Skill and consistency of ECMWF forecasts of Atlantic tropical cyclone genesis

Skill and Consistency of ECMWF Forecasts of Atlantic Tropical Cyclone Genesis

DAVID S. RICHARDSON,[a,b] HANNAH L. CLOKE,[a,c] LINUS MAGNUSSON,[b] SHARANYA J. MAJUMDAR,[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 Atmospheric Sciences, University of Miami, Miami, Florida*

(Manuscript received 27 June 2024, in final form 29 November 2024, accepted 7 February 2025)

ABSTRACT: We evaluate the skill and jumpiness of the ECMWF medium-range ensemble (ENS) in predicting tropical cyclone genesis in the Atlantic basin. Focusing on the probabilistic performance of the ENS, we assess how far in advance the ENS can predict genesis, quantify the consistency (jumpiness) from run to run, and investigate what factors influence the skill and consistency. We find that first indications of genesis are picked up at least 7 days ahead in 50% of the observed cases, although strong signals often only appear less than 3 days before genesis. There are significant regional differences, with observed genesis events predicted 2–3 days earlier in the eastern Atlantic than in other areas. The genesis probabilities can be jumpy from run to run, and the jumpiest cases are in the more skillful regions (central and eastern Atlantic) and for situations where the initial signal for genesis appears at longer lead time. In the eastern Atlantic, there is a tendency for the ENS tracks to reach tropical storm strength earlier and further east than observed; this model bias can affect both skill and jumpiness of the genesis forecasts. Our results provide guidance to forecasters on how to use and interpret the ENS predictions. Areas for future work include the link between early intensification in the eastern Atlantic and African easterly wave activity, the relationship between skill and the TC development pathways, and the impact of systematic analysis differences between 0000 UTC and 1200 UTC on forecast intensity.

SIGNIFICANCE STATEMENT: Forecasting where and when tropical cyclones will appear increases the lead time at which decision-makers can begin to take preparatory mitigating action. Numerical weather prediction models can provide important guidance but sometimes are not consistent from one run to the next. We evaluate the skill and consistency of a state-of-the-art global model in predicting the formation of tropical cyclones up to 10 days ahead and provide guidance to forecasters on how to use and interpret the model predictions. We show that the formation of tropical cyclones can be predicted 2–3 days earlier in the eastern Atlantic than in the western Atlantic and identify some of the factors influencing both skill and consistency.

KEYWORDS: Tropical cyclones; Ensembles; Forecast verification/skill; Numerical weather prediction/forecasting; Probability forecasts/models/distribution; Model evaluation/performance

1. Introduction

Following significant progress in forecasting tropical cyclone (TC) tracks (Landsea and Cangialosi 2018) and intensity (Cangialosi et al. 2020), there is increasing focus on predicting TC genesis (Hon et al. 2023). For the Atlantic basin, the U.S. National Hurricane Center (NHC) Tropical Weather Outlook provides forecasts of TC genesis for 2 and 7 days ahead (Hon et al. 2023). By providing information about the likely development of TCs before they have formed, skillful genesis forecasts can effectively increase the lead time at which decision-makers can begin to take preparatory mitigating action.

Numerical weather prediction (NWP) forecasts including ensemble forecasts are used in operational genesis forecasts (Titley et al. 2019; Hon et al. 2023), often in combination with statistical methods (Halperin et al. 2017). Use and verification of NWP genesis forecasts has focused on deterministic aspects, assessing hits and false alarms using standard contingency-table measures such as hit rate or probability of detection, success ratio, and the threat score or critical success index (Wilks 2020). These have been applied to the high-resolution global forecasts from different centers (Halperin et al. 2016, 2013; Liang et al. 2021) to ensemble mean forecasts (Li et al. 2016; Wang et al. 2018) and to individual ensemble members (Zhang et al. 2023).

Recently, there has been increasing development of probabilistic TC genesis forecast products for operational centers (Hon et al. 2023). For example, Halperin et al. (2017) developed a statistical–dynamical tool to generate TC genesis probabilities using logistic regression models applied to the outputs from several high-resolution global NWP models. A consensus probability is also provided when more than one model predicts a genesis event. Verification using Brier scores and reliability diagrams showed that these provide useful guidance (Halperin et al. 2017), and the products are regularly used in the NHC (Hon et al. 2023). The use of probabilistic information from the

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

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

DOI: 10.1175/WAF-D-24-0115.1

ensembles is more limited, although ensemble forecasts have been shown to have skill in predicting TC genesis (Komaromi and Majumdar 2014, 2015; Majumdar and Torn 2014; Yamaguchi and Koide 2017; Yamaguchi et al. 2015).

One of the key issues limiting the uptake of ensemble TC forecasts is the run-to-run jumpiness that can occur in some situations (Dunion et al. 2023; Magnusson et al. 2021). Large jumps in the predicted probability of TC genesis between successive ensemble forecasts present a significant challenge to forecast centers and lessen users' confidence in the prediction system (McLay 2008; Elsderry and Dobos 1990; Hewson 2020; Dunion et al. 2023; Pappenberger et al. 2011). Although approaches such as multimodel combinations or lagged ensembles can help mitigate such jumpiness, it is important to identify and understand the underlying causes of such jumpy behavior. Quantifying the level of jumpiness in an ensemble system provides valuable information to the forecast user. This can be important, for example, in helping the user to decide between acting now or waiting for the next forecast (Regnier and Harr 2006; Jewson et al. 2022, 2021). Identifying the circumstances in which jumpiness occurs is an important step toward addressing the underlying cause—is it related to model or analysis uncertainty (lack of spread in the ensemble perturbations) or model bias, or is it an indication of insufficient ensemble size to give a reliable uncertainty estimate? Jumpiness of TC track forecasts has been investigated for the western North Pacific (Elsderry and Dobos 1990) and the Atlantic (Fowler et al. 2015; Richardson et al. 2024). However, there has been no corresponding assessment of TC genesis forecasts. In this study, we conduct a first assessment of the jumpiness of the European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble (ENS) forecasts for TC genesis.

Another factor limiting the use of ensemble TC genesis forecasts is the lack of routine evaluation of the products provided by the global centers. Although ECMWF regularly publishes verification results for ensemble forecasts of the track and intensity of existing TCs (Haiden et al. 2023), it does not routinely evaluate genesis forecasts, so users do not have a clear picture of ENS performance (Magnusson et al. 2021).

These knowledge gaps are addressed in this study which evaluates the skill and jumpiness of the ECMWF medium-range ENS in predicting TC genesis in the Atlantic basin. We address the following questions:

- How far in advance can the ENS forecast TC genesis in the Atlantic basin?
- How consistent from run to run are the forecasts of the observed genesis events?
- What are the factors that influence the skill and consistency of the ENS genesis forecasts and what future work will help to improve these forecasts?

In each case, we focus on the probabilistic performance of the ENS. The data we use in this study and the methods we apply to identify genesis events are described in section 2, with verification scores and consistency measures introduced in section 3. Results are presented in section 4, addressing each of the three key questions in turn. We conclude with a summary and discussion of directions for future work in section 5.

2. Data

We investigate the ability of the ECMWF ENS to predict the genesis of tropical cyclones over the Atlantic. ENS comprises 50 perturbed members integrated on ~18-km grid until 27 June 2023 and thereafter on ~9-km grid. The ECMWF tropical cyclone tracker (Magnusson et al. 2021) identifies and tracks both existing TCs and those that develop during the forecast. The tracker is applied to all ensemble members. These operational forecast tracks are archived on the TIGGE database (Bougeault et al. 2010; Swinbank et al. 2016). We retrieve the operational forecast tracks for ENS forecasts initialized at 0000 and 1200 UTC from May to December 2019–23 and consider forecast lead times from 1 to 10 days ahead.

We evaluate the forecasts against the observed TC data from the International Best Track Archive for Climate Stewardship (IBTrACS; Knapp et al. 2018, 2010). We extract the observed positions and maximum winds from all named Atlantic tropical storms (i.e., tropical cyclones that reach tropical storm strength during their life cycle). We focus our evaluation on the first time the observed system is reported as a tropical system of at least tropical storm strength (winds at least 34 kt; 1 kt $\approx$ 0.51 m s^{-1}), which we define as the genesis time for the tropical storm (TS) (Magnusson et al. 2021; Zhang et al. 2023). To ensure a consistent set of forecast lead times throughout the evaluation, we limit the verification times to also be 0000 and 1200 UTC and so the observed genesis time is the first 0000 or 1200 UTC time with wind $>$ 17 m s^{-1}. There were 98 observed tropical storms in the Atlantic basin during the 5-yr study period. However, TS Imelda (2019) was a TS for less than 12 h and was not included in the verification; therefore, we used 97 observed TS genesis in this work.

To investigate how well and how consistently the ENS can forecast the observed TS genesis events, we compute the probability of TS genesis or TS activity at the observed genesis time and location for each of the 97 observed TS.

For a given verification time t_v, we refer to an ensemble forecast f valid for this time and initialized h hours earlier as $f(t_v, h)$ and write the individual ensemble members as $f_m(t_v, h)$. Given the inherent limitations of predictability as well as uncertainties in both forecasts and observations (Landsea and Franklin 2013; Torn and Snyder 2012), we do not expect the forecast to predict genesis at exactly the time and location of the reported observed genesis event. Therefore, we define tolerances in both space and time. Several different choices have been used in previous studies (Halperin et al. 2016, 2013; Zhang et al. 2023; Magnusson et al. 2021; Yamaguchi et al. 2015). For each observed TS genesis event, we use the following procedure where t_v represents the observed genesis time:

- For the ENS forecast $f(t_v, h)$, we count how many members m have TC tracks that pass within 500 km of the observed genesis location at any time between $t_v - 24$ h and $t_v + 24$ h. We define the proportion of members m/M as the forecast probability of TC activity at the observed

TABLE 1. Different forecast sets considered in this study. Identifier used to refer to each set of forecast probabilities.

Identifier	Set	Description
FG17	Forecast TS genesis 17 m s^{-1}	Forecast TC track passes within 500 km and 24 h of given location, and the first time that wind is > 17 m s^{-1} along this track is within this time/location tolerance.
FA17	Forecast TS activity 17 m s^{-1}	Forecast TC track passes within 500 km and 24 h of given location and has wind > 17 m s^{-1}. But forecast genesis may have occurred earlier (i.e., first step with wind > 17 m s^{-1} may have occurred more than 24 h before t_v) and more than 500 km from the given location.
FA15	Forecast TC activity 15 m s^{-1}	Forecast TC track passes within 500 km and 24 h of given location and has wind > 15 m s^{-1}. But forecast genesis may have occurred earlier (i.e., first step with wind > 15 m s^{-1} may have occurred more than 24 h before t_v) and more than 500 km from the given location. Accounts for overall lower intensity in forecasts.
FATC	Forecast TC activity	Forecast TC track passes within 500 km and 24 h of given location (forecast wind may not reach TS strength).

genesis event. This gives the probability for TC but does not address the intensity or the location of genesis in the forecast. We refer to this set of forecast probabilities as FATC.

- To address the intensity, we select the subset of the forecast tracks that have maximum wind greater than a given threshold. We use 17 m s^{-1} for a direct comparison with the observed intensity but also consider lower thresholds (e.g., 15 m s^{-1}) to account for potential differences in intensity in the forecasts. We refer to these forecast activity probabilities as FA17 and FA15, respectively.
- Finally, to address the timing of the genesis, we again subset the forecast tracks to keep only those that have forecast genesis within 24 h and 500 km of the observed genesis event. We define the forecast genesis event as the first point on the track with wind greater than 17 m s^{-1} and refer to this set of forecast probabilities as FG17.

Table 1 summarizes the different sets of forecast probabilities that we consider in this study and the naming convention that we use.

For a broader perspective, to consider the overall forecast probabilities of TC genesis and to include assessment of false alarms, we also conduct some evaluation on a regular $1° \times 1°$ latitude–longitude grid. At each grid point, the forecast TS genesis probability is defined as the proportion of ENS members that predict a TS genesis event to occur within 500 km of that grid point (center of the $1° \times 1°$ box) and between 24 and 216 h ahead. Similarly, we define TS genesis to occur if there is an observed TS genesis event within 500 km of the grid point and within the same 192-h (8 day) time window.

3. Verification and consistency measures

We evaluate the ENS forecasts of TC activity and genesis using the Brier score (BS) (Wilks 2020), which is a measure of the mean-squared error of the forecast probability:

$$b = \frac{1}{N}\sum_{i=1}^{N}(p_n - y_n)^2, \tag{1}$$

where p_n is the forecast probability (proportion of ENS members that predict the event), y_n is 1 if the event occurs and 0 otherwise, and N is the total number of cases.

In the assessment of overall performance using the gridded data (section 4d), we use the observed sample climate probability of genesis $\bar{y}$ as a reference forecast:

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i. \tag{2}$$

This sample climate includes all dates in our evaluation data and is computed separately for each grid point. By construction, the sample climate has the lowest Brier score of any fixed reference forecast and so is harder to beat than a long-term climate; using this as a reference for the Brier skill score hence provides a conservative indication of forecast skill.

The Brier score of the climate forecast is given as

$$b_c = \frac{1}{N}\sum_{i=1}^{N}(\bar{y} - y_i)^2, \tag{3}$$

and the Brier skill score is then given as

$$B = \frac{b_c - b}{b_c}. \tag{4}$$

Positive values of B indicate positive skill relative to the sample climate. Maximum skill $B = 1$ is achieved for perfect deterministic forecasts.

We evaluate the hits and false alarms associated with different forecast probability thresholds using the relative operating characteristic (ROC) (Mason 1982; Ben Bouallègue and Richardson 2022) and performance diagram (Roebber 2009). The ROC is a plot of the hit rate (proportion of observed events correctly forecast) against false alarm rate (proportion of observed nonevents where genesis was forecast). The performance diagram plots the hit rate against the success ratio (proportion of genesis forecasts that were correct); the performance diagram also shows the frequency bias (number of forecast events divided by number of observed events); and the threat score (number of hits divided by the sum of hits, misses, and false alarms).

To measure the jumpiness or consistency over a sequence of forecasts, we measure the difference (divergence) d in probability between consecutive forecasts.

Here, we consider the forecasts initialized at 12-h intervals between 24 and 216 h before a given verification time t_v. The

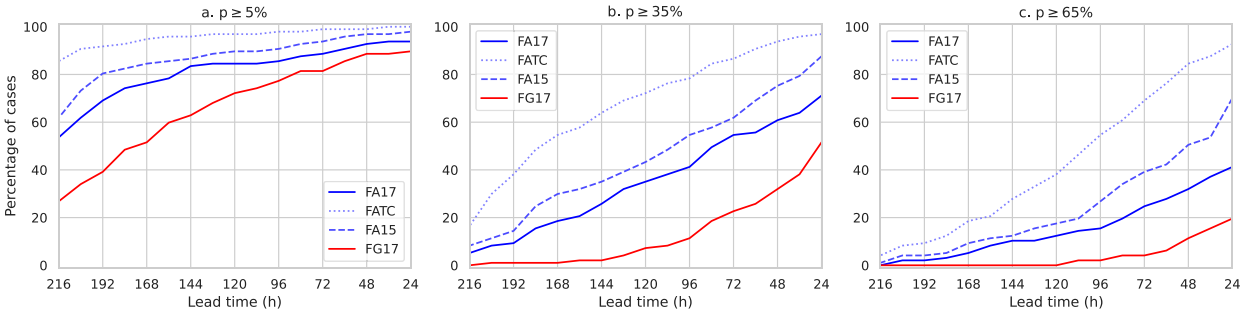

FIG. 1. Lead time of ENS forecasts of TS genesis. The percentage of cases predicted with probability of at least (a) 5% (low), (b) 35% (medium), and (c) 65% (high) at lead times from 216 to 24 h before the observed TS genesis time.

probability of the given event (TC activity or TS genesis) in the ENS forecast initialized at $t_v - h$ is written as $p(t_v, t_v - h)$, and the difference between consecutive forecasts is

$$D(t_v, h) = d[f(t_v, h), f(t_v, h - 12)] = |p(t_v, h) - p(t_v, h - 12)|. \quad (5)$$

The mean divergence over the full sequence of $L = 17$ initial times is

$$\overline{D(t_v)} = \frac{1}{L-1}\left[\sum_{l=1}^{L-1} D(t_v, 24 + 12l)\right]. \quad (6)$$

The minimum value of $\overline{D}$ is zero, indicating that the forecast probability does not change over the set of forecasts, while larger values indicate greater differences in probability between successive forecasts in the sequence.

For each observed genesis event, we expect that the forecast probability will be low at the longest forecast ranges (close to the climatological probability) and will increase, ideally reaching close to 100% at the shortest forecast ranges. To account for the expected increase in probability over the sequence of forecasts, we use the difference between the probabilities from the first and last forecasts of the sequence to represent this overall trend. We then subtract this difference from $\overline{D}$ to give the divergence index (DI; Richardson et al. 2020, 2024):

$$\mathrm{DI}(t_v) = \overline{D(t_v)} - \frac{1}{L-1}|p[t_v, 24 + 12(L-1)] - p(t_v, 24)|. \quad (7)$$

DI summarizes the jumpiness about the overall trend over the sequence of forecasts, with larger values of DI indicating more jumpy forecasts (bigger difference in probabilities).

4. Results

First, we evaluate how far in advance the ENS can predict the observed genesis events with low, medium, and high probability. Next, we assess how consistent these probabilities are in the sequence of consecutive forecasts leading up to each observed genesis event. We then consider potential factors that may affect the jumpiness and skill of these forecasts. Finally, we assess the overall skill of the ENS probability forecasts for TC genesis and activity.

a. *How far in advance can we predict the observed Atlantic TS genesis events?*

Figure 1 shows the percentage of the 97 observed genesis events that were forecast with at least 5%, 35%, and 65% probabilities at or before each forecast lead time from 216 to 24 h in advance. The probability thresholds were chosen to be consistent with the categories used to indicate low, medium, and high probability, respectively, in the NHC Tropical Weather Outlook: NHC genesis probabilities are given in 10% intervals, and their low, medium, and high probability categories are 10%–30%, 40%–60%, and 70%–100%, respectively.

The red curve shows the results for the FG17 probabilities where the forecast is required to match the observed genesis in both timing and intensity (within the specified 500-km and 24-h tolerances). Few cases are predicted with high probability, and only 20% of cases can be predicted with medium probability more than 72 h ahead. The low probability threshold is reached in over 50% of cases at 168-h lead time, indicating that the ENS is capable of generating tropical storms a week in advance although the predictability is low.

The three blue curves in Fig. 1 help to identify some of the reasons for this poor performance in the direct forecasting of the observed genesis. The solid blue curve shows the results for the FA17 probabilities. As well as the hits included in FG17, these allow for early genesis in the forecasts and indicate the proportion of ENS members that have TS activity at the observed genesis time and location. Many more cases are predicted for all three probability categories for FA17 than for FG17: More than 20% of observed events are predicted with high probability at least 72 h ahead, with the proportion increasing to over 50% for the medium probability threshold and over 80% for low probability. The 25% of cases are predicted with medium probability at least 6 days (144 h) ahead. The higher probabilities for FA17 compared to FG17 show that the timing of TS genesis is one significant difference between ENS and observed genesis, with a substantial number of forecast tracks reaching TS strength before the observed genesis time. Comparing FA17 and FA15 (solid and dashed blue lines) shows that the choice of wind threshold for the forecast tracks also affects the performance. The relatively

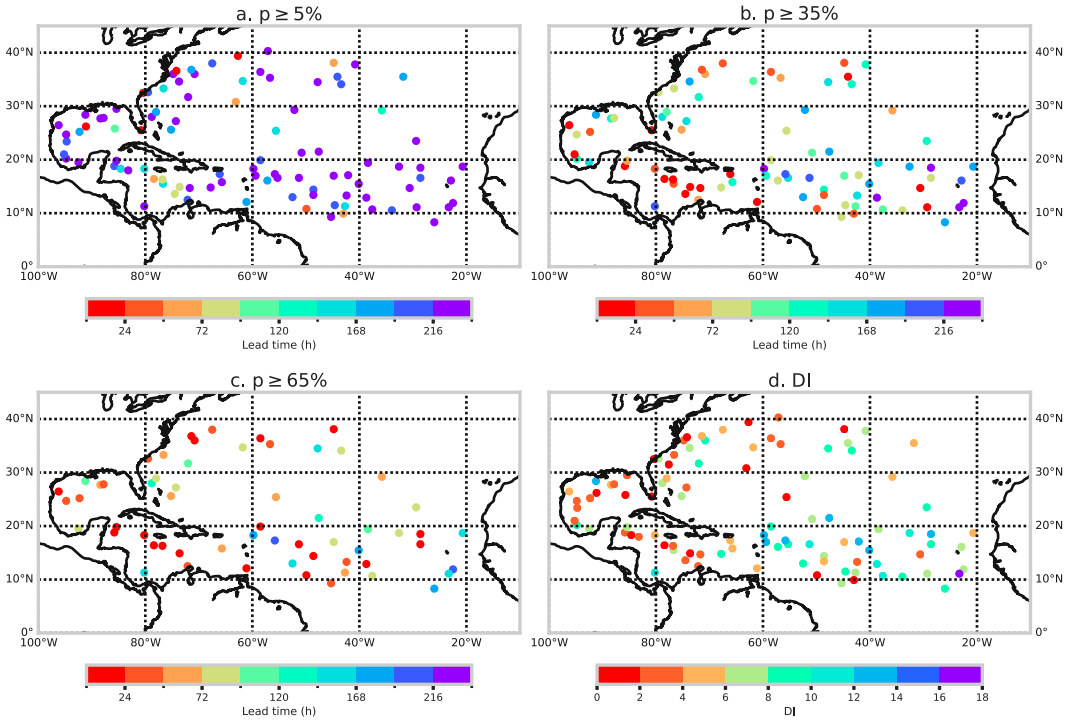

FIG. 2. Lead time of ENS forecasts of observed TS genesis events. Longest lead time (hours) at which the probability of TS activity (FA17) at the observed TS genesis location was predicted with probability of at least (a) 5%, (b) 35%, and (c) 65%. (d) The jumpiness in forecast probability for these cases, as measured by DI.

minor change of wind threshold from 17 to 15 m s^{-1} increases the proportion of correctly forecast cases by around 10% points. Larger improvements are achieved when considering all forecast tracks without specifying a minimum wind speed (FATC; dotted blue line): Around 60% of cases are predicted with medium probability at least 6 days (144 h) ahead and with high probability at least 4 days (96 h ahead). The sensitivity to wind thresholds agrees with results from other studies (Yamaguchi et al. 2015; Zhang et al. 2023).

The geographical distribution of the FA17 results is shown in Fig. 2 for each of the low/medium/high probability thresholds. The TS in the eastern Atlantic tend to be predicted earlier than those in the Caribbean and the Gulf of Mexico. In the central and eastern Atlantic (east of 60°W and south of 30°N), the median lead time for the first indications of TS activity (low 5% probability threshold) is 228 h (the longest lead we have considered here). For medium and high probability thresholds, the corresponding median lead times are 132 and 72 h, respectively. In contrast, the equivalent median lead times for the western Atlantic, Caribbean, and Gulf of Mexico (south of 30°N, west of 60°W) are 204, 48, and 36 h, respectively. In other words, the observed genesis events in the eastern Atlantic are predicted 2–3 days earlier than those in further west. The predictability for the genesis > 30°N is generally similar to that for the western Atlantic. The consistency or jumpiness of these forecasts as measured by DI is shown in Fig. 2d. Again, there are strong geographical variations, with the highest DI (jumpiest cases) in the central and east Atlantic. The median DI for this region is 8.75, more than twice the median DI value of the western and northern regions (3.5 and 4.0, respectively).

The regional differences may be associated with different tropical cyclogenesis pathways (McTaggart-Cowan et al. 2013, 2008). The more predictable (and also more jumpy) cases tend to occur in regions dominated by nonbaroclinic developments, although some of the most predictable and jumpiest genesis events occur in the Cape Verde region associated with the low-level baroclinic pathway (baroclinic development under the African easterly jet). The less predictable cases further west and north are in regions where other baroclinic pathways [tropical transition (TT); Davis and Bosart 2003, 2004; trough interaction] are more common developments. This is consistent with results from Wang et al. (2018) who found lower predictability in the TT pathways in an evaluation of reforecasts from the NCEP GEFS ensemble. It is, however, notable that there are very few predictable cases in the Caribbean and Gulf of Mexico despite the nonbaroclinic pathway also being a significant development category in this region. These nonbaroclinic pathways often originate from barotropic breakdown of vorticity along stalled fronts, which are smaller and could be less predictable, especially for a lower-resolution model. Environmental factors influencing TC genesis in the western Atlantic have been discussed by Klotzbach et al. (2022) and (in the wider context of cyclonic circulations over Central America) by Papin et al. (2017). Additional factors, such as land interactions, may also affect the model ability to

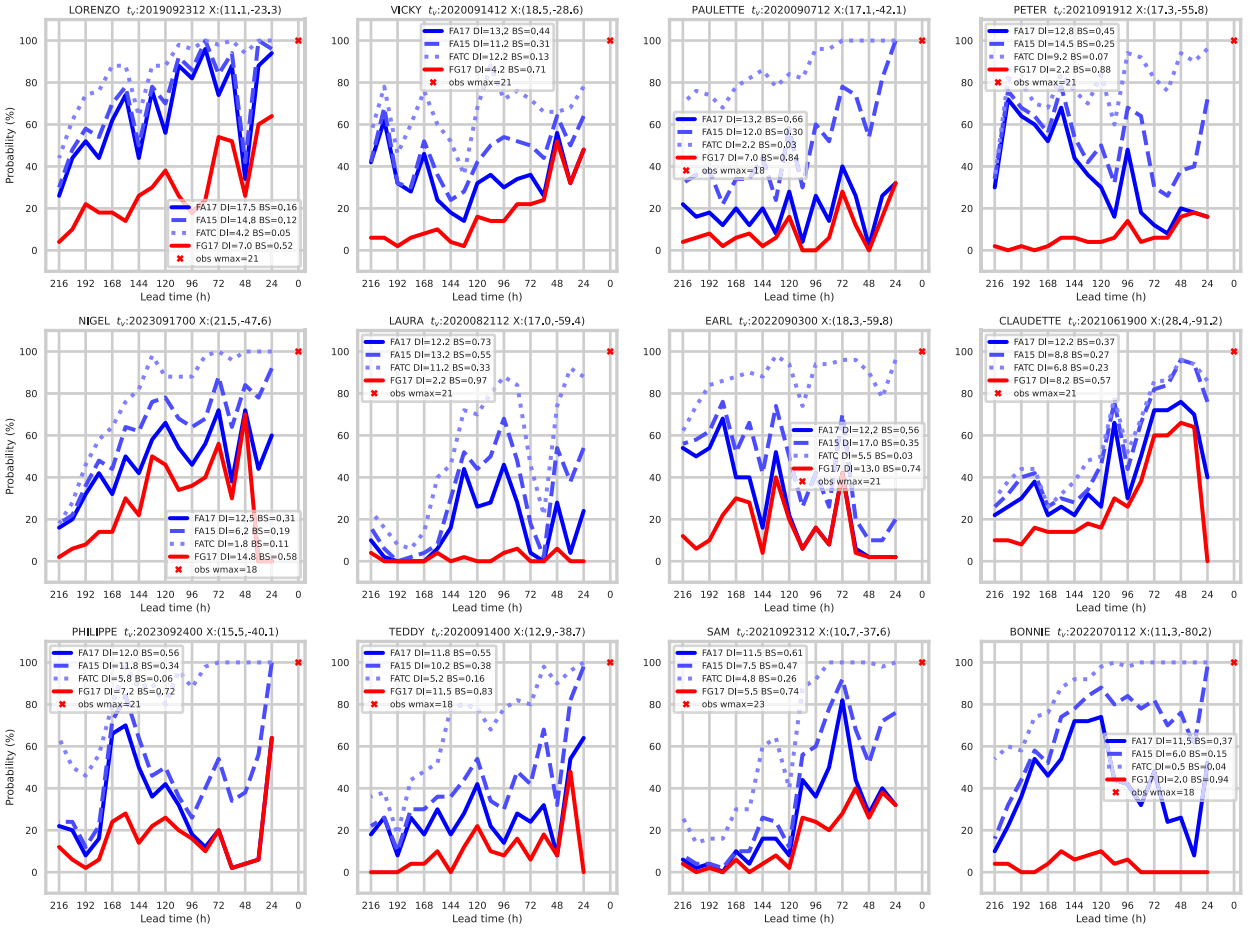

FIG. 3. Forecast probability of TS activity for the jumpiest FA17 cases. Curves show the forecast probability of TC activity at the observed genesis time t_v and location X (latitude, longitude) for forecasts initialized at 12-h intervals from 216 to 24 h before the observed genesis time. The probability for genesis (FG17) is shown by the red line, while the three blue curves show the probability of TC activity with different wind intensity thresholds FA17 (solid dark blue), FA15 (dashed blue), and FATC (dotted light blue). The legend shows the jumpiness (DI) and error (BS) for each.

correctly predict genesis and would have a more significant impact on genesis forecasts in the western Atlantic and Caribbean rather than eastern Atlantic; this is an area for future research.

b. *Consistency—The jumpiest forecasts of observed TS genesis events*

The run-to-run consistency of the ENS forecast probabilities is shown in Fig. 3 for the 12 cases with highest DI for the FA17 forecasts. For each case, the forecast probabilities from the forecasts initialized every 12 h from 24 to 216 h before the observed genesis event are shown for each forecast set FG17, FA17, FA15, and FATC.

Most of these jumpy cases occur in September (August for Laura), and there are cases for each of the 5 years in our sample. As seen from Fig. 2, the jumpy cases are typically in the central to east Atlantic and between 10° and 20°N. The two exceptions to both time and location are Bonnie and Claudette which were both early season TCs in the west of the basin. Claudette was the only one of these cases that did not originate from an African easterly wave.

In most cases, the jumpiness is related to the forecast intensity: The FATC probabilities are much more consistent from run to run than the FA17 probabilities, and the corresponding DI is consequently much lower. The two notable exceptions to this are Laura and Vicky, which both have substantial jumpiness for the lower wind thresholds. Interactions between African easterly waves or between these waves and other low pressure systems have also been noted to affect the forecast probabilities of genesis for cases including Laura and Paulette (Magnusson et al. 2021). In the case of Vicky, we note that Teddy and Vicky originated from successive easterly waves that developed off the coast of Africa on 10 and 11 September 2020. The earlier ENS forecasts tended to favor a development associated with Vicky with tracks moving north-westward away from the coast of Africa, while later forecasts produced more westward tracks associated with Teddy. This uncertainty about which would be the stronger development, together with potential interactions between the two, may

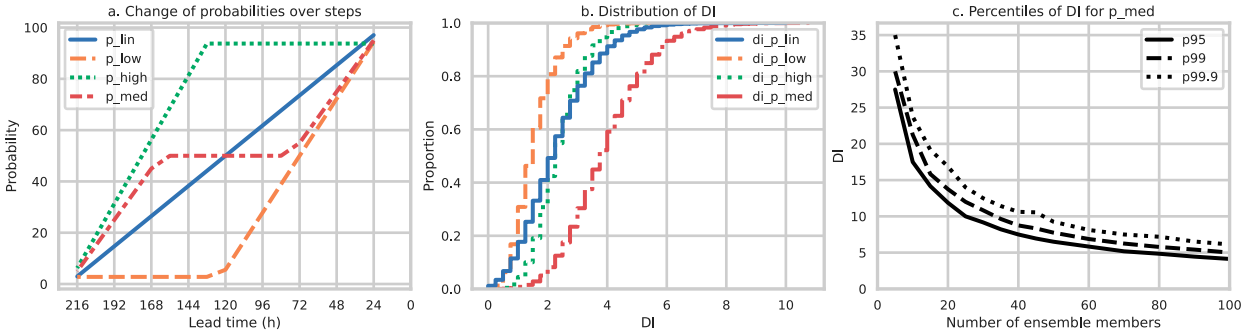

FIG. 4. Effect of ENS size on forecast jumpiness. (a) Four idealized examples of how the probability of TC genesis might evolve over a sequence of seventeen 50-member ENS forecasts initialized, e.g., every 12 h from 216 to 24 h before a given verification time. (b) The empirical cumulative distribution of DI for each of the probability sets shown in (a) based on 10 000 cases. (c) The effect of ENS size (number of members) on the extreme percentiles (95% solid, 99% dashed, and 99.9% dotted) of the DI distribution for the probability set leading to the jumpiest cases (p_med).

account for the jumpiness seen in the predictions for both Vicky and Teddy.

A notable feature of several cases is the high probability for TS activity (FA17) at longer range that is not maintained in the following forecasts made closer to the observed genesis time. Peter, Earl, Philippe, and Bonnie all have high probability (>65%) at some time five or more days ahead but then have much lower probabilities for later forecasts. However, in all these cases, the probability for TC activity (FATC) remains consistently high (well above 65%).

The jumpiest case in this sample is Hurricane Lorenzo. There is a clear flip-flopping in the FA17 probabilities between the forecasts started at 0000 UTC and those started at 1200 UTC: The forecasts from 1200 UTC tend to have lower probability for TS activity than the forecasts from 0000 UTC made 12 h earlier and later. This suggests some systematic difference between the analyses for 0000 and 1200 UTC that affects the forecast intensity. Similar flip-flops, though not as large or long lasting, can be seen in some other cases (e.g., Nigel, Paulette).

These cases illustrate a number of different behaviors in the run-to-run consistency of the forecasts. In the next section, we consider some of the factors that may contribute to these distinctive characteristics.

c. *Factors affecting forecast jumpiness and skill*

In this section, we consider three factors that may affect the forecast jumpiness results discussed in the previous section. We look at the effect of ensemble size and the issue of flip-flops between 0000 and 1200 UTC analysis times and finally consider the early genesis noted in all results and how this model bias may affect the results for both jumpiness and skill. Although a detailed analysis of causes is beyond the scope of the present study, the aim of this initial assessment is to identify avenues for further research.

1) THE EFFECT OF ENSEMBLE SIZE

We compute the forecast probabilities as the proportion of ensemble members that predict TC activity at a given time and location. How much does the finite ensemble size affect the jumpiness in these probabilities? In this section, we use a simple idealized framework to illustrate sampling effects and show the levels of jumpiness that might be expected in an ensemble of 50 members.

Figure 4a shows four idealized examples of how the probability of a TC increases over a set of 17 consecutive forecasts (such as the sequences of forecasts initialized every 12 h from 216 to 24 h before a given observed genesis time, as used in this study). For each set of probabilities, we generate an idealized M-member ensemble by drawing a random sample with the given probability p at each step (Bernoulli process such that each member is either 1, representing forecast of genesis or 0, indicating genesis not forecast) and then compute the DI for this sequence of 17 ensemble forecasts. We repeat this to generate 10 000 cases and summarize the distribution of DI over these 10 000 cases in Fig. 4b.

The four examples represent different predictability: linear increase in probability with forecast lead time (p_lin); a high predictability situation (p_high) in which the genesis event is forecast with high probability from 5 days ahead; a low predictability situation (p_low) where there is no signal at longer range, and medium probability (35%) is reached only around 3–4 days ahead; and finally an intermediate situation (p_med) where the signal for genesis is captured with medium probability more than 7 days ahead, and this level of predictability is maintained until the probability increases again closer to the event.

The expected jumpiness for a 50-member ensemble varies depending on the underlying predictability (Fig. 4b). The low predictability situation is also the least jumpy of the four examples—when the probability of the event is low, there is little variability in the ensemble probability due to sampling (i.e., the finite ensemble size) and the jumpiness (DI) is also low. The intermediate predictability (p_med) situation is the jumpiest, with expected DI substantially higher than for the other examples. In general, the sampling effects due to limited ensemble size are largest for probabilities close to 50%.

We have seen that the jumpiness of the ENS genesis forecasts is higher in the central and eastern Atlantic where the predictability is also higher than in other parts of the basin.

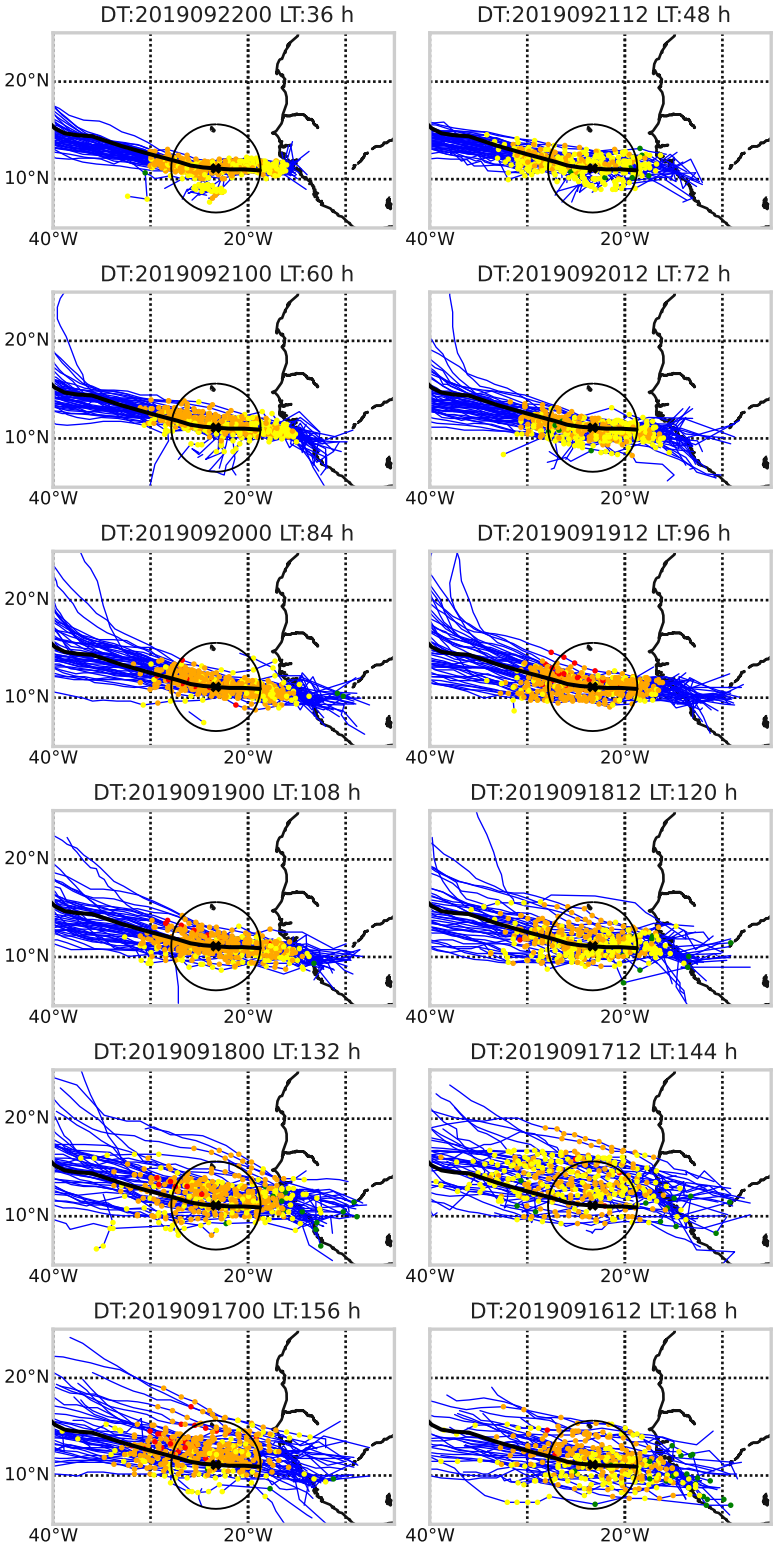

FIG. 5. ENS forecasts for the genesis of Lorenzo at 1200 UTC 23 Sep 2019. ECMWF ENS forecast tracks (blue) and observed track (black). Forecast start dates (DT) from 1200 UTC 16 Sep to 0000 UTC 22 Sep 2019 (LT: forecast lead time in hours to observed genesis time). Colored symbols show forecast intensity (maximum wind speed) at all times within 24 h of the observed genesis time (1200 UTC 21 Sep–1200 UTC 23 Sep); \colors represent the maximum wind speed: yellow (<17 m s^{-1}), orange (17–32 m s^{-1}), and red (>32 m s^{-1}). Observed genesis location at 1200 UTC 23 Sep marked x, and circle indicates locations within 500-km radius of this location.

This is consistent with the above results—the low predictability (p_low) situation is more typical in the west of the basin, while the intermediate (p_med) is more representative of the central and eastern Atlantic. Users should be aware that more predictable situations are likely to be more jumpy because of sampling effects from the finite size of the ensemble.

For all four idealized distributions, the maximum DI is less than 10. In section 4a, we noted that the median DI for the observed genesis events in the eastern Atlantic was 8.75. This is much higher than would be expected from any of the idealized cases considered here. While still high compared to these idealized results, the median DI in the other parts of the Atlantic basin (3.5–4) is closer to the values suggested by these idealized cases.

Figure 4c shows how the ensemble size affects the results for the probability distribution that gives the jumpiest results overall (p_med; Fig. 4b). As noted above, for a 50-member ensemble, the probability of DI $>$ 10 is extremely small. However, for a 20-member ensemble, the chance of having DI $>$ 10 is not negligible: We should expect that more than 5% of cases will have DI $>$ 10. In general sampling, uncertainties will be larger for smaller ensembles (the proportion of members predicting genesis will be a less reliable estimate of the true underlying probability) and, therefore, the jumpiness from run to run will increase and more cases should be expected with large DI. Conversely, there is a steady decrease in the chances of high jumpiness as the ensemble size increases from 20 to 100 members: For a 100-member ensemble, the maximum DI is not likely to be above 5.

Overall, these idealized results suggest that for the ENS and the set of observed cases considered here, values of DI greater than 10 are unlikely to be due purely to ensemble size. The high median value of DI (8.75) for the cases in the eastern Atlantic suggests there are a substantial number of cases where factors other than pure sampling contribute to the jumpiness.

However, it should be noted that if the ensemble is underdispersive, the effective ensemble size could be lower than the nominal 50 members and this could significantly affect the DI. These idealized results also show that increasing ensemble size would be expected to reduce overall jumpiness and improve the overall consistency of the ENS predictions. This may be important for some decision-making applications (Jewson et al. 2022) such as deciding when to plan and initiate evacuation from areas at potential risk (Regnier and Harr 2006) or rerouting of transportation to avoid adverse weather (McLay 2008).

2) Analysis impacts—Flip-flop between 0000 and 1200 UTC initial conditions

The case of Lorenzo demonstrated a marked jumpiness between the forecasts initialized at 0000 UTC and at 1200 UTC. Figure 5 shows the forecast tracks for Lorenzo initialized from 36 to 168 h before the observed genesis time. The circle indicates locations within 500 km of the observed genesis location. The potential for TS activity is predicted at all lead times, and the earliest forecast with high probability was initialized 7 days before the observed genesis time (Fig. 3). Most of the forecast TCs intensify to TS strength very soon after the track leaves land and moves over the sea off the African coast. This is generally earlier than the observed genesis, consistent with the low probabilities shown in the FG17 curve in Fig. 3. A notable feature of the forecast probabilities (both FA17 and FA15) is the long sequence of flip-flops in the probabilities between successive forecasts: The forecasts started from 0000 UTC have higher probability than those started 12 h earlier and 12 h later at 1200 UTC.

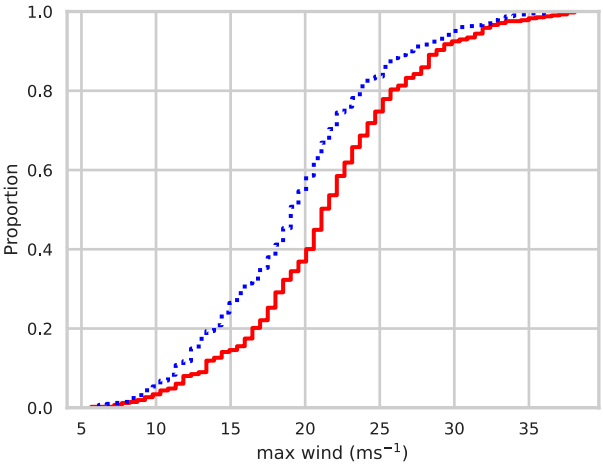

FIG. 6. Sensitivity of TC intensity to analysis time in ENS forecasts for the genesis of Lorenzo at 1200 UTC 23 Sep 2019. Empirical cumulative distribution functions of maximum wind speed for ENS TC forecasts initialized at 0000 UTC (solid red line) and at 1200 UTC (dotted blue line) that are within 500 km and 24 h of the observed genesis event of Lorenzo (at 11.1°N, 23.3°W). All forecasts start between 0000 UTC 14 Sep and 1200 UTC 22 Sep.

We extracted the maximum wind for each forecast TC position within 500 km and 24 h of the observed genesis position and time of Lorenzo for all ENS forecasts started from 0000 UTC and compared the distribution of these winds with those from the forecasts started at 1200 UTC. There is a statistically significant shift toward stronger winds in the forecasts from 0000 UTC analysis times (Fig. 6). This suggests that there is some systematic difference in the assimilation at 0000 and 1200 UTC that affects the intensification of the forecasts in this case. One possibility is the analysis over West Africa where a systematic difference in analysis increments has been identified in the ECMWF assimilation system (Bormann et al. 2023). The reasons for this are not yet understood and are the subject of further investigation.

While some other cases in the same region also have some flip-flops between 0000 and 1200 UTC initial conditions, this is not a common occurrence. Therefore, while assimilation differences may be one factor, it is likely that a combination of factors may be involved to make the large and significant impact found in this Lorenzo case. Further evaluation of this case is beyond the scope of this paper, but the results suggest that additional investigation into the differences between 0000 and 1200 UTC analyses may be relevant.

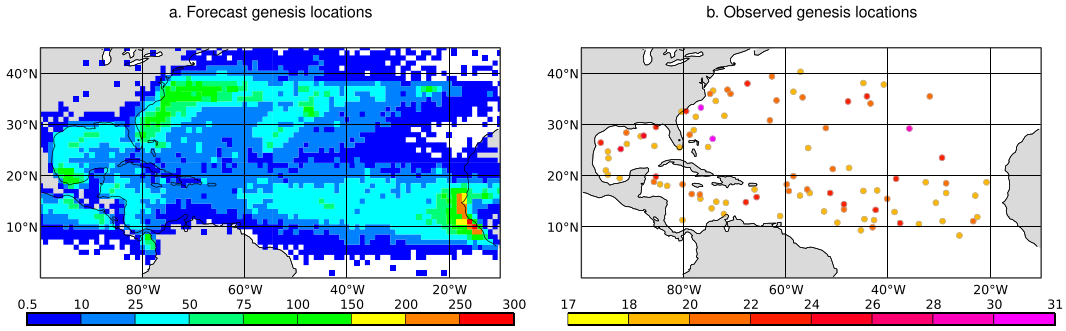

FIG. 7. Locations of TS genesis in forecasts and observations. (a) Forecast genesis: location of the first point on each forecast track with maximum wind speed > 17 m s^{-1}; map shows the total number of forecast genesis events in each $1° \times 1°$ grid box over the full set of forecasts during May–December 2019–23. (b) Observed TS genesis locations for all 97 observed cases; color indicates the reported maximum wind at genesis time in the IBTrACS data (m s^{-1}).

3) MODEL BIAS (SYSTEMATIC ERROR)

In many cases that develop from tropical waves over Africa, the forecast tracks intensified to TS strength before the observed TS genesis time. The example of Lorenzo above shows that the forecast tracks often intensified to TS strength immediately after leaving the African continent and moving over sea.

To investigate how typical this early intensification is, we consider all forecast tracks in the 5-yr sample. Figure 7a shows the location of the first time each forecast track reaches TS strength, accumulated on a $1° \times 1°$ grid. Figure 7b shows the observed locations for the equivalent first time that the observed TC is reported as TS. There is a substantial peak in the number of forecast TCs that intensify to TS strength immediately after leaving the African coast. In contrast, none of the observed cases are reported to reach TS intensity east of 20°W. There are fewer forecast TS genesis events in the central and western areas (60°–80°W, 10°–20°N). Overall, there is a shift eastward of the genesis locations in the forecasts. A similar bias in overforecasting TC genesis was found in the NCEP GEFS reforecasts, associated with overactivity of African easterly waves in that system (Li et al. 2016; Wang et al. 2018).

Overdevelopment of initial wave activity over Africa and the quick intensification to TS soon after the waves move over the open sea may also account for some of the high DI cases shown in Fig. 3. Peter and Philippe were two cases predicted with high probability at longer lead times, but for both, the probability for TS intensity dropped at shorter leads. In each case, the higher probabilities occurred for forecasts initialized when the wave activity was still over the African continent, and TS genesis occurred soon after the system left the coast. In the later forecasts where the forecast TC developed further to the west, the probabilities for more intense developments (both FA17 and FA15) were lower.

In summary, there is a tendency in the ENS for TC development to occur too quickly in TCs that develop from African easterly waves and for the intensification to TS to occur soon after the wave moves over the ocean, often before the TC reaches 20°W. This may be a cause of the jumpy behavior seen in some cases.

We hypothesize that this bias is associated with overdevelopment of African easterly wave activity in the ENS and identify this as an important area for future research.

d. Overall skill of TS genesis forecasts

So far, we have focused on the results for observed TS genesis events. Although these results show the performance for hits and misses of observed events, they do not take account of false alarms in the forecasts.

To assess the overall performance of the ENS genesis probability forecasts, we now include all forecast tracks, including those false alarm cases where a TS did not actually occur. For each case, and at each grid point, the forecast is the probability that a TS genesis event will occur within 500 km and between 24 and 216 h ahead.

Figure 8 shows the Brier skill score [B, Eq. (4)] of these ENS forecasts of TS genesis. This shows that there is skill in some areas. The highest skill is in the eastern Atlantic, consistent with the regions where genesis was found to be more predictable at longer lead for the observed cases (Fig. 2). Although Brier skill score (BSS) is lower in more western areas, there are still some regions with positive skill. The low overall skill is consistent with the findings in the earlier sections that FG17 skill is limited because of the tendency in the ENS to predict TS genesis earlier than observed.

Figure 9a shows the reliability diagram for the TS genesis forecasts; the ENS probabilities are grouped into 10% probability intervals and accumulated over all grid points and over the full 5-yr sample. The curve is below the diagonal, indicating that the genesis forecasts are overconfident and lack reliability. While this can be a result of lack of spread in the ensemble, it is also consistent with our results that the ENS tends to predict TS genesis earlier than observed. A similar overconfidence is also found in the operational ECMWF verification of TC activity (Haiden et al. 2023) and in corresponding TC activity forecasts from other ensemble systems (Magnusson et al. 2021).

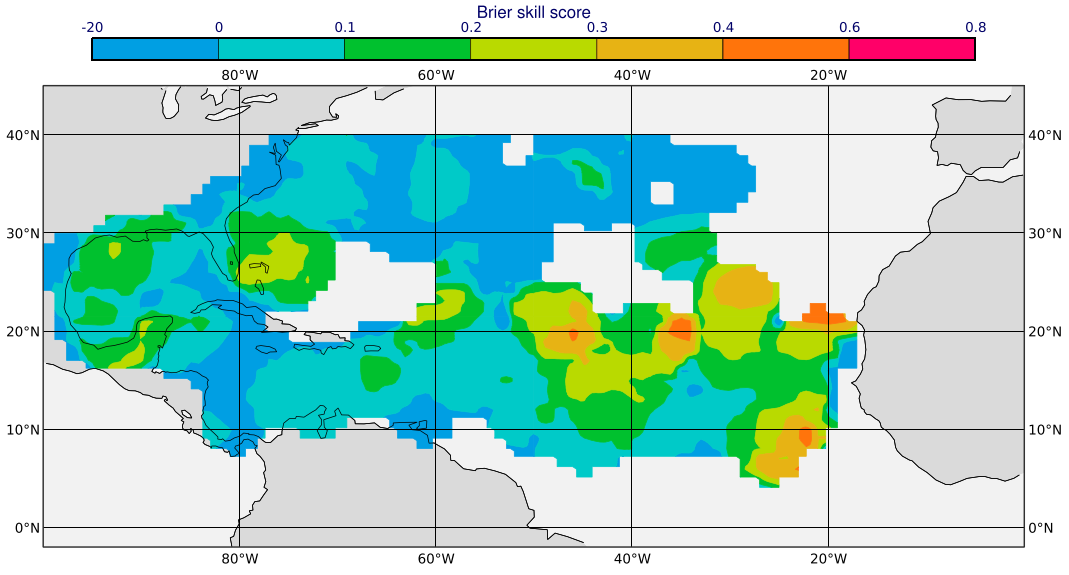

FIG. 8. Skill of ENS forecasts of TS genesis. BSS for the forecast probability that TS genesis will occur within 500 km of each grid point during the forecast, between 24- and 216-h lead time; score computed over all forecasts in a 5-yr sample (2019–23).

The positive slope of the reliability curve shows that, while lacking reliability, the forecasts do have some resolution: the ability to distinguish between more and less likely genesis events. This discrimination ability is confirmed in Fig. 9b which shows the ROC diagram for the genesis forecasts. In the ROC computation, all possible forecast probabilities are considered (Ben Bouallègue and Richardson 2022). In Fig. 9b, the ROC for all grid points is compared with the corresponding ROC curves for three subregions: The skill is greater in the eastern Atlantic (east of 60°W and south of 30°N) and lower in the western (west of 60°W and south of 30°N) and northern (north of 30°N) areas. This confirms the regional differences in skill noted in the evaluation of the observed cases (Fig. 2). Although the reliability diagrams for the subareas are more noisy due to the smaller sample size in each subarea, they also indicate better performance for the eastern region and lowest reliability in the northern region.

To highlight the false alarms as a proportion of the genesis forecasts, the skill of the genesis forecasts for the low, medium, and high probability thresholds in the eastern and western regions is shown on a performance diagram in Fig. 9c. As for the reliability diagram and ROC, Fig. 9c shows a substantial difference in performance between eastern and western areas, especially for the low and medium probabilities, with substantially better hit rate for a similar false alarm ratio. As for the other performance measures, the northern region has the poorest performance (not shown).

Figure 9d shows the ROC curves for the FG17 forecasts for days 3, 5, and 7 (72, 120, and 168 h in gray) together with the overall ROC (same as in Fig. 9b). The discrimination skill decreases at longer lead, although there is still substantial discrimination ability at 168 h. The overall ROC (for genesis between 24 and 216 h) lies between the curves for 120 and 168 h, suggesting the overall results are reasonably indicative of the medium-range performance.

The results in this section have been based on the comparison of the forecast and observed genesis of tropical storms, defined as the first point on forecast or observed track with wind speed of 17 m s^{-1}. To investigate the sensitivity of the results to the forecast wind speed threshold, we recomputed the ROC results using alternative forecast wind speed thresholds of 8, 15, and 19 m s^{-1}, all verified against the operational genesis of TS (17 m s^{-1}). We found that the results are relatively insensitive to small changes ($\pm$2 m s^{-1}) in the forecast wind speed threshold, but a large reduction in the forecast threshold (to 8 m s^{-1}) substantially reduces the forecast skill. This section has focused on whether TS genesis will occur at some point during the forecast, and this may be why these results are not too sensitive to the wind threshold—a given threshold will likely be exceeded as the tropical cyclone intensifies during the forecast. A more detailed investigation of the definition of genesis in the forecast and the effect on forecast skill will be a topic for future research.

5. Conclusions

We have investigated the ability of the ECMWF ensemble forecasts ENS to predict the genesis of tropical cyclones in the Atlantic basin up to 10 days ahead. We compared the ENS operational TC track forecasts to observed tracks from the IBTrACS archive for all named tropical storms for the 5 years 2019–23. We focused on the probabilistic performance of the ENS rather than the evaluation of deterministic forecasts that has been more typically the subject of previous studies.

Defining a genesis event as the first time the TC reached tropical storm strength (winds at least 17 m s^{-1}), the ENS

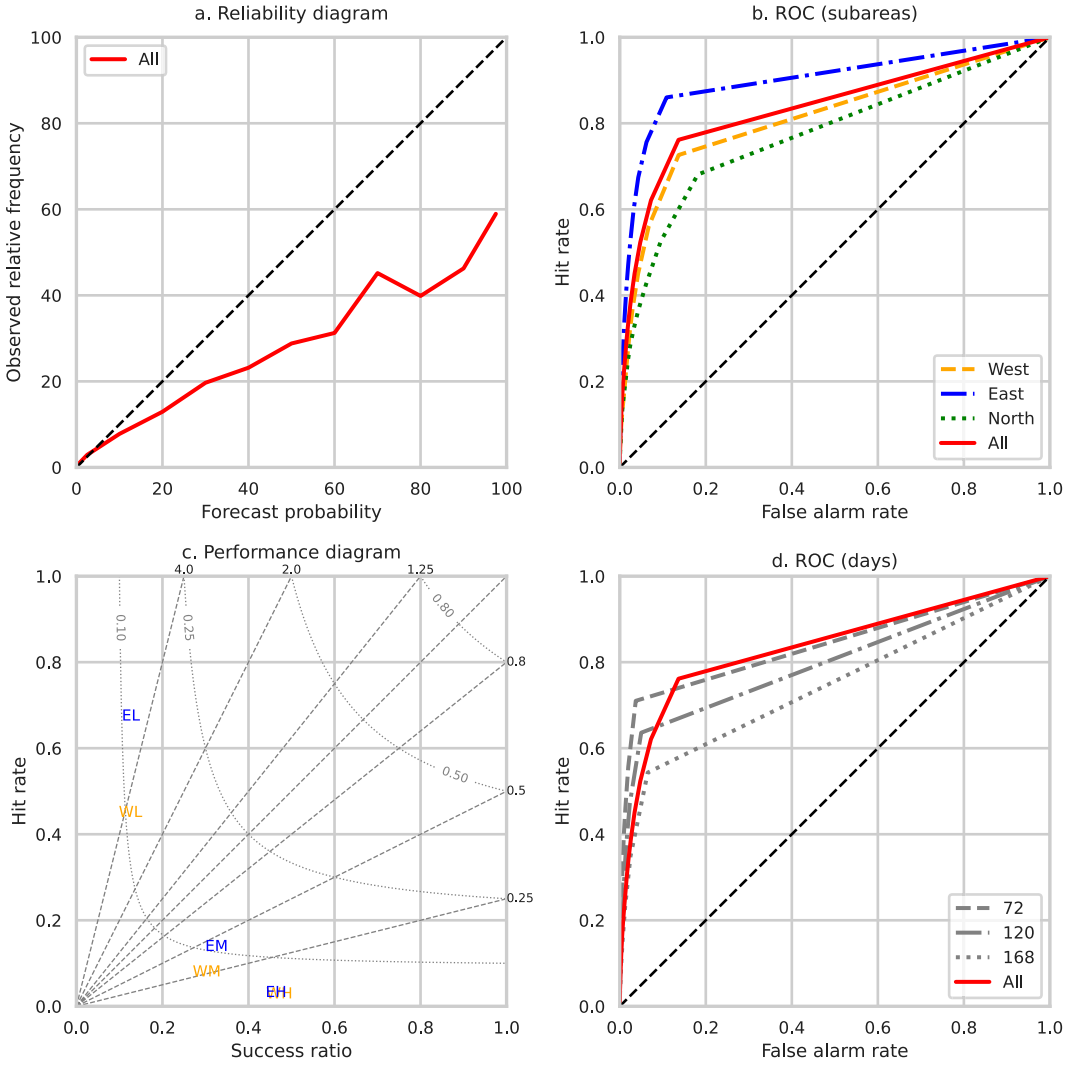

FIG. 9. Evaluation of ENS forecasts of TS genesis to occur between 24- and 216-h lead time; scores computed over all forecasts in a 5-yr sample (2019–23). (a) Reliability diagram, results accumulated over all grid points; (b) ROC diagram for all grid points (solid red) and for western (orange dashed), eastern (blue dash–dotted), and northern (dotted green) subregions (see text for details); (c) performance diagram for eastern (E) and western (W) regions and for the low (L), medium (M), and high (H) probability thresholds (first letter indicates region and second letter indicates the probability threshold), gray diagonal lines show bias, and gray curved lines show threat score; (d) ROC diagram comparing overall results [all, solid red, same as in (b)] with FG17 forecasts of TS genesis at lead times of 72, 120, and 168 h.

probability forecasts (FG17; Table 1) of the observed genesis events had relatively low skill with only 20% of the observed cases predicted with medium or high probability (probability 35% or more) more than 72 h ahead. In many cases, the forecast track reached TS strength more than 24 h before the observed TS genesis time. Allowing for this early genesis in the forecasts increased the forecast probabilities (FA17; Table 1) for the observed event.

In part, this may reflect differences between the IBTrACS reports and the ECMWF TC tracker—the ECMWF tracker tends to pick up the TC at an earlier stage than the official designation as a TC. Differences in feature identification between different TC trackers can have a significant impact on the number of TCs identified by a forecast model (Conroy et al. 2023), and there is currently no generally agreed best practice for the definition and evaluation of TC genesis (Dunion et al. 2023).

We also found substantial geographical variation in the performance of the ENS probabilities: Observed genesis events were predicted 2–3 days earlier in the central and eastern Atlantic than in other parts of the basin. The regional differences may be associated with intrinsic differences in predictability in different tropical cyclogenesis pathways (McTaggart-Cowan et al. 2013, 2008; Wang et al. 2018). Investigation of the ENS skill and jumpiness in the different pathways is an area of future research.

We assessed the run-to-run consistency of the ENS probabilities of genesis using the divergence index (DI) (Richardson et al. 2020, 2024). The DI also varied between different regions, with the jumpiest cases being in the central and eastern Atlantic. The median DI here was more than twice that was found in the western and northern parts of the basin. The most jumpy cases occurred in different years but almost always in late August or September. In most of these cases, the jumpiness depended on the forecast intensity: The forecasts were consistent in predicting the existence of the TC, but the probability for the TC to be at tropical storm strength varied from run to run.

Understanding the causes of jumpiness is important to inform both users and model developers. Forecast jumpiness is a measure of the internal consistency of the forecasting system. Although we used the observed genesis events as reference, the computation of DI does not depend on the observations. Hence, the results for jumpiness are not directly affected by the differences between the model and observed definitions of genesis discussed above. Examining the issues affecting jumpiness can therefore help to identify potential weakness in the modeling system. Based on consideration of the most jumpy cases in our sample, we considered a number of factors that could affect the ENS jumpiness in predicting TC genesis.

One possible cause of large jumpiness is the sampling uncertainty associated with the limited ensemble size. We found that the DI for the most jumpy cases is significantly higher than should be expected for a well-constructed 50-member ensemble. However, jumpiness is sensitive to ensemble size, and the highest values of DI found in our results may occur for ensembles with around 20 members. While ENS track forecasts are well calibrated, the forecast intensity is overall underdispersive (Haiden et al. 2023), and in some situations, this may reduce the effective ensemble size, contributing to increased jumpiness. In certain situations with intrinsically low predictability, there may be particular sensitivity to ensemble size and substantially more than 50 members may be needed to properly represent the underlying distribution (Leutbecher 2019; Craig et al. 2022; Kondo and Miyoshi 2019). This may be important in some genesis situations involving complex interactions between waves, where the ENS showed large jumpiness.

In some cases, there was a notable sequence of flip-flops between the forecasts started from 0000 to 1200 UTC analyses. Lorenzo was a particularly strong example, and for this case, we found a significant difference between the forecast maximum winds associated with the TCs initialized at the two analysis times, with higher winds from the 0000 UTC analysis. We hypothesize that this may be associated with a known systematic difference in analysis increments at 0000 and 1200 UTC over West Africa in the ECMWF assimilation system (Bormann et al. 2023). However, this flip-flop behavior was not a common feature across cases, suggesting that a combination of factors in addition to the analysis differences may be involved to make the large and significant impact found in this case. This is an area requiring further investigation.

A significant difference between the observed and forecast TS genesis is that the ENS TC tracks tend to intensify to TS strength earlier than the observed TS genesis event. ENS tracks that develop from African easterly waves often reach TS soon after the wave moves over the ocean, often before the TC reaches 20°W. This may be a cause of the jumpy behavior seen in some cases (e.g., Peter and Philippe) where earlier forecasts had high probability for TS development, while later forecasts that were initialized after the disturbance moved over the ocean had lower probability. The association with jumpy behavior lends weight to this being a systematic error in the forecasting system and not just an artifact of the differences between forecast and observed genesis identification methods. We hypothesize that this bias is associated with overdevelopment of African easterly wave activity in the ENS and identify this as an important area for future research.

Finally, we provided a baseline evaluation of the skill of the ENS TS genesis forecasts including all forecasts from the 5-yr sample to take account of both hits and false alarms. Overall, forecasts were overconfident but showed good discrimination ability, with higher skill in the east of the basin (particularly for low to medium probabilities) consistent with the results for the observed genesis cases. The ECMWF forecasting system is typically upgraded annually, and some of these changes affect the tropical cyclone performance, for example, the increase in ensemble resolution in 2023 (Haiden et al. 2023). Given that TS genesis is a relatively rare event, skill evaluation generally needs to be carried out over a sample of several seasons, inevitably covering a number of different model versions (Leonardo and Colle 2021). We found cases of large jumpiness in each year of our sample, and this suggests that the underlying causes still need to be addressed. The overall results can be seen as a general assessment of recent model performance and provide a benchmark against which to evaluate future model developments.

Acknowledgments. This work is based on TIGGE data. TIGGE (The International Grand Global Ensemble) is an initiative of the World Weather Research Programme (WWRP). David Richardson is supported by a Wilkie Calvert PhD Studentship at the University of Reading. Sharanya J. Majumdar gratefully acknowledges support from National Science Foundation Grant AGS-1747781 and the University of Miami and ECMWF for jointly supporting a sabbatical year at ECMWF. We thank three anonymous referees 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

Ben Bouallègue, Z., and D. S. Richardson, 2022: On the ROC area of ensemble forecasts for rare events. *Wea. Forecasting*, **37**, 787–796, https://doi.org/10.1175/WAF-D-21-0195.1.

Bormann, N., L. Magnusson, D. Duncan, and M. Dahoui, 2023: Characterisation and correction of orbital biases in AMSU-A and ATMS observations in the ECMWF system. ECMWF Tech. Memo. 912, 27 pp., https://doi.org/10.21957/d281dc221a.

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.

Cangialosi, J. P., E. Blake, M. DeMaria, A. Penny, A. Latto, E. Rappaport, and V. Tallapragada, 2020: Recent progress in tropical cyclone intensity forecasting at the National Hurricane Center. *Wea. Forecasting*, **35**, 1913–1922, https://doi.org/10.1175/WAF-D-20-0059.1.

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.

Davis, C. A., and L. F. Bosart, 2003: Baroclinically induced tropical cyclogenesis. *Mon. Wea. Rev.*, **131**, 2730–2747, https://doi.org/10.1175/1520-0493(2003)131<2730:BITC>2.0.CO;2.

——, and ——, 2004: The TT problem: Forecasting the tropical transition of cyclones. *Bull. Amer. Meteor. Soc.*, **85**, 1657–1662, https://doi.org/10.1175/BAMS-85-11-1657.

Dunion, J. P., and Coauthors, 2023: Recommendations for improved tropical cyclone formation and position probabilistic Forecast products. *Trop. Cyclone Res. Rev.*, **12**, 241–258, https://doi.org/10.1016/j.tcrr.2023.11.003.

Elsderry, 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.

Haiden, T., M. Janousek, F. Vitart, Z. Ben-Bouallegue, and F. Prates, 2023: Evaluation of ECMWF forecasts, including the 2023 upgrade. ECMWF Tech. Memo. 911, 60 pp., https://doi.org/10.21957/d47ba5263c.

Halperin, D. J., H. E. Fuelberg, R. E. Hart, J. H. Cossuth, P. Sura, and R. J. Pasch, 2013: An evaluation of tropical cyclone genesis forecasts from global numerical models. *Wea. Forecasting*, **28**, 1423–1445, https://doi.org/10.1175/WAF-D-13-00008.1.

——, ——, ——, and ——, 2016: Verification of tropical cyclone genesis forecasts from global numerical models: Comparisons between the North Atlantic and eastern North Pacific basins. *Wea. Forecasting*, **31**, 947–955, https://doi.org/10.1175/WAF-D-15-0157.1.

——, R. E. Hart, H. E. Fuelberg, and J. H. Cossuth, 2017: The development and evaluation of a statistical–dynamical tropical cyclone genesis guidance tool. *Wea. Forecasting*, **32**, 27–46, https://doi.org/10.1175/WAF-D-16-0072.1.

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

Hon, K. K., and Coauthors, 2023: Recent advances in operational tropical cyclone genesis forecast. *Trop. Cyclone Res. Rev.*, **12**, 323–340, https://doi.org/10.1016/j.tcrr.2023.12.001.

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.

Klotzbach, P. J., and Coauthors, 2022: A hyperactive end to the Atlantic hurricane season October–November 2020. *Bull. Amer. Meteor. Soc.*, **103**, E110–E128, https://doi.org/10.1175/BAMS-D-20-0312.1.

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 data. *Bull. Amer. Meteor. Soc.*, **91**, 363–376, https://doi.org/10.1175/2009BAMS2755.1.

——, H. J. Diamond, M. C. Kossin, and C. J. Schreck, 2018: International Best Track Archive for Climate Stewardship (IBTrACS) project, version 4.01. NOAA National Centers for Environmental Information, accessed 2 February 2024, https://doi.org/10.25921/82ty-9e16.

Komaromi, W. A., and S. J. Majumdar, 2014: Ensemble-based error and predictability metrics associated with tropical cyclogenesis. Part I: Basinwide perspective. *Mon. Wea. Rev.*, **142**, 2879–2898, https://doi.org/10.1175/MWR-D-13-00370.1.

——, and ——, 2015: Ensemble-based error and predictability metrics associated with tropical cyclogenesis. Part II: Wave-relative framework. *Mon Wea. Rev.*, **143**, 1665–1686, https://doi.org/10.1175/MWR-D-14-00286.1.

Kondo, K., and T. Miyoshi, 2019: Non-Gaussian statistics in global atmospheric dynamics: A study with a 10 240-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.

Landsea, C. W., and J. L. Franklin, 2013: Atlantic hurricane database uncertainty and presentation of a new database format. *Mon. Wea. Rev.*, **141**, 3576–3592, https://doi.org/10.1175/MWR-D-12-00254.1.

——, 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, 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**, 107–128, https://doi.org/10.1002/qj.3387.

Li, W., Z. Wang, and M. S. Peng, 2016: Evaluating tropical cyclone forecasts from the NCEP Global Ensemble Forecasting System (GEFS) reforecast version 2. *Wea. Forecasting*, **31**, 895–916, https://doi.org/10.1175/WAF-D-15-0176.1.

Liang, M., J. C. L. Chan, J. Xu, and M. Yamaguchi, 2021: Numerical prediction of tropical cyclogenesis part I: Evaluation of model performance. *Quart. J. Roy. Meteor. Soc.*, **147**, 1626–1641, https://doi.org/10.1002/qj.3987.

Magnusson, L., and Coauthors, 2021: Tropical cyclone activities at ECMWF. ECMWF Tech. Memo. 888, 140 pp., www.ecmwf.int/en/elibrary/81277-tropical-cyclone-activities-ecmwf.

Majumdar, S. J., and R. D. Torn, 2014: Probabilistic verification of global and mesoscale ensemble forecasts of tropical cyclogenesis. *Wea. Forecasting*, **29**, 1181–1198, https://doi.org/10.1175/WAF-D-14-00028.1.

Mason, I., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.*, **30**, 291–303.

McLay, J. G., 2008: Markov chain modeling of sequences of lagged NWP ensemble probability forecasts: An exploration of model properties and decision support applications. *Mon. Wea. Rev.*, **136**, 3655–3670, https://doi.org/10.1175/2008MWR2376.1.

McTaggart-Cowan, R., G. D. Deane, L. F. Bosart, C. A. Davis, and T. J. Galarneau Jr., 2008: Climatology of tropical cyclogenesis in the North Atlantic (1948–2004). *Mon. Wea. Rev.*, **136**, 1284–1304, https://doi.org/10.1175/2007MWR2245.1.

——, T. J. Galarneau Jr., L. F. Bosart, R. W. Moore, and O. Martius, 2013: A global climatology of baroclinically influenced tropical cyclogenesis. *Mon. Wea. Rev.*, **141**, 1963–1989, https://doi.org/10.1175/MWR-D-12-00186.1.

Papin, P. P., L. F. Bosart, and R. D. Torn, 2017: A climatology of Central American gyres. *Mon. Wea. Rev.*, **145**, 1983–2000, https://doi.org/10.1175/MWR-D-16-0411.1.

Pappenberger, F., H. L. Cloke, A. Persson, and D. Demeritt, 2011: HESS opinions "on forecast (in)consistency in a hydrometeorological 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., 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.

——, ——, J. A. Methven, and F. Pappenberger, 2024: Jumpiness in ensemble forecasts of Atlantic tropical cyclone tracks. *Wea. Forecasting*, **39**, 203–215, https://doi.org/10.1175/WAF-D-23-0113.1.

Roebber, P. J., 2009: Visualizing multiple measures of forecast quality. *Wea. Forecasting*, **24**, 601–608, https://doi.org/10.1175/2008WAF2222159.1.

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.

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.

Torn, R. D., and C. Snyder, 2012: Uncertainty of tropical cyclone best-track information. *Wea. Forecasting*, **27**, 715–729, https://doi.org/10.1175/WAF-D-11-00085.1.

Wang, Z., W. Li, M. S. Peng, X. Jiang, R. McTaggart-Cowan, and C. A. Davis, 2018: Predictive skill and predictability of North Atlantic tropical cyclogenesis in different synoptic flow regimes. *J. Atmos. Sci.*, **75**, 361–378, https://doi.org/10.1175/JAS-D-17-0094.1.

Wilks, D. S., 2020: *Statistical Methods in the Atmospheric Sciences.* 4th ed. Elsevier, 840 pp.

Yamaguchi, M., and N. Koide, 2017: Tropical cyclone genesis guidance using the early stage Dvorak analysis and global ensembles. *Wea. Forecasting*, **32**, 2133–2141, https://doi.org/10.1175/WAF-D-17-0056.1.

——, F. Vitart, S. T. K. Lang, L. Magnusson, R. L. Elsberry, G. Elliott, M. Kyouda, and T. Nakazawa, 2015: Global distribution of the skill of tropical cyclone activity forecasts on short- to medium-range time scales. *Wea. Forecasting*, **30**, 1695–1709, https://doi.org/10.1175/WAF-D-14-00136.1.

Zhang, X., J. Fang, and Z. Yu, 2023: The forecast skill of tropical cyclone genesis in two global ensembles. *Wea. Forecasting*, **38**, 83–97, https://doi.org/10.1175/WAF-D-22-0145.1.