12.2 PROBABILISTIC GUIDANCE OF AVIATION HAZARDS FOR TRANSOCEANIC FLIGHTS

K. A. Stone, M. Steiner, J. O. Pinto, C. P. Kalb, C. J. Kessinger
NCAR, Boulder, CO

M. Strahan
Aviation Weather Center, Kansas City, MO

1. INTRODUCTION

An important aspect in planning for transoceanic flights is the acquisition and interpretation of forecasts of aviation weather hazards that might be encountered along a desired flight path. Because of the long distances flown, look-ahead times of 24 hours or more are needed for planning purposes. The current ICAO-sanctioned significant weather forecasts (SIGWX) serving this purpose are manually generated, with coarse resolution in space and time, and deterministic. The next generation of products will likely be gridded and probabilistic forecasts, derived from a combination of numerical weather prediction (NWP) model outputs, possibly generated by different centers (e.g., World Area Forecast Center (WAFC) London and WAFC Washington).

Here, we present an effort aimed at leveraging multiple global ensemble forecasts to generate globally harmonized, probabilistic forecast guidance products. The methodology developed will help improve the process of predicting significant weather in the strategic planning timeframe for transoceanic flights. Emphasis in this study is placed on calibration of individual models and methods for the subsequent fusion of probabilistic forecasts (i.e., harmonization).

Several techniques for the calibration and combination of global ensemble models are presented, and we show how improvements in the harmonized forecast are possible even with spatially and temporally varying performance of each individual model.

Perhaps more important than improved forecast performance, the approach presented allows qualified international providers to collaboratively contribute to a single, global forecast, supporting the forecast's acceptance and use in oceanic air route planning. During long-distance flights across multiple countries, it is critical that the international aviation community agrees to use such a single source of weather information, which will make air traffic safer and more efficient to manage.

Corresponding author address: Ken Stone, National Center for Atmospheric Research, PO BOX 3000, Boulder, CO 80307. Email: kstone@ucar.edu

2. DATA AND CASE STUDY SET-UP

The case study was conducted using data from climatologically distinct seasons (March – May (MAM) and June – August (JJA)) and regions (the two domains outlined in Figure 1, Caribbean and Pacific), and data from four global ensemble forecasts generated by the National Centers for Environmental Prediction (“NCEP”), Canadian Meteorological Center (“CMCE”), The Met Office (“UKMO”), and the European Center for Medium Range Forecasts (“ECMF”). The latter two forecasts were obtained from the TIGGE database (Bougeault et al., 2010) and the former two from the NOAA Operational Model Archive and Distribution System (NOMADS) data feed.

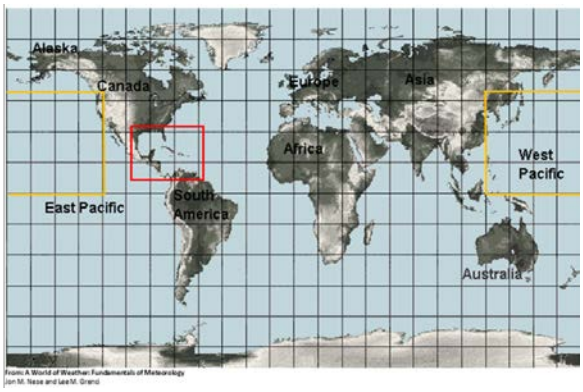

Figure 1. Pacific (0N – 50N, 120E – 240E) and Caribbean (5N – 35N, 260E – 310E) domain used in study.

The goal was to examine regional and seasonal differences in performance when using different weighting schemes to combine the calibrated probabilities obtained from each ensemble model, particularly when using different calibration approaches.

Probabilities from each model were computed for every pixel using a relative frequency of occurrence based on ensemble member values exceeding a selected hazard threshold. For example, if 10 of 20 members indicated a quantity greater than a given threshold, then p=0.5. Only the perturbed members of the ensemble were used and not the control; however, we anticipate this would have only minor (if any) impacts on the results.

The focus during this initial methodology development effort was on convective storm hazards, but the approach is applicable for other weather hazards as well.

2.1 Verification Field

As a starting point, and for simplicity in interpreting comparisons, we used accumulated liquid water precipitation over a 6-hour period (APCP6hr) as the predictand. APCP6hr is readily available in all four models, and a reasonably well-understood observation field (in this case CMORPH 3-hour rain rates (Joyce et al., 2004)) can be used to generate a physically equivalent validation field. The CMORPH data were down-sampled to the horizontal resolution of the model grid (1 degree x 1 degree), and integrated over two 3-hour rain rate periods to obtain a 6 hour accumulated value, which is comparable to that available from the global model ensembles. In this study we used a threshold of 2 mm on the observed precipitation as a proxy for significant weather.

Although the subject of on-going research, we believe the overall approach will extend to convection once a verification field focused specifically on convection is more readily available. Also, recall that convective hazards for aviation include turbulence, wind shear, lightning, heavy precipitation, hail, icing, and visibility.

2.2 Calibration Methodology

For the calibration of each model, an optimal threshold was selected (to the nearest 0.1 mm) to minimize a cost function for the region and season. We examined three cost functions: 1) unconditional bias, 2) Brier Score, and 3) the reliability component of the Brier Score (Murphy and Winkler, 1992; Brier, 1950).

For the first method, the unconditional bias was computed by differencing the average forecast probability and the verification field's average observed occurrence rate. For the other two cost functions we used the "CR" decomposition of the mean square error, as outlined by Murphy and Winkler (1992). Figure 2 shows the resulting thresholds (y-axis) for each ensemble, organized into columns of region and season, using the three methods. The average probability (unconditional bias) approach produces the smallest calibration correction, while the reliability approach produces the largest.

To combine models, we used the unconditional bias approach for threshold selection, as it was easy to implement and conceptually easy to explain as producing a probability forecast that, on average, matches the observed rate of occurrence. For the other two methods, performance can influence the calibration threshold selection, particularly at high probabilities, typically resulting in much higher thresholds than the purely unconditional bias approach.

2.3 Fusion Methodology

To examine the impact of the approach used to combine the calibrated probability field from two or more ensemble models, we examine the performance of three different weighting schemes: 1) Equal weights, 2) Optimal weights based on using Brier Skill Score (BSS), and 3) Optimal weights determined with multiple linear regression.

The combined probability was computed using a weighted average of each ensemble's calibrated probability. For the equal weights approach, each weight was set to 0.25. For weights determined from the BSS, which is given as

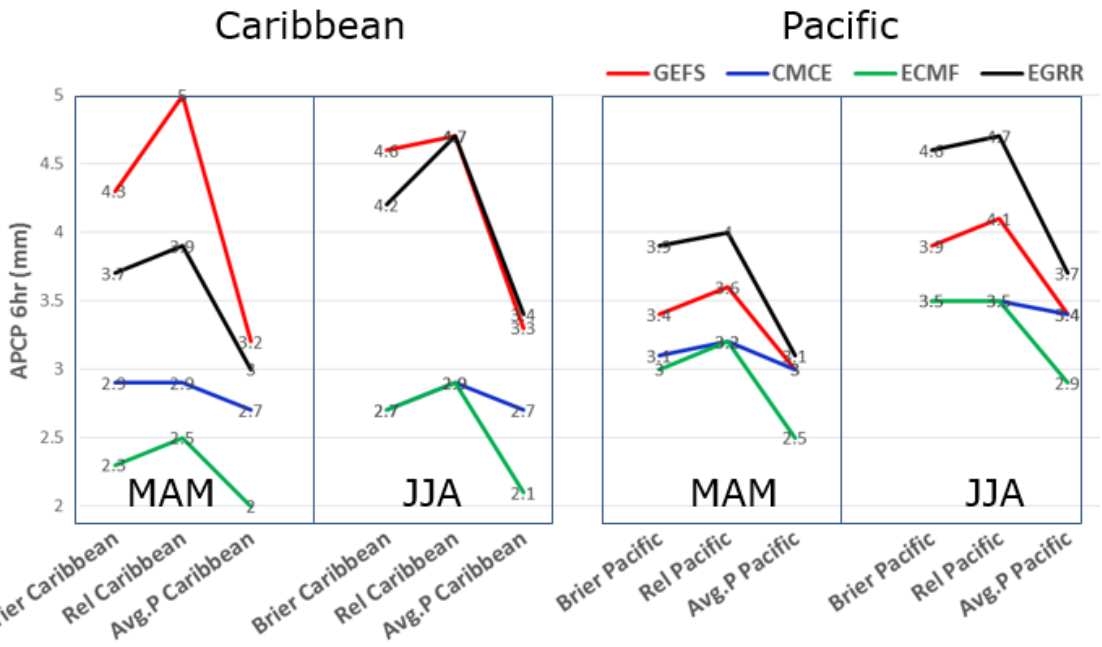

Figure 2. Resulting thresholds (APCP6hr in mm) for the two regions (Caribbean left; Pacific right) and two seasons (MAM, JJA). NCEP in Red; CMCE in blue; ECMF in green; and UKMO in black. For each sub-column, the first value is derived using the Brier Score, the 2nd value from the Reliability component alone of the Brier Score and the 3rd value is derived using the unconditional bias.

$$BSS = \frac{BrierScore - BrierScore_{REF}}{BrierScore_{REF}}, \qquad (1)$$

we used the NCEP model's calibrated probability as the reference (REF). Following Hamill and Juras (2006) scores were computed and compared only for identical observations, to avoid base rate differences influencing the metrics. Each model was given an equal starting weight of 0.25, then adjusted depending on the relative performance compared to NCEP. The weights, normalized across the four models, are shown in Table 1. Note that ECMF weights are highest, indicating a better Brier Score (at the selected threshold providing the best unconditional bias).

For the regression approach, a multiple linear regression fit was performed (e.g., see Exelis IDL "regress.pro") on the set of calibrated probabilities and the verification field to determine the coefficients. Table 2 shows the resulting weights after the coefficients were normalized.

Regression-derived weights tend to provide a wider range of weights compared to those obtained when optimizing the BSS. For the Pacific region the pattern of lowest to highest weight is similar to the BSS-derived values seen in Table 1; however, for the Caribbean, NCEP and CMCE reverse positions in the ordering and ECMF weighting is amplified, keeping it the highest.

Table 1. Normalized Brier Skill Score derived weights for each region and season, for the four models in the case study. Due to rounding in the table presentation, the sum of weights for each column doesn't always add-up to one.

Model	Pacific		Caribbean	
	MAM	**JJA**	**MAM**	**JJA**
NCEP	0.18	0.19	0.25	0.24
CMCE	0.28	0.30	0.26	0.26
ECMF	0.33	0.32	0.30	0.31
UKMO	0.22	0.19	0.19	0.18

Table 2. Derived weights as in Table 1, except using the regression approach.

Model	Pacific		Caribbean	
	MAM	**JJA**	**MAM**	**JJA**
NCEP	0.14	0.19	0.25	0.27
CMCE	0.27	0.28	0.22	0.19
ECMF	0.36	0.34	0.41	0.42
UKMO	0.23	0.18	0.13	0.12

3. RESULTS

Using the resulting weights with the underlying calibrated probabilities, performance metrics were computed. Namely, BSS compared to NCEP and the reliability component of the Brier Score, which is most indicative of calibration (Murphy and Winkler, 1992). Figure 3 summarizes the results over the two regions, seasons, and weight-derivation approaches.

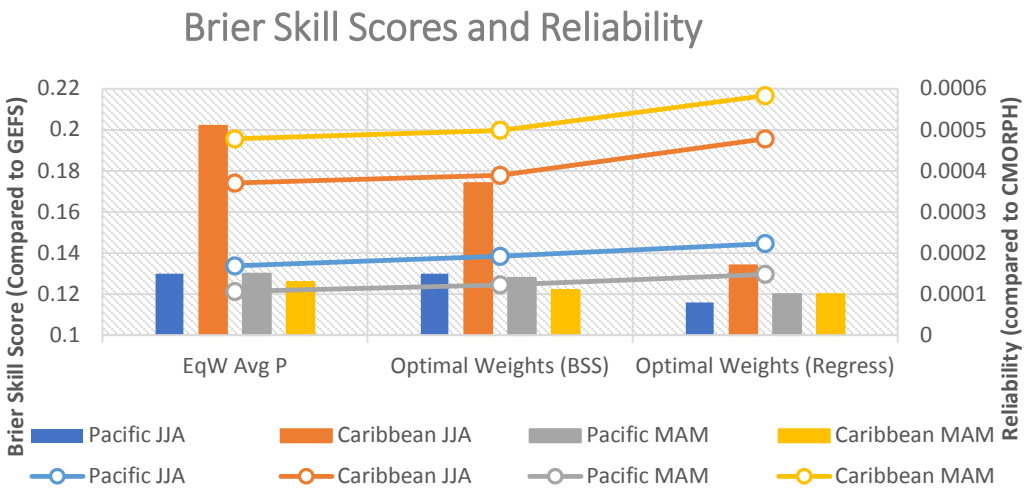

Figure 3. Resulting reliability and Brier Skill Scores (BSS) for the three weighting approaches (Equal Weights on the left, BSS selected in the middle, Regression on the right). The connected lines are BSS (higher is better) and the bars are reliability (lower is better). Each color represents a different season/region (blue = Pacific JJA; orange = Caribbean JJA; grey = Pacific MAM; yellow =Caribbean MAM).

There is some benefit of the BSS compared with equal weights as seen in both the increase in BSS (compared to GEFS) and decrease in the reliability score, while the regression approach offers some improvement over the BSS approach. While the regression approach appears most promising, it is more difficult to implement and more difficult to explain compared to the straightforward mean-squared error approach of the BSS. Perhaps more importantly, the BSS-derived weights can be determined independently by each potential international provider of weather information without requiring data from all other modeling centers.

To further explore the performance of the individual and combined probabilities, we also looked at the components of the Brier Score (reliability and resolution) for several combinations of equally weighted models and multiple calibration approaches. These results are summarized in Figure 4 for the Caribbean domain for JJA. Other periods and regions appeared to give qualitatively similar results although inter-comparison is made challenging by the underlying differences in the observed base rate (Hamill and Juras, 2006). The different base rates give offsets in the Brier Score due to the uncertainty component and also impact the resolution component. Using a single domain and season with the same verification field (forecast and observation pairs) makes the interpretation less problematic as differences in resolution and Brier Score are due to forecasts alone and not the observed frequency.

For the models shown, the actual calibration approach (reliability, Brier Score, bias) doesn't appear to significantly influence overall performance. As long as the threshold used to calibrate the model forecasts is "in the ballpark" the results are similar when combining models. Performance improves as probabilistic forecasts from skillful ensembles are included in the final forecast probability with the best performance being achieved when combining all four models.

4. SUMMARY

A methodology was presented on how to combine multiple ensemble forecasts to develop probabilistic, globally harmonized aviation weather hazard guidance products for transoceanic flight planning. Various options for the calibration and fusion of ensemble forecasts were discussed, with an initial focus on convective weather hazards. APCP6hr forecasts varied in comparison to CMORPH, while a fused product of all bias-corrected models provided the best performance. This is reasonably consistent with other TIGGE-related findings reported by Park (2008) and Hamill (2012). Even an equally weighted average provided good results, likely because the four models studied are mature and highly refined, so equal weighting is a good first-guess. As Hamill (2012) pointed out, examining a multi-model combination with less mature models may not provide the same level of combined forecast benefits. The BSS-based weighting approach used here could lend itself to those cases where individual ensemble's performance show more extreme spatial or temporal variation, and thus provide a scalable approach for combining ensembles.

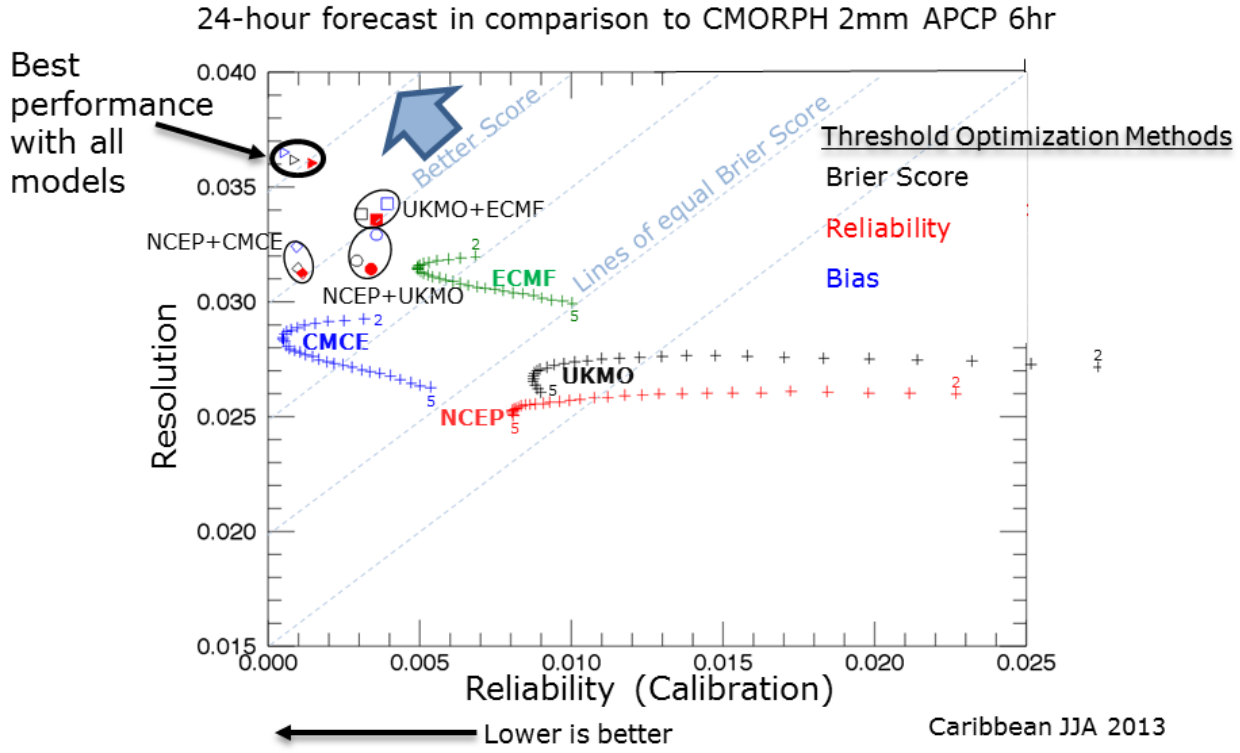

Figure 4. Brier Score components (reliability and resolution) for the Caribbean (JJA). Reliability is better toward the left of the chart, and resolution is better toward to the top. The + symbols are individual model results for thresholds of 2.0 mm to 5.0 mm, as labeled. The large black circles encapsulate combined ensemble performance. Best performance is achieved using all four model ensembles. Within each combined probability, results are shown for different calibration approaches (blue = avg. probability; red = reliability; black=Brier Score). The diamonds are NCEP+CMCE; the circles are NCEP+UKMO; the squares are UKMO+ECMF, and the triangles are all models combined. The diagonal dashed lines are isopleths of Brier Score (better score to the upper left).

The case study was limited to forecasting probability of significant precipitation (over a 6-hour period), an initial proxy field for convective hazards. We anticipate the overall approach could work for forecasts of other aviation hazards as well. However, further investigation is needed to assess how to treat regional variations in performance among the models and to assess how skill might change as probabilistic forecasts from additional ensembles of varying levels of skill are included in the harmonization process. Moreover, defining the "truth" for assessing the predictions against remains challenging due to sparse observations of relevant atmospheric quantities.

5. ACKNOWLEDGEMENTS

This research is in response to requirements and funding by the Federal Aviation Administration (FAA). The views expressed are those of the authors and do not necessarily represent the official policy or position of the FAA.

6. REFERENCES

Bougeault, P., and Co-authors, 2010: The THORPEX Interactive Grand Global Ensemble. *Bull. Amer. Meteor. Soc*., **91**, 1059 – 1072.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1 – 3.

Hamill T. M., and J. Juras, 2006: Measuring forecast skill: is it real skill or is it the varying climatology? Q. J. R. Meteorol. Soc., **132**, 2905 – 2923.

Hamill T. M., 2012: Verification of TIGGE Multimodel and ECMWF Reforecast-Calibrated Probabilistic Precipitation Forecasts over the Contiguous United States. *Mon. Wea. Rev.*, **140**, 2232 – 2252.

Joyce, R. J., J. E. Janowiak, P. A. Arkin, and P. Xie, 2004: CMORPH: A method that produces global precipitation estimates from passive microwave and infrared data at high spatial and temporal resolution. *J. Hydromet*., **5**, 487 – 503.

Murphy, A. H., and R. L. Winkler, 1992: Diagnostic verification of probability forecasts. *Int. J. Forecast.*, **7**, 435 – 455.

Park, Y.-Y., R. Buizza, and M. Leutbecher, 2008: TIGGE: Preliminary results on comparing and combining ensembles. Q.J.R. Meteorol. Soc., **134**: 2029 – 2050.