Extended Prediction of North Indian Ocean Tropical Cyclones

James I. Belanger*, Peter J. Webster, Judith A. Curry, and Mark T. Jelinek

School of Earth and Atmospheric Sciences, Georgia Institute of Technology, Atlanta, GA

February 2012

Submitted in revised form to Weather & Forecasting

*Corresponding Author Address:
James I. Belanger
School of Earth & Atmospheric Sciences
Georgia Institute of Technology
311 Ferst Drive
Atlanta, GA 30332-0340
Email: james.belanger@eas.gatech.edu

Abstract

This analysis examines the predictability of several key forecasting parameters using the ECMWF Variable Ensemble Prediction System (VarEPS) for tropical cyclones (TCs) in the North Indian Ocean (NIO) including: tropical cyclone genesis, pre-genesis and post-genesis track and intensity projections and regional outlooks of tropical cyclone activity for the Arabian Sea and the Bay of Bengal. Based on the evaluation period from 2007 to 2010, the VarEPS TC genesis forecasts demonstrate low false alarm rates and moderate to high probabilities of detection for lead-times of one to seven days. In addition, VarEPS pre-genesis track forecasts on average perform better than VarEPS post-genesis forecasts through 120 hrs and feature a total track error growth of 41 nm per day. The VarEPS provides superior post-genesis track forecasts for lead-times greater than 12 hrs compared to other models including: UKMET, NOGAPS, and GFS, and slightly lower track errors than the Joint Typhoon Warning Center. We conclude with a discussion of how the VarEPS can provide much of this extended predictability in a probabilistic framework for the region.

1. Introduction:

Tropical cyclones (TCs) in the North Indian Ocean have a profound impact on the littoral countries of the Arabian Sea and the Bay of Bengal. The combination of a shallow coastal plain along with a thermodynamically favorable environment allow TCs to impart high surface winds, torrential rains and significant wave heights (wave setup plus storm surge) as these systems move inland. In addition, the world's highest population density coupled with low socioeconomic conditions in the region has resulted in several landfalling TCs becoming devastating natural disasters. In fact, eight of the ten deadliest TCs of all time have occurred in the Bay of Bengal and the Arabian Sea with five impacting Bangladesh and three making landfall in India (WMO-TD No. 84). Furthermore, the highest storm tide ever recorded by a TC (45 feet) occurred in the North Indian Ocean near the Meghna Estuary, Bangladesh in 1876 (WMO-TD No. 84). These occurrences highlight the need to provide regional governments and populace in the region as much advance warning as possible.

The Indian Meteorological Department (IMD), which is the WMO–designated Regional Specialized Meteorological Centre (RSMC), provides the official tropical cyclone forecasts and warnings in the North Indian Ocean. As mandated by the WMO, the IMD is required to coordinate and release their forecasts daily with each member country within the North Indian Ocean. However, ultimate responsibility for forecast development and warning dissemination lies with each country's national meteorological service. As part of the daily operational procedure when a TC is not present in the region, the IMD is required to prepare a daily tropical weather outlook, which assesses the possibility of tropical depression development in the Bay of Bengal and the Arabian Sea. Unlike the National Hurricane Center's Tropical Weather Outlook, which provides the likelihood of TC genesis during the next 48 hours (Rappaport et al. 2009), the

WMO provides no mandatory time constraint for this outlook product as the forecast time period is determined separately by each RSMC. In addition, this product does not provide any quantitative, probabilistic information about the potential for tropical cyclone formation or the track for the system if formation occurs.

After a TC has reached depression status, the IMD begins issuing forecast advisories, which describes the system's past movement, current location and intensity and its future location, translation speed, wind intensity, maximum average surface wind speed including the highest surface wind gust. However, the IMD does not produce storm surge forecasts, even though this region has historically experienced devastating impacts from TC-induced storm surge (Webster 2008); storm surge forecasts remain the responsibility of each country's national meteorological service (WMO-TD No. 48). In addition, the WMO only requires that these forecasts cover a time horizon of three days, while most operational numerical guidance in other basins spans a five-day forecast window (Rappaport et al. 2009).

A comprehensive literature review of American Geophysical Union and American Meteorological Society journals found no articles published in the last five years that examine the predictability of NIO TCs using the latest generation of global numerical weather prediction systems. In addition, very few studies have been devoted to assessing the performance of ensemble prediction systems for tropical cyclones. Recently, Dupont et al. (2011) have assessed how well ensemble-based tropical cyclone track forecasts perform in the South Indian Ocean. In particular, they show that calibrated probabilistic forecasts from the European Centre for Medium-Range Weather Forecasts (ECMWF) Variable Ensemble Prediction System (VarEPS) perform better than climatology in assessing track uncertainty for a lead-time of three days.

Results from this analysis has led RSMC La Réunion to develop their uncertainty cones using the VarEPS and to extend their track forecasts from a lead-time of three to five days.

The purpose of this study, similar to the Dupont et al. (2011) analysis, is to assess how well TC forecasts from the VarEPS perform in the North Indian Ocean for the period 2007–2010. In essence, forecast performance will be evaluated through an assessment of probabilistic tropical cyclone genesis forecasts, pre-genesis track and intensity forecasts. This assessment is followed by a more typical comparison of post-genesis forecast track and intensity performance of the VarEPS relative to other global modeling forecast systems. After evaluating the performance of TC genesis, track, and intensity forecasts from the VarEPS in Section 3, recommendations on how this model guidance may be used to produce extended-range probabilistic tropical cyclone forecasts are presented in Section 4.

2. Data and Methods

a) ECMWF Variable Ensemble Prediction System

The predictability of tropical cyclones in the North Indian Ocean is evaluated using the ECMWF VarEPS (hereafter; VarEPS). During the period 2007–2010, the VarEPS has undergone a number of important changes[1] that include increasing the horizontal and vertical resolution of the modeling system, expanding data assimilation procedures to include a greater number of satellite radiance measurements, updating model physics (including cumulus convection parameterization schemes) and changing how initial and stochastic perturbations are generated using singular vectors.

As of 26 January 2010, the VarEPS includes the ECMWF global model that is run at TL1279 spectral truncation (horizontal resolution ~ 16 km) with 91 vertical levels out to ten days along

[1] A complete description of these changes may be found at: http://www.ecmwf.int/products/data/technical/model_id/index.html.

with 51 ensemble members (50 perturbed members + 1 control run) at TL639 spectral truncation (horizontal resolution ~ 32 km) with 62 vertical levels. For days 11–15, the 51-member ensemble is processed at a reduced TL319 spectral truncation (horizontal resolution about 63 km).

To represent the uncertainty in initial conditions, ensemble perturbations are constructed using singular vectors which capture the fastest growing errors in the first 48 hours (Buizza and Palmer 1995). Stochastic perturbations are also added during the model integration to account for the uncertainty in parameterized physical processes. Five additional singular vectors are computed and perturbed in the six grid spaces enclosing each TC using a diabatic, adjoint version of the ECMWF global atmospheric model at TL42 spectral truncation with 42 vertical levels (Barkmeijer et al. 2001, Puri et al. 2001). The 15-day VarEPS 00UTC forecasts for the period 1 January 2007 to 31 December 2010 were obtained through the THORPEX Interactive Grand Global Ensemble (TIGGE; http://tigge.ecmwf.int/) project where the North Indian Ocean domain included the region: 0–30N; 40–110E and were at a horizontal resolution of 0.25^{o} x 0.25^{o}.

b) Tropical Cyclone Tracking Scheme

To isolate tropical cyclones in the VarEPS analysis and forecast fields, we use a modified version of the Suzuki-Parker tracking scheme (Holland et al. 2010). To increase the intensity retrievals from the VarEPS, 10 m winds are replaced with winds averaged in the lower troposphere (10 m, 925 hPa, and 850 hPa). The tracking scheme processes each ensemble member for tropical cyclones by first identifying candidate vortices that exhibit a local minimum in mean sea level pressure. This initial set of vortices is filtered by removing systems that do not have a maximum lower tropospheric wind speed greater than 16 kts (8.2 m s^{-1}) and an 850 hPa

relative vorticity maximum greater than $1x10^{-4}$ s^{-1}. The next component of the tracking scheme confirms that each identified vortex possess a warm core as defined using the Hart (2003) phase-space method. The cyclone phase analysis quantifies the thermal structure by assessing the lower to middle troposphere's thickness gradient across the cyclone and the magnitude of the cyclone's lower troposphere and middle to upper troposphere's thermal wind. After the tracking scheme has been implemented for each ensemble member, any tracks that originate over land are removed unless the ensemble tracks are within 300 n mi of an observed tropical cyclone. Furthermore, all ensemble forecast tracks must have a lifetime of at least one day. The scheme does have limitations with respect to exceedances in lower troposphere winds and relative vorticity which can produce unrealistic track forecasts, such as in the vicinity of 10-17N; 40-55W. Accordingly, all ensemble tracks from this region were eliminated unless they were in association with an observed tropical cyclone.

c) Filtering Tropical Cyclone Forecast Tracks and Determining False Alarms

After applying the tropical cyclone tracking scheme and post-processing routines to the VarEPS forecasts, a tropical cyclone filtering algorithm is used to determine which ensemble track forecasts for a particular VarEPS integration are associated with an observed tropical cyclone (Figure 1). First, the starting time and initial position of each ensemble track forecast, $\mathbf{E}_j(x_o, y_o, t_o)$, is compared to the initial location and time of each observed tropical cyclone, $\mathbf{O}_j(x_o, y_o, t_o)$. If the ensemble forecast track is within the spatial and temporal thresholds set by the filtering algorithm, then the ensemble forecast track is associated with an observed tropical cyclone. The spatial limit is 500 n mi at t_o with an increase of 100 n mi for each 24 hour increase in forecast lead-time. The temporal limit is ± 5 days at t_o with an increase of one day for each 24 hour increase in forecast lead-time. For ensemble track forecasts within the limits, these

forecasts are further subdivided into pre-genesis forecast tracks, $\mathbf{E}_{PRE\text{-}TC\,j}(\mathbf{x})$, and post-genesis forecast tracks, $\mathbf{E}_{POST\text{-}TC\,j}(\mathbf{x})$. If the ensemble forecast track precedes the date of observed tropical cyclone genesis, defined as the date when the first tropical depression advisory is issued by the Joint Typhoon Warning Center (JTWC), then the forecast track is classified as a pre-genesis forecast track, $\mathbf{E}_{PRE\text{-}TC\,j}(\mathbf{x})$. Post-genesis forecasts tracks, $\mathbf{E}_{POST\text{-}TC\,j}(\mathbf{x})$, begin with the first tropical depression advisory and continue until the last tropical advisory has been issued for an observed tropical cyclone.

After determining which ensemble forecast tracks are associated with an observed tropical cyclone, there is an implication that all other ensemble forecast tracks must be false alarms. However, to take advantage of the probabilistic framework of the VarEPS, we have developed a false-alarm clustering algorithm such that a false alarm occurs for a particular model integration when a cluster of VarEPS ensembles produces a localized set (in space and time) of forecast tracks. The method uses the spatial and temporal thresholds from the tropical cyclone filtering routine and "k–means clustering" (Mirkin 1996). K-means clustering is an algorithm designed to divide a set of points (i.e. initial ensemble forecast coordinates) into k–clusters whose membership is based on the distance between each ensemble's forecast point and each cluster's mean or centroid location.

First, the definition of a false alarm cluster is when the normalized number of ensemble tracks within a cluster is in excess of this false alarm ensemble probability, P_F. The ensemble probability threshold, P_F, is pre-defined by the user and should be determined based on the end-user's needs (i.e. for few false alarms, select a high P_F). Next, all ensemble tracks, $\mathbf{E}_j(\mathbf{x})$, for a particular model integration are compared with the set of ensemble forecast tracks that were defined as either pre-genesis, $\mathbf{E}_{PRE\text{-}TC\,j}(\mathbf{x})$, or post-genesis, $\mathbf{E}_{POST\text{-}TC\,j}(\mathbf{x})$, tropical cyclone forecast

tracks. Those ensemble forecast tracks that are not in the set of tropical cyclone forecast tracks become candidate false alarms, $\mathbf{E}_{CFA\,j}(\mathbf{x})$, creating a new set of ensemble forecast tracks for further analysis, i.e. $\mathbf{E}_j(\mathbf{x}) \rightarrow \mathbf{E}_{CFA\,j}(\mathbf{x})$. Thereafter, the clustering routine begins first with the assumption that all candidate false alarm tracks belong to the same cluster, i.e. k = 1, so that an ensemble mean initial location $\mathbf{F}_k(\boldsymbol{x}_0)$ and forecast time $\mathbf{F}_k(t_0)$ is defined. This coordinate information then becomes the "clustering point" that the false alarm filtering routine uses in conjunction with the spatial and temporal thresholds defined previously to determine which ensemble forecast tracks are close to the cluster's ensemble mean starting time and initial location. For ensemble forecast tracks that are not within the distance and time thresholds of the cluster, these tracks are removed from further false alarm consideration unless the number of clusters changes.

After the subset of false alarm ensemble forecasts for the cluster is identified (i.e., $\mathbf{E}_{CFA\,j}(\mathbf{x}) \rightarrow \mathbf{E}_{FAC\,j}(\mathbf{x})$), then if the total number of forecast tracks within this set is in excess of the false alarm ensemble threshold P_F, then the process described previously is repeated except that the number of clusters is increased by one (k = 1 $\rightarrow$ k = 2), so that now, two k-means clusters are created from the initial set of ensemble forecast tracks, $\mathbf{E}_{CFA\,j}(\mathbf{x})$. Each cluster's mean coordinate information is then used in the false alarm filtering routine to identify which ensemble forecast tracks are within the cluster's starting time and initial location. Afterwards, if ensemble track membership of both clusters is in excess of the false alarm probability threshold, then the complete process is repeated except that three clusters (k = 3) are defined. This iterative process continues until the number of ensemble tracks for any cluster drops below the false alarm ensemble probability threshold. When this occurs, the finalized number of false alarm clusters for a particular model integration is then k–1 clusters.

3. Results and Discussion

In Section 3a we present an analysis of the VarEPS predictions of tropical cyclone formation in the North Indian Ocean for the period 2007–2010. The analysis is accomplished by evaluating the probability of detection and the false alarm rate as a function of forecast lead-time and increasing probability threshold. We also include an evaluation of how well the pre-genesis tropical cyclone forecast tracks and intensities from the VarEPS perform relative to observations. In Section 3b, the forecast skill for track and intensity forecasts post-genesis is evaluated. Finally, in Section 3c we examine the regional predictability of TC activity in the North Indian Ocean by evaluating the VarEPS forecasts separately for the Arabian Sea and the Bay of Bengal.

a) Tropical Cyclone Formation and Pre-genesis Forecasts

Figure 2a shows the spatial distribution of tropical cyclone forecast tracks for Severe Cyclone Nargis from the VarEPS initialized on 23 April 2008 00UTC, which is about four days prior to the initiation of tropical depression advisories by the JTWC. This case illustrates some of the forecast information that is contained within the tropical cyclone VarEPS forecasts. The VarEPS forecasts are in good agreement that the pre-tropical vortex that would become Nargis would reach advisory criteria around 27 April 2008 in the central Bay of Bengal. Thereafter, based on forecasts on 23 April, the tropical cyclone is forecast to move generally towards the east-northeast on a track that would cause the system to make landfall in Myanmar around 30 April or 1 May with a high (60%+) probability as a hurricane and a much lower (5%) probability as a severe cyclone/major hurricane. Nargis was observed to intensify to category 4 hurricane level on the Saffir-Simpson intensity scale with maximum sustained winds of 115 kts. The system made landfall in southern Myanmar on 2 May 12UTC, propagating eastward across the Irrawaddy delta (Webster 2008). The forecast performance of the VarEPS for Severe Cyclone

Nargis is exceptional in nearly all facets of prediction: tropical cyclone genesis was forecast correctly eight days in advance, ensemble mean track errors never exceeded 375 n mi even based on pre-genesis track forecasts, and intensity forecasts, although underestimated, indicated a moderate (30%+) probability of Nargis reaching hurricane intensity nearly six days in advance of TC formation.

A more systematic study of all tropical cyclones in the North Indian Ocean is now conducted to determine whether the Nargis forecast performance is characteristic of the VarEPS. Figure 3 is a relative operating characteristic (ROC) evaluation of the VarEPS forecasts for NIO tropical cyclone genesis using the metrics of probability of detection (POD; also known as hit rate) and false alarm rate (FAR; also known as probability of false detection) for all 23 tropical cyclones occurring during the 2007–2010 period using the false alarm clustering methodology from Section 2c. The POD is a measure of the fraction of observed tropical cyclones where TC genesis was forecasted correctly relative to all observed TCs. The FAR is the proportion of all forecasts where a forecast of TC genesis was issued and did not occur along with the number of correct rejections (i.e. TC genesis was not forecast to occur and it did not occur). The false alarm rate should not be confused with the false alarm ratio, which is the proportion of all forecasts where TC formation is forecast to occur but did not (Barnes et al. 2009). Since the false alarm rate is a function not only of the number of false alarms but also the number of correct rejections, the number of VarEPS forecasts included in this evaluation will modulate the false alarm rate mainly through the number of correct rejections. If all VarEPS forecasts from 2007–2010 are included, the false alarm rate as a function of forecast probability threshold is significantly lower than if only the months traditionally associated with NIO TC activity are

219 included. Therefore, for this analysis, the VarEPS evaluation is restricted to the months of April
220 to June and August to December for the 2007–2010 period.

221 The POD and FAR statistics are also sensitive to the time window that is used for verification.
222 For instance, if one were to verify all one-day lead-time TC genesis forecasts, but did not require
223 that the VarEPS correctly forecast the actual date and time of TC genesis (i.e. a 360-hr time
224 window is used), the POD would be much larger and the FAR much smaller than if a short
225 window centered on the time of TC genesis is required. Figure 3a-c shows the POD and FAR of
226 the VarEPS forecasts using 48-hr, 96-hr, and 360-hr time windows calculated at a 10% forecast
227 probability interval ranging from 0 to 100%. The time window is defined with respect to the
228 time of TC genesis, t_o (e.g. a 48-hr window is ± 24 hrs from t_o). Using the 48-hr time window,
229 the VarEPS forecasts for lead-times of one to five days in advance exhibit moderate probabilities
230 of detection (0.4 to 0.7) with very low false alarm rates (0.1–0.2) mainly for forecast probability
231 thresholds of 10–40% (Figure 3a). Although the false alarm rate never exceeds 0.2 even at
232 extended lead-times, VarEPS forecasts made over seven days in advance tend to have low POD
233 once forecast probability thresholds surpass 20%. This observation reflects a decreased
234 frequency of VarEPS forecasts at long lead-times where the forecast probability exceeds 20%. If
235 a 96-hr time window is used, the probability of detection of the VarEPS on average increases by
236 about 0.2 and the false alarm rate by about 0.1 for forecasts of TC genesis made less than seven
237 days in advance (Figure 3b). Forecasts greater than seven days in advance benefit even more
238 from the increase in time window, as the POD for a ten-day lead-time forecast increases on
239 average by about 0.15, and the FAR increases by only 0.05 for a forecast probability threshold of
240 10–30%. Finally, using the full 15-day period of the VarEPS to define the time window of TC

241 genesis, the average POD exceeds 0.7 and the FAR ranges from 0.3 to 0.6 based on forecast
242 probability thresholds of 10–40% (Figure 3c).
243 A key component of any operational forecasting system for TC genesis is knowing what
244 combination of forecast lead-time and time window maximizes the probability of detection while
245 minimizing the false alarm rate. If the ROC score is calculated as a function of forecast lead-time
246 for various time windows, the optimum time window is found by determining at which time
247 window the ROC score obtains a maximum value. It should be noted that the area under the
248 ROC curve, known as the ROC score, varies from 0 to 1, where 0.5 indicates no forecast skill,
249 and a value of 1 indicates a perfect forecast system (Mason and Graham 1999). For TC genesis
250 forecasts with lead-times of one to five days in advance, a 48-hr time window around the
251 forecast date of genesis maximizes the probability of detection while minimizing the false alarm
252 rate. For TC genesis forecasts beyond a lead-time of five days, the ROC score is maximized if
253 the full 15-days of the VarEPS integration is used to determine the forecast time of TC genesis.
254 The following analysis allows a greater understanding of the NIO false alarms that occur in
255 the VarEPS. The spatial distribution of each false alarm cluster's ensemble starting location
256 from 2007 to 2010 is shown in Figure 4a. Although the false alarms in the Arabian Sea and
257 southern Bay of Bengal tend to be distributed uniformly during the 2007–2010 period, there is a
258 relatively high concentration of false alarm clusters that stretch from the northwestern Bay of
259 Bengal into the extreme northeastern Arabian Sea. The false alarm clusters in the northwestern
260 Bay of Bengal tend to occur in a localized region where there is a large gradient in topography
261 between the Bay of Bengal and the northern portion of the Eastern Ghats mountains. This
262 concentration of false alarms may be defining the preferential track of pre-existing cyclonic
263 vortices that move through the Bay of Bengal and into India but never become TCs. The

localized nature of this false alarm concentration may indicate a relationship between current convective parameterization schemes, topography, and the forecast frequency of TC genesis in the VarEPS.

Figure 4b shows the false alarm ratio for the period 2007–2010 with Figure 4c indicating the cumulative distribution function of all false alarms as a function of forecast lead-time. The false alarm ratio obtains a peak value around 0.5 for a forecast probability level of 10% and decays nearly exponentially as the forecast probability threshold increases. At a forecast probability threshold of 25%, the false alarm ratio and the forecast probability threshold are equivalent. Figure 4c shows that around half of all false alarms during the 2007–2010 period occurred at a forecast lead-time of 120 to 240 hours.

We now evaluate how well the VarEPS forecasts perform for track and intensity prior to TC genesis. The average ensemble mean track error at a lead-time of 24 hrs is 82 n mi with a 50% interval of 51–106 n mi, and at 120 hrs it is 224 n mi with a 50% interval of 116–292 n mi for all pre-genesis VarEPS forecasts during the period 2007–2010 (Figure 5a). From a lead-time of 24 hrs to 240 hrs, the mean ensemble track error growth is nearly linear at 41 n mi per day, so that by a lead-time of 240 hrs, the total mean ensemble error is 409 n mi with a 50% percent interval of 186–498 n mi. In addition, the ensemble track error distribution becomes increasingly non-Gaussian as forecast lead-times increases. The implication is that beyond 72 hrs, the mean ensemble track error grows larger than the maximum likelihood of the pre-genesis track error distribution. Figure 5a also shows that the VarEPS track forecasts perform similarly regardless of year, indicating that even though the VarEPS has undergone several major changes during the 2007-2010 period, there has not been a substantial change in forecast track performance in the NIO. To place these pre-genesis track errors in perspective, the Indian Meteorological

287 Department's post-genesis track errors at a lead-time of 48 hrs and 72 hrs typically average 162 n
288 mi and 270 n mi, respectively[2]. This is equivalent to VarEPS ensemble mean *pre-genesis*
289 forecasts at lead-times of 90 and 138 hours, respectively.
290 Figure 5b shows the absolute intensity error for all pre-genesis ensemble forecasts during the
291 period 2007–2010. Forecasts with a lead-time of 24 hrs have a mean absolute intensity error of
292 11 kts (6 m s^{-1}) with a 50% percent interval of 6–15 kts (3–8 m s^{-1}), and by 120 hrs the mean
293 absolute intensity error grows to 23 kts (12 m s^{-1}) with a 50% interval of 7–27 kts (4–15 m s^{-1}).
294 However, unlike the pre-genesis TC track forecasts where the 95th percentile interval is
295 approximately 2.5 times as large as the mean ensemble track error at 120 hrs, the 95th percentile
296 interval for absolute intensity error is 3.4 times as large, reflecting a substantial negative intensity
297 bias for several of the most intense tropical cyclones during the 2007–2010 period (not shown).
298 In terms of the interannual variation in forecast performance of intensity, Figure 5b indicates that
299 the VarEPS forecasts for 2008–2010 have on average performed substantially better than the
300 VarEPS forecasts from 2007, with 2008 and 2009 showing three times the improvement relative
301 to 2007. This marked change in forecast skill of intensity may be due in part to the horizontal
302 and vertical resolution increase that occurred after 2007 in the VarEPS.
303 Forecasting tropical cyclone formation requires an estimate of not only the likely location of
304 TC genesis, but also the time when a system is likely to reach advisory thresholds. Figure 5c
305 shows the relative error (in days) in the VarEPS's forecast timing of TC genesis. Positive values
306 indicate the VarEPS TC genesis forecasts occur sooner than observations, while negative values
307 indicate a later genesis date. For the first 120 hours, the ensemble spread becomes more
308 dispersive as forecast lead-time increases, such that at a lead-time of 24 hours, the VarEPS mean

[2] http://www.imd.gov.in/section/nhac/dynamic/faq/FAQP.htm

ensemble error is -0.5 days with a 50% ensemble interval of -1 to 0.5 days. By 120 hrs, the VarEPS mean ensemble error is 0.5 days but with a 50% ensemble interval of -0.9 to 1.9 days. After a lead-time of 168 hrs, the VarEPS mean ensemble bias begins to increase more rapidly but with little change in ensemble spread, such that at 240 hrs the VarEPS mean ensemble bias grows to 2.2 days with a 50% interval of 0.8 to 3.3 days. Although the ensemble spread is several times larger than the VarEPS mean error, the systematic growth in the mean bias as a function of forecast lead-time is a robust feature of Figure 5c. If one compares the mean error for the first 72 hours (-0.2 days) relative to 168–240 hours (1.3 days), this difference is statistically significant at the 99% confidence level using a bootstrap resampling test.

One possible explanation of why the VarEPS's forecasts for timing of TC genesis are well-constrained through a lead-time of 168 hrs is the dispersion or spread among the VarEPS' ensembles. Figure 5d shows the distribution of ensemble spread in genesis time for each TC from 2007 to 2010 as a function of forecast lead-time. The ensemble spread is calculated as the difference in time (in days) of TC genesis between the earliest and latest ensemble member. As shown in Figure 5d, the ensemble spread in TC genesis time grows rapidly for the first seven days of forecast lead-time, then begins to grow less rapidly reaching a peak spread around 12 days after a forecast lead-time of 168 hrs. While the 15-day integration period has the effect of artificially limiting the spread, the limited TC sample size from 2007-2010 makes it impossible to evaluate the statistical significance of the bias (c.f. Figure 5c).

b) Post-genesis Tropical Cyclone Forecasts

Figure 6a shows the error distribution of all VarEPS track forecasts during the 2007–2010 period indicating how the total track error statistic varies as a function of lead-time after TC genesis has occurred. From Figure 6a it is seen that at a lead-time of 24 hrs, the mean track error

is 71 n mi with a 50% interval of 35–98 n mi. From 24 to 120 hrs, the VarEPS track errors increase linearly at a rate of 58 n mi day^{-1} such that by a lead-time of 120 hrs, the mean track error is 325 n mi with a 50% interval of 148–427 n mi. When compared to the pre-genesis track forecasts, the post-genesis track errors are larger by 15% to 30% depending on the lead-time and quantile considered of the track error distribution. Although a topic of future work, this finding suggests that the VarEPS post-genesis procedure of increasing ensemble spread through moist singular vectors in a sub-region enclosing a TC leads to greater track forecast degradation through a lead-time of 120 hrs than if the procedure was not used. Similar to the VarEPS pre-genesis track forecasts, the VarEPS post-genesis track forecasts show no significant improvement in annual performance for the period 2007–2010.

To place the VarEPS's post-genesis track forecasts in perspective, Figure 6b compares the VarEPS control and ensemble mean forecasts with other forecasting agencies including the JTWC, the U.S. Navy's version of the GFDL (GFDN), the United Kingdom Meteorological Office's global model (UKMET), the National Center for Environmental Prediction's Global Forecast System model (GFS) and the U.S. Navy's NOGAPS model. Since these forecasts were obtained through the U.S Navy's Automated Tropical Cyclone Forecasting System, most of the forecast guidance is limited to 72-hrs, similar to the temporal limit of the JTWC's forecasts prior to 2010. Although the VarEPS control and ensemble mean forecasts on average begin with the largest initial track error, 12-hrs later and beyond the VarEPS control and ensemble mean exhibit the lowest track errors among all other model forecasts. In addition, the VarEPS control and ensemble mean on average exhibit slightly lower track errors than the JTWC through a lead-time of 72-hrs, although this difference is not statistically significant at the 95% confidence level. Relative to the next best performing forecast model, the VarEPS ensemble mean's 24-hr, 48-hr,

and 72-hr track forecast error is on average 10%, 19%, and 27% smaller than NCEP's GFS. Since the linear track error growth per day for the VarEPS ensemble mean (41 n mi/day) is considerably smaller than the GFS (66 n mi/day), greater track forecast utility is obtained at longer lead-times with the VarEPS in comparison to other model forecasts.

Figure 6c shows the distribution of absolute intensity error for all VarEPS post-genesis forecasts during the 2007–2010 period. At the analysis time step (0 hrs), the mean absolute intensity error of all VarEPS forecasts is high at 21 kts (11 m s^{-1}) with a 50% interval of 5–27 kts (3–14 m s^{-1}). By a lead-time of 72 hrs the mean absolute intensity error reaches 28 kts (15 m s^{-1}) with a 50% interval of 8–47 kts (4–25 m s^{-1}), and by 108 hrs, the VarEPS mean forecasts reach their maximum intensity error of 32 kts (18 m s^{-1}) with a 50% interval of 10–46 kts (6–25 m s^{-1}). Unlike the intensity error statistics for the pre-genesis VarEPS forecasts, the interannual variation of post-genesis intensity forecasts reflects a more substantial improvement for the 2008–2010 period compared to 2007. Using the first 72 hrs of lead-time as a reference, an average improvement of 67% relative to 2007 is evident. Finally, Figure 6d compares the VarEPS control and ensemble mean absolute intensity error to other forecast models and the JTWC after removing the initial intensity bias. Generally, for the 2007–2010 period, the VarEPS begins with much higher initial intensity error than any other forecast model (not shown). In addition, the growth rate in mean absolute intensity error among the global forecast models is similar through the first 48 hours. Overall, the Navy's version of the GFDL features the lowest mean absolute intensity error among the models considered here and is very similar to the forecast performance of the JTWC beyond 36 hrs.

c) Regional Outlooks of Tropical Cyclone Activity

We now consider the VarEPS's ability to produce skillful regional outlooks of TC activity (i.e. the probability that a tropical depression strength or greater vortex will be located within a region) by dividing the North Indian Ocean into two sub-domains: the Arabian Sea and the Bay of Bengal. Figure 7a shows the ROC for the Arabian Sea using the VarEPS forecasts from April–June and August–December of 2007–2010. Similar to the ROC analysis of the VarEPS TC genesis forecasts, the VarEPS forecasts for TC activity in the Arabian Sea exhibit moderate probabilities of detection with very low false alarm rates, even as the forecast decision threshold decreases to lower probabilities. As forecast lead-time increases, a transition in forecast performance occurs after a lead-time of ten days. The ROC curve for forecasts with lead-times five to ten days in advance have similar POD and FAR with forecasts at shorter lead-times. However, comparing the ROC curve for forecasts with lead-times of ten to fifteen days relative to five to ten days in advance reveals a large decrease in the probability of detection although the false alarm rate essentially remains constant. From this analysis it is unclear whether this decrease in POD is due an inherent lack of predictability at this longer-time scale or is functionally dependent on the current configuration of the VarEPS with reduced horizontal resolution at lead-times greater than 240 hrs. However, it is clear that the current configuration of the VarEPS is incapable of generating forecast probabilities of TC activity in the Arabian Sea greater than 30% for lead-times of ten to fifteen days, which is one reason why the POD is so much lower than for forecasts with lead-times less than ten days.

Figure 7b shows the ROC for the Bay of Bengal using the same set of VarEPS forecasts as in Figure 7a. Relative to the Arabian Sea, the VarEPS's forecasts of TC activity in the Bay of Bengal have lower probabilities of detection for shorter lead-times, but higher POD values at longer lead-times. Although for lead-times of ten to fifteen days in advance, forecasts in the Bay

400 of Bengal also exhibit slightly higher false alarm rates than the Arabian Sea, the relative change
401 in POD compared with FAR indicates that overall the VarEPS forecasts are more skillful in the
402 Bay of Bengal than the Arabian Sea at extended lead-times.
403 To quantify how well the VarEPS forecasts of TC activity perform in the Arabian Sea and
404 the Bay of Bengal as a function of forecast lead-time, two skill score metrics are used: the Brier
405 skill score (BSS) and the ROC score. The BSS measures the accuracy or relative skill of a
406 forecast over climatology by comparing whether or not an event is forecast to occur relative to
407 observations. A BSS greater than zero implies forecast skill beyond climatology. In this
408 analysis, a 30-yr climatology (1980–2009) of tropical cyclones was developed from the JTWC
409 best-track dataset. Although the BSS metric reveals how skillful a forecast system is relative to
410 climatology, the BSS is regarded as a harsh forecast standard, as it can often hide useful
411 formation information even when the BSS is less than 0 (Mason 2004). Therefore, we use the
412 ROC score as another skill score metric. Table 1 provides the BSS and ROC scores for the
413 Arabian Sea and the Bay of Bengal based on the VarEPS forecasts from April–June and August–
414 December from 2007 to 2010 as a function of forecast lead-time. It should be noted that a
415 BSS/ROC score of one indicates a perfect set of forecasts. To establish statistical significance at
416 the 95% confidence level, a nonparametric bootstrap test was used. In this case, statistical
417 significance of the BSS (ROC score) is determined if the 95% confidence interval of the BSS
418 exceeds 0 (0.50). From Table 1, the BSS metric indicates that the VarEPS forecasts for TC
419 activity in the Arabian Sea are skillful beyond climatology for forecasts up to ten days in
420 advance. Beyond ten days, however, forecasts for TC activity are not skillful relative to
421 climatology. In contrast to the BSS metric, the ROC score is well-above 0.50 out to 15 days,

422 which indicates that the VarEPS provides skillful forecasts of TC activity through two weeks in
423 the Arabian Sea.

424 Table 1 also provides the BSS and ROC scores for TC forecasts in the Bay of Bengal.
425 According to the BSS, forecasts less than five days in advance in the Bay of Bengal are nearly
426 60% less skillful than in the Arabian Sea, which is likely due to the higher frequency of false
427 alarms in the Bay of Bengal (c.f. Figure 4c and Figure 7a-b). However, consistent with the ROC
428 interpretation of Figure 7a-b, the BSS is positive and statistically greater than 0 at the 95%
429 confidence level through a lead-time of ten days. For the ten to fifteen day forecast period, the
430 BSS indicates that the VarEPS performs as well as climatology in the Bay of Bengal. Although
431 TC forecasts from the VarEPS in the Arabian Sea are more skillful than those in the Bay of
432 Bengal for lead-times less than ten days, this result does not hold at longer time scales where the
433 Bay of Bengal forecasts have on average an 18% higher BSS. If the ROC score metric is used to
434 determine forecast skill, the VarEPS forecasts for TCs in the Bay of Bengal are skillful through
435 fifteen days, which is similar to the ROC score results for the Arabian Sea.

436 The results in Table 1 may be compared to a similar analysis that Belanger et al. (2010)
437 performed for the tropical North Atlantic using the ECMWF Monthly Forecast System during
438 the hurricane seasons of 2008 and 2009. They find that the most predictable region for TC
439 activity in the North Atlantic is the Main Development Region, as ROC scores for forecast days
440 8-14 are 0.81 and 0.75 for forecast days 15-21. Relative to these findings, the results presented
441 here indicate similar predictability in the Bay of Bengal as in the North Atlantic Main
442 Development Region, with TC predictability in the Arabian Sea more characteristic of that of the
443 Caribbean Sea at extended forecast lead-times. Belanger et al. (2010) attribute the predictability
444 at these time scales to the ability of the VarEPS to predict accurately the magnitude of deep-layer

445 (850-200 hPa) vertical wind shear as well as the correct frequency of pre-existing cyclonic
446 vortices such as African easterly waves. However unlike the Caribbean Sea, where predictability
447 is modulated largely by the variability in intensity and location of the tropical upper-tropospheric
448 trough, in the Arabian Sea it is likely the combination of dry environmental air and changes in
449 deep-layer vertical wind shear in association with the onset and end of the south Asian monsoon.
450 In addition, Vitart (2009) along with Belanger et al. (2010) show that regional TC predictability
451 in the tropical Atlantic is strongly modulated by the phase and amplitude of the Madden-Julian
452 Oscillation (MJO). Given the more pronounced impact that the MJO has in the NIO (Webster
453 and Hoyos 2004, Hoyos and Webster 2007), we expect this finding to be even more applicable to
454 TCs in the NIO.

455 **4. Summary and Conclusions**

456 The performance of the ECMWF VarEPS in forecasting tropical cyclones in the North Indian
457 Ocean has been examined for the period 2007–2010. The VarEPS is shown to have low false
458 alarm rates and moderate to high probabilities of detection for forecast lead-times through seven
459 days. The VarEPS TC genesis forecast performance is sensitive to the time window that is used
460 to define whether or not an event is forecast to occur. Based on an optimization procedure to
461 achieve the highest ROC score (i.e. maximizing probability of detection and minimizing the false
462 alarm rate), the optimum forecasting combination to predict TC genesis is a 48-hr time window
463 for a forecast lead-time through five days. Thereafter, the full time period of the VarEPS
464 integration should be used to generate TC genesis forecasts five to ten days in advance.

465 Analysis of the VarEPS forecasts from 2007 to 2010 shows that tropical cyclone track
466 forecasts made prior to TC genesis perform 15–30% better than track forecasts produced after
467 TC genesis has occurred. For a lead-time of 24 to 240 hours, the total mean track error grew at a

rate 41 n mi per day such that by a lead-time of 120 hours (240 hours), the average track error of all VarEPS forecast is 224 n mi (409 n mi) with a 50% interval of 116–324 n mi (186–498 n mi). The performance of these track forecasts prior to TC genesis is remarkable considering that the Indian Meteorological Department's average 72-hr forecast track error is 500 km (270 n mi), and these forecasts are issued only after TC genesis has occurred. In terms of post-genesis TC track forecasting, the VarEPS forecasts for a lead-time of 24 hrs average 71 n mi with a 50% interval of 35 to 98 n mi, and by 120 hours the mean track error is 325 n mi with a 50% interval of 148–427. In addition, both the pre-genesis and post-genesis track analyses show that the distribution of ensemble track error becomes increasingly non-Gaussian as forecast lead-time increases. Therefore, to maximize forecast track predictability using the VarEPS, the full distribution of VarEPS track forecasts should be considered and not the mean VarEPS alone. Although the VarEPS control and ensemble mean forecast on average starts with the largest initial track error when compared to other forecast models (i.e. NOGAPS, UKMET, GFS, GFDN), for forecast lead-times 12-hrs later and beyond, the VarEPS control and ensemble mean show the lowest track errors among all other model forecasts. In fact, the VarEPS control and ensemble mean on average exhibit slightly lower track errors than the JTWC through a lead-time of 72-hrs. Since these forecasts are not bias-adjusted to account for the difference in the starting location of the observed TC relative to the initialized location in the VarEPS model or systematic along-track or cross-track biases, additional statistical post-processing steps could be applied to significantly lower the average track errors of the VarEPS at extended lead-times.

In addition, the VarEPS track forecasts were translated into regional outlooks to provide the likelihood of TC activity in the Arabian Sea and the Bay of Bengal. Skill score metrics including the Brier Skill Score and the Relative Operating Characteristic Score were used to evaluate the

VarEPS forecasts. The BSS statistic indicates that the VarEPS TC forecasts for the Arabian Sea and Bay of Bengal are predictable relative to climatology through ten days, whereas the ROC score statistic show that TC activity for both regions are predictable through two weeks.

Based on this evaluation of the VarEPS TC forecasts, it appears feasible for warning agencies in the NIO to begin providing a probabilistic TC formation outlook that assesses the potential for TC development through a lead-time of seven days. When the probability of formation is within moderate (30–60%) levels, the VarEPS's probability of detection will average around 60% with a false alarm rate of about 30% for a lead-time of seven days. In addition, since the distribution of the VarEPS forecasts provides a dynamical measure of the forecast uncertainty in the atmosphere's future state (Dupont et al. 2011), some TCs will be more predictable than others. Therefore, operational forecasts could include a probabilistic outlook including both TC track and intensity derived from the VarEPS. Although the VarEPS tends to be slightly underdispersive at longer forecast lead-times (Majumdar et al. 2010), some additional statistical post-processing steps including bias-correction and probability calibration could be incorporated to ensure that the final forecast track and intensity probabilities are well-conditioned relative to observations.

Acknowledgments

The authors would like to thank the TIGGE project for supplying the ECMWF VarEPS data. Also, we appreciate continuing interactions with Dr. Frederic Vitart and his colleagues at ECMWF. The Climate Dynamics Division of the National Science Foundation under grant NSF 0826909 provided funding support for this research.

513 **References**

514 Barkmeijer, J., R. Buizza, T. N. Palmer, K. Puri, and J.-F. Mahfouf, 2001: Tropical singular
515 vectors computed with linearized diabatic physics. *Quart. J. Roy. Meteor. Soc.*, **127**, 685–708.
516 Barnes, Lindsey R., David M. Schultz, Eve C. Gruntfest, Mary H. Hayden, Charles C. Benight,
517 2009: Corrigendum: false alarm rate or false alarm ratio? *Wea. Forecasting*, **24**, 1452–1454.
518 doi: 10.1175/2009WAF2222300.1
519 Belanger, J. I., J. A. Curry, and P. J. Webster, 2010: Predictability of North Atlantic tropical
520 cyclones on intraseasonal time scales, *Mon. Wea. Rev.*, **138**, 4393–4401.
521 Buizza, R., and T. N. Palmer, 1995: The singular vector structure of the atmospheric global
522 circulation. *J. Atmos. Sci.*, **52**, 1434–1456.
523 Dupont, T., M. Plu, P. Caroff, and G. Faure, 2011: Verification of ensemble-based uncertainty
524 circles around tropical cyclone track forecasts, *Wea. Forecasting,* **26**, 664–676. doi:
525 10.1175/WAF-D-11-00007.1
526 Hart, R.E., 2003: A cyclone phase space derived from thermal wind and thermal asymmetry.
527 *Mon. Wea. Rev.*, **131**, 585–616.
528 Holland, G. J., J. Done, C. Bruyere, C. Cooper, and A. Suzuki-Parker, 2010: Model
529 investigations of the effects of climate variability and change on future Gulf of Mexico tc
530 activity. *Offshore Technology Conference*, **20690**, 1–13.
531 Hoyos, C. D. and P. J. Webster, 2007: The Role of intraseasonal variability in the nature of
532 Asian monsoon precipitation. *J. Climate*, **20**, 4402–4424.
533 Mason, S. J., and N. E. Graham, 1999: Conditional probabilities, relative operating
534 characteristics, and relative operating levels. *Wea. Forecasting*, **14**, 713–725.

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.

Majumdar, Sharanya J., Peter M. Finocchio, 2010: On the Ability of Global Ensemble Prediction Systems to Predict Tropical Cyclone Track Probabilities. *Wea. Forecasting*, **25**, 659–680. doi: 10.1175/2009WAF2222327.1

Mirkin, B., 1996: *Mathematical Classification and Clustering*. Kluwer Academic, 428 pp.

Puri, K., J. Barkmeijer, and T. N. Palmer, 2001: Ensemble prediction of tropical cyclones using targeted diabatic singular vectors. *Quart. J. Roy. Meteor. Soc.*, **127**, 709–734.

Rappaport, Edward N., and Coauthors, 2009: Advances and challenges at the national hurricane center. *Wea. Forecasting*, **24**, 395–419. doi: 10.1175/2008WAF2222128.1

Tropical cyclone operational plan for the Bay of Bengal and Arabian Sea, World Meteorological Organization Tech. Document WMO/TD-No. 84, 1–97. [Available from World Meteorological Organization, Case Postale No. 5, HC-1211, Geneva 20, Switzerland.]

Vitart, F., 2009: Impact of the Madden-Julian Oscillation on tropical storms and risk of landfall in the ECMWF forecast system. *Geophys. Res. Lett.*, **36**, L15802, doi:10.1029/2009GL039089.

Webster, P. J. and C. Hoyos, 2004: Prediction of monsoon rainfall and river discharge on 15-30 day time scales. *Bull. Amer. Met. Soc.*, **85**, 1745–1765.

Webster, P. J., 2008: Myanmar's deadly daffodil. *Nature Geoscience*, **1**, 488–490, doi:10.1038/ngeo257

555 **Figure Captions**

556 Figure 1: Schematic of the false alarm clustering routine. Input variables include P_F, t_t, x_t, $\mathbf{O}_i(\mathbf{x})$,

557 and $\mathbf{E}_j(\mathbf{x})$. P_F is the false alarm probability threshold, t_t is a time threshold of 120 hrs,

558 $x_t = \left(\frac{100}{24}\right)t + 500 \text{ (n mi)}$, $\mathbf{O}_i(\mathbf{x})$ contains location and time information for the i^{th} observed TC

559 during a particular season, $\mathbf{E}_j(\mathbf{x})$ is the ECMWF forecast track for the j^{th} ensemble member.

560 Other variables include $\mathbf{E}_{TC\,j}(\mathbf{x})$ which is the ECMWF forecast track for the j^{th} ensemble member

561 that corresponds to an observed TC. $\mathbf{E}_{TC\,j}(\mathbf{x})$ is made up of pre-genesis TC forecast tracks,

562 $\mathbf{E}_{PRE\text{-}TC\,j}(\mathbf{x})$ and post-genesis TC forecast tracks, $\mathbf{E}_{POST\text{-}TC\,j}(\mathbf{x})$. K is the number of clusters

563 employed in the k–means clustering algorithm, $\mathbf{F}_k(\mathbf{x})$ contains the k^{th} false-alarm cluster's mean

564 location and starting time information. $\mathbf{E}_{CFA\,j}(\mathbf{x})$ is the 'candidate false-alarm' for each j^{th}

565 ECMWF forecast track, $\mathbf{E}_{FAC\,k}(\mathbf{x})$ is the final false-alarm grouping for the k^{th} cluster.

566

567 Figure 2a: Example of the VarEPS forecasts for Severe Cyclone Nargis on 23 April 2008

568 00UTC, which was 3.75 days prior to tropical cyclone genesis according the JTWC Best Track

569 dataset. The black line with red dots denotes the observed track of Nargis. The thin grey lines

570 indicate unique ensemble track forecasts from the VarEPS with the thick black (blue) line

571 denoting the VarEPS ensemble mean (control) track. Figure 2b: The intensity forecast from the

572 VarEPS is shown as a probabilistic time series in which color shading denotes percentile

573 intervals of the VarEPS forecasts ranging from 10 to 90%. Note: The black line with red dots

574 denotes the observed intensity of Nargis.

575

576 Figure 3: Relative operating characteristic of the VarEPS forecasts for tropical cyclone genesis
577 during the period 2007–2010. The color-coded dots correspond to VarEPS tropical cyclone
578 genesis probabilities (%) ranging from 0 to 100%. a) ROC is determined as a function of
579 forecast days in advance using a 48-hour window on the date of TC genesis. b) Similar to a),
580 except for a 96-hour window on the date of TC genesis. c) Similar to a), except the full 360-hour
581 forecast period is used for the TC genesis evaluation.
582
583 Figure 4: a) Initial ensemble location of each false alarm cluster from 2007 to 2010 using a false
584 alarm ensemble threshold of 25%. b) False alarm ratio as a function of increasing false alarm
585 ensemble probability threshold from 2007 to 2010. The red line indicates the location where the
586 false alarm ratio is equivalent to the probability threshold. c) Cumulative distribution functions
587 of the false alarm initiation time as a function of forecast lead-time in hours using a false alarm
588 ensemble threshold of 25%. The legend in Fig. 4c provides the number of false alarm clusters
589 that occurred each year.
590
591 Figure 5: a) VarEPS pre-genesis track errors (in n mi) and b) absolute intensity errors (in kts) for
592 all ensemble forecasts during 2007–2010. c) VarEPS forecasts of the difference between
593 forecast time of TC genesis and observations (in days) with values greater (less) than 0
594 indicating the VarEPS forecasts are sooner (later) than observations. Note: Values have been
595 filtered using a 1-day running mean. d) Ensemble spread in the forecast time of TC genesis.
596 Color shading indicates the percentile ranges for the VarEPS forecasts and the black line is the
597 VarEPS ensemble mean. The total number of ensemble forecasts included in the verification is
598 listed above the abscissa.

599 Figure 6: a) VarEPS post-genesis track errors (in n mi) and c) absolute intensity errors (in kts)
600 for all ensemble forecasts during 2007–2010. Color shading indicates the percentile ranges for
601 the VarEPS forecasts and the black line is the VarEPS ensemble mean. The total number of
602 ensemble forecasts included in the verification is listed above the abscissa. b) Comparison of
603 ECMWF control and ensemble mean track errors (in n mi) and d) absolute intensity errors (in
604 kts) to other global weather models and the Joint Typhoon Warning Center for the period 2007–
605 2010. Initial intensity bias has been removed from each model and the JTWC in Figure 6d.
606
607 Figure 7: Relative operating characteristic for the a) Arabian Sea and b) Bay of Bengal using the
608 VarEPS forecasts from 2007 to 2010 during the months of April–June and August–December for
609 various forecast lead-times.
610

611 **Tables**

612 Table 1. Brier skill scores (BSS) and relative operating characteristic scores (ROCS) for the
613 Arabian Sea and the Bay of Bengal based on VarEPS forecasts for tropical cyclone activity
614 during the months of April–June and August–December for 2007–2010. BSS (ROCS) in bold
615 are statistically different from 0 (0.5) at the 95% confidence level.

Arabian Sea	Brier Skill Score	Relative Operating Characteristic Score
All Forecast Days	**0.17**	**0.82**
Forecast Days ≤ 2	**0.47**	**0.85**
Forecast Days 2–5	**0.32**	**0.87**
Forecast Days 5–10	0.04	**0.82**
Forecast Days 10–15	-0.14	**0.70**
Bay of Bengal	**Brier Skill Score**	**Relative Operating Characteristic Score**
All Forecast Days	**0.09**	**0.80**
Forecast Days ≤ 2	**0.30**	**0.77**
Forecast Days 2–5	**0.16**	**0.79**
Forecast Days 5–10	**0.16**	**0.82**
Forecast Days 10–15	-0.02	**0.74**

616

617

Figures

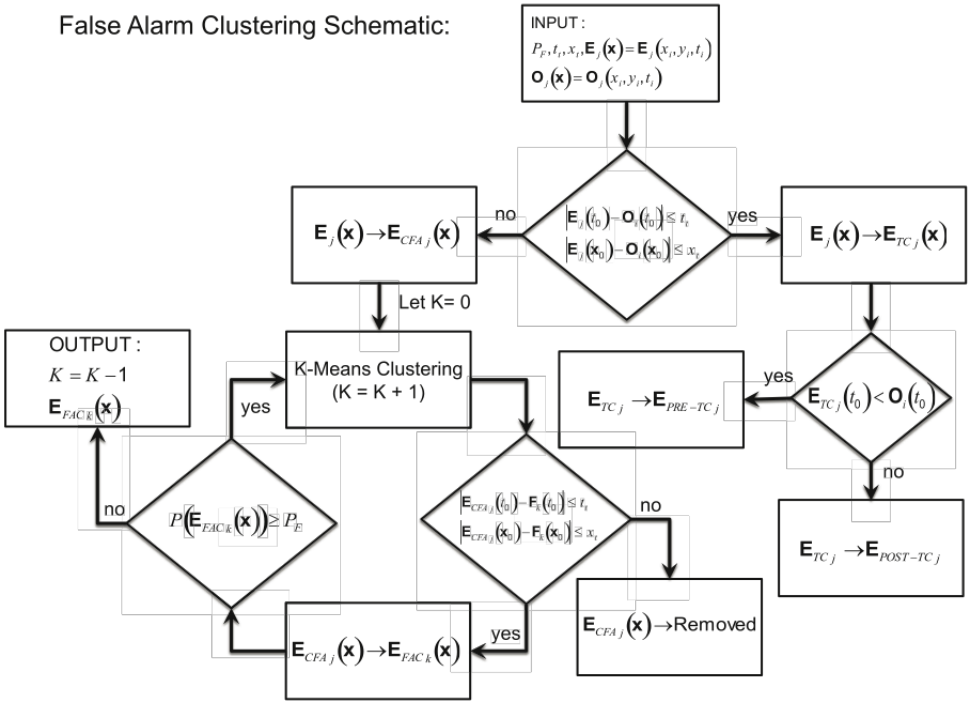
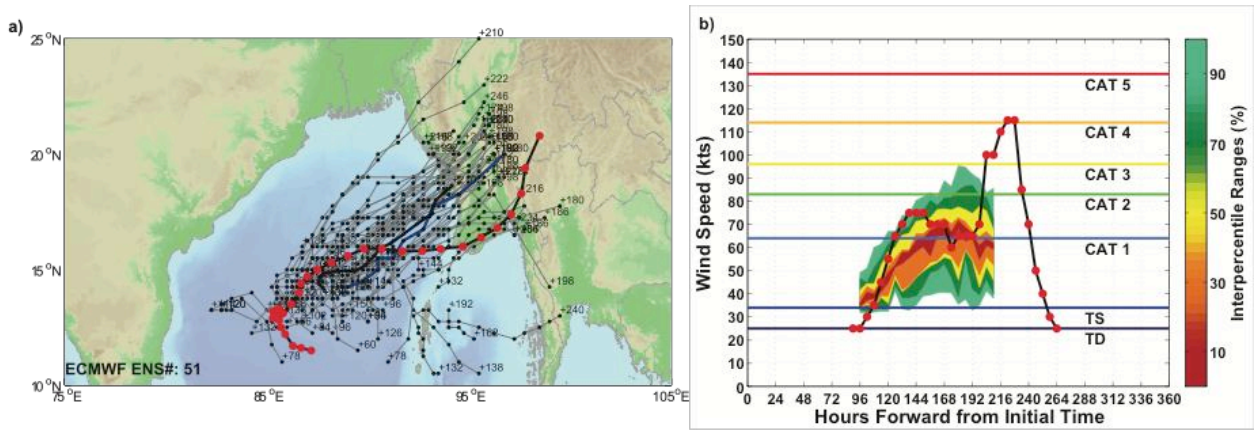


Figure 1: Schematic of the false alarm clustering routine. Input variables include P_F, t_t, x_t, $\mathbf{O}_i(\mathbf{x})$, and $\mathbf{E}_j(\mathbf{x})$. P_F is the false alarm probability threshold, t_t is a time threshold of 120 hrs, $x_t = \left(\frac{100}{24}\right)t + 500$ (n mi), $\mathbf{O}_i(\mathbf{x})$ contains location and time information for the i^{th} observed TC during a particular season, $\mathbf{E}_j(\mathbf{x})$ is the ECMWF forecast track for the j^{th} ensemble member. Other variables include $\mathbf{E}_{TC\,j}(\mathbf{x})$ which is the ECMWF forecast track for the j^{th} ensemble member that corresponds to an observed TC. $\mathbf{E}_{TC\,j}(\mathbf{x})$ is made up of pre-genesis TC forecast tracks, $\mathbf{E}_{PRE\text{-}TC\,j}(\mathbf{x})$ and post-genesis TC forecast tracks, $\mathbf{E}_{POST\text{-}TC\,j}(\mathbf{x})$. K is the number of clusters employed in the k–means clustering algorithm, $\mathbf{F}_k(\mathbf{x})$ contains the k^{th} false-alarm cluster's mean location and starting time information. $\mathbf{E}_{CFA\,j}(\mathbf{x})$ is the 'candidate false-alarm' for each j^{th} ECMWF forecast track, $\mathbf{E}_{FAC\,k}(\mathbf{x})$ is the final false-alarm grouping for the k^{th} cluster.



630

631 Figure 2a: Example of the VarEPS forecasts for Severe Cyclone Nargis on 23 April 2008

632 00UTC, which was 3.75 days prior to tropical cyclone genesis according the JTWC Best Track

633 dataset. The black line with red dots denotes the observed track of Nargis. The thin grey lines

634 indicate unique ensemble track forecasts from the VarEPS with the thick black (blue) line

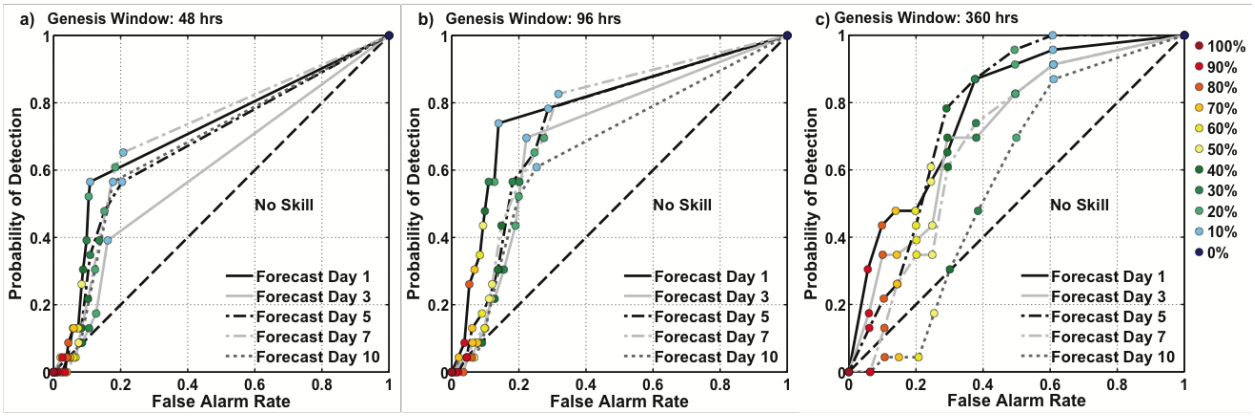
635 denoting the VarEPS ensemble mean (control) track. Figure 2b: The intensity forecast from the

636 VarEPS is shown as a probabilistic time series in which color shading denotes percentile

637 intervals of the VarEPS forecasts ranging from 10 to 90%. Note: The black line with red dots

638 denotes the observed intensity of Nargis.

639

640 

641 Figure 3: Relative operating characteristic of the VarEPS forecasts for tropical cyclone genesis
642 during the period 2007–2010. The color-coded dots correspond to VarEPS tropical cyclone
643 genesis probabilities (%) ranging from 0 to 100%. a) ROC is determined as a function of
644 forecast days in advance using a 48-hour window on the date of TC genesis. b) Similar to a),
645 except for a 96-hour window on the date of TC genesis. c) Similar to a), except the full 360-hour
646 forecast period is used for the TC genesis evaluation.

647

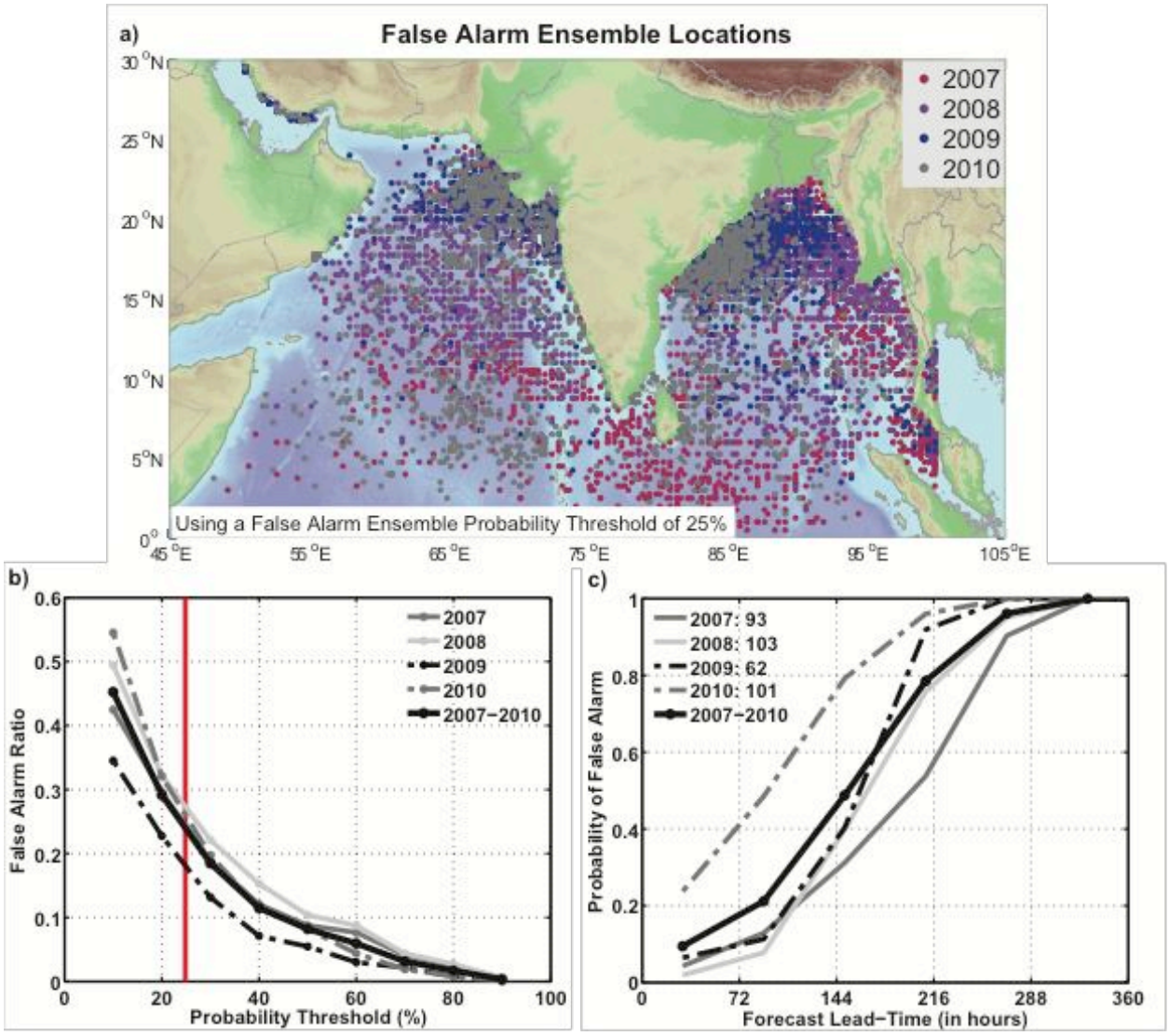


Figure 4: a) Initial ensemble location of each false alarm cluster from 2007 to 2010 using a false alarm ensemble threshold of 25%. b) False alarm ratio as a function of increasing false alarm ensemble probability threshold from 2007 to 2010. The red line indicates the location where the false alarm ratio is equivalent to the probability threshold. c) Cumulative distribution functions of the false alarm initiation time as a function of forecast lead-time in hours using a false alarm ensemble threshold of 25%. The legend in Fig. 4c provides the number of false alarm clusters that occurred each year.

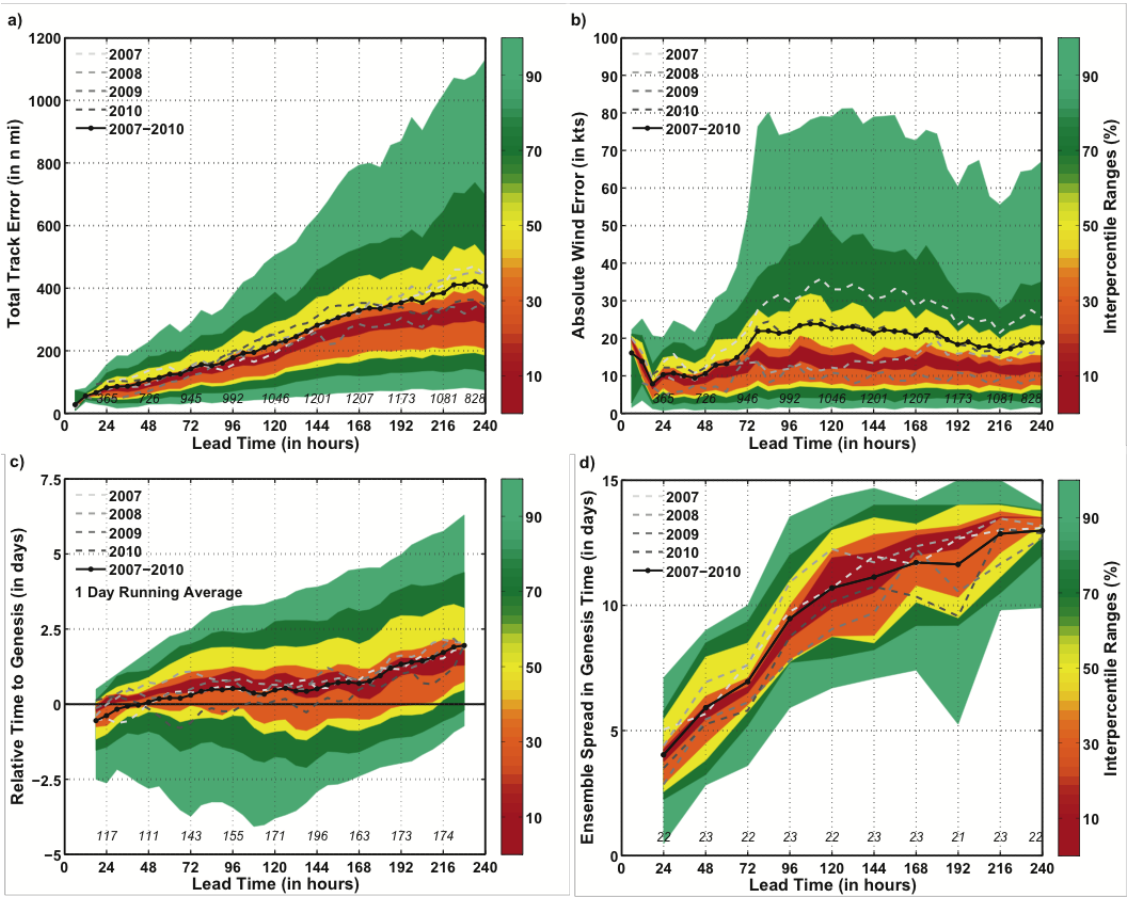


Figure 5: a) VarEPS pre-genesis track errors (in n mi) and b) absolute intensity errors (in kts) for all ensemble forecasts during 2007–2010. c) VarEPS forecasts of the difference between forecast time of TC genesis and observations (in days) with values greater (less) than 0 indicating the VarEPS forecasts are sooner (later) than observations. Note: Values have been filtered using a 1-day running mean. d) Ensemble spread in the forecast time of TC genesis. Color shading indicates the percentile ranges for the VarEPS forecasts and the black line is the VarEPS ensemble mean. The total number of ensemble forecasts included in the verification is listed above the abscissa.

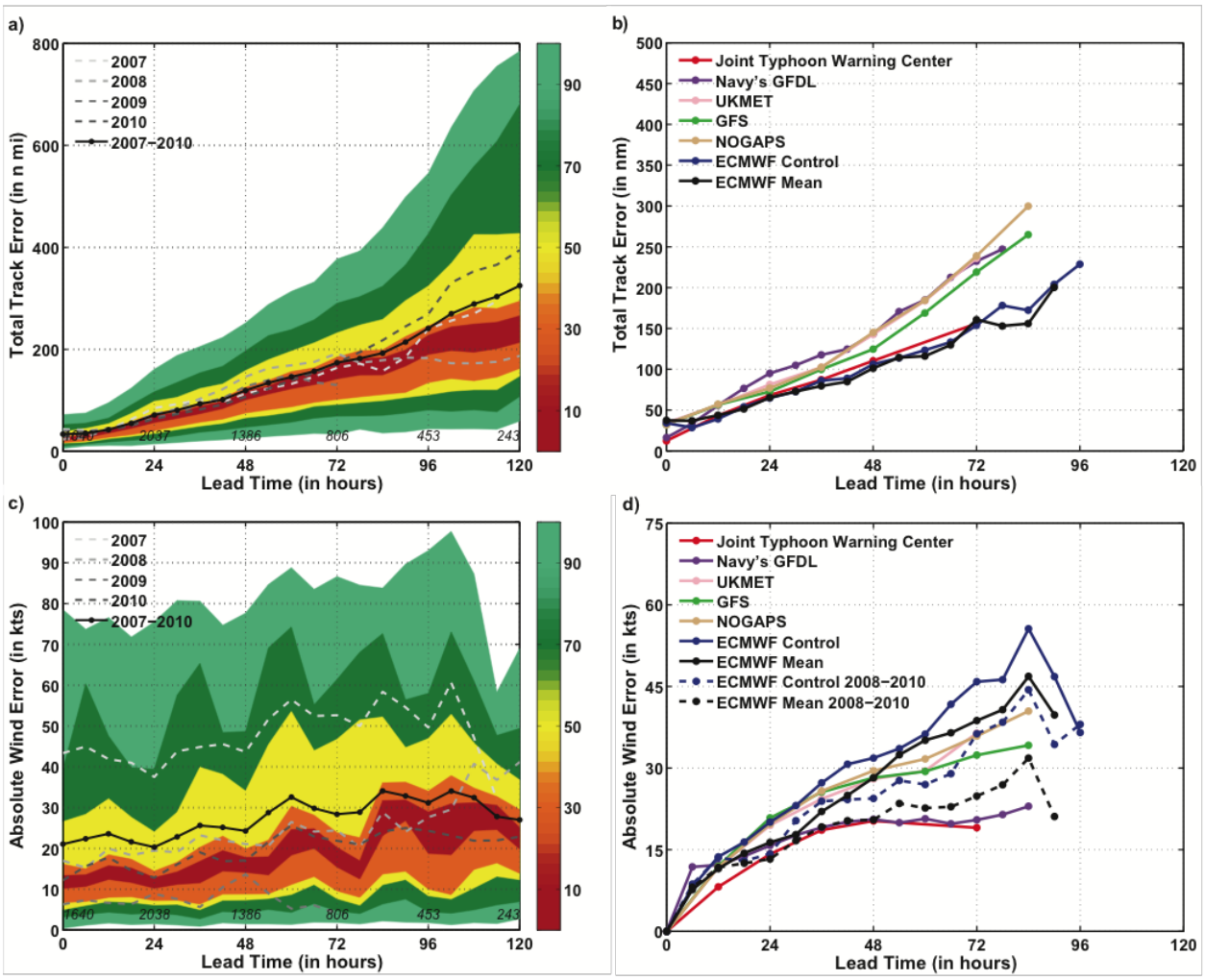
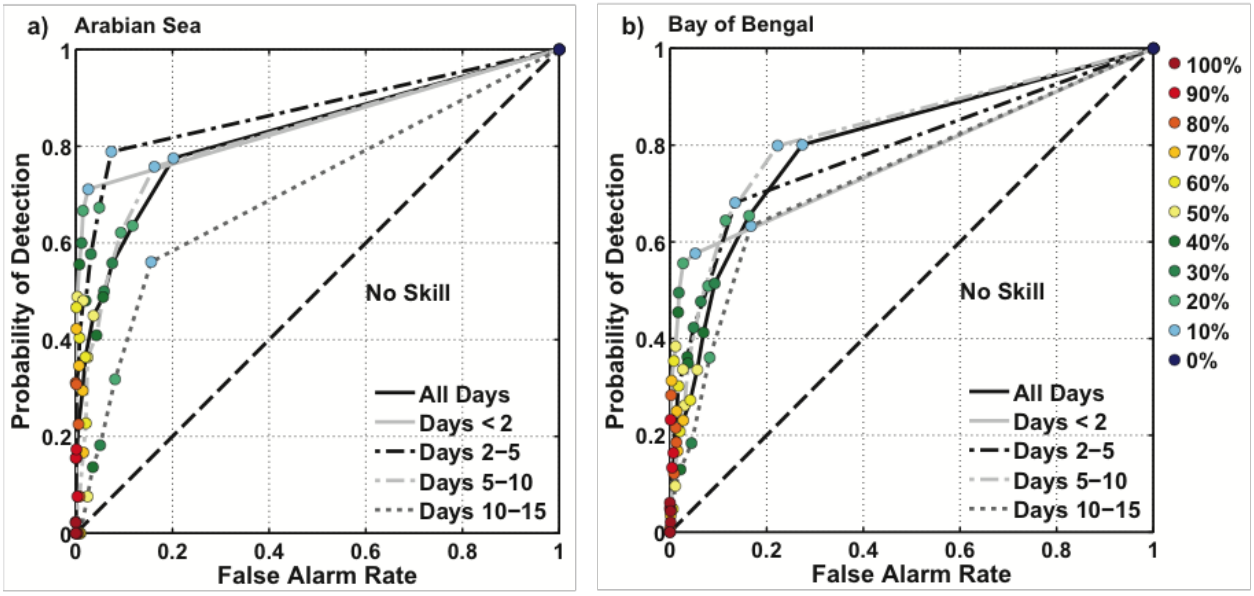


Figure 6: a) VarEPS post-genesis track errors (in n mi) and c) absolute intensity errors (in kts) for all ensemble forecasts during 2007–2010. Color shading indicates the percentile ranges for the VarEPS forecasts and the black line is the VarEPS ensemble mean. The total number of ensemble forecasts included in the verification is listed above the abscissa. b) Comparison of ECMWF control and ensemble mean track errors (in n mi) and d) absolute intensity errors (in kts) to other global weather models and the Joint Typhoon Warning Center for the period 2007–2010. Initial intensity bias has been removed from each model and the JTWC in Figure 6d.



675

676 Figure 7: Relative operating characteristic for the a) Arabian Sea and b) Bay of Bengal using the
677 VarEPS forecasts from 2007 to 2010 during the months of April–June and August–December for
678 various forecast lead-times.