Tellus SERIES A DYNAMIC METEOROLOGY AND OCEANOGRAPHY
PUBLISHED BY THE INTERNATIONAL METEOROLOGICAL INSTITUTE IN STOCKHOLM

Tellus (2011), 63A, 727–745

TELLUS

Hurricane forecasts using a suite of large-scale models

By T. N. KRISHNAMURTI[1*], MRINAL K. BISWAS[1], BRIAN P. MACKEY[2], ROBERT G. ELLINGSON[1] and PAUL H. RUSCHER[1], [1]*Department of Earth, Ocean and Atmospheric Sciences, The Florida State University, Tallahassee, FL 32306-4520, USA;* [2]*Weather Predict Inc, Raleigh, NC, USA*

(Manuscript received 12 June 2010; in final form 24 January 2011)

ABSTRACT

This paper provides an account of the performance of a multimodel ensemble for real time forecasts of Atlantic tropical cyclones during 2004, 2005 and 2006. The Florida State University (FSU) superensemble is based on a suite of model forecasts and the interpolated official forecast that were received in real time at the National Hurricane Center. The FSU superensemble is a multimodel ensemble that utilizes forecasts from the member models by removing their individual biases based on a recent past history of their performances. This superensemble carries separate statistical weights for track and intensity forecasts for every 6 h of the member model forecasts.

The real time results from 2004 show an improvement up to 15% for track forecasts and up to 11% for intensity forecasts for the superensemble compared to other models and consensus aids. During 2005, the superensemble intensity performance was best for most lead times. The consistency of the superensemble forecasts of track are also illustrated for several storms of 2004 season. The superensemble methodology produced impressive intensity forecasts for Rita and Wilma during 2005. The study shows the capability of the superensemble in predicting rapidly intensifying storms when most member models failed to capture their strengthening.

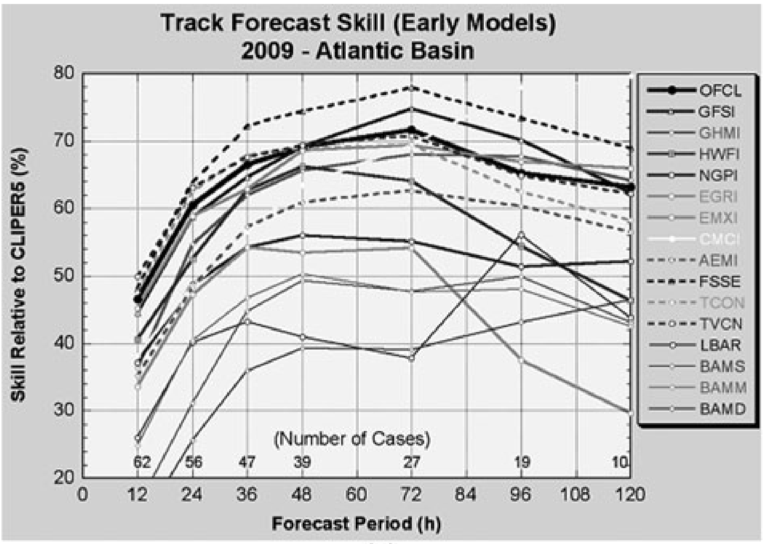
1. Introduction

This paper is a sequel, to a recent study, Krishnamurti et al. (2010), on hurricane track and intensity forecasts from a suite of multimodels. The previous study was based on a suite of mesoscale models. This study utilizes a suite of operational large-scale models. This present study was motivated from the performance of a multimodel superensemble (Krishnamurti et al., 1999) during the current years operations for the hurricane season of 2009. Figure 1a, was provided to us by the National Hurricane Center (NHC), from their post season skill evaluations. This illustration includes the track forecast skills from as many as 15 models and the official forecast. Here the ordinate denotes the percentage improvement on tack forecasts above those obtained from CLIPER, which is based on a combination of climatology and persistence. The abscissa denotes hours of forecast. This is a summary of an entire season's skill, which season included 9 named storms and 62 forecasts. This illustration essentially shows that the Florida State University (FSU) multimodel superensemble carried the most improvement over the CLIPER. In this paper, we describe this methodology and skills of forecast for several recent past years.

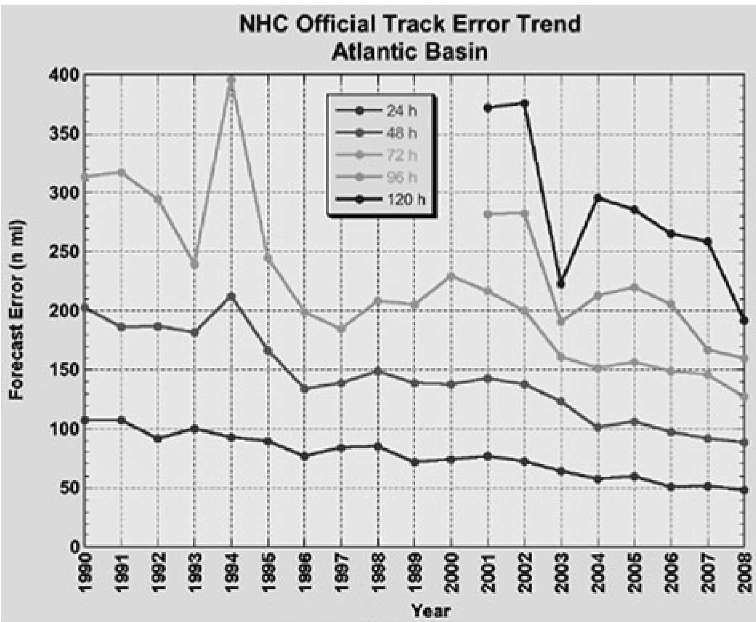
*Corresponding author.
e-mail: tkrishnamurti@fsu.edu
DOI: 10.1111/j.1600-0870.2011.00519.x

Multimodel ensemble forecasting has been recently facilitated by rapid data exchange among various operational modelling groups. It has now become possible to use multimodel datasets to construct ensemble forecasts for weather and seasonal climate and most major operational forecast agencies have ensemble forecasts as an essential component for their operational forecasting activities (e.g. Toth and Kalnay, 1993, 1997; Molteni et al., 1996; Buizza et al., 2005). The idea of a multimodel superensemble was first reported in Krishnamurti et al. (1999). The current superensemble methodology takes the forecasts of track and intensity every 6 h from each of the member models, and provides a collective, bias-corrected consensus forecast, by removing the individual biases of the member models. This superensemble carries separate statistical weights for track and intensity forecasts for every 6 h of the member model forecasts. This study focuses on real time Atlantic tropical cyclone forecast performance, based on FSU superensemble guidance provided to the NHC for the 2004 and 2005 seasons. The 2006 forecasts shown here were not disseminated to the NHC but were done on realtime for research purposes. The forecasts provided to NHC during 2006 were from a private company and not from FSU. Thus, we are verifying the results based on the research done at FSU.

Results on tropical cyclone forecast performance, from the use of this multimodel superensemble, were reported in a series of papers by Krishnamurti et al. (1999, 2000), Williford et al.



(a)



(b)

Fig. 1. (a) NHC official Track forecast skills for 2009 season. (b) NHC official annual average track errors (nm) in Atlantic basin for hurricanes and tropical storms. (c) Official Intensity error trend (in kt) in Atlantic Basin from 1990 (Adapted from NHC Annual Verification Report 2008).

(2003), Kumar et al. (2003) and Cane and Milelli (2006). In this series of papers, a superensemble methodology was developed that required hindcasts for recent years storms. Goerss et al. (2004) and Sampson et al. (2005) have used ensemble-average (consensus) forecast for western North Pacific and Southern Hemisphere and have shown great improvement over the member models. Several ensemble averages or consensus aids based on combinations of good models, have been used as models in the Atlantic basin such as the GUNA [Franklin (2006) and Goerss (2000) (introduced in 2001 by NHC which provides an ensemble mean forecast of **G**FDL (Kurihara et al., 1993, 1995, 1998), **U**. K. Met Office model (UKMO) (Cullen, 1993; Heming et al., 1995), **N**OGAPS (Hogan and Rosmond, 1991; Goerss and Jeffries, 1994) and **A**viation now Global Forecast System (GFS)]. Similarly GUNS, introduced in 2000 by NHC is extracted from the GFDL, UKMO model, NOGAPS. An ensemble model CONU (Goerss, 2000) is derived from at least two of the five models: the National Centers for Environmental Predictions (NCEP) GFS and GFDL model, the Navy's version of the GFDL model, NOGAPS, and the UKMO model was started in 2004. These consensus aids are also used by NHC for track forecasts of tropical cyclones (A complete list of acronyms appears in Table 1).

In the last 5 yr considerable progress in forecast improvements has been made from ensemble forecasting efforts, especially in of global Numerical Weather Prediction (NWP) following

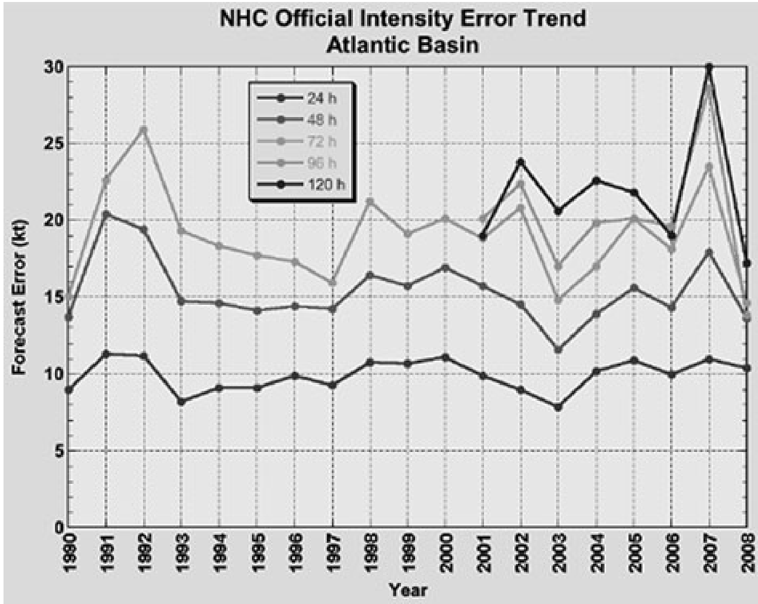


Fig. 1. Continued.

Table 1. Acronyms used for the description of hurricane superensemble forecasts

Acronyms	Descriptions	Acronyms	Descriptions
GFDL	Geophysical Fluid Dynamics Laboratory model	GFSI (formerly AVNI)	A time interpolated data set of the US National Weather Service Global Forecast System (GFS)
GFDI	A time interpolated data set of the GFDL model	GUNA	Ensemble mean of GFDI, UKMI, NGPI and GFSI
UKMI	A time interpolated data set of the U.K. Met Office model (UKMO)	SHIFOR5	A statistical model of HRD for hurricane intensity forecasts
NOGAPS	(U. S.) Navy Operational Global Atmospheric Prediction System	SHIPS	A statistical model of HRD for hurricane intensity forecasts
NGPI	A time interpolated data set of the NOGAPS model	OFCI	A time interpolated data set of the NHC forecast
EMC	Environmental Meteorological Center	BAMS	Beta Advection Model (Shallow Layer)
SHF5	SHIFOR5	HRD	Hurricane Research Division
ENSM	Ensemble mean of member models	FSSE	Florida State University Superensemble
NHC	National Hurricane Center	OFCL	Official forecasts of NHC
DSHP	SHIPS with inland decay	JTWC	Joint Typhoon Warning Center

Krishnamurti et al. (2000), Kumar et al. (2003) used FSU superensemble for western North western Pacific typhoons, and Krishnamurti et al. (2001) showed that forecast improvements are possible for forecasting of heavy rains. Vialard et al. (2005) and Chakraborty and Krishnamurti (2006) used ensemble forecasting for climate forecasts. Raftery et al. (2003) used Bayesian Model Averaging used for combining ensemble members for weather prediction.

In the past two decades there has been a remarkable improvement in NHC's hurricane track predictions (Franklin, 2009). Figure 1b shows the trend in NHC track forecast error during the last 18 yr. The 72-h track errors of NHC were reduced from around 460 nm (∼850 km) to 160 nm (∼296 km) in 2005. Track forecast errors are defined as the great-circle distance between a cyclone's forecasted position and the best track position at the forecast verification time (Järvinen et al., 1984; Franklin, 2005a).

Whereas substantial improvements in track forecasts have been demonstrated in the last two decades, intensity forecasts remain problematic. Figure 1c shows the improvements in NHC intensity forecasts during the last 15 yr. Only a slight improvement in intensity forecasts during the last 5 yr is evident. For better intensity forecasts, very high-resolution modelling may be needed in order to capture the intricate details of the eye wall.

Phenomena like eye wall replacement, its initiation and life cycles, require grid resolutions of 1–2 km for its proper simulation (Houze et al., 2007). Statistical modelling may also provide for some improvement in this area, as well.

In this paper, the superensemble methodology for hurricane forecasts is described along with some examples of how the superensemble works. Our primary focus is on the performance of the pseudo-operational FSU superensemble during the 2004 and 2005 Atlantic hurricane season, which was delivered for the first time on a regular and timely basis to operational forecasters at NHC. The 2006 season statistics shown here are the forecasts done by FSU and are not part of the ATCF (http://www.nrlmry.navy.mil/ atcf_web/ index1.html) 'a deck' files for NHC. Also discussed are some insights into the evolving science of the superensemble technique.

2. Superensemble methodology

Three separate superensemble forecasts, one each for latitude, longitude and intensity, are combined to predict the track and intensity of a tropical cyclone. The superensemble procedure includes a training phase and a forecast phase. The training phase utilizes track and intensity forecasts from a suite of models and multimodels. The training used for Atlantic storm forecasts are solely from model outputs in the Atlantic basin. This is done since we have seen from our past experience that model behaviour changes from basin to basin. The superensemble methodology works best when the whole of Atlantic basin is taken into account instead of dividing it into small basins. The biases are also calculated from model forecasts from the Atlantic storms only and hence they do not vary in space. But they do vary for different forecast times. Usually, we use the previous year storms to construct the training data set in the beginning of a season. As a storm is declared non-tropical, they were added to the training dataset. Thus, the length of the training database is not fixed for the entire season. As an example the training length (no of cases) for Frances are 433, 394, 350, 299, 259, 224, 135, 115, 99 and 91 from 12 to 120 h lead time whereas for Ivan it was 481, 440, 394, 341, 299, 260, 167, 143, 125 and 115 h. The tropical cyclones chosen for the training period includes all cyclones of tropical storm strength (17 m s^{-1}) or greater.

The equation below is solved for the construction of the superensemble:

$$S_j = \bar{O}_j + \sum_{i=1}^{n} a_{ij}(F_{ij} - \bar{F}_{ij}), \tag{1}$$

where S_j is the superensemble forecast increment (12-h changes in latitude, longitude or intensity) for time j (e.g. 60 h), $\bar{O}_j$ is the observed mean increment at time j, a_{ij} are the regression coefficients for the ith member model at time j, F_{ij} is the ith member model forecast increment at time j, $\bar{F}_{ij}$ is the member model forecast mean increment at time j during the training period, and n is the number of member models used. The increments are the difference in latitude, longitude and intensity values from the last forecast hour. The term 'increment' used here should not be confused with the 'increment' used for data assimilation purposes. The time-mean values are obtained from the training period. In the training phase the member model forecasts are regressed against the corresponding observed values to obtain the regression coefficients (or weights) for each individual model. The regression coefficients are obtained through the least square minimization technique (Stefanova and Krishnamurti, 2002). This is done independently for the 12-h latitude increment, the 12-h longitude increment and the 12-h intensity increment. For example to calculate the latitudinal/longitudinal/intensity increment at 48 h, the difference between the 36-h latitude/longitude/intensity and the 48-h latitudinal/longitudinal/intensity value is taken. The observed mean increment $\bar{O}$ values are calculated from the 'final best track' or 'working/operational best track' (during the season) that is obtained from the NHC's post-storm analysis. The training period at the start of the season is usually the forecast cases from the past season. To summarize, the superensemble tends to correct the biases and to assign higher weight to a better performing model while assigning less weight to a worse-performing member model. The superensemble technique differs from the simple ensemble mean consensus in that gives equal weights (a_{ij}) to all member models, does not consider bias, and does not use the increment approach.

The member models used in the training and forecast phase for the construction of the superensemble are all 'early' models, meaning those models which are available to the forecasters in time to meet forecast deadlines, that is, 0300 UTC, 0900 UTC, 1500 UTC and 2100 UTC. The term 'early' or 'late' models is determined by the availability of a particular model to the forecaster during the synoptic times given above. All dynamic models take a few hours to integrate and hence are not available to the forecaster at the initial forecast time. They are generally available after the forecaster has done the forecast and hence called a 'late model'. The late-model forecasts come from the latest available cycle of a model and are smoothed and shifted in time to apply to the forecast synoptic time, for example, the 0000 UTC GFS tracker is not available until 0500 UTC, so to make a late-model forecast starting at 0600 UTC (due at 0900 UTC), the 0000 UTC run is interpolated so that the $t = 0$ at 0600 UTC is the $t = 6$ forecast from 0000 UTC. This version of a particular model becomes the primary guidance of the forecaster at that particular synoptic time and for historical reason this new version is called 'interpolated' models. Franklin (2005a, 2006) and Rhome (2007) provide further details of 'early' and 'late' models.

The model forecasts are obtained from the 'A DECKS' contained in the NHC ATCF (http://www.nrlmry.navy.mil/atcf_web/index.html) data set. The FSSE forecast usually takes 10 minutes to run and depends on the length of the training data set. Since we use GUNA as a member model, if one member model is missing, automatically we

Table 2. Member models used for the construction of the superensemble

Member models for tracks	Member models for intensity
OFCI	OFCI
UKMI	UKMI
NGPI	GFSI
GFDI	SHF5
GFSI	DSHP
GUNA	

lose another one. But if three or more models are missing, the forecast for that particular hour is not performed. Thus the availability of FSSE is restricted, we have experienced some bad superensemble based forecasts when we used two or three member models as input.

The suite ensemble of models/forecast aids used in the construction of the superensemble includes those that already provide real time guidance to the NHC and the official forecast, interpolated to the forecasted synoptic time (OFCI). The member models for the construction of the superensemble are provided in Table 2. Included among the member models are some specific statistical models that were tailored for improving hurricane intensity. We have found that inclusion of OFCI in the suite of member models (forecast aids) carried a demonstrable positive impact on the performance of the superensemble. It was noted that if one excludes the superensemble and examines the performance of the member models and the OFCI, one finds that the OFCI does outperform most models on a regular basis. This was noted during the 2004, 2005 and the 2006 seasons (Franklin, 2005a, 2006, 2007). The consistent superior performance of the OFCI does rank it among the best models and for that reason we had simply included it as a member model. It is worth noting that the verification done by NHC includes the OFCL and not OFCI, which is used by FSU superensemble as one of its member models. However, the important message that emerges after its inclusion is that the superensemble is able to make use of the skill of OFCI and outperform it on most occasions (shown later).

3. Walkthrough of a superensemble forecast

The latitude, longitude and intensity forecasts are computed separately while doing a superensemble forecast. The mathematics of a superensemble forecast for longitude is illustrated in Table 3. Here we show how the superensemble is constructed for a 72-h longitudinal forecast for Ivan. S_j denotes the superensemble longitudinal increment from 60 to 72 h and $\overline{O}$ is the mean of all observed longitudinal increments between those hours 60 and 72 from the observed best tracks in the training period. $\bar{F}_j$ denotes the average longitudinal increment for a single forecast model for hours 60–72 calculated from the training period and

Table 3. Walkthrough a superensemble forecast for longitude

$$S_j = \bar{O}_j + \sum_{i=1}^{n} a_{ij}(F_{ij} - \bar{F}_i)$$

Observed mean increment $\overline{O} = 0.44$
S_j (j: 60 → 72 h) = 0.44 + 0.45(1.10 – (0.38)) [Model 1]
+ −0.07(0.40–0.62) [Model 2]
+ 0.08[1.00–(0.55)] [Model 3]
+ 0.40[1.20–(0.41)] [Model 4]
+ 0.26[0.90–(0.45)] [Model 5]
S (60 → 72 h) = 1.24 [Superensemble increment from 60 → 72 h]
S (at 72 h) = Superensemble forecast at 60 h + S (60 → 72 h)
Superensemble forecast at 72 h = 79.30 + 1.24 = 80.54°W
Observed (at 72 h) = 80.40°W
Ensemble mean = 79.66°W

Table 4. Walkthrough example of superensemble for latitude

$$S_j = \bar{O}_j + \sum_{i=1}^{n} a_{ij}(F_{ij} - \bar{F}_i)$$

Observed mean increment $\overline{O} = 1.34$
S_j (j: 84 → 96 h) = 1.34 + 0.27(0.80–1.3) [Model 1]
+ 0.53(0.80–1.21) [Model 2]
+ 0.18(1.40–1.52) [Model 3]
+ 0.14(2.4–1.47) [Model 4]
+ −0.1(1.3–1.42) [Model 5]
S (108 → 120 h) = 1.10 [Superensemble increment from 108 → 120 h]
S (at 120 h) = Superensemble forecast at 108 h + S (108 → 120 h)
Superensemble forecast at 120 h = 25.53 + 1.10 = 26.63°N
Observed (at 120 h) = 27.10°N

i denotes a tag for a member model. F_{ij} is the instantaneous forecast value of the longitudinal increment (60→72 h) for the ith model. The entries below the equation in Table 3 provide the contributions for the summation from the respective models. We have included five models for this illustration of the superensemble methodology. The superensemble forecast for the longitudinal increment came to –1.24°. When that is added to the 60-h superensemble forecasted value of –79.30° (79.3°W), it provided a forecast of –80.54° (80.54°W) for the predicted longitude for Hurricane Ivan at 72 h. That verifies very closely to the observed best track longitude of 80.40°W. The ensemble mean consensus fares poorly in comparison. A similar example is provided for latitudinal forecasts in Table 4. This shows a superensemble latitude forecast for 96 h. The same exercise was also carried out for the intensity forecasts shown in Table 5 for the intensity superensemble. The 108-h forecast for Rita valid 0600 UTC 19 September 2005 is shown in Table 5. The superensemble forecast was the best among the member models and the ensemble mean, which was a very typical outcome for the 2005 season.

The availability of the superensemble is bit limited in the sense that we start making forecasts when a system reaches tropical storm intensity. And though the cases may be limited

Table 5. Walkthrough example of superensemble for intensity (initialized at 0600 UTC 19 September 2005)

$S_j = \bar{O}_j + \sum_{i=1}^{n} a_{ij}(F_{ij} - \bar{F}_i)$

96 h lead time with initial time 19 September 2005 0600 UTC

Models	Coef. α_i	$F_i(t)$	$\bar{F}_i$
Model 1	0.62	0.0	−3.84
Model 2	0.69	2.3	−0.99
Model 3	−0.43	2.3	0.27
Model 4	0.19	−4.6	−2.02
Model 5	0.05	−2.3	−8.27

Notes: $\bar{O} = -3.93$; S (84 → 96) $= \bar{O} + \alpha_i(F_i(t) - \bar{F}) = -3.93 + 3.58 = -0.35$; S (at 96 h) = Superensemble forecast at 84 h + S (84 → 96) = 130.36 − 0.35 = 130.01 mph = 57.2 m s^{-1}; observed: 138.00 mph = 60.72 m s^{-1}; ensemble mean: 89.47 mph = 39.36 m s^{-1}; superensemble: 130.01 mph = 57.2 m s^{-1}.

Table 6. Forecast availability of FSSE with respect to OFCI and GUNA

Year	No of forecasts for OFCI	No of forecasts for GUNA	No of forecasts for FSSE	No of 12 h cases selected for Annual Summary (NHC report)
2004	434	363	323	294
2005	588	519	453	383

to a mere couple, we do not make a forecast if there are only two models present at a particular time. The availability of FSSE with OFCI and GUNA is included in the following Table 6. This contains all forecasts made during 2004 and 2005 season on real time. The OFCI and GUNA forecasts are more because it contains the INVEST cases done for a particular named storm or a depression. But still FSSE has enough number of cases comparable to OFCI and GUNA.

4. Mean track errors

The verification presented here for the FSSE and other models does not include the extratropical stage and non-tropical. The system must be a tropical (depression, storm and hurricane) at the initial and verification time including depressions and covers even after it makes landfall. These are in accordance with the NHC verification conventions.

The mean track errors for different models including the superensemble for the entire 2004 hurricane season are shown in Fig. 2a. Among eight entries, the two on the far right are the forecasts from the ensemble mean consensus of the member models (Table 2) and the superensemble. There were 300 entries for 12-h forecasts, which drops to 129 entries for 120-h forecasts since several models terminated their forecasts prior to day 5 if the tropical cyclone dissipates in the model forecast or in the observations. The multimodel superensemble has the smallest errors for the track forecast through day 5 of the forecasts (Fig. 2a). It reduces the errors by as much as 39 km for 120 h forecasts compared to the best model. Among the member models, the forecasts from the GUNA were somewhat superior compared to all others in this suite. The ensemble mean of the member models was better than member models, even beating the GUNA ensemble at 12–120 h except 72 and 84 h. The forecast errors of the ensemble mean suffer from the fact that the statistical weights of the poorer models and the best model are all assigned the same weight. The superensemble is more selective in this regard since it assigns fractional positive or even negative weights. In the context of the multiple regressions, the negative weights imply negative correlations between the observed and the predicted values (of track and intensity) during the training phase (Chakraborty et al., 2007). In the construction of the superensemble these negative weights provide the means to correct for models that perform opposite to those of the observed changes.

The track errors for superensemble during 2005 (Fig. 2c) were the smallest among the models for 12–48 h of the forecasts but were ranked second or third for 72, 96, 108 and 120 h forecasts from the superensemble. GUNA performed better than the FSU superensemble after 60-h forecasts. After a lead-time of 120 h, the UKMI was the only dynamic model that performed as well as GUNA.

The track errors for the 2006 forecasts for the superensemble are shown in Fig. 2e. The superensemble had the least error except for 12- and 24-h forecasts where GUNA was slightly better. The consensus aids performed better than the dynamic models at all forecast times.

The track errors for the 2007 and 2008 seasons are shown in Figs 2g and i, respectively. The figures shows that the FSU superensemble was able to outperform all the member models and the ensemble mean at all forecast times. Among the dynamic models GFDI performed better than GFSI, UKMI and NGPI for the 2008 season.

5. Mean absolute intensity errors

The mean absolute intensity errors for the 2004 season are shown in Fig. 2b. Here the lowest errors, at all forecast hours, are for the multimodel superensemble. There has not been much improvement in the NHC's intensity forecasts during the last decade (see Fig. 1b) but the superensemble results from 2004, 2005 and 2006 seasons shown in Figs 2b, d and f, respectively seem to suggest that some improvements are possible with respect to the best member model. Many of these are large-scale models and further improvements in intensity forecasts will most likely come from the deployment of a suite of mesoscale models, as suggested in the work of Krishnamurti et al. (2010) and Cartwright (2004). The FSU superensemble tends to perform

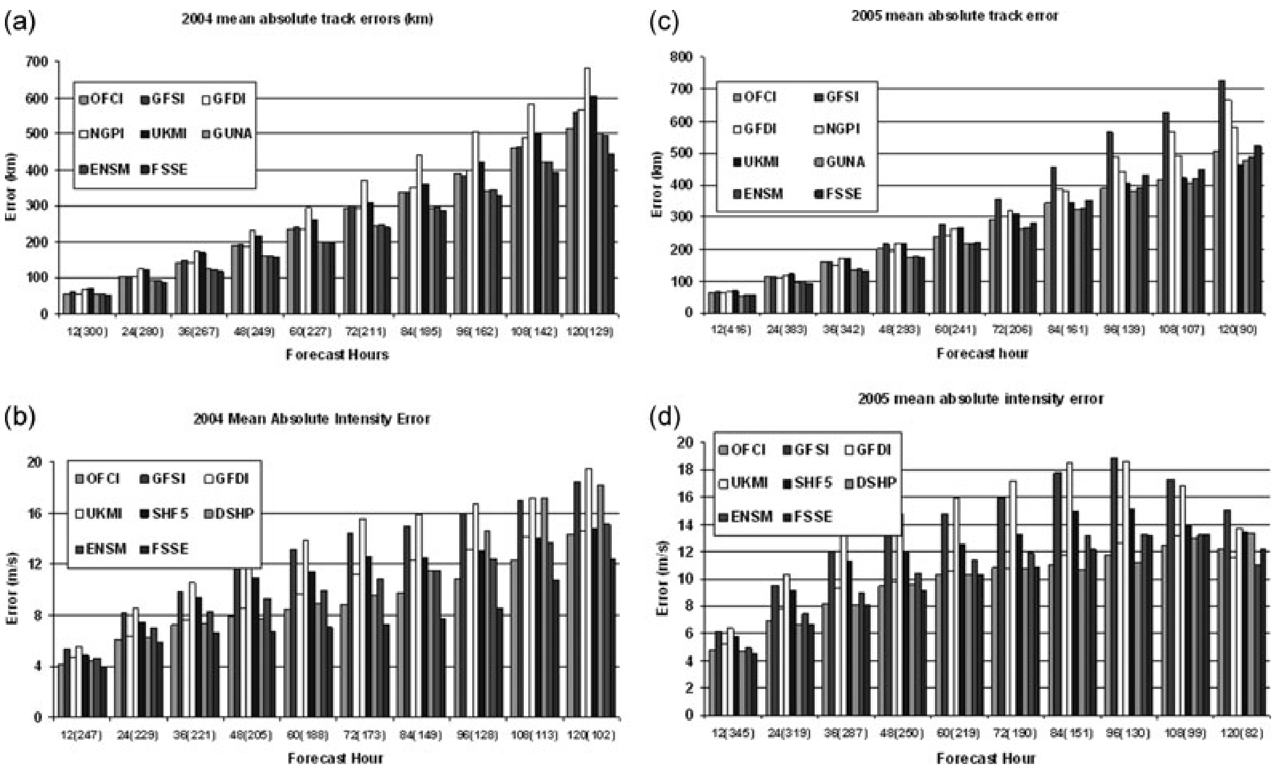


Fig. 2. Homogeneous comparisons of (a) Mean absolute track errors (in km) for 2004 season. (b) Mean absolute intensity errors (in ms^{-1}) for 2004. (c) Mean absolute track errors (in km) for 2005. (d) Mean absolute intensity errors (in ms^{-1}) for 2005. (e) Mean absolute track errors (in km) for 2006. (f) Mean absolute intensity errors (in ms^{-1}) for 2006. (g) Mean absolute track errors (in km) for 2007. (h) Mean absolute intensity errors (in ms^{-1}) for 2007. (i) Mean absolute track errors (in km) for 2008. (j) Mean absolute intensity errors (in ms^{-1}) for 2008. The number within the parentheses in the abscissa shows the number of cases for that particular forecast hour.

better for stronger systems compared to the weaker ones (not shown here). A large percentage improvement is seen for storms greater than Category 2 storms compared to the best model.

The corresponding intensity forecast errors of different models for the year 2005 are presented in Fig. 2d. The intensity errors for the superensemble during 2005 (Fig. 2d) were lowest at some forecast lead times, with respect to the participating models, except for 72–108 h of the forecasts where DSHP (a statistical model) was better than the superensemble. However, the 120 h intensity error for the 2005 season was more than the ensemble mean of the member models. This may be due to the fact that the training period of the 2005 season was affected by model changes that took place during the 2005 hurricane season. Several storms were difficult to forecast and the dynamic models all had widely divergent tracks for many forecast times, which affects the relative performance of ensembles derived from them. Starting 2005 the NHC introduced a consensus forecast ICON (Franklin, 2006; Sampson et al., 2008), which is a simple ensemble mean of various intensity models. It contains some of the best intensity models. The member models of ICON are reviewed each year by the NHC and changed accordingly. During 2005 it outperformed all of the dynamic and statistical model forecasts except at the 12-h lead time. The performance of ICON is not presented here since it was not computed operationally in real time and was not a part of the superensemble suite.

The 2006 intensity forecast errors for the superensemble along with other models is shown in Fig. 2f. The errors for the superensemble were lowest for all forecast times except for 120-h forecasts where SHF5 carried the best results. Among the dynamic models GFDI performed reasonably well during the 2006 season. The large-scale models still had problems in predicting the intensity. The statistical models have less mean absolute errors compared to the dynamic models (Fig. 2f).

The 2007 and 2008 mean absolute intensity errors are shown in Figs 2h and j, respectively. The forecast error reduced to approximately 6 ms^{-1} at 120 h for the 2007 season. Significant error reduction was noted for the 2008 season where the errors of the member models grew to more than 15 ms^{-1} for the 5-d forecast. The FSU superensemble was able to correct the model biases and produce a consistent forecast compared to the member models and their ensemble mean.

6. Storm-by-storm performance

A few illustrations on forecasts for specific storms are provided here. The choice of these illustrations was not intended to show

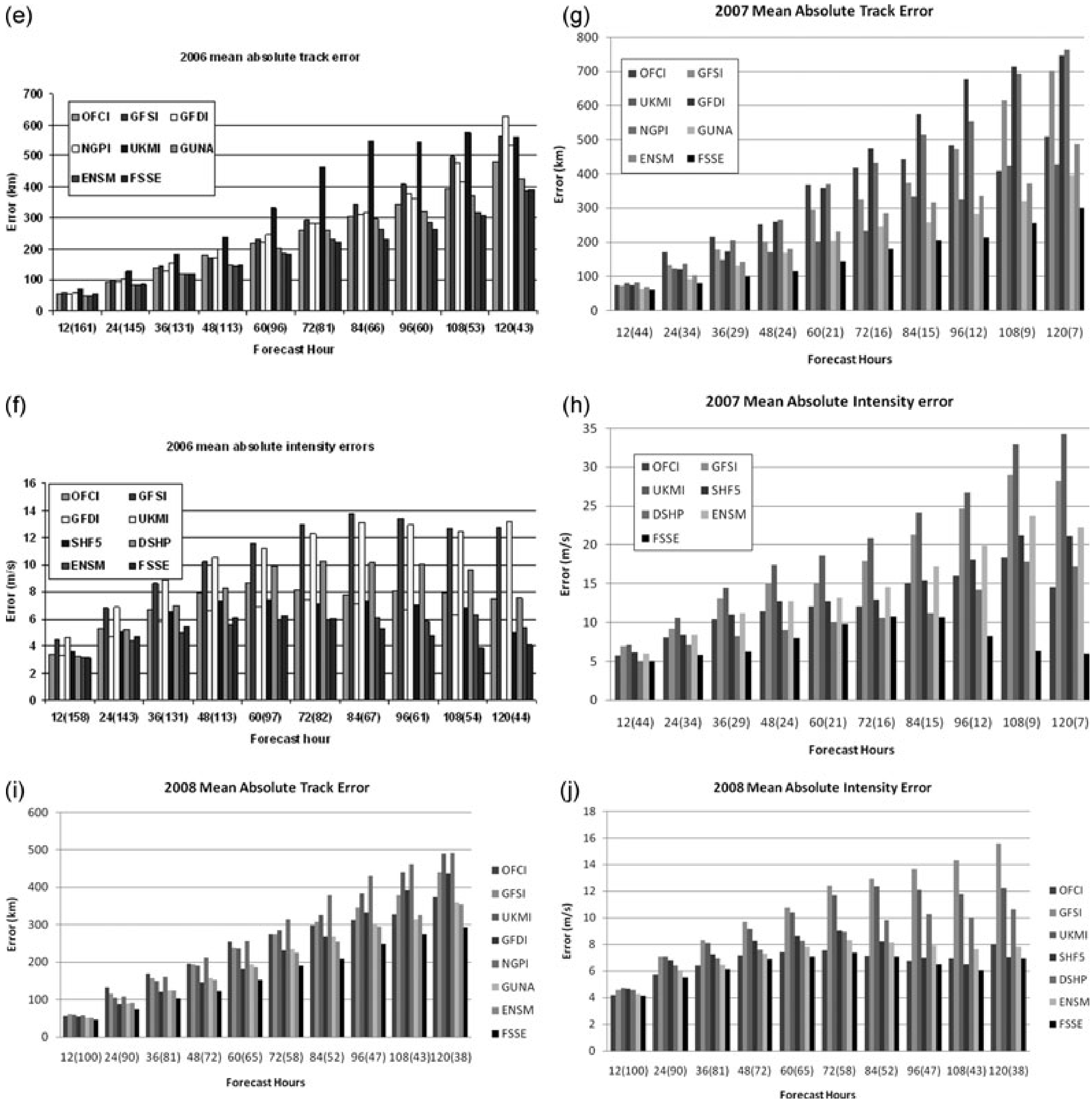


Fig. 2. Continued.

the very best results; the results show how the superensemble performed for prominent storms at different forecast times.

6.1. The 2004 season

In Fig. 3a we show the track errors for 72-h forecasts for some of the major storms of the 2004 season. The large variations in the number of cases verified for a particular storm depends on the model availability and the lifecycle of the storm. The superensemble performed better than other models for the storms Frances and Ivan. GUNA forecast errors are slightly less than the superensemble for Jeanne. For Charley the GFDI had the smallest errors at 72 h, the superensemble performed better than GUNA and had similar errors as GFDI.

Several of these models carried their forecasts out to 120 h. Those track errors are shown in Fig. 3b. The 120-h forecast errors again confirm the overall strength of the multimodel superensemble. The 120 h mean absolute track errors range from 250 km to as high as 830 km for various storms of the 2004

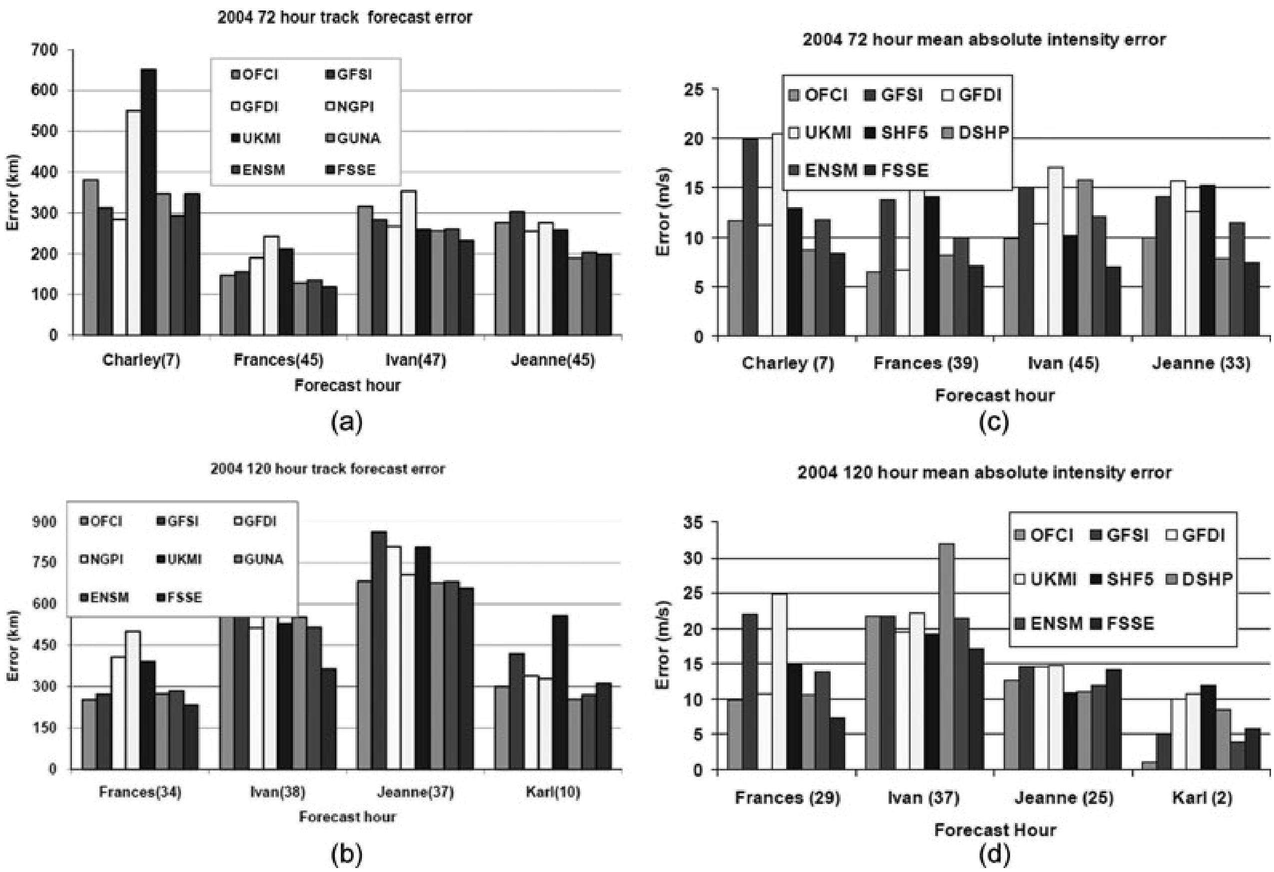


Fig. 3. Mean absolute track errors in km for (a) 72-h forecast for Hurricanes Alex, Charley, Danielle, Frances, Ivan, and Jeanne (in km) and (b) 120-h forecast for Hurricanes Frances, Ivan, Jeanne and Karl (in km). (c) Same as Fig. 4(a) except for intensity (in ms^{-1}). (d) Same as Fig. 4(b) except for intensity (in ms^{-1}). The number within the parentheses in the abscissa shows the number of cases for that particular storm.

season. The large forecast errors at a lead-time of 120 h for Jeanne (Fig. 3b) relate largely to its post-landfall phase when the storm was affected by steering due to fast west–southwest upper level winds. Those forecast errors provided a large dispersion of the tracks with larger track errors. An exception was Karl where at 120 h, the track errors for the superensemble were larger than those of GUNA, GUNS and ENSM.

The member models for intensity forecasts include a few different models from those used for track forecasts. These are listed in Table 2. The results on intensity forecasts error for the major storms of the 2004 season are summarized in Fig. 3c (for 72 h) and Fig. 3d (for 120 h). Although there was a consistent improvement in the forecasts from the superensemble, there were a few exceptions at particular lead times. When we examined each of the storms, we note that the performance of the superensemble was always among the top three models for 72-h forecasts (Fig. 3c). The superensemble had the smallest errors for Charley, Ivan and Jeanne. The performance of the superensemble was best for Frances and Ivan for intensity forecasts at 120-h lead time as shown in Fig. 3d, the errors were however larger for Jeanne and Karl.

6.2. *Right or left bias for model forecasts*

In Figs 4–7 several track forecast initialized every 6 h and show positions at 12-h intervals are shown for several models. Within these sets of illustrations we show a collection of forecast tracks from the superensemble, the OFCI, the best performing member model and the worst performing member model. This sequence of forecasts also includes the official best track from the NHC. What stands out in these forecasts is how a strong left or right bias from one forecast to the next for many models is minimized by the superensemble technique.

Except for two initial forecasts for Charley, the superensemble was consistent in predicting the landfall position (Fig. 4d). The track biases were reduced by the superensemble methodology, biases that were clearly seen in OFCI, GFSI and the UKMI forecasts.

A large right bias in forecasts was noted for several member models during forecasts of hurricane Frances (Figs 5a–d). Ensemble averaging, such as those implicit in GUNA does not remove a consistent track bias error. The superensemble recognizes such features of forecasts from consistent errors

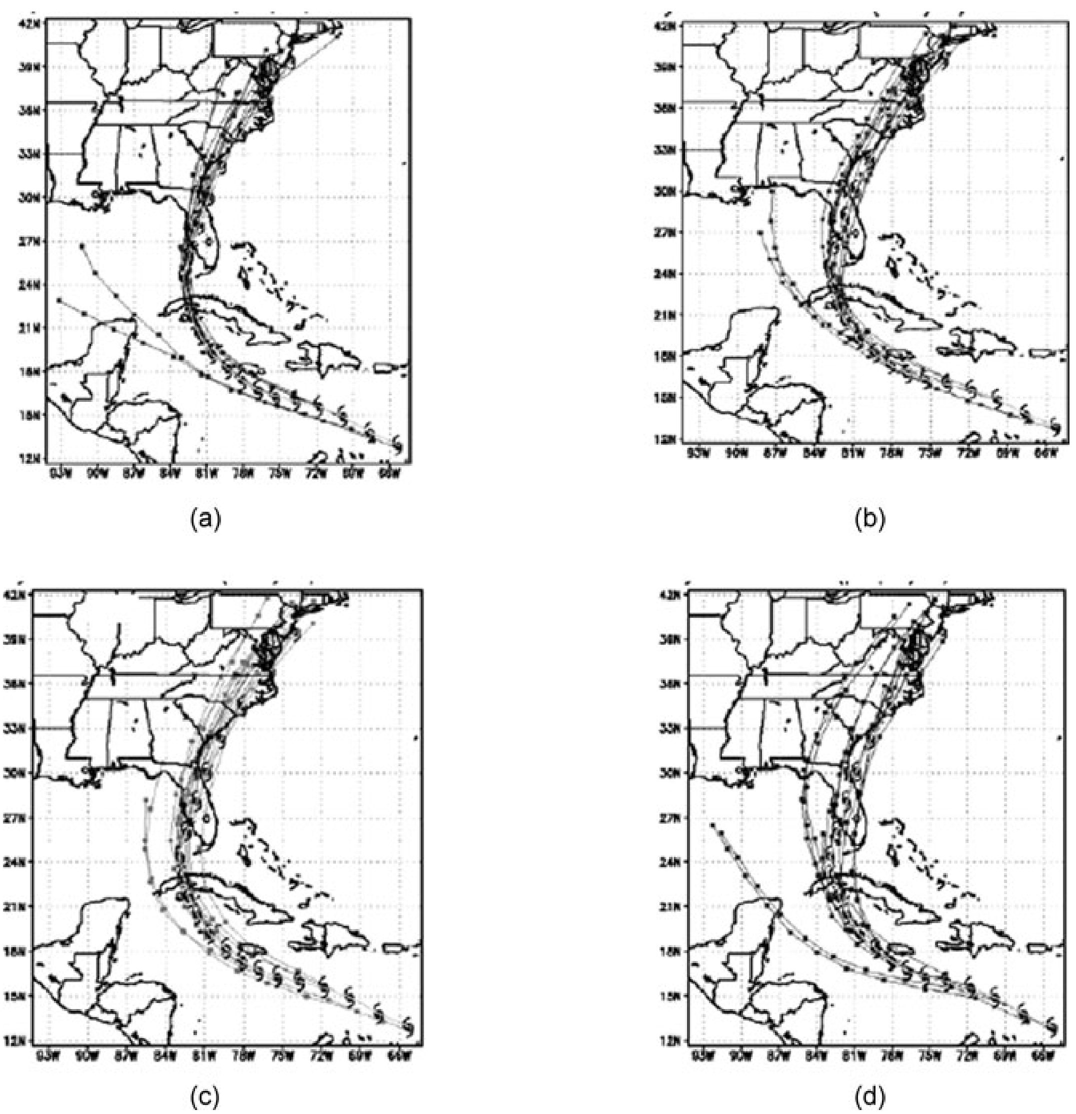


Fig. 4. Successive 6-hourly forecast tracks for the (a) FSU superensemble, (b) OFCI, (c) GFSI and (d) UKMI models for Hurricane Charley of 2004. The observed best track is shown by the yellow line with hurricane symbols representing positions plotted every 12 h. Mean absolute errors (in km) are shown in the inset along with the forecast hour.

during its training phase and does correct it to a large extent. The superensemble was able to reduce this type of bias drastically by removing the individual biases of the member models.

The forecasts for Ivan for OFCI, UKMI and NGPI, Figs 6a–d, show the track biases; this right bias was especially pronounced during its recurvature over the eastern Gulf of Mexico. The bias was clearly much less for the multimodel superensemble. The superensemble had the lowest errors for the track forecasts for Ivan by reducing the right bias from one forecast to the next. The superensemble forecasts were consistent throughout the storm lifecycle.

One of the most difficult storms to forecast was Jeanne, since it had a looping track. Forecast errors of member models for 120 h forecasts were as large as 808 km (Fig. 3b). A succession of forecasts from the superensemble, OFCI, GUNA and the GFSI model are shown in Figs 7a–d. As is evident, large track biases led to large errors in several forecasts. Most models failed to capture the looping track. The superensemble had the smallest errors for this storm. The mean absolute error for the superensemble for 84 h forecast was 262 km and those for other models were larger and are shown within the figures Figs 7a–d.

The track biases for 120 h forecasts for Frances, Ivan and Jeanne (Franklin, 2005b) of the member models and those of the superensemble are shown in Figs 8a–c. Smaller biases are shown to be closer to the centre of the polar coordinate diagram. As is seen from the figure, the track biases were the smallest for

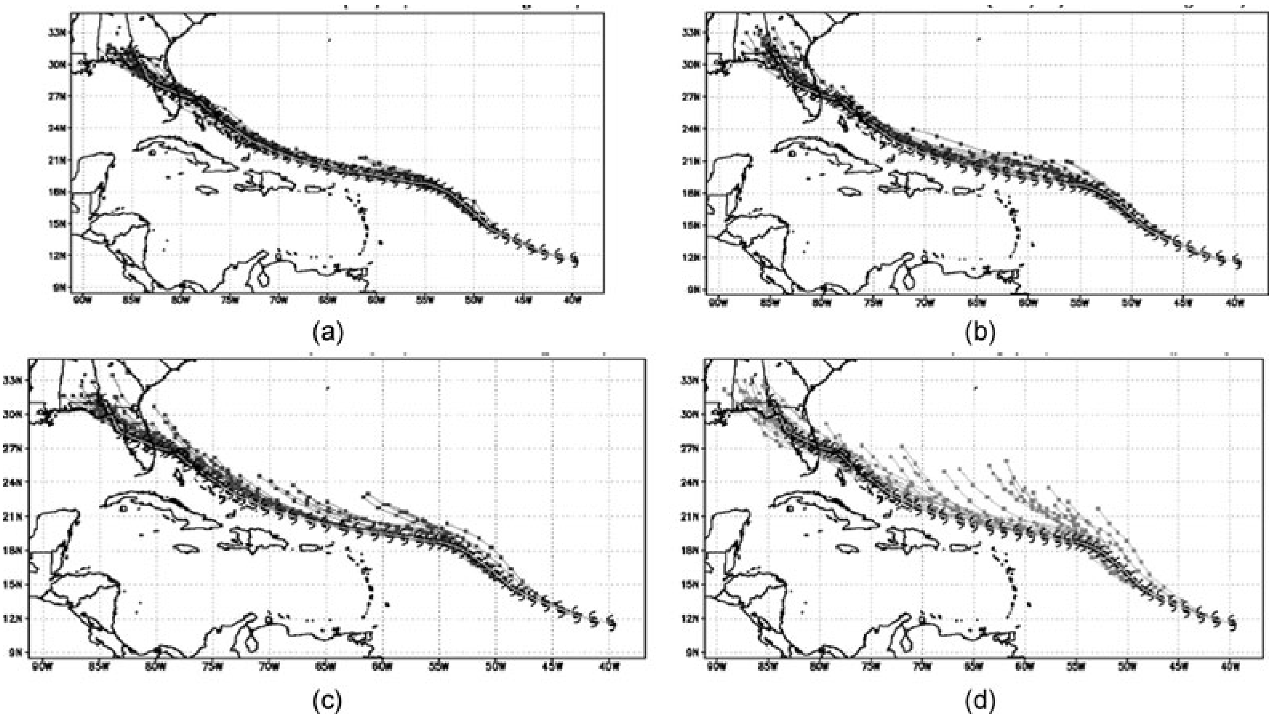


Fig. 5. As in Figure 5 except for the (a) FSU superensemble, (b) OFCI, (c) GUNA and (d) UKMI models for Hurricane Frances (2004).

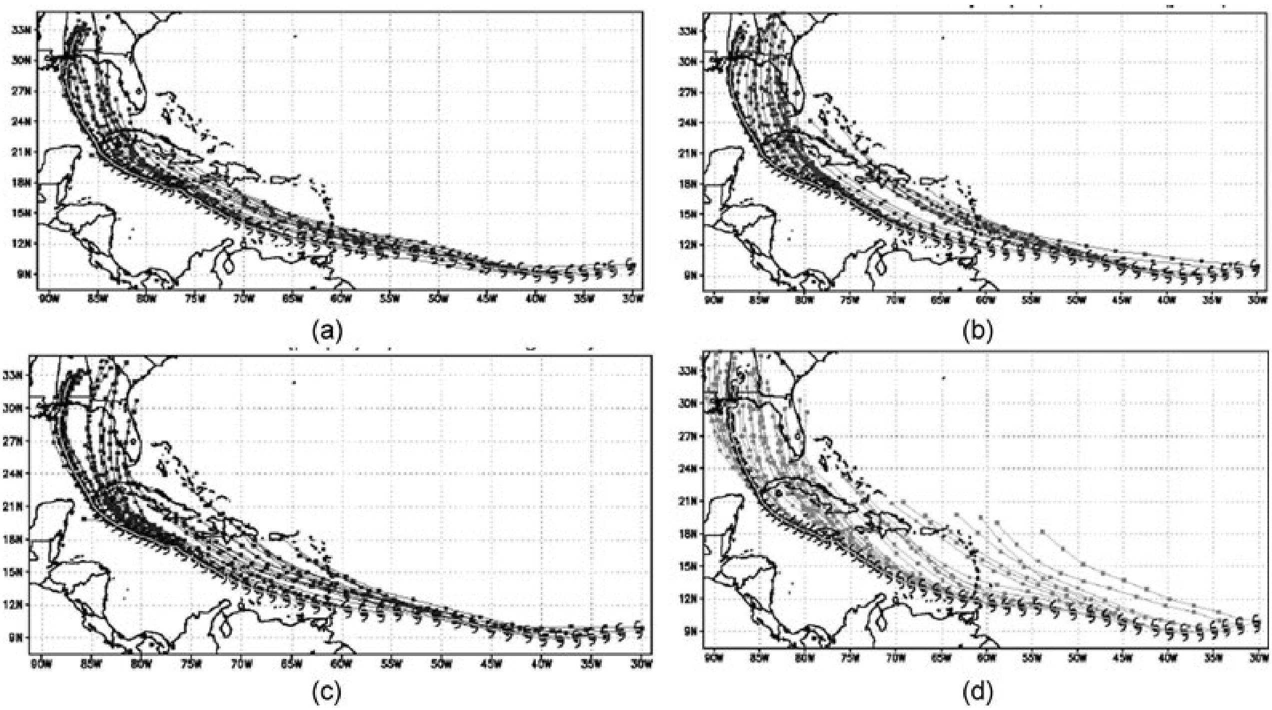


Fig. 6. As in Figure 5 except for the (a) FSU superensemble, (b) OFCI, (c) UKMI and (d) NGPI models for Hurricane Ivan (2004).

hurricanes Frances, Ivan and Jeanne during the 2004 season for the superensemble. This figure shows that even if the member models have a specific bias the superensemble is able to detect this bias and reduce it.

6.3. Is superensemble adding value to the forecast?

A Student's *t*-test is done to show that indeed the superensemble forecasts (combined for 2004, 2005 and 2006) are adding value

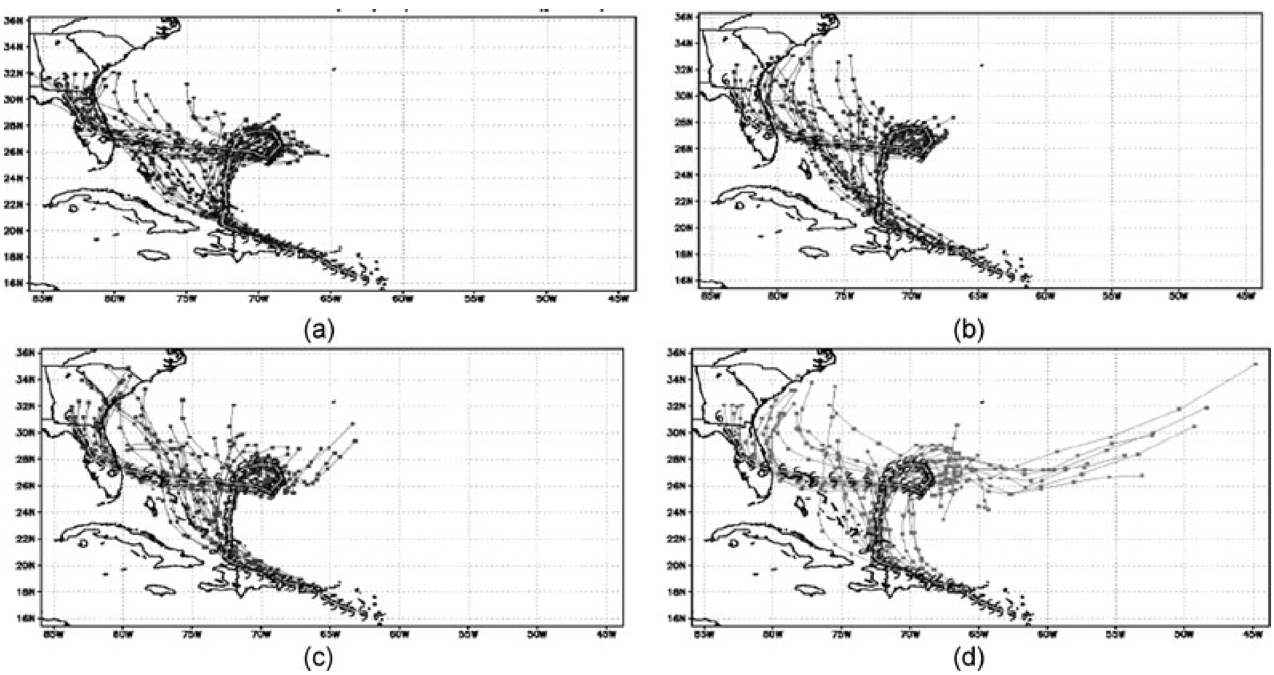


Fig. 7. As in Figure 5 except for the (a) FSU superensemble, (b) OFCI, (c) GUNA and (d) GFSI models for Hurricane Jeanne (2004).

and not only reducing the errors among the member models. Figure 9 shows the t-test results for track and intensity errors. Figure 9 show that at all the forecast lead times, the superensemble track and intensity forecasts are statistically significant. The confidence levels are shown for each forecast times. The significance level of the superensemble forecast with respect to the ensemble mean forecast was calculated using a Student's t-test. The null hypothesis is expressed by

$$H() = F_{\rm em} \neq F_{\rm se},$$

where $F_{\rm em}$ and $F_{\rm se}$ are ensemble mean and superensemble forecasts.

To calculate the t-value, the mean and the standard deviation of the forecasts of the ensemble members were considered along with the superensemble forecast

$$t = \frac{(F_{\rm em} - F_{\rm se})}{\overline{S_d}},$$

where $\overline{S_d}$ is $\frac{S_d}{\sqrt{n}}$, and S_d is the standard deviation of the model ensemble forecasts and n denotes the number of models. The significance level was calculated using the above equation and a Student's t-table. This calculation is similar to that described in Chakraborty et al. (2007).

7. The major storms of the 2005 season: Katrina, Rita and Wilma

Katrina was one of the greatest natural disasters in United States history. Around 1833 fatalities have been reported and most of them were due to storm surge flooding in the New Orleans area (Knabb et al., 2006a) when artificial levees designed to protect the city failed. Rita became a Category 5 hurricane from a tropical storm in less than 36 h with the fourth lowest sea level pressure ever recorded in the Atlantic basin at 895 hPa (Knabb et al., 2006b). Both Katrina and Rita had immense social, economic and environmental impacts. Oil production and refineries were badly affected during hurricanes Katrina and Rita (Knabb et al., 2006a,b). Wilma also had devastating effects on the residents of the Florida Keys, and was an exceptionally difficult storm to forecast (Pasch et al., 2006), including observed 24-h intensity changes that set records for the Atlantic basin, and a record lowest-ever sea level pressure estimate of 882 hPa. Models were generally in agreement about a south Florida landfall, but erred in particular in terms of their estimates of forward motion of the storm.

Early on, a number of model forecasts including GFSI and NGPI were calling for Katrina to move over the eastern Gulf of Mexico, with landfall at Apalachee Bay, Florida on 29 August. For Katrina it was difficult for most models including superensemble to accurately predict the 96- and 120-h positions leading up to landfall. The GFDI model was one among the suite of models that provided a more accurate guidance for the 96 h forecasts. It is difficult to assess what data or other common modelling problems were present within a number of models that failed at these longer ranges (96 and 120 h). Figures 10a and b shows 120 and 84 h forecasts for Katrina. Most models were able to forecast the landfall position for Katrina in their 60 h forecasts. The multimodel intensity forecasts initialized at 1800 UTC 26 August for 60 h leading up to landfall are shown

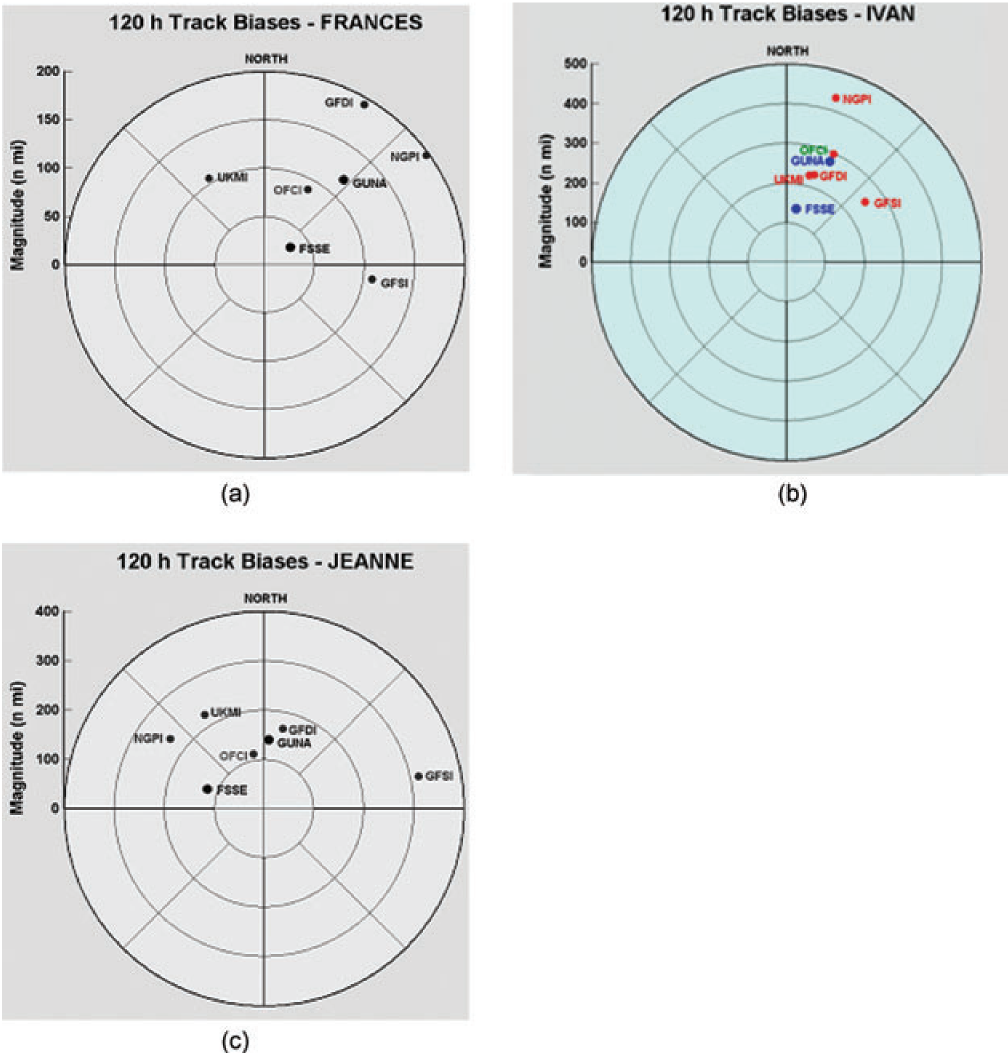


Fig. 8. 120-h track biases for the 2004 storms (a) Frances, (b) Ivan and (c) Jeanne. The polar diagram shows the magnitude and direction of biases of the member models in nm. The mean model errors are also plotted in the figure (Adapted from IHC presentation by James Franklin 2005).

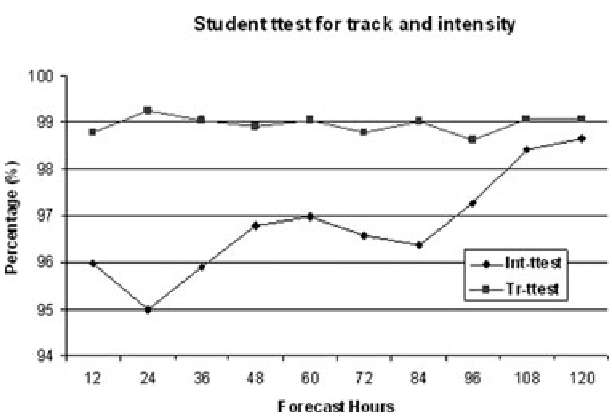


Fig. 9. FSU superensemble Student's *t*-test for (a) track errors and (b) intensity errors. The Tr-ttest shows the confidence level for track forecasts and the in Int-ttest shows the confidence level for intensity forecasts.

in Fig. 10c. The intensification of Katrina (up to category 4 status) prior to its landfall was well predicted by the multimodel superensemble. That forecast stood out in the sense that most of the member models underestimated the intensity of Katrina during the landfall period. However, it is to be noted that none of the member models nor the superensemble predicted the ultimate intensification up to category 5 of Katrina prior to its weakening just before landfall.

Track forecasts for Rita encountered problems with track shifts over time, as was the case with Katrina. Longer-range 120 h forecasts before landfall showed a westward shift for the member models that called for a landfall over the central Texas coast. An abrupt eastward shift in the forecasts towards the Mississippi coast was seen for the 72 h forecasts for landfall.

In Fig. 11, mean absolute errors for intensity are shown as a function of forecast hour for Rita. These intensity forecasts

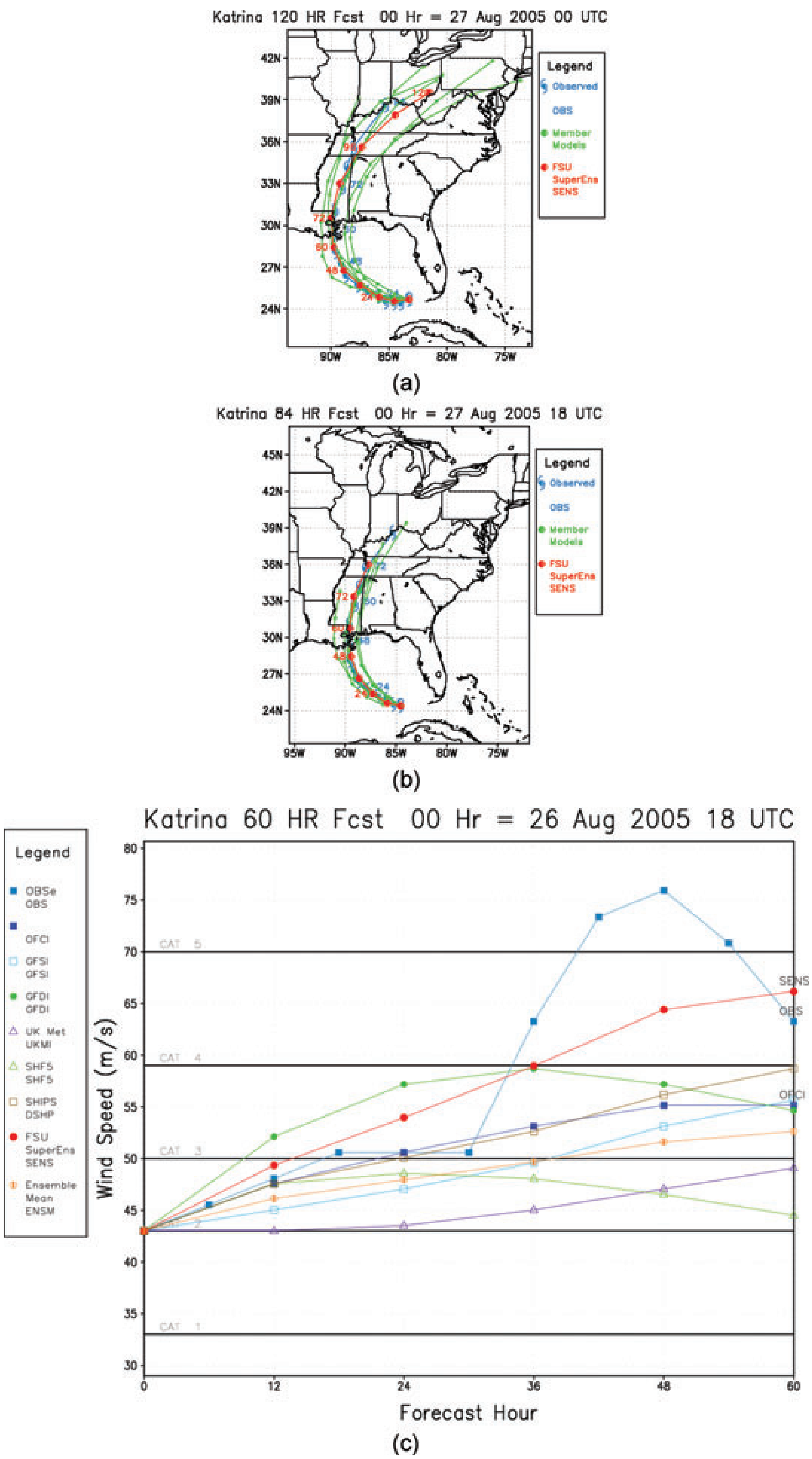


Fig. 10. (a) 120-h track forecasts for Hurricane Katrina with initial time at 27 August 2005 00 UTC. (b) 84-h track forecasts for hurricane Katrina with initial time at 27 August 2005 18 UTC. The red, blue and green lines show the superensemble track, the observed best track and the member model forecast tracks at 12-h intervals. (c) 60-h intensity forecasts for Katrina at the initial time 26 August 2005 18 UTC. The red line indicates the superensemble forecast and the blue line denotes the observed intensity at 6-h intervals.

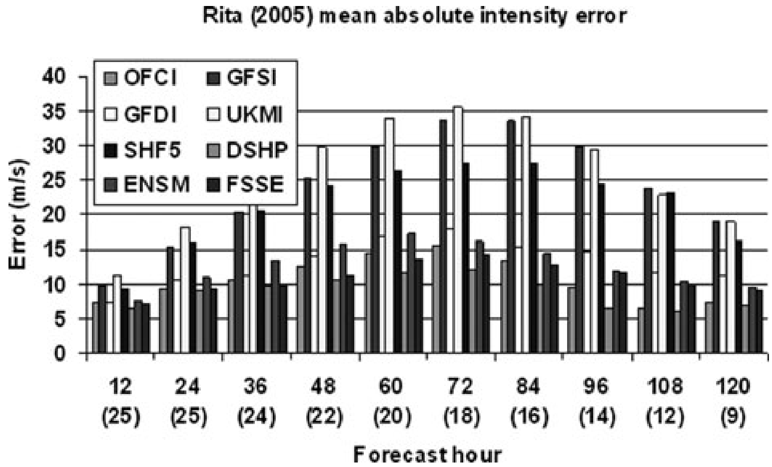


Fig. 11. Mean absolute intensity errors (in m s^{-1}) for Hurricane Rita.

for Rita from the multimodel superensemble were in fact more quite impressive. Although the absolute intensity error of the superensemble increased between 12 h (9.5 m s^{-1}) and 72 h (14 m s^{-1}), thereafter the superensemble forecast errors decreased drastically with forecast errors of only 9 m s^{-1} for the 120 h forecast (Fig. 11). Some of the well-known models had rather large errors (15–20 m s^{-1}) between 72 and 120 h. The intensity errors of the superensemble were tied with DSHP during 24- and 36-h forecasts.

Figure 12a illustrates a representative example of landfall forecasts initialized at 1800 UTC 21 October for Wilma. The spread of the forecast tracks from the participating models was very small. Steering by strong winds in the middle and upper troposphere appear to have minimized the large spread of the tracks although the same strong winds caused timing errors.

The corresponding spread for the intensity forecasts for Wilma is shown in Fig. 12b. The intensity reached Category 5 (close to 82 m s^{-1}) on October 19 at 1200 UTC, 2005 (Pasch et al., 2006). Almost all models underestimated the intensity. High-resolution models may be required for improvements for predicting intense tropical cyclones (Bengtsson et al., 1995). The superensemble predicted a slightly higher intensity than all other models but it was still not close to the observed intensity. It was during this period that Wilma strengthened to a Category 5 hurricane from barely a Category 1 hurricane in just 24 h. It is to be noted that the ensemble mean of the member models failed to show the intensification thus showing the strength of the superensemble in correcting the biases of the individual models. The prediction of rapid intensification of tropical cyclones is one of the problems with low-resolution models (Park et al., 2009). Wilma's intensity and track forecasts created problems for the residents of south Florida who in general reported not being prepared for its intensity.

The mean errors for all of the track forecasts that were made for hurricane Wilma are shown in Fig. 13a. The forecast errors of the superensemble for 24-, 48- and 72-h forecasts are 40, 117 and 262 km, respectively. The superensemble had the smallest errors during the first 36 h of the track forecasts among all of the participating models. Between 48 and 84 h the UKMI, the consensus ensemble mean GUNA and the superensemble had the smallest track errors. A summary of the mean absolute intensity errors for all of the forecasts for hurricane Wilma is presented in Fig. 13b. The summary shows a consistent reduction of the intensity forecast errors for the superensemble compared to all of the other member models through the 96-h forecasts.

8. 2006 results

The 2006 season was more quite than was predicted with only 10 named storms and only two tropical storms Alberto and Ernesto made landfall in US. The damages from these two tropical storms were minimal.

The track errors for Ernesto are shown in Fig. 14a. The superensemble forecast was best from 36 to 108 h. At 120 h the superensemble ranked second to NGPI, which performed best among the member models. After 84-h NGPI was the best among the dynamic models and has less error than GUNA. A representative forecast track initialized at 0000 UTC 29 August 2006 is shown in Fig. 14b. The figure shows the superensemble track in red along with other member model forecasts and the observed. The superensemble was able to predict the 24-h first landfall near Adams Beach, Florida quite accurately. The timing was also fairly accurate.

Isaac was the tenth named storm of the 2006 season. A sample forecast track initialized at 1200 UTC 29 September 2006 is shown in Fig. 14c. The figure shows the left and right biases of the member models and how the superensemble (in red) corrects the individual biases of the models and provides a superior forecast. The track errors for the member models and the superensemble Isaac is shown in Fig. 14d. The superensemble had the least errors up to 60-h forecasts. The 72-h error was worse than GFDI whereas the superensemble ranked second. Here also the consensus models performed better than the individual

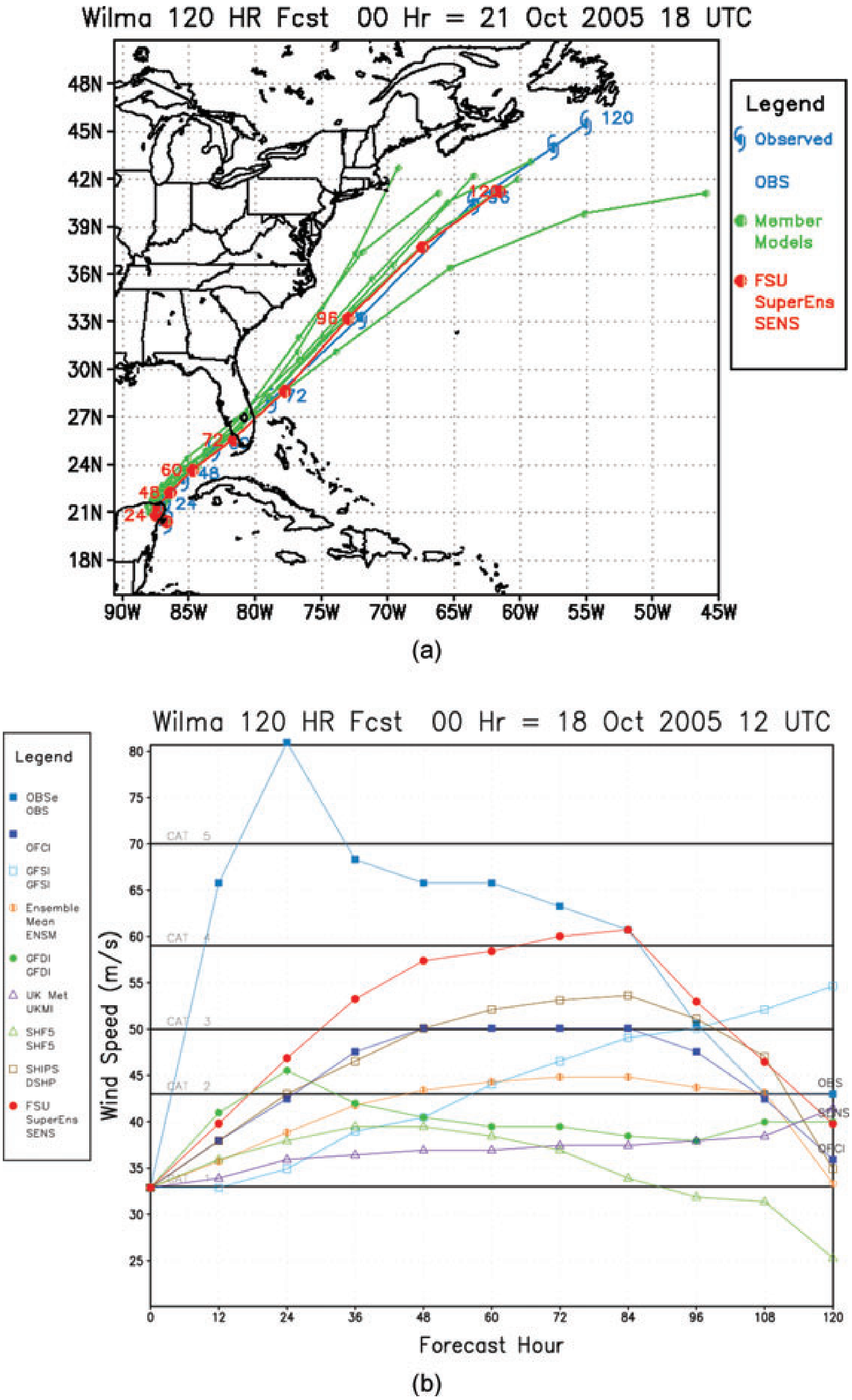


Fig. 12. (a) 120-h forecast for hurricane Wilma with initial time at 21 October 2005 18 UTC. The red, blue and green lines show the superensemble track, the observed best track and the member model forecast tracks at 12-h intervals. (b) 120-h intensity forecasts for Wilma at the initial time 18 October 2005 12 UTC. The red line indicates the superensemble forecast and the blue line denotes the observed at 12-h intervals.

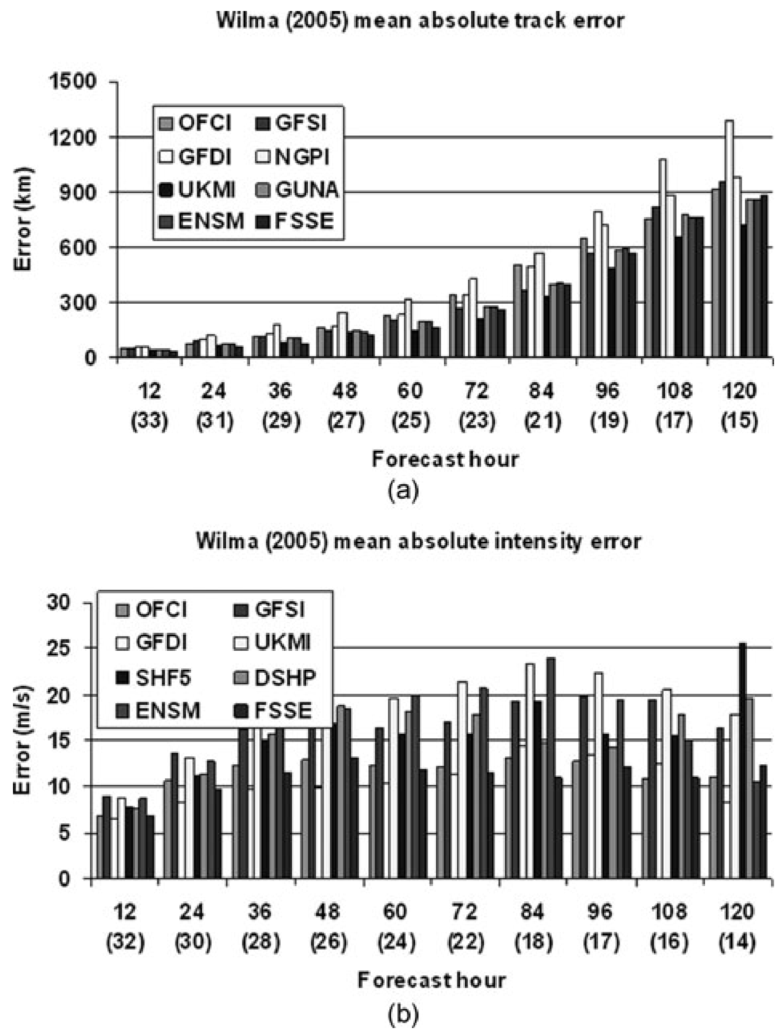


Fig. 13. (a) Mean absolute track errors (in km). (b) Mean absolute intensity errors (in m s^{-1}) for hurricane Wilma. The number within the parentheses in the abscissa shows the number of cases for that particular forecast hour.

dynamic models. The UKMI and the GFSI were the worst models for this particular storm.

9. Concluding remarks

Since the first report of multimodel superensemble-based hurricane forecasts for track and intensity in Krishnamurti et al. (1999), considerable improvements have been made. The main limitation of the superensemble is that the changes made to the member models affect the relevance of the statistics calculated from the training period. The beginning of a hurricane season always starts with the statistical coefficients of the previous season's storms for the training phase of the superensemble. Model changes between seasons obviously affects performance if they affect model biases used in superensemble. As the season progresses, the training is updated due to the inclusion of storms from the current season. Thus, the superensemble statistics are continually evolving during a season as newer forecasts are completed and verified. Changes in the member models of the superensemble just prior to the 2004 season were minimal and as a result the multimodel superensemble outperformed all of the participating models for the real time forecasts during the 2004 season. The 2005 season commenced with some large changes for the GFS, UKMO and NOGAPS models. The errors for FSSE during the 2005 was large and the training cases had to be taken from the 2004 season forecasts and the change of biases in the member models was not accounted.

Our 2005 intensity forecasts for Hurricanes Katrina, Wilma, and especially Rita stand out in terms of smallest errors for the real time superensemble. The superensemble had a mean absolute error of only 9 m s^{-1} for 120-h forecasts of Rita but for 72-h FSSE absolute mean intensity error was 22 m s^{-1}. In these same forecasts the member models' errors were two to three times larger. The NHC evaluation of long-term trends in skill are based on a large number of years of data sets. The superensemble's performance has only been tested in real time over a few years. It is clear that we need to evaluate the superensemble over several more years. The NHC operational forecasts of intensity in

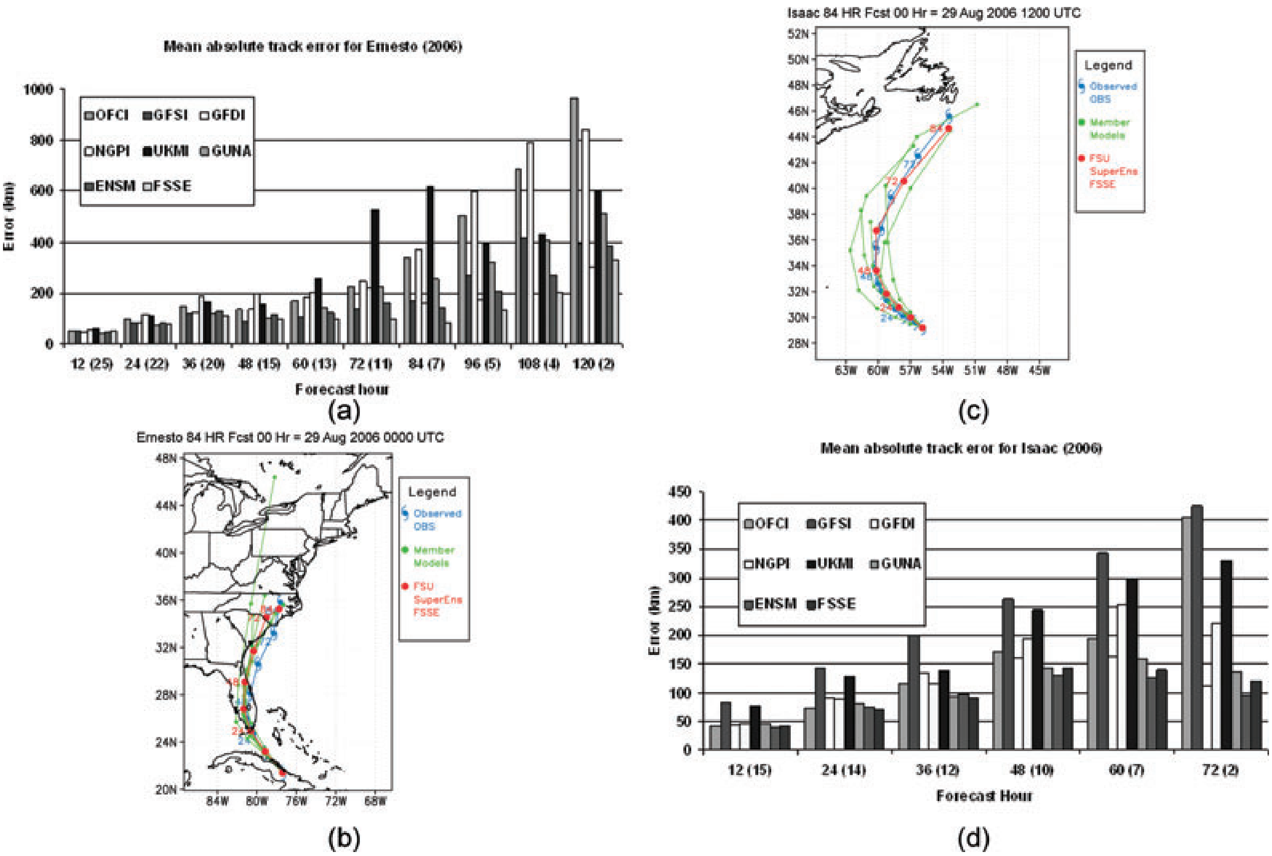


Fig. 14. (a) Mean absolute track errors (in km) for Ernesto. (b) 84-h track forecast of Ernesto at initial time 0000UTC 29 August 2006. The red line indicates the superensemble forecast, the green lines are the forecasts from the member models and the blue line denotes the observed at 12-h intervals. (c) 84-h track forecast of Isaac at initial time 1200UTC 29 September 2006. The red line indicates the superensemble forecast, the green lines are the forecasts from the member models and the blue line denotes the observed at 12-h intervals. (d) Mean absolute track errors (in km) for Isaac.

recent years clearly show an improvement over all member models. The forecasts provided by NHC, are indeed very impressive. It is felt that improvement of the cloud microphysics (requiring very high resolution modelling) will be an important ingredient for future improvements in hurricane intensity forecasting. A suite of mesoscale non-hydrostatic microphysical models may be needed to address superensemble methodologies for intensity forecasts in the future.

10. Acknowledgments

The authors are thankful to James Franklin for helping in validating the results. This research work is support by the following grants: NSF Grant ATM-0419618, NASA Grant NAG8–1848, NOAA Grant 12000586–09 and ACS Inc. (Southcom) Award No. ACSD-04–036.

References

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

Buizza, R., Houtekamer, P. L., Toth, Z., Pellerin, G., Wei, M. and co-authors. 2005. A comparison of the ECMWF, MSC and NCEP global ensemble prediction systems. *Mon. Wea. Rev.* **133**, 1076–1097.

Cane, D. and Milelli, M. 2006. Weather forecasts obtained with a multimodel superensemble technique in a complex orography region. *Meteorol. Z.* **15**, 207–214.

Cartwright, T. J. 2004. *Warm season mesoscale superensemble precipitation forecasts*. Ph.D. Dissertation, The Florida State University, Tallahassee, FL, 104 pp. Available at: http://etd.lib.fsu.edu/theses/available/etd-10252004-132554/. Last accessed Mar 2011.

Chakraborty, A. and Krishnamurti, T. N. 2006. Improved seasonal climate forecasts of the south Asian summer monsoon using a suite of 13 coupled ocean–atmosphere models. *Mon. Wea. Rev.* **134**, 1697–1721.

Chakraborty, A., Krishnamurti, T. N. and Gnanaseelan, C. 2007. Prediction of the Diurnal Cycle Using a Multimodel Superensemble. Part II: clouds. *Mon. Wea. Rev*. **135**, 4097–4116.

Cullen, M. J. P. 1993. The Unified Forecast/Climate Model. *Meteor. Mag.* **122**, 81–94.

Franklin, J. 2005a. 2004 National Hurricane Center Forecast Verification Report. Available at: http://www.nhc.noaa.gov/verification/pdfs/Verification_2004.pdf. Last accessed Mar 2011.

Franklin, J. 2005b. 2004 NHC verification report, 59th Interdepartmental Hurricane Conference, Jacksonville, FL. Available at: http://www.ofcm.gov/ihc05/linking_file_ihc05.htm. Last accessed Oct 2008.

Franklin, J. 2006. 2005 National Hurricane Center Forecast Verification Report. Available at: http://www.nhc.noaa.gov/verification/pdfs/Verification_2005.pdf. Last accessed Mar 2011.

Franklin, J. 2009. 2008 National Hurricane Center Forecast Verification Report. Available at: http://www.nhc.noaa.gov/verification/pdfs/Verification_2008.pdf. Last accessed Mar 2011.

Goerss, J. S. and Jeffries, R. A. 1994. Assimilation of synthetic tropical cyclone observations into the Navy Operational Global Atmospheric Prediction System. *Wea. Forecast.* **9**, 557–576.

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

Goerss, J. S., Sampson, C. R. and Gross, J. M. 2004. A History of Western North Pacific Tropical Cyclone Track Forecast Skill. *Wea. Forecast.* **19**, 633–638.

Heming, J., Chan, J. and Radford, A. 1995. A new scheme for the initilisation of the tropical cyclones in the UK Meteorological Office global model. *Meteorol. Appl.* **2**, 171–184.

Hogan, T. and Rosmond, T. 1991. The description of the Navy Operational Global Atmospheric prediction systems spectral forecast model. *Mon. Wea. Rev.* **119**, 1786–1815.

Houze R. A., Chen, S. S., Smull, B. F., Lee, W-C. and Bell, M. M. 2007. Hurricane intensity and eyewall replacement. *Science* **315**, 1235–1239.

Järvinen, B. R., Neumann, C. J. and Davis, M. A. S. 1984. A tropical cyclone data tape for the North Atlantic Basin, 1886–1983: contents, limitations, and uses. NOAA Technical Memorandum, NWS NHC 22, Coral Gables, Florida, 21 pp.

Knabb, R. D., Rhome, J. R. and Brown, D. P. 2006a. Tropical cyclone report, Hurricane Katrina, 23–30 August 2005. Available at: http://www.nhc.noaa.gov/pdf/TCR-AL122005_Katrina.pdf. Last accessed Mar 2011.

Knabb, R. D., Rhome, J. R. and Brown, D. P. 2006b. Tropical cyclone report, Hurricane Rita, 18–26 September 2005. Available at: http://www.nhc.noaa.gov/pdf/TCR-AL182005_Rita.pdf. Last accessed Mar 2011.

Krishnamurti, T. N., Surendran, S., Shin, D. W., Torres, R. C., Vijaya Kumar, T. S. V. and co-authors. 2001. Real-Time multianalysis–multimodel superensemble forecasts of precipitation using TRMM and SSM/I Products. *Mon. Wea. Rev.* **129**, 2861–2883.

Krishnamurti T. N., Kishtawal, C. M., LaRow, T., Bachiochi, D., Zhang, Z. and co-authors. 1999. Improved skills for weather and seasonal climate forecasts from multimodel superensemble. *Science* **285**, 1548–1550.

Krishnamurti, T. N., Kishtawal, C. M., LaRow, T., Bachiochi, D., Zhang, Z., and co-authors. 2000. Multimodel superensemble forecasts for weather and seasonal climate. *J. Climate* **13**, 4196–4216.

Krishnamurti, T. N., Pattnaik, S., Biswas, M. K., Kramer, M., Bensman, Ed and co-authors. 2010. Hurricane forecasts with a mesoscale suite of models. *Tellus A* **62A**, 633–646.

Kumar, T. S. V., Krishnamurti, T. N., Fiorino, M. and Nagata, M. 2003. Multimodel superensemble forecasting of tropical cyclones in the Pacific. *Mon. Wea. Rev.* **131**, 574–583.

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

Kurihara, Y., Bender, M., Tuleya, R. and Ross, R. 1995. Improvements in the GFDL hurricane prediction system. *Mon. Wea. Rev.* **123**, 2791–2801.

Kurihara, Y., Tuleya, R. and Bender, M. 1998. The GFDL hurricane prediction system and its performance in 1995 hurricane season. *Mon. Wea. Rev.* **126**, 1306–1322.

Molteni, R., Buizza, Palmer, T. N. and Petroliagis, T. 1996. The ECMWF ensemble prediction system: methodology and validation. *Quart. J. Roy. Meteor. Soc.* **122**, 73–120.

Park, K., Zou, X. and Li, G. 2009. A numerical study on rapid intensification of Hurricane Charley (2004) near landfall. *Front. Earth Sci. China* **3**, doi: 10.1007/s11707-009-0048-y.

Pasch, R. J., Blake, E. S., Cobb, H. D. and Roberts, D. V. 2006. Tropical cyclone report, Hurricane Wilma, 15–25 October 2005. Available at: http:// www.nhc.noaa.gov/pdf/TCR-AL252005_Wilma.pdf. Last accessed Mar 2011.

Raftery, A. E., Balabdaoui, F., Gneiting, T. and Polakowski, M. 2003. Using Bayesian model averaging to calibrate forecast ensembles. Tech. Rep. No. 440, Dept. of Stat., University of Washington, 32 pp.

Rhome, J. R. 2007. Technical Summary of the National Hurricane Center Track and intensity Models. Available at: http://www.nhc.noaa.gov/modelsummary.shtml. Last accessed Mar 2011.

Sampson, Goerss, J. S. and Schrader, A. J. 2005. A consensus track forecasts for southern hemisphere tropical cyclones. *Aust. Met. Mag.* **54**, 115–119.

Sampson, C. R., Franklin, J. L., Knaff, J. A. and DeMaria, M. 2008. Experiments with a simple tropical cyclone intensity consensus. *Wea. Forecast.* **23**, 304–312.

Stefanova, L. and Krisnamurti, T. N. 2002. Interpretation of seasonal climate forecast using Brier skill score, the Florida State University superensemble, and the AMIP-I dataset. *J. Clim.* **15**, 534–544.

Toth, Z. and Kalnay, E. 1993. Ensemble forecasting at NMC: the generation of perturbations. *Bull. Amer. Meteor. Soc.* **74**, 2317–2330.

Toth, Z. and Kalnay, E. 1997. Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.* **125**, 3297–3319.

Vialard, J., Vitart, F., Balmaseda, M. A., Stockdale, T. N. and Anderson, D. L. T. 2005. An ensemble generation method for seasonal forecasting with an ocean–atmosphere coupled model. *Mon. Wea. Rev*. **133**, 441–453.

Williford, C. E., Krishnamurti, T. N., Torres, R. C., Cocke, S., Christidis, Z. and co-authors. 2003. Real-time multimodel superensemble forecasts of Atlantic tropical systems of 1999. *Mon. Wea. Rev.* **131**, 1878–1894.