Jumpiness in ensemble forecasts of Atlantic tropical cyclone tracks

Richardson, D. S., Cloke, H. L. ORCID: https://orcid.org/0000-0002-1472-868X, Methven, J. A. ORCID: https://orcid.org/0000-0002-7636-6872 and Pappenberger, F. (2024) Jumpiness in ensemble forecasts of Atlantic tropical cyclone tracks. Weather and Forecasting, 39 (1). pp. 203-215. ISSN 1520-0434 doi: https://doi.org/10.1175/WAF-D-23-0113.1 Available at https://centaur.reading.ac.uk/114359/

Jumpiness in Ensemble Forecasts of Atlantic Tropical Cyclone Tracks

DAVID S. RICHARDSON,[a,b] HANNAH L. CLOKE,[a,c,d] JOHN A. METHVEN,[c] AND FLORIAN PAPPENBERGER[b]

[a] *Department of Geography and Environmental Science, University of Reading, Reading, United Kingdom*
[b] *ECMWF, Reading, United Kingdom*
[c] *Department of Meteorology, University of Reading, Reading, United Kingdom*
[d] *Department of Earth Sciences, Uppsala University, Uppsala, Sweden*

(Manuscript received 4 July 2023, in final form 25 October 2023, accepted 25 November 2023)

ABSTRACT: We investigate the run-to-run consistency (jumpiness) of ensemble forecasts of tropical cyclone tracks from three global centers: ECMWF, the Met Office, and NCEP. We use a divergence function to quantify the change in cross-track position between consecutive ensemble forecasts initialized at 12-h intervals. Results for the 2019–21 North Atlantic hurricane season show that the jumpiness varied substantially between cases and centers, with no common cause across the different ensemble systems. Recent upgrades to the Met Office and NCEP ensembles reduced their overall jumpiness to match that of the ECMWF ensemble. The average divergence over the set of cases provides an objective measure of the expected change in cross-track position from one forecast to the next. For example, a user should expect on average that the ensemble mean position will change by around 80–90 km in the cross-track direction between a forecast for 120 h ahead and the updated forecast made 12 h later for the same valid time. This quantitative information can support users' decision-making, for example, in deciding whether to act now or wait for the next forecast. We did not find any link between jumpiness and skill, indicating that users should not rely on the consistency between successive forecasts as a measure of confidence. Instead, we suggest that users should use ensemble spread and probabilistic information to assess forecast uncertainty, and consider multimodel combinations to reduce the effects of jumpiness.

SIGNIFICANCE STATEMENT: Forecasting the tracks of tropical cyclones is essential to mitigate their impacts on society. Numerical weather prediction models provide valuable guidance, but occasionally there is a large jump in the predicted track from one run to the next. This jumpiness complicates the creation and communication of consistent forecast advisories and early warnings. In this work we aim to better understand forecast jumpiness and we provide practical information to forecasters to help them better use the model guidance. We show that the jumpiest cases are different for different modeling centers, that recent model upgrades have reduced forecast jumpiness, and that there is not a strong link between jumpiness and forecast skill.

KEYWORDS: Tropical cyclones; Ensembles; Forecast verification/skill

1. Introduction

Official forecasts of tropical cyclone (TC) tracks are typically based on guidance from numerical weather prediction (NWP) models (Conroy et al. 2023). NWP ensemble forecasts are increasingly being used. Although their use in official forecasts is often limited to the ensemble mean (EM) track, there is increasing evidence of the benefits of using more of the ensemble probabilistic information (Titley et al. 2019, 2020; Kawabata and Yamaguchi 2020; Leonardo and Colle 2017). One benefit of using ensembles is the increased consistency between consecutive forecasts (Buizza 2008; Zsoter et al. 2009). There are nevertheless occasions where an ensemble is unexpectedly jumpy with the predicted TC locations flip-flopping over several consecutive forecasts (Magnusson et al. 2021). Such cases can be difficult to interpret, complicating the creation of consistent forecast advisories and early warning communications. Understanding the frequency and reasons for these cases as well as information about the overall levels of consistency in operational ensemble forecasts can help forecasters to better use the available ensemble track data.

As new forecast information arrives (usually every 6–12 h for global NWP models), forecasters need to decide how to revise their forecasts to take account of the new forecast information. National Hurricane Center (NHC) Tropical Cyclone Advisories often discuss the change in forecast track due to updated guidance, making adjustments to the path depending on the new information. There is a balance to be struck between closely following the changed model guidance and taking a more conservative approach of making a smaller change to minimize the potential need to make a change in the opposite direction later, that is to avoid a so-called windshield-wiper effect (Broad et al. 2007). Contradictory messages from such jumpiness can cause difficulties for decision-makers and reduce users' confidence in the forecasts (Hewson 2020; Pappenberger et al. 2011b; McLay 2011;

Denotes content that is immediately available upon publication as open access.

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/WAF-D-23-0113.s1.

Corresponding author: David S. Richardson, d.s.richardson@pgr.reading.ac.uk

DOI: 10.1175/WAF-D-23-0113.1

Elsberry and Dobos 1990). Information quantifying the consistency between successive probabilistic forecasts can be important to inform optimal decision-making, such as whether to act now or wait for the next forecast (Regnier and Harr 2006; Jewson et al. 2021, 2022). Both noted that such information is not readily available to users.

Evaluation of operational ensemble TC track forecasts includes EM track errors, ensemble spread, and strike probability (e.g., Cangialosi 2022; Haiden et al. 2022; Titley et al. 2020; Heming et al. 2019; Leonardo and Colle 2017). However, few authors have addressed the jumpiness of TC track forecasts. Elsberry and Dobos (1990) investigate consistency of TC guidance for the western North Pacific by using the difference in cross-track errors between successive forecasts. Fowler et al. (2015) assess consistency of Atlantic TC track forecasts by counting forecast crossovers—how often in a sequence of forecasts the predicted position changes from one side to the other of a fixed reference track, for example the observed track. However, they caution that biased forecasts may appear to be consistent since successive forecasts may jump considerably without crossing the observed track. Both Elsberry and Dobos (1990) and Fowler et al. (2015) recommend the regular evaluation of forecast consistency in addition to the standard assessments of forecast accuracy.

More generally, there has been limited investigation of forecast jumpiness, especially for ensemble forecasts. Zsoter et al. (2009) considered flip-flops in sequences of forecasts all valid for a given time and showed that EM forecasts are more consistent than the corresponding ensemble control forecasts. Griffiths et al. (2019) introduced a flip-flop index to compare the consistency of automated and manual forecasts, while Ruth et al. (2009) assessed how model output statistics improved forecast consistency. Forecast consistency has been considered for rainfall (Ehret 2010) and river flow (Pappenberger et al. 2011a).

These previous studies were mainly focused on deterministic forecasts (either single runs or EM) and the methods are not directly applicable to assess the jumpiness in sequence of ensemble forecasts taking account of the full ensemble distribution. Recently, Richardson et al. (2020) introduced a measure of forecast jumpiness based on forecast divergence that accounts for all aspects of the ensemble empirical distribution. They used this to investigate jumpiness of ensemble forecasts for the large-scale flow over the Euro-Atlantic region.

In the present study we apply the forecast jumpiness measure introduced by Richardson et al. (2020) to ensemble forecasts of Atlantic TCs, focusing on the run-to-run consistency in the cross-track direction which is most important in determining the location of TC landfall. The aim is to provide forecasters and model developers with information about the jumpiness of ensemble TC forecasts. This will help forecasters and decision-makers better understand the expected changes between successive forecasts. We address the following questions:

- How does run-to-run jumpiness vary from case to case and between the ensemble systems of different NWP centers?
- Is there a common cause of "jumpy" cases—are the ensembles from different centers particularly jumpy for the same TC cases and if so what is the reason?
- Have recent ensemble model upgrades had a noticeable effect on the forecast consistency?
- What guidance should be provided to forecasters and decision-makers on the ensemble jumpiness – what information is practically useful? Is there any useful link between jumpiness and skill?

We investigate these questions using ensemble forecast data from three global NWP centers. The data used in this study and the methods to assess forecast jumpiness are introduced in sections 2 and 3. Results are presented in section 4. We start with a case study to illustrate the issues of ensemble TC track jumpiness. Then we look at the overall jumpiness over the 2019, 2020 and 2021 Atlantic hurricane seasons. Finally, we consider the relationship between jumpiness, error and spread. We conclude with a summary, recommendations for forecasters and avenues for future work in section 5.

2. Data

In this study we investigate the run-to-run consistency of ensemble tropical cyclone track forecasts from three global centers: the European Centre for Medium-Range Weather Forecasts (ECMWF), the U.S. National Centers for Environmental Prediction (NCEP) and the Met Office. Each center runs its own tropical cyclone tracker (Conroy et al. 2023) and the resulting track forecasts are archived on the TIGGE database (Bougeault et al. 2010; Swinbank et al. 2016). We retrieve the TIGGE forecast tracks for all available dates from the Atlantic basin for 2019, 2020 and 2021 for forecasts initialized at 0000 and 1200 UTC from the ECMWF ensemble (ENS, 51 members integrated on ~18-km grid), NCEP ensemble (GEFS, 21 members, ~34-km grid until 22 September 2020; 31 members, ~25-km grid from 23 September 2020 onward), and Met Office ensemble (MOGREPS-G, 36 members, ~20-km grid). A given TC is not always tracked in every ensemble member (e.g., because the system dissipates in that member or the forecast intensity is below the threshold used in the tracking algorithm) and we exclude cases where a center has fewer than 10 members that track the TC at each forecast step.

We use the observed TC positions from International Best Track Archive for Climate Stewardship (IBTrACS; Knapp et al. 2010, 2018). We concentrate our analysis on named Atlantic tropical cyclones and for each cyclone include all 0000 and 1200 UTC verification times when the observed system is at least tropical storm strength (winds at least 34 kt; 1 kt $\approx$ 0.51 m s^{-1}) and the system is reported as tropical in IBTrACS (Titley et al. 2020; Goerss 2000). For each of these verification times we consider all available TIGGE forecasts. These include forecasts initialized when the TC is still a tropical depression. However, TIGGE forecast tracks are only generated for existing TCs, so longer lead-time forecasts are not always available for verification times close to when the TC is first analyzed as a tropical storm. This means that overall there are fewer forecasts for longer lead times than for shorter lead times in our sample.

We make a homogeneous sample by only including a case if the ensemble data are available from each of the three centers. This ensures that we are comparing the different centers

over the same set of cases. The total number of cases decreases with forecast lead time from 356 for 12-h forecasts to 91 for 120-h forecasts. To maintain a reasonable sample we restrict the study to forecasts of 120 h or less.

Our focus is on the changes between successive forecasts for a given verification time. We therefore need to set a minimum number of consecutive initial times over which we can assess these changes. For a given verification time t_v, we require a minimum of six consecutive forecasts, initialized at (t_v − 12 h), (t_v − 24 h), up to (t_v − 72 h), all valid for t_v. To ensure homogeneity, the same cases must be available from all three centers. With these conditions, the total number of available cases to assess the run-to-run jumpiness is 139 over the 3-yr period.

Each NWP center has made upgrades to their operational ensemble system during the 2019–21 period used in this study. A major upgrade to the GEFS was implemented on 23 September 2020, including the introduction of a new forecast model and an increase in the number of ensemble members from 20 to 30 (Zhou et al. 2022). This upgrade brought significant improvements to the ensemble performance, including for tropical cyclone forecasts. The MOGREPS-G ensemble was upgraded on 4 December 2019, including a major change to the generation of the ensemble perturbations (Inverarity et al. 2023) and revised model physics (Walters et al. 2019). This upgrade improved TC track errors (Met Office 2019).

Upgrades to the ECMWF ENS in June 2019 (Haiden et al. 2019), June 2020 (Haiden et al. 2021), and May 2021 (Rodwell et al. 2021) were neutral in terms of TC track performance, although the latter two brought improvements to intensity forecasts (Bidlot et al. 2020; Rodwell et al. 2021). A later upgrade in October 2021 did also improve TC track forecasts (Haiden et al. 2022); however, there was only one Atlantic TC in 2021 after this date. Overall, the ECMWF ensemble track forecast performance can be considered relatively stable over the period of this study. We therefore use the ENS as a reference against which to evaluate the impact of the upgrades of the other centers on ensemble jumpiness.

3. Methods

For each tropical cyclone, the observed track provides a convenient frame of reference. We consider jumpiness in a sequence of forecasts in terms of changes in the predicted cross-track location (Elsberry and Domos 1990). A positive cross-track position indicates that the forecast is to the right of observed track (facing the observed direction of travel). We also consider the links between jumpiness, ensemble error and spread. All scores—error, spread, and jumpiness—are computed in terms of the cross-track distance and are defined below.

We measure the cross-track error of the ensemble forecasts using the continuous ranked probability score (CRPS). The CRPS is widely used for evaluation of ensemble forecasts. It is a so-called proper score: if the "true" forecast probability distribution is F, a proper score ensures that the best expected score will be achieved using the forecast F rather than any other forecast distribution $G \neq F$. Hence forecasters are rewarded for honest forecasts reflecting their true beliefs. As a proper score, CRPS discourages hedging (Gneiting and Raftery 2007) and rewards both reliability and resolution (Hersbach 2000).

For an ensemble of M members f_i, $i = 1, \ldots M$ the CRPS is given in its kernel representation by

$$\mathrm{CRPS}(f) = \frac{1}{M}\sum_{i=1}^{M}|f_i - y| - \frac{1}{2M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}|f_i - f_j|, \qquad (1)$$

where y is the verifying observation (Gneiting and Raftery 2007). The first term is the mean of the absolute error of the individual ensemble members and the second term is the mean of the distances between the different ensemble members, which accounts for the ensemble spread.

The ensemble mean forecast is given by

$$\bar{f} = \frac{1}{M}\sum_{i=1}^{M} f_i. \qquad (2)$$

For a single deterministic forecast, the CRPS is equal to the mean absolute error, so the error of the ensemble mean is

$$\mathrm{CRPS}(\bar{f}) = |\bar{f} - y|. \qquad (3)$$

To allow us to compare the mean spread and error over the sample of cases, we use a measure of ensemble spread that is also based on the mean absolute difference. The spread measure which corresponds to the mean absolute error of the ensemble mean is the mean absolute deviation of ensemble members from the ensemble mean:

$$s = \frac{1}{M}\sum_{i=1}^{M}|f_i - \bar{f}|. \qquad (4)$$

On average over a large sample of cases the ensemble mean error [Eq. (3)] and spread [Eq. (4)] should be equal for a well-tuned ensemble system.

To measure the "jump" from one forecast to the next we follow Richardson et al. (2020) and use the divergence function d associated with the CRPS. For two ensembles f and g with M and N members, respectively, d is given by

$$d(f, g) = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}|f_i - g_j| - \frac{1}{2M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}|f_i - f_j| - \frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}|g_i - g_j|. \qquad (5)$$

The first term measures the distance between the two ensembles f and g, while the second and third terms reflect the variability (spread) in each ensemble, f and g, respectively. Comparing Eq. (5) to Eq. (1) shows that the divergence reduces to the CRPS if either M or N is equal to one. If both M and N are one, then d is the absolute distance $|f - g|$. The divergence d takes account of both location and spread differences between f and g and, like the CRPS, d is a proper score (Gneiting and Raftery 2007; Thorarinsdottir et al. 2013) which discourages hedging.

Consider a given verification time t_v: an ensemble forecast f valid for this time and initialized h hours before is written

$f(t_v, h)$ and individual ensemble members are $f_i(t_v, h)$. In this study $f_i(t_v, h)$ represents the distance (in km) in the cross-track direction from the observed TC location at verification time t_v. The difference between two consecutive ensemble forecasts initialized at time $(t_v - h)$ and $[t_v - (h - 12)]$ and valid for the same time t_v is

$$D(t_v, h) = d[f(t_v, h), f(t_v, h - 12)], \tag{6}$$

where d is the divergence function [Eq. (5)].

To measure the overall divergence between the sequence of L forecasts valid for a given time we use the mean divergence between successive pairs of forecasts:

$$\overline{D(t_v)} = \frac{1}{L-1}\left[\sum_{l=2}^{L} D(t_v, 12l)\right]. \tag{7}$$

Larger values of $\overline{D}$ indicate greater change (in position, spread or both) between successive forecasts in the sequence. However, it does not necessarily indicate jumpiness in the sense of flip-flopping back and forth between different solutions. For example, if in the initial ensemble forecast all members are far to the right of the observed position and subsequent forecasts become progressively closer to the observed location, this will result in large $\overline{D}$. To distinguish between "trend" cases and "flip-flop" cases, we use the difference between the first and last forecasts of the sequence to represent this overall change (trend). Subtracting this difference from $\overline{D}$ gives the divergence index (DI) introduced by Richardson et al. (2020), which highlights jumpiness (flip-flops) in the sequence:

$$\mathrm{DI}(t_v) = \overline{D(t_v)} - \frac{1}{L-1} d[f(t_v, 12L), f(t_v, 12)]. \tag{8}$$

In this way, DI will be less sensitive than $\overline{D}$ to trends caused by bias or to cases with single large jumps (resulting for example from a sudden increase in predictability). This means that the larger values of DI will be more closely related to flip-flops in the sequence of forecasts.

Our focus is on the performance of the ensemble forecast distribution and both D and DI are computed using all available ensemble members. However, because the ensemble mean (EM) track is also often used in operational forecasting we also compute the same measures for the ensemble mean. Note that for tropical cyclone tracks, the ensemble mean refers to the Euclidean mean position of the tracks from the individual ensemble members and not to a track calculated from the ensemble mean spatial fields.

The statistical significance of differences between the different centers' distributions of $\overline{D}$ and DI are assessed using the Kolmogorov–Smirnov (KS) and Mann–Whitney U (MWU) tests (Wilks 2019). Both tests are nonparametric statistical methods to compare the empirical cumulative distributions of two samples. The MWU test is mainly sensitive to differences in location (e.g., differences in the median), while the KS test is sensitive to differences in both location and shape of the distributions.

4. Results

We start with an example to illustrate the issues of jumpiness and sampling. Then we look at the overall jumpiness over 2019, 2020 and 2021 seasons. Finally, we consider the relationship between jumpiness, error and spread.

a. *Example: Hurricane Laura, August 2020*

Hurricane Laura formed initially as a tropical storm in the western tropical Atlantic on 20 August 2020 and affected several Caribbean countries. After traveling across the Caribbean, it reached hurricane strength on 25 August as it entered the Gulf of Mexico. It made landfall in Louisiana at 0600 UTC 27 August. Here we focus on the ECMWF ensemble (ENS) forecasts for 0000 UTC 27 August, just before the Louisiana landfall. Figure 1 shows the ENS tracks for Laura from forecasts initialized every 12 h between 21 and 25 August. The earliest forecasts, from 1200 UTC 20 August (not shown) to 0000 UTC 21 August were almost all to the northeast (right-hand side) of the observed track throughout the forecast, and predicted landfall most likely along the central and eastern Gulf coast. From 1200 UTC 21 August, the forecasts showed a higher probability for landfall further west, although with a large uncertainty as shown by the distribution of the tracks from the individual ensemble members. Between 0000 UTC 22 August and 0000 UTC 24 August, successive forecasts exhibited a "flip-flop" behavior, alternating between the western or more central Gulf coast as the most likely landfall location. Finally, from 1200 UTC 24 August onward, the forecasts more consistently indicated the western solution as most likely and it turned out that the observed track was at the eastern (right-hand) end of the range of predicted locations.

We can summarize the variations in successive forecasts for a fixed valid time in a box-and-whisker meteogram (Fig. 2). This shows the distribution of the position in the cross-track direction for all ensemble members valid for 0000 UTC 27 August, from forecasts initialized every 12 h between 1200 UTC 20 August (the first available forecast) and 1200 UTC 26 August. Each ENS forecast has one control forecast and 50 perturbed members. However, the number of members that successfully track Laura until 27 August is substantially below this, especially for the earlier forecasts. Figure 2 clearly shows the jumpiness of the ENS forecasts. The earlier forecasts are mainly to the right of the observed track (too far east), while the shorter-range forecasts are too far west (left of observed track). Intermediate forecasts flip-flop between left and right of the observed position. For each lead time (except the 48-h forecast from 0000 UTC 25 August), the observed track does lie within the ensemble distribution. However, the jumpiness (lack of consistency) between successive forecasts poses a challenge for forecasters trying to assess the most likely location of landfall.

This was a particularly jumpy case for the ENS (Magnusson et al. 2021) which merits further investigation. Comparing with other ensemble forecasts may help to identify possible causes. For example, if all centers display the same flip-flop behavior it might suggest a common cause, such as changes in available observational data between the different analysis times.

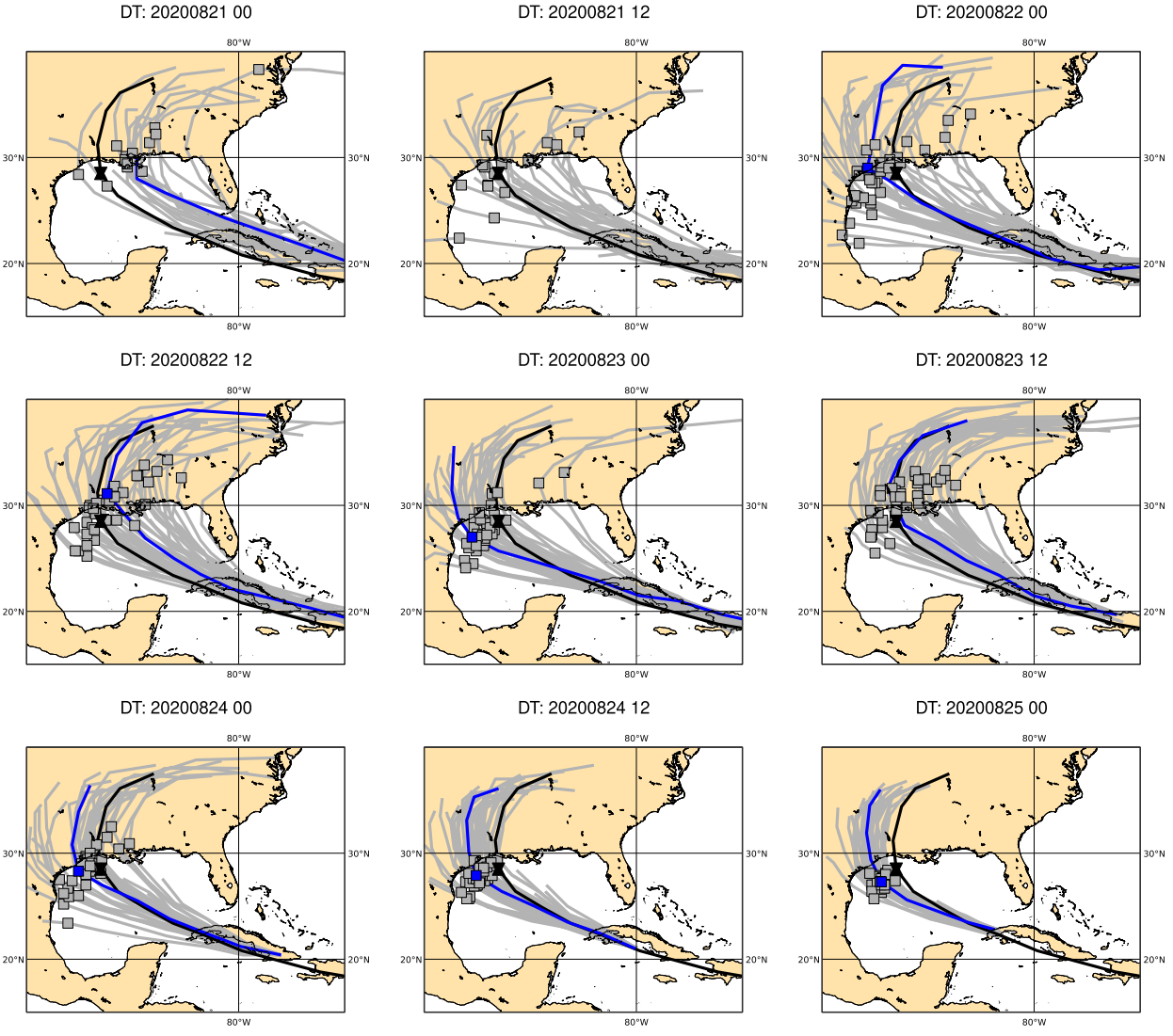

FIG. 1. Hurricane Laura: ECMWF ensemble forecast tracks (blue: control; gray: perturbed members) and observed track (black). Forecast start dates (DT) from 0000 UTC 21 Aug to 0000 UTC 25 Aug 2020. Colored symbols show forecast and observed (hourglass) position at 0000 UTC 27 Aug.

Figures 2b and 2c show the corresponding cross-track position forecasts for the MOGREPS-G and GEFS ensembles. Note that the MOGREPS-G ensemble data are missing from the TIGGE archive for forecast start times 1200 UTC 21 August and 0000 UTC 22 August. There are some similarities between all three centers: a general right bias for earlier forecasts (initialized at 0000 UTC 21 August and earlier), with a substantial proportion of members not able to track Laura as far as the verification time of 0000 UTC 27 August. Short-range forecasts for all centers are slightly left of the observed position. However, neither MOGREPS-G nor GEFS shows the same degree of flip-flop behavior as ENS.

The MOGREPS-G forecasts are the most consistent from 1200 UTC 22 August onward, with relatively small changes between successive forecasts. The GEFS forecasts maintain the initial right-hand bias for several successive forecasts, with a notable jump between 0000 and 1200 UTC 21 August. There is a second noticeable jump between 1200 UTC 23 August and 0000 UTC 24 August, after which the GEFS forecasts are generally close to the observed position, although with a small left bias. It is also worth noting that both MOGREPS-G and GEFS track Laura in all members for forecasts initialized from 1200 UTC 23 August onward, while the ECMWF ensemble does not, even for the shorter ranges. The three centers use different tracking algorithms, and this suggests differences in the sensitivity and robustness of the different trackers (Conroy et al. 2023).

This example was chosen to illustrate jumpiness in the ECMWF ENS, and in particular the flip-flops between successive forecasts. Comparison with the other centers shows that

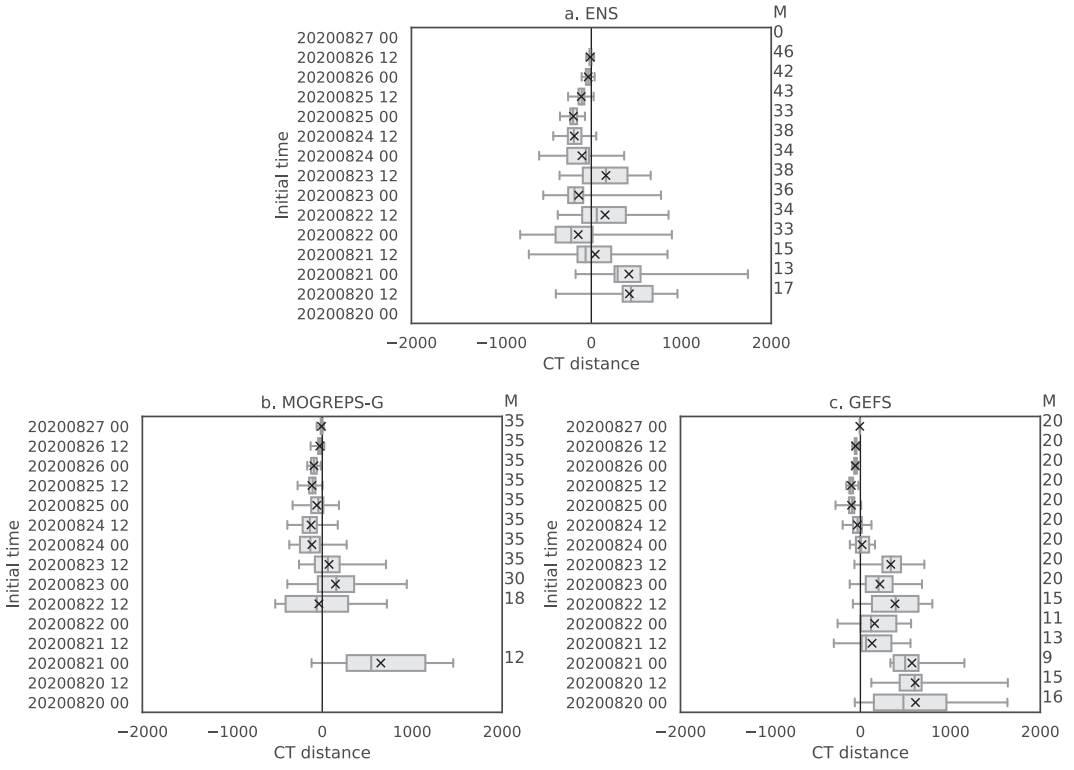

FIG. 2. Jumpiness of ensemble forecasts for hurricane Laura, valid at 0000 UTC 27 Aug 2020. Each boxplot summarizes the distribution of the cross-track (CT) errors (error at right angles to the observed direction of travel; negative values indicate left-of-track error) for one ensemble forecast (distance measured in km). Forecasts started every 12 h from 1200 UTC 20 Aug; the *y* axis shows the forecast initial time. The box-and-whisker plot shows the min, max and 25th, 50th, and 75th percentiles of the ensemble distribution (number of members shown to right of plot). The ensemble mean is shown as *X*. (a) ECMWF ENS, (b) Met Office MOGREPS-G, and (c) NCEP GEFS.

this was not a feature common to all centers. The ENS jumpiness may be related to possible issues with the data assimilation or initial perturbations, but further work is needed to investigate this (Magnusson et al. 2021). Alternatively, this could be just a chance occurrence due to the limited number of ensemble members. For each of the initial times before 25 August, 20%–30% of the ENS members did not track Laura as far as the verification time of 0000 UTC 27 August. In some cases, especially for initial times on 24 and 25 August, the ECMWF tracker misassigned some of the later forecast steps to hurricane Marco. However, this does not account for the majority of the missing tracks. These may be related to difficulties in initializing the cyclone due to the land interactions as Laura passed Puerto Rico, Hispaniola, and Cuba, while at earlier initial times, Laura was a relatively weak tropical storm and there was relatively large uncertainty in the initial analyzed position (Magnusson et al. 2021). We have recomputed the results including the corrected misassigned tracks and confirmed that this does not affect any of our conclusions.

How typical is this Laura case? To investigate how often such jumpy cases occur and whether jumpiness tends to occur for the same or different cases in different ensemble systems, the following sections consider the run-to-run consistency over all Atlantic tropical cyclones from 2019 to 2021.

b. *Ensemble jumpiness 2019–21*

To summarize the run-to-run inconsistency for a single case, we use the mean divergence $\overline{D}$ and DI, both computed over all forecasts verifying at a given time for a given tropical cyclone. The mean divergence $\overline{D}$ measures the overall change in each sequence of forecasts, while DI accounts for the trend over the sequence and highlights any flip-flop behavior.

Figure 3 shows the distribution of $\overline{D}$ and DI over all available cases for Atlantic tropical cyclones from 2019 to 2021 for the ENS, MOGREPS-G and GEFS ensembles. For $\overline{D}$, ENS has the lowest median value and smallest interquartile range, while the distribution for GEFS is noticeably broader than for the other centers. The difference between the distributions of GEFS and the other centers are statistically significant at the 1% level for both the KS and MWU tests. Although much closer to each other, the difference between ENS and MOGREPS-G distributions is significant at the 5% level for MWU test (but not significant for KS). For DI, GEFS also has the broadest distribution and ENS has the narrowest distribution. The difference between MOGREPS-G and GEFS is not statistically significant. ENS is significantly different from both MOGREPS-G and GEFS at the 5% level.

In general, a larger ensemble should give a more robust representation of the predicted distribution while a smaller

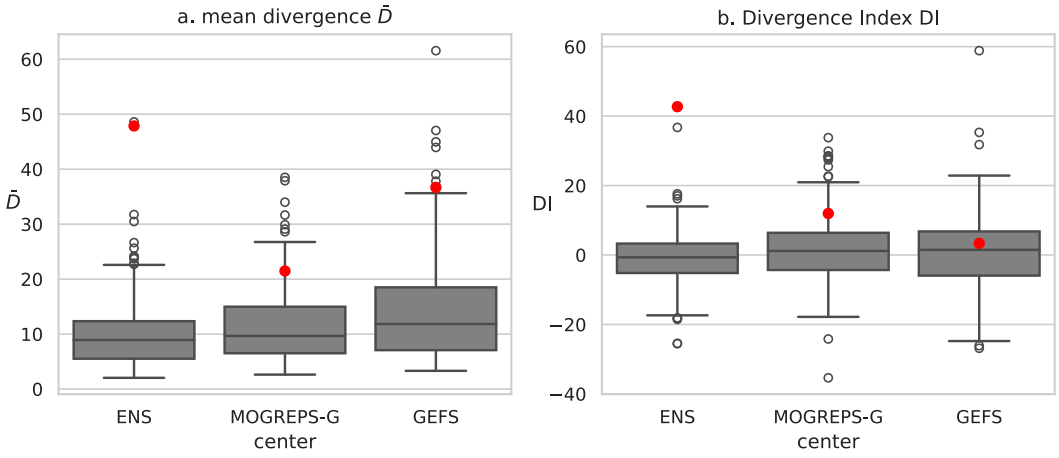

FIG. 3. Run-to-run inconsistency (jumpiness) of ensemble forecasts for Atlantic tropical cyclone tracks (2019–21). Boxplots show the distribution over all cases for the two divergence-based measures: (a) mean divergence ($\overline{D}$) and (b) divergence index (DI). Boxplots show the interquartile range and the median; the whiskers indicate the minimum and maximum values that are within 1.5 times the interquartile range; any more extreme points are shown with open circles as outliers. For both $\overline{D}$ and DI, larger positive values indicate the most inconsistent cases. The points for the example case of Hurricane Laura shown in Figs. 1 and 2 (verification time at 0000 UTC 27 Aug 2020) are marked as red filled circles.

ensemble will be more susceptible to sampling uncertainties and therefore may be expected to jump more from run to run. The above results are therefore consistent with the GEFS ensemble having fewer members than the other centers, especially before the upgrade to 31 members in September 2020. However, other factors can also influence the run-to-run consistency of the ensemble. For example, a lack of spread due to underrepresentation of either initial condition or model uncertainties would also tend to make the ensemble more jumpy. The impact of the upgrade is considered in the next subsection.

High positive values indicate the most inconsistent cases for both $\overline{D}$ and DI. For each center, points that are more than 1.5 times the interquartile range above the upper quartile are classed as outliers (marked with open circles in Fig. 3). The example case for Hurricane Laura discussed in the previous section is highlighted—this is an extreme outlier for ENS for both measures, highlighting the unusually large jumpiness for this case.

For MOGREPS-G and GEFS, this case was not an outlier for DI, consistent with the absence of flip-flops that characterized the ENS forecasts. Although not the most extreme case, this case was an outlier for GEFS using the $\overline{D}$ measure. This was due to the large right bias in the earlier GEFS forecasts. This example illustrates the difference between $\overline{D}$ and DI: ENS had several flip-flops between successive forecasts, while changes between GEFS forecasts were more associated with a trend away from the initial right bias. Both centers had large mean divergence $\overline{D}$, but the underlying cause was different. MOGREPS-G was more consistent than the other centers.

We have seen that while Laura was an example of extreme jumpiness for ENS, this was not such an extreme case for the other centers, especially for DI. Scatterplots of $\overline{D}$ and DI for pairs of centers (Fig. 4) show that this is a typical example. For each pair of centers, the number of cases that are outliers (high positive values, the most inconsistent cases) for either one center or both centers are indicated in the figure. The dashed lines in the figures indicate the threshold used for the outliers (1.5 times the interquartile range above the upper quartile). The jumpiest cases (high positive DI) for one center are in general not extremes for the other centers. For DI, none of the other ENS outliers are also outliers for either of the other centers. The results are similar for the outliers from MOGREPS-G and GEFS. There is only one case which is an outlier for more than one center, MOGREPS-G and GEFS, but that case is not an outlier for ENS. For $\overline{D}$, the highlighted Laura case is unusual in that it has high $\overline{D}$ for both ENS and GEFS, although the cause is different for each center as discussed above. However, more typically the cases of high $\overline{D}$ for one center are not exceptional for the other centers. In the scatterplots, the outliers with high $\overline{D}$ tend to lie away from the diagonal so that there are substantially more cases in the upper-left and lower-right quadrants than in the upper right.

These results suggest that the ensemble jumpiness is not strongly linked to the atmospheric situation or to the availability of observations. Rather, they suggest that individual model deficiencies or sampling uncertainties are more likely causes for the jumpiness. Sampling uncertainties will lead to run-to-run jumpiness if the ensemble is not large enough to fully represent the distribution of possible outcomes; a larger ensemble would better sample this underlying distribution and improve consistency from run to run. Alternatively, an ensemble may fail to properly represent the range of possible outcomes because the perturbations to initial conditions are not adequate or because the uncertainties in the model formulation are not sufficiently represented. Either of these will result in the ensemble spread being too small and may lead to jumpy behavior.

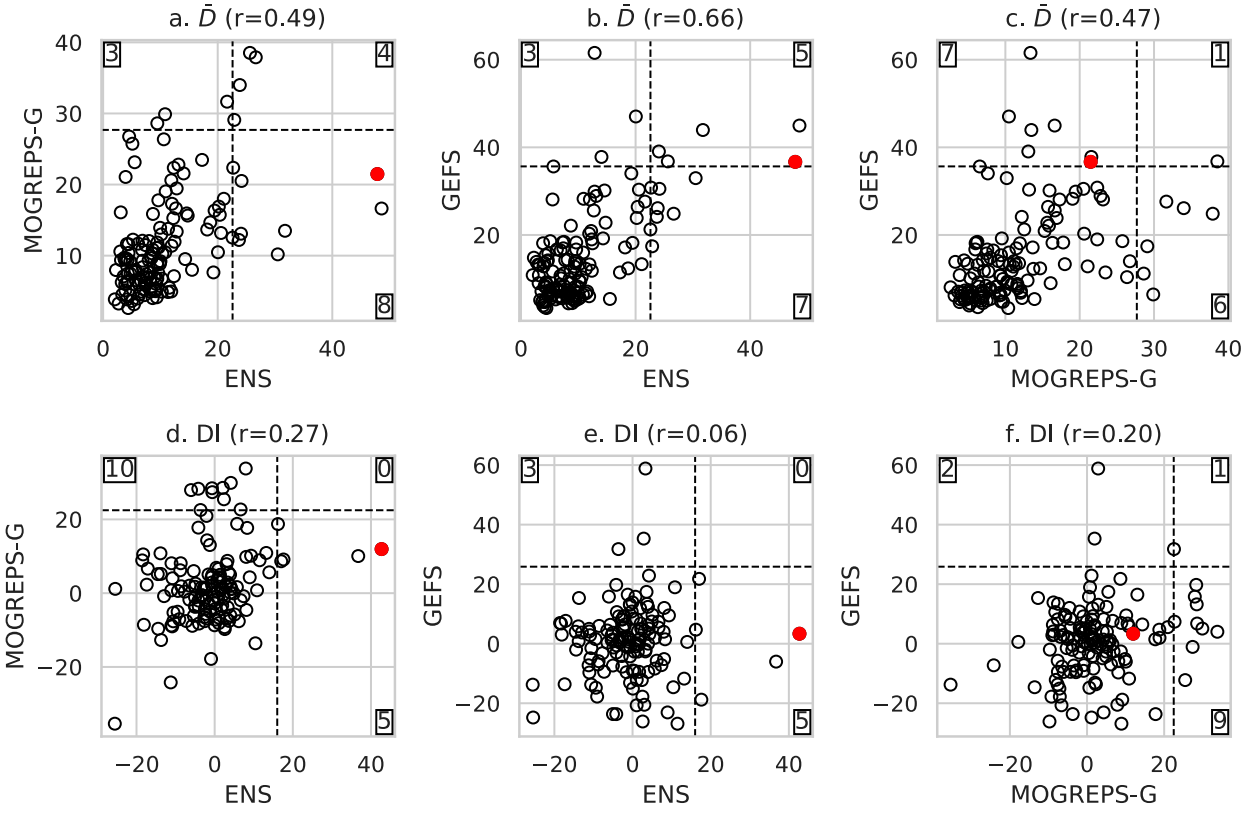

FIG. 4. Comparison of jumpiness between different centers' ensemble forecasts for Atlantic tropical cyclone tracks (2019–21). Scatterplots show the distribution of the two divergence-based measures: (top) mean divergence ($\overline{D}$) and (bottom) divergence index (DI) over all cases for pairs of centers. For both $\overline{D}$ and DI, larger positive values indicate the most inconsistent cases. Dashed lines mark the threshold for the most inconsistent outliers (1.5 times the interquartile range above the upper quartile). In each panel, the number of cases that are outliers for both centers or just one of the centers is indicated in the corresponding quadrant. The points for the example case of Hurricane Laura shown in Figs. 1 and 2 (verification time at 0000 UTC 27 Aug 2020) are marked as red filled circles.

c. *The effect of recent NWP system upgrades on ensemble jumpiness*

The results of the previous section showed that overall GEFS was more jumpy than the other centers. The GEFS upgrade in September 2020 was the most substantial upgrade of any of the centers during the study period, including a new forecast model, changes to the ensemble perturbations and an increase in the number of ensemble members. It brought a substantial improvement in the spread of tropical cyclone track forecasts (Zhou et al. 2022). Here we consider the impact of the upgrade on the jumpiness of ensemble track forecasts.

We separate our sample into two subsets initialized before (64 cases) and after (75 cases) the GEFS upgrade. In Fig. 5 we compare the empirical cumulative distribution of the mean divergence $\overline{D}$ for the three centers before (Fig. 5a) and after (Fig. 5b) the upgrade. Overall, $\overline{D}$ is significantly lower after the upgrade (comparing Figs. 5a,b). However, this applies also to the results from the other centers, suggesting that the difference is at least partly due to the differences between the observed samples. To mitigate this sampling effect, we focus on the difference between the GEFS ensemble and the other centers for the two subsets of cases.

Before the upgrade, the GEFS had substantially more cases with high values of $\overline{D}$ compared to ENS and MOGREPS (Fig. 5a). The difference in distribution compared to the other centers is highly significant at well below the 1% level for both KS and MWU tests. Differences in the distributions for ENS and MOGREPS-G are not statistically significant. After the upgrade, the GEFS distribution was much closer to those of the other centers (Fig. 5b) and there were no statistically significant differences between the distributions of any of the centers. These results show that the upgrade to the GEFS did make a significant difference to the consistency in terms of mean divergence $\overline{D}$. As for the full sample, differences in the distributions of DI are smaller (not shown); the only statistically significant difference between GEFS and either of the other centers is with ENS before the GEFS upgrade.

The GEFS upgrade brought a substantial improvement in the spread of tropical cyclone track forecasts. This was considerably underdispersive in the previous version and the upgrade resulted in a much better spread–error relationship, due to the upgrade to the stochastic model perturbations (Zhou et al. 2022). The change in $\overline{D}$ is consistent with this increase in spread for the GEFS system. In general, a larger

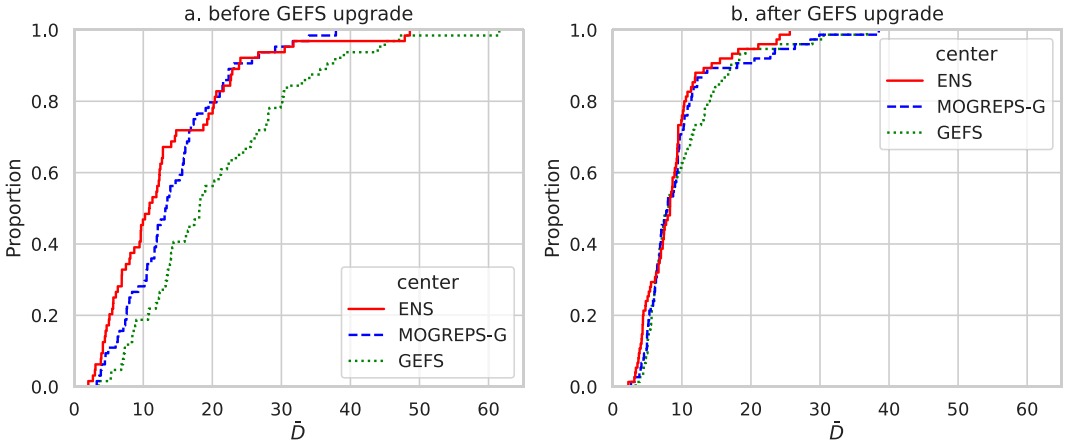

FIG. 5. Effect of GEFS v12 cycle upgrade, 23 Sep 2020. Empirical cumulative distribution function of $\overline{D}$ for subsamples of cases (a) before and (b) after the upgrade.

spread will give a broader distribution of tropical cyclone positions and the change between the set of positions for successive forecasts would tend to be less than for a less dispersive ensemble. For the same reason, the improved spread might also be expected to affect DI. Although there was some indication of this in our results (the ENS and GEFS distributions were closer and not significantly different after the upgrade), it was not such a clear change as for $\overline{D}$.

It is possible that additional factors as well as the increased spread also helped to improve $\overline{D}$. For example, a reduction in cross-track bias in the longer-lead forecasts would help to reduce $\overline{D}$, but would not tend to affect DI. Leonardo and Colle (2021) showed that the GEFS had larger cross-track errors than ENS in a large sample of Atlantic tropical cyclones for 2008–16. We were not able to identify any significant changes in the GEFS bias after the upgrade in our sample of cases. While the change in ensemble spread was large enough to identify in our sample, it may be that other differences require larger samples. Leonardo and Colle (2021) also noted that large year-to-year variability made it difficult to identify any changes due to model upgrades.

The MOGREPS-G upgrade in December 2019 also improved TC track errors and spread (Met Office 2019; Titley et al. 2020). Taking the same approach as above we found that for the subset of cases before the MOGREPS-G upgrade there was a significant difference between the ENS and MOGREPS-G distributions for both $\overline{D}$ and DI (with the MOGREPS-G having overall higher jumpiness). After the upgrade there was no significant difference between the two centers. See Fig. S1 in the online supplemental material.

We conclude that the recent upgrades to the MOGREPS-G and GEFS systems both improved the run-to-run consistency of the ensemble track forecasts, and that since these upgrades the overall jumpiness is similar for the three ensemble systems.

d. Comparison of error, spread, and divergence

We now compare the mean scores over all cases for the three different aspects of ensemble performance: error, spread and divergence. The upper panel of Fig. 6 shows the ensemble error (CRPS, left), divergence (D, center) and spread (s, right) at lead times out to 5 days ahead for the three centers. The vertical bars indicate the bootstrapped 95% confidence intervals for each center's scores. Overall, the three centers have similar performance and most differences between scores are not statistically significant.

The larger divergences in the short range for ENS and GEFS (Fig. 6b) are consistent with the lower spread (Fig. 6c) at these time steps for these centers. MOGREPS-G has larger initial spread (maybe partly due to the time-lagging of the initial conditions of the MOGREPS-G system), and this will tend to reduce the difference (divergence) between consecutive forecasts as seen in Fig. 6b.

For each center, the mean ensemble divergence (Fig. 6b) is approximately equal to the mean difference in CRPS between consecutive forecasts (difference between successive points on the curves in Fig. 6a). The agreement is particularly strong at short range for all centers, and for ENS at all forecast ranges. In other words, on average the divergence gives an indication of the expected change in error for the next forecast. However, this does not apply in individual cases.

Table 1 shows the Pearson correlation between divergence and CRPS across all available cases for each forecast lead time. For comparison, the correlation between ensemble spread and CRPS is also shown. Corresponding scatterplots are shown in Figs. S2–S5 in the online supplemental material. The association between divergence and error is in general substantially weaker than the link between spread and error. These results are consistent with previous studies that show the benefit of using spread as a measure of forecast uncertainty (Majumdar and Finocchio 2010; Yamaguchi et al. 2009; Kawabata and Yamaguchi 2020; Titley et al. 2019). However, the low correlation for divergence suggests that it does not provide useful case-to-case guidance: there is no indication that users should expect less jumpy cases to be more skillful.

Table 2 shows the Pearson correlation over all cases between the two overall measures, $\overline{D}$ and DI, and the corresponding mean error over all forecast lead times $\overline{\text{CRPS}}$. Although for $\overline{D}$ the correlation is somewhat higher than for the individual forecast steps (Table 1), the corresponding scatterplots show large

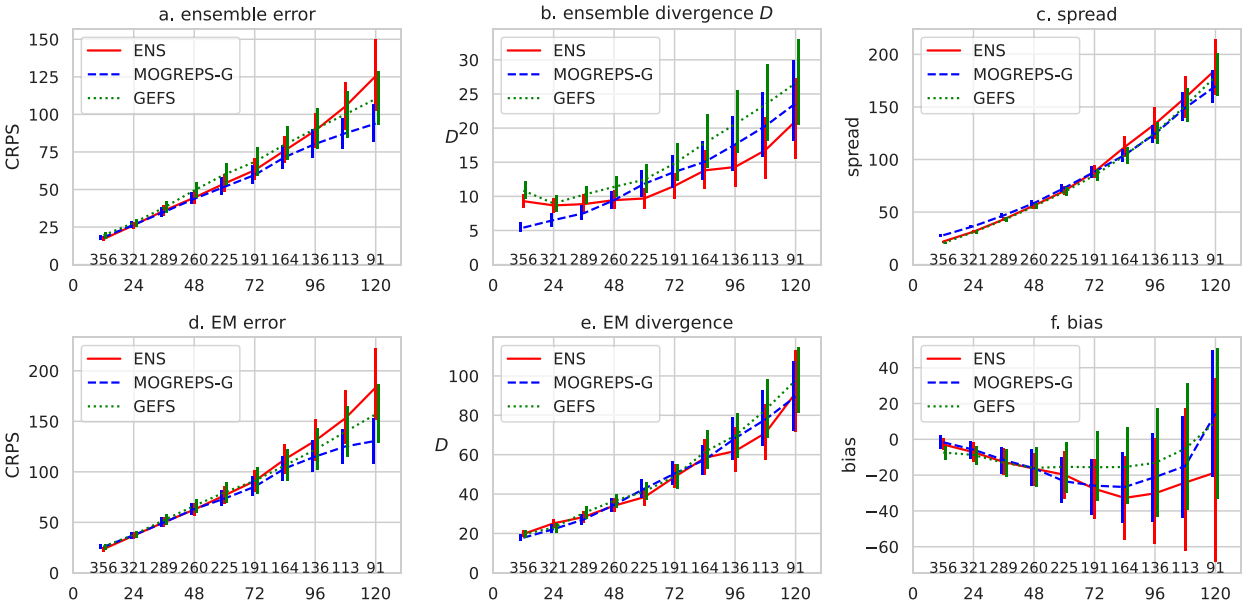

FIG. 6. Error, spread, and divergence for forecast lead time from 12 to 120 h. Scores for the (top) full ensemble and (bottom) corresponding error and divergence for the ensemble means. (a),(d) CRPS error; (b),(e) divergence; (c) ensemble spread; and (f) bias. Vertical bars indicate 95% confidence intervals. Mean scores over all available cases for each forecast lead time: number of cases indicated above the x axis.

variations in error for cases of both low and high $\overline{D}$. This again suggests that users should be cautious in individual cases—a consistent case with relatively low jumpiness may still have large overall error.

We can do the same analysis for the ensemble-mean forecasts, which are often used in operational TC forecasting (Figs. 6d,e; lower panel). Again, the divergence gives useful additional information for forecast users. For example, for ENS the ensemble mean cross-track error is around 175 km for 120-h forecasts (Fig. 6d), and the ensemble spread is similar (showing that the ensemble system is overall well-tuned; Fig. 6c). The mean expected change in cross-track EM position between $T + 120$ and $T + 108$ is ~80 km (Fig. 6e). This is similar for all three centers.

The forecast systematic error (bias) is shown in Fig. 6f. Overall, each center has a negative bias, that is the forecast positions tend to be to the left of the observed position. However, there is large uncertainty as indicated by the large confidence intervals shown on the plot. Magnusson et al. (2021) show that the ENS tends to have a left-of-track bias for northward-moving TCs, but a right-of track bias for westward moving systems and this situation-dependent variation in bias may partly explain the large confidence intervals at longer lead times. As for the other scores, the confidence intervals indicate that there is no significant difference between the biases of the different centers. Comparing Figs. 6d and 6f shows that for all centers the bias is relatively small compared to the total error.

TABLE 1. Correlation between divergence and error. Each row shows the correlation between the CRPS error at a given forecast lead time h and the divergence D between h- and $(h + 12)$-h forecasts. For comparison the correlation between the CRPS and the ensemble spread for the h-h forecasts is shown in parentheses.

Step (h)	ENS	MOGREPS-G	GEFS
72	0.18 (0.45)	0.22 (0.38)	0.07 (0.29)
84	0.25 (0.56)	0.32 (0.47)	0.05 (0.27)
96	0.19 (0.58)	0.36 (0.47)	−0.01 (0.32)
108	0.29 (0.67)	0.42 (0.41)	0.19 (0.44)

5. Conclusions

We have carried out an investigation of the jumpiness or run-to-run consistency of ensemble forecasts of tropical cyclone tracks. We used ensemble forecasts from the TIGGE tropical cyclone track archive for three global centers: ECMWF (ENS), Met Office (MOGREPS-G), and NCEP (GEFS). The forecasts were compared to the observed tracks for all named tropical cyclones from the IBTrACS archive for the Atlantic basin for 2019, 2020, and 2021.

We looked at the change in the distribution of cross-track position (relative to the observed track) for tropical cyclones in consecutive ensemble forecasts initialized at 12-h intervals.

TABLE 2. Correlation between overall jumpiness and error ($\overline{\text{CRPS}}$).

Center	$\overline{D}$ vs $\overline{\text{CRPS}}$	DI vs $\overline{\text{CRPS}}$
ENS	0.54	−0.30
MOGREPS-G	0.56	−0.01
GEFS	0.67	−0.30

This was quantified using the divergence function D associated with the CRPS error score following Richardson et al. (2020). The overall jumpiness of a sequence of forecasts all verifying at the same time was summarized using the mean divergence $\overline{D}$ and the divergence index (DI).

We present our conclusions in the framework of the questions posed in the introduction.

a. How does run-to-run jumpiness vary from case to case and between the ensemble systems of different NWP centers?

The distribution of DI was similar for each center, showing substantial variation between centers with a few significant outliers. There was no strong agreement between the centers on which cases were most jumpy. The case shown for Hurricane Laura was a typical example: this was the most extreme case of jumpiness (largest DI) for the ECMWF ENS, showing a clear flip-flopping of the ensemble between being left and right of the observed track in successive forecasts. This behavior was not apparent in either the MOGREPS or GEFS ensembles. This case also illustrated the difference between the two summary measures $\overline{D}$ and DI. Earlier GEFS forecasts were substantially to the right of the observed track and this right-of-track bias decreased in later forecasts. The large trend over successive forecasts is indicated in the relatively high mean divergence. However, the absence of the flip-flop behavior seen in the ECMWF ENS results in the DI being close to the overall median value. Using the combination of both $\overline{D}$ and DI can help to distinguish these different behaviors in a sequence of forecasts.

b. Is there a common cause of "jumpy" cases—Are the ensembles from different centers particularly jumpy for the same cases and if so, what is the reason?

The jumpiest cases were different for each center for both $\overline{D}$ and DI, indicating that there is not a common cause of jumpiness across the different ensemble systems. This suggests that the ensemble jumpiness is not strongly related to the prevailing atmospheric conditions or to the available observations.

Outliers for the different centers may be due more to specific issues in the data assimilation, models or ensemble configurations. Recent studies highlight both continuing progress and ongoing challenges in each of these areas (e.g., Magnusson et al. 2019, 2021). However, a deeper analysis of outliers would require a substantially larger sample than we have used and is beyond the scope of the present work. Leonardo and Colle (2021) used 9 years (2008–16) of Atlantic TC data to investigate the causes of large cross-track errors in the GEFS and ENS. However, we have also seen that recent upgrades to ensemble systems have led to a significant reduction in the ensemble jumpiness and therefore including a longer sample of earlier years may not be representative of the current ensemble capabilities.

Another possible reason for the occasional cases of large jumpiness is sampling uncertainty due to finite ensemble size. This would be consistent with outliers occurring at different times for the different centers. Richardson (2001) showed how even a well-tuned ensemble will appear unreliable if it has insufficient members and that the required number of ensemble members depends on both the underlying distribution and the needs of the users. Leutbecher (2019) and Craig et al. (2022) have demonstrated substantial sensitivity to ensemble size in studies using large ensembles of 200 members and 1000 members, respectively. Kondo and Miyoshi (2019) suggest that up to 1000 ensemble members are necessary to represent important aspects of some forecast distributions. The impact of ensemble size on forecast jumpiness has not been investigated and is a topic for future work.

c. Have recent ensemble model upgrades had a noticeable effect on the forecast jumpiness?

In this study we used a 3-yr period to provide a sufficient number of cases to assess. During this period upgrades to both the MOGREPS-G and GEFS ensembles resulted in substantial improvements to their predictions of TC tracks. Using the ECMWF ENS as a reference, we found that both these upgrades significantly reduced the jumpiness of the ensembles. Before the upgrades the ENS was significantly less jumpy than the other centers. However, after the upgrades there was no significant different between the centers. Both upgrades increased the spread of the ensembles, and the improved jumpiness is consistent with this change. These results suggest that it is the overall level of ensemble spread that is important and that differences in initialization and perturbation methodology between the current systems are not a major factor in determining the overall level of ensemble jumpiness.

The more recent upgrade to the ENS at the end of 2021 improved TC track errors by 10% but had little impact on the overall spread (Haiden et al. 2022). This improved the statistical reliability of the TC track. The impact on jumpiness of this upgrade has not been assessed but can be done once a sufficient sample of cases is available.

d. What guidance should be provided to forecasters and decision-makers on the ensemble jumpiness—What information is practically useful? Is there any useful link between jumpiness and skill?

The divergence D gives an indication of the expected change in cross-track position from one forecast to the next. For example, a user should expect on average that the ensemble mean position will change by around 80–90 km in the cross-track direction between a forecast for 120 h ahead and the 108-h forecast for the same time made 12 h later. The expected change between a 72- and 60-h forecast is around 50 km. These expected changes were similar for all three centers. Corresponding values for the expected divergence for the full ensemble distributions are 20–25 and 10–15 km, respectively. These results address the user requirements identified for example by Regnier and Harr (2006) and Jewson et al. (2022) to provide objective measures of the expected change from run to run so that users can take account of this in their decision-making.

We did not find any strong link between either $\overline{D}$ or DI and error (CRPS). This indicates that users should not rely on the jumpiness or consistency between successive forecasts as measure of confidence in the forecasts. This is consistent with the

work of Zsoter et al. (2009) who found only a weak link between jumpiness and error in ensemble forecasts for Europe. In contrast, ensemble spread and the ensemble probabilistic information (e.g., strike probabilities) have been shown to provide useful situation-dependent guidance on forecast uncertainty (Majumdar and Finocchio 2010; Leonardo and Colle 2017; Titley et al. 2020; Kawabata and Yamaguchi 2020).

Although we note that the effect of more recent system upgrades has not yet been evaluated, users should expect generally similar levels of jumpiness in the three ensemble systems considered in this study. The jumpiest cases will tend to be different for the different centers, likely to be a result of sampling uncertainties or specific deficiencies in the individual ensemble configurations.

One practical approach for users to adopt to address both these potential sources of jumpiness would be to combine the ensemble forecasts from the different centers into multimodel ensembles. Such multimodel combinations have already been shown to improve probabilistic TC track prediction (Yamaguchi et al. 2012; Leonardo and Colle 2017; Titley et al. 2020; Kawabata and Yamaguchi 2020). Another option would be to use lagged ensembles, combining consecutive forecasts from one center. By construction this will reduce jumpiness and this is already used in the MOGREPS-G system to increase ensemble size. Although our aim in this study was to evaluate and compare the jumpiness in the individual systems, the effect of multimodel combinations on ensemble jumpiness is an area for future work.

Acknowledgments. This work is based on TIGGE data. The International Grand Global Ensemble (TIGGE) is an initiative of the World Weather Research Programme (WWRP). David Richardson is supported by a Wilkie Calvert Ph.D. studentship at the University of Reading. We thank Linus Magnusson, Sharanya Majumdar, and two anonymous reviewers for their valuable comments.

Data availability statement. The forecast data used in this study are available from The International Grand Global Ensemble (TIGGE) Model Tropical Cyclone Track Data, Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory at https://doi.org/10.5065/D6GH9GSZ (Bougeault et al. 2010; Swinbank et al. 2016). The observed tropical cyclone tracks are available from NOAA's International Best Track Archive for Climate Stewardship (IBTrACS) archive at https://doi.org/10.25921/82ty-9e16 (Knapp et al 2010, 2018).

REFERENCES

Bidlot, J.-R., F. Prates, R. Ribas, A. Mueller-Quintino, M. Crepulja, and F. Vitart, 2020: Enhancing tropical cyclone wind forecasts. *ECMWF Newsletter*, No. 164, ECMWF, Reading, United Kingdom, 33–37, https://www.ecmwf.int/en/elibrary/81182-enhancing-tropical-cyclone-wind-forecasts.

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

Broad, K., A. Leiserowitz, J. Weinkle, and M. Steketee, 2007: Misinterpretations of the "cone of uncertainty" in Florida during the 2004 hurricane season. *Bull. Amer. Meteor. Soc.*, **88**, 651–668, https://doi.org/10.1175/BAMS-88-5-651.

Buizza, R., 2008: The value of probabilistic prediction. *Atmos. Sci. Lett.*, **9**, 36–42, https://doi.org/10.1002/asl.170.

Cangialosi, J. P., 2022: National Hurricane Center forecast verification report: 2021 hurricane season. NOAA/National Hurricane Center Rep., 76 pp., https://www.nhc.noaa.gov/verification/pdfs/Verification_2021.pdf.

Conroy, A., and Coauthors, 2023: Track forecast: Operational capability and new techniques—Summary from the Tenth International Workshop on Tropical Cyclones (IWTC-10). *Trop. Cyclone Res. Rev.*, **12**, 64–80, https://doi.org/10.1016/j.tcrr.2023.05.002.

Craig, G. C., M. Puh, C. Keil, K. Tempest, T. Necker, J. Ruiz, M. Weissmann, and T. Miyoshi, 2022: Distributions and convergence of forecast variables in a 1,000-member convection-permitting ensemble. *Quart. J. Roy. Meteor. Soc.*, **148**, 2325–2343, https://doi.org/10.1002/qj.4305.

Ehret, U., 2010: Convergence index: A new performance measure for the temporal stability of operational rainfall forecasts. *Meteor. Z.*, **19**, 441–451, https://doi.org/10.1127/0941-2948/2010/0480.

Elsberry, R. L., and P. H. Dobos, 1990: Time consistency of track prediction aids for western North Pacific tropical cyclones. *Mon. Wea. Rev.*, **118**, 746–754, https://doi.org/10.1175/1520-0493(1990)118<0746:TCOTPA>2.0.CO;2.

Fowler, T. L., B. G. Brown, J. H. Gotway, and P. Kucera, 2015: Spare change: Evaluating revised forecasts. *MAUSAM*, **66**, 635–644, https://doi.org/10.54302/mausam.v66i3.572.

Gneiting, T., and A. E. Raftery, 2007: Strictly proper scoring rules, prediction, and estimation. *J. Amer. Stat. Assoc.*, **102**, 359–378, https://doi.org/10.1198/016214506000001437.

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.

Griffiths, D., M. Foley, I. Ioannou, and T. Leeuwenburg, 2019: Flip-flop index: Quantifying revision stability for fixed-event forecasts. *Meteor. Appl.*, **26**, 30–35, https://doi.org/10.1002/met.1732.

Haiden, T., M. Janousek, F. Vitart, L. Ferranti, and F. Prates, 2019: Evaluation of ECMWF forecasts, including the 2019 upgrade. ECMWF Tech. Memo. 853, 56 pp., https://doi.org/10.21957/mlvapkke.

——, ——, ——, Z. Ben-Bouallegue, L. Ferranti, C. Prates, and D. Richardson, 2021: Evaluation of ECMWF forecasts, including the 2020 upgrade. ECMWF Tech. Memo. 880, 56 pp., https://doi.org/10.21957/6njp8byz4.

——, ——, ——, ——, ——, F. Prates, and D. Richardson, 2022: Evaluation of ECMWF forecasts, including the 2021 upgrade. ECMWF Tech. Memo. 902, 56 pp., https://doi.org/10.21957/xqnu5o3p.

Heming, J. T., and Coauthors, 2019: Review of recent progress in tropical cyclone track forecasting and expression of uncertainties. *Trop. Cyclone Res. Rev.*, **8**, 181–218, https://doi.org/10.1016/j.tcrr.2020.01.001.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting*, **15**, 559–570, https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Hewson, T., 2020: Use and verification of ECMWF products in member and co-operating states (2019). ECMWF Tech. Memo. 860, 42 pp., https://doi.org/10.21957/80s471ib1.

Inverarity, G. W., and Coauthors, 2023: Met Office MOGREPS-G initialisation using an Ensemble of Hybrid Four-Dimensional Ensemble Variational (En-4DEnVar) data assimilations. *Quart. J. Roy. Meteor. Soc.*, **149**, 1138–1164, https://doi.org/10.1002/qj.4431.

Jewson, S., S. Scher, and G. Messori, 2021: Decide now or wait for the next forecast? Testing a decision framework using real forecasts and observations. *Mon. Wea. Rev.*, **149**, 1637–1650, https://doi.org/10.1175/MWR-D-20-0392.1.

——, ——, and ——, 2022: Communicating properties of changes in lagged weather forecasts. *Wea. Forecasting*, **37**, 125–142, https://doi.org/10.1175/WAF-D-21-0086.1.

Kawabata, Y., and M. Yamaguchi, 2020: Probability ellipse for tropical cyclone track forecasts with multiple ensembles. *J. Meteor. Soc. Japan*, **98**, 821–833, https://doi.org/10.2151/jmsj.2020-042.

Knapp, K. R., M. C. Kruk, D. H. Levinson, H. J. Diamond, and C. J. Neumann, 2010: The International Best Track Archive for Climate Stewardship (IBTrACS): Unifying tropical cyclone best track data. *Bull. Amer. Meteor. Soc.*, **91**, 363–376, https://doi.org/10.1175/2009BAMS2755.1.

——, H. J. Diamond, J. P. Kossin, M. C. Kruk, and C. J. Schreck III, 2018: International Best Track Archive for Climate Stewardship (IBTrACS) Project, version 4. NOAA/National Centers for Environmental Information, accessed 26 May 2022, https://doi.org/10.25921/82ty-9e16.

Kondo, K., and T. Miyoshi, 2019: Non-Gaussian statistics in global atmospheric dynamics: A study with a 10240-member ensemble Kalman filter using an intermediate atmospheric general circulation model. *Nonlinear Processes Geophys.*, **26**, 211–225, https://doi.org/10.5194/npg-26-211-2019.

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.

——, and ——, 2021: An investigation of large cross-track errors in North Atlantic tropical cyclones in the GEFS and ECMWF ensembles. *Mon. Wea. Rev.*, **149**, 395–417, https://doi.org/10.1175/MWR-D-20-0035.1.

Leutbecher, M., 2019: Ensemble size: How suboptimal is less than infinity? *Quart. J. Roy. Meteor. Soc.*, **145** (Suppl. 1), 107–128, https://doi.org/10.1002/qj.3387.

Magnusson, L., and Coauthors, 2019: ECMWF activities for improved hurricane forecasts. *Bull. Amer. Meteor. Soc.*, **100**, 445–458, https://doi.org/10.1175/BAMS-D-18-0044.1.

——, and Coauthors, 2021: Tropical cyclone activities at ECMWF. ECMWF Tech. Memo. 888, 140 pp., https://www.ecmwf.int/sites/default/files/elibrary/2021/20228-tropical-cyclone-activities-ecmwf.pdf.

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.

McLay, J. G., 2011: Diagnosing the relative impact of "sneaks," "phantoms," and volatility in sequences of lagged ensemble probability forecasts with a simple dynamic decision model. *Mon. Wea. Rev.*, **139**, 387–402, https://doi.org/10.1175/2010MWR3449.1.

Met Office, 2019: Parallel Suite 43 release notes. Met Office, accessed 18 May 2023, https://www.metoffice.gov.uk/services/data/met-office-data-for-reuse/ps43_ftp.

Pappenberger, F., K. Bogner, F. Wetterhall, Y. He, H. L. Cloke, and J. Thielen, 2011a: Forecast convergence score: A forecaster's approach to analysing hydro-meteorological forecast systems. *Adv. Geosci.*, **29**, 27–32, https://doi.org/10.5194/adgeo-29-27-2011.

——, H. L. Cloke, A. Persson, and D. Demeritt, 2011b: HESS opinions "On forecast (in)consistency in a hydro-meteorological chain: Curse or blessing?" *Hydrol. Earth Syst. Sci.*, **15**, 2391–2400, https://doi.org/10.5194/hess-15-2391-2011.

Regnier, E., and P. A. Harr, 2006: A dynamic decision model applied to hurricane landfall. *Wea. Forecasting*, **21**, 764–780, https://doi.org/10.1175/WAF958.1.

Richardson, D. S., 2001: Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quart. J. Roy. Meteor. Soc.*, **127**, 2473–2489, https://doi.org/10.1002/qj.49712757715.

——, H. L. Cloke, and F. Pappenberger, 2020: Evaluation of the consistency of ECMWF ensemble forecasts. *Geophys. Res. Lett.*, **47**, e2020GL087934, https://doi.org/10.1029/2020GL087934.

Rodwell, M. J., and Coauthors, 2021: IFS upgrade provides more skilful ensemble forecasts. *ECMWF Newsletter*, No. 168, ECMWF, Reading, United Kingdom, 18–23, https://www.ecmwf.int/sites/default/files/elibrary/2021/20115-ifs-upgrade-provides-more-skilful-ensemble-forecasts.pdf.

Ruth, D. P., B. Glahn, V. Dagostaro, and K. Gilbert, 2009: The performance of MOS in the digital age. *Wea. Forecasting*, **24**, 504–519, https://doi.org/10.1175/2008WAF2222158.1.

Swinbank, R., and Coauthors, 2016: The TIGGE project and its achievements. *Bull. Amer. Meteor. Soc.*, **97**, 49–67, https://doi.org/10.1175/BAMS-D-13-00191.1.

Thorarinsdottir, T. L., T. Gneiting, and N. Gissibl, 2013: Using proper divergence functions to evaluate climate models. *SIAM/ASA J. Uncertainty Quantif.*, **1**, 522–534, https://doi.org/10.1137/130907550.

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.tcrr.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.

Walters, D., and Coauthors, 2019: The Met Office Unified Model Global Atmosphere 7.0/7.1 and JULES Global Land 7.0 configurations. *Geosci. Model Dev.*, **12**, 1909–1963, https://doi.org/10.5194/gmd-12-1909-2019.

Wilks, D. S., 2019: *Statistical Methods in the Atmospheric Sciences*. 4th ed. Elsevier, 840 pp.

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, 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.

Zhou, X., and Coauthors, 2022: The development of the NCEP Global Ensemble Forecast System version 12. *Wea. Forecasting*, **37**, 1069–1084, https://doi.org/10.1175/WAF-D-21-0112.1.

Zsoter, E., R. Buizza, and D. Richardson, 2009: "Jumpiness" of the ECMWF and Met Office EPS control and ensemble-mean forecasts. *Mon. Wea. Rev.*, **137**, 3823–3836, https://doi.org/10.1175/2009MWR2960.1.