9B.6 A PROBABILISTIC EVALUATION OF GLOBAL TROPICAL CYCLONE FORECASTS FROM THE UPGRADED MET OFFICE MOGREPS-G ENSEMBLE, AND THE VALUE OF MULTI-MODEL ENSEMBLES

Helen A. Titley * and Rebecca Stretton
Met Office, Exeter, UK

1. INTRODUCTION

Ensemble forecast models have a vital role to play in tropical cyclone forecasting, through their ability to highlight the situation dependent uncertainty and provide probabilistic forecast information to help inform decision makers. At the Met Office, tropical cyclone tracking (Heming, 2017) is run in real time on the Met Office MOGREPS-G ensemble, and the ECMWF ENS and NCEP GEFS ensembles. The three ensembles are also combined in to a 108-member multi-model ensemble. A range of products, including track and intensity forecasts for both named and forming storms, are produced and distributed to several operational tropical cyclone forecasting centres. The forecasts are evaluated using a probabilistic verification framework.

2. JULY 2017 UPGRADE TO MOGREPS-G

In July 2017, MOGREPS-G was upgraded from N400 (~33km) to N640 (~20km) grid resolution. The number of ensemble members run each time was also increased from 12 to 18. The model runs four times a day (00/06/12/18UTC) out to T+168, with the most-recent two runs being combined so that each new MOGREPS-G cycle consists of 36 members from the time-lagged ensemble.

In an ideal reliable ensemble, the root mean square error of the ensemble mean forecast should be equal to the ensemble spread. Verification of a 3 month trial period (July to September 2016) showed a significant improvement in the error/spread relationship for both tropical cyclone track and intensity in the upgraded model i.e. a narrowing of the gap between the ensemble spread and error through both an increased spread and lower error.

3. 2017 HURRICANE SEASON: CASE STUDIES

The active and high-impact 2017 North Atlantic hurricane season has provided many interesting cases to evaluate the performance of the upgraded MOGREPS-G ensemble forecasts, and compare to other global ensembles, both individually and through multi-model ensemble combination with ECMWF ENS and NCEP GEFS.

For Hurricane Harvey, the multi-model ensemble captured the possibility of a Texas landfall up to seven days ahead, with probabilities slowly increasing over time. MOGREPS-G began to give useful guidance for the possibility of re-intensification over the Gulf of Mexico six days prior to landfall (Figure 1). Although the majority of the ensemble members forecast a landfall to the south of the observed track, there is good spread, with the possibility of a Texas landfall shown by several members. Over the next 12 hours subsequent forecasts significantly increased the probability for the Texas landfall. Although after landfall the uncertainty in the ensemble storm track forecasts grew in the weak steering flow, there was a strong signal for stalling, which led to the extreme multi-day precipitation event.

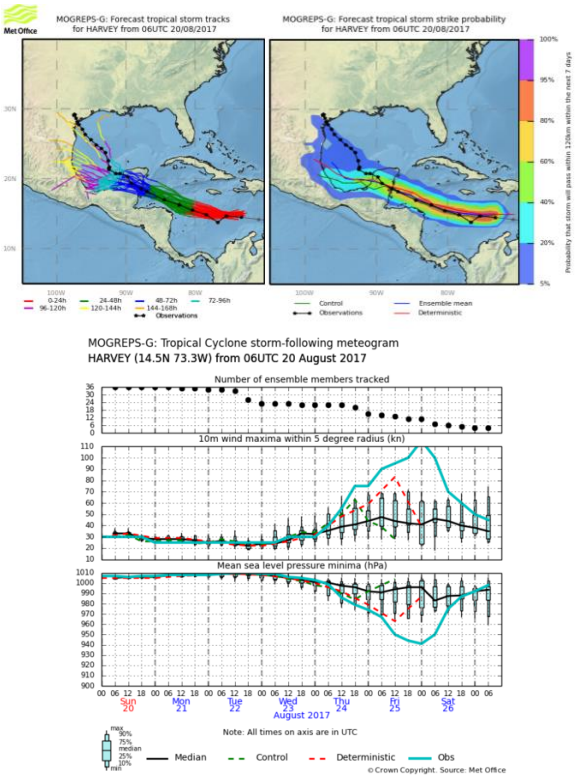

Figure 1. MOGREPS-G forecast products for Hurricane Harvey from 06UTC 20/08/2017: Forecast tropical storm tracks out to seven days from each MOGREPS-G member (top left), and a strike probability plot showing forecast probability that the storm will pass within 120km within the next seven days (top right); Storm-following meteogram summarising the forecasts of storm intensity (bottom).

For Hurricane Irma, MOGREPS-G was signalling a 40-60% probability of tropical storm genesis in the correct area west of Cape Verde (Figure 2), four days ahead of the first tropical storm advisory of Irma, and eleven days ahead of the Barbuda landfall.

Irma was an interesting case in the perception of forecast skill for different users. More perceived weight appeared to be given to whether a storm will impact a location (arguably of more importance for warning and preparedness) rather than when, therefore allowing more tolerance for along-track errors compared to cross-track errors. MOGREPS-G provided useful guidance for the turn to the north, and the landfall in the Florida Keys and the Gulf Coast of

* Corresponding author address: Helen A. Titley, Met Office, FitzRoy Road, Exeter, Devon, EX1 3PB, UK; email: helen.titley@metoffice.gov.uk

Florida (giving relatively low cross-track errors) in several forecast runs where other forecasts were concentrated much further east (Figure 3). The forward speed of Irma, however, was underestimated due to more southerly tracks and increased land interaction over Cuba. Although this negatively impacted overall track errors, the model received very favourable feedback from a user-orientated perspective.

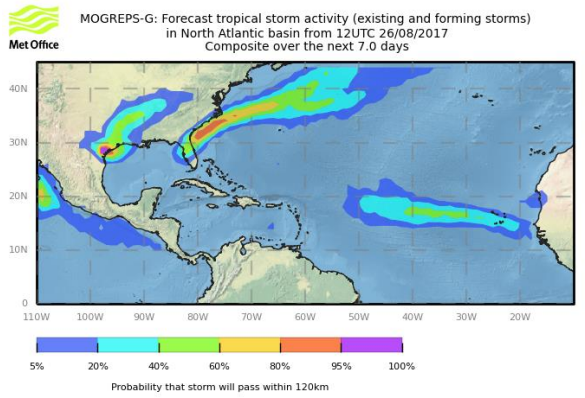

Figure 2. MOGREPS-G forecast of tropical storm activity from 12UTC 26/08/2017, defined as the probability that any storm (whether it is an existing named storm, or one in the pre-genesis phase that is forecast to form during the forecast), will pass within 120km over the next 7 days.

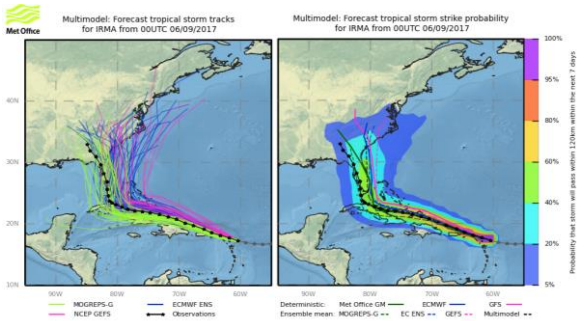

Figure 3. Multi-model ensemble forecasts for Hurricane Irma from 00UTC 09/09/2017: Tracks coloured according to model (left), and multi-model ensemble strike probability with deterministic (solid), ensemble mean (dashed) tracks (right).

4. OBJECTIVE VERIFICATION

To fully assess the skill and value of ensemble tropical cyclone forecasts, a framework to produce objective verification has been developed at the Met Office. This includes the verification of probabilistic forecasts from each global ensemble, and the various multi-model combinations, for named storm strike probability forecasts and tropical cyclone activity forecasts (for named and forming storms). The verification is produced twice yearly, at the end of the northern and southern hemisphere seasons, for all basins over the previous 12 months.

A range of probabilistic verification statistics are calculated to assess the skill, reliability and value of the forecasts. These include the Relative Operating Characteristic (ROC), reliability diagrams, relative economic value and Brier Skill Score.

The ROC plot assesses the skill of the forecast at discriminating between events and non-events. The points along the curve are the hit rates and false alarm rates for each probability bin. Perfect skill would produce a curve from bottom left to top left to top right.

Reliability diagrams display how well the predicted probabilities correspond to their observed frequencies. Perfect reliability would be a diagonal line from (0,0) to (1,1), a line above the diagonal indicates under-forecasting and below the diagonal shows over-forecasting.

For a given user, their cost-loss ratio is the term given to the ratio of the cost of a preventative measure to the loss averted, and can be used to guide the probability threshold above which to take action. The relative economic value plot displays the relative improvement in economic value between the sample climatology and a perfect forecast for all cost loss ratios. It shows how value varies depending on the user's specific cost-loss ratio. An ideal forecast would have an economic value of 1 for all cost loss ratios.

The Brier Skill Score assesses the relative skill of the probabilistic forecast over that of climatology, in terms of predicting whether an event occurred. A score of 0 indicates no skill when compared to the reference forecast and a score of 1 would be a perfect score. In this report the reference forecasts are CLIPER (CLImatology and PERsistence) forecasts for named storm strike probability and the sample forecast climatology for tropical cyclone activity forecasts.

For more information on these scores see the WWRP/WGNE Joint Working Group on Forecast Verification Research forecast verification web page) http://www.cawcr.gov.au/projects/verification/

4.1 Strike probability forecasts for named storms

Verification results for the strike probability forecasts for all named storms in the 12 month period January to December 2017 are shown in Figures 4 and 5. As the forecast is for the probability that the storm will pass within 120km within the next seven days, there is no lead time component in this verification.

Verification of the three ensembles and multi-model ensemble combination in Figure 4 displays good reliability for all models, with ECMWF ENS showing near perfect reliability for all probabilities. MOGREPS-G and NCEP GEFS both show over-forecasting for probabilities 50% and greater, with NCEP GEFS contrastingly showing slight under-forecasting for 0-20% probabilities. In the relative economic value plot, the multi-model ensemble value curve fully encompasses the three individual models showing the multi-model ensemble combination gives the greatest economic value for all cost-loss ratios. All the models display the greatest relative economic value (over 0.7) for very small cost loss ratios (0 to 0.1). For tropical cyclones, user's cost-loss ratios vary significantly but are often very low due to high potential losses. The ROC plot shows similar false alarm rates for all the models with more variation in the hit rates. Overall, all the models have good skill in forecasting tropical cyclones, with the greatest skill shown by the multi-model ensemble.

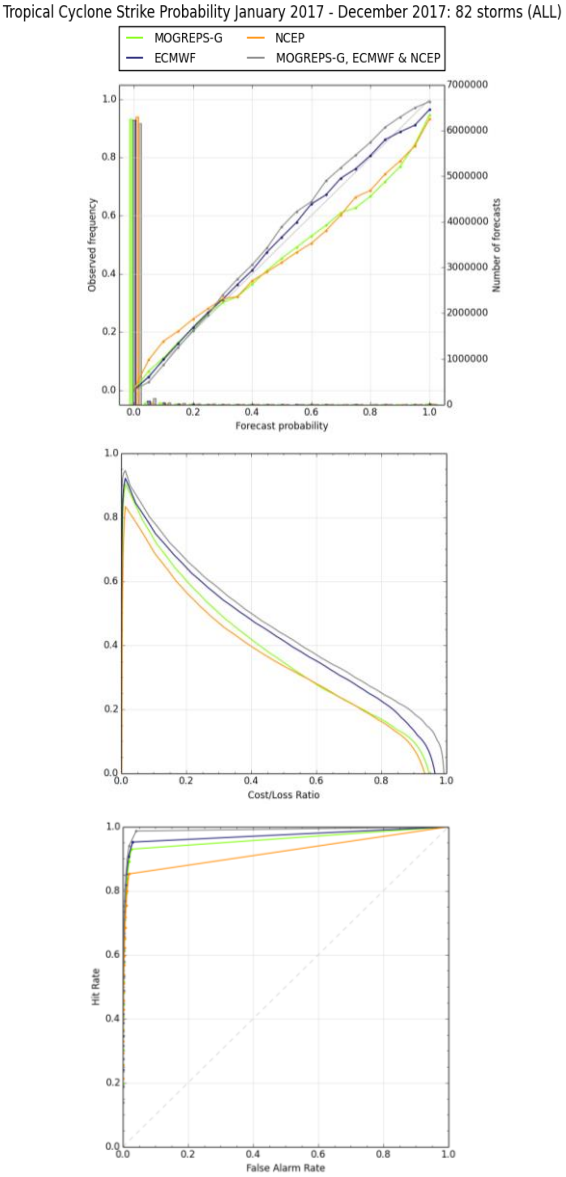

Figure 4. Verification plots comparing MOGREPS-G, ECMWF ENS, NCEP GEFS and multi-model ensemble forecasts of named storm strike probability, from January to December 2017. Reliability diagram (top), relative economic value plot (middle) and ROC plot (bottom).

Figure 5 compares the Brier Skill Score for the strike probability forecasts between model, basin and storm. ECMWF ENS is the most skilful of the three included global ensembles in all basins, with the relative performance of MOGREPS-G and NCEP GEFS varying from basin to basin. However, in all basins additional skill is gained using the multi-model ensemble. The storm-based verification of two high-profile storms (Hurricanes Irma and Matthew) further demonstrates the value in the multi-model ensemble as for each storm a different ensemble displays the highest skill (ECMWF ENS for Irma and MOGREPS-G for Matthew). In both cases, the multi-model ensemble is of comparative skill to the strongest performing model (which would not be known at the time of the forecast).

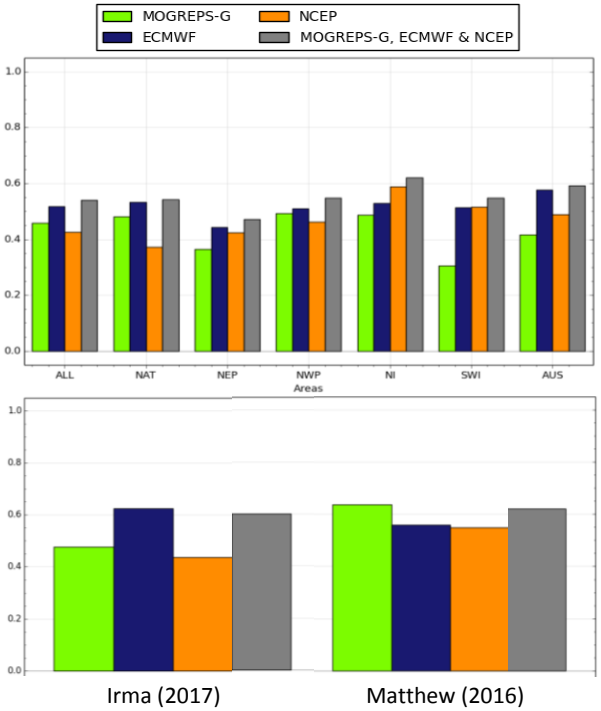

Figure 5. Brier Skill Score of MOGREPS-G, ECMWF ENS, NCEP GEFS and multi-model ensemble forecasts of named storm strike probability: All storms from January to December 2017 and split by tropical cyclone basin (top) and a comparison of Hurricanes Irma and Matthew (bottom).

4.2 24-hour tropical cyclone activity forecasts

Unlike the 7-day strike probability forecasts evaluated in Section 4.1, the tropical cyclone activity forecasts are the forecast probability that a tropical storm will pass within 120km in a given 24-hour period (from T+12 to T+156). Therefore, there is a lead time component in the verification. The forecast verification can be stratified to evaluate three types of activity forecasts: i) forecasts for storms that are named at forecast run time (named); ii) forecasts for storms that undergo genesis during the forecast (forming); and iii) forecasts for all storms (named and forming). Only the results of the first of these (named storms) are presented here, for the North Atlantic and North Pacific basins from July 2017 to December 2017.

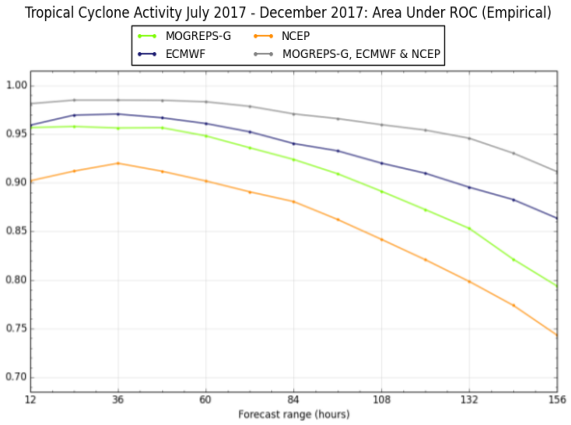

Figure 6. Area under the ROC against lead time for forecasts from MOGREPS-G, ECMWF ENS, NCEP GEFS and the multi-model ensemble.

The area under the ROC curve plotted against lead time displayed in Figure 6 provides a useful summary of the forecast skill between models and across all lead times. In cases such as tropical cyclone activity where there are many correct rejections, this score will be strongly dependent on the hit rate and false alarm rate of the highest probability bin. ROC areas range from 0 to 1, with 0.5 indicating no skill, and a score of 1 showing perfect skill. Figure 6 shows that although skill reduces with lead time, there is significant skill at all lead times, and that additional skill is gained at all lead times by using a multi-model ensemble in named storm tropical cyclone activity forecasts. The drop off in skill with lead time is also lowest for the multi-model ensemble forecasts.

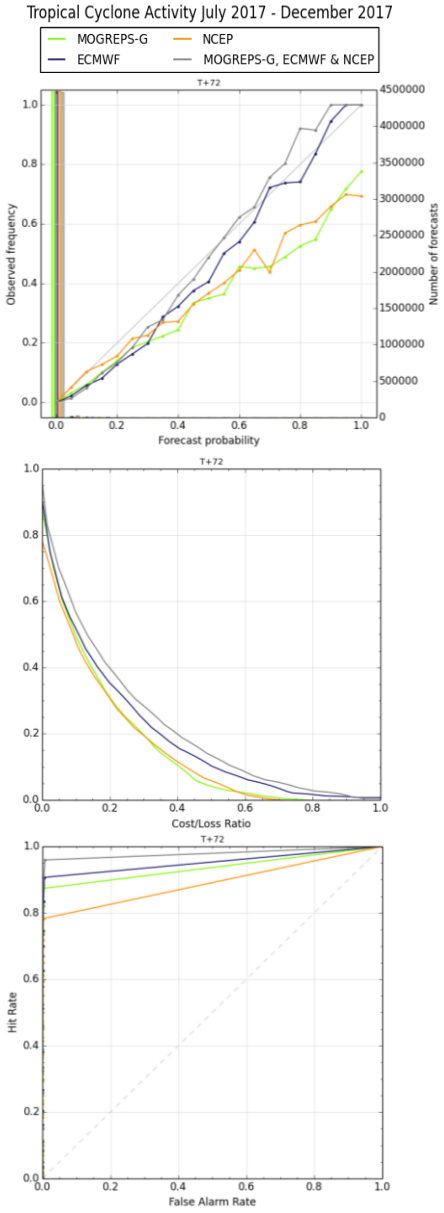

Figure 7. Verification plots comparing the MOGREPS-G, ECMWF ENS, NCEP GEFS and multi-model ensemble forecasts of tropical cyclone activity forecasts for named storms for the 24-hour forecast period centred on T+72: Reliability diagram (top), relative economic value plot (middle) and ROC plot (bottom).

Figure 7 provides a more detailed picture of the verification for one particular lead time, the 24-hour forecast period centred on T+72. Overall, the results are similar to those seen in the 7-day strike probability verification in Section 4.1, with additional skill, reliability and value being gained by using the multi-model ensemble.

Initial analysis of the other two types of tropical cyclone activity verification, (forming storms and all storms (named and forming)) indicates that in many cases the models are over-forecasting the probability of developing forming storms, which is hampering overall forecast skill.

5. FUTURE PLANS

5.1 MOGREPS-G model upgrades in 2018/19

Several changes are planned to the MOGREPS-G ensemble over the next two years. A change scheduled for July 2018 will include improvements to stochastic physics. Verification of trials of the updated model has shown an improvement in the ensemble spread in the tropics.

A major change is then scheduled to go live in late 2018 or early 2019, as the ensemble perturbation system used in MOGREPS-G is changed from Ensemble Transform Kalman Filter (ETKF) to an ensemble of data assimilations (En-4DEnVar, Bowler et al., 2017). In the new system, data assimilation is performed for each member, creating increments relative to its own background trajectory. Figure 8 shows that for 850hPa winds in the tropics although in the current ensemble system ETKF gives good spread at initial time, this spread grows too slowly compared to the root mean square error. Comparative trials of the new En-4DEnVar have shown much faster spread growth, with a much better match to errors. A partial re-centring around the deterministic analysis gives an additional increase in skill and reduces jumpiness. The effect on tropical cyclone track and intensity is currently being evaluated using trial data.

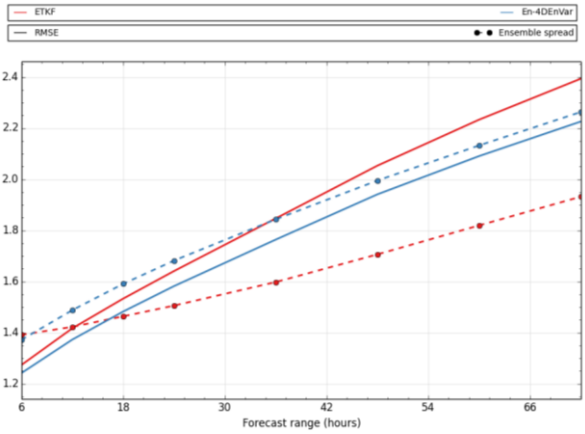

Figure 8. RMSE (solid) and spread (dashed) of wind at 850hPa for the tropics from the current MOGREPS-G ensemble with ETKF perturbations (red), and the planned MOGREPS-G upgraded ensemble using En-4DEnVar (blue). Both are verified against ECMWF analyses.

5.2 HIWeather collaboration on the ensemble forecasting of tropical cyclones

A new international collaboration on the topic of the ensemble forecasting of tropical cyclones, as a sub-project of the WMO's HIWeather project, aims to enhance collaboration amongst the research and operational community on this topic. It aims to evaluate and demonstrate the benefits of using ensemble forecasts, and gather current and future user requirements, with a view to developing new and user-orientated ways to display and verify probabilistic tropical cyclone forecasts, and increase the use of ensemble forecasts in tropical cyclone forecasting.

Following recent resolution increases in several global ensembles, it is important to re-evaluate current levels of probabilistic forecasting skill for tropical cyclone intensity. Therefore, one task as part of the HIWeather collaboration is to carry out a number of verification activities identifying the current level of forecasting skill for tropical cyclone intensity from global ensemble forecasts. At the Met Office that work is focussing on evaluating the ability of global ensemble forecasts to verify intensity trends. Initial results show that although in many cases (e.g. Hurricane Harvey in Figure 9) good guidance is given of the forecast intensity trend, there are still many cases where the ensemble resolution is insufficient to predict the timing and extent of intensification (e.g. Hurricane Irma in Figure 9).

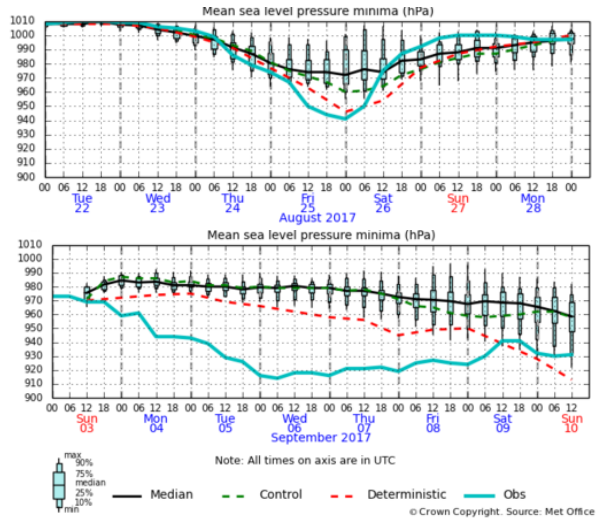

Figure 9. MOGREPS-G Intensity forecasts (measured by the mean sea level pressure minima of the storm) for Hurricane Harvey from 00UTC 22nd Aug 2017 (top) and Hurricane Irma from 12UTC 3rd Sep 2017 (bottom). The observed intensity is overlain in cyan.

6. CONCLUSIONS

- To maximise the benefit of ensemble forecasts in tropical cyclone forecasting, and to fairly evaluate model upgrades and inter-model comparisons, it is essential to use the full probabilistic information in forecast products and ensemble forecast verification.
- Ensemble forecast models are shown to have provided useful guidance in many of the high-profile hurricanes of 2017.
- The relative skill between ensemble forecast models in named storm tropical cyclone strike probability forecasts varies from storm to storm and basin to basin.
- A clear benefit is shown in using multi-model ensemble forecasts over the most skilful individual ensemble forecast model, both in named storm 7-day strike probability forecasts and in tropical cyclone activity forecasts.

7. REFERENCES

Bowler, N.E., Clayton, A.M., Jardak, M., Jermey, P.M., Lorenc, A.C., Wlasak, M.A., Barker, D.M., Inverarity, G.W. and Swinbank, R., 2017: The effect of improved ensemble covariances on hybrid variational data assimilation. *Q.J.R. Meteorol. Soc.*, **143**: 785–797.

Heming, J.T., 2017: Tropical cyclone tracking and verification techniques for Met Office numerical weather prediction models. *Met. Applications,* **24**: 1-8.