A short-range weather prediction system for South Africa based on a multi-model approach

Stephanie Landman[1,2*], Francois A. Engelbrecht[3,4], Christien J. Engelbrecht[5], Liesl L. Dyson[2], Willem A. Landman[2,3]

[1]*South African Weather Service, Pretoria, 0001, South Africa,*
[2]*Department of Geography. Geoinformatics and Meteorology, University of Pretoria, 0002*
[3]*CSIR Natural Resources and the Environment: Climate Studies, Modelling and Environmental Health, Pretoria 0001, South Africa*
[4]*Climatology Research Group, GAES, University of the Witwatersrand, South Africa*
[5]*Agricultural Research Council – Institute for Soil, Climate and Water, Pretoria, 0001, South Africa*

ABSTRACT

Predicting the location and timing of rainfall events has important social and economic impacts. It is also important to have the ability to predict the amount of rainfall accurately. At many operational centres, such as the South African Weather Service, forecasters use deterministic model output data as guidance to produce subjective probabilistic rainfall forecasts. The aim of this research is to determine the skill of a new objective multi-model, multi-institute probabilistic ensemble forecast system for South Africa. This was achieved by obtaining and combining the rainfall forecasts of two high-resolution regional atmospheric models operational in South Africa. The first model is the Unified Model (UM), which is operational at the South African Weather Service. The UM contributed three ensemble members which differ in physics, data assimilation techniques and horisontal resolution. The second model is the conformal-cubic atmospheric model (CCAM) which is operational at the Council for Scientific and Industrial Research, which in turn contributed two members to the ensemble system differing in horisontal resolution. A single-model ensemble forecast, with each of the ensemble members having equal weights, was constructed for the UM and CCAM models, respectively. The UM and CCAM single-model ensemble predictions have been used in turn to construct a multi-model ensemble prediction system, using simple un-weighted averaging. The probabilistic forecasts produced by single-model system as well as the multi-model system are here tested against observed rainfall data over three austral summer half-years from 2006/07 to 2008/09, by using verification metrics such as the Brier skill score, relative operating characteristics, and the reliability diagram. The forecast system is found to be skillful. Moreover, the system outscores the forecast skill of the individual models.

Keywords: short-range, ensemble, forecasting, precipitation, multi-model, verification

* Tel: +27 12 367 6054
Fax: +27 12 367 6189
e-mail: stephanie.landman@weathersa.co.za

1. Introduction

Precipitation is of high relevance to users of meteorological data in South Africa, but it is also highly variable in time and space, making it one of the most difficult meteorological variables to predict skillfully. Nonetheless, skillful precipitation forecasts are essential especially in order to provide early warning of heavy rainfall and floods that may lead to loss of life and property. Most operational weather centers rely on limited-area numerical weather prediction (NWP) models in order to generate reliable and accurate weather forecasts (Stensrud *et al.*, 1999; Toth *et al.*, 2001). At short-range time scales (from 12-hours up to 3 days ahead) predicting the location of a precipitation event in general has a greater error than the prediction of the pattern and amount of precipitation (Theis *et al*, 2005). The large spatial and temporal variability in rainfall together with some internal NWP model restrictions contribute to the uncertainties and low skill associated with rainfall predictions (Ebert, 2001; Theis *et al*, 2005; Roy Bhowmik and Durai, 2010).

Precipitation forecast from NWP models are often provided in a deterministic manner. An inherent characteristic of deterministic forecasts is that the future state of the atmosphere is conditional on the present state of the atmosphere, and its evolution is governed by deterministic equations. Accurate short-range numerical forecasting is therefore dependent on accurately describing the initial conditions (Kalnay, 2003). The reason for this dependency on accurate initial conditions stems from the chaotic and non-periodic characteristics of the atmosphere (Lorenz, 1963). Forecasts that are initialized with only slightly different initial states diverge increasingly as a function of model integration time. Deterministic or best-guess forecasts are therefore considered to become less reliable as the model integration-time increases, due to the uncertainties that exist in the initial conditions as well as the internal error (physics and dynamics) of the numerical models (Lorenz, 1963; Ebert, 2001; Stensrud *et al.*, 2005; Theis *et al*, 2005).

Most National Meteorological Services (NMS) issue precipitation forecasts in terms of subjective probabilities, whereby the user can get additional information regarding the uncertainty pertaining to the specific forecast (Staël von Holstein, 1971). Forecasters have long been aware of the fact that different models can produce a variety in the outcome of the predicted weather (Ebert, 2001). These probability forecasts issued by forecasters are subjective because they are based on the forecaster's own insights and experience (Staël von Holstein, 1971). The issuing of objective probability forecasts at the short-range time-scale has the potential to objectively addresses, to some extent, the uncertainties associated with describing the initial state of the atmosphere as well as the uncertainties induced by internal model errors (Theis *et al*, 2005).

Ensemble prediction systems (EPS) represent a stochastic approach which couples probability with determinism (Lewis, 2005), and which has the specific aim of predicting the probability of future weather events to occur; in turn addressing the uncertainty of a deterministic forecast (Stensrud *et al.*, 1999). Theis *et al* (2005) concluded that precipitation forecasts should be addressed in a probabilistic

manner in order to account for the chaotic nature of NWP forecasts. An important goal of ensemble prediction is to provide estimations of the reliability of the forecast being made (Kalnay, 2003; Grimit and Mass, 2005). The ensemble of forecast outputs from single or multiple numerical weather prediction models provides detail of the forecast regarding the confidence, possible errors and probability outcomes (Bakhshaii and Stull, 2009).

Ensemble forecasts may be constructed in various ways (e.g. Kalnay, 2003). The traditional approach is to perform multiple model runs using the same model, by initializing each run from differently constructed initial conditions. Single-model ensemble systems effectively inhibit the description of the forecast uncertainty associated with model error, and this may lead to underestimation of the forecast error (Clark *et al*, 2008). However, multi-model ensemble forecasting succeeds in addressing the uncertainties that exists in the systematic errors of each numerical model, as well as the uncertainties within the initial conditions (Ebert, 2001). Clark *et al* (2008) noted that in addition to addressing uncertainties in initial conditions, ensemble forecasting, more specifically multi-model systems, will also inadvertently address errors related to lateral boundary conditions. At the short-range time-scale, synoptic and mesoscale features are less predictable due to their more chaotic features than the features of the planetary scale (Hamill and Colucci, 1997; Friederichs and Hense, 2008; Roy Bhowmik and Durai, 2010). For this reason, ensemble methodologies will improve primarily in describing the uncertainty and model error that exists in relation to these shorter length scales. The model uncertainty can be accounted for by running the same model with different physical parameterizations or analysis times or by using model runs from different numerical models (Bowler *et al*, 2008b, Wandishin *et al.,* 2001). Typically, the errors and uncertainties in each individual member of the ensemble cancel when calculating the ensemble average, making the ensemble average appearing relatively smooth (Bowler *et al.*, 2008b; Kalnay, 2003). The combination of a multi-model system from a multi-institute ensemble has the advantage of effectively sharing the computational power needed to construct large ensembles amongst different institutions.

Even though research has shown that the ensemble mean forecast outperforms the single deterministic forecast (Ebert, 2001), operational use of short-range ensemble systems has lagged behind that of long-range and medium-range forecasting (Eckel and Mass, 2005). However, there is a number of NMS's that uses short-range ensemble prediction systems operationally or quasi-operationally. These include NCEP (USA), INM (Spain), NMI (Norway), the Met Office (UK), DWD (Germany), BoM (Australia) and recently also SAWS (South Africa).

The objectives of this paper are to investigate the skill of a multi-model ensemble in South Africa in predicting 24-hour probabilistic precipitation for South Africa, and to compare the skill of the multi-model ensemble to that of the single-model ensemble systems.

The outline of this paper is as follows. In the next section, the observed data sets and forecasting systems applied in the paper are discussed. The construction of a new multi-model ensemble system

based on two operational NWP models in South Africa, and forecast verification methods applied to describe the accuracy and skill of the system, are described in section 3. Verification results are presented in section 4, and conclusions are drawn in section 5.

2. Data

2.1. Rainfall data

Three austral summer half-years were selected for the purpose of performing hindcasts. These are the months of October through to March for the years 2006/07, 2007/08 and 2008/09. South Africa is primarily a summer rainfall region, with only the southwest being a winter rainfall region (Tyson, 1986). For these three seasons, 24-hour rainfall totals were calculated from numerical weather prediction data, as well as from observations stemming from automatic and manual weather stations of SAWS and the Agricultural Research Council (ARC). Figure 1 indicates the distribution of the combined observation network of SAWS and the ARC over South Africa, as used in this study. The rainfall totals were accumulated over the 24-hour periods from 06:00 UTC on a given day, to 06:00 UTC the next day, in correspondence to the time of observation made at manual weather stations managed by SAWS. Therefore, the corresponding forecast hours of the NWP data for this accumulation are for forecast hours 6 to 30 (all the NWP forecasts utilized in this study were initialised at 00:00 UTC).

Figure 1:

In order for the numerical precipitation forecasts to be directly compared to the observed rainfall, the rainfall totals recorded at the weather station locations are interpolated to a grid field. Due to the sparse distribution of stations, it is not meaningful to construct a country wide grid field at a resolution finer than about 0.25º. An average rainfall value was calculated for each grid box defined by a grid of this resolution, using a box-average technique (Peel and Wilson, 2008). This procedure has been shown to successfully represent station data on the same grid field as that of the numerical weather prediction output (Peel and Wilson, 2008). Varying numbers of rainfall stations were used in the calculation of the average grid box value, depending on the availability of rainfall stations in the geographical area demarcated by the grid box. If no stations were present within a specific grid box, the grid box was excluded from the subsequent verification calculations. The results of the verification will be sensitive to the number of stations per grid box. For this study, the minimum number of stations required to be present within a grid box was chosen to be one. With the minimum number of at least two stations per grid box, the grid would have fewer samples, particularly in sparsely covered regions (i.e. Northern Cape), and the results would be skewed toward more populous regions of South Africa (i.e. Gauteng and Western Cape). Even with a grid box represented by only one station, which in turn has the characteristics of a point measurement, making comparison with a model grid box average a bit more problematic, the greater number of observational grid boxes was chosen to be the better option for the purpose of this study.

2.2. Model data

2.2.1. Unified Model

The Unified Model (UM) is a nonhydrostatic model developed at the UK Met Office (Davies *et al*, 2005). Its vertical coordinate is based on geometric height. The UM can in principle be applied at time-scales ranging from weather forecasting to climate projection, and at resolutions ranging from relatively low to very high resolutions beyond the validity of the hydrostatic assumption (Davies *et al*, 2005). The UK Met Office runs the UM at global scale with horisontal resolution of 40 km, four times per day providing initial and boundary conditions for a regional version of the UM. Since May 2006, the UM version 6.1 has been running operationally at SAWS with different configurations, including various horisontal resolutions, parameterizations schemes and data assimilation processes (Tennant, 2007). The three configurations used in this study are described in detail. All three of the configurations run in-house at the SAWS on a NEC SX-8 supercomputer.

2.2.1.1 12 km no Data Assimilation (no-DA)

The 12 km no-DA UM forecast covers the sub-continent of southern Africa as well as large areas of the surrounding oceans (Figure 2a). This configuration runs once a day with 38 levels in the vertical, and produces forecasts 48-hours ahead from the initialized field at 00:00 UTC (Tennant, 2007). The forecast output fields are written every hour. This run uses the 18:00 UTC forecast from the UM Global Model to provide initial conditions to the 12 km run at 00:00 UTC, as well as lateral boundary condition fields.

2.2.1.2 12 km DA

This configuration field has the same domain (Figure 2a) and resolution as the 12 km no-DA run, but incorporates continuous three-dimensional variational (3DVAR) DA. DA is a statistical method of combining the latest observational data and the first guess field from the previous short-range forecast for the same period (Kalnay, 2003). The assimilation process is repeated every six hours, forecasting six hours ahead, i.e. four times a day, but at the 00:00 UTC assimilation update, the model continues to forecast 48-hours ahead.

2.2.1.3. 15 km no-DA

The 15 km horisontal resolution run has a much smaller domain (Figure 2b), than the 12 km resolution runs. It is set-up to cover only the South African domain, from 22ºS to 35ºS and 15ºE to 34ºE, making it computationally less expensive. This configuration uses no data assimilation, but also uses the 18:00 UTC forecast from the UM Global Model to provide 00:00 UTC initial conditions.

2.2.2. CCAM

The Conformal-cubic atmospheric model (CCAM) was developed by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) in Australia (McGregor, 19962005a, 2005b; McGregor and Dix, 2001, 2008). CCAM is a variable-resolution global model, that may be applied either in quasi-uniform mode to function as a global circulation model, or alternatively in stretched-grid mode to provide high resolution forecasts over an area of interest. The model solves the hydrostatic primitive

equations using a semi-implicit semi-Lagrangian method (McGregor, 2005). It has been illustrated by Engelbrecht *et al*. (2009) and Engelbrecht *et al*. (2012) that CCAM is capable of satisfactorily simulating many attributes of the present-day climatological conditions over southern and tropical Africa. The model has also been shown to produce skillful short-range and seasonal forecasts over the southern African region (Potgieter, 2006; Ghile and Schulze, 2010; Landman et al., 2009; Landman *et al*., 2010, Engelbrecht et al, 2011). CCAM became operational at the Council for Scientific and Industrial Research (CSIR) in 2010, so that hindcast data were created in order to perform verification studies for the three summer half-years relevant to this study. In operational mode, the CCAM is initialized at 00:00 UTC, using initial condition fields obtained from the Global Forecast System (GFS). Two different 7-day forecasts are produced daily using the 00:00 UTC initial state. A forecast that has a resolution of about 60 km over southern and tropical Africa is performed first (Figure 2c). In order to obtain this forecast the model is applied in stretched-grid mode over southern and tropical Africa, with the resolution decreasing to about 400 km in the far-field. A high-resolution forecast is subsequently performed using a more strongly-stretched grid that provides resolution of 15 km over southern Africa, with this run nudged within the 60 km forecast. Hindcasts for the three half-years under consideration were performed using a set-up that mirrors the operational forecasting system. For both the 60 km and 15 km hindcasts, model output is available at six-hourly time-steps over a domain that covers southern and tropical Africa. All the hindcasts were performed on the Sun Hybrid System of the Centre for High Performance Computing (CHPC) in South Africa.

Figure 1.

3. Methodology

3.2. Construction of the Multi-Model Ensemble Prediction System

An ensemble of forecasts to some extent describes the uncertainties pertained in single-model forecasts (Zongjian, 2008; Kalnay, 2003). The multi-model ensemble system (MMENS) presented here is formulated with the purpose of predicting the probability of precipitation exceeding predetermined thresholds, over a 24-hour period, from 06:00 UTC to 06:00 UTC. Although each of the individual members of the mult-model ensemble described in the following sections covers a bigger domain than South African (22º to 35ºS and 16ºE to 33ºE), the spatial extent of the SAWS and ARC observational network limits the verification analysis to South Africa. The model output was regridded to the same horisontal resolution of 0.25° over the South African domain as applicable to the observational grid.

Different 24-hour rainfall total thresholds are considered in order to formulate dichotomous forecasts for each threshold value. That is, for a given threshold a value of zero is assigned to the forecast if the threshold is not exceeded and a value of one if the threshold is exceeded. In this paper the threshold values of daily rainfall totals considered are 1 mm and 10 mm, with the latter representing significant rainfall events. Ebert (2001) noted that 1 mm/day threshold is useful in the construction of gridded

rainfall fields, in order to eliminate dew and insignificant rain. The forecast accuracy and skill in predicting rainfall occurring at or above each of the various thresholds are subsequently investigated.

The MMENS is constructed from forecasts for the UM and CCAM. The skill of the single-model ensemble forecasts is compared to the MMENS and the influences of each of the single-model ensemble systems on the MMENS accuracy and skill is described. Only for days where all of the ensemble members are available are used in the analysis.

The individual members of each of the single-models contribute equal weights to the respective single-model ensemble systems. The UM ensemble (UMENS) is created by adding the dichotomous forecasts at the grid points for each of the individual members, and to then dividing by N (the number of model configurations). Symbolically,

The same process is repeated for the CCAM ensemble (CCAMENS), by adding the dichotomous forecasts of the two deterministic forecasts and to then divide by :

The MMENS is then created by applying equal weights to the two single-model ensemble systems described above.

That is, the output from the UMENS is added to that of the CCAMENS and the total is then averaged, so that both models contribute equally to the MMENS.

3.3. Verification metrics

The score presented for the two thresholds are calculated over all of the 18 months of the three summer half-years. For a more detailed description of the results for individual months see Landman (2012).

The various statistical scores to be presented for the two thresholds were calculated over all of the 18 months of the three summer half-years. For a more detailed description of the results for individual months see Landman (2012).

The forecast bias is calculated in order to determine whether the ensemble systems have a wet (positive) or a dry (negative) bias. The forecast bias (–) explores whether a variable under consideration is systematically over-forecast or under-forecast, and is a measure of forecast accuracy. The perfect forecast would have a bias of 0. Here represents the total number of forecasts issued for the summer half-year.

The statistics used here for determining the performance of the dichotomous forecasts by using a contingency table (Table 1) are the Frequency Bias Index (FBI; *FBI* ——), Probability of Detection (POD; ——), False Alarm Rate (F; ——) and the Critical Success Index (CSI; ——). These verification scores are calculated to explore the accuracy of the dichotomous forecasts based on different thresholds. For each forecast or observation where the threshold is exceeded, the deterministic forecast becomes "yes (or 1)" and "no (or 0)" if the threshold is not exceeded. This forecast is then analysed using the contingency table which shows the frequency of "yes" and "no" forecasts relative to the observed occurrences (Joliffe and Stephenson, 2003; Wilks, 2006, Fawcett, 2008 – see Table 1). The series of verification statistics obtained in this way, for various threshold values, give an indication of the forecast to correctly predict the occurrence as well as the amount of rainfall (Ebert, 2001). This process was applied separately to the different ensembles that were formulated. Usually, the contingency tables are set-up to explore the average forecast performance over a model domain. This has the disadvantage that the verification scores represents an area average (Ebert, 2001) and cannot distinguish between different geographical locations of the domain or different weather regimes. For this reason, a contingency table was set-up for each grid box in the domain, and the scores calculated to present forecast performance at each grid box. In this manner the spatial patterns of the forecast performance can be evaluated. In this paper, only the area average values will be presented, with the spatial details provided by Landman (2012).

TABLE 1: The contingency table for the analysis of dichotomous forecasts

		OBSERVED		
		YES	**NO**	
FORECAST	**YES**	***a***	***b***	*forecast yes*
	NO	***c***	***d***	*forecast no*
		observed yes	*observed no*	**TOTAL**

a : HITS – the event was forecast and observed

b : FALSE ALARMS – the event was forecast but not observed

c : MISSES – the event was not forecast but observed

d : CORRECT NEGATIVES – the event was neither forecast not observed

For probability forecasts of dichotomous events the verification metrics consisted of the Brier skill score (BSS; Stanski et al (1990)), the reliability diagram (Jollife and Stephenson (2003); Wilks (2006)) as well as the relative operating characteristic (ROC; Jollife and Stephenson (2003); Wilks (2006)). These metrics give almost a complete diagnostic evaluation of forecast performance (Peel and Wilson, 2008).

The BSS (Stanski *et al*. 1990) is derived from the Brier score (BS; Wilks, 2006, Fawcett, 2008). The BSS answers the question of the relative skill the probability forecast (predicting whether the event occurred or not) has over that of the persistence (reference) forecast (Mason, 2004). The BS consists of the mean squared error in the probability forecast

– (1)

Here assumes a value of 1 if the event was forecast and 0 if the event was forecast not to occur. Similarly, a value of 1 is assigned to if the event did occur and 0 if the event did not occur. The three independent terms of the Brier score are also calculated and are indicated in (2).

Here is the number of times the events was observed for each of the forecasts made for each probability, , is the mean of all the observations and is the mean of all the forecasts (Wilks, 2006).

– – (2)

The reliability term needs to be as small as possible, which will indicate a well calibrated forecast because it summarizes the conditional bias of the forecast. The resolution term needs to be as large as possible, which will indicate that the forecast resolves the event strongly because it summarizes the ability of the forecasts to discern between events. The uncertainty term is only dependent on the climatological frequency of an event occurring and therefore is not influenced by the forecast.

In this paper the BSS is obtained by (3) since it has the advantage of being independent of how the forecasts are binned, where is the Brier score with persistence as the reference forecast.

—— (3)

The relative operating characteristics determines the discrimination of the forecast between events and non-events. The area under the ROC curve is calculated here with the trapezoid method and this value depends on the degree of separation of distribution of forecast probabilities conditional on the

occurrence of the event from the distribution conditional on non-events. (Wilks, 2006; Clark *et al*, 2008; Peel and Wilson, 2008).

The reliability diagram represents the relationship between the observed frequency and the forecast probability of an event (Joliffe and Stephenson, 2003, Wilks, 2006). The reliability diagram is a good companion to the ROC curve, where the reliability diagram is conditioned on the forecast. The reliability diagram shows what the observed frequency is given the forecast probability for that event to occur.

Together with the reliability diagram a sharpness or frequency diagram is constructed where the forecast probability bins are plotted against the frequency of the event forecast within each probability bin (over the verification period and at all the gridpoints). The sharpness diagram is in indication of the confidence of the forecast system under investigation.

4. Results

The average bias (calculated over the three summer half-years) for each of the ensemble systems in predicting 24-hour rainfall totals during is shown Figure 3 (a) to (c). . The map of the bias provides insight into the location of areas with relatively high and low as well as positive (blue shades) and negative (red shades) biases. Considering the three maps, it is noticeable that all three ensembles systems generally have positive biases over the entire domain. The spatial average bias was calculated for each of the ensemble systems and the value is given on the maps in Figure 3 (a) to (c). It is seen in Figure 3 (c) that the CCAMENS has the lowest average bias (0.63 mm/day) of all three systems, whereas the UMENS has the highest daily average bias of ~0.98 mm/day.

Figure 3

Considering the contingence table related scores, the UMENS generally outscores the MMENS with the lower threshold value of 1 mm/day, whereas the MMENS outscores both the single model ensemble systems with the 10 mm/day threshold values. The exception to this is with the POD values where the UMENS has a slightly higher detection rate at of the 10 mm/day events than the MMENS (Table 2).

TABLE 2:

In Figure 4, the BSS is presented spatially with a score value at each of the grid boxes. Figure 4 (a) – (c) represent the BSS for each of the three ensemble systems for the 1 mm/day threshold and Figure 4 (d) – (f) the BSS for the 10 mm/day threshold. On each of the maps, the percentage of grid points with positive BSS values is provided. This number gives an indication of the percentage of grid points

over the domain that has skill over that of the persistence (reference) forecast. Therefore, the greater this number, the greater the skill of the forecast is for the 18-month period.

Similar to the scores calculated with the contingency table, the MMENS is outscored on the area average percentage positive grid boxes for the case of the 1 mm/day threshold, but is more skillful than the single-model systems forthe 10 mm/day threshold. All three systems have no skill over the interior and west-coast of the country in predicting rainfall totals greater than 1 mm/day, with the remaining coastal regions having some skill over persistence. The MMENS only has 16.6% skillful grid boxes compared to 21.6% of the UMENS at 1 mm/day threshold, but for the important 10 mm/day threshold the MMENS outscores the two single-model systems, having skill over 72.2% of the total number of grid boxes. However, all three systems have difficulties in correctly forecasting the ocurrence of events greater than 10 mm/day over the Free State and parts of the northeast of the country. The results as depicted in Figure 3 has a significant bearing on operational weather forecasting, since it shows that the forecasters have a better chance of success for 10 mm/day thresholds as opposed to the low threshold of 1 mm/day. Weather forecasters and other users of NWP rainfall forecasts should therefore be careful with the interpretation and use of low threshold value forecasts.

FIGURE 4:

The ROC curves for all three ensemble systems are represented in Figure 5. In contrast to the low skill determined by the BSS for the 1 mm/day threshold events, the MMENS shows the best discrimination for these events (compared to the 10 mm/day thresholds) indicating that the ensemble can distinguish between rainfall and non-rainfall events (ROC areas > 0.6). The scores obtained by the MMENS for the 1 mm/day and 10 mm/day thresholds are in general very similar. However, the MMENS can skillfully distinguish between events and non-events for all the different thresholds.

Considering the ROC values in Table 3, it is shown that the single-model ensemble systems also display good discrimination abilities for both thresholds and both systems are skillful for these threshold during the summer half-year. It is interesting to note that although the CCAMENS scores systematically lower than the UMENS, the MMENS forecasts are more skillful than those of both the constituting single-model ensembles. This result may be due to the CCAMENS having only three possible outcomes for each probabilistic forecast at a given location (0, 0.5 or 1), whilst there are 4 possible outcomes for the UMENS and 12 for the MMENS. It is a significant result, which indicates the value of a multi-model ensemble system over single-model systems.

The ROC analysis has shown that the MMENS system is the most suitable to discriminate between rainfall events exceeding predetermined thresholds, from non-events. Hence, reliability diagrams are only presented for this system (Figure 6). The diagram however shows that the MMENS system exhibit over-confidence for both thresholds. Considering the 1 mm/day threshold (blue line) the

system is under-forecasting the events with low probabilities and over-forecasting for higher probabilities (< 70%). The 10 mm/day threshold (green line) has slightly better reliability but are over-forecasting the events > 30%. The sharpness diagrams in Figure 6 show that for both thresholds the MMENS has high confidence. In all the threshold events, the highest number of forecasts is made in the lower probability bins, with some increase with the 10 mm/day threshold events in the higher probabilities.

In order to accurately determine the difference between the three systems, the reliability, resolution and uncertainty are calculated for both threshold values and represented in Table 3. For the events exceeding the 10 mm/day threshold, the MMENS has a better resolution, but the UMENS has better resolution for 1 mm/day events. For reliability, the MMENS outscores at the lower threshold, but the CCAMENS is the most reliable of the three systems with 10 mm/day events. The same holds true for the uncertainty of the systems, except that the uncertainty is the lowest with the UMENS at 10 mm/day threshold.

In terms of the skill for each of the ensemble systems, the three systems are skillful in predicting rainfall for the South African domain. All of the systems are however less skillful in predicting low threshold events (1 mm/day). Looking at the multi-models' ability to distinguish between events and non-events, the multi-model has better discrimination abilities than the two single-model ensemble systems. The multi-model ensemble system can possibly be improved by removing the model errors within ensemble members as well as through the use of a weighted combination method that considers the relative skill of the individual contributing ensemble members.

5. Discussion and Conclusions

Weather forecasters at operational centres such as SAWS are often faced with the challenge of making reliable probabilistic forecasts for rainfall to occur over certain areas for the next day or two. Even though they are presented with forecasts from a number of configurations from the same weather forecast model, and/or forecasts from other models, for the forecaster to combine the various forecast output into a probability statement is done highly subjectively and is often based on a forecaster's own experience or preference of a particular model. Combining forecasts through a simple un-weighted approach into a single objective probability forecast was presented here, and the forecasts were verified over three summer half-years.

The results show that in general combining forecasts from different forecast systems outscores forecasts from the individual models. Care should however be taken when using this multi-model system in predicting low threshold values (i.e. 1 mm/day). In fact, the systematic overestimation of rainfall by all three ensemble systems over the interior of South Africa, the absence of skill in predicting the occurrence of rainfall above the 1 mm threshold event, and relatively poor performance of all systems in predicting events above the 10 mm threshold over the central interior of South Africa,

warrants research into the improvement of convective rainfall parameterizations, and perhaps the application of non-hydrostatic models at very high resolution, over South Africa (e.g. Engelbrecht et al., 2007). The paper has also demonstrated the attributes of combining forecasts produced by different institutions running different forecast models, and therefore suggests that additional models' output may be considered for including in a multi-model system for operational weather forecasting practices. Additional forecast output to consider includes forecasts from the Weather and Research Forecast model (to be used as operational model at SAWS and at the University of Pretoria), the NCEP ensemble, and possibly forecasts from the European Centre for Medium-Range Weather Forecasts.

Apart from improving on model physics and numerics, future NWP research in South Africa should address the best way to weigh forecasts form different models, downscaling or recalibrating forecasts (since it was shown here that the different models have different systematic errors), and the use of larger forecast ensembles. In addition, the use of even higher resolution forecasts beyond the hydrostatic limit should be considered.

6. Acknowledgements

The South African Weather Service who funded this research study. The Water Research Commission, through project K5/1646 funded the creation of the CCAM hindcast data used in this study. We would like to thank the CHPC for their excellent support whilst performing the large set of CCAM hindcasts.

7. References

BAKHSHAII A and STULL R (2009) Deterministic Ensemble Forecasting Using Gene-Expression Programming. *Weather Forecast.* **24** 1431-1451

BOWLER NE, ARRIBAS A, MYLNE KR, ROBERTSON KB and BEARE SE (2008) The MOGREPS short-range ensemble prediction system. *Quart.J.R. Meteorol. Soc.* **134** 703-722

CLARK AJ, GALLUS WA and CHEN T (2008) Contributions of Mixed Physics versus Perturbed Initial/Lateral Boundary Conditions to Ensemble-Based Precipitation Forecast Skill. *Mon. Weather Rev.* **136** 2140-2156

DAVIES T, CULLEN MJP, MALCOLM AJ, MAWSOM MH, STANIFORTH A, WHITE AA and WOOD N (2005) A new dynamical core for the MetOffice's global and regional modelling of the atmosphere. *Quart. J. R., Meteor. Soc.* **131** *1759-1782*

EBERT EE (2001) Ability of a Poor Man's Ensemble to Predict the Probability and Distribution of Precipitation. *Mon. Weather Rev.* **129** 2461-2480

ENGELBRECHT FA, MCGREGOR JL and ENGELBRECHT CJ (2009). Dynamics of the conformal-cubic atmospheric model projected climate-change signal over southern Africa. *Int. J. Climatol.* **29** 1013-1033. DOI: 10/1002/joc.1742. 29 1013-1033.

ENGELBRECHT CJ., ENGELBRECHT FA and DYSON LL (2012). High-resolution model-projected changes in mid-tropospheric closed-lows and extreme rainfall events over southern Africa. *Int. J. Climatol*. DOI: 10.1002/joc.3240.

ENGELBRECHT FA, LANDMAN WA, ENGELBRECHT CJ, LANDMAN S, BOPAPE MM, ROUX B, MCGREGOR JL and THATCHER M (2011) Multi-scale climate modelling over Southern Africa using a variable-resolution global model. *WaterSA*. **37** No. 5, WRC 40-Year Celebration Special Edition 2011

ECKEL FA and MASS CF (2005) Aspects of Effective Mesoscale, Short-Range Ensemble Forecasting. *Weather Forecast.* **20** 328-350

FAWCETT R (2008) Verification Techniques and Simple Theoretical Forecast Models. *Weather Forecast.* **23** 1049-1068

FRIEDERICHS P and HENSE A (2008) A Probabilistic Forecast Approach for Daily Precipitation Totals, *Weather and Forecasting.* **23** 659-673

GHILE YB, and SCHULZE RE (2009) Evaluation of Three Numerical Weather Prediction Models for Short and Medium Range Agrihydrology Applications. *Water Resource Management*, **24** 1005-1028

GRIMIT EP and MASS CF (2002) Initial Result of a Mesoscale Short-Range Ensemble Forecasting System over the Pacific Northwest, *Weather Forecast.* **17** 192-205

HAMILL TM and COLUCCI SJ (1997) Verification of Eta-RSM Short-Range Ensemble Forecasts. *Mon. Weather Rev.* **125** 1312-1327

JOLIFFE IT and STEPHENSON DB (2003) Forecast Verification: A Practitioner's Guide in Atmospheric Sciences. John Wiley & Sons Ltd, England, 1-1630pp

KALNAY E (2003) Atmospheric Modeling, Data Assimilation and Predictability. Cambridge University Press, UK

LANDMAN S, ENGELBRECHT FA, ENGELBRECHT CJ, LANDMAN WA and DYSON L.L (2010) A Multi-Model Ensemble System for Short-Range Weather Prediction in South Africa. Proc.*26th Annual Conference of the South African Society for Atmospheric Sciences.* September 2010, Gariep Dam. ISBN 978-0-620-47333-0

LANDMAN S (2012) A multi-model ensemble system for short-range weather prediction in South Africa. M.Sc. Thesis. University of Pretoria, Pretoria. 137 pp.

LEWIS JM (2005) Roots of Ensemble Forecasting. *Mon. Weather Rev.* **133**, 1865-1885

LORENZ EN (1963) Deterministic Nonperiodic Flow. *J. Atmos.. Sci.* **20** 130-141

MASON SJ (2004) On Using “Climatology” as a Reference Strategy in the Brier and Ranked Probability Skill Scores. *Notes and Correspondence*, **132** 1891-1895

MCGREGOR JL (2003) A new convection scheme using simple closure. In: Current Issues in the Parameterizations of Convection. BMRC Research Report 93. 36-36

MCGREGOR JL (2005) C-CAM: Geometric aspects and dynamical formulation. CSIRO Atmospheric Research Tech. Paper No. 70. 43 pp.

PEEL S and WILSON LJ (2008) A Diagnostic Verification of the Precipitation Forecasts Produced by the Canadian Ensemble Prediction System. *Weather Forecast*. **23** 569-616

POTGIETER CJ (2006) Short-range weather forecasting over southern Africa with the conformal-cubic atmospheric model. M.Sc. Thesis University of Pretoria, Pretoria. 172 pp.

ROY BHOWMIK SK and DURAI VR (2010) Application of multimodel ensemble techniques for real time district level rainfall forecasts in short range time scale over Indian region. *Meteorol. Atmos. Phys.* **106** 19-35

STAËL VON HOLSTEIN CS (1971) An Experiment in Probabilistic Weather Forecasting. *J. Applied Meteor*. **10** 635-645

STANSKI H.R, WILSON LJ and BURROWS WR (1990) Survey of common verification methods in meteorology. WMO World Weather Watch Tech. Rep. 8 WMO TD 358, 114 pp.

STENSRUD DJ and YUSSOUF N (2005) Bias-corrected short-range ensemble forecasts of near surface variables. *Meteor. Appl.* **12** 217-230

STENSRUD DJ, BROOKS HE, DU J, TRACTON S and ROGERS E (1999) Using Ensembles for Short-Range Forecasting. *Mon. Weather Rev.* **127** 433-446

TENNANT WJ, TOTH Z and RAE KJ (2007) Application of the NCEP Ensemble Prediction Syustem to Medium-Range Forecasting in South Africa: New Products, Benefits, and Challenges. *Weather Forecast.* **22** 18-35

THEIS SE, HENSE A and DAMRATH U (2005) Probabilistic precipitation forecasts from a deterministic model: a pragmatic approach. *Meteor. Appl*., **12** 257-268

TOTH Z, XHU Y and MARCHOK T (2001) The Use of Ensembles to Identify Forecasts with Small and Large Uncertainties. *Weather Forecast.* **16** 463-477

TYSON PD (1986) Climate Change and Variability over southern Africa. Oxford University Press, Cape Town

WANDISHIN MS, MULLEN SL, STENSRUD DJ and BROOKS HE (2001) Evaluation of a Short-Range Ensemble System. *Mon. Weather Rev.* **129** 729-747

WILKS D (2006) Statistical methods in the atmospheric sciences. 2nd Edition, Elsevier Academic Press, California, United States of America

ZONGJIAN KE, DONG W and ZHANG P (2008) Multimodel Ensemble Forecasts for Precipitation in China in 1998. *Adv. Atmos. Sci*. **25** 72-82

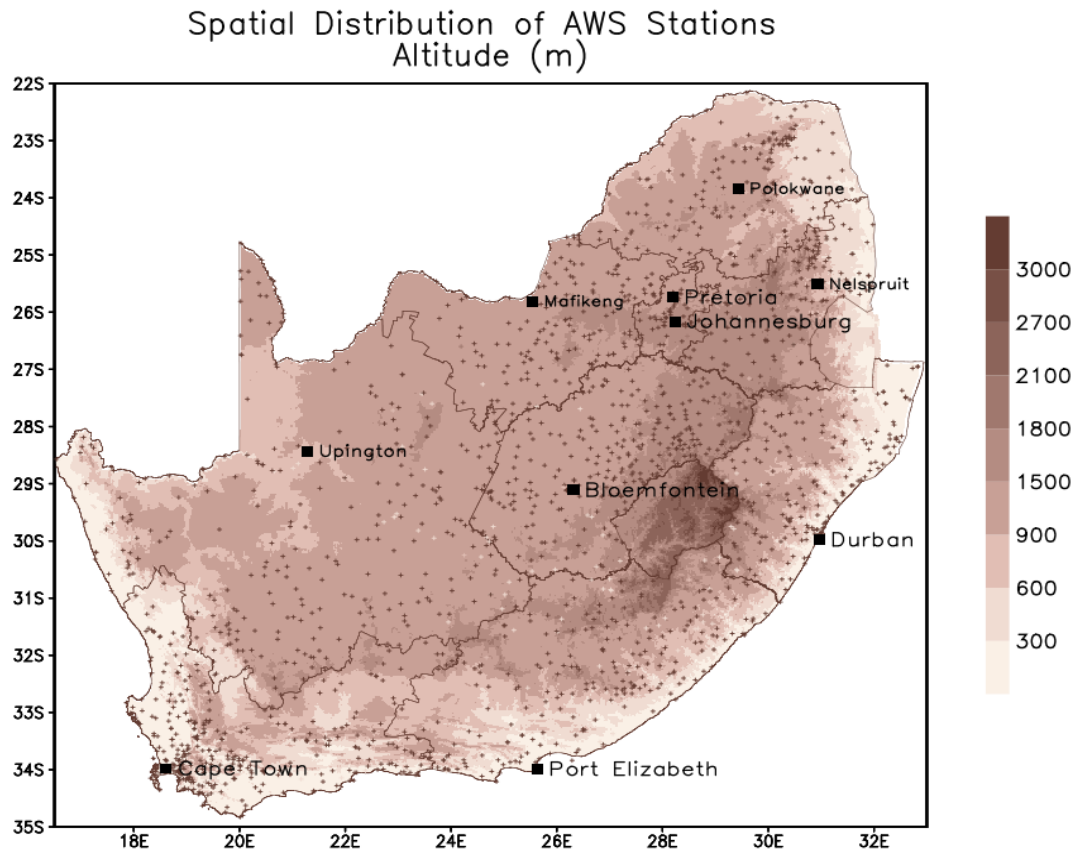

FIGURE 1: Location of rainfall stations in the combined SAWS-ARC data set.

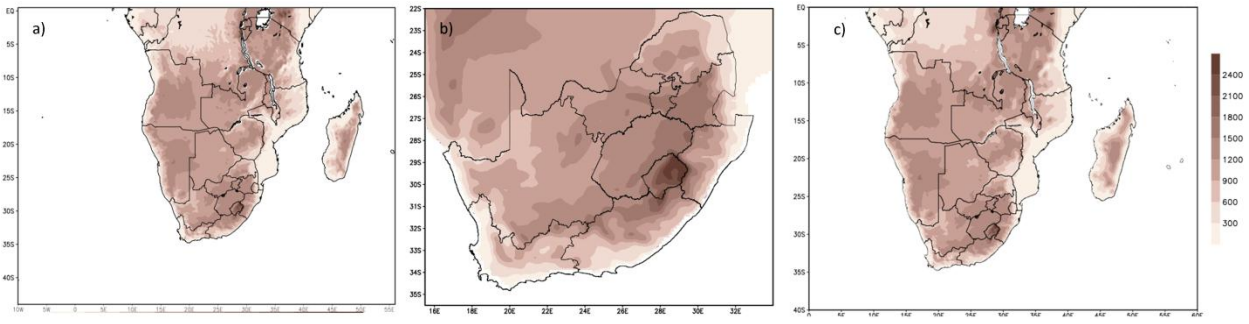

FIGURE 2: Domain size maps of the individual members of the multi-model ensemble system; (a) UM 12 km resolution members, (b) UM 15 km resolution member and (c) CCAM members.

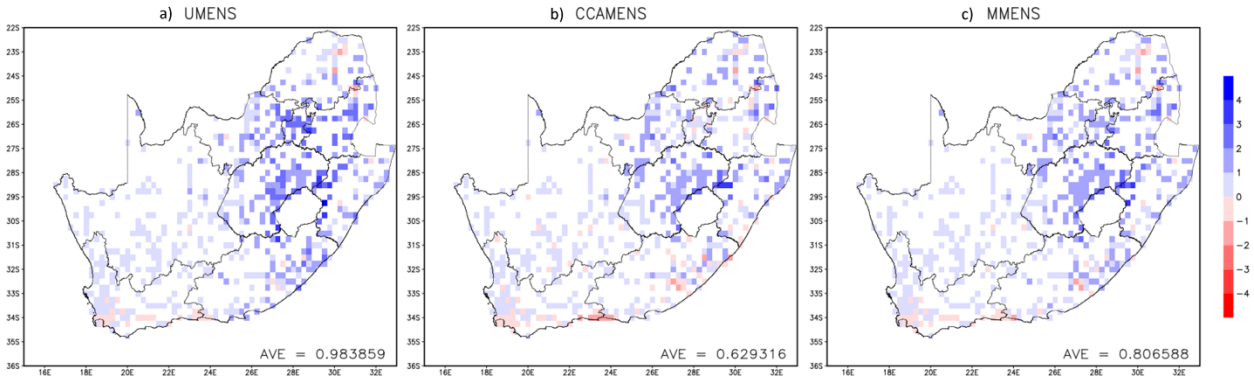

FIGURE 3: The spatial maps for the daily bias for the three ensemble systems; (a) UMENS, (b) CCAMENS and (c) MMENS.

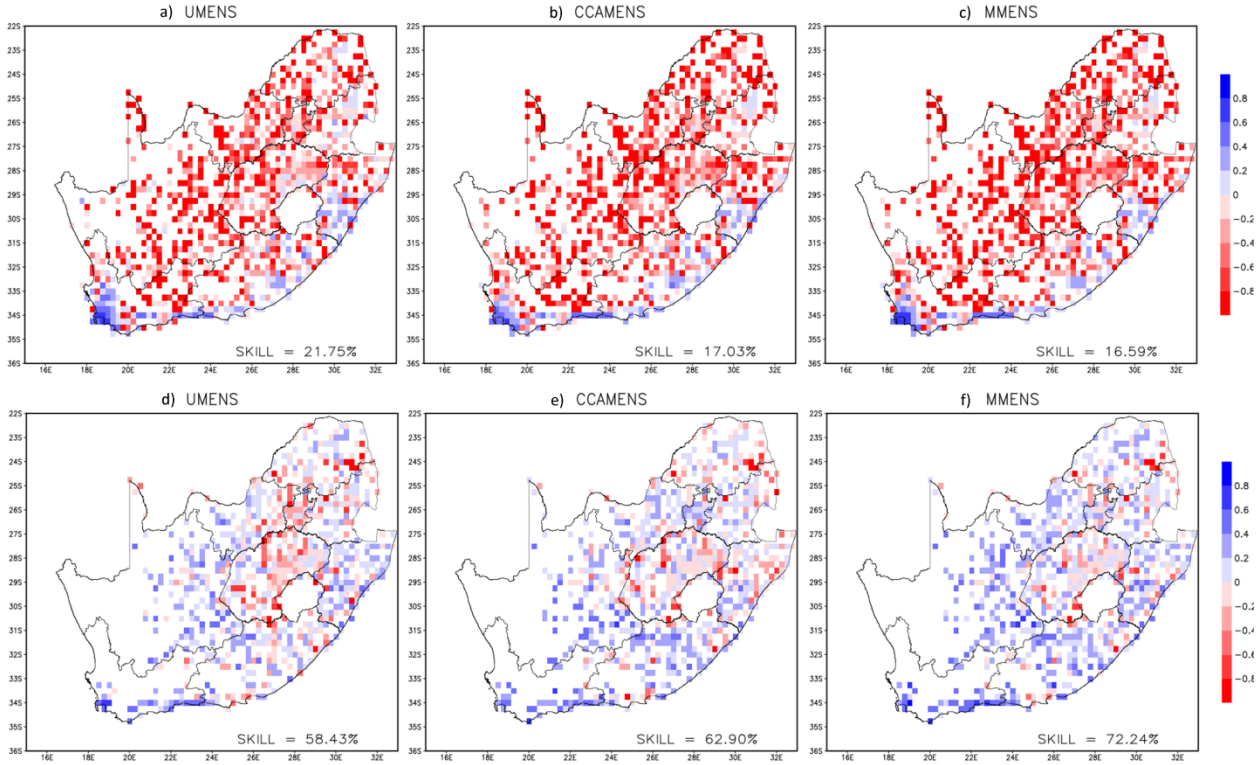

FIGURE 4: The spatial maps for the Brier skill score for the three ensemble systems. a) UMENS, (b) CCAMENS and (c) MMENS represent the Brier skill score with 1 mm/day threshold and (d) UMENS, (e) CCAMENS and (f) MMENS 10 mm/day threshold.

TABLE 2: Summary of the area average values of the verification scores calculated from the contingency table for both the threshold values as well as for all three ensemble systems. The best verification score is indicated in bold for each of the systems

VERIFICATION METRIC	UMENS		CCAMENS		MMENS	
	1 mm/day	10 mm/day	1 mm/day	10 mm/day	1 mm/day	10 mm/day
FBI – 1 mm/day	**2.515**	1.417	2.629	1.213	2.797	**1.211**
CSI – 1 mm/day	**0.362**	0.216	0.340	0.187	0.352	**0.225**
FAR – 1 mm/day	**0.309**	0.057	0.330	0.049	0.38	**0.047**
POD – 1 mm/day	0.823	**0.393**	0.807	0.323	**0.862**	0.371

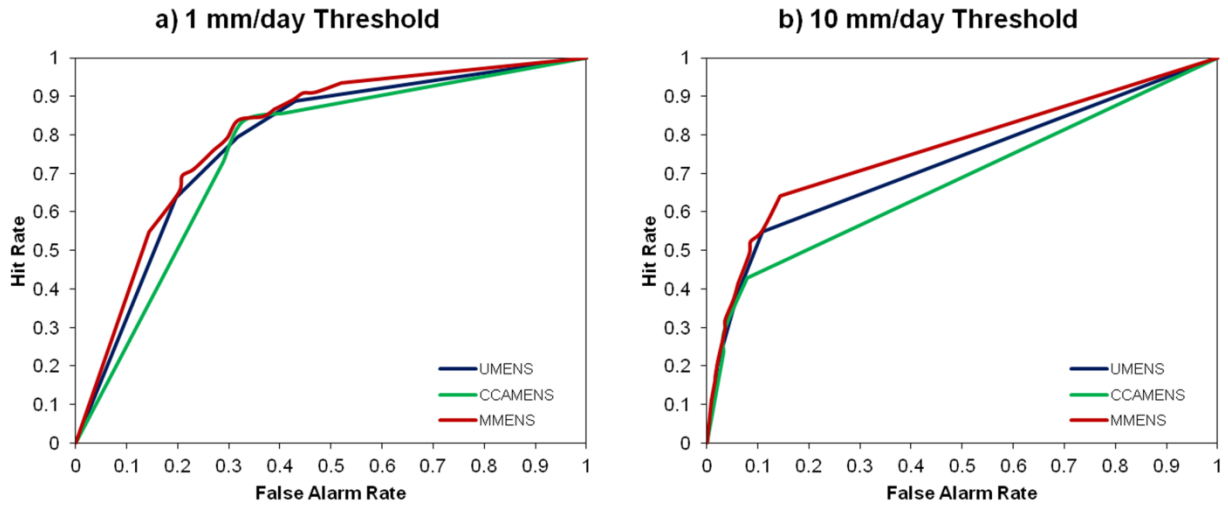

FIGURE 5: The ROC curves for all three ensemble systems. The UMENS curve is in blue, the CCAM in green and the MMENS in red. The two threshold values are represented by (a) 1 mm/day and (b) 10 mm/day respectively.

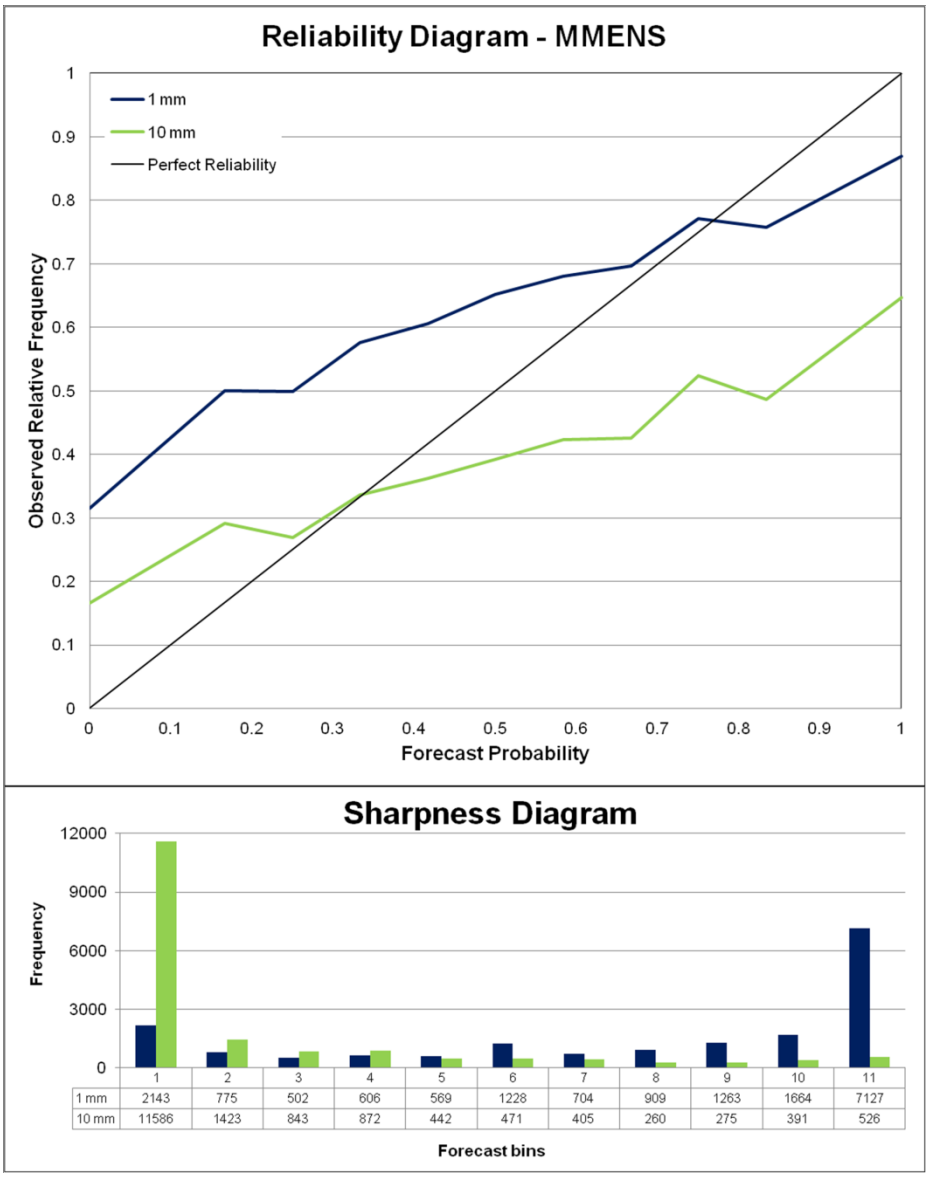

FIGURE 6: The reliability and sharpness diagram for the MMENS. The two threshold values are represented by 1 mm/day the blue line/bar and 10 mm/day the green line/bar.

TABLE 3: Summary of the resolution, reliability and uncertainty scores as calculated by the three terms of the Brier skill score for both the threshold values as well as for all three ensemble systems. The best score is indicated in bold for each of the systems

VERIFICATION METRIC	UMENS		CCAMENS		MMENS	
	1 mm/day	10 mm/day	1 mm/day	10 mm/day	1 mm/day	10 mm/day
ROC Score	0.719697	0.724556	0.648406	0.674402	**0.765278**	**0.762357**
Resolution	**0.370072**	0.076532	0.306917	0.178127	0.130799	**0.248717**
Reliability	0.069963	0.158378	0.079857	**0.044348**	**0.04625**	0.128062
Uncertainty	0.236747	**0.224744**	0.209404	0.235932	**0.178549**	0.241282