NOTES AND CORRESPONDENCE

Operational Evaluation of a Selective Consensus in the Western North Pacific Basin

CHARLES R. SAMPSON

Naval Research Laboratory, Monterey, California

JOHN A. KNAFF

Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, Colorado

EDWARD M. FUKADA

Joint Typhoon Warning Center, Pearl Harbor, Hawaii

(Manuscript received 9 February 2006, in final form 10 August 2006)

ABSTRACT

The Systematic Approach Forecast Aid (SAFA) has been in use at the Joint Typhoon Warning Center since the 2000 western North Pacific season. SAFA is a system designed for determination of erroneous 72-h track forecasts through identification of predefined error mechanisms associated with numerical weather prediction models. A metric for the process is a selective consensus in which model guidance suspected to have 72-h error greater than 300 n mi (1 n mi = 1.85 km) is first eliminated prior to calculating the average of the remaining model tracks. The resultant selective consensus should then provide improved forecasts over the nonselective consensus. In the 5 yr since its introduction into JTWC operations, forecasters have been unable to produce a selective consensus that provides consistent improved guidance over the nonselective consensus. Also, the rate at which forecasters exercised the selective consensus option dropped from approximately 45% of all forecasts in 2000 to 3% in 2004.

1. Introduction

The Systematic Approach Forecast Aid (SAFA) is a knowledge-based tropical cyclone track forecast system developed to assist the forecaster in the information management, visualization, and proactive investigation of error mechanisms associated with numerical weather prediction (NWP) models (Carr et al. 2001). Those NWP models are the U.S. Navy Operational Global Atmospheric Prediction System (Hogan and Rosmond 1991; Goerss and Jeffries 1994), the Geophysical Fluid Dynamics Laboratory Hurricane Prediction System (Kurihara et al. 1993, 1995, 1998; Rennick 1999), the Japan Meteorological Agency global and typhoon models (Kuma 1996), and the Met Office global model (Cullen 1993; Heming et al. 1995). Forecasts from these five NWP models are used to compute a simple nonselective consensus (an average of the forecast positions from the available NWP model forecasts through 72 h) named NCON. NCON is an extension of a consensus method that had been installed and used intermittently at JTWC since 1998. The original consensus method (Goerss 2000) formed a consensus of the two regional models for 0600 and 1800 UTC and then formed a consensus of the three global models for 0000 and 1200 UTC. NCON extended the work of Goerss (2000) in that it employed a method to relabel and extrapolate the NWP model forecast tracks so that they are available every 6 h (Elsberry and Carr 2000). The method used is similar to the interpolator developed and used in operations at the National Hurricane Center in the late 1990s (Goerss et al. 2004). According to Carr et al. (2001), a key metric for evaluating the value added of SAFA is a selective consensus (SCON) whereby one or more NWP model forecast tracks suspected of having a 72-h forecast position error greater than 300 n mi (1 n

Corresponding author address: Charles R. Sampson, NRL, 7 Grace Hopper Ave., Stop 2, Monterey, CA 93943-5502.
E-mail: sampson@nrlmry.navy.mil

DOI: 10.1175/WAF991.1

mi = 1.85 km) is eliminated prior to computing SCON. NCON is then used as a baseline for measuring SCON forecast improvement. The feasibility for this is shown in Elsberry and Carr (2000) where removing the poorest performer improved the consensus in almost all cases when the known error of the model removed was larger than 300 n mi. The 300 n mi limit is used because the SAFA conceptual models are based on years of evaluation using real NWP model track forecast errors larger than 300 n mi (Carr et al. 2001). Although a suspected 300 n mi forecast error is a necessary condition to eliminate an NWP model forecast, a large spread (greater than 250 n mi) and an error mechanism also must be present. Thus, an independent SCON is not necessarily computed for every case.

In a 1999 beta test on available 72-h forecast track guidance for JTWC western North Pacific storms 19–30, Carr et al. (2001) found that for 14 out of 31 cases the developers were able to improve the selective consensus by an average of 10% over the nonselective consensus. At the time, this was considered an underestimate of the improvement that could be achieved because forecast fields for two of the models were missing during the beta test. Based on the results achieved during the beta test, SAFA was installed for operational evaluation at the Joint Typhoon Warning Center (JTWC) in 2000 and evaluation continued through 2004. This paper presents the results of the operational evaluation and offers some comments on the utility of SAFA.

2. Training and operational test procedures

Many recent JTWC forecasters who attended the Naval Postgraduate School in Monterey, California, had some exposure to SAFA during their education. For the rest, SAFA training and testing software modules were included as part of the Typhoon Duty Officer qualification regimen. Although these forecasters did not obtain knowledge and experience equal to SAFA developers and other experienced researchers involved with the beta test, a concerted effort to train forecasters on SAFA was undertaken.

For the 2000 western North Pacific season, the JTWC forecast procedures were modified to include a mandatory SAFA analysis immediately after the objective aid (i.e., numerical model forecast) ingest and display. For the years 2000–04, the SCON and NCON 72-h forecast tracks were saved to the objective aid database on the Automated Tropical Cyclone Forecast System (ATCF; Sampson and Schrader 2000).

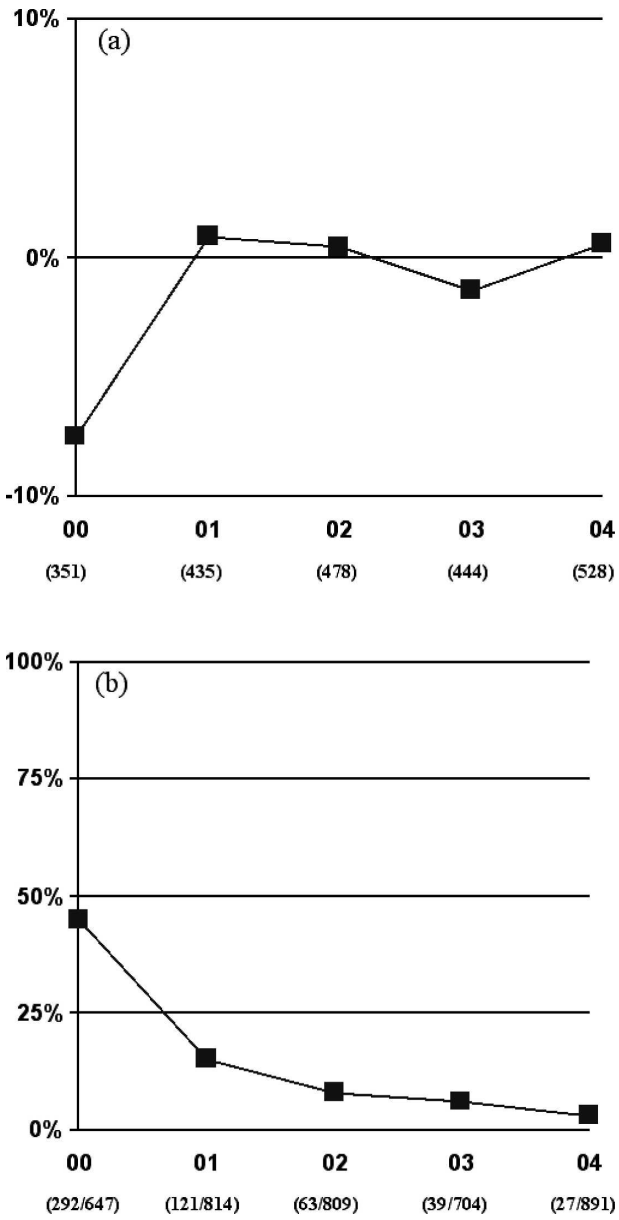

FIG. 1. Performance of the selective consensus for all SAFA analysis cases in the years 2000–04. (a) The 72-h forecast improvement (%) of selective consensus over nonselective consensus. The number of cases is shown in parentheses. (b) Percentage of total forecasts for which a selective consensus was produced. The total number of selective consensus forecasts and SAFA analyses, respectively, are shown in parentheses.

3. Results and conclusions

Statistics resulting from the 72-h forecasts are presented since that was the forecast hour on which the SCON methodology and application focused. The overall 72-h forecast results for the years 2000–04 are shown in Fig. 1. Analyses are created using all forecasts where

both the forecasts and best tracks are west of the international date line. JTWC defines the western North Pacific season as including all sequentially numbered storms assigned between 0000 UTC 1 January and 1800 UTC 31 December and that convention will be followed here. The 2002 data include forecasts for two central Pacific storms (storm numbers 2 and 3) that moved into the western North Pacific. For the purpose of fairly evaluating SAFA, eight SCON forecasts with gross errors (24-h forecast errors larger than 1000 n mi) were removed before the analysis: five for storm number 15 of year 2000 (1200 and 1800 UTC 9 August, 0000 and 1800 UTC 10 August, and 0600 UTC 16 August), one for storm number 10 (0600 UTC 20 July), and two for storm number 19 (1800 UTC 20 August and 0000 UTC 21 August). Gross errors such as these were not found in subsequent years. Even with the removal of these gross errors, a 7.5% degradation of SCON forecasts with respect to the NCON forecasts is immediately apparent in the 2000 season results. This degradation was partially attributed to overuse of SCON (i.e., creating SCON when it was not justified) and training issues (Jeffries and Fukada 2002). However, a postseason evaluation of the 2000 season results by developers showed that the SCON results would not have improved on NCON results even if the individual error mechanisms had been selected correctly (L. Carr 2001, personal communication). This led to an additional step in the SAFA decision process to account for situations with compensating error mechanisms since forecasters had focused only on one error mechanism during 2000. With compensating error mechanisms, suspected large errors of one group of aids might offset the errors of another group so that removal of any aids results in a selective consensus that underperforms the nonselective consensus.

The 2001 results were more successful than those of 2000, due in part to restricting the creation of SCON forecasts (Cantrell and Jeffries 2002). SCON forecasts were created for 14% of the attempts as compared to 45% in 2000, and an overall improvement of 0.9% was attained over the corresponding NCON forecasts. Likewise, the 2002 results show a slight improvement for the entire year, but the creation of SCON forecasts had dropped to 8% of all cases so the overall improvement is small. The 2003 and 2004 seasons also had decreased creation of SCON forecasts and consequently very little overall improvement or degradation. The decreased creation of the SCON forecasts was partly a result of model improvements. For 2001–04 the 72-h spread of consensus member forecast positions (consensus spread) exceeded 250 n mi in less than 15% of the forecasts, so there were fewer chances to form a SCON forecast as prescribed by Carr et al. (2001). The consensus spread restriction was relaxed in order to create more chances to form the SCON forecast, but this also may have degraded the results, as was the case in the beta test (Carr et al. 2001). Cantrell and Jeffries (2002) also found that in 2000 and 2001 it was difficult for JTWC to improve on NCON when it contained four or five model tracks. Based on their results, SCON forecasting was further restricted in JTWC operations. Other issues that may have affected the number of SCON forecast attempts through the period were training and forecast time constraints. It should also be noted that other NWP models became available to forecasters during the years 2001–04 that were not incorporated into SAFA. In this same period, objective consensus aids that were developed to include the new models were run operationally at JTWC. These new consensus aids were known to outperform NCON (Goerss et al. 2004; Sampson et al. 2006) in 2001. Over the period of this study, these new consensus forecasts gradually became JTWC forecasting baselines, so production of SCON became less relevant.

TABLE 1. Homogeneous comparison of selective consensus forecast errors and nonselective consensus forecast errors for the period 2000–04. The period 2001–04 is also evaluated because modifications were made to both the operational procedures and the identification of error mechanisms after the 2000 season.

Years	2000–04	2001–04
SCON mean improvement (%)	−6.8	2.4
Probability of model differences	0.94	0.67
Cases (adjusted for 30-h serial correlation)	159.4	103.8
Superior performance (%)	47	53
Total cases	300	148

A Student's *t* test performed with a confidence level of 95% shows that differences between NCON and SCON average errors for the year 2000 are significant, but differences for other years are not. Statistical significance calculation accounts for serial correlation within 30 h (von Storch and Zwiers 1999). The differences are generally small. The small differences in the 2001–04 results exist because SCON does not differ from NCON in the majority of the SAFA analyses. The SCON produced for a forecast is equal to NCON when the consensus spread meets the requirements for SCON creation, but the forecaster finds no error mechanism.

The cases for which the SCON forecast differs from the NCON forecast are gathered for further analysis. Doing so results in 300 cases for the period 2000–04 and 141 cases for the period 2001–04 (Table 1). Analysis for the period 2001–04 is presented separately, as there was

too frequent creation of SCON during the 2000 season (Jeffries and Fukada 2002). For the entire period 2000–04, the selective consensus is worse, but not quite at the 95% confidence level. Very little difference exists in the number of SCON forecasts that outperform the NCON forecasts (denoted as superior performance in Table 1). For the period 2001–04 the SCON forecasts outperform the corresponding NCON forecasts by 2.4%, but the results are not statistically significant (the probability of model differences is 0.67). Very little difference exists in the number of SCON forecasts that outperform the matching NCON forecasts in 2001–04.

To evaluate whether or not the SCON performance in 2001–04 was degraded by relaxing the requirement that the consensus spread be larger than the 250 n mi specified in Carr et al. (2001), statistical analysis for cases in which the consensus spread was larger than 250 n mi are analyzed. For the 30 cases that verified (1.5% of the total number of verifying SCON forecasts), SCON outperformed NCON by about 30%. The results are significant at the 95% level. It is worth noting that the other skillful NWP models and consensus aids available to JTWC forecasters may have influenced the forecasters while performing the SAFA analysis.

In summary, forecasters have been unable to produce a 72-h selective consensus that provides overall consistent improved guidance over a nonselective consensus. Even when the rate at which forecasters created SCON forecasts dropped from approximately 45% of all forecasts in 2000 to less than 15% in 2001–04, overall significant improvement over the NCON forecasts was not achieved. The number of opportunities to create SCON forecasts decreased through the evaluation period as NWP model forecasts improved. It is suspected that one reason it is difficult to select out "bad guidance" is shown in Goerss et al. (2004) where adding more guidance to a consensus improved the overall results. This implies that forecasters who mistakenly eliminate "good guidance" from their consensus would generally pay a penalty for doing so, especially in cases with small consensus spread (Elsberry and Carr 2000). This was a lesson learned in both the 1999 beta test and the 2000 season when the SCON was overproduced to the point of markedly degrading the consensus. Analysis of a subset of 30 SCON forecasts in the 2001–04 dataset with large (>250 n mi) consensus spreads confirms that the methodology may have been skillful, but limited in application since it only applies in approximately 1.5% of all SAFA analyses.

During the 2005 season, SCON production was dropped from the JTWC operations. This change was due in part to the lack of success and opportunity to form SCON forecasts noted in section 3, and the large amount of staff required for training and use of this process. JTWC decided to shift the analysis and forecast emphasis toward intensity and wind field determination. An objective and automated technique that employed the SCON methodology, which but used all available NWP models and extended to 120 h, might be appropriate for use in operations.

Acknowledgments. The authors would like to acknowledge Ann Schrader for her work with the ATCF and the entire staff at JTWC for their efforts with SAFA. The ensemble work of Jim Goerss is deeply appreciated, as are the comments by Mark DeMaria, John Cook, Ted Tsui, Simon Chang, Russ Elsberry, Jim Goerss, and two anonymous reviewers. The NOAA Office of Research and Applications (Grant NA05AANEG0221) provided funding for this research.

REFERENCES

Cantrell, C. E., and R. A. Jeffries, 2002: Analysis of the first operational-test of the systematic approach to tropical cyclone forecasting aid at the Joint Typhoon Warning Center during the 2000 and 2001 tropical cyclone season. Preprints, *25th Conf. on Hurricanes and Tropical Meteorology,* San Diego, CA, Amer. Meteor. Soc., 305–306

Carr, L. E., III, R. L. Elsberry, and J. E. Peak, 2001: Beta test of the systematic approach expert system prototype as a tropical cyclone forecasting aid. *Wea. Forecasting,* **16,** 355–368.

Cullen, M. J. P., 1993: The unified forecast/climate model. *Meteor. Mag.,* **122,** 81–122.

Elsberry, R. L., and L. E. Carr III, 2000: Consensus of dynamical tropical cyclone track forecasts—Errors versus spread. *Mon. Wea. Rev.,* **128,** 4131–4138.

Goerss, J. S., 2000: Tropical cyclone track forecasts using an ensemble of dynamical models. *Mon. Wea. Rev.,* **128,** 1187–1193.

——, and R. A. Jeffries, 1994: Assimilation of synthetic tropical cyclone observations into the Navy Operational Global Atmospheric Prediction System. *Wea. Forecasting,* **9,** 557–576.

——, C. R. Sampson, and J. M. Gross, 2004: A history of western North Pacific tropical cyclone track forecast skill. *Wea. Forecasting,* **19,** 633–638.

Heming, J. T., J. C. L. Chan, and A. M. Radford, 1995: A new scheme for the initialization of tropical cyclones in the UK Meteorological Office global model. *Meteor. Appl.,* **2,** 171–184.

Hogan, T. F., and T. E. Rosmond, 1991: The description of the Navy Operational Global Atmospheric Prediction System's spectral forecast model. *Mon. Wea. Rev.,* **119,** 1786–1815.

Kuma, K., 1996: NWP activities at Japan Meteorological Agency. Preprints, *11th Conf. on Numerical Weather Prediction,* Norfolk, VA, Amer. Meteor. Soc., J15–J16.

Kurihara, Y., M. A. Bender, and R. J. Ross, 1993: An initialization scheme of hurricane models by vortex specification. *Mon. Wea. Rev.,* **121,** 2030–2045.

——, ——, R. E. Tuleya, and R. J. Ross, 1995: Improvements in the GFDL hurricane prediction system. *Mon. Wea. Rev.,* **123,** 2791–2801.

——, R. E. Tuleya, and M. A. Bender, 1998: The GFDL hurricane prediction system and its performance in the 1995 hurricane season. *Mon. Wea. Rev.,* **126,** 1306–1322.

Jeffries, R. A., and E. J. Fukada, 2002: Consensus approach to tropical cyclone forecasting. *Proc. Fifth Int. Workshop on Tropical Cyclones,* Cairns, Australia, WMO. [Available online at http://www.aoml.noaa.gov/hrd/iwtc/index.html.]

Rennick, M. A., 1999: Performance of the Navy's tropical cyclone prediction model in the western North Pacific basin during 1996. *Wea. Forecasting,* **14,** 3–14.

Sampson, C. R., and A. J. Schrader, 2000: The Automated Tropical Cyclone Forecasting System (version 3.2). *Bull. Amer. Meteor. Soc.,* **81,** 1231–1240.

von Storch, H., and F. W. Zwiers, 1999: *Statistical Analysis in Climate Research*. Cambridge University Press, 484 pp.