OBEST: An Observation-Based Ensemble Subsetting Technique for Tropical Cyclone Track Prediction

LIN DONG

National Meteorological Center, Beijing, China, and Department of Meteorology, The Pennsylvania State University, University Park, Pennsylvania

FUQING ZHANG

Department of Meteorology, and Center for Advanced Data Assimilation and Predictability Techniques, The Pennsylvania State University, University Park, Pennsylvania

(Manuscript received 29 April 2015, in final form 10 November 2015)

ABSTRACT

An observation-based ensemble subsetting technique (OBEST) is developed for tropical cyclone track prediction in which a subset of members from either a single- or multimodel ensemble is selected based on the distance from the latest best-track position. The performance of OBEST is examined using both the 2-yr hindcasts for 2010–11 and the 2-yr operational predictions during 2012–13. It is found that OBEST outperforms both the simple ensemble mean (without subsetting) and the corresponding deterministic high-resolution control forecast for most forecast lead times up to 5 days. Applying OBEST to a superensemble of global ensembles from both the European Centre for Medium-Range Weather Forecasts and the National Centers for Environmental Prediction yielded a further reduction in track forecast errors by 5%–10% for lead times of 24–120 h.

1. Introduction

During the past couple of decades, significant progress has been made in tropical cyclone (TC) track forecasts. For the western North Pacific basin, the 24-h track forecast error today is 100 km less than that of 20 years ago while the forecast of a 48-h track is as accurate as that of a 24-h track forecast 20 years ago (Qian et al. 2012). For the Atlantic basin, compared to track forecast errors 15–20 years ago, the track forecast errors from day 1 to day 5 in 2013 have been reduced by more than 50% (Cangialosi and Franklin 2014). Similarly impressive improvements in TC forecasts have also been observed in other basins (Mohapatra et al. 2013; WMO 2007; Chan 2010).

Advances in numerical weather prediction (NWP) are the driving force for the decrease in the TC track forecast error. At operational centers around the world, TC model guidance evolved from classic statistical models, to hybrid statistical–dynamical models, and then to deterministic fully dynamical regional and global NWP in the 1990s (e.g., Goerss et al. 2004; Sampson et al. 2005; Elsberry 2007, 2014; Qian et al. 2012). In the past decade or so, there has been a major push toward a multimodel consensus of deterministic NWP models, which is better than the forecasts from any of the individual component models owing to offsetting random errors (Elsberry 2014) and represents the state-of-the-art in operational TC track forecasting (Burton 2006).

Consensus models are not forecast models per se, but are instead combinations of forecasts from multiple models (Cangialosi and Franklin 2014). Therefore, the performance of a consensus model is determined by two factors: the consensus technique and the consensus components. Consensus technique refers to how the weight of each component is assigned, and can be divided into two categories: equal weights and unequal weights. Equal weights are calculated by a simple arithmetic average, whereas unequal weights are determined using more complex techniques such as multiple regressions (which require a long training phase). Because the upgrade cycles for NWP models are short and it is difficult to retrain the data before every

Denotes Open Access content.

Corresponding author address: Prof. Fuqing Zhang, Dept. of Meteorology, The Pennsylvania State University, University Park, PA 16802.
E-mail: fzhang@psu.edu

DOI: 10.1175/WAF-D-15-0056.1

hurricane season (Williford et al. 2003), operational centers prefer to use the faster, equal-weight consensus technique (Goerss 2000; Goerss et al. 2004; Sampson et al. 2005; Burton 2006; Krishnamurti et al. 2010). The research on consensus models started in the 1970s, and the components of the consensus forecast have evolved in the same manner as the guidance models described in the previous paragraph. Before the 1990s, the consensus components primarily consisted of subjective forecasts, statistical models, and statistical–dynamical models (Sanders 1973; Thompson 1977; Danard 1977; Leslie and Fraedrich 1990). Thereafter, the components have mainly consisted of deterministic NWP models (Elsberry and Carr 2000; Goerss 2000; Krishnamurti et al. 2010; Kumar et al. 2003; Sampson et al. 2005; Weber 2003; Williford et al. 2003; Elsberry et al. 2008).

Over the past couple of decades, operational NWP has evolved from using single deterministic forecasts toward ensemble prediction systems (EPSs), whereby an ensemble of forecasts is generated, often by making slight perturbations to the model's initial conditions and model physics (e.g., Toth and Kalnay 1993). The observation-based ensemble subsetting technique (OBEST) presented herein follows the work of others (Lee and Wong 2002; Yamaguchi et al. 2012; Jun et al. 2014; Qi et al. 2014) by using EPS members as components in a consensus model. The multimodel, equal-weight approaches were used in Lee and Wong (2002) and Yamaguchi et al. (2012), while Jun et al. (2014) proposed the use of multiple regressions for determining the respective weight of the mean of each EPS but each ensemble member within each EPS is still weighted equally. Qi et al. (2014) first proposed the use of the average of a subset of an EPS that includes all members with a short-term track error less than the average track error of the whole ensemble. They examined two methods of obtaining an ensemble mean track forecast: one in which all subset members are weighted equally and the other in which the chosen members are weighted as a function of their short-term track forecasted error. The current study also complements the work of Qi et al. [(2014); which applied to a single ensemble] through modification in the subset selection and expansion to the use of a superensemble (combination of multiple ensemble prediction systems).

When China Meteorological Administration (CMA) forecasters used the EPS data, they found that for a given forecast ensemble, some ("good") members gave small track forecast errors whereas other ("bad") members gave poor forecast tracks that deviated substantially from the truth and the good members. Unfortunately, the absolute and relative accuracies of any given ensemble member will change from run to run, making it impossible for a forecaster to subjectively determine whether that member is good or bad. An objective method of identifying the good members would provide tremendous benefit to forecasters.

This paper is organized as follows. The datasets and methodology are presented in section 2. The application and verification of OBEST for CMA operational TC forecasts are described in section 3. The development and performance of alternative OBEST approaches are shown in section 4. An exemplar demonstration of OBEST for Tropical Storm (TS) Cimaron (2013) is presented in section 5. Concluding remarks are given in section 6.

2. Datasets and methodology

a. *Datasets for ensemble predictions and best-track observations*

The tropical cyclone track predictions by two operational ensemble prediction systems (hereafter EC-EPS and NCEP-EPS) used in this study were downloaded from the THORPEX Interactive Grand Global Ensemble (TIGGE) website (http://apps.ecmwf.int/datasets/data/tigge). The position, minimum sea level pressure, and maximum sustained wind speed near the center of each TC from each ensemble member are included in this dataset (Bougeault et al. 2010). The EC-EPS is from the European Centre for Medium-Range Weather Forecasts (ECMWF) and comprises 50 perturbed ensemble members plus a control forecast initialized at 0000 and 1200 UTC every day at T639 (~50 km) resolution (Buizza et al. 2007). The NCEP-EPS is from the National Centers for Environmental Prediction (NCEP) and comprises 20 perturbed ensemble members plus a control forecast initialized four times per day (the EPS resolution is T190 for the hindcast period and T254 for the real-time applications discussed in this study). Unlike EC-EPS, a tropical cyclone relocation procedure is used in NCEP-EPS (Wei et al. 2008). To temporally align the data of the two ensemble prediction systems, only forecasts out to 5 days and initialized at 0000 and 1200 UTC are used in this study. In addition, forecasts of the deterministic higher-resolution models from both centers are also used for comparison.

The best-track dataset over the western North Pacific from the Japan Meteorological Agency (JMA; http://www.jma.go.jp/jma/jma-eng/jma-center/rsmc-hp-pub-eg/trackarchives.html) has a total of 91 TCs in the domain 0°–50°N and 100°E–180° from 2010 to 2013.[1] During the

[1] Although it is beyond the scope of the current study to pinpoint the uncertainties in the best-track estimates of JMA, a crude estimate can be achieved by computing the average distance between the JMA best-track estimate and another independent best-track position estimate issued by the Joint Typhoon Warning Center (JTWC). The average distance between these two operational centers for all 994 of the homogeneous samples over 2012–13 is 23 km.

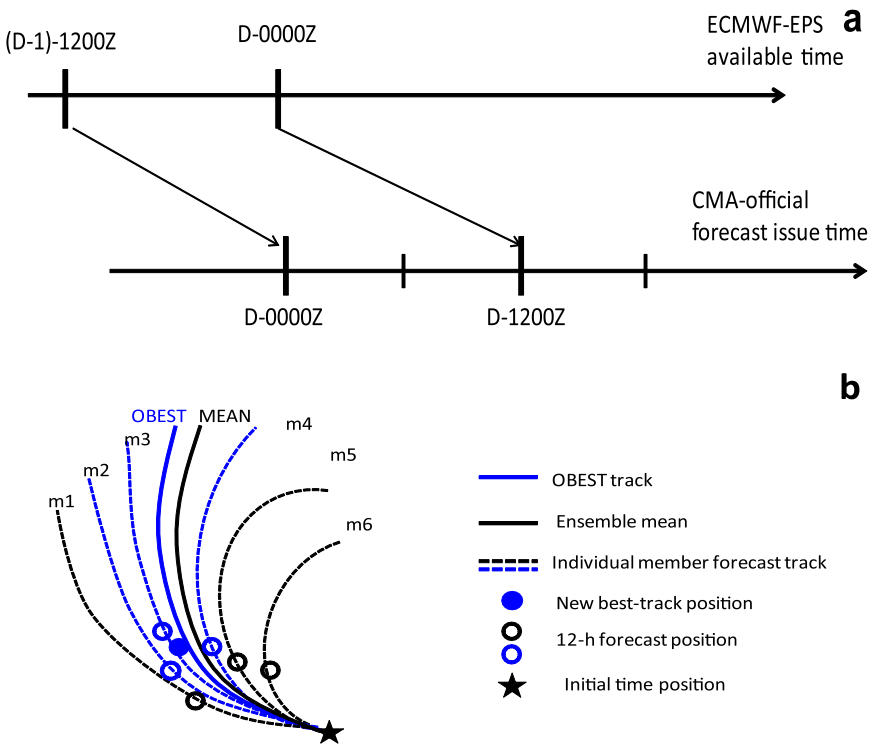

FIG. 1. (a) Schematic illustration of the 12-h difference between the time when EC-EPS forecasts arrive at CMA and the time when official forecasts are issued. (b) Schematic illustration of the OBEST algorithm where the blue dashed lines represent the track of selected members for subsetting (see text for details).

development of OBEST from 2010 to 2011, hindcast experiments were performed for a total of 35 TCs with 302, 242, 190, 140, and 98 homogeneous samples for lead times of 24, 48, 72, 96, and 120 h, respectively. For the operational testing of OBEST during 2012–13, forecasts for a total of 56 TCs with 473, 372, 273, 182, and 119 homogenous samples are compared for lead times of 24, 48, 72, 96, and 120 h respectively. All forecasts with initial intensity reaching tropical storm strength or stronger are included in the verification. No additional requirements are imposed on the forecasts of TC intensity at the verifying time, as long as there are best-track data to be verified against, including those degraded to tropical depressions (TDs) or under extratropical transition.

Track forecast error is defined as the great-circle distance between a TC's forecast position and the best-track position at the forecast verification time. Expressed as a percentage improvement over the baseline, reduction in track forecast errors of a forecast S_f is defined as a percentage by

$$S_f(\%) = 100(e_b - e_f)/e_b, \quad (1)$$

where e_b is the error of the baseline model and e_f is the error of the forecast being evaluated. It can be seen that the percentage is positive when the forecast error e_f is smaller than the error from the baseline e_b (Cangialosi and Franklin 2014).

b. *Methodology: Observation-based ensemble subsetting technique*

There is nearly a 12-h lag between the initialization time of the EC-EPS and the time that TC track prediction data are first available to CMA operational forecasters (Fig. 1a). In contrast, observational best-track data are available in near–real time, which allows for comparison with 12-h forecast data from the EC-EPS. A schematic diagram of OBEST is illustrated in Fig. 1b. Out of an EPS of N total members (e.g., $N = 51$ for EC-EPS), the OBEST consensus consists of the M members ($M \leq N$) that have the smallest 12-h track errors as verified against the latest best-track observed position. The subsequent OBEST consensus forecast track is a simple arithmetic average of the position forecasts from the M selected members.

The inherent premise of this technique is that members with a more accurate short-term forecast (smaller 12-h track error) will on average give better performance at longer forecast lead times. An important advantage of OBEST is that it can adjust quickly to the

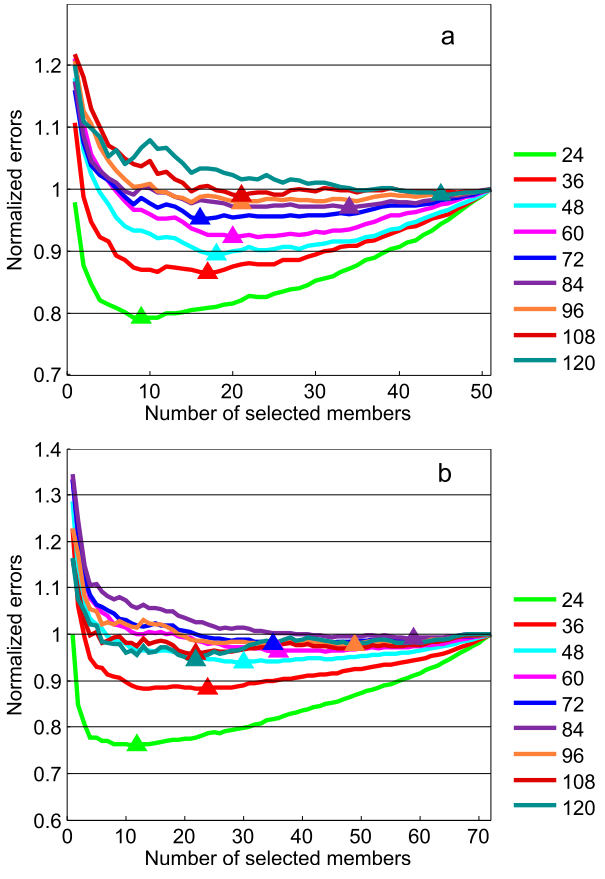

FIG. 2. The consensus mean track errors (normalized by total ensemble mean error without subsetting) as a function of selected good members for (a) the EC-EPS ensemble and (b) the combined ECMWF–NCEP superensemble for all western North Pacific storms during 2010–11. Colored lines represent the normalized errors at different lead times. Triangles are plotted at locations corresponding to the smallest mean errors at different lead times.

updated model configuration and performance characteristics. A schematic representation of the subsetting technique is depicted in Fig. 1b in which ensemble members 2, 3, and 4 (out of six total depicted; i.e., $M = 3$ and $N = 6$ in this example) are selected based on their verification against the best-track observations at 12 h after model initial time. In OBEST, the exact number of selected good members (M) in the ensemble subset is based on past performance with all available ensemble members. This strategy is different from that of Qi et al. (2014), in which M varies according to the number of members whose errors are smaller than the simple average error. Figure 2 shows the consensus mean track errors (normalized by total ensemble mean error, which is the case where $M = N$; a normalized error <1 indicates that the OBEST method for a given M outperforms the full ensemble mean) as a function of selected good members M for EC-EPS for all western North Pacific storms during 2010 and 2011. It is clear from Fig. 2a that for most lead times, the smallest track errors can be obtained when only $M = 16$–21 good members (out of $N = 51$ total) were selected. For simplicity and without loss of generality, we use $M = 20$ members in OBEST for the EC-EPS, which represents about 40% of the total available members. Nevertheless, future studies will explore the use of a time-variant M in selecting the ensemble subset.

3. Applications and verification

a. *Performance of hindcasts with OBEST during 2010–11*

We first evaluate the OBEST using the track forecasts from the operational ECMWF ensemble (EC-EPS) for all western North Pacific tropical cyclones during 2010 and 2011 archived in the TIGGE dataset. Figure 3 shows the performance of OBEST in terms of absolute error with the mean of the entire EC-EPS ensemble (ECM) as well as the high-resolution ECMWF deterministic model prediction (ECD), all of which were verified against the JMA best-track estimates. It can be seen from Fig. 3a that OBEST has considerably smaller error than ECM for all lead times until 96 h [although only the difference up to 36 h passes the 95% statistical significance using the Student's t test; Wilks (2006)]. In particular, OBEST has track errors of 59 (118) km at 24 (48) h, which represents an 18.5% (10.1%) error reduction from the simple ensemble mean of ECM. The advantage of OBEST over ECM diminishes at 108 h, and ECM performs better at 120-h lead time. OBEST also performs better than the deterministic forecast (ECD) at almost all lead times except for 84 and 96 h. It is also worth noting that ECD is comparable to or slightly better than ECM up to 96 h, but considerably worse than ECM at longer lead times of 108 and 120 h. The reasons that lead to the different performance of ECD versus ECM are beyond the scope of the current study.

b. *Operational performance of OBEST during 2012–13*

Our OBEST approach was motivated by a similar subsetting method proposed by Qi et al. (2014). However, the number of subsetting ensemble members was not fixed in Qi et al. (2014) while the weight of their subsetting ensemble mean is inversely proportional to the respective errors of each ensemble member. A preliminary comparison of our new approach to the original method of Qi et al. (2014) for 2012–13 shows a persistent reduction of absolute mean forecast error at nearly all times; however, none of these differences is statistically

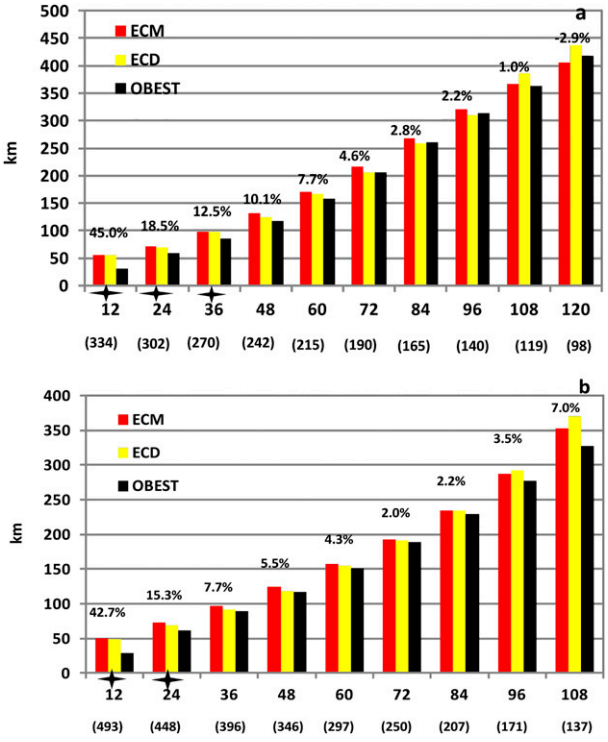

FIG. 3. Performance of OBEST in terms of mean absolute error (black) at different forecast lead times (x axis; h), in comparison with ECM (red) as well as ECD (yellow), verified against the JMA best-track estimates for (a) hindcasts of 2010 and 2011 and (b) operational forecasts of 2012 and 2013. Best-track positions are replaced by CMA operational estimates for selecting OBEST members during operational implementation of 2012–13, but forecasts are still verified against the JMA best track for independent validation. Percentages of relative error reduction from ECM to OBEST at different times are marked above the bars while the sample sizes in the comparison are noted in parentheses. Forecast mean error difference between ECM and OBEST significant above the 95% confidence level is noted with a star.

significant and thus is not shown. Therefore, thanks to the collective efforts by several operational TC forecasters at CMA, OBEST was implemented operationally by the start of the 2012 western North Pacific typhoon season.

Figure 3b shows the operational performance of OBEST during 2012 and 2013 in terms of absolute mean track forecast error in comparison with ECM and ECD at different lead times. Note that because none of the three forecast products (OBEST, ECM, and ECD) is available at the model initialization time, the effective operational lead times for each of the products have been subtracted by 12 h. In addition, the initial positions of the TCs (which are used for selecting the OBEST ensemble subset in real time) are based on the CMA operational best-track estimate while the forecast error is verified against the (independent) JMA postseason best-track estimate.

It can be seen from Fig. 3b that OBEST gives considerably better performance than both ECM and ECD at almost all lead times (though again the statistical significance at the 95% confidence level can only be established for shorter lead times). More specifically, the mean track forecast errors for OBEST at 24-, 48-, 72-, and 96-h forecast times (which equal lead times plus 12 h) are 61, 117, 189, and 277 km, respectively, which are 15.3%, 5.5%, 2.0%, and 3.5% smaller than the ECM errors of 73, 124, 193, and 288 km, respectively. Note that while CMA issues forecasts every 6 h, the verifications shown here are only for forecasts issued at 0000 and 1200 UTC. Operationally OBEST is applied at 0600 and 1800 UTC cycles using 18-h EC-EPS forecasts in member selection (not shown).

Figure 4 shows scatterplots of track errors from OBEST versus those from ECM for all valid forecast samples at different forecast hours for 2012 and 2013 [different from Fig. 3b in which CMA best-track estimates were used, the results displayed in Fig. 4 (and later; see Figs. 6–9) are derived using the JMA best-track estimates]. At 24 h (Fig. 4a), OBEST provides a more accurate forecast than ECM by a large margin (specifically, OBEST has a smaller error in over 77.6% of all 24-h forecast samples). At forecast hours 48 and 72 (Figs. 4b,c), OBEST is more accurate than ECM approximately 60% of the time. At the longest lead times (96 and 120 h; see Figs. 4d and 4e, respectively) OBEST outperforms ECM in just over 50% of all cases. The decreasing advantage of OBEST at increasingly longer lead times is a sign of nonlinear error growth and saturation when initially small-scale and/or small-amplitude initial condition or forecast model errors begin to alter the larger-scale environmental flow that can significantly affect the track forecasts (Lorenz 1969; Zhang et al. 2007; Gilmour et al. 2001). Nevertheless, at both these two long lead times, OBEST has noticeably smaller track errors for the ECM forecasts that have extremely large forecast errors (subjective thresholds at 800 km for the 96-h forecast and at 1000 km for the 120-h forecast). Combining all five of these lead times results in better performance of OBEST (compared with ECM) for 63.9% of the total 1419 valid forecast samples (Fig. 4f). Moreover, if we consider any forecast by either OBEST or ECM to be an outlier (very bad forecast) when it exceeds the 95th percentile threshold value out of the total combined forecast error samples at each respective lead time, the percentage of outliers is larger for ECM than OBEST at almost all lead times (except for at 120 h). The sum of the outliers for all five lead times is 164 for ECM and 122 for OBEST (Fig. 4f).

The use of ensemble forecasts, especially those from the ECMWF, has contributed to significant progress in CMA official track forecast accuracy[2] over the past decade (Fig. 5a) while improvement in intensity

[2] Verified against an independent best-track estimate from JMA.

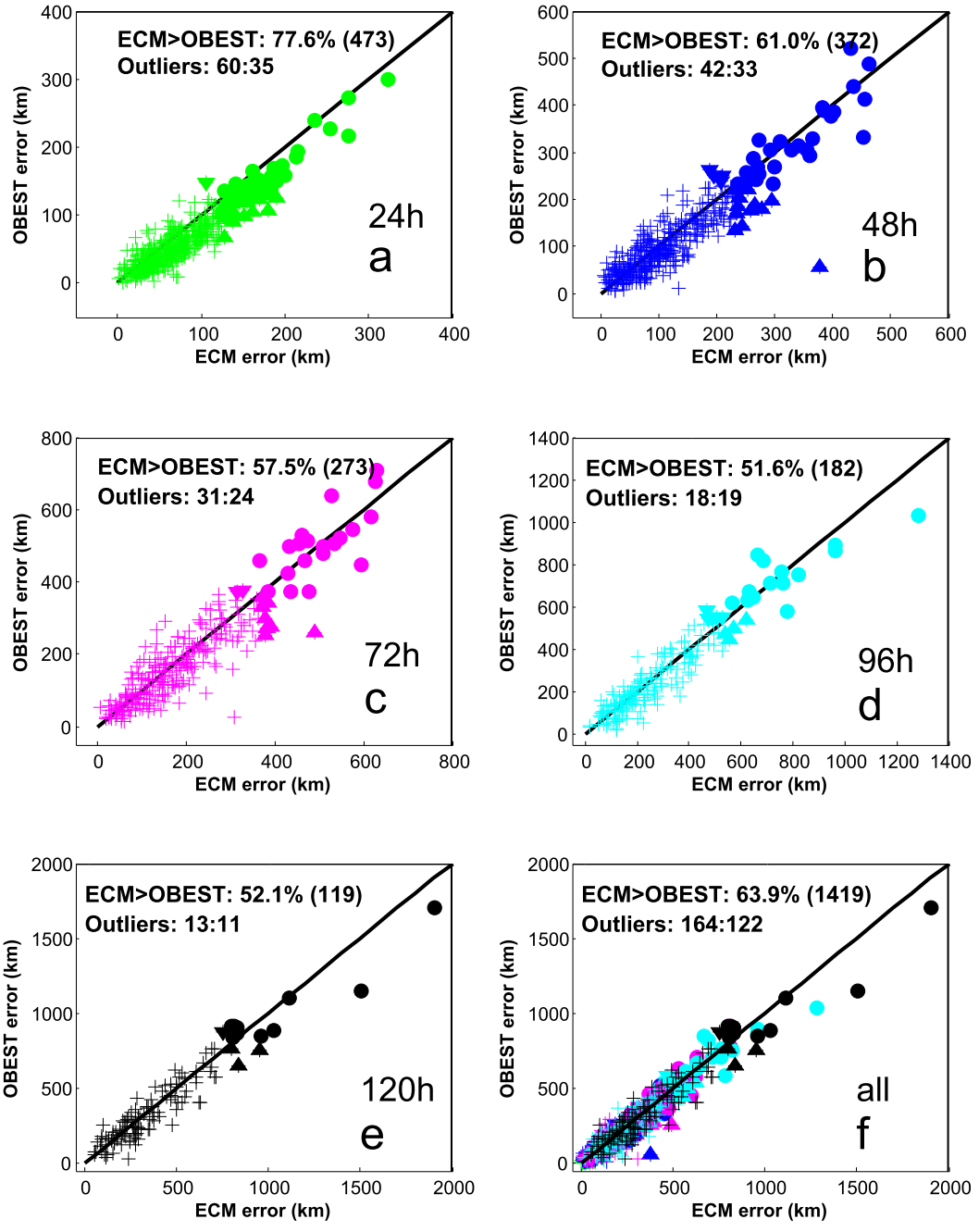

FIG. 4. Scatterplots of consensus track errors from OBEST with EC-EPS vs those from ECM for all valid forecast samples for 2012 and 2013 at different forecast hours: (a) 24, (b) 48, (c) 72, (d) 96, and (e) 120 h, as well as (f) all five times (numbers in the parentheses are the total number of samples). The number of outliers at each lead time, and the percentage of samples where ECM error is greater than OBEST, are marked in each panel. Solid dots indicate that both ECM and OBEST errors are outliers, upward (downward)-pointing triangles are used when only ECM (OBEST) error is an outlier, and crosses are for forecasts where neither ECM nor OBEST is an outlier.

forecasting has been slow to materialize (Fig. 5b), similar to other TC forecast centers around the world (Cangialosi and Franklin 2014). More specifically in Fig. 5a, the track forecast error has been decreasing steadily but slowly from 2004 to 2009 while there is a sharp reduction since 2010 when the ECMWF EPS was introduced as the primary forecast guidance. Also remarkable is that Fig. 5a suggests there may have been

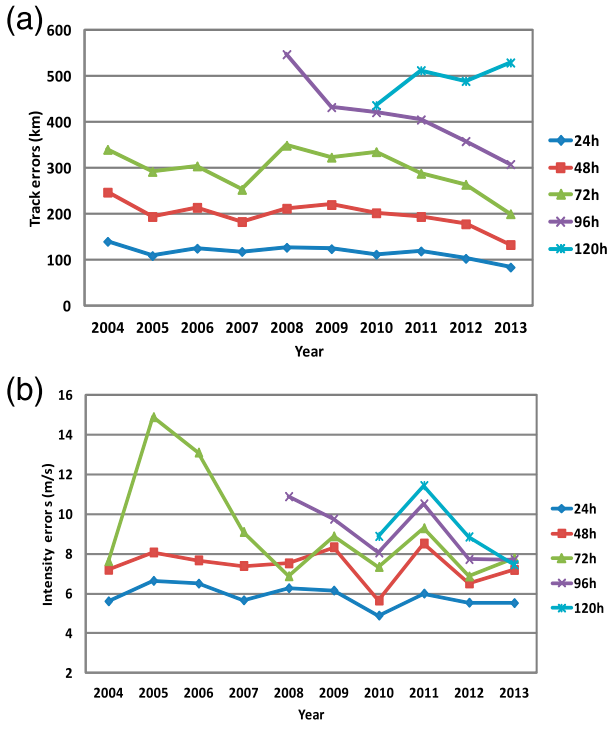

FIG. 5. Evolution of annual mean absolute CMA official TC forecast errors from 2004 to 2013 for (a) track errors (km) and (b) intensity errors ($m\,s^{-1}$).

further track forecast improvement through the introduction of the OBEST method since 2012. Although there is considerable variability at the track forecast accuracy from year to year, it is certainly encouraging to observe that the CMA official track forecast errors at 24-, 48-, 72-, and 96-h lead times were on average reduced from 120, 195, 289, and 406 km in 2011 to 85, 134, 201, and 309 km in 2013, leading to improvements of 29%, 31%, 30%, and 24% over the 2-yr period, respectively.

4. Super-OBEST

The success of OBEST following its operational implementation at CMA since 2012 encouraged us to investigate how this EPS-based consensus forecast model can be further improved. One way to do so is through multimodel ensembles (sometimes called superensembles), a topic that has seen more research and operational attention in recent years. For example, Pearman (2011) combined the EPSs from ECMWF, NCEP, and the Met Office to form a superensemble and found that track forecasts from the superensemble mean (using equal weights) were generally more accurate than the forecasts from any of the individual component models.

Motivated by Pearman (2011) and other recent successes in using multimodel ensembles, we explore the potential benefit of a 72-member superensemble version of OBEST (Super-OBEST), by adding the 21-member NCEP-EPS to the 51-member EC-EPS[3] already used in the operational OBEST. We only selected the NCEP and ECMWF EPSs because these two provide the most reliable TC track guidance out of all EPSs available in real time to operational TC forecasters at CMA.

We first examine the performance of the NCEP-EPS ensemble mean without subsetting in comparison with the full EC-EPS ensemble mean and the superensemble mean (SUPERM) that includes both NCEP and ECMWF ensembles without subsetting; the control high-resolution deterministic forecasts from both centers (ECD and NCEPD) are also shown in Fig. 6a. It is found that ECM performs similarly to NCEP-EPS ensemble mean (NCEPM) at ensemble forecast times[4] before 60 h, whereas NCEP-EPS outperforms at longer lead times. As a reference, the ensemble mean of each EPS performs similarly to its respective high-resolution deterministic forecast for forecast lead times up to 72 h, but better at longer lead times.

We also compare the performance of the track forecasts from both ECM and NCEPM for all storms at all lead times (Fig. 7). Except at lead time 96 h, ECM is slightly more often accurate than NCEPM [ranging from just over 50.3% of all 24-h forecasts (Fig. 7a) to 53.8% of all 120-h forecasts (Fig. 7e)]; overall, ECM yields the more accurate forecast nearly 50.6% of the time (Fig. 7f). The marginally better forecast skill of EC-EPS over NCEP-EPS is consistent with similar findings of Lee and Wong (2002), despite the fact that considerable changes have been made to both EPSs in the past decade. That being said, however, Fig. 7 indicates that the NCEP ensemble mean has fewer outliers of extremely erroneous track forecasts, most evident at 72, 96, and 120 forecast hours (Figs. 7c–e).

Given the rather comparable performance of the two ensembles (NCEP vs EC) in terms of TC track forecasts (Figs. 6a and 7), it is natural to combine them into a superensemble that will not only enlarge the ensemble size but also potentially compensate for the model errors in each ensemble. As shown in Fig. 6a, given a slightly

[3] We only include the 0000 and 1200 UTC NCEP-EPS forecasts to be compatible with EC-EPS. The NCEP ensembles initialized at 0600 and 1800 UTC will be considered in future operational implementations.

[4] The first effective forecast lead time is 24 h because the ensemble prediction systems are only available in real time at 12 h of ensemble integration time, which is used as the x axis in Fig. 6a.

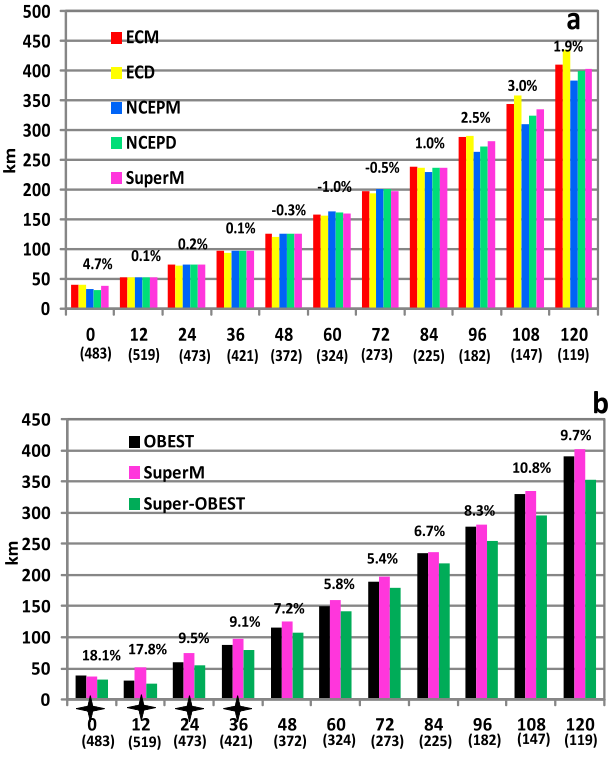

FIG. 6. (a) Comparison of forecast performance in terms of mean absolute track error verified against JMA best track among ECD, ECM, NCEPD, NCEPM, and SUPERM at different forecast lead times for 2012–13. (b) Comparison of the performance with Super-OBEST vs SUPERM and the control OBEST with EC-EPS only. The percentage values are improvements in the superensemble mean against ECM in (a) and Super-OBEST against control OBEST in (b), with those differences significant above the 95% confidence level noted by a star. The sample sizes in the comparison are noted in parentheses.

smaller average track error for ECM than NCEPM at shorter lead times and a slightly larger error for ECM than NCEPM at longer lead times, SUPERM will have a slightly larger average track error than ECM at shorter lead times and a slightly smaller error than ECM at longer lead times, though none of the differences are statistically different (not shown).

Does the advantage of the OBEST still apply for the combined NCEP and ECMWF superensemble? Using the combined superensemble (which has a total of $N = 72$ members), we tested different sizes of the ensemble subset (i.e., the value of M) using a procedure similar to that in section 2b and Fig. 2a; we found that the best performance for Super-OBEST is obtained when $M = 28$ members are selected in the ensemble subset (Fig. 2b). Figure 6b directly compares the means of the track errors for the control OBEST (EC-EPS only as in sections 2 and 3), the superensemble version of OBEST (Super-OBEST), and the superensemble mean without subsetting (SUPERM), Super-OBEST has considerably smaller mean track errors than the control OBEST at almost all lead times (though again only at shorter lead times are the results statistically significant at the 95% significance level, as marked with stars in Fig. 6b). More specifically, the mean position forecast errors are reduced from the control OBEST of 60, 116, 189, 277, and 390 km to 55, 108, 179, 254, and 352 km for model forecast times of 24, 48, 72, 96, and 120 h (subtract 12 h to have the effective lead forecast times), respectively, which represent an improvement over the control OBEST in the track forecast accuracy by averages of 9.5%, 7.2%, 5.4%, 8.3%, and 9.7%, respectively.

Compared to SUPERM, Super-OBEST is superior at all lead times. Despite this modest improvement in the mean sense (Fig. 6b), the percentage of forecasts among all valid samples for which Super-OBEST has a lower track error than SUPERM ranges from 84.1% at hour 24 (Fig. 8a) to around 65% at hours 96 and 120 (Figs. 8d,e); aggregating over all forecast lead times, Super-OBEST yields a more accurate track than SUPERM in 73.1% of all samples (Fig. 8f). Super-OBEST also has a noticeably smaller number of forecast outliers than SUPERM. Moreover, there is also a much greater chance for Super-OBEST to have smaller track errors than for control OBEST with EC-EPS only when evaluating individual forecasts at different lead times (Fig. 9). There is also noticeable improvement in accuracy for forecasts with very large errors (outliers).

5. An example of an OBEST forecast: Tropical Storm Cimaron (2013)

The use of OBEST for TC forecast tracks is exemplified in the forecast of TS Cimaron (2013). Cimaron was a relatively short-lived tropical cyclone that developed from a tropical disturbance on 16 July 2013 and that subsequently developed into a tropical storm with a 10-min maximum sustained wind of 20 m s^{-1} by 0000 UTC 17 July 2013. At that time, Cimaron was moving northwestward under the influence of the western North Pacific subtropical high. The operational challenge in Cimaron's track forecast was to evaluate whether the subtropical high would weaken, which would lead Cimaron northward and eventually recurve, and potentially make landfall on the southeastern coast of China, or if the subtropical high would strengthen, resulting in a west-southwestward track toward Vietnam. The track uncertainty is partially reflected by the divergence of ensemble tracks within the EC-EPS single ensemble (Fig. 10a), and even more so among the members of the superensemble with the addition of NCEP-EPS (Fig. 10b). In reality, Cimaron took the track

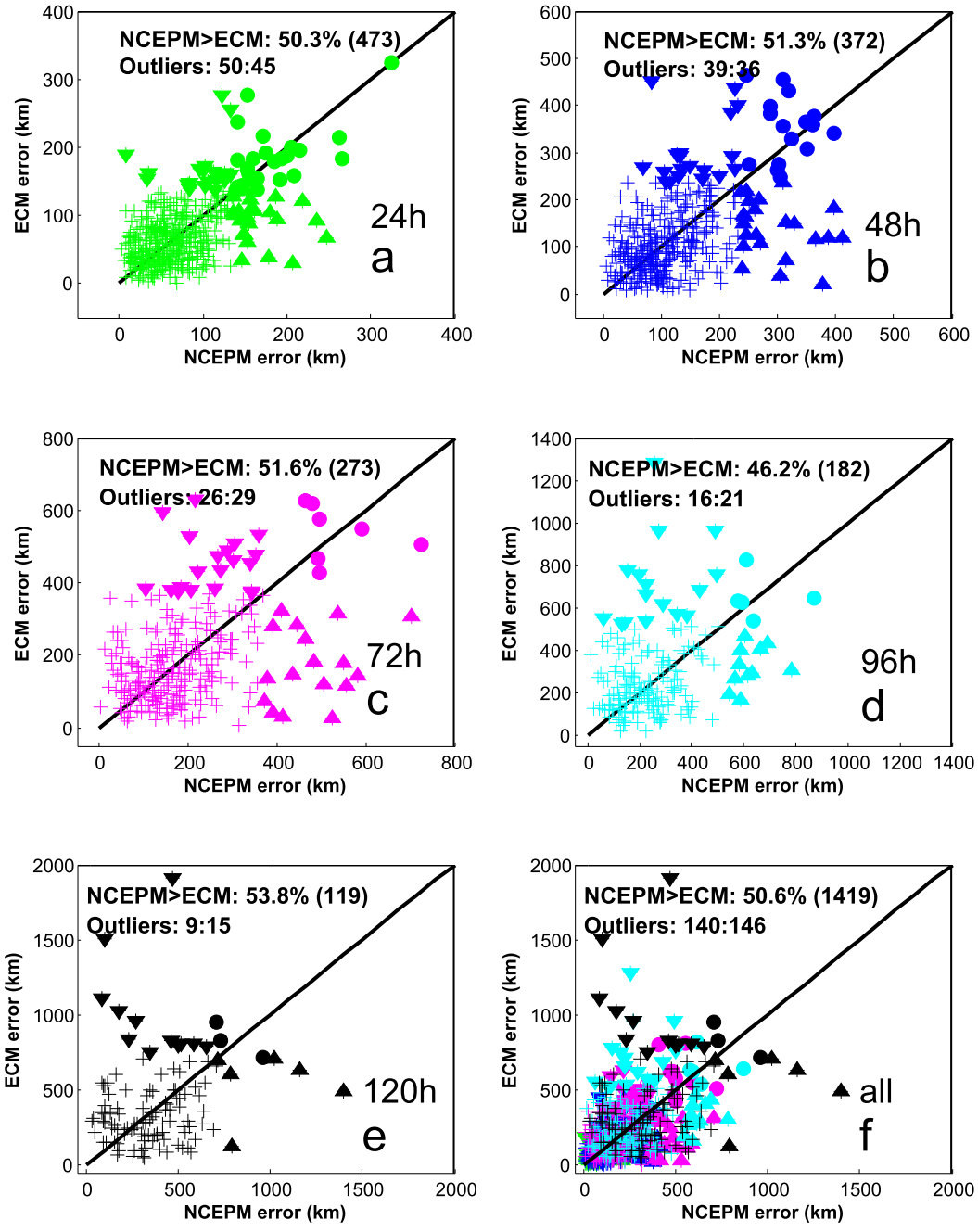

FIG. 7. As in Fig. 4, but for scatterplots of consensus track errors from the ensemble mean with NCEPM vs ECM for all valid forecast samples for 2012 and 2013.

toward the southeastern seaboard of China, as a result of the weakening of the subtropical high over the western North Pacific region.

For this particular event at this forecast time, a majority of EC-EPS members (and thus ECM) predicted a west-southwestward track of the TC toward Vietnam, as a consequence of the strengthening subtropical high that stretches farther westward. The OBEST with the EC-EPS improves upon the ECM considerably, especially for the first 48 h, through the selection of ensemble members that have smaller 12-h forecast errors (Fig. 10a).

In contrast, a majority of NCEP-EPS members (and thus NCEPM) predicted northward-curved tracks

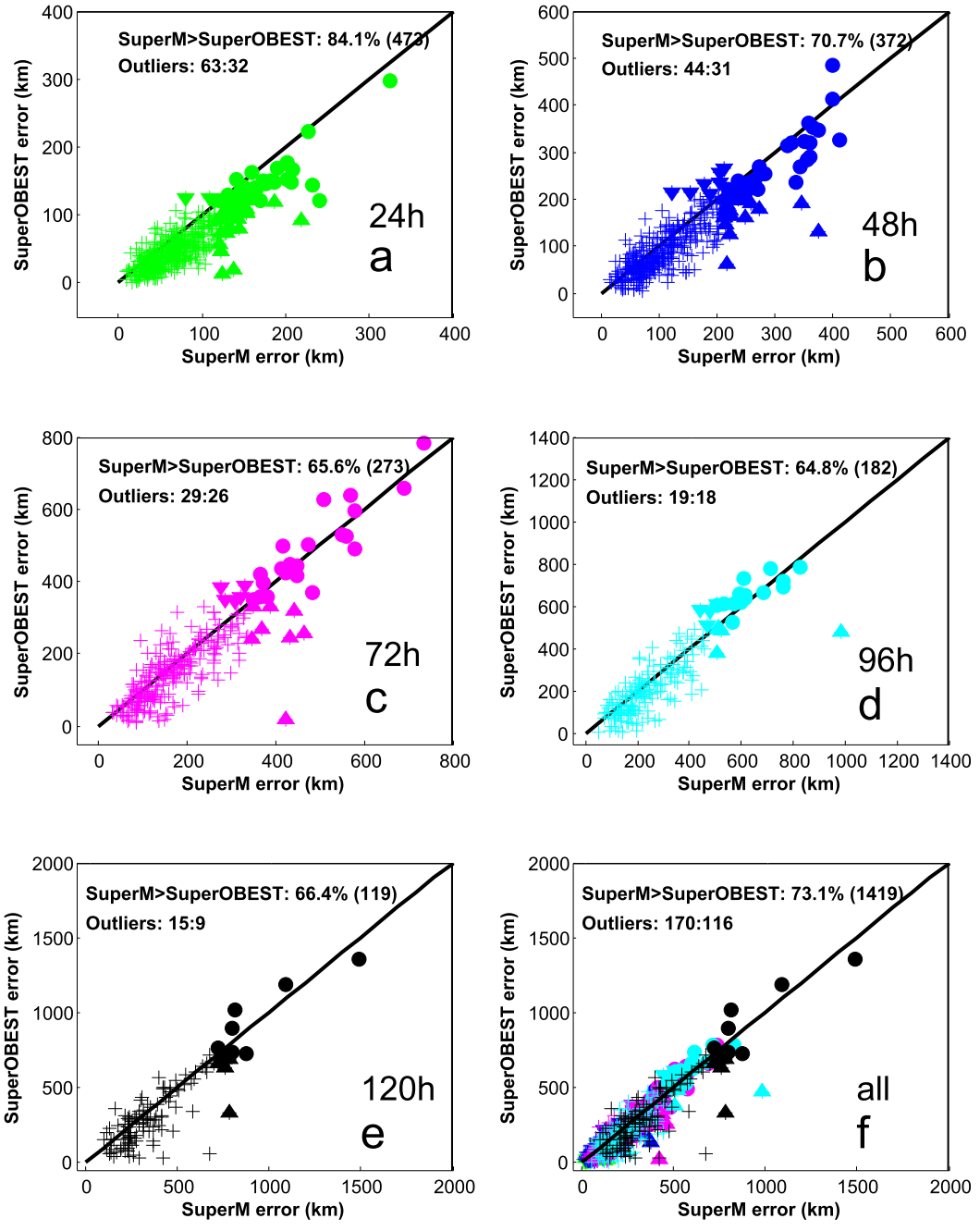

FIG. 8. As in Fig. 4, but for scatterplots of consensus track errors from SUPERM vs Super-OBEST for all valid forecast samples for 2012 and 2013.

(Fig. 10b). The track of SUPERM is closer to the best track than either ECM or NCEPM, which highlights the importance of including more than one ensemble prediction system to form a superensemble. Furthermore, Super-OBEST utilizing the EC-EPS and NCEP-EPS combined superensemble correctly predicted the northern turn of the TC that had the least mean error (Fig. 10b). In addition to the reduction in track error, another benefit of superensembles is that they provide more uncertainty information to forecasters and decision-makers (e.g., Elsberry and Carr 2000; Yamaguchi et al. 2009; Majumdar and Finocchio 2010; Yamaguchi et al. 2012). It is also worth noting that the track forecast of Cimaron or similar storms in operations

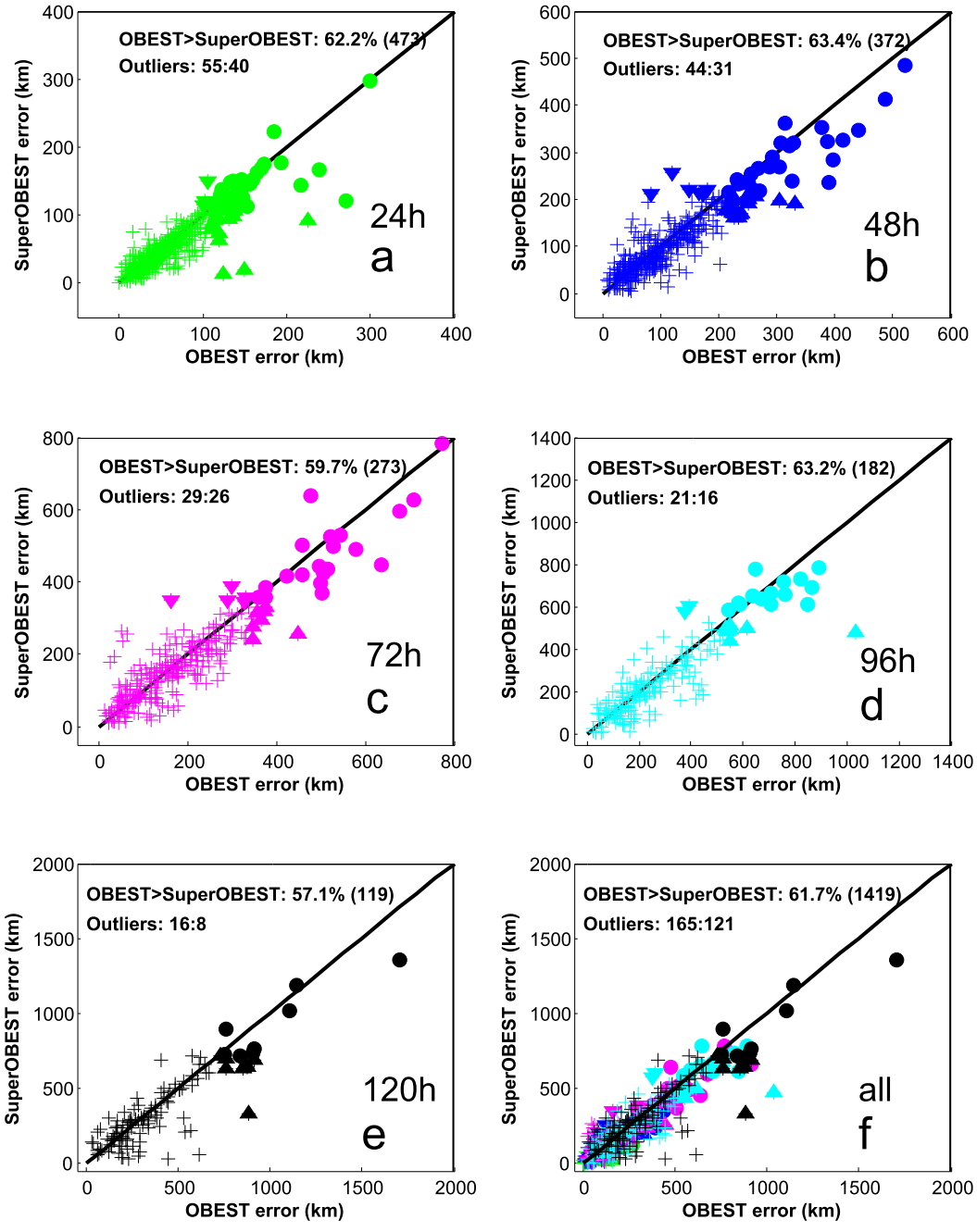

FIG. 9. As in Fig. 4, but for scatterplots of consensus track errors from Super-OBEST vs the control OBEST with EC-EPS only for all valid forecast samples for 2012 and 2013.

is particularly challenging because of its weak intensity, which will be further explored in future studies.

6. Concluding remarks

An observation-based ensemble subsetting technique (OBEST) is developed for tropical cyclone (TC) track prediction in which a subset of members from either a single- or multimodel ensemble is selected based on the distance from the latest best-track position. The performance of OBEST is examined using the THORPEX Interactive Grand Global Ensemble (TIGGE) dataset as archived by the China Meteorological Administration (CMA).

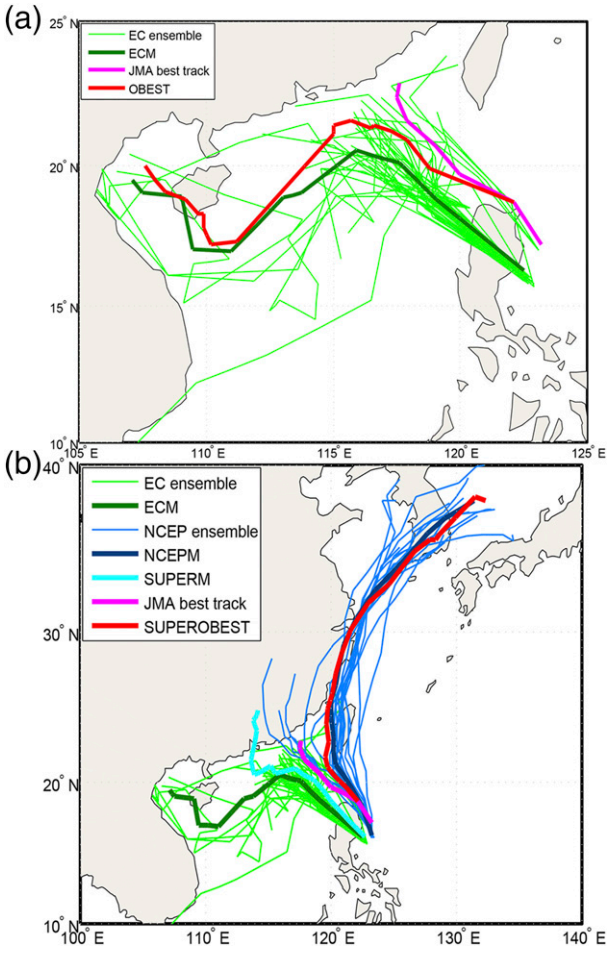

FIG. 10. The JMA best-track observations vs different forecast tracks for TS Cimaron initialized at 1200 UTC 16 Jul 2013. Shown are (a) the ensemble track forecasts of EC-EPS and OBEST with EC-EPS only and (b) the ensemble track forecasts of EC-EPS, NCEP-EPS, and OBEST with the combined superensemble.

Results from both the 2-yr hindcasts for 2010–11 and the 2-yr operational predictions during 2012–13 show that OBEST outperforms both the simple ensemble mean (without subsetting) and the corresponding deterministic high-resolution control prediction at most of the forecast lead times up to 5 days. The operational implementation of OBEST using EC-EPS may have led to considerable track forecast improvements at CMA since 2012.

Further reduction in track forecast errors by as much as 5%–10% for 24–120-h forecasts in experiments was found with the application of OBEST to a superensemble consisting of global ensembles from both the European Centre for Medium-Range Weather Forecasts (ECMWF) and the National Centers for Environmental Prediction (NCEP). The illustration and comparison of OBEST with the EC-EPS single ensemble versus the ECMWF–NCEP combined superensemble is exemplified in the forecast of Tropical Storm (TS) Cimaron (2013). Super-OBEST performed better than either ensemble mean or the single-ensemble OBEST.

However, despite the promising results shown in this study, there are potential drawbacks of this newly proposed technique, which include, but are not limited to, 1) a relatively larger number of ensemble members is required to apply OBEST, 2) this technique may be more limited for those EPSs whose individual member storms have gone through the vortex relocation procedure (although limited tests have shown there is still advantage to applying the OBEST for NCEP-EPS, which has vortex relocation), 3) OBEST may be affected by initial model spinup–spindown issues, and 4) there will be some delay in the forecast delivery time. Moreover, it is unclear whether the dispersion of the member ensemble(s) matters when applying OBEST. For instance, if subsampling from an underdispersive (and limited sized) ensemble, there are likely to be cases in which the method performs very well but also cases where the nonlinearity of the system causes the method to perform not so well (e.g., good short-term forecasts but poor long-term forecasts). These issues will be further examined in future studies as well as in practical operations. Future studies will also systematically examine the potential of using OBEST for probabilistic forecasts, along with the inclusion of more ensemble prediction systems into a greater superensemble.

Acknowledgments. The authors are grateful to the TIGGE team for constructing useful and user-friendly portal sites and providing analysis and forecast data of operational ensemble prediction systems. LD thanks her colleagues in the typhoon forecasting team under the leadership of Chuanhai Qian at CMA for their contributions to the development and implementation of the OBEST method. The authors are also grateful to Qifeng Qian and Yonghui Weng for their generous help in processing the TIGGE dataset and to Benjamin Green for proofreading an earlier version of the manuscript. We also benefited greatly from review comments by three anonymous reviewers. FZ was partially supported by ONR Grant N000140910526 and NOAA HFIP. Computing at the Texas Advanced Computing Center (TACC) is acknowledged.

REFERENCES

Bougeault, P., and Coauthors, 2010: The THORPEX Interactive Grand Global Ensemble. *Bull. Amer. Meteor. Soc.*, **91**, 1059–1072, doi:10.1175/2010BAMS2853.1.

Buizza, R., J.-R. Bidlot, N. Wedi, M. Fuentes, M. Hamrud, G. Holt, and F. Vitart, 2007: The new ECMWF VAREPS (Variable Resolution Ensemble Prediction System). *Quart. J. Roy. Meteor. Soc.*, **133**, 681–695, doi:10.1002/qj.75.

Burton, A., 2006: Sharing experiences in operational consensus forecasting. *Proc. Sixth Int. Workshop on Tropical Cyclones*, San Jose, Costa Rica, WMO/CAS/WWW, Topic 3a. [Available online at http://severe.worldweather.org/iwtc/document/Topic_3a_Andrew_Burton.pdf.]

Cangialosi, J. P., and J. L. Franklin, 2014: 2013 National Hurricane Center Forecast verification report. NOAA/National Hurricane Center, 84 pp. [Available online at http://www.nhc.noaa.gov/verification/pdfs/Verification_2013.pdf.]

Chan, S. T., 2010: Tropical cyclone operational warning strategies (IWTC-7). WMO/CAS/WWW, 27 pp. [Available online at http://www.wmo.int/pages/prog/arep/wwrp/tmr/otherfileformats/documents/4_3.pdf.]

Danard, M., 1977: Comments on ''How to improve accuracy by combining independent forecasts.'' *Mon. Wea. Rev.*, **105**, 1198–1199, doi:10.1175/1520-0493(1977)105<1198:COTIAB>2.0.CO;2.

Elsberry, R. L., 2007: Advances in tropical cyclone motion prediction and recommendations for the future. *WMO Bull.*, **56**, 131–135.

——, 2014: Advances in research and forecasting of tropical cyclones from 1963–2013. *Asia-Pac. J. Atmos. Sci.*, **50**, 3–16, doi:10.1007/s13143-014-0001-1.

——, and L. E. Carr III, 2000: Consensus of dynamical tropical cyclone track forecasts—Errors versus spread. *Mon. Wea. Rev.*, **128**, 4131–4138, doi:10.1175/1520-0493(2000)129<4131:CODTCT>2.0.CO;2.

——, J. R. Hughes, and M. A. Boothe, 2008: Weighted position and motion vector consensus of tropical cyclone track prediction in the western North Pacific. *Mon. Wea. Rev.*, **136**, 2478–2487, doi:10.1175/2007MWR2262.1.

Gilmour, I., L. A. Smith, and R. Buizza, 2001: Linear regime duration: Is 24 hours a long time in synoptic weather forecasting? *J. Atmos. Sci.*, **58**, 3525–3539, doi:10.1175/1520-0469(2001)058<3525:LRDIHA>2.0.CO;2.

Goerss, J. S., 2000: Tropical cyclone track forecasts using an ensemble of dynamical models. *Mon. Wea. Rev.*, **128**, 1187–1193, doi:10.1175/1520-0493(2000)128<1187:TCTFUA>2.0.CO;2.

——, C. R. Sampson, and J. M. Gross, 2004: A history of western North Pacific tropical cyclone track forecast skill. *Wea. Forecasting*, **19**, 633–638, doi:10.1175/1520-0434(2004)019<0633:AHOWNP>2.0.CO;2.

Jun, S., J. Kim, W. J. Lee, K. Y. Byun, K. H. Chang, and D. S. Shin, 2014: Objective consensus typhoon track forecasting using multimodel superensemble. *Proc. 31st Conf. on Hurricanes and Tropical Meteorology*, San Diego, CA, Amer. Meteor. Soc., 87. [Available online at https://ams.confex.com/ams/31Hurr/webprogram/Paper243915.html.]

Krishnamurti, T. N., S. Pattnaik, M. K. Biswas, E. Bensman, M. Kramer, N. Surgi, and T. S. V. V. Kumar, 2010: Hurricane forecasts with a mesoscale suite of models. *Tellus*, **62A**, 633–646, doi:10.1111/j.1600-0870.2010.00469.x.

Kumar, T. S. V. V., T. N. Krishnamurti, M. Fiorino, and M. Nagata, 2003: Multimodel superensemble forecasting of tropical cyclones in the Pacific. *Mon. Wea. Rev.*, **131**, 574–583, doi:10.1175/1520-0493(2003)131<0574:MSFOTC>2.0.CO;2.

Lee, T. C., and M. S. Wong, 2002: The use of multiple-model ensemble techniques for tropical cyclone track forecast at the Hong Kong Observatory. Preprints, *Tech. Conf. on Data Processing and Forecasting Systems*, Cairns, QLD, Australia, WMO, Topic 2. [Available online at http://www.wmo.int/pages/prog/www/DPS/TC-DPFS-2002/Papers-Posters/Topic2-Lee.pdf.]

Leslie, L. M., and K. Fraedrich, 1990: Reduction of tropical cyclone position errors using an optimal combination of independent forecasts. *Wea. Forecasting*, **5**, 158–161, doi:10.1175/1520-0434(1990)005<0158:ROTCPE>2.0.CO;2.

Lorenz, E. N., 1969: The predictability of a flow which possesses many scales of motion. *Tellus*, **21**, 289–307, doi:10.1111/j.2153-3490.1969.tb00444.x.

Majumdar, S. J., and P. 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.

Mohapatra, M., D. P. Nayak, R. P. Sharma, and B. K. Bandyopadhydy, 2013: Evaluation of official tropical cyclone track forecast over north Indian Ocean issued by India Meteorological Department. *J. Earth Syst. Sci.*, **122**, 589–601, doi:10.1007/s12040-013-0291-1.

Pearman, D. W., 2011: Evaluating tropical cyclone forecast track uncertainty using a grand ensemble of ensemble prediction systems. M.S. thesis, Dept. of Meteorology, Naval Postgraduate School, 63 pp. [Available online at http://calhoun.nps.edu/handle/10945/5465.]

Qi, L., H. Yu, and P. Chen, 2014: Selective ensemble-mean technique for tropical cyclone track forecast by using ensemble prediction systems. *Quart. J. Roy. Meteor. Soc.*, **140**, 805–813, doi:10.1002/qj.2196.

Qian, C., Y. Duan, S. Ma, and Y. Xu, 2012: The current status and future development of China operational typhoon forecasting and its key technologies. *Adv. Meteor. Sci. Technol.*, **2** (5), 36–43.

Sampson, C. R., J. S. Goerss, and A. J. Schrader, 2005: A consensus track forecast for Southern Hemisphere tropical cyclones. *Aust. Meteor. Mag.*, **54**, 115–119.

Sanders, F., 1973: Skill in forecasting daily temperature and precipitation: Some experimental results. *Bull. Amer. Meteor. Soc.*, **54**, 1171–1179, doi:10.1175/1520-0477(1973)054<1171:SIFDTA>2.0.CO;2.

Thompson, P. D., 1977: How to improve accuracy by combining independent forecasts. *Mon. Wea. Rev.*, **105**, 228–229, doi:10.1175/1520-0493(1977)105<0228:HTIABC>2.0.CO;2.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 2317–2330, doi:10.1175/1520-0477(1993)074<2317:EFANTG>2.0.CO;2.

Weber, H. C., 2003: Hurricane track prediction using a statistical ensemble of numerical models. *Mon. Wea. Rev.*, **131**, 749–770, doi:10.1175/1520-0493(2003)131<0749:HTPUAS>2.0.CO;2.

Wei, M., Z. Toth, R. Wobus, and Y. Zhu, 2008: Initial perturbations based on the ensemble transform (ET) technique in the NCEP Global Operational Forecast System. *Tellus*, **60A**, 62–79, doi:10.1111/j.1600-0870.2007.00273.x.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences.* 2nd ed. Academic Press, 648 pp.

Williford, C. E., T. N. Krishnamurti, R. C. Torres, S. Cocke, Z. Christidis, and T. S. V. V. Kumar, 2003: Real-time multimodel superensemble forecasts of Atlantic tropical

systems of 1999. *Mon. Wea. Rev.*, **131**, 1878–1894, doi:10.1175//2571.1.

WMO, 2007: Sixth WMO International Workshop on Tropical Cyclone (IWTC-VI). WMO Rep. WWRP 2007-1, 92 pp. [Available online at http://www.aoml.noaa.gov/hrd/Landsea/WWRP2007_1_IWTC_VI.pdf.]

Yamaguchi, M., R. Sakai, M. Kyoda, T. Komori, and T. Kadowaki, 2009: Typhoon ensemble prediction system developed at the Japan Meteorological Agency. *Mon. Wea. Rev.*, **137**, 2592–2604, doi:10.1175/2009MWR2697.1.

——, T. Nakazawa, and S. Hoshino, 2012: On the relative benefits of a multi-centre grand ensemble for tropical cyclone track prediction in the western North Pacific. *Quart. J. Roy. Meteor. Soc.*, **138**, 2019–2029, doi:10.1002/qj.1937.

Zhang, F., N. Bei, R. Rotunno, C. Snyder, and C. C. Epifanio, 2007: Mesoscale predictability of moist baroclinic waves: Convection-permitting experiments and multistage error growth dynamics. *J. Atmos. Sci.*, **64**, 3579–3594, doi:10.1175/JAS4028.1.