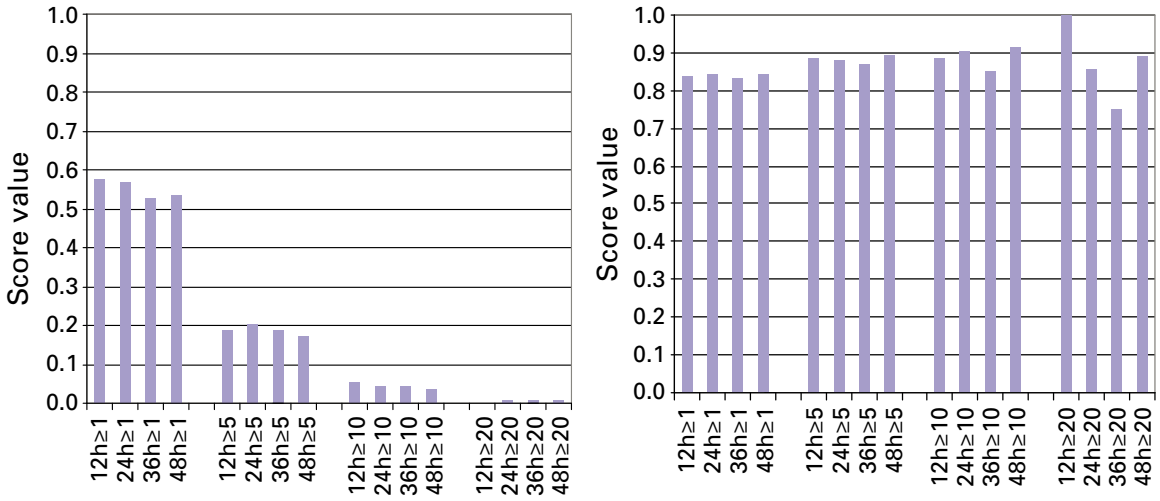

Figure 20. Hit rate (left) and false alarm ratio (right) for the Met Office (UK) forecasts of 6-h total precipitation total, for each 12 hours out to 48 hours

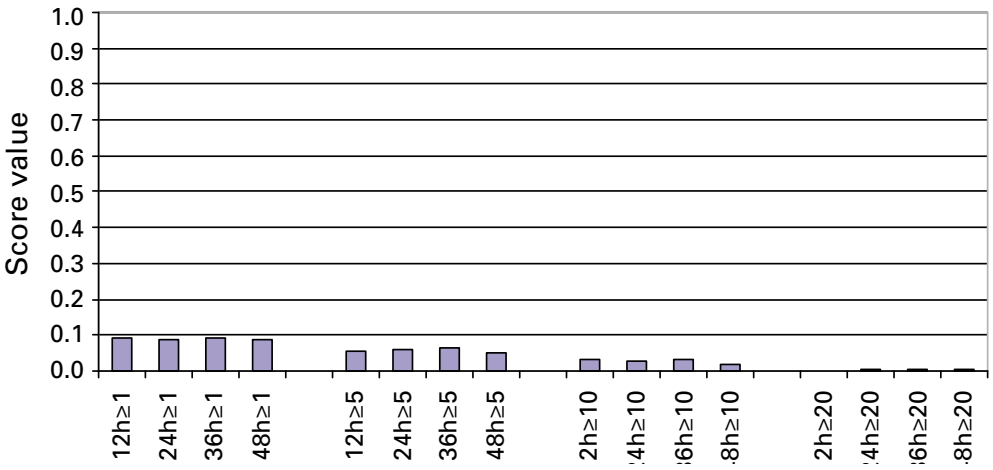

Figure 21. Equitable threat score for the 2010–2011 MET Office (UK) forecasts of 6-h precipitation amounts valid for 12-, 24-, 36- and 48-h forecast projections

6.2.3 *Equitable threat score*

The ETS is shown in Figure 21 partly because it is a commonly used score for summary verification, and therefore provides a better idea of the numerical values of the score vis-à-vis accuracy in practice. For these forecasts, the ETS values are very low, less than 0.1 for all thresholds and projections. This is expected, especially when the verification is for 6-h accumulations at specific locations (points); models are generally unable to predict the location of convection-dominated precipitation events, and normally underestimate the peak intensity of such events. This is at least partly due to the relatively low spatial resolution of global models; higher resolution regional models might do a better job of predicting the intensity of events, but spatial precision still presents difficulties. Categorical precipitation forecasts then must be taken with "several grains of salt", that is, to be used carefully. The patterns seen on forecast charts give only a general idea of where precipitation might occur; the amounts predicted cannot be expected to be accurate.

6.2.4 *Pierce skill score (Hanssen–Kuipers score; true skill statistic)*

This score is included because it measures a different attribute of the forecasts: discrimination, the ability of the forecast to differentiate situations preceding the event from those preceding the non-event, as mentioned above. Figure 22 shows the KSS for the Met Office (UK) forecasts.

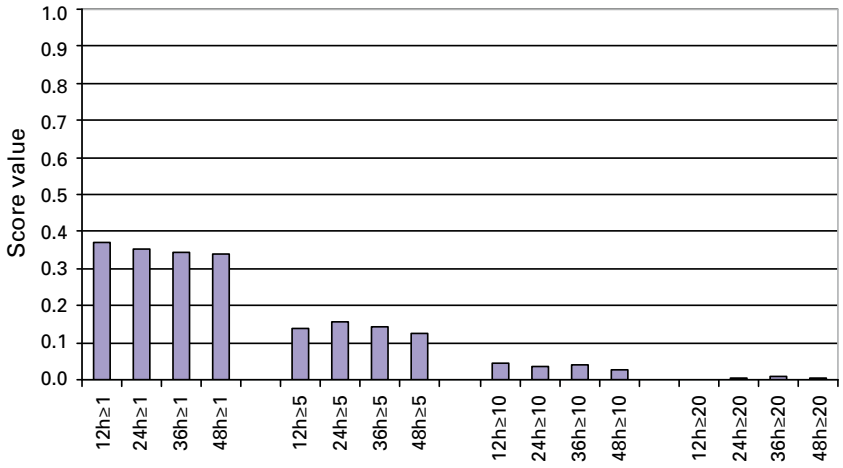

Figure 22. Pierce skill score for the 2010–2011 Met Office (UK) forecasts of 6-h precipitation accumulation valid 12, 24, 36 and 48 hours in advance

Figure 22 shows that there is some discriminating ability in the forecasts for the lowest threshold, but not for any of the other thresholds. The forecasts have little accuracy in prediction at any threshold, so it would seem likely that the discrimination comes from the ability of the forecasts to separate those days when convection does not form from those when it does, and to separate locations in the verification domain where convection occurs less often from those where it is more frequent.

7. CONCLUSION

This publication describes the verification activities and methodology for the southern and eastern African SWFDPs. Since the forecast programme focuses on warnings of severe weather events defined in terms of threshold exceedance, the most appropriate set of verification tools is considered to be that associated with contingency tables. These are described and examples of their application are shown using African station data. In a recent initiative, graphical severe weather products issued by the RSMCs are being verified using a spatial analogue to the contingency table.

This publication is accompanied by the datasets used in the verification of the global forecast products, and three different Excel spreadsheets set up for the computation of contingency tables and their scores, available with the electronic version of this publication.

Much remains to be done. Future efforts should concentrate on the following aspects:

(a) Continue to encourage the NMHSs to save all their severe weather data, both forecasts and observations, using the SWFDP event table, and to apply the verification tools described in this publication, along with the contingency table calculator spreadsheets.

(b) Emphasize efforts to use all datasets that are available inside African countries to improve the verification:

  (i) Obtain another dataset of matched forecasts and observations from the three global centres, in the same format as was used for the verification shown in Figures 11 to 18, but covering both southern Africa and eastern Africa, for one full year. Global centres should be asked to provide the forecast data only as nearest gridpoint values for station locations as specified by the NMHSs, and the NMHSs and/or RSMCs would be expected to add the verifying observations to the datasets. Then, the data can be entered into the spreadsheets available with the electronic version of this publication for computation of verification scores;

  (ii) Exploit the use of technology to collect non-standard data for use in verification (for example, mobile phones) in collaboration with forecast dissemination efforts;

  (iii) Ground-truth satellite-based systems such as the hydro-estimator in order to gain confidence in the quality of remotely sensed precipitation estimates. For example, Lake Victoria, which is important to the project, is a complete data void at present.

(c) So far verification has been carried out for some warnings in some countries, for some RSMC forecasts and some of the global model forecasts. Nothing has been done yet for the local area models, and many other products issued from the regional centres. Some promising work is in progress on verifying graphical products from the RSMCs, but very little work has been done on verifying ensemble forecasts. Work should continue towards verification of all products from the SWFDPs including the local area models and ensemble forecasts. In the case of the RSMC graphical products, the newer spatial verification methods should be explored, and, once the work is published, it should be extended to the eastern African SWFDP.

8. WEB RESOURCES FOR FURTHER INFORMATION

The European Virtual Organization for Meteorological Training (EUMETCAL) training site on verification – computer-aided learning:

http://www.eumetcal.org/resources/ukmeteocal/temp/msgcal/www/english/courses/msgcrs/crsindex.htm

The website of the Joint WWRP/WGNE Working Group on Forecast Verification Research:

http://www.cawcr.gov.au/projects/verification/

This contains definitions of all the basic scores and links to other sites for further information.

For more information, please contact:

World Meteorological Organization

7 bis, avenue de la Paix – P.O. Box 2300 – CH 1211 Geneva 2 – Switzerland

Communications and Public Affairs Office

Tel.: +41 (0) 22 730 83 14/15 – Fax: +41 (0) 22 730 80 27

E-mail: cpa@wmo.int

www.wmo.int

JN 14402