Description and Skill Evaluation of Experimental Dynamical Seasonal Forecasts of Tropical Cyclone Activity at IRI

TECHNICAL REPORT 08-02

The International Research Institute for Climate and Society

Lamont Campus 61 Route 9W, Monell Building
Palisades, NY 10964-8000, USA
http://iri.columbia.edu

Photo Credit

NASA image created by Jesse Allen, Earth Observatory, using data provided by the MODIS Rapid Response team.

Hurricane Ioke at the time of this image had a well-defined round shape, clear spiral-arm structure, and a distinct but cloud-filled (or "closed") eye. The University of Hawaii's Tropical Storm Information Center reported that Hurricane Ioke had sustained winds of around 255 kilometers per hour (160 miles per hour) at the time this satellite image was acquired.

About the IRI

The IRI supports sustainable development by bringing the best science to bear on managing climate related risks in sectors such as agriculture, food security, water resources, and health. By providing practical advancements that enable better management of climate risks and opportunities in the present, the IRI is creating solutions that will increase adaptability to long term climate change. A member of the Earth Institute, the IRI was established as a cooperative agreement between the National Oceanic and Atmospheric Administration's Office of Global Programs and Columbia University.

Description and Skill Evaluation of Experimental Dynamical Seasonal Forecasts of Tropical Cyclone Activity at IRI

Suzana J. Camargo* and Anthony G. Barnston,
International Research Institute for Climate and Society,
The Earth Institute at Columbia University,
Lamont Campus, Palisades, NY

Abstract

The International Research Institute for Climate and Society has been issuing experimental seasonal tropical cyclone activity forecasts for several ocean basins since early 2003. In this paper we describe the method used to obtain these forecasts, and evaluate their performance. The forecasts are based on tropical cyclone-like features detected and tracked in a low-resolution climate model, namely ECHAM4.5. The simulation skill of the model using historical observed sea surface temperatures (SSTs) over several decades, as well as with SST anomalies persisted from the month ending at the forecast start time, is discussed. These simulation skills are compared with skills of purely statistically based hindcasts using as predictors observed SSTs preceding the forecast start time. For the recent 6-year period during which real-time forecasts have been made, the skill of the raw model output is compared with that of the subjectively modified probabilistic forecasts actually issued.

Despite variations from one basin to another, the hindcast skills of the dynamical and statistical forecast approaches are found, overall, to be approximately equivalent. The dynamical forecasts require statistical post-prossessing (calibration) to be competitive with, and in some circumstances superior to, the statistical models. Hence, during the recent period of real-time forecasts, the subjective forecasts are found to have resulted in probabilistic skill better than that of the raw model output, primarily because of the forecasters' elimination of the systematic bias of "overconfidence" in the model's forecasts. Prospects for the future improvement of dynamical tropical cyclone prediction are considered.

*Current address: Lamont-Doherty Earth Observatory, The Earth Institute at Columbia University, 61 Route 9W, PO Box 1000, Palisades, NY 10964-8000,
E-mail: suzana@ldeo.columbia.edu

1 Introduction

Tropical cyclones (TCs) are one of the most devastating type of natural disaster. Seasonal forecasts of TC activity could help the preparedness of coastal populations for an upcoming TC season and reduce economical and human losses.

Currently, many institutions issue operational seasonal TC forecasts for various regions. In most cases, these are statistical forecasts, such as the North Atlantic hurricane outlooks produced by NOAA[1], the seasonal typhoon activity forecasts of the City University of Hong Kong (Chan et al. 1998, 2001), the Atlantic hurricane forecasts of Colorado State University (Gray et al. 1993; Klotzbach 2007a,b), and Tropical Storm Risk (Saunders and Lea 2004). A review of TC seasonal forecasts is found in Camargo et al. (2007b), and the skills of some of them were discussed in Owens and Landsea (2003).

Since April 2003 the International Research Institute for Climate and Society (IRI) has been issuing experimental dynamical seasonal forecasts for 5 ocean basins [2]. In this paper, we describe how these forecasts are produced and discuss their skills when the atmospheric general circulation model (AGCM) is forced by predicted sea surface temperature (SST) in a two-tiered prediction system.

The possible use of dynamical climate models to forecast seasonal TC activity has been explored by various authors, e.g. Bengtsson et al. (1982). Although the low horizontal resolution ($2^\circ - 3^\circ$) of climate general circulation models of the early 2000s is not adequate to realistically reproduce the structure and behavior of individual cyclones, such models are capable of forecasting with some skill several aspects of the general level of TC activity over the course of a season (Bengtsson 2001; Camargo et al. 2005). Dynamical TC forecasts can serve specific applications, e.g. early warning for TC landfall activity over Mozambique (Vitart et al. 2003) or the Philippines (Lyon and Camargo 2008). The level of performance of dynamical TC forecasts depends on many factors, including the model used (Camargo et al. 2005), the model resolution (Bengtsson et al. 1995), and the inherent predictability of the large-scale circulation regimes (Vitart and Anderson 2001), including those related to the ENSO condition (Wu and Lau 1992; Vitart et al. 1999).

In addition to IRI's dynamically based experimental TC forecasts, such forecasts are also produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) (Vitart 2006), the UK Meteorological Office and the EUROpean Seasonal to Interannual Prediction (EUROSIP - super-ensemble of ECMWF, UK Met Office and Metéo France coupled models) (Vitart et al. 2007) . An important consideration is the dynamical design used to produce the forecasts. The European dynamical TC forecasts are produced using fully coupled atmosphere-ocean models (Vitart and Stockdale 2001; Vitart 2006). At IRI, a two-tiered (Bengtsson et al. 1993), multi-model (Rajagopalan et al. 2002; Robertson et al. 2004) procedure is used to produce temperature and precipitation forecasts once a SST forecast (or set of them) is first established (Mason et al. 1999; Goddard et al. 2003; Barnston et al. 2003, 2005). The IRI experimental TC forecasts use a subset of the IRI two-tier forecast system,

[1] `http://www.cpc.noaa.gov/products/outlooks/hurricane.shtml`

[2] `http://portal.iri.columbia.edu/forecasts`

in that only a single AGCM is used, compared with several AGCMs for surface climate. As described below, more than one SST forcing scenario is used.

TCs in low-resolution models have many characteristics comparable to those observed, but at much lower intensity and larger spatial scale (Bengtsson et al. 1995; Vitart et al. 1997). The climatology, structure and interannual variability of model TCs have been examined (Bengtsson et al. 1982, 1995; Vitart et al. 1997; Camargo and Sobel 2004). A successful aspect of this work has been that, over the course of a TC season in a statistical sense, the spatial and temporal distributions, as well as interannual anomalies of number and total energy content, of model TCs roughly follow those of observed TCs (Vitart et al. 1997; Camargo et al. 2005). There have been two general methods in which climate models are used to forecast TC activity. One method is to analyze large-scale variables known to affect TC activity (Ryan et al. 1992; Thorncroft and Pytharoulis 2001; Camargo et al. 2007c). Another approach, and the one used here, is to detect and track the cyclone-like structures in climate models (Manabe et al. 1970; Broccoli and Manabe 1990; Wu and Lau 1992), coupled ocean-atmosphere models (Matsuura et al. 2003; Vitart and Stockdale 2001), and regional climate models (Landman et al. 2005; Knutson et al. 2007). These methods have also been used in studies of possible changes in TC intensity due to global climate change using AGCMs (Bengtsson et al. 1996; Royer et al. 1998; Bengtsson et al. 2007a,b) and regional climate models (Walsh and Ryan 2000; Walsh et al. 2004).

In Section 2 we describe how the real-time seasonal tropical forecasts are produced at IRI. The model performance over a multi-decadal hindcast period and over the recent 6-year period of real-time forecasting is are discussed in Section 3. A comparison of the AGCM performance result with that of simple SST-based statistical forecasts is shown in Section 4. The conclusions are given in Section 5.

2 Description of the real-time forecasts

The IRI climate forecast system (Mason et al. (1999)) is two-tiered: SSTs are first forecasted, and then each of a set of atmospheric models is forced with several tropical SST forecast scenarios. Many ensemble members of atmospheric response are produced from each model forced with the SST scenarios. For the TC seasonal forecasts, just one atmospheric model is used: ECHAM4.5, which is run on a monthly basis. Six-hourly output data is used, as this fine temporal resolution makes possible detection of the needed TC characteristics. The ECHAM4.5 was developed at the Max-Planck Institute for Meteorology in Hamburg (Roeckner et al. 1996), and has been studied extensively for various aspects of seasonal TCs activity (Camargo and Zebiak 2002; Camargo and Sobel 2004; Camargo et al. 2005, 2007c).

The integrations of the ECHAM4.5 model are subject to differing tropical SST forcing scenarios (Table 1). In all scenarios, the extratropical SST forecasts consist simply of damped persistence of the anomalies from the previous month's observation (added to the forecast season's climatology), with an anomaly e-folding time of 3 months (Mason et al. 1999). In the tropics, multi-model, mainly dynamical SST forecasts are used for the Pacific, while statistical and dynamical forecasts are combined for the Indian and Atlantic Oceans. Statis-

Table 1: Tropical Pacific SST forecast types used in this study. The concurrent, observed SST (OSST) data, for non-forecast simulations (lead time less than zero), denoted as "S" in table is the Reynolds version 2(Reynolds et al. 2002); the real time persisted SST ($FSST_p$) and hindcast persisted SST ($HSST_p$) are undamped and then damped persisted anomalies initialized from anomalies observed the previous month. The evolving anomalous SST (FSST) was initially from the NCEP coupled model ($FSST_1$)(Ji et al. 1998) and more recently was the mean of 3 models ($FSST_2$): the CFS(Saha et al. 2006), the LDEO-5 (here called LDEO)(Chen et al. 2004) and the statistical constructed analogue (CA)(van den Dool 1994, 2007) models. Still later, multiple predicted SSTs were used: first as the above three models separately ($FSST_3$), and most recently as the 3-model mean, plus and minus a perturbation field representing the dominant principal component (PC) of the three models' errors ($FSST_4$).

SST type	Period	Ensembles	Lead	Pacific SSTs
OSST	01/1950 - 12/2005	24	S	Reynolds
OSSTr	01/1970 - 12/2005	24	S	Reynolds
$HSST_p$	01/1968 - 05/2003	12	4	persisted
$FSST_p$	08/2001 - 08/2007	24	4	persisted
$FSST_1$	08/2001 - 05/2004	24	6	NCEP
$FSST_2$	06/2004 - 01/2007	24	6	Mean: CFS, LDEO, CA
$FSST_3$	06/2004 - 04/2007	8,8,8	6	CFS, LDEO, CA
$FSST_4$	05/2007 - 08/2007	8,8,8	6	PC: CFS, LDEO, CA

tical forecasts play the greatest role in the tropical Atlantic. The models contributing to the tropical SST forecasts, particularly for the Pacific, have changed during our study period as forecast producing centers have introduced newer, more advanced prediction systems. In the non-Pacific tropical basins during seasons having near-zero apparent SST forecast predictive skill, damped persisted SST anomalies are used, but at a lower damping rate than that used in the extratropics. (No damping occurs in the first 3 months, followed by linear damping that reaches zero by month 8.) However, for seasons in which SST predictive skill is found beyond that of damped persistence, CCA models are used in the Indian Ocean (Mason et al. 1999) and tropical Atlantic Ocean (Repelli and Nobre 2004).

Globally undamped anomalous SST persisted from the previous month, applied to the climatology of the months being forecast, is used as an additional SST forcing scenario (called $FSST_p$). In this case the 24 ensemble members of ECHAM4.5 are integrated using persisted SST anomaly out to 5 months beyond the previous month. For example, for a mid-January forecast, the model is forced from January - May using undamped persisted SST anomalies from December globally[3].

[3] The $FSST_p$ runs have been produced in real time from August 2001 to the present and are available also in hindcast mode ($HSST_p$) over the period January 1968 to May 2003, with 12 ensemble members (Table 1).

In the case of the non-persisted, evolving forecasted SST anomalies, the AGCM is run out to 7 months beyond the previous month's observed SST (e.g., for a mid-January forecast, observed SST exists for December, and the model is forced from January to July with evolving SST predictions). Four versions of the forecasted SST anomalies have been used since 2001, denoted by FSST_v, v indicating the version number, from 1 to 4. The FSST_v forecasts initially consisted of two versions of a single deemed best estimated forecast SST scenario for tropical Pacific SST ($\mathrm{FSST}_{1,2}$) through May 2004, after which time multiple SST scenarios began being used ($\mathrm{FSST}_{3,4}$). The single-forecast $\mathrm{FSST}_{1,2}$ fields from August 2001 to March 2007, with 24 ensemble members (Table 1) are available on line at the IRI website.[4].

From June 2004 until April 2007 the IRI used the ensemble mean SST scenarios predicted by each of the three forecast models footnoted above for the tropical Pacific SST individually, without averaging them, in separate runs of the AGCMs (FSST_3). This design was believed to better represent the uncertainty expressed by the spread of the ensemble mean forecast among the three models, whose suggested ENSO states matter critically to the TC forecast responses in most of the ocean basins. The FSST_3 runs thus consist of 8 ensemble members apiece for the tropical Pacific ensemble mean forecast scenario predicted by each of the three respective above-mentioned models.

The idea of forcing the AGCM with multiple SST forcing scenarios was refined further starting in May 2007, upon noting that on some occasions the ensemble mean forecasts of the three models agreed closely with one another, while on other occasions they differed wildly. The degree of disagreement was not believed to be more than weakly related to forecast uncertainty, as suggested in several studies (e.g. (Kharin and Zwiers 2002; Tippett et al. 2007)). To ensure more approximately comparable scenario differences from one year to the next for the same forecast start time and lead time, the three scenarios were derived based on the historical error of the 3-way superensemble mean of the models, for hindcast verifications relative to the observed SSTs over the global tropics. The tropical SST is used to ensure that uncertainty of the known impacts of tropical SST on atmospheric responses is captured. The structures of the scenarios are found applying principal components analysis (PCA) to the historical record of the error of the multi-model mean SST forecast. The three scenarios used are (1) the 3-way multi-model ensemble mean SST forecast itself, and that forecast (2) plus, and (3) minus, the first PC of the historical error. For this set of forecast SST scenarios (FSST_4), the perturbations to the mean vary by start month and SST forecast lead time, but not by year.

Because use of the differing versions of the FSST forcing design was limited to only a few years apiece, we combine the tropical cyclone forecast skill analyses using FSST_p, FSST_1, FSST_2, FSST_3 and FSST_4 into a single real-time forecast sample based on some type of

[4] From August 2001 to May 2004 the tropical Pacific SST for the FSST forecast was based solely on the NCEP-MRF9 dynamical model (Ji et al. 1998) (FSST_1); from June 2004 to February 2007, the forecast was assembed using the average of 3 SST model predictions for the tropical Pacific: (1) the NCEP coupled forecast system (NCEP-CFS) (Saha et al. 2006), (2) the Lamont-Doherty Earth Observatory intermediate model version 5 (LDEO-5) (Chen et al. 2004) (an improved version of the Cane-Zebiak model Cane and Zebiak (1985); Cane et al. (1986); Zebiak and Cane (1987)), and (3) the statistical constructed analog (CA) model (van den Dool 1994, 2007) (FSST_2)

forecast SST, generalized as FSST in the skill results shown below.

The ECHAM4.5 was also forced with the actually observed SSTs (OSST) (Reynolds et al. 2002) prescribed during the period 1950 to the present. These AMIP-type runs provide estimates of the upper limit of skill of the model in forecasting TC activity, discussed in previous studies (Camargo and Zebiak 2002; Camargo and Sobel 2004; Camargo et al. 2005, 2007c). The skills presented below are broken out into three SST forcing types: (1) FSST (for real-time forecasts), (2) HSST_p (long-term hindcast anomally-persisted SST), and (3) OSST (long-term observed SST for AMIP-type AGCM simulations).

For any type of SST forcing, we analyze the output of the AGCM for TC activity. To define and track TCs in the models, we used objective algorithms (Camargo and Zebiak 2002) based in large part on prior studies (Vitart et al. 1997; Bengtsson et al. 1995). The algorithm has two parts: detection and tracking. In the detection part, storms that meet environmental and duration criteria are identified. A model TC is identified when chosen dynamical and thermodynamical variables exceed thresholds calibrated to the observed tropical storm climatology. Most studies (Bengtsson et al. 1982; Vitart et al. 1997) use a single set of threshold criteria globally. However, to take into account model biases and deficiencies, we use basin- and model-dependent threshold criteria, based on analyses of the correspondence between the model and observed climatologies (Camargo and Zebiak 2002). Thus, we use a threshold exclusive to ECHAM4.5. Once detected, the TC tracks are obtained from the vorticity centroid, defining the center of the TC, using relaxed criteria appropriate for the weak model storms. The detection and tracking algorithms have been applied to regional climate models (Landman et al. 2005; Camargo et al. 2007a) and to multiple AGCMs (Camargo and Zebiak 2002; Camargo et al. 2005).

Following detection and tracking, we count the number of TCs (NTC) and compute the model accumulated cyclone energy (ACE) index (Bell et al. 2000) over a TC season. ACE is defined as the sum of the squares of the wind speeds in the TCs active in the model at each 6-hour interval. For the observed ACE, only TCs of tropical storm intensity or greater are included.

The model ACE and NTC results are then corrected for bias, based on the historical model and observed distributions of NTC and ACE over the 1971-2000 period, on a per-basin basis. Corrections yield matching values in a percentile reference frame (i.e., a correspondence is achieved non-parametrically). Using 1971-2000 as the climatological base period, tercile boundaries for model and observed NTC and ACE are then defined, since the forecasts are probabilistic with respect to tercile-based categories of the climatology (below, near, and above normal).

For each of the SST forcing designs, we count the number of ensemble members having their NTC and ACE in a given ocean basin in the below-normal, normal and above-normal categories, and divide by the total number of ensembles. These constitute the "raw", objective probability forecasts. In a final stage of forecast production, the IRI forecasters examine and discuss these objective forecasts and develop subjective final forecasts that are issued on the IRI website. The most typical difference between the raw and the subjective forecasts is that the latter have weaker probabilistic deviations from climatology, given the knowledge

that the models are usually too "confident". The overconfidence of the model may be associated with too narrow an ensemble spread, too strong a model signal (deviation of ensemble mean from climatology), or both of these. The subjective modification is intended to increase the probabilistic reliability of the predictions. Another consideration in the subjective modification is the degree of agreement among the forecasts, in which less agreement would suggest greater uncertainty and thus more caution with respect to the amount of deviation from the climatological probabilities.

The raw objective forecasts are available since August 2001. The first subjective forecast for the western North Pacific basin was produced in real-time in April 2003. However, subjective hindcasts were also produced for August 2001 through April 2003 without knowledge of the observed result, making for 6 years of experimental forecasts.

Table 2: Ocean basins in which IRI experimental TC forecasts are issued: Eastern North Pacific (ENP), Western North Pacific (WNP), North Atlantic (ATL), Australia (AUS) and South Pacific (SP). Date of the first issued forecast; seasons for which TC forecasts are issued (JJAS: June to September, ASO: August - October, JASO: July to October, ASO: August to October, JFM: January to March, DJFM: December to March); months in which the forecasts are issued; and variables forecasted — NTC (number of TCs), ACE (accumulated cyclone energy).

Basin	First Forecast	Season	Months Forecasts Are Issued	Variables
ENP	Mar. 2004	JJAS	Mar., Apr., May, Jun.	NTC, ACE
WNP	Apr. 2003	JASO	Apr., May, Jun., Jul.	NTC, ACE
ATL	Jun. 2003	ASO	Apr., May, Jun., Jul., Aug.	NTC, ACE
AUS	Sep. 2003	JFM	Sep., Oct., Nov., Dec., Jan.	NTC
SP	Sep. 2003	DJFM	Sep., Oct., Nov., Dec.	NTC

For each ocean basin, forecasts are produced only for the peak TC season, from certain initial months prior to that season (Table 2), and updated monthly until the first month of the peak season[5]. The lead time of this latest forecast is defined as being zero, and the lead times of earlier forecasts are defined by the number of months earlier that they are issued.

The basins in which forecasts are issued are shown in Fig. 1, and the numbers of years available for each SST scenario and basin are indicated in Table 3. In the southern hemisphere (South Pacific and Australian regions), only forecasts for NTC are produced, while in the northern hemisphere basins both NTC and ACE forecasts are issued. ACE is omitted for the southern hemisphere because ACE is more sensitive to data quality than NTC, and the observed TC data from the southern hemisphere is known to be of somewhat questionable quality, particularly in the earlier half of the study period (e.g. Buckley et al. (2003)).

Forecasts for the South Pacific are a special case in terms of bias correction. The TC

[5] The data available for the forecast released during the first month of the TC peak season cover only through the end of the previous month.

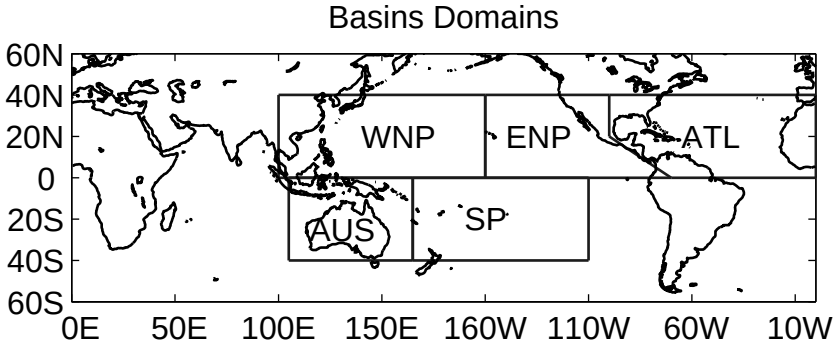

Figure 1: Definition of the ocean basin domains used in this study: Australian (AUS), $(105°E-165°E)$; South Pacific (SP), $165°E-110W$; western North Pacific (WNP), $100°E-160°W$, eastern North Pacific (ENP), $160°W-100°W$; and Atlantic (ATL), $100°W-0$. All latitude boundaries are along the equator and $40°N$ or $40°S$. Note the unique boundary paralleling Central America for ENP and ATL basins.

Table 3: Number of years for each lead and SST type. S denotes simulations, with a negative lead time.

SST Type	ENP					WNP					ATL					
Lead	3	2	1	0	S	3	2	1	0	S	4	3	2	1	0	S
FSST	6	6	6	6	—	6	6	6	6	—	6	6	6	6	7	—
$HSST_p$	—	—	35	35	—	—	—	35	35	—	—	—	35	35	35	—
OSST	—	—	—	—	56	—	—	—	—	56	—	—	—	—	—	56
OSSTr	—	—	—	—	36	—	—	—	—	36	—	—	—	—	—	36

SST Type	AUS						SP				
Lead	4	3	2	1	0	S	3	2	1	0	S
FSST	6	6	6	6	6	—	6	6	6	6	—
$HSST_p$	—	—	34	34	35	—	—	—	34	34	—
OSST	—	—	—	—	—	56	—	—	—	—	55
OSSTr	—	—	—	—	—	36	—	—	—	—	35

forecast season is DJFM, but we use the model output for NDJF to forecast DJFM, including the bias correction. It was found that hindcast skill levels are appreciably higher with this one-month offset, which we consider to be a temporal aspect of the bias correction.

The observed TC data used to correct historical model biases and for verification of the model forecasts is the best-track data from the National Hurricane Center (Atlantic and eastern North Pacific[6]) and the Joint Typhoon Warning Center (western North Pacific and southern hemisphere [7]).

3 Performance in hindcasts and real-time forecasts

NTC or ACE historical simulation and real-time predictive skill results are computed for each ocean basin for their respective peak TC seasons. Both deterministic and probabilistic skills are examined.

3.1 Deterministic skills

Temporal anomaly correlation skills are shown in Table 4 for NTC by lead time, for each type of SST forcing, and likewise for ACE in Table 5. The simulation skills are shown both for the full period of 1950-2005 and for 1970-2005, during which the TC data are known to be of higher quality, particularly for the southern hemisphere basins. The correlations for the real-time predictions are uncentered[8]. Simulation skills (OSST) are seen to be statistically significant levels for most of the ocean basins. Skills for the 1970-2005 period (OSST_r in Table 4) tend to exceed those for 1950-2005, due both to better average data quality and the greater ENSO variability following 1970. Consistent with Camargo et al. (2005), highest skills occur in the Atlantic basin with correlations of roughly 0.50, with more modest skill levels in the other basins. Skills for zero-lead forecasts using SST anomalies persisted from those of the most recent month (HSST_p, lead 0), as expected, are usually lower than those of observed simultaneous SSTs. For the three northern hemisphere basins, for which both NTC and ACE are simulated, simulation skills are higher for ACE than for NTC, as noted also in Camargo et al. (2005). This may be related to the continuous nature of ACE as opposed to the discrete, more nonparametric, character of NTC.

A reference forecast more difficult to beat than a random or a climatology forecast is that of simple persistence of observed TC observation from the previous year. The correlation score for such a reference forecast is just the 1-year autocorrelation coefficient over the 1971-2005 base period, and is shown at the bottom of Tables 4 and 5 as "Pers". The persistence

[6] `http://www.nhc.noaa.gov`

[7] `https://metocph.nmci.navy.mil/jtwc.php`

[8] In computing the correlation skill for forecasts for much shorter periods than the climatological base period, the sub-period means are not removed, and are not used for computing the standard deviation terms. Instead, the longer base period means are used. This is done so that, for example, if in the sub-period the forecasts and observations have small-amplitude out-of-phase variations but both are generally on the same side of the longer period mean, a positive correlation would result, and we believe justifiably.

Table 4: Correlations ($\times 10^2$) with observations for NTC, per basin, by lead time and SST forecast scenario. S denotes simulations, whose lead time is negative. "Pers" denotes one year simple persistence, with a lead potentially longer than 4 months, but shown in the column of the longest lead. Statistically significant skills are shown in bold.

SST Type	Eastern Pacific					Western Pacific					Atlantic					
Lead	3	2	1	0	S	3	2	1	0	S	4	3	2	1	0	S
FSST	53	58	58	46	—	-84	-82	-68	2	—	13	48	65	35	11	—
$HSST_p$	—	—	32	**50**	—	—	—	11	7	—	—	—	7	**33**	**36**	—
OSST	—	—	—	—	**32**	—	—	—	—	21	—	—	—	—	—	**40**
OSSTr	—	—	—	—	**37**	—	—	—	—	28	—	—	—	—	—	**55**
Pers	9	—	—	—	—	**33**	—	—	—	—	14	—	—	—	—	—

SST Type	Australia						South Pacific				
Lead	4	3	2	1	0	S	3	2	1	0	S
FSST	52	40	7	-1	21	—	44	72	63	34	—
$HSST_p$	—	—	-4	26	16	—	—	—	**42**	**42**	—
OSST	—	—	—	—	—	**38**	—	—	—	—	**43**
OSSTr	—	—	—	—	—	**40**	—	—	—	—	**42**
Pers	11	—	—	—	—	—	-26	—	—	—	—

Table 5: Correlations ($\times 10^2$) with observations for ACE, per basin, by lead time and SST forecast scenario. S denotes simulations, whose lead time is negative. "Pers" denotes one year simple persistence, with a lead potentially longer than 4 months, but shown in the column of the longest lead. Statistically significant skills are shown in bold.

SST Type	Eastern Pacific					Western Pacific					Atlantic					
Lead	3	2	1	0	S	3	2	1	0	S	4	3	2	1	0	S
FSST	-16	47	48	62	—	69	2	21	36	—	-4	25	17	61	14	—
$HSST_p$	—	—	20	23	—	—	—	6	23	—	—	—	18	**43**	**37**	—
OSST	—	—	—	—	7	—	—	—	—	26	—	—	—	—	—	**57**
OSSTr	—	—	—	—	**45**	—	—	—	—	**33**	—	—	—	—	—	**60**
Pers	21	—	—	—	—	20	—	—	—	—	**35**	—	—	—	—	—-

correlation scores are usually lower than those of the AGCM's forecast using observed or persisted SST, with the one exception of the NTC forecasts in Northwestern Pacific.

Real-time predictive verification skills (FSST in Tables 4 and 5) over the basins not only have lower expected values than those using simultaneous observed SST due to the imperfection of the predicted SST forcing, but also much greater sampling errors given only 6 to 7 cases per lead time per basin (Table 3). These skills range from near or below zero for the western North Pacific NTC, to approximately 0.5 for the three shortest leads for the eastern North Pacific ACE. For all basins collectively and for NTC and ACE together, the skills approximate those of HSST_p, individual differences likely due foremost to sampling variability. Consistent with the small sample problem, the correlations for FSST for all of the basin-lead time combinations are statistically non-significant, as nearly 0.8 is required for significance.

A look at the possible impact of differing SST forcing scenarios and lead times on the real-time forecast skills is more meaningful when results for all oceans basins are combined, lessening the sampling problem. Basin-combined skill results by lead time and SST forcing type are shown in Table 6 for NTC and ACE. The middle panel shows NTC results for northern hemisphere basins only, allowing a direct comparison between NTC and ACE. Results show higher skills for forecasts of ACE than NTC, and only a very weak tendency for decreasing skill with increasing lead time. This is summarized still further in the bottom row of the table, showing results for NTC and ACE combined.

Skills were evaluated using additional deterministic verification measures: the Spearman rank correlation, the Heidke skill score, and the mean squared error skill score (MSESS). Table 7 provides an example of the four scores together, for ACE in the Northwestern Pacific Basin. The rank correlation and Heidke skill scores are roughly consistent with the correlation skill, allowing for expected scaling difference where the Heidke is roughly one-half of the correlation (Barnston 1992). The MSESS, however, which uses the 1971-2000 climatology as the zero-skill reference forecast, is comparatively unfavorable: some of the cases having positive correlation and Heidke skills have negative MSESS results. This outcome is attributable to a marked tendency of the model forecasts toward too great a departure from climatological forecasts, given the degree of inherent uncertainty and thus the relatively modest level of true predictability. Such "overconfidence" in the model forecasts, which can be adjusted for statistically, will be discussed in more detail below in the context of probabilistic verification, where a detrimental effect on scores comparable to that seen in MSESS will become apparent.

3.2 Probabilistic skills

The TC forecasts were verified probabilistically using the ranked probability skill score (RPSS), likelihood skill score, and, for the real-time forecasts, the relative operating characteristics (ROC) score. All of the above measures were computed with respect to the tercile-based categories with 1971-2000 as the base period.

RPSS (Epstein 1969; Goddard et al. 2003) measures the sum of squared errors between categorical forecast probabilities and the observed categorical probabilities, cumulative over

Table 6: Correlations ($\times 10^2$) for all basins combined, by lead time and SST forecast scenario. S denotes simulations, whose lead time is negative. Statistically significant skills are shown in bold. The sample size is doubled for significance evaluations for all basins combined, and increased by 60% for the northern hemisphere basins combined, relative to the single basin sample.

Lead	4	3	2	1	0	S
SST Type	NTC all basins					
FSST	38	18	35	28	27	—
$HSST_p$	—	—	1	**25**	**27**	—
OSST	—	—	—	—	—	**25**
OSSTr	—	—	—	—	—	**46**
SST Type	NTC northern hemisphere					
FSST	13	-25	17	-3	27	—
$HSST_p$	—	—	7	23	**28**	—
OSST	—	—	—	—	—	**31**
OSSTr	—	—	—	—	—	**42**
SST Type	ACE northern hemisphere					
FSST	-4	41	29	46	41	—
$HSST_p$	—	—	17	**28**	**29**	—
OSST	—	—	—	—	—	**39**
OSSTr	—	—	—	—	—	**48**
SST Type	NTC & ACE all basins					
FSST	31	29	33	36	33	—

categories, relative to a reference (or standard baseline) forecast. The observed probabilities are 1 for the observed category and 0 for the other two categories, and the baseline forecast used here is the climatology forecast of a 1/3 probability for each category.

The likelihood score, related to the concept of maximum likelihood estimation (Aldrich 1997), is based on the nth root of the product of the probabilities given to the tercile category that was indeed observed, spanning temporally over all n forecasts. This "geometric average" probability is then compared with the same calculation done for a reference forecast using climatological probabilities of 1/3, to produce a likelihood skill score. Perpetual climatological probabilities would produce a likelihood skill score of zero, and forecasts of 100% for the correct tercile would produce a score of unity[9].

Probabilistic verification using the RPSS and likelihood scores are shown for NTC and

[9] Because the probabilities for the "raw" model predictions are based on the number of ensemble members in a given tercile category divided by the total number of ensembles, and occasionally this number is zero for the observed category, a likelihood score of 0 is produced by just one such occurrence. To circumvent this, zero probabilities for the correct category are set to 0.01 when computing the score, which would severely penalize the score but not "destroy" it.

Table 7: Comparison of four skill measures ($\times 10^2$) for real-time forecast SST_s (FSST), multidecadal hindcasts using persisted SST forcing ($HSST_p$) and simulations using observed SST forcing (OSST and $OSST_r$), for ACE in the Northwest Pacific basin.

Western North Pacific ACE		Lead		
SST Type	Score	1	0	S
FSST	Correl	11	34	—
FSST	Rank Cor	**44**	26	—
FSST	Heidke	3	7	—
FSST	MSESS	-63	-37	—
HSSTp	Correl	6	23	—
HSSTp	Rank Cor	-5	**34**	—
HSSTp	Heidke	1	12	—
HSSTp	MSESS	-61	-27	—
OSST	Correl	—	—	26
OSST	Rank Cor	—	—	27
OSST	Heidke	—	—	8
OSST	MSESS	—	—	4
OSSTr	Correl	—	—	33
OSSTr	Rank Cor	—	—	**39**
OSSTr	Heidke	—	—	11
OSSTr	MSESS	—	—	10

ACE in Tables 8 - 9 and Tables 10 - 11, respectively, for both the multi-decadal simulations and hindcasts (OSST and $HSST_p$), and the real-time forecasts forced by the multiple SST prediction scenarios (FSST). These skills are mainly near or below zero. This poor result can be attributed to the lack of probabilistic reliability of the ECHAM4.5 ensemble-based TC predictions as is seen in many predictions made by individual AGCMs—not just for TC activity but for most climate variables (Anderson 1996; Barnston et al. 2003; Wilks 2006). Climate predictions by AGCMs have model-specific systematic biases, and their uncorrected probabilities tend to deviate too strongly from climatological probabilities due to too small an ensemble spread and/or too large a mean shift from climatology. This problem leads to comparably poor probability forecasts, despite positive correlation skills for the ensemble means of the same forecast sets. Positive correlations, but negative probabilistic verification is symptomatic of poorly calibrated probability forecasts—a condition that can be remedied using objective statistical correction precedures.

Probabilistic persistence may be a more competitive simple reference forecast than forecasts of climatological probabilities. Based on the weak but generally positive year-to-year autocorrelations shown in Tables 4 and 5, we designed the persistence probabilistic forecasts to be 0.4 for the tercile-based category observed the previous year, and 0.3 for the

Table 8: Ranked probability skill scores ($\times 10^2$) for NTC, per basin, by lead time and SST forecast scenario. S denotes simulations, whose lead time is negative. "Pers" denotes one year weak probabilistic persistence (see text), with a lead potentially longer than 4 months, but shown in the column of the longest lead. Statistically significant skills are shown in bold.

SST Type	Eastern Pacific					Western Pacific					Atlantic					
Lead	3	2	1	0	S	3	2	1	0	S	4	3	2	1	0	S
FSST	-12	8	7	-60	—	-76	-52	-63	-34	—	-51	-8	3	-13	-61	—
$HSST_p$	—	—	-92	-80	—	—	—	-35	-36	—	—	—	-107	-70	-68	—
OSST	—	—	—	—	-18	—	—	—	—	-5	—	—	—	—	—	3
OSSTr	—	—	—	—	-24	—	—	—	—	3	—	—	—	—	—	-21
Issued	13	13	9	8	—	8	0	1	15	—	-5	6	11	15	-4	—
Pers	-2	—	—	—	—	-2	—	—	—	—	-2	—	—	—	—	—

SST Type	Australia						South Pacific				
Lead	4	3	2	1	0	S	3	2	1	0	S
FSST	-24	-36	-15	-74	-68	—	1	-27	-38	-27	—
$HSST_p$	—	—	-40	-34	-51	—	—	—	-28	-42	—
OSST	—	—	—	—	—	-32	—	—	—	—	-93
OSSTr	—	—	—	—	—	-21	—	—	—	—	-88
Issued	5	5	15	2	5	—	10	15	10	3	—
Pers	1	—	—	—	—	—	-1	—	—	—	—

Table 9: Ranked probability skill scores ($\times 10^2$) for ACE, per basin, by lead time and SST forecast scenario. S denotes simulations, whose lead time is negative. "Pers" denotes one year weak probability persistence (see text), with a lead potentially longer than 4 months, but shown in the column of the longest lead. Statistically significant skills are shown in bold.

SST Type	Eastern Pacific					Western Pacific					Atlantic					
Lead	3	2	1	0	S	3	2	1	0	S	4	3	2	1	0	S
FSST	-30	9	9	-3	—	26	-8	-1	-11	—	-45	-6	-16	-6	-31	—
$HSST_p$	—	—	-35	-24	—	—	—	-40	-9	—	—	—	-33	2	-6	—
OSST	—	—	—	—	-26	—	—	—	—	**6**	—	—	—	—	—	**11**
OSSTr	—	—	—	—	-7	—	—	—	—	**13**	—	—	—	—	—	**16**
Issued	2	0	6	10	—	15	5	5	4	—	-7	10	12	15	6	—
Pers	5	—	—	—	—	1	—	—	—	—	2	—	—	—	—	—

Table 10: Likelihood skill scores ($\times 10^2$) for NTC, per basin, by lead time and SST forecast scenario. S denotes simulations, whose lead time is negative.

SST Type	Eastern Pacific					Western Pacific					Atlantic					
Lead	3	2	1	0	S	3	2	1	0	S	4	3	2	1	0	S
FSST	0	11	15	-14	—	-23	-25	-21	-3	—	-27	-12	0	-7	-22	—
$HSST_p$	—	—	-29	-26	—	—	—	-17	-17	—	—	—	-38	-29	-28	—
OSST	—	—	—	—	-2	—	—	—	—	-2	—	—	—	—	—	-14
OSSTr	—	—	—	—	-15	—	—	—	—	-15	—	—	—	—	—	-28
Issued	7	7	6	6	—	3	-1	-1	7	—	-1	3	5	6	-1	—

SST Type	Australia						South Pacific				
Lead	4	3	2	1	0	S	3	2	1	0	S
FSST	-2	-4	3	-9	-6	—	11	3	2	0	—
$HSST_p$	—	—	-13	-11	-20	—	—	—	-8	-11	—
OSST	—	—	—	—	—	-8	—	—	—	—	-21
OSSTr	—	—	—	—	—	-16	—	—	—	—	-26
Issued	4	5	9	5	5	—	6	10	7	5	—

Table 11: Likelihood skill scores ($\times 10^2$) for ACE, per basin, by lead time and SST forecast scenario. S denotes simulations, whose lead time is negative.

SST Type	Eastern Pacific					Western Pacific					Atlantic					
Lead	3	2	1	0	S	3	2	1	0	S	4	3	2	1	0	S
FSST	-10	1	9	7	—	14	-3	2	3	—	-8	6	1	3	-2	—
$HSST_p$	—	—	-12	-8	—	—	—	-21	-6	—	—	—	-14	8	2	—
OSST	—	—	—	—	-8	—	—	—	—	3	—	—	—	—	—	5
OSSTr	—	—	—	—	2	—	—	—	—	6	—	—	—	—	—	9
Issued	-1	-1	3	5	—	7	2	2	2	—	-2	6	7	7	4	—

other two categories. Resulting RPSS are shown at the bottom of Tables 8 and 9. These weakly persistent probabilistic forecasts often have better RPSS scores than those of the AGCM forced with persisted SST ($HSST_p$), and sometimes as good as or better than those forced with observed SSTs. Rather than showing that use of the AGCM with observed or predicted SST is unsuccessful, this outcome again shows that probabilities that deviate only mildly from climatological probabilities, even if derived from something as simple as the TC activity of the previous year, fare better under calibration-sensitive probabilistic verification measures (here, RPSS) than the higher amplitude probability shifts from climatology typically produced by today's AGCMs without proper statistical calibration.

The probability forecasts actually issued by IRI begin with the "raw" AGCM probabilities, modified to what the forecasters judge to have better probabilistic reliability. This nearly universally involves damping the amplitude of the model's deviation from climatological probabilities. A typical adjustment might be to modify the model's predicted probabilities of 5%, 10% and 85% to 20%, 30% and 50% for the below-, near- and above-normal categories, respectively. A less common adjustment is that of "rounding out" a bimodal probability forecast such as 35%, 5%, and 60% to a more Gaussian distribution such as 25%, 30% and 45%. Part of the reason for sharply bimodal distributions is assumed to be the limited (24-member) ensemble size. A still less common case for modification, and one that does not always improve the forecast quality, is that of the forecasters' judgement against the model forecasts, believing there is a model bias. Such doubt can pertain also to the SST forecast used to force the AGCM.

Tables 8, 9, 10, 11, indicate that the actually issued forecasts have better probabilistic reliability than the forecasts of the model output. Likelihood skill scores, and especially RPSS, are mainly positive for the issued forecasts, although modest in magnitude. This implies that the probability forecasts of the AGCM are potentially useful, once calibrated to correct for overconfidence or an implausible distribution shape. Such calibration could be done objectively, based on the longer hindcast history, rather than subjectively by the forecasters as done to first order here.

Figure 2 shows the approximately 6-year record of AGCM ensemble forecasts of NTC and ACE at all forecast lead times for each of the ocean basins. The vertical boxes show the inter-quartile range among the ensemble members, and the vertical dashed lines ("whiskers") extend to the ensemble member forecasts outside of that range. The asterisk indicates the observation value. Favorable and unfavorable forecast outcomes can be identified, such as, respectively, the ACE forecasts for the western North Pacific for 2002, and the ACE forecasts for the North Atlantic for 2004.

Figure 3 shows the same forecasts, except probabilistically for each of the tercile-based categories, both for the AGCM's forecasts (the "x" symbols) and for the subjectively modified publicly issued forecasts (the "o" symbols connected by lines). The AGCM's probability forecasts often deviate by large amounts from climatology, while the issued forecasts remain closer to climatology. Figure 4 shows the ranked probability skill score (RPSS) of these probability forecasts in the same format. The AGCM's probability forecasts result in highly variable skill (including both strongly negative and positive cases), leading to a somewhat

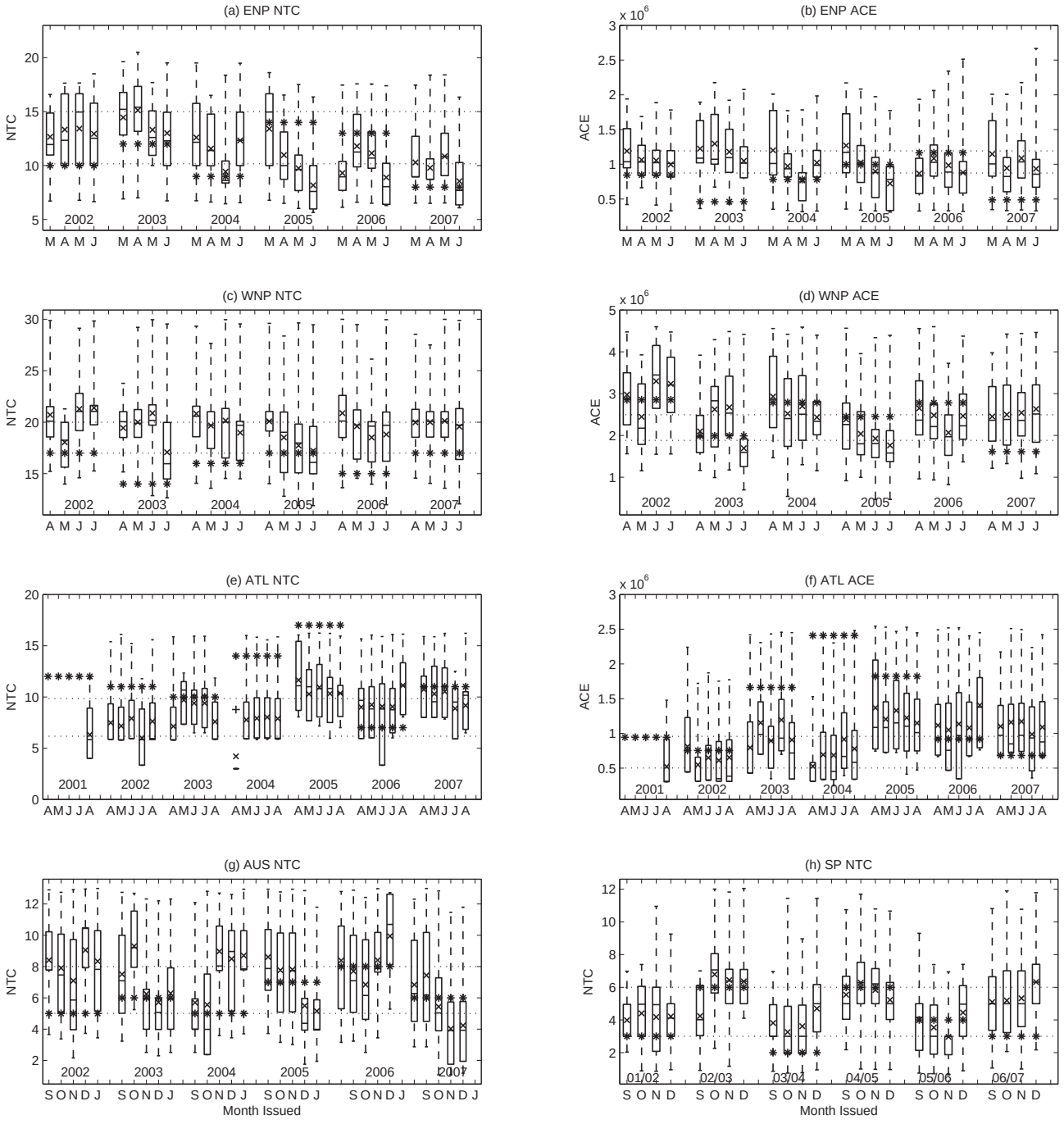

Figure 2: Model (raw) forecasts (box plots and whiskers) and observations (asterisks) of number of TCs (NTC) and accumulated cyclone energy (ACE) for all basins and leads. The cross inside the box shows the ensemble mean, and the horizontal line shows the median. Also shown by dotted horizontal lines are the boundaries between the tercile categories. Panels (a) - (f) are for the northern hemisphere basins, with NTC on the left panels and ACE on right panels, for ENP, ATL, WNP in each row, respectively. The two bottom panels are for NTC in the southern hemisphere basins: AUS (g) and SP (h).

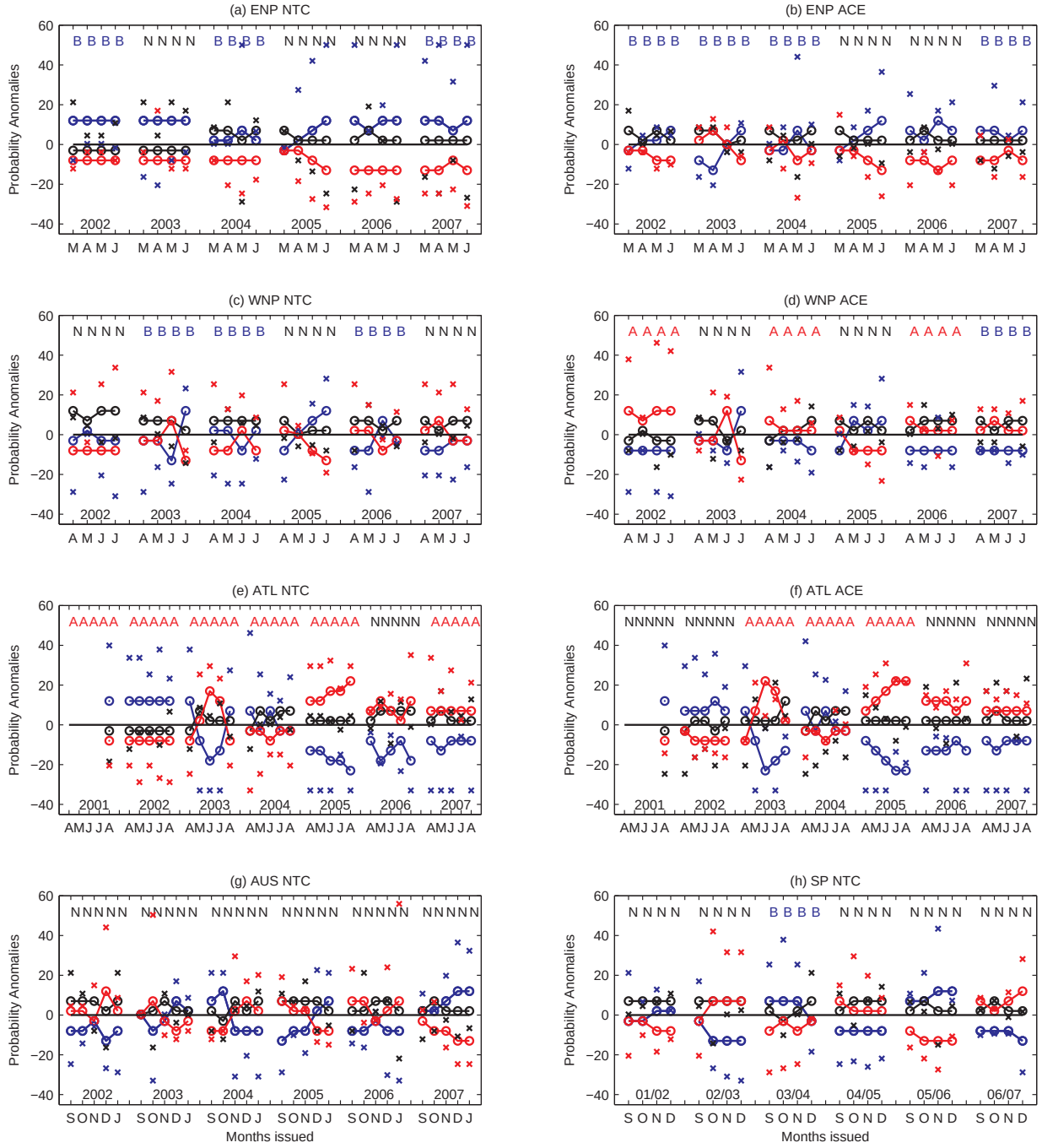

Figure 3: Issued (circles) and model (crosses) probability anomalies (difference of probability from 33.3% climatological probability values, X100) for all leads and years in each basin. The above (below) normal category probability anomalies are given in red (blue), and the near normal anomalies in black. The observed category is shown near top by the letters B (below normal), N (near normal) and A (above normal). Panels (a) to (h) are arranged as in Fig. 2,

negative overall skill. The issued forecasts, while never reaching positive magnitudes as great as those of some of the AGCM forecasts, also avoid negative overall skills of more than small magnitude[10]. Hence, the humanly modified TC forecasts have a higher average probabilistic skill level using RPSS.

The "over-confidence" of the AGCM forecasts is shown in more concrete terms in a reliability (or attributes) diagram (Hsu and Murphy 1986), shown in Fig. 5. Here the correspondence of the forecast probabilities with the observed relative frequency of occurrence is shown for the above normal and below normal categories. When the forecast probabilities closely match the observed relative frequencies, as would be desired, the lines approximate the dotted 45° line. Parts (a) and (b) show, for the 6-year period of forecasts, reliabilities for the issued forecasts and for the AGCM's forecasts prior to subjective modification, respectively. Despite the "jumpy" lines due to the small sample sizes, the lines for the issued forecasts are seen to have slopes roughly resembling the 45° line, indicating favorable reliability, while the lines for the AGCM's forecasts have a less obvious upward slope. The AGCM's forecast probabilities for the above or below normal categories of TC activity deviate from the climatological probabilities of 1/3 by much greater amounts than do their corresponding observed relative frequencies (see lower inset in panels of Fig. 5), resulting in low probabilistic forecast skill. The issued forecasts' deviations from climatological probabilities are limited by the forecasters according to the perceived level of uncertainty, and within the restricted probability ranges an approximate correspondence to the observed relative frequencies is achieved. The more reliable issued forecasts carry appropriately limited utility as represented by the lack of forecast sharpness—i.e. that the forecast probabilities rarely deviate appreciably from climatology, and from one another.

The lower panels of Fig. 5 show reliabilities for the longer historical period of AGCM hindcasts using prescribed observed SST ($OSST_r$; part(c)) and persisted SST anomaly ($HSST_p$; part (d)). Here the lines are smoother due to the larger sample sizes. Both diagrams show forecasts having some information value, as the lines have positive slope, but the slopes are considerably shallower than the 45° line, indicating forecast overconfidence. The slopes for forecasts using observed SST are slightly steeper than those for forecasts using persisted SST anomaly, as would be expected with the higher skill realized in forecasts forced by the actually observed lower boundary conditions.

That the TC activity forecasts of the AGCM have mainly positive correlation skill is consistent with their positive slopes in Fig. 5(b,c,d). Additionally, their mainly negative RPSS (Tables 8 - 11) is expected when the positive slopes on the reliability diagram (Fig. 5) are shallower than one-half of the ideal 45° slope=1 line (i.e. slope < 0.5) because then the forecasts' potential information value is more than offset by the miscalibration of the forecast probabilities (Hsu and Murphy 1986; Mason 2004). This is consistent with the deterministic TC forecasts having positive correlation skill but negative MSESS using climatology as the

[10] Because the RPSS is computed as a sum of squares of cumulative (over tercile categories) differences between forecast and observed probabilities, the lower limit of RPSS (-3.5) is farther below zero than the upper limit (+1.0) is above zero. Thus, high probabilities forecasted for an incorrect category outweigh high probabilities forecast for the correct category, and "over-confident" forecasts result in severe penalties even when the forecasts have some positive level of information value.

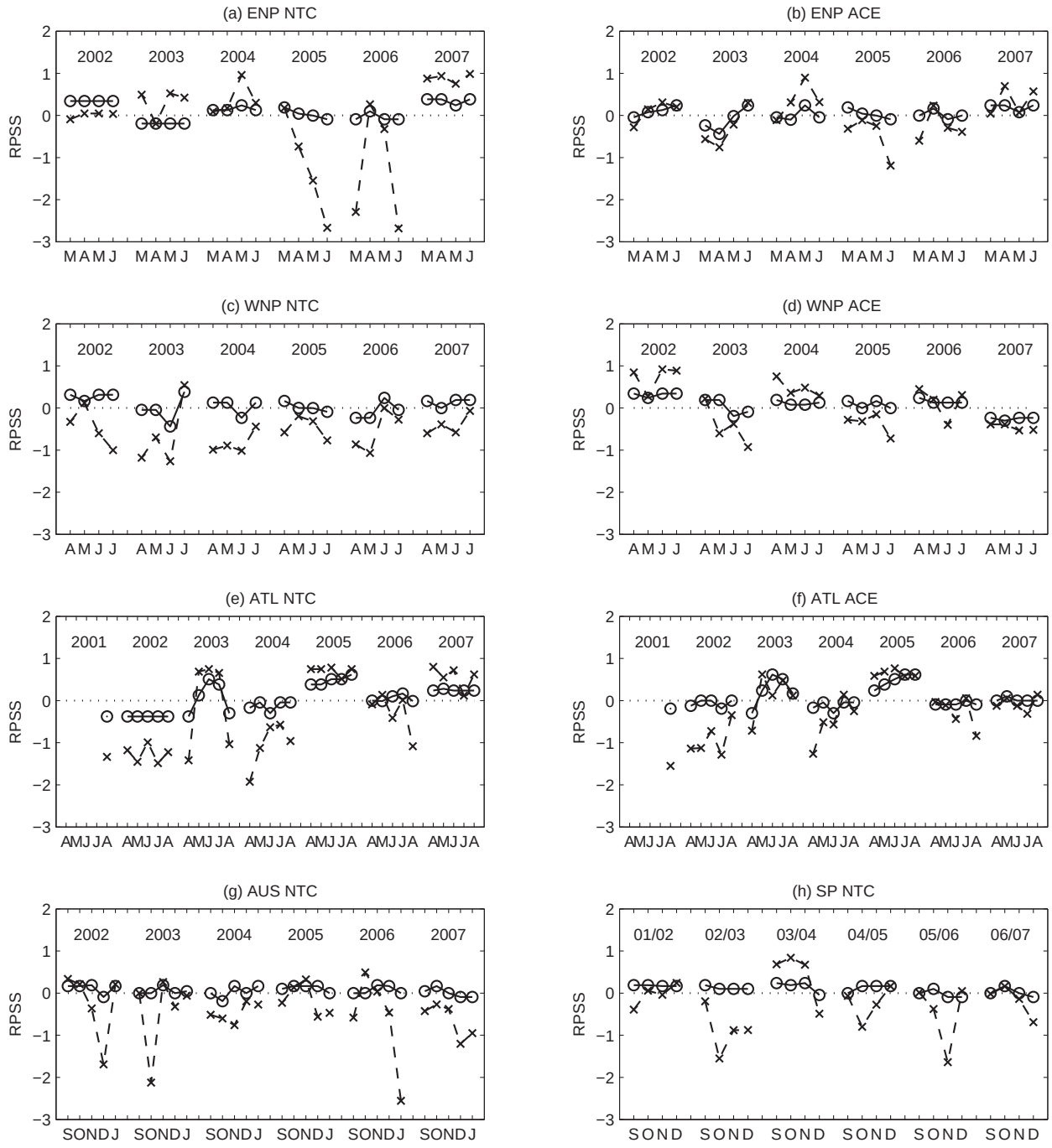

Figure 4: Ranked probability skill score (RPSS) for the issued (circles) and model (crosses) forecasts for all leads and years in each basin. Panels (a) to (h) are arranged as in Fig. 2

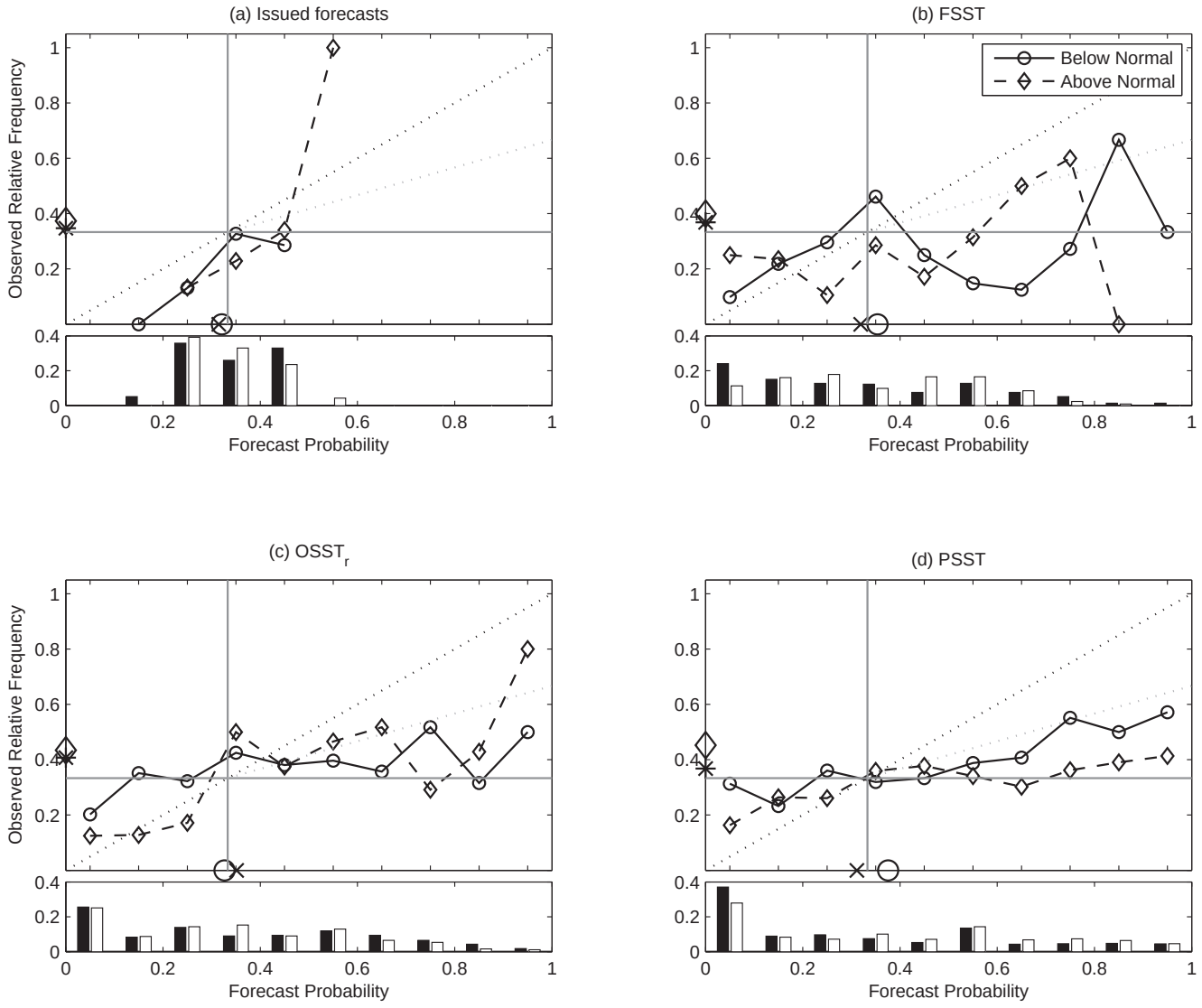

Figure 5: Reliability diagrams for above (circles ∘) and below (diamonds ⋄) normal categories: (a) issued forecasts, (b) FSST, (c) OSST_r, (d) HSST_p. The histograms (black bars - below normal, white bars - above normal) below each plot show the percentage frequency with which each category of probability was forecast. The circle (cross) indicates the overall mean of the forecast probabilities for the above (below) normal categories, and the diamond (asterisk) indicates likewise for observed relative frequencies. The vertical and horizontal lines indicate the climatologically expected foreast probability and observed relative frequency, respectively. The ideal realiability is shown by the dotted 45° diagonal line. The dotted line with shallower slope is the slope above which positive skill would be realized in RPSS, and below which (for positive slope) RPSS would be negative but correlation skill for corresponding deterministic forcasts would usually be positive, suggesting information value.

reference forecast, due to forecast anomalies stronger than warranted for the expected skill level.

Table 12: Comparison of skill ($\times 10^2$) between real-time probability forecasts for TC activity based directly on the AGCM (FSST), and those issued by IRI forecasters following subjective modification. Skills for forecasts for all five ocean basins for both NTC and ACE over the approximately 6-year period are aggregated using RPSS, likelihood and scaled ROC area for forecasts when above normal (AN) or below normal (BN) TC activity was observed. The ROC score is scaled as 2 x (area - 0.5) for increased comparability to the other skill measures.

Number of forecasts		48	48	48	50
Skill Score	Type				
Lead		3	2	1	0
RPSS	FSST	-24	-12	-22	-37
RPSS	Issued	9	9	8	6
Likelihood	FSST	2	-1	-1	-5
Likelihood	Issued	4	5	4	4
ROC (AN)	FSST	37	27	28	10
ROC (AN)	Issued	50	42	59	47
ROC (BN)	FSST	15	14	21	15
ROC (BN)	Issued	24	25	21	15

The skills of the real-time probabilistic forecasts over the approximately 6-year period are summarized in full aggregation (over basins and TC variable) in Table 12 using the RPSS, likelihood, and relative operating characteristics (ROC) (Mason 1982)) verification measures. The comparisons between the objective AGCM forecast output and the actually issued forecasts again underscore the need for calibration of AGCM forecasts that greatly underestimates the real-world forecast uncertainty. The AGCM's non-trivially positive scaled ROC areas for both above and below normal observed outcomes reveal their ability to provide useful information, as the ROC lacks sensitivity to calibration in a manner analogous to correlation for deterministic, continuous forecasts. In this particular set of forecasts, greater capability to discriminate above than below normal TC activity is suggested by the ROC skills.

3.3 A favorable and an unfavorable real-time forecast

Identification of "favorable" or "unfavorable" forecasts, while straightforward when considered deterministically, is less clear when comparing an observed outcome with its corresponding probability forecast. Probabilistic forecasts implicitly contain expressions of uncertainty. The position of an observed outcome within the forecast distribution is expected to vary across cases, and many cases are required to confirm that this variation is well described

by the widths of the probability forecast distributions. When a particular observation lies on a tail of the forecast distribution, it is impossible to determine whether this represents an unfavorable forecast or if it is an expected rare case, without examining a large set of such forecasts. The forecast distribution may be fully appropriate given the known forcing signals (Barnston et al. 2005). Here, we identify "favorable" and "unfavorable" cases in terms of the difference between the deterministic forecast (the model ensemble mean—which usually also approximates the central tendency of the forecast probability distribution) and the corresponding observation.

A critical aspect of the SST forcing to be forecast is the ENSO state during the peak season. Figure 6 shows the IRI's forecasts of the seasonal Niño3.4 index at 2 months lead time (e.g., a forecast for Aug-Sep-Oct SST issued in mid-June, with observed data through May) during the period of issued TC forecasts, with the corresponding observed seasonal SST. A moderate EN occurred during 2002-03, with weak ENs in 2004-05 and late 2006. A weak, brief La Niña (LN) condition was observed in very late 2005 and early 2006, and a stronger LN developed during mid-2007. The average of the observed Niño3.4 SST anomaly over the approximately 5 year period is 0.45, compared with an average 2-month lead forecast anomaly of 0.37, indicating a small forecast bias. The uncentered correlation coefficient for the period (where the means over the period are not removed) is in the 0.70s for forecasts for the northern hemisphere peak seasons, and 0.80s for forecasts for the southern hemisphere peak seasons. These correlations suggest moderately skillful forecasts of tropical Pacific SST fluctuations for the peak TC seasons in both hemispheres.

A favorable forecast for ACE in the western North Pacific took place in 2002. Figure 2(d) shows that the observation was in the above-normal category, and that the AGCM forecasts were not far from this number for the four lead times. For ACE in the western North Pacific, the ENSO condition is key, with EN (LN) associated with higher (lower) ACE. Between April through June of 2002 it became increasingly clear that an EN was developing, although the SST predictions contained a weaker EN that that observed (Fig. 6). Nonetheless, the SST predictions contained ENSO-related anomaly patterns of sufficient ampliltude to force an above normal ACE prediction that verified positively. The favorable AGCM forecasts are shown probabilistically in Fig. 3(d), with a positive RPSS verification shown in Fig. 4(d).

An unfavorable forecast outcome occurred for ASO 2004, when the ACE in the North Atlantic was observed to be 2.41×10^6 kt^2, the highest on record after 1970 for this season, but the AGCM forecasts from all five lead times were for between 0.5 and 1.0×10^6 kt^2, only in the near-normal category. A weak EN developed just prior to the peak season, which, while somewhat underpredicted, was present in the SST forecasts. But despite weak EN conditions during the 2004 peak season, NTC and especially ACE were well above normal (Fig. 2(e,f)). A feature of the EN that likely weakened its inhibiting effect on Atlantic TC development was its manifestation mainly in the central part of the tropical Pacific, and underrepresentation in the Niño3 region that appears more critical. Coupling of the warmed SSTs to the overlying atmosphere was also modest in ASO. Aspects of the SST that were less well predicted than ENSO were those that mattered more critically in this case: the Atlantic Multi-decadal Oscillation (AMO) (Goldenberg et al. 2001) in the North Atlantic,

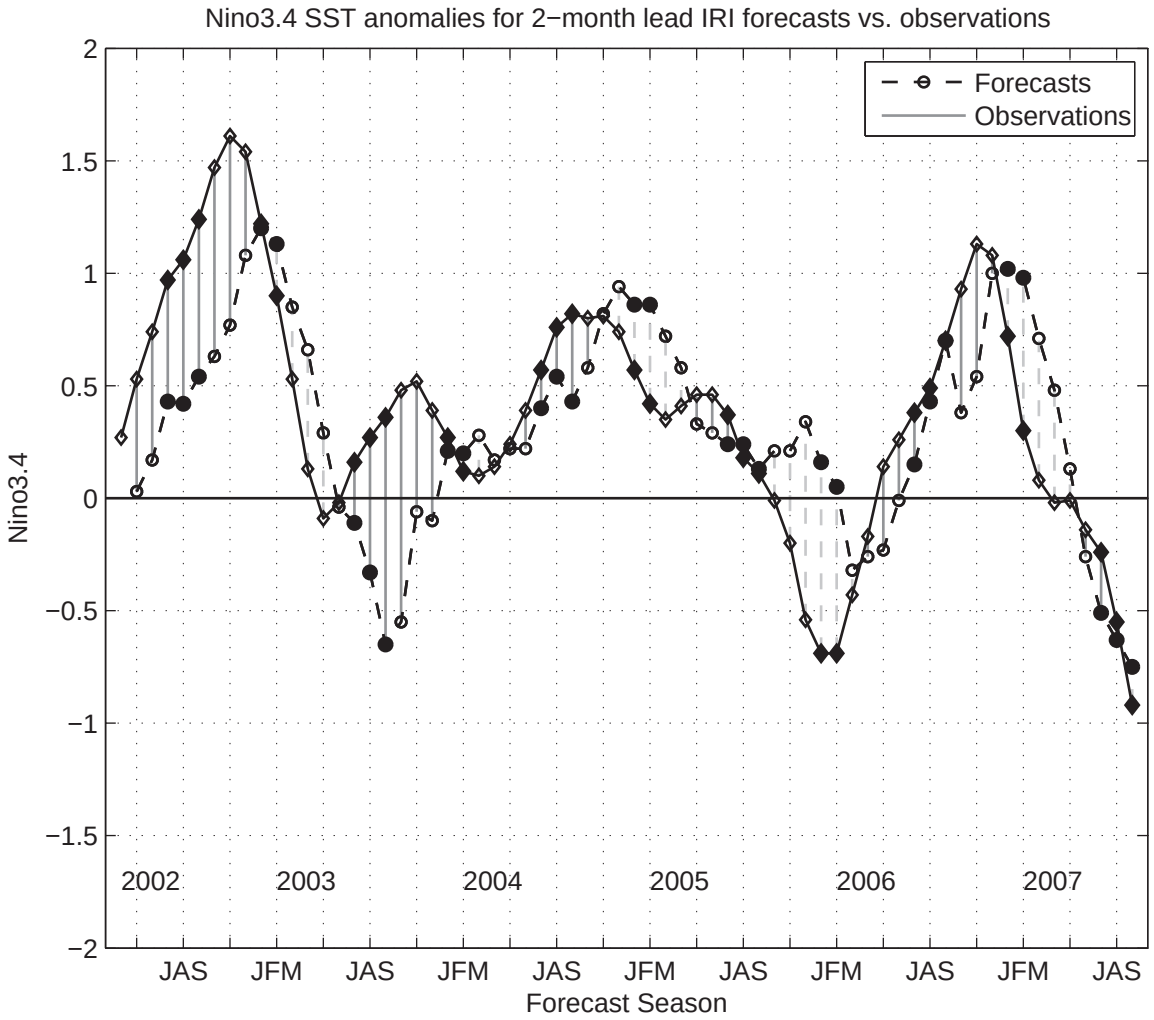

Figure 6: Two-month lead IRI forecasts (circles) and observations (diamonds) of Niño3.4 SST index anomalies. Seasons for which TCs are forecast are shown with filled symbols. The difference between the forecasts and observations are shown by solid (dotted) gray vertical lines when the observation was warmer (colder) than the forecast.

the Atlantic Meridional Mode (AMM) (Vimont and Kossin 2007) in the tropical Atlantic and in the main development region (MDR) (Goldenberg and Shapiro 1996) in the Caribbean. These regions developed markedly stronger positive anomalies than had been observed in April and May or forecast for the forthcoming peak season months, and are believed to have been a major cause of the high 2004 Atlantic TC activity level.

In both examples described above, the importance of the quality of the SST forecast for the peak TC season in the relevant tropical and subtropical ocean regions is clear. While ENSO-related Pacific SST is known to have some predictability, there is room for improvement in capturing it; and seasonal prediction of SST in the equatorial and North Atlantic is a yet more serious challenge.

4 Comparison with simple statistical predictions

One reasonably might ask whether the skill levels of the AGCM simulations and predictions described above are obtainable using statistical models derived purely from the historically observed TC data and the immediately preceding environmental data such as the wind, sea level pressure, or SST conditions the month prior to the forecast. How much the dynamical approach to TC prediction offers that is not obtainable using much less costly empirical approaches is explored here for deterministic skill, using environmental predictors in multiple regression. To minimize artificial skill associated with "fishing" for accidentally skillful predictors, the following restrictions are imposed: (i) A maximum of two predictors is used for each basin (although a predictor may be a fixed combination of more than one component, such as a current state plus a recent rate of change); (ii) for each ocean basin, the same predictors are used for NTC as for ACE; (iii) except for the case of Australia in which sea level pressure is used as a predictor, all predictors are SSTs averaged over rectangular index regions; (iv) all predictors must have a plausible physical linkage with the TC activity. "Leave-out-one" cross-validation is applied to assess the expected real-time predictive skill of the statistical models. We use mainly SST because of the well documented influence of SST anomaly patterns, including in particular the state and direction of evolution of ENSO, on the interannual variability of TC activity in most ocean basins. Statistical predictions are made at a lead time of one-month (e.g. June SST predicting the Atlantic peak season of ASO. A similar "prediction" is done for simulation of TC activity using predictors simultaneous with the center month of the peak TC season.[11] The simulation predictors are usually the same as those used for the one-month lead prediction.

The selection of the predictor SST indices is based both on previous studies and on examination of the geographical distribution of interannual correlation between SST and the given TC variable using 1970-2005 data. For example, Figure 7 shows the correlation field for SST in June versus Atlantic NTC during the ASO peak season, indicating the well-known inverse relationship with warm ENSO and positive association with the Atlantic Multidecadal Oscillation (Goldenberg et al. 2001)(via Atlantic SST near and North of 40°N), which

[11]The second month is used for both 3-month and 4-month peak seasons.

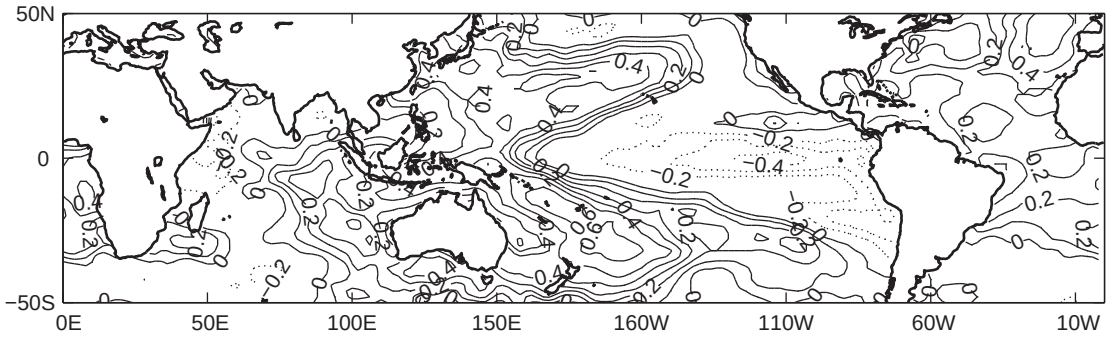

Figure 7: Correlations between June SST and Aug-Sep-Oct Atlantic NTC, 1970-2006. Contour interval 0.1. Zero and positive (negative) contours represented by solid (dotted) lines. The -0.1 contour is not shown.

is associated with the Atlantic Meridional Model (AMM) on interannual time-scales (Vimont and Kossin 2007; Kossin and Vimont 2007). When September SST is used, simultaneous with the Atlantic TC activity, these same two key regions remain important, but even stronger correlations appear for SST in the Main Development Region (Goldenberg and Shapiro 1996) in the North tropical Atlantic.

Table 13 identifies the two predictors used for each ocean basin for 1-month lead forecasts and for simultaneous simulations. For non-standard SST index regions, the boundaries of the rectangle are given. In the case of the Atlantic and western North Pacific forecasts, the first predictor contains both a recent SST level and a recent time derivative for the same region, to capture the ENSO status and direction of evolution. The model is applied to data spanning the 1970-2005 period. MDR is the main development region of the Atlantic (10°-20°N, 82°-20°W). AMO is the Atlantic multi-decadal oscillation region (here, 40°-50°N, 75°W-0°). Darwin is located in northern Australia (12.4°S, 130.9°E).

Many of the statistical predictors are ENSO-related. The Niño3 SST as the east-central tropical Pacific is found more relevant to Atlantic TC activity ((Gray et al. 1993)) than the location most central to ENSO itself (i.e. Niño3.4; (Barnston et al. 1997)). However, Niño3.4 is used as the ENSO index for Western North Pacific TC activity, the second predictor being SST in the subtropical northeastern Pacific associated with the North Pacific atmospheric circulation pattern found linked with that TC activity (Barnston and Livezey 1987; Chan et al. 1998, 2001). For northeast Pacific TC activity, SST regions are selected that highlight an ENSO-related east-west dipole in a northern subtropical latitude band, while the South Pacific and Australia regional TC predictions, also ENSO-keyed, are tailored to their locations south of the equator.

Table 14 indicates the strengths of relationship between each of the predictors and the predictand, the predictors' correlations with one another, and the resulting multiple correlation coefficient first within the model development sample, and then upon using one-year-out cross-validation. The cross-validated result—a more realistic estimate of expected skill in real-time forecasting—can be compared with the AGCM-derived skill shown in the subse-

Table 13: Predictors for statistical tropical cyclone 1-month lead forecasts and simulations. The month of the two SST predictors is indicated.

Basin	Season	1-month Lead Forecasts	Simultaneous Simulations
ENP	JJAS	1. 20°-30°N, 110°-130°W (Apr) (N subtropical version of Nino3) 2. 20°-30°N, 180°-140°E (Apr) (N part of outer ENSO horseshoe)	1. Same as forecasts, but for Jul 2. Same as forecasts, but for Jul
WNP	JASO	1. Niño3.4 (May) + Niño3.4 change from Feb-Mar to May 2. 20°-30°N, 110°-150°W (May) (NE subtropical Pacific)	1. Niño3.4 (Aug) 2. Same as forecasts, but for Aug
ATL	ASO	1. Niño3 (Jun) + Niño3 change from Mar-Apr to Jun 2. MDR (weighted 1/3) + AMO (weighted 2/3)	1. Niño3 (Sep) 2. MDR (Sep)
AUS	JFM	1. Niño3.4 to Niño4 (Nov) (5°N-5°S, 120°W-160°E) 2. Darwin SLP (SON)	1.Same as forecasts, but for Feb 2. Same as forecasts, but for JFM
SP	DJFM	1. Niño3 to Niño3.4 (Oct) (5°N-5°S, 90°-170°W) 2. 25°-35°S, 170°W-170°E (Oct) (S part of outer ENSO horseshoe)	1. Same as forecasts, but for Jan 2. Same as forecasts, but for Jan

quent column. Results for this comparison are mixed. The dynamical forecasts and simulations are slightly more skillful for south Pacific NTC, and in the eastern North Pacific basin in most cases. The statistical model produces higher skills in most cases in the Atlantic and western North Pacific for forecasts and, for some cases, simulations. Statistical tests indicate that none of the dynamical versus statistical skill differences are significant for the 36-case sample size. Considering this, and the alternation of skill rank between the approaches over the basins, there is no clear suggestion that one approach is generally superior to the other. That the dynamical approach tended to yield higher skills in the South Pacific, and no lower than the statistical method in the Australian region, could be related to the comparatively lower quality of SST predictor data south of the equator, particularly in the 1970s. It is possible that less accurate SST data would degrade the statistical forecasts more than the AGCM forecasts forced by the SST because the SST indices used in the statistical forecasts represent relatively smaller regions than the aggregate of the SST regions influencing the behavior of the AGCM. The larger areas of SST influencing the model may allow opportunity for opposing error impacts, leading to smaller net impacts.

Some points worth noting about this methodological comparison are: (i) the statistical models used here were kept fairly simple (limited to two adjustable coefficients, mainly

Table 14: Diagnostics for the 2-predictor statistical tropical cyclone forecasts and simulations, 1970-2005. The pr1 (pr2) columns show the correlations ($\times 10^2$) between the first (second) predictor and the observed TC variable (NTC or ACE) in 2-predictor multiple regression predictions at 1-month lead time, per ocean basin. Sim1 (sim2) are likewise, except for simulations in which the SST forcing is prescribed as that observed during peak TC season. Correlation between the two predictors is indicated in the "1vs2" column. The next two columns show the full sample and the one-year-out cross-validated multiple regression coefficients, the latter to be regarded as the skill estimate for real-time forecasts for comparison with dynamical (AGCM-based) skills, shown in the subsequent column. Dynamical predictive skill comes from the HSST_p at 1 month lead, and simulation skill from OSST_r.

		pr1	pr2	1vs2	R	Rcv	Dyn	sim1	sim2	1vs2	R	Rcv	Dyn
ENP	NTC	29	-30	**-33**	**37**	14	32	32	-28	-18	**39**	20	**37**
ENP	ACE	**40**	**-33**	**-33**	**45**	31	20	**47**	-27	-18	**50**	**35**	**45**
WNP	NTC	19	**41**	**47**	**41**	23	11	23	14	**54**	23	-6	28
WNP	ACE	**60**	**53**	**47**	**66**	**58**	6	**67**	**56**	**54**	**71**	**66**	**33**
ATL	NTC	**-38**	**49**	-19	**57**	**46**	**33**	**-45**	**58**	-2	**72**	**66**	**55**
ATL	ACE	-27	**60**	-19	**62**	**55**	**43**	**-34**	**47**	-2	**57**	**50**	**60**
AUS	NTC	**-40**	-30	**83**	**41**	23	26	**-44**	**-53**	**81**	**53**	**42**	**40**
SP	NTC	**38**	-28	**-55**	**39**	20	**42**	**42**	**-37**	**-48**	**46**	30	**42**

SST, and constant for NTC versus ACE), and may not be near-optimum; (ii) despite the use of cross-validation, some "fishing" may still have occurred in selecting the locations of the predictor SSTs, and there may be some artificial skill; (iii) the one-year-out cross-validation design has a negative skill bias in truly low predictability situations (Barnston and van den Dool 1993). Together, such caveats of opposing implications imply that the skill comparison results should be considered as rough estimates, intended to detect obvious skill differences. Such differences are not revealed here. One might expect that much of the skill of a near-perfect dynamical model would be realizable by a sophisticated (e.g., containing nonlinearities) statistical model if accurate observed data were available, since the observations should occur because of, and be consistent with, the dynamics of the ocean-atmosphere system with noise added. Seasonal climate has been shown to be statistically modeled fairly well using only linear relationships (Peng et al. 2000). However, linearity may compromise statistical skill in forecasting some seasonal phenomena, such as TC activity with its highly nonlinear hydrodynamics in individual storms that may not be reduced to linear behavior even upon aggregating over an entire season.

5 Conclusions

The International Research Institute for Climate and Society (IRI) has been issuing experimental TC activity forecasts for several ocean basins since early 2003. The forecasts are based on TC-like features detected and tracked in the ECHAM4.5 atmospheric general circulation model (AGCM), at low horizontal resolution (T42). The model is forced at its lower boundary by sea surface temperatures (SSTs) that are predicted first, using several other dynamical and statistical models. The skill of the model's TC predictions using historical observed SSTs are discussed as references against which skills using several types of predicted SSTs (including persisted SST anonmalies) are compared. The skill of the raw model output is also compared with that of subjective probabilistic forecasts actually developed since mid-2001, where the subjective forecasts attempt to correct the "overconfident" probabilistic forecasts from the AGCM. The skills of the AGCM-based forecasts are also compared with those from simple statistical forecasts based on observed SSTs preceding the period being forecast.

Results show that low-resolution AGCMs deliver statistically significant, but fairly modest, skill in predicting the interannual variability of TC activity. The levels of correlation skill are comparable to the levels obtained with simple empirical forecast models—here, models employing 2-predictor multiple regression using preceding area-average SST anomalies and their recent time derivative. In ocean basins where observed SST predictor data is of questionable quality, statistical prediction is less effective. Despite that this same SST is used as the boundary forcing for the AGCMs, the dynamical predictions tend to slightly outperform the statistical predictions in this circumstance.

In a 2-tiered dynamical prediction system such as that used in this study, the effect of imperfect SST prediction is noticeable in skills of TC activity compared with skills when the model is forced with historically observed SSTs.

Similar to climate forecasts made by AGCMs, probabilistic reliability of the AGCM's forecasts for TC activity forecasts is not favorable in that the model ensemble forecasts usually deviate too strongly from the climatological distribution, due sometimes to too narrow an ensemble spread but more often to too large a shift in the ensemble mean from climatology. This "overconfidence" of the AGCM forecasts is partly due to their being based on specific representations of the physics, including abbreviations through parametrization, and their own hindcast performance is not taken into account in forming ensemble forecasts. Upon subjective human intervention the forecasts are made more conservative and reliability is improved, leading to higher probabilistic verification scores than (but similar correlation scores to) the uncalibrated AGCM forecasts.

We plan to examine skill of other models in hopes of being able to add more information, and hopefully skill, to our seasonal TC forecast. The problem of "overconfidence" in AGCMs is relieved to some extent with the use of multi-model ensembles: adding additional models should help restrain the probabilistic amplitude exhibited by a single model. Issues not examined here are the role of AGCM spatial resolution in governing predictive skill, and the impact of using a fully coupled dynamical system rather than a 2-tiered system as used here. Although prospects for the future improvement of dynamical TC prediction are

uncertain, it appears likely that additional improvements in dynamical systems will make possible better TC predictions. As is the case for dynamical approaches to ENSO and near-surface climate prediction, future improvements will depend on better understanding of the underlying physics, more direct physical representation through higher spatial resolution, and substantial increases in computer capacity. Hence, improved TC prediction should be a natural by-product of improved prediction of ENSO, global tropical SST, and climate across various spatial scales.

References

Aldrich, J., 1997: R. A. Fisher and the making of maximum likelihood 1912-1922. *Statist. Sci.*, **12**, 162 – 176.

Anderson, J. L., 1996: A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate*, **9**, 15181530.

Barnston, A. G., 1992: Correspondence among the correlation, RMSE, and Heidke forecast verification measures - refinement of the Heidke score. *Wea. Forecasting*, **7**, 699-709.

— and R. E. Livezey, 1987: Classification, seasonality and persistence of low-frequency atmospheric circulation patterns. *Mon. Wea. Rev.*, **115**, 1083–1126.

— and H. M. van den Dool, 1993: A degeneracy in cross-validated skill in regression-based forecasts. *J. Climate*, **6**, 963 – 977.

— , M. Chelliah, and S. B. Goldenberg, 1997: Documentation of a highly ENSO-related SST region in the equatorial Pacific. *Atmosphere-Ocean*, **35**, 367–383.

— , A. Kumar, L. Goddard, and M. P. Hoerling, 2005: Improving seasonal prediction practices through attribution of climate variability. *Bull. Amer. Meteor. Soc.*, **86**, 59–72, doi: 10.1175/BAMS-86-1-59.

— , S. J. Mason, L. Goddard, D. G. DeWitt, and S. Zebiak, 2003: Multimodel ensembling in seasonal climate forecasting at IRI. *Bull. Amer. Meteor. Soc.*, **84**, 1783–1796.

Bell, G. D., et al., 2000: Climate assessment for 1999. *Bull. Amer. Meteor. Soc.*, **81**, S1–S50.

Bengtsson, L., 2001: Hurricane threats. *Science*, **293**, 440–441.

—, H. Böttger, and M. Kanamitsu, 1982: Simulation of hurricane-type vortices in a general circulation model. *Tellus*, **34**, 440–457.

— , M. Botzet, and M. Esch, 1995: Hurricane-type vortices in a general circulation model. *Tellus*, **47A**, 175–196.

— , —, and —, 1996: Will greenhouse gas-induced warming over the next 50 years lead to higher frequency and greater intensity of hurricanes? *Tellus*, **48A**, 57–73.

— , K. I. Hodges, and M. Esch, 2007a: Tropical cyclones in a T159 resolution global climate model: comparison with observations and re-analysis. *Tellus*, **59 A**, 396 – 416.

— , — , —, N. Keenlyside, L. Kornblueh, J.-J. Luo, and T. Yamagata, 2007b: How may tropical cyclones change in a warmer climate? *Tellus*, **59 A**, 539 – 561.

— , U. Schlese, E. Roeckner, M. Latif, T. P. Barnett, and N. E. Graham, 1993: A two tiered approach to long-range climate forecasting. *Science*, **261**, 1026–1029.

Broccoli, A. J. and S. Manabe, 1990: Can existing climate models be used to study anthropogenic changes in tropical cyclone climate? *Geophys. Rev. Lett.*, **17**, 1917–1920.

Buckley, M. J., L. M. Leslie, and M. S. Speer, 2003: The impact of observational technology on climate database quality: Tropical cyclones in the Tasman Sea. *J. Climate*, **16**, 2650–2645.

Camargo, S. J. and A. H. Sobel, 2004: Formation of tropical storms in an atmospheric general circulation model. *Tellus*, **56A**, 56–67.

— and S. E. Zebiak, 2002: Improving the detection and tracking of tropical storms in atmospheric general circulation models. *Wea. Forecasting*, **17**, 1152–1162.

—, A. G. Barnston, and S. E. Zebiak, 2005: A statistical assessment of tropical cyclones in atmospheric general circulation models. *Tellus*, **57A**, 589–604.

—, H. Li, and L. Sun, 2007a: Feasibility study for downscaling seasonal tropical cyclone activity using the NCEP regional spectral model. *Int. J. Clim.*, **27**, 311–325, doi 10.1002/joc.1400.

—, A. G. Barnston, P. J. Klotzbach, and C. W. Landsea, 2007b: Seasonal tropical cyclone forecasts. *WMO Bull.*, **56**, 297–309.

—, A. H. Sobel, A. G. Barnston, and K. A. Emanuel, 2007c: Tropical cyclone genesis potential index in climate models. *Tellus*, **59A**, 428–443.

Cane, M. A. and S. E. Zebiak, 1985: A theory for El-Niño and the Southern Oscillation. *Science*, **228**, 827–832.

—, —, and S. C. Nolan, 1986: Experimental forecasts of El Niño. *Nature*, **321**, 827–832.

Chan, J. C. L., J. E. Shi, and C. M. Lam, 1998: Seasonal forecasting of tropical cyclone activity over the western North Pacific and the South China Sea. *Wea. Forecasting*, **13**, 997–1004.

—, —, and K. S. Liu, 2001: Improvements in the seasonal forecasting of tropical cyclone activity over the western North Pacific. *Wea. Forecasting*, **16**, 491–498.

Chen, D., M. A. Cane, A. Kaplan, S. E. Zebiak, and D. Huang, 2004: Predictability of El Niño over the past 148 years. *Nature*, **428**, 733–735.

Epstein, E. S., 1969: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.*, **8**, 985–987.

Goddard, L., A. G. Barnston, and S. J. Mason, 2003: Evaluation of the IRI's "Net Assessment" seasonal climate forecasts: 1997-2001. *Bull. Amer. Meteor. Soc.*, **84 (12)**, 1761–1781.

Goldenberg, S. B. and L. J. Shapiro, 1996: Physical mechanisms for the association of El Niño and West African rainfall with Atlantic major hurricane activity. *J. Climate*, **9**, 1169–1187.

—, C. W. Landsea, A. M. Mestas-Nuñez, and W. M. Gray, 2001: The recent increase in Atlantic hurricane activity: Causes and implications. *Science*, **293**, 474–479.

Gray, W. M., C. W. Landsea, P. W. Mielke Jr., and K. J. Berry, 1993: Predicting Atlantic basin seasonal tropical cyclone activity by 1 August. *Wea. Forecasting*, **8**, 73–86.

Hsu, W. -R., and A. H. Murphy, 1986: The attributes diagram: A geometirical framework for assessing the quality of probability forecasts. *Int. J. Forecasting*, **2**, 285-293.

Ji, M., D. W. Behringer, and A. Leetmaa, 1998: An improved coupled model for ENSO prediction and implications for ocean initialization. Part II: The coupled model. *Mon. Wea. Rev.*, **126**, 1022–1034.

Kharin, V. V., and F. W. Zwiers, 2002: Climate predictions with multimodel ensembles. *J. Climate*, **15**, 793–799.

Klotzbach, P. J., 2007a: Recent developments in statistical prediction of seasonal Atlantic basin tropical cyclone activity. *Tellus*, **59A**, 511–518.

—, 2007b: Revised prediction of seasonal Atlantic basin tropical cyclone activity from 1 August. *Wea. Forecasting*, **22**, 937–949.

Knutson, T. R., J. J. Sirutis, S. T. Garner, and R. E. T. I. M. Held, 2007: Simulation of the recent multidecadal increase of Atlantic hurricane activity using an 18-km-Grid regional model. *Bull. Amer. Meteor. Soc.*, **88**, 1549-1565.

Kossin, J. P. and D. J. Vimont, 2007: A more general framework for understanding Atlantic hurricane variability and trends. *Bull. Amer. Soc.*, **88**, 1767–1781.

Landman, W. A., A. Seth, and S. J. Camargo, 2005: The effect of regional climate model domain choice on the simulation of tropical cyclone-like vortices in the southwestern Indian ocean. *J. Climate*, **18**, 1263–1274.

Lyon, B., and S. J. Camargo, 2008: The seasonally-varying influence of ENSO on rainfall and tropical cyclone activity in the Philippines. *Clim. Dyn.*, **26**, doi: 10.1007/s00382-008-0380-z, online first.

Manabe, S., J. L. Holloway, and H. M. Stone, 1970: Tropical circulation in a time-integration of a global model of the atmosphere. *J. Atmos. Sci.*, **27**, 580–613.

Mason, I., 1984: A model for assessment of weather forecasts. *Aus. Meteor. Mag.*, **30**, 291–303.

Mason, S. J., 2004: On using climatology as a reference strategy in the Brier and ranked probability skill scores. *Mon. Wea. Rev.*, **132**, 1891–1895.

—, L. Goddard, N. E. Graham, E. Yulaeva, L. Q. Sun, and P. A. Arkin, 1999: The IRI seasonal climate prediction system and the 1997/98 El Niño event. *Bull. Amer. Meteor. Soc.*, **80**, 1853–1873.

Matsuura, T., M. Yumoto, and S. Iizuka, 2003: A mechanism of interdecadal variability of tropical cyclone activity over the western North Pacific. *Clim. Dyn.*, **21**, 105–117.

Owens, B. F. and C. W. Landsea, 2003: Assessing the skill of operational Atlantic seasonal tropical cyclone forecasts. *Wea. Forecasting*, **18**, 45–54.

Peng, P. T., A. Kumar, A. G. Barnston, and L. Goddard, 2000: Simulation skills of the SST-forced global climate variability of the NCEP-MRF9 and the Scripps-MPI ECHAM3 models. *J. Climate*, **13**, 3657–3679.

Rajagopalan, B., U. Lall, and S. E. Zebiak, 2002: Categorical climate forecasts through regularization and optimal combination of multiple GCM ensembles. *Mon. Wea. Rev.*, **130**, 1792–1811.

Repelli, C. A. and P. Nobre, 2004: Statistical prediction of sea-surface temperature over the tropical Atlantic. *Int. J. Climatol.*, **24**, 45–55.

Reynolds, R. W., N. A. Rayner, T. M. Smith, D. C. Stokes, and W. Wang, 2002: An improved in situ and satellite SST analysis for climate. *J. Climate*, **15**, 1609–1625.

Robertson, A. W., U. Lall, S. E. Zebiak, and L. Goddard, 2004: Optimal combination of multiple atmospheric GCM ensembles for seasonal prediction. *Mon. Wea. Rev.*, **132**, 2732–2744.

Roeckner, E., et al., 1996: The atmospheric general circulation model ECHAM-4: Model description and simulation of present-day climate. Tech. Rep. 218, Max-Planck Institute for Meteorology, Hamburg, Germany. 90 pp.

Royer, J.-F., F. Chauvin, B. Timbal, P. Araspin, and D. Grimal, 1998: A GCM study of the impact of greenhouse gas increase on the frequency of occurrence of tropical cyclone. *Climatic Change*, **38**, 307–343.

Ryan, B. F., I. G. Watterson, and J. L. Evans, 1992: Tropical cyclone frequencies inferred from Gray's yearly genesis parameter: Validation of GCM tropical climate. *Geophys. Res. Lett.*, **19**, 1831–1834.

Saha, S., S. Nadiga, C. Thiaw, J. Wang, W. Wang, Q. Zhang, H.M. Van den Dool, H.-L. Pan, S. Moorthi, D. Behringer, D. Stokes, M. Peña, S. Lord, G. White, W. Ebisuzaki, P. Peng, and P. Xie, 2006: The NCEP climate forecast system. *J. Climate*, **19**, 34833517.

Saunders, M. A. and A. S. Lea, 2004: Seasonal prediction of hurricane activity reaching the coast of the United States. *Nature*, **434**, 1005–1008, doi:10.1038/nature03454.

Thorncroft, C. and I. Pytharoulis, 2001: A dynamical approach to seasonal prediction of Atlantic tropical cyclone activity. *Wea. Forecasting*, **16**, 725–734.

Tippett, M. K, A. G. Barnston, and A. W. Robertson, 2007: Estimation of seasonal precipitation tercile-based categorical probabilities from ensembles. *J. Climate*, **20**, 22102228.

Van den Dool, H. M., 1994: Searching for analogues, how long must we wait? *Tellus*, **46A**, 314–324.

—,2007: *Empirical Methods in Short-Term Climate Prediction*, Chapter 7, Oxford University Press, 215 pp.

Vimont, J. P. and J. P. Kossin, 2007: The Atlantic meridional mode and hurricane activity. *Geophys. Res. Lett.*, **34**, L07 709, doi:10.1029/2007GL029683.

Vitart, F., 2006: Seasonal forecasting of tropical storm frequency using a multi-model ensemble. *Q. J. R. Meteorol. Soc.*, **132**, 647–666, doi:10.1256/qj.05.65.

— and J. L. Anderson, 2001: Sensitivity of Atlantic tropical storm frequency to ENSO and interdecadal variability of SSTs in an ensemble of AGCM integrations. *J. Climate*, **14**, 533–545.

— and T. N. Stockdale, 2001: Seasonal forecasting of tropical storms using coupled GCM integrations. *Mon. Wea. Rev.*, **129**, 2521–2537.

—, D. Anderson, and T. Stockdale, 2003: Seasonal forecasting of tropical cyclone landfall over Mozambique. *J. Climate*, **16**, 3932–3945.

—, J. L. Anderson, and W. F. Stern, 1997: Simulation of interannual variability of tropical storm frequency in an ensemble of GCM integrations. *J. Climate*, **10**, 745–760.

—, —, and —, 1999: Impact of large-scale circulation on tropical storm frequency, intensity and location, simulated by an ensemble of GCM integrations. *J. Climate*, **12**, 3237–3254.

—, M. R. Huddleston, M. Déqué, D. Peake, T. N. Palmer, T. N. Stockdale, M. K. Davey, S. Inenson, and A. Weisheimer, 2007: Dynamically-based seasonal forecasts of Atlantic tropical storm activity issued in June by EUROSIP. *Geophys. Res. Lett.*, **34**, L16 815, doi:10.1029/2007GL030740.

Walsh, K. J. E., K. C. Nguyen, and J. L. McGregor, 2004: Fine-resolution regional climate model simulations of the impact of climate change on tropical cyclones near Australia. *Clim. Dyn.*, **22**, 47–56.

— and B. F. Ryan, 2000: Tropical cyclone intensity increase near Australia as a result of climate change. *J. Climate*, **13**, 3029–3036.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*, 2nd Ed. International Geophysics Series, Vol. 59, Academic Press, 627 pp.

Wu, G. and N. C. Lau, 1992: A GCM simulation of the relationship between tropical storm formation and ENSO. *Mon. Wea. Rev.*, **120**, 958–977.

Zebiak, S. E. and M. A. Cane, 1987: A model El-Niño Southern Oscillation. *Mon. Wea. Rev.*, **115**, 2262–2278.