Use of Multiensemble Track Clustering to Inform Medium-Range Tropical Cyclone Forecasts

ALEX M. KOWALESKI^a

*Department of Meteorology and Atmospheric Science, The Pennsylvania State University,
University Park, Pennsylvania*

JENNI L. EVANS

*Department of Meteorology and Atmospheric Science, and Institute for CyberScience,
The Pennsylvania State University, University Park, Pennsylvania*

(Manuscript received 6 January 2020, in final form 7 May 2020)

ABSTRACT

Tropical cyclone ensemble track forecasts from 153 initialization times during 2017–18 are clustered using regression mixture models. Clustering is performed on a four-ensemble dataset [ECMWF + GEFS + UKMET + CMC (EGUC)], and a three-ensemble dataset that excludes the CMC (EGU). For both datasets, five-cluster partitions are selected to analyze, and the relationship between cluster properties (size, ensemble composition) and 96–144-h cluster-mean error is evaluated. For both datasets, small clusters produce very large errors, with the least populous cluster producing the largest error in more than 50% of forecasts. The mean of the most populous EGUC cluster outperforms the most accurate (EGU) ensemble mean in only 43% of forecasts; however, when the most populous EGUC cluster from each forecast contains $\geq 30\%$ of the ensemble population, its average cluster-mean error is significantly reduced compared to when the most populous cluster is smaller. Forecasts with a highly populous EGUC cluster also appear to have smaller EGUC-, EGU-, and ECMWF-mean errors. Cluster-mean errors also vary substantially by the ensembles composing the cluster. The most accurate clusters are EGUC clusters that contain threshold memberships of ECMWF, GEFS, and UKMET, but *not* CMC. The elevated accuracy of EGUC CMC-excluding clusters indicates the potential utility of including the CMC in clustering, despite its large ensemble-mean errors. Pruning ensembles by removing members that belong to small clusters reduces 96–144-h forecast errors for both EGUC and EGU clustering. For five-cluster partitions, a pruning threshold of 10% affects 49% and 35% of EGUC and EGU ensembles, respectively, improving 69%–74% of the forecasts affected by pruning.

1. Introduction

The dramatic reduction of tropical cyclone (TC) track forecast errors during the last 25 years is among the most striking meteorological forecasting successes. In the North Atlantic, mean 72-h track forecast error from the National Hurricane Center (NHC) decreased from approximately 460 km (248 nm; 1 nm = 1.852 km) in 1995 to 290 km in 2005 and 185 km in 2018. Current 72-h track forecasts have similar errors to 48-h forecasts from the mid-2000s and 24-h forecasts from the early to mid-1990s (NHC 2019). These track forecast improvements

prompted the NHC to begin issuing 5-day forecasts in 2003 (Rappaport et al. 2009), and to extend watch and warning lead times by 12 h in 2010, providing greater time to prepare for TC hazards.

Near-continuous improvements in numerical weather prediction models have driven the increase in track forecast accuracy. These improvements include higher model resolution, improved data assimilation, better physical parameterizations, and increasing quantity and quality of observations used to initialize the models (Landsea and Cangialosi 2018). An additional contributor to TC track error reductions is the use of consensus models (e.g., Goerss 2000; Goerss et al. 2004; Krishnamurti et al. 2011), which are a simple or weighted average of track forecasts from multiple dynamical models.

Despite the long-term downward trend in TC track forecast error, there is some evidence that forecast skill

^a Current affiliation: Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, Colorado.

Corresponding author: Jenni L. Evans, jle7@psu.edu

has leveled off during the last five years, at least in the North Atlantic and eastern North Pacific basins, suggesting that operational forecasts may be approaching the limit of TC track predictability (Landsea and Cangialosi 2018). Furthermore, at current track forecast skill, track predictability remains a major limiter of TC hazard predictability at individual locations; additional track forecast improvements would yield tangible benefits for TC hazard forecasts. The extreme winds of an intense TC typically cover only a small area near its center, so predicting if and when a location will experience these extreme winds requires a highly accurate track forecast. For storm surge, Fossell et al. (2017) demonstrated that with current track forecast errors, storm surge inundation at individual locations is predictable no more than 12–24 h before landfall. Greater track forecast accuracy would allow useful wind and surge hazard forecasts at longer lead times, giving stakeholders more time for protective actions.

While previous TC track forecast improvements have primarily stemmed from superior deterministic models, better utilization of global ensemble forecasts may also have potential for reducing operational TC track forecast errors. In ensemble forecasts, the initial atmospheric conditions of each ensemble member are perturbed to reflect atmospheric uncertainty at the time of model initialization. These perturbations cause the atmosphere to evolve differently in each member, generating a set of potential atmospheric scenarios. An ensemble may also employ multiphysics (e.g., Berner et al. 2011; Jankov et al. 2017) or stochastic physics (e.g., Berner et al. 2017; Melhauser et al. 2017) to represent uncertainty in parameterized atmospheric physical processes. Ensemble forecasts are typically run at a lower spatial resolution than their corresponding deterministic forecast due to computational limitations.

Ensemble forecasts of TC track are typically described by the ensemble mean, the ensemble spread, and strike probabilities. The mean is the arithmetic average of the TC latitude and longitude among all ensemble members at each time point, while the ensemble spread represents the set of possible track scenarios. The ensemble spread also provides a measure of forecast confidence; forecasts with larger track spread tend to have larger ensemble-mean track errors (Yamaguchi et al. 2009; Majumdar and Finocchio 2010). Strike probabilities employ the ensemble track distribution to compute the probability that the TC will pass within a specified distance (e.g., 120 km) from each location, at a specific time or within a time range (Majumdar and Finocchio 2010; Dupont et al. 2011; Yamaguchi et al. 2012; Titley et al. 2020).

Although TC ensemble track forecasts provide skillful probabilistic guidance (Titley et al. 2020), and some

forecasting centers have begun to use probabilistic ensemble output in their track forecasts (Titley et al. 2019), TC forecasters still operate primarily in a deterministic framework, as their primary responsibility is to generate deterministic forecasts of TC position and intensity. In ensembles, a deterministic track forecast can be represented by the ensemble mean. The ensemble mean is a simple position average; thus, excepting cases such as ensemble bifurcation [e.g., Hurricane Joaquin (2015); Nystrom et al. 2018] it is more likely to be physically realistic than ensemble means of fields with large spatial variance, such as convective precipitation. Leonardo and Colle (2017) examined ensemble-mean track forecasts out to 120 h from the European Centre for Medium-Range Weather Forecasts (ECMWF), the U.S.'s National Centers for Environmental Prediction, and the Met Office (UKMET) for North Atlantic TCs from 2008 to 2015. They found that ensemble-mean forecasts had slightly larger average errors than their deterministic counterparts; however, they were less often the poorest performing forecast compared to their deterministic counterparts. The ECMWF ensemble mean, in particular, showed substantial accuracy.

For the North Atlantic and eastern North Pacific basins, the NHC has primarily relied on consensus track guidance computed from operational deterministic model output; however, in recent years it has begun to include ensemble-mean forecasts in its highly skilled consensus guidance, indicating their operational forecasting utility. The HFIP-Corrected Consensus Model (HCCA; Simon et al. 2018), which includes the ECMWF and GEFS ensemble means, debuted in 2016 (Cangialosi and Franklin 2017). The Variable Consensus (TVCA) Model added the ECMWF ensemble mean in 2019 (Cangialosi 2019).

The recent inclusion of ensemble-mean forecasts in consensus TC track guidance demonstrates why improving ensemble forecasts can have practical implications for operational track forecasting. Toward this end, substantial effort in recent years has focused on extracting additional information from ensembles. One of these research avenues seeks to improve ensemble-mean forecasts by selecting or weighting individual members by their performance at short lead times (e.g., 12 h), near the time that the forecasts become available. Qi et al. (2014), Dong and Zhang (2016), and Zhang and Yu (2017) each demonstrated a variation of this method, and each study produced track forecast error reductions. Dong and Zhang (2016) found that for a combined ECMWF + GEFS ensemble, selecting the 28 of 72 members with the smallest 12-h track errors yielded the most accurate mean track forecasts, outperforming the

unweighted ensemble mean on average at all lead times and improving 73.1% of forecasts.

Clustering of ensemble track forecasts provides another method to extract information from a single- or multimodel ensemble. By partitioning ensemble members to minimize intracluster spread and maximize intercluster spread, clustering highlights TC track groupings within the data. Kowaleski and Evans (2016) applied regression mixture-model track clustering (Gaffney et al. 2007) to forecasts of Hurricane Sandy from four global ensemble prediction systems (EPSs) to show how clustering illustrates distinct scenarios of Sandy's evolution. Kowaleski and Evans (2018) clustered Sandy's track among 72 regional WRF simulations, derived from ECMWF and GEFS ensemble initial conditions, to illustrate the relationship between Sandy's track and its structural evolution prior to landfall.

In this study, we apply track clustering to 153 TC track forecasts from ensembles that comprise multiple EPSs. We evaluate the relationship between cluster properties (size, ensemble composition) and medium-range (96–144 h) cluster-mean track forecast error. The 96–144-h lead time is near and beyond the edge of current operational TC forecasts; the challenge of improving operational guidance is why we choose to focus on this lead time. Finally, we demonstrate how “pruning” small clusters, or excluding ensemble members that belong to clusters below a threshold size, typically reduces ensemble-mean forecast error. Although we focus on deterministic ensemble-mean forecasts, the techniques applied here may also be relevant to scenario-based and probabilistic forecasting.

The rest of the paper proceeds as follows: section 2 describes the dataset used, EPS performance, clustering methodology, and selection of the mixture-model characteristics used in clustering. Clustering results are presented in section 3, and ensemble pruning is described in section 4. Results are summarized and conclusions presented in section 5.

2. Data and methodology

a. Dataset

Track forecasts from 153 initialization times for 36 TCs occurring in 2017–18 are obtained via the UCAR TIGGE database (NOAA/NWS/NCEP 2020; <https://rda.ucar.edu/datasets/ds330.3>). TIGGE data employed in this study are provided by the ECMWF (51 members), the NCEP's Global Ensemble Forecast System (GEFS; 21 members), the Met Office Global Ensemble Prediction System (UKMET; 36 members), and the Environment

Canada Global Ensemble Prediction System (CMC; 21 members). Each initialization time is only included if the TC is tracked out to 144 h in at least 50% of each EPS and in the best track. The study includes TC forecasts from the North Atlantic (60), eastern North Pacific (36, including central Pacific), western North Pacific (42), South Pacific (11), and south Indian (4) basins (Table 1). Forecasts from individual basins are aggregated to produce a larger population to evaluate.

Best track data from the North Atlantic and eastern North Pacific are obtained from NHC poststorm reports. These reports are identical to HURDAT2 data (HRD 2020), except for storm positions west of 140°W, which were provisional at the time they were obtained. Western North Pacific best track data are obtained from the Japan Meteorology Agency best track archive (JMA 2020). Southern Hemisphere data are obtained from IBTrACS (NCEI 2020), via the UNC Asheville IBTrACS site (UNCA 2020).

Storm positions in each ensemble member are obtained at 6-h intervals out to 144 h. Among the four EPSs, varying vortex trackers are used to track the TC, but all utilize sea level pressure and lower-tropospheric vorticity. The ECMWF tracker allowing a storm to “disappear” for up to 24 h (Vitart et al. 2012). In these cases, TC position is interpolated from values at the time steps bracketing the missing time(s). In rare cases of missing 0-h ECMWF position data, TC position in that member is estimated as the mean initial position from the other ECMWF members.

b. Ensemble performance

Across the 153 forecasts, the ECMWF ensemble produces the most accurate ensemble-mean forecast at all lead times beyond 12 h (Fig. 1), consistent with the results of Leonardo and Colle (2017) for North Atlantic TC forecasts from 2008 to 2015. Among the four EPSs, the ECMWF has the lowest ensemble-mean total error, absolute along-track (AT) error, and absolute cross-track (XT) error averaged across the 9 time steps from 96 to 144 h (96–144-h error; Table 2). The UKMET generates the second-best ensemble mean, a departure from Leonardo and Colle (2017), who found that the GEFS outperformed the UKMET. On average, the CMC is the worst-performing EPS (total, AT, and XT) at all lead times. Among the mean tracks of all multi-EPS combinations, the ECMWF+GEFS+UKMET (EGU) ensemble mean produces the most accurate 96–144-h forecasts, with an average total error of 268 km (Table 2). Therefore, we perform clustering on the ECMWF+GEFS+UKMET+CMC (EGUC) ensemble, because it contains forecasts from full dataset,

TABLE 1. List of the 36 tropical cyclones and 153 forecasts clustered in this study. Storm cases are drawn from five ocean basins.

Tropical cyclone	Basin	Year	No. of forecasts	Tropical cyclone	Basin	Year	No. of forecasts
Harvey	North Atlantic	2017	2	Rosa	East Pacific	2018	1
Irma	North Atlantic	2017	8	Sergio	East Pacific	2018	7
Jose	North Atlantic	2017	13	Lan	West Pacific	2017	2
Maria	North Atlantic	2017	7	Noru	West Pacific	2017	10
Florence	North Atlantic	2018	11	Talim	West Pacific	2017	4
Helene	North Atlantic	2018	4	Jebi	West Pacific	2018	2
Leslie	North Atlantic	2018	13	Kong-rey	West Pacific	2018	2
Michael	North Atlantic	2018	2	Mangkhut	West Pacific	2018	4
Fernanda	East Pacific	2017	2	Man-yi	West Pacific	2018	1
Greg	East Pacific	2017	1	Maria	West Pacific	2018	2
Hilary	East Pacific	2017	2	Shanshan	West Pacific	2018	3
Irwin	East Pacific	2017	3	Soulik	West Pacific	2018	2
Otis	East Pacific	2017	1	Trami	West Pacific	2018	4
Fabio	East Pacific	2018	2	Yutu	West Pacific	2018	6
Hector	East Pacific	2018	7	Berguita	South Indian	2018	1
Lane	East Pacific	2018	4	Marcus	South Indian	2018	3
Norman	East Pacific	2018	2	Gita	South Pacific	2018	4
Olivia	East Pacific	2018	4	Iris	South Pacific	2018	7

and the EGU ensemble, because it has the smallest average ensemble-mean error across all forecasts.

c. Clustering methodology

Each EGUC and EGU forecast is clustered using the regression mixture-model clustering technique developed by Gaffney et al. (2007), as applied by Kowaleski and Evans (2016, 2018). In regression mixture-model clustering, each ensemble track is assigned probabilistically to each model (cluster) by its fit to the central trajectories (latitude and longitude with time), error covariance matrix, and the weight (population) defining that model. Cluster assignments are performed through an expectation-maximization (E-M) algorithm (see Gaffney et al. 2007 and Camargo et al. 2007 for details) that (i) calculates each cluster's parameters (polynomial coefficients, error covariance matrix, and cluster size), and (ii) calculates each track's membership probabilities to maximize the likelihood (fit) of the mixture model. Because the iterative E-M algorithm may converge on a local, rather than global, likelihood maximum, this process is repeated 1000 times using random initial cluster membership probabilities to increase the probability that at least one repetition converges on the global likelihood maximum. The solution with the highest likelihood (best overall fit) is then selected. An example of EGUC clustering using forecasts of Tropical Storm Soulik initialized at 0000 UTC 16 August 2018 is shown in Fig. 2.

Regression mixture-model clustering requires initial assignment of the number of clusters and polynomial order of the central cluster trajectories. To assess the effectiveness of different number/polynomial values, we cluster all forecasts using polynomial orders 1–6 and

number of clusters 2–9. Although optimal cluster characteristics vary forecast-to-forecast, results across all forecasts are examined to select a single combination of polynomial order and number of clusters to evaluate. For mixture-model clustering of ensemble TC tracks, Kuruppumullage Don et al. (2016) and Kowaleski and Evans (2016) found that selecting the optimal polynomial order is more straightforward than selecting the optimal number of clusters; therefore, we evaluate polynomial order first.

Relative cluster compactness (RCC) is the metric used to select the optimal polynomial order. For each forecast, RCC is defined as the ratio of the mean distance between each pair of members that are assigned to same cluster and each pair that are assigned to separate clusters, averaged from 0 to 144 h.¹ A lower RCC value indicates a better clustering partition. Averaged across all forecasts and number of clusters 2–9, RCC for EGUC clustering decreases by 1.58% between first and second order, then increases slightly for polynomial orders beyond 2 (Fig. 3a). For EGU clustering, RCC decreases by 2.1% between first and second order, and by 0.18% between second and third order, before increasing slightly for polynomial orders beyond 3. Taking both results into account, and seeking a single specification for EGUC and EGU clustering, we choose second-order polynomials for all clustering results presented in this study.

It is notable that RCC suggests a lower-order polynomial than the Bayesian information criterion (BIC)—a

¹For RCC calculations and throughout this paper, ensemble members are deterministically assigned to the cluster with their highest probability of assignment.

FIG. 1. Temporal evolution of the mean (a) total, (b) absolute along-track, and (c) absolute cross-track error for the ensemble-mean of each of the four EPSs (Table 2), along with the mean error for the multi-EPS ensembles EGU and EGUC.

measure based on log-likelihood that penalizes more complex solutions (e.g., higher-order polynomials, more clusters)—one of the metrics previously used by Kowaleski and Evans (2016, 2018). For EGUC and EGU clustering, BIC peaks for fifth-order solutions (Fig. 3b). We hypothesize that RCC recommends lower-order polynomials because it optimizes fit between

ensemble members only, whereas BIC optimizes fit between ensemble members and polynomial trajectories.

Neither RCC nor BIC unambiguously indicates an optimal number of clusters, as each metric decreases monotonically with additional clusters 3–9. Because this study evaluates the relationship between cluster size and cluster-mean error (section 3), we explore how the

TABLE 2. Mean total, along-track (bolded), and cross-track (italicized) error (reported to the nearest km) averaged over 0–144 and 96–144 h for each EPS and select multi-EPS combinations. Ensemble spread is computed as the average distance between ensemble members and the ensemble mean.

	ECMWF	GEFS	UKMET	CMC	EGU	EGUC
Mean 0–144-h error	162 116 <i>90</i>	194 126 <i>119</i>	183 129 <i>104</i>	257 177 <i>149</i>	149 106 <i>81</i>	152 108 <i>84</i>
Mean 96–144-h error	292 215 <i>152</i>	357 238 <i>215</i>	320 224 <i>185</i>	461 328 <i>255</i>	268 195 <i>143</i>	278 203 <i>148</i>
Mean 0–144-h ensemble spread	163 110 <i>96</i>	141 96 <i>84</i>	134 91 <i>80</i>	250 174 <i>144</i>	180 120 <i>109</i>	203 137 <i>121</i>
Mean 96–144-h ensemble spread	281 193 <i>164</i>	246 171 <i>141</i>	236 164 <i>136</i>	440 312 <i>245</i>	313 212 <i>186</i>	352 243 <i>205</i>

FIG. 2. Example of 144-h EGUC clustering using forecasts of Tropical Storm Soulik initialized at 0000 UTC 16 Aug 2018. The (a) 129 ensemble tracks are assigned to clusters with (b) central polynomial trajectories shown. Ensemble tracks are (c) color-coded by cluster with (d) cluster-mean tracks shown. The observed track of Soulik in (d) is shown in black. Markers in (d) indicate Soulik's position at 96, 120, and 144 h.

linear correlation between cluster fractional size (relative to ensemble size) and cluster-mean 96–144-h track error varies with number of clusters across all forecasts. Cluster-mean track error is defined as the displacement between the mean track of a cluster (Fig. 2d) and the observed track, averaged across the nine time steps from 96 to 144 h. For both EGUC and EGU clustering, five-cluster solutions produce the strongest relationship between cluster size and cluster-mean error across the full set of forecasts (Fig. 3c). Therefore, we primarily interpret results from five-cluster, second-order solutions.

3. Results

a. Cluster distribution by ensemble prediction system

Across all forecasts, the average number of clusters to which each EPS contributes varies substantially between

EPSs. For five-cluster EGUC partitions (hereafter EGUC5), the ECMWF ensemble (51 members), on average, contributes at least one member to the most clusters (4.64). Though the CMC ensemble is tied for the smallest population (21 members), and is the least accurate (Fig. 1), it contributes to the second-most clusters (4.35). This is likely due to its large ensemble spread, which greatly exceeds that of the other three EPSs (Table 2). The UKMET ensemble (36 members) contributes to an average of 3.97 clusters, while the GEFS ensemble (21 members) contributes to the fewest (3.60). Unsurprisingly, the UKMET and GEFS have the smallest average ensemble spreads (Table 2).

The widespread distribution of the ECMWF, especially compared to the GEFS, is further highlighted by evaluating how frequently each EPS contributes to each number of clusters (Fig. 4). In EGUC5 clustering, the ECMWF contributes members to all five clusters

FIG. 3. Across all forecasts, percentage decrease (improvement) in (a) relative cluster compactness and (b) Bayesian information criterion with each additional polynomial order, averaged for 2–9-cluster solutions of EGUC and EGU clustering; (c) correlation (R^2) between cluster fractional size and 96–144-h cluster-mean track error for second-order EGUC and EGU polynomial trajectories.

in 65% of forecasts, compared to 16% for the GEFS. The UKMET and CMC contribute to five clusters in 33% and 51% of forecasts, respectively (Fig. 4a). Calculating EPS contributions using a higher threshold, 10% of each EPS's population, further emphasizes these differences. The ECMWF ensemble contributes at least 10% of its members to four or more clusters in 42% of forecasts, compared to 12% for GEFS, 22% for UKMET, and 23% for CMC. The GEFS, in contrast, meets the 10% threshold for one or two clusters in 52% of forecasts; no other EPS exceeds 30% (Fig. 4b). Therefore, the GEFS most frequently fails to capture one or more potential scenarios (clusters) within the EGUC ensemble. This is due to both its small average spread, and its comparatively small population. Though the UKMET has a slightly smaller average spread than the GEFS, its larger population allows it to capture more scenarios, on average.

Similar results are observed for five-cluster EGU clustering (hereafter EGU5; Figs. 4c,d). On average, the GEFS contributes at least one member to 3.79 clusters, compared to 4.21 for UKMET and 4.81 for ECMWF. The GEFS provides at least one member to all five clusters in 24% of forecasts, much less often than UKMET (40%), and especially ECMWF (82%; Fig. 4c). The GEFS also contributes at least 10% of its population to only one or two clusters in 50% of forecasts,

substantially more than UKMET (19%), and ECMWF (3%; Fig. 4d).

The comparatively small average number of clusters that include GEFS members is consistent with previous studies that have found the GEFS to produce underdispersed TC track forecasts (Yamaguchi and Majumdar 2010; Hamill et al. 2011; Magnusson et al. 2014). This problem appears to have persisted through 2017–18. In 2020 the GEFS is expected to be revamped with an FV3 dynamical core, higher horizontal resolution, upgraded stochastic physics, and 10 additional ensemble members (Zhou et al. 2018). Further ensemble analysis will be necessary to determine whether these upgrades improve ensemble dispersion of TC track forecasts.

b. Relationship between cluster characteristics and cluster-mean 96–144-h track error

For EGUC5 clustering, a modest ($R^2 = 0.27$), but highly statistically significant linear correlation is observed between cluster fractional size relative to the total ensemble population and 96–144-h cluster-mean track error across all forecasts (Fig. 5a). Correlation is somewhat lower for EGU5 clustering ($R^2 = 0.18$; Fig. 5b), but still highly significant. For both ensembles, correlation improves markedly when correlating using

FIG. 4. Distribution of each EPS across track clusters generated from 0 to 144-h forecasts. For (a),(b) EGUC and (c),(d) EGU 5-cluster solutions: (left) fraction of forecasts in which each EPS contributes at least one member to each number of clusters 1–5; and (right) fraction of forecasts in which each EPS contributes at least 10% of its membership to each number of clusters 1–5.

a two-term exponential function ($R^2 = 0.43$ for EGUC and 0.30 for EGU), as the smallest clusters tend to generate very large cluster-mean errors.

For both EGUC5 and EGU5, cluster-mean errors decrease rapidly with increasing cluster size at small sizes, then less rapidly at larger sizes. For EGUC5, clusters with <10% of the total ensemble population (i.e., clusters having fractional size < 10%) have an average 96–144-h error of 1105 km, decreasing rapidly to 576 km for clusters with fractional size 10%–14.9%, and 456 km for 15%–19.9%. This monotonic decrease continues at larger cluster sizes; for example, clusters with fractional size $\geq 30\%$ have a mean error of 271 km (Fig. 5c). EGU5 clusters show a somewhat less pronounced decreasing error trend, with mean errors of 994 km for fractional size < 10%, 489 km for 10%–14.9%, 419 km for 15%–19.9%, and 288 km for clusters with fractional size $\geq 30\%$ (Fig. 5d).

The observed increase in cluster-mean accuracy with cluster size is consistent with the fact that larger ensembles will, on average, outperform smaller ensembles, given equal accuracy among ensemble members (e.g., Buizza and Palmer 1998; Goerss 2000). However, the very large cluster-mean errors of the smallest clusters

indicate that members of these clusters typically have particularly large errors. For example, the average 96–144-h cluster-mean error of EGUC5 clusters with fractional size < 10% (1105 km) is more than twice as large as the average error of individual ensemble members across the 75 forecasts that contain those clusters (460 km).

Evaluating the cluster-mean error ranks 1–5 (best–worst) by the cluster size ranks 1–5 (most populous–least populous) across all forecasts indicates that the least populous cluster in each forecast often produces the largest 96–144-h error (Fig. 6, rightmost bar groups). For EGUC5 clustering (Fig. 6a), 61% (89/145) of the least populous clusters (size rank 5) generate the largest error (error rank 5), and an additional 16% (23/145) generate a rank 4 error. Size rank 5 EGUC5 clusters generate the smallest error in only 3% (5/145) of cases. EGU5 clustering (Fig. 6b) produces a slightly weaker signal, but with the same trend: size rank 5 clusters generate the largest error in 56% (81/144) of cases, and a rank 4 error in an additional 15% (22/144), while generating the smallest error in only 8% (11/144) of cases.

As expected from the inverse relationship between cluster size and cluster-mean error, the most populous

FIG. 5. Scatterplot of cluster-mean 96–144-h track error with cluster fractional size for all forecasts using (a) EGUC5 clustering and (b) EGU5 clustering. Also plotted are linear (blue) and two-term exponential (red) best-fit lines. Mean total, along-track, and cross-track error with cluster fractional size are shown for (c) EGUC5 clustering, and (d) EGU5 clustering. Error bars in (c) and (d) denote total error standard deviation in each cluster size bin.

EGUC5 and EGU5 clusters are preferentially associated with smaller errors (Fig. 6, leftmost bar groups), though with a weaker relationship than that observed for the least populous clusters. In EGUC5 clustering, size rank 1 clusters produce the smallest error in 37% (58/157) of occurrences and a rank 2 error in an additional 31% (48/157), while generating a rank 4–5 error in only 12% (19/157) of occurrences. EGU5 clustering again yields a similar, but weaker signal, consistent with its smaller correlation between cluster size and cluster-mean error: EGU5 clusters of size rank 1 produce the smallest error in 29% (46/160) of cases, a rank 2 error in 31% (49/160), and a rank 4–5 error in 19% (30/160) of occurrences. Examining clusters by AT and XT errors yields qualitatively similar results (not shown).

c. Comparison between cluster-mean and ensemble-mean 96–144-h track errors

The relative accuracy of size rank 1 (most populous) clusters motivate investigation into how their track errors compare to the ensemble-mean track errors generated from various EPSs and multi-EPS ensembles. Results are summarized in Table 3. Across the 153 forecasts, the most populous EGUC5 and EGU5 clusters have nearly identical average 96–144-h cluster-mean

errors: 306 and 309 km, respectively.² These average errors are smaller than those of the GEFS, UKMET, and CMC ensembles means, but larger than the ECMWF mean and the means of the EGU and EGUC multi-EPS combinations (Table 2).

Across all forecasts, the mean of the most populous EGUC5 cluster outperforms the EGUC mean in 48% (74/153) of forecasts, and the most accurate EGU mean in 43% (66/153). Though the EGU ensemble mean is on average more accurate than the EGUC mean (Table 2), its most populous cluster bests the multi-EPS combinations slightly less frequently, outperforming the EGUC mean in 43% (66/153) of forecasts and the EGU mean in 42% (64/153; Table 3). Comparable results are found for AT and XT error components (Table 3). The performance of the most populous EGUC5 and EGU5 clusters may be contrasted with that of the remaining (i.e., not-most populous) clusters, which outperform the EGUC and EGU means substantially less frequently (Table 3).

² When two or more clusters in a forecast tie for the most populous, the average cluster-mean error of these clusters is used.

FIG. 6. For 96–144-h cluster-mean track errors: fraction of clusters of each size rank 1–5 (from most populous to least populous; abscissa) with each error rank 1–5 (from largest to smallest; color bars) using (a) EGUC5 clustering and (b) EGU5 clustering. The number of clusters of each size rank across all forecasts are in parentheses. Numbers vary due to ties in size rank. Within each size rank, the error rank fractions (color bars) sum to 1.

Across the full set of ensemble forecasts, the most populous EGUC5 and EGU5 clusters produce larger average errors than the EGUC and EGU ensemble means. However, the accuracy of the most populous EGUC5 cluster varies by its size, with particularly large

clusters showing greater accuracy than smaller clusters. When the most populous EGUC5 cluster comprises $\geq 30\%$ of the ensemble population (76 cases), it has a mean 96–144-h error of 262 km, which is 25% less than the 349 km mean error when it has $<30\%$ of all

TABLE 3. Comparison of the total, along-track (bolded) and cross-track (italicized) mean 96–144-h track errors of the most populous EGUC5 and EGU5 clusters relative to the EGUC and EGU multi-EPS means, along with ECMWF, GEFS, UKMET, and CMC ensemble means. Errors of the remaining (i.e., not-most populous) clusters are presented for comparison. Ensemble-mean errors from each EPS and ensemble combination are found in Table 2.

Cluster	Mean error 96–144-h	Better than EGUC mean (%)	Better than EGU mean (%)	Better than ECMWF mean (%)	Better than GEFS mean (%)	Better than UKMET mean (%)	Better than CMC mean (%)
Most populous cluster EGUC5	306	48.4	43.1	48.4	56.9	53.6	72.5
	215	45.8	43.8	49.7	51.0	56.2	66.7
	<i>170</i>	<i>46.4</i>	<i>43.1</i>	<i>46.4</i>	<i>62.7</i>	<i>55.6</i>	<i>71.2</i>
Most populous cluster EGU5	309	43.1	41.8	45.1	57.5	52.3	71.2
	212	49.0	43.1	51.0	48.4	56.9	67.3
	<i>177</i>	<i>42.5</i>	<i>37.9</i>	<i>45.1</i>	<i>57.5</i>	<i>52.9</i>	<i>73.2</i>
Remaining clusters EGUC5	542	21.1	20.4	21.7	31.3	30.1	42.6
	380	25.5	26.5	28.0	31.9	33.7	45.9
	<i>308</i>	<i>23.7</i>	<i>23.2</i>	<i>24.2</i>	<i>38.5</i>	<i>32.9</i>	<i>42.6</i>
Remaining clusters EGU5	467	26.6	24.6	26.8	37.9	34.7	52.9
	330	29.1	28.9	31.6	37.2	39.0	54.5
	<i>262</i>	<i>28.1</i>	<i>25.6</i>	<i>26.8</i>	<i>40.8</i>	<i>38.3</i>	<i>53.1</i>

TABLE 4. Comparison of mean 96–144-h track errors of the most populous cluster in EGUC5 and EGU5 clustering based on (i) population of the most populous cluster relative to the total ensemble, and (ii) population of the most populous cluster relative to the next-most populous cluster. Bolded values in columns 4–6 indicate that differences in cluster-mean error with cluster size (e.g., $\geq 30\%$ vs $< 30\%$) are statistically significant at 95% confidence. Columns 7–8 show the fraction of clusters in each subset that outperform the EGUC and EGU ensemble means in total, AT (bolded), and XT (italicized) errors.

Ensembles and No. of clusters	Subset of most populous clusters	No. of forecasts	Total error (km) mean and std dev	AT error (km) mean and std dev	XT error (km) mean and std dev	Better than EGUC mean (%)	Better than EGU mean (%)
EGUC5	$\geq 30\%$	76	262 (152)	182 (109)	152 (128)	53.9	47.4
						47.4	44.7
						55.3	50.0
EGUC5	$< 30\%$	77	349 (212)	248 (211)	188 (130)	42.9	39.0
						44.2	42.9
						37.7	36.4
EGUC5	$\geq 1.2x$	70	259 (158)	175 (132)	153 (117)	54.3	47.1
						54.3	47.1
						51.4	50.0
EGUC5	$< 1.2x$	83	345 (205)	249 (192)	185 (139)	43.4	39.8
						38.6	41.0
						42.2	37.3
EGU5	$\geq 30\%$	64	293 (177)	202 (141)	169 (135)	48.4	42.2
						51.6	48.4
						46.9	39.1
EGU5	$< 30\%$	89	320 (165)	219 (163)	182 (119)	39.3	34.8
						47.2	39.3
						39.3	37.1
EGU5	$\geq 1.2x$	70	305 (184)	205 (150)	181 (142)	44.3	38.6
						50.0	47.1
						44.3	37.1
EGU5	$< 1.2x$	83	311 (159)	217 (158)	174 (110)	42.2	37.3
						48.2	39.8
						41.0	38.6

ensemble members (Table 4). A similar difference is obtained using a threshold of 20% more members than the next-most populous cluster ($1.2x$). Both differences are significant at 95% confidence (two-tailed Student's t test; used for all significance tests below). In contrast to EGUC5, the accuracy of the most populous EGU5 cluster varies less with its size. The 8.5% difference in track error between the most populous EGU5 clusters with $\geq 30\%$ and $< 30\%$ of the ensemble population does not reach statistical significance (Table 4).

Despite the improved performance of most-populous EGUC5 clusters that comprise $\geq 30\%$ of the ensemble population, they still produce larger average errors (262 km) than the EGU mean (230 km) and EGUC5 mean (245 km) for that subset of 76 forecasts. Therefore, our results indicate that the mean of the most populous cluster should not replace the ensemble mean as a deterministic forecast. However, clustering does provide information about which clusters are more likely to have comparatively small errors. Cluster analysis, if applied operationally, could be used to distill an ensemble into a few scenarios, providing objective information about which clusters in the ensemble are more likely to verify.

A skilled forecaster could combine this objective information with knowledge of the synoptic scenario and ranges of TC characteristics associated with each cluster to produce a superior deterministic or scenario-based forecast.

In addition to the enhanced accuracy of EGUC5 clusters that comprise $\geq 30\%$ of the total ensemble population, forecasts containing these clusters also tend to have smaller *ensemble-mean* track errors. Forecasts in which the largest EGUC5 cluster comprises $\geq 30\%$ of the ensemble population have EGUC-mean and EGU-mean errors that are 21% and 25% smaller, respectively, than forecasts in which the largest EGUC5 cluster is smaller (Fig. 7). The ensemble-mean errors of the constituent ECMWF and UKMET ensembles also vary substantially with the size of the largest EGUC5 cluster; differences in the EGUC-, EGU-, ECMWF-, and UKMET-mean errors all reach 95% statistical significance. These differences appear to be driven by the lower frequency of large ensemble-mean errors in forecasts with a dominant EGUC5 cluster. For example, the 96–144-h EGUC-mean error exceeds 500 km for 1/76 forecasts in which the largest EGUC5 cluster

FIG. 7. Mean and standard deviation of (a) total, (b) along-track, and (c) cross-track 96–144-h multi-EPS-mean and ensemble-mean errors for forecasts in which the most populous EGUC5 cluster contains $\geq 30\%$ of the total ensemble population and forecasts in which it contains $< 30\%$. Asterisks indicate statistically significant differences at 95% confidence.

comprises $\geq 30\%$, compared to 12/77 forecasts in which it comprises $< 30\%$. However, different size cutoffs (e.g., 27.5% and 32.5%) yield less clear relationships between cluster size and ensemble-mean error. Therefore, a larger dataset is needed to more fully evaluate whether larger EGUC5 clusters are associated with smaller ensemble-mean errors.

Notably, the enhanced accuracy of the EGUC-mean in forecasts with a particularly large EGUC5 cluster does not appear related to smaller EGUC ensemble spread. The mean 96–144-h distance between each EGUC member and the ensemble-mean track is 350 km for forecasts in which the most populous cluster is $\geq 30\%$ of the ensemble population, nearly identical to the 355 km when it is $< 30\%$. Similar results are obtained when comparing mean 0–144-h ensemble spread. Therefore, while ensemble spread is inversely related to ensemble-mean accuracy (Yamaguchi et al. 2009; Majumdar and Finocchio 2010), the size of the most populous EGUC5 cluster may be an additional useful predictor for assessing confidence in ensemble-mean forecasts.

d. Relationship between cluster composition and forecast error

In addition to cluster size, EGUC5 and EGU5 cluster-mean track errors vary substantially by which EPSs contribute to the cluster. For EGUC5 and EGU5, we evaluate ensemble-mean error across all forecasts by

grouping clusters based on which EPSs contribute a threshold fraction of each cluster's population. We select a threshold that varies linearly with EPS population and corresponds to 7.5% of a cluster's population from GEFS and CMC (21 members each), 12.9% from UKMET (36 members), and 18.2% from ECMWF (51 members).³ Applying these thresholds to EGUC5 shows that clusters containing threshold membership from each of ECMWF, GEFS, and UKMET (hereafter E + G + U clusters; 197 cases) have a significantly smaller mean error (340 km; Fig. 8a) than that of all other clusters (non-E + G + U; 547 km; 568 cases). E + G + U clusters are the most accurate cluster (error rank 1) 36% of the times that they occur, compared to 14% for non-E + G + U clusters (Fig. 8b).

For EGUC5 clustering, the subset of E + G + U clusters that contain threshold memberships of ECMWF, GEFS, and UKMET, but *not* CMC (hereafter E + G + U – C; 97 cases) are even better performing, with a mean error of 287 km and an error rank 1 in 44% of occurrences. The E + G + U – C clusters also outperform the E + G + U clusters generated in EGU5 clustering (349 km; 229 cases; Fig. 8a), which yield an error

³ Thresholds that correspond to 5% and 10% of cluster population from the GEFS and CMC, and proportionally higher from UKMET and ECMWF, yield qualitatively similar results (not shown).

FIG. 8. For EGUC5 and EGU5 clustering: (a) mean and standard deviation of 96–144-h cluster-mean error by the EPSs that contribute to each cluster at or above a variable threshold, and (b) fraction of clusters with error ranks 1 and 2 by the EPSs that contribute at or above that threshold. Error ranks 1 and 2 indicate the smallest and second-smallest cluster-mean error, respectively, for each forecast. The EPS threshold varies with ensemble population, corresponding to 7.5% of cluster population from GEFS and CMC, 12.9% from UKMET, and 18.2% from ECMWF. For EGUC5 clustering, $E + G + U - C$ is a subset of $E + G + U$.

rank 1 in 30% of occurrences (Fig. 8b). These results illustrate the potential advantage of including the CMC in clustering, though it has by far the largest average ensemble-mean 96–144-h error (Table 2). Inclusion of the CMC in EGUC5 clustering highlights the superior accuracy of $E + G + U - C$ clusters, which either contain no CMC members or contain CMC membership below the 7.5% threshold. These clusters are not highlighted in EGU5 clustering.

While $E + G + U - C$ clusters tend to be more populous, on average, than other clusters, their superior performance is not simply due to their larger size. For example, within the subset of clusters with $\geq 20\%$ fractional size, $E + G + U - C$ clusters are, on average, only slightly larger than all other clusters (26.8% versus 25.9%). However, among this subset, they have substantially smaller average 96–144-h errors (261 km versus 378 km). If a cluster has substantial contributions from the three more accurate EPSs, but not the CMC, it has a relatively high probability of producing an accurate forecast. Thus, cluster composition provides an additional metric beyond cluster size for operational forecasters to assess the likelihood of track scenarios within a multi-EPS ensemble.

Although cluster composition provides information about cluster-mean accuracy independently of cluster size, $E + G + U - C$ clusters are especially accurate when they are the most populous cluster in a forecast. Of the 149 forecasts for which a single most populous

EGUC5 cluster exists, it is an $E + G + U - C$ cluster in 34 cases. Averaged across these 34 forecasts, the $E + G + U - C$ cluster-mean 96–144-h error (230 km) is similar to the EGU-mean (230 km) and the EGUC-mean (238 km) errors, and modestly smaller than the ECMWF-mean error (267 km). The $E + G + U - C$ cluster mean outperforms the ECMWF mean and EGU mean in 18/34 forecasts, and bests the EGUC mean in 21/34 forecasts. Thus, if the most populous EGUC5 cluster contains threshold memberships from the ECMWF, GEFS, and UKMET, but *not* the CMC, it can be expected to produce a forecast with similar errors to the most accurate multi-EPS means, and more accurate than the ECMWF and GEFS ensemble means currently used in consensus track models. If EGUC5 clustering were to be employed in operational forecasting, a large $E + G + U - C$ cluster would indicate a potentially realistic track scenario that should be further investigated.

Across the 153 forecasts examined here, cluster size and composition are meaningful predictors of cluster-mean track forecast error in the 96–144-h period, near and beyond the limit of operational tropical cyclone forecasting. Although the most populous EGUC5 and EGU5 cluster means are less accurate, on average, than the ECMWF mean and the more accurate multi-EPS means, the most populous EGUC5 cluster from each forecast has elevated accuracy when it either (i) is especially populous, and/or (ii) contains substantial

TABLE 5. Changes in ensemble-mean 96–144-h track error from EGUC5 and EGU5 pruning using different pruning thresholds. A pruning threshold means that all ensembles that belong to clusters smaller than the threshold are discarded. Changes in error distribution from bolded pruning thresholds are shown in Figs. 10 and 11.

Ensembles and No. of clusters	Pruning threshold	Forecasts affected (% of all forecasts)	Forecasts improved (% of affected forecasts)	Change in mean error of all affected forecasts (km and %)	Change in mean error of improved forecasts (km and %)	Change in mean error of degraded forecasts (km and %)
EGUC5	12.5%	116 (75.8%)	78 (67.2%)	-17.0 (-6.2%)	-47.4 (-17.0%)	45.5 (17.6%)
EGUC5	10%	75 (49.0%)	52 (69.3%)	-20.0 (-7.0%)	-45.1 (-15.0%)	36.9 (15.0%)
EGUC5	40% of most populous cluster	100 (65.4%)	65 (65.0%)	-17.2 (-6.3%)	-48.0 (-17.0%)	39.9 (15.6%)
EGU5	12.5%	101 (66.0%)	59 (58.4%)	-7.9 (-3.0%)	-37.4 (-13.8%)	33.7 (13.0%)
EGU5	10%	53 (34.6%)	39 (73.6%)	-20.7 (-7.3%)	-37.6 (-12.9%)	26.4 (10.4%)
EGU5	25% of most populous cluster	34 (22.2%)	26 (76.5%)	-25.1 (-8.9%)	-38.0 (-12.7%)	16.9 (7.3%)

membership from the ECMWF, GEFS, and UKMET, but *not* the CMC. Cluster analysis, if used to inform operational TC forecasting, could help forecasters to evaluate which scenarios within an ensemble are more likely to verify.

4. Cluster pruning to reduce ensemble-mean errors

Perhaps the clearest result from this study is that the smallest EGUC5 and EGU5 clusters tend to produce very large cluster-mean errors. This finding leads us to investigate whether pruning these clusters from the ensemble decreases ensemble-mean forecast error. In pruning, all ensemble members that belong to each cluster smaller than a threshold. (e.g., 10% of the ensemble population) are discarded. Then, the ensemble-mean track is recomputed using the remaining ensemble members. Ensemble-mean forecast errors before and after pruning are compared to evaluate whether pruning improves or degrades the forecast. If small clusters are routinely poorly performing outliers, pruning may be expected to reduce the ensemble-mean track error. Some key results are highlighted here, with additional results provided in Table 5.

Examples of EGU5 pruning with a pruning threshold of 10% are shown in Fig. 9. In each forecast shown, the pruned cluster comprises a small set of ensemble members near the edge of the ensemble spread. For Soulik (Fig. 9a), Jose (Fig. 9b), and Yutu (Fig. 9c), pruning improves the 96–144-h ensemble-mean forecast by discarding poorly performing members. However, the observed track of Noru (Fig. 9d) falls nearer to the pruned cluster; discarding these ensemble members increases 96–144-h ensemble-mean track error.

For EGUC5 clustering, multiple pruning thresholds decrease mean error across the subset of forecasts in which one or more clusters is pruned (Table 5). Pruning all clusters that comprise <12.5% of the ensemble population (Figs. 10a and 11a) affects 116/153 forecasts, improving the ensemble mean of 67% and decreasing average 96–144-h error across all affected forecasts by 17 km (6.2% of the mean unpruned error). Using a 10% pruning threshold (Figs. 10b and 11b) affects fewer forecasts (75), but produces a slightly larger average error reduction (20 km; 7.0%).

Pruning the more accurate EGU ensemble by cluster size also reduces ensemble-mean error, though it affects fewer forecasts than EGUC pruning (Table 5). For EGU5, a 10% pruning threshold (Figs. 10c and 11c) affects 53/153 forecasts, improving 74% and reducing average error across the 53 forecasts by 21 km (7.3%). Pruning EGU5 clusters by size relative to the most populous cluster also shows promising results; a 25%

FIG. 9. Examples of ensemble pruning using EGU5 clustering and a pruning threshold of 10% of the total ensemble population. In each case, ensemble tracks are shown in gray, with pruned ensemble tracks in red. The unpruned and pruned ensemble means are blue and magenta, respectively. The observed track is black. Circles, squares, and triangles indicate TC position at 96, 120, and 144 h, respectively. Forecasts shown are (a) Soulik at 0000 UTC 16 Aug 2018, (b) Jose at 0000 UTC 13 Sep 2017, (c) Yutu at 0000 UTC 25 Oct 2018, and (d) Noru at 0000 UTC 30 Jul 2017.

pruning threshold relative to the most populous cluster (Figs. 10d and 11d) improves 26 of 34 forecasts affected, with an average error reduction of 25 km (8.9%) across the 34 forecasts. Notably, for every five-cluster pruning threshold shown in Table 5, the average error reduction for improved forecasts is similar to or greater than the average error increase for degraded forecasts. Thus, across all forecasts, the numerous small error reductions are not offset by catastrophic error increases in a small number of forecasts.

Pruning using a variety of thresholds decreases ensemble-mean error across the affected forecasts; however, ideally cluster and/or ensemble characteristics could be used to evaluate a priori whether pruning will likely decrease or increase the ensemble-mean error. For 10% and 12.5% EGUC5 and EGU5 pruning thresholds, the mean 96–144-h distance between the pruned ensemble members and the ensemble mean is computed for each forecast. Then, the distances are compared for improved and degraded forecasts.

No strong signal is observed for EGUC5. However, for EGU5, pruning is more likely to produce improvement when it removes ensemble members that are especially far from the ensemble mean (Table 6). For a 10% pruning threshold, the mean 96–144-h distance between pruned members and the unpruned ensemble mean in degraded forecasts (675 km) is 30% smaller than the mean distance in improved forecasts (960 km). A 25% difference (552 km versus 740 km) is found using a 12.5% pruning threshold. Both differences are significant at 95% confidence.

Much of the difference between improved and degraded EGU5 pruning cases is due to differences in the along-track displacement between pruned ensemble members and the unpruned ensemble mean (Table 6). For both 10% and 12.5% thresholds, the average difference in along-track displacement between improvement and degradation cases (36% and 33%, respectively; Table 6) is somewhat larger than the average difference in cross-track displacement (21% and

FIG. 10. Histograms and kernel density plots of 96–144-h ensemble-mean errors before and after pruning, where the pruning threshold is characterized by cluster size: (a) EGUC5 12.5% of the ensemble population, (b) EGUC5 10% of the ensemble population, (c) EGU5 10% of the ensemble population, and (d) EGU5 25% of the population of the most populous cluster. Gray and red bars indicate error bin frequencies pre- and postpruning, respectively. Percentages refer to the percentages of forecasts in which at least one cluster is pruned. Only forecasts affected by pruning are included in each plot.

17%, respectively). Therefore, if a small EGU5 cluster is particularly far from the ensemble mean during the 96–144-h period, especially if it has a large along-track displacement, pruning it will likely improve the EGU-mean forecast. In an operational forecasting context, automated pruning of the EGU mean could be performed prior to its inclusion in deterministic consensus track guidance.

5. Summary and conclusions

Tropical cyclone forecasts with lead times of 96–144-h are near the current edge of operational forecasts. For this reason, we investigate whether ensemble clustering can inform track forecasts at these lead times. Regression mixture-model track clustering is applied to 153 tropical cyclone ensemble forecasts from 2017 and 2018. Clustering is performed using a combined four-EPS dataset [ECMWF + GEFS + UKMET + CMC (EGUC), 129 members], and a three-EPS dataset [ECMWF + GEFS + UKMET (EGU), 108 members]

that excludes the least accurate CMC. Five-cluster partitions generated from second-order polynomials are selected to analyze for EGUC and EGU clustering (EGUC5 and EGU5).

For both multi-EPS ensembles, the smallest clusters tend to produce very large 96–144-h cluster-mean errors. For EGUC5 and EGU5 clustering, clusters comprising <10% of the total ensemble population generate mean 96–144-h errors of 1105 and 994 km, respectively. Average cluster-mean error decreases with increasing cluster size, though this decrease becomes more gradual at larger sizes. While the most populous EGUC5 and EGU5 clusters in each forecast are, on average, less accurate than the EGU and EGUC ensemble means, the most populous EGUC5 clusters show additional accuracy when they contain $\geq 30\%$ of the ensemble population. Forecasts in which the most populous EGUC5 cluster contains $\geq 30\%$ of the ensemble population also appear to have smaller EGUC, EGU, ECMWF, and UKMET ensemble-mean errors, though a larger sample size is needed to confirm this.

FIG. 11. Histograms and kernel density plots of changes in ensemble-mean 96–144-h error from each pruning threshold across the set of forecasts affected by pruning. Negative values indicate forecast improvements.

Beyond cluster size, cluster-mean error also varies substantially with the EPS members that a cluster contains. For both EGUC5 and EGU5 clustering, clusters that contain substantial contributions from each of the ECMWF, GEFS, and UKMET significantly outperform clusters that do not contain contributions from these three EPSs. Furthermore, clusters that contain substantial contributions from each of ECMWF, GEFS, and UKMET, but *not* CMC, are even more accurate. When these clusters are the most populous in an ensemble, their cluster-mean errors are competitive with the most accurate EGU ensemble mean. These results suggest that including the CMC in clustering may be beneficial, despite its large average ensemble-mean error. Across

the full set of 153 forecasts, including the CMC in clustering highlights the elevated accuracy of clusters that have few or no CMC members, but have substantial contributions from each of the other three EPSs.

The large 96–144-h errors produced by small clusters motivates investigation into whether removing the members of these small clusters (“pruning”) improves ensemble-mean forecasts. For EGUC5, pruning clusters with fewer than 12.5% or 10% of the ensemble population affects 76% and 49% of forecasts, respectively, improving slightly more than 2/3 of the forecasts affected in each case. Pruning EGU clusters with fewer than 10% of the ensemble population affects fewer forecasts (35%), but improves 74% of affected forecasts.

TABLE 6. Average of the total, AT, and XT mean displacement from 96 to 144 h between pruned ensemble members and the ensemble mean for pruning cases that improve and degrade the ensemble-mean forecast. Bolded values indicate differences that are statistically significant at 95% confidence.

Ensembles and No. of clusters	Pruning threshold	No. of improved forecasts	No. of degraded forecasts	Total displacement (km) of improved (degraded) forecasts	AT displacement (km) of improved (degraded) forecasts	XT displacement (km) of improved (degraded) forecasts
EGUC5	12.5%	78	38	822 (781)	575 (524)	480 (467)
EGUC5	10%	52	23	1043 (935)	739 (644)	611 (539)
EGU5	12.5%	59	42	740 (552)	511 (341)	440 (364)
EGU5	10%	39	14	960 (675)	661 (423)	569 (449)

Pruning appears particularly likely to improve the EGU mean when it removes ensemble members that are especially distant from the mean.

Ensemble TC track forecasts have become increasingly used in operational consensus track guidance. Results from this paper show how clustering can be used to identify clusters, based on size and EPS composition, that are more likely to have relatively small errors at 4–6-day lead times, near and beyond the edge of operational TC forecasts. Cluster pruning, as demonstrated here, can improve medium-range ensemble-mean forecasts by excluding members that belong to small clusters, which we have shown typically have larger errors. If a multi-EPS ensemble were to be included in a consensus track model, pruning could easily be applied objectively prior to its inclusion. Because cluster pruning is effective at reducing 96–144-h ensemble-mean errors, it complements other ensemble selection methods (Qi et al. 2014; Dong and Zhang 2016; Zhang and Yu 2017) that primarily produce forecast improvements at shorter lead times.

On the whole, the improvements that clustering yields for deterministic forecasts are fairly modest. Individual clusters, even populous ones, are only competitive with the most accurate EGU-mean forecast in specific circumstances. While pruning typically improves the ensemble mean, only a portion of forecasts are affected. Furthermore, global EPSs are frequently upgraded, and the results presented here may be sensitive to the specifications of individual EPSs. Were the methods here to be employed operationally, frequently updated statistical analysis would be required to evaluate how clustering statistics change as the component EPSs undergo upgrades.

Another important caveat to the results in this paper is that five clusters and second-order polynomial trajectories were used for every forecast. While this was done for consistency across the dataset, optimal cluster characteristics vary between storms and even forecast-to-forecast. If cluster analysis were to use the optimal cluster characteristics for each forecast, clustering might produce more useful forecast guidance.

Beyond the deterministic-focused results presented in this paper, ensemble forecasts yield valuable probabilistic information about TC track (Titley et al. 2020), though this information remains underutilized in operational TC forecasts (Titley et al. 2019). Therefore, while clustering can guide deterministic forecasts in certain circumstances, it may have greater use in bridging the deterministic-probabilistic gap by distilling a large ensemble into a small set of scenarios and helping forecasters to determine which scenario(s) are most likely to verify. These scenarios could then be connected

to synoptic patterns, TC evolution (Kowaleski and Evans 2018), and TC-related hazards (Kowaleski et al. 2020). Additionally, further research is necessary to determine whether the pruning methods demonstrated in this paper can also improve probabilistic TC track guidance.

Acknowledgments. We thank Judith Berner, whose ideas motivated some of the research in this paper. We are also grateful to Julian Heming at the Met Office for providing deterministic UKMET track forecasts. We acknowledge high-performance computing support from the Penn State Institute for Computational and Data Sciences (ICDS) Advanced CyberInfrastructure (ICDS-ACI).

Data accessibility statement: Ensemble track forecasts used in this study may be accessed via the UCAR TIGGE database: <https://rda.ucar.edu/datasets/ds330.3>. The curve clustering toolbox used in this study can be downloaded via the CC Toolbox website at <http://www.datalab.uci.edu/resources/CCToolbox/#Download>. Best track data can be obtained from HURDAT 2 (https://www.aoml.noaa.gov/hrd/hurdat/Data_Storm.html), the Japan Meteorological Agency (<https://www.jma.go.jp/jma/jma-eng/jma-center/rsmc-hp-pub-eg/trackarchives.html>), and from IBTrACS via the University of North Carolina Asheville (<http://ibtracs.unca.edu/>). Some track data used in this study were provisional and have since been updated online; the provisional data will be provided upon request.

REFERENCES

- Berner, J., S.-Y. Ha, J. P. Hacker, A. Fournier, and C. Snyder, 2011: Model uncertainty in a mesoscale ensemble prediction system: Stochastic versus multiphysics representations. *Mon. Wea. Rev.*, **139**, 1972–1995, <https://doi.org/10.1175/2010MWR3595.1>.
- , and Coauthors, 2017: Stochastic parameterization: Toward a new view of weather and climate models. *Bull. Amer. Meteor. Soc.*, **98**, 565–588, <https://doi.org/10.1175/BAMS-D-15-00268.1>.
- Buizza, R., and T. N. Palmer, 1998: Impact of ensemble size on ensemble prediction. *Mon. Wea. Rev.*, **126**, 2503–2518, [https://doi.org/10.1175/1520-0493\(1998\)126<2503:IOESOE>2.0.CO;2](https://doi.org/10.1175/1520-0493(1998)126<2503:IOESOE>2.0.CO;2).
- Camargo, S. J., A. W. Robertson, S. J. Gaffney, P. Smythe, and M. Ghil, 2007: Cluster analysis of typhoon tracks. Part I: General properties. *J. Climate*, **20**, 3635–3653, <https://doi.org/10.1175/JCLI4188.1>.
- Cangialosi, J. P., 2019: 2018 hurricane season. National Hurricane Center Forecast Verification Rep., 73 pp., https://www.nhc.noaa.gov/verification/pdfs/Verification_2018.pdf.
- , and J. Franklin, 2017: 2016 hurricane season. National Hurricane Center Forecast Verification Rep., 72 pp., https://www.nhc.noaa.gov/verification/pdfs/Verification_2016.pdf.
- Dong, L., and F. Zhang, 2016: OBEST: An observation-based ensemble subsetting technique for tropical cyclone track prediction. *Wea. Forecasting*, **31**, 57–70, <https://doi.org/10.1175/WAF-D-15-0056.1>.

- Dupont, T., M. Plu, P. Caroff, and G. Faure, 2011: Verification of ensemble-based uncertainty circles around tropical cyclone track forecasts. *Wea. Forecasting*, **26**, 664–676, <https://doi.org/10.1175/WAF-D-11-00007.1>.
- Fossell, K. R., D. Ahijevych, R. E. Morss, C. Snyder, and C. Davis, 2017: The practical predictability of storm tide from tropical cyclones in the Gulf of Mexico. *Mon. Wea. Rev.*, **145**, 5103–5121, <https://doi.org/10.1175/MWR-D-17-0051.1>.
- Gaffney, S. J., A. W. Robertson, P. Smith, S. J. Camargo, and M. Ghil, 2007: Probabilistic clustering of extratropical cyclones using regression mixture models. *Climate Dyn.*, **29**, 423–440, <https://doi.org/10.1007/s00382-007-0235-z>.
- Goerss, J. S., 2000: Tropical cyclone track forecasts using an ensemble of dynamical models. *Mon. Wea. Rev.*, **128**, 1187–1193, [https://doi.org/10.1175/1520-0493\(2000\)128<1187:TCTFUA>2.0.CO;2](https://doi.org/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, [https://doi.org/10.1175/1520-0434\(2004\)019<0633:AHOWNP>2.0.CO;2](https://doi.org/10.1175/1520-0434(2004)019<0633:AHOWNP>2.0.CO;2).
- Hamill, T. M., J. S. Whitaker, M. Fiorino, and S. G. Benjamin, 2011: Global ensemble predictions of 2009's tropical cyclone initialized with an ensemble Kalman filter. *Mon. Wea. Rev.*, **139**, 668–688, <https://doi.org/10.1175/2010MWR3456.1>.
- HRD, 2020: Hurricane database. Accessed 23 March 2020, https://www.aoml.noaa.gov/hrd/hurdat/Data_Storm.html.
- Jankov, I., and Coauthors, 2017: A performance comparison between multiphysics and stochastic approaches within a North American RAP ensemble. *Mon. Wea. Rev.*, **145**, 1161–1179, <https://doi.org/10.1175/MWR-D-16-0160.1>.
- JMA, 2020: Best track data. Accessed 23 March 2020, <https://www.jma.go.jp/jma/jma-eng/jma-center/rsmc-hp-pub-eg/trackarchives.html>.
- Kowaleski, A. M., and J. L. Evans, 2016: Regression mixture model clustering of multimodel ensemble forecasts of Hurricane Sandy: Partition characteristics. *Mon. Wea. Rev.*, **144**, 3825–3846, <https://doi.org/10.1175/MWR-D-16-0099.1>.
- , and —, 2018: Relationship between the track and structural evolution of Hurricane Sandy (2012) using a regional ensemble. *Mon. Wea. Rev.*, **146**, 4279–4302, <https://doi.org/10.1175/MWR-D-18-0121.1>.
- , R. E. Morss, D. Ahijevych, and K. R. Fossell, 2020: Using a WRF-ADCIRC ensemble and track clustering to investigate storm surge hazards and inundation scenarios associated with Hurricane Irma. *Wea. Forecasting*, **35**, 1289–1315, <https://doi.org/10.1175/WAF-D-19-0169.1>.
- Krishnamurti, T. N., M. K. Biswas, B. P. Mackey, R. G. Ellingson, and P. H. Ruscher, 2011: Hurricane forecasts using a suite of large-scale models. *Tellus*, **63**, 727–745, <https://doi.org/10.1111/j.1600-0870.2011.00519.x>.
- Kurupmullage Don, P., J. L. Evans, F. Chiaromonte, and A. M. Kowaleski, 2016: Mixture-based path clustering for synthesis of ECMWF ensemble forecasts of tropical cyclone evolution. *Mon. Wea. Rev.*, **144**, 3301–3320, <https://doi.org/10.1175/MWR-D-15-0214.1>.
- Landsea, C. W., and J. P. Cangialosi, 2018: Have we reached the limits of predictability for tropical cyclone track forecasting? *Bull. Amer. Meteor. Soc.*, **99**, 2237–2243, <https://doi.org/10.1175/BAMS-D-17-0136.1>.
- Leonardo, N. M., and B. A. Colle, 2017: Verification of multimodel ensemble forecasts of North Atlantic tropical cyclones. *Wea. Forecasting*, **32**, 2083–2101, <https://doi.org/10.1175/WAF-D-17-0058.1>.
- Magnusson, L., J.-R. Bidlon, S. T. K. Kang, A. Thorpe, N. Wedi, and M. Yamaguchi, 2014: Evaluation of medium-range forecasts for Hurricane Sandy. *Mon. Wea. Rev.*, **142**, 1962–1981, <https://doi.org/10.1175/MWR-D-13-00228.1>.
- 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, <https://doi.org/10.1175/2009WAF2222327.1>.
- Melhauser, C., F. Zhang, Y. Weng, Y. Jin, H. Jin, and Q. Zhao, 2017: A multiple-model convection-permitting ensemble examination of the probabilistic prediction of tropical cyclones: Hurricanes Sandy (2012) and Edouard (2014). *Wea. Forecasting*, **32**, 665–688, <https://doi.org/10.1175/WAF-D-16-0082.1>.
- NCEI, 2020: IBTrACS. NOAA/NCDC, accessed 23 March 2020, <https://www.ncdc.noaa.gov/ibtracs/>.
- NHC, 2019: National Hurricane Center forecast verification. NOAA/NHC, accessed 1 April 2019, <https://www.nhc.noaa.gov/verification/>.
- NOAA/NWS/NCEP, 2020: THORPEX Grand Global Ensemble (TIGGE) model tropical cyclone track data. National Center for Atmospheric Research, Computational and Information Systems Laboratory, accessed 3 January 2020, <https://doi.org/10.5065/D6GH9GSZ>.
- Nystrom, R. G., F. Zhang, E. B. Munsell, S. A. Braun, J. A. Sippel, Y. Weng, and K. Emanuel, 2018: Predictability and dynamics of Hurricane Joaquin (2015) explored through convection-permitting ensemble sensitivity experiments. *J. Atmos. Sci.*, **75**, 401–424, <https://doi.org/10.1175/JAS-D-17-0137.1>.
- 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**, 805–815, <https://doi.org/10.1002/QJ.2196>.
- Rappaport, E. N., and Coauthors, 2009: Advances and challenges at the National Hurricane Center. *Wea. Forecasting*, **24**, 395–419, <https://doi.org/10.1175/2008WAF2222128.1>.
- Simon, A., A. B. Penny, M. DeMaria, J. L. Franklin, R. J. Pasch, E. N. Rappaport, and D. A. Zelinsky, 2018: A description of the real-time HFIP Corrected Consensus Approach (HCCA) for tropical cyclone track and intensity guidance. *Wea. Forecasting*, **33**, 37–57, <https://doi.org/10.1175/WAF-D-17-0068.1>.
- Titley, H. A., M. Yamaguchi, and L. Magnusson, 2019: Current and potential use of ensemble forecasts in operational TC forecasting: Results from a global forecaster survey. *Trop. Cyclone Res. Rev.*, **8**, 166–180, <https://doi.org/10.1016/j.tcr.2019.10.005>.
- , R. L. Bowyer, and H. L. Cloke, 2020: A global evaluation of multi-model ensemble tropical cyclone track probability forecasts. *Quart. J. Roy. Meteor. Soc.*, **146**, 531–545, <https://doi.org/10.1002/qj.3712>.
- UNCA, 2020: IBTrACS. Accessed 23 March 2020, <http://ibtracs.unca.edu/>.
- Vitart, F., F. Prates, A. Bonet, and C. Sahin, 2012: New tropical cyclone products on the web. *ECMWF Newsletter*, No. 130, ECMWF, Reading, United Kingdom, 17–23, <https://www.ecmwf.int/en/elibrary/14592-newsletter-no-130-winter-2011-12>.
- Yamaguchi, M., and S. Majumdar, 2010: Using TIGGE data to diagnose initial perturbations and their growth for tropical cyclone ensemble forecasts. *Mon. Wea. Rev.*, **138**, 3634–3655, <https://doi.org/10.1175/2010MWR3176.1>.
- , R. Sakai, M. Kayoda, T. Komori, and T. Kadowaki, 2009: Typhoon ensemble prediction system developed at the Japan

- Meteorological Agency. *Mon. Wea. Rev.*, **137**, 2592–2604, <https://doi.org/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, <https://doi.org/10.1002/qj.1937>.
- Zhang, X., and H. Yu, 2017: A probabilistic tropical cyclone track forecast scheme based on the selective consensus of ensemble prediction systems. *Wea. Forecasting*, **32**, 2143–2157, <https://doi.org/10.1175/WAF-D-17-0071.1>.
- Zhou, Z., B. Fu, D. Hou, W. Li, J. Peng, Y. Luo, and E. Sinksy, 2018: The development of next NCEP GEFS. *43rd NOAA Annual CDPW*, Santa Barbara, CA, NWS STI Climate Bulletin, 22 pp., <https://www.cpc.ncep.noaa.gov/products/outreach/CDPW/43/oral-sessions/presentations/thurs/CDPW-2018-Zhou.pdf>.