HOW SKILLFUL ARE THE MULTIANNUAL FORECASTS OF ATLANTIC HURRICANE ACTIVITY?

Louis-Philippe Caron, Leon Hermanson, Alison Dobbin, Jara Imbers, Llorenç Lledó, and Gabriel A. Vecchi

Multiannual Atlantic hurricane forecasts are one area that can benefit from the recent development of initialized climate predictions.

Despite high-profile storms like Sandy and Matthew, the 2012–2016 period has been relatively quiet in terms of hurricane activity over the northern Atlantic compared to the previous two decades. Because the available observational record of Atlantic hurricane shows this basin alternating between decadelong periods of high and low activity since the end of the nineteenth century (Vecchi and Knutson 2011; Chenoweth and Divine 2014), this has led to some speculation as to whether we have entered into a new prolonged period of low hurricane activity similar to what was observed from the early 1970s to the mid-1990s (Klotzbach et al. 2015). This question is of great interest, not only for the academic community, but also for other sectors, such as policy-makers, nongovernmental organizations (i.e., disaster relief agencies), and the insurance industry. For example, in the case of the property and casualty (PC) insurance industry that typically underwrites annual contracts and may have several automatic renewals, the quantification of hurricane risk on short to medium time scales is of big economic relevance.

In theory, multiannual forecast systems could be used to give the odds of basinwide hurricane activity remaining low for the foreseeable future. However, in comparison to seasonal hurricane forecasts, which originated in the mid-1980s (Gray 1984), the field of multiannual forecasting is very much in its infancy. Until recently, this type of long-term forecast was exclusively produced using statistical models, wherein hurricane activity is first derived by forecasting a subset of the climate conditions deemed to control hurricane activity (e.g., sea surface temperature over certain key regions) and then combining that forecast with a statistical model linking past climate conditions and past hurricane activity to produce a prediction of upcoming hurricane activity (Jewson et al. 2009). Statistical approaches of varying complexity have been adopted by the risk modeling industry

(Bonazzi et al. 2014) because, up until very recently, no viable alternatives existed.

The advent of climate prediction (also referred to as decadal forecasting) (Doblas-Reyes et al. 2013; Meehl et al. 2014), wherein climate models are initialized with the contemporaneous states of the atmosphere, ocean, and sea ice, has allowed the development of similar multiannual forecasts based on climate model simulations. These climate simulations can be used either to replace the first step of a statistical forecast (Vecchi et al. 2013; Caron et al. 2014, 2015) (so-called hybrid forecasts) or to do without empirical models altogether (Smith et al. 2010; Hermanson et al. 2014) (so-called dynamical forecasts). The latter technique involves directly tracking tropical cyclone–like disturbances in climate output using an automated detection and tracking algorithm. These dynamical forecasts are the most demanding in terms of resources because they require an infrastructure in place to detect and track the storms (Ullrich and Zarzycki 2017) as well as high-frequency data, which, in a decadal forecasting context, can be prohibitive in terms of the amount of data storage required for such analysis. These restrictions also limit the possibilities for multimodel ensemble analyses.

By combining aspects of both dynamical and statistical forecasts, the hybrid forecast offers a compromise between the first two approaches. In such forecasts, the large-scale conditions expected to modulate hurricane activity are derived from climate model simulations, and hurricane activity is inferred using a statistical relationship between these large-scale fields and past hurricane activity. Although hurricane activity is implicit in this case, hybrid forecasts have the advantage of relying on large-scale features of the atmosphere–ocean system (usually large areas of sea surface temperature), which the climate models can be expected to be better at simulating and forecasting than smaller-scale features, such as hurricanes. Furthermore, such forecasts are usually computed using seasonal or yearly means, thus greatly reducing the amount of the data required and, incidentally, making desirable multimodel analyses more affordable. Both the dynamical and the hybrid approaches are used regularly in the seasonal forecasting and climate communities in order to derive hurricane statistics from climate model simulations (Vecchi et al. 2011; Vitart et al. 2007; Camargo et al. 2007).

Two hybrid techniques have so far been investigated to forecast hurricane activity at the multiannual time scale. The first of these techniques relies on predicting the weighted difference in sea surface temperature (SST) of the tropical Atlantic with respect to that of the wider tropics (Vecchi et al. 2013; Caron et al. 2014). In this case, a relatively warm (cold) Atlantic with respect to the rest of the tropics will lead to higher (lower) hurricane activity due to more (less) conducive dynamic and thermodynamic conditions over the Atlantic. The second technique relies on forecasting a proxy index for the Atlantic multidecadal oscillation (AMO) (Klotzbach and Gray 2008; Caron et al. 2015), a slow oscillation in Atlantic SST that is thought to modulate hurricane activity at long time scales (Zhang and Delworth 2006; Knight et al. 2006; Goldenberg et al. 2001). A positive index is usually associated with increased hurricane activity.

Here, we present and compare the different approaches (statistical, hybrid, dynamical) currently available to provide multiyear forecasts of North Atlantic hurricane activity, starting with a short description of the different forecast systems. These systems are also summarized in Table 1.

FORECASTING SYSTEMS. *Climate model data.* All climate simulations used here are initialized using contemporaneous observations, thus aligning the simulated natural variability with the observed variability. External forcing (greenhouse gases, solar activity, stratospheric aerosols associated with volcanic eruptions and anthropogenic aerosols) are taken from observations for start dates ranging from 1961 (first forecast period: 1961–65) to 2005 and the representative concentration pathway (RCP) 4.5 scenario (Meinshausen et al. 2011) from 2006 to 2014 (last forecast period: 2010–14). Systematic climate model drift in these simulations is addressed by computing a lead-time-dependent climatology for each individual model by first averaging the predicted variable for all of its members across the start-date dimension and then subtracting that climatology from each hindcast to obtain the anomalies over the whole predicted period (García-Serrano and Doblas-Reyes 2012).

Observational data. The hurricane time series used as reference is derived from the revised National Hurricane Center "best track" hurricane database (HURDAT2) (Landsea and Franklin 2013) and includes only hurricanes forming between 5° and 25°N during the period 1961–2014 and which survived at least 48 h at tropical storm strength (or above). The geographical limitation is introduced in order to allow for comparison with the dynamical forecasts, which limits tracking to that region.

TABLE 1. Short descriptions of the techniques used to derive hurricane numbers from the climate simulations.

Technique	Type	Summary
Cyclone tracking	Dynamical	Hurricane numbers are obtained directly by tracking local minima in surface pressure over the tropical Atlantic.
Relative SST	Hybrid	Hurricane numbers are estimated through a statistical model with two predictors: mean Jun–Nov SST over 1) the tropical Atlantic and 2) the entire tropics. The parameters are derived from the sensitivity of a high-resolution atmospheric GCM to a number of SST perturbations.
AMO index	Hybrid	Hurricane activity is estimated using a climate index correlated with low-frequency North Atlantic hurricane variability. Positive (negative) index values are associated with conditions more (less) conducive to hurricane formation.
Statistical combination	Statistical	Basin hurricane numbers are estimated by combining six statistical models that either use 1) regression relationships between mean Jun and Nov SST over the MDR or between the MDR and tropical Pacific region or 2) averages of historical hurricanes counts in the basin in active or inactive periods.

Dynamical forecast systems. With this technique, long-lived local minima in daily mean sea level pressure are tracked over the tropical North Atlantic (5°–25°N) during June–November in initialized climate simulations performed with three different versions of the Met Office decadal prediction systems:

- 20 simulations based on the Hadley Centre Coupled Model, version 3 (HadCM3) (Smith et al. 2014) as submitted to phase 5 of the Coupled Model Intercomparison Project (CMIP5);
- 9 simulations also based on HadCM3 but using nine variants obtained by perturbing poorly constrained atmospheric and surface parameters in order to sample modeling uncertainty (Smith et al. 2010); and
- 4 simulations based on the Hadley Centre Global Environment Model, version 3 (HadGEM3) (Knight et al. 2014).

Initial conditions are generated every year between 1961 and 2010 by relaxing the coupled model to analyses of atmosphere and ocean following the anomaly initialization approach, except for 10 members of the CMIP5 ensemble, which rely on full field initialization.

The number of long-lived minima is then counted for each year, and the anomalies are subsequently computed by removing the mean and dividing by the standard deviation. To allow for comparison with observations, we then multiply the time series with the standard deviation of the observed time series. Variance adjustment is necessary to account for the much larger number of tropical disturbances detected by this technique compared to observations. The

three model means are then averaged together and the variance is adjusted a second time. Additional information on this technique can be found in Smith et al. (2010), Dunstone et al. (2011), and Hermanson et al. (2014).

Hybrid forecast systems. Both hybrid forecasts rely on a multimodel ensemble (MME) of multiannual reforecasts performed within the context of CMIP5 [Geophysical Fluid Dynamics Laboratory Climate Model, version 2.1 (GFDL CM2.1) (Dunne et al. 2014) (10 members); HadCM3 (Smith et al. 2014) (20 members); and Model for Interdisciplinary Research on Climate, version 5 (MIROC5) (AORI/NIES/JAMSTEC 2015) (6 members)] and the European Seasonal-to-Decadal Climate Prediction for the Improvement of European Climate Services (SPECS) project [Max Planck Institute Earth System Model (MPI-ESM) (Matei et al. 2012) (10 members)], for a total of four forecast systems. The systems were selected from a larger pool of available systems by choosing those with start dates available every year from 1961 to 2010. The multimodel ensemble-mean hurricane anomalies are computed by giving an equal weight to each model mean, regardless of the number of ensemble members available for a particular model, and the variance of the ensemble mean of both series of reforecasts has been adjusted to match that of the observed time series.

HURRICANE NUMBERS FROM RELATIVE SEA SURFACE TEMPERATURE. With this technique (Vecchi et al. 2011), frequencies of North Atlantic hurricanes are estimated based on the weighted difference in sea surface temperature between the tropical Atlantic

and the tropics at large. More specifically, the annual Atlantic hurricane frequency λ is derived from a statistical model formulated as a Poisson regression model with two predictors and is given by

$$\lambda = \exp(1.707 + 1.388 \text{SST}_{\text{Atl}} - 1.521 \text{SST}_{\text{Trop}}), \quad (1)$$

where SST_{Atl} and SST_{Trop} are the mean SST anomalies over the tropical Atlantic (in the region 10°–25°N, 80°–20°W) and of the entire tropics (between 30°N and 30°S), respectively, during the period June–November. In this model, an increase in SST over the main development region (MDR) leads to an increase in Atlantic hurricane numbers, while an increase in SST over the tropics at large leads to a decrease in hurricane activity. The parameters in Eq. (1) are derived from the sensitivity of the hurricane response to a number of SST perturbations in a high-resolution atmospheric GCM (Vecchi et al. 2011). To be commensurable with the other techniques, the variance of the ensemble-mean reforecasted time series is adjusted to that of the hurricane time series.

AMO INDEX. In this case, we make use of the relationship between Atlantic hurricane activity and the AMO, also referred to as the Atlantic multidecadal variability (AMV), at decadal time scales and estimate hurricane activity using an AMO-proxy index developed by Klotzbach and Gray (2008). The index is constructed using the difference in standardized SST anomalies over the North Atlantic subpolar gyre (50°–60°N, 50°–10°W) and the standardized mean sea level pressure anomalies over the tropical and extratropical Atlantic (0°–50°N, 70°–10°W). To translate the forecasted index values into hurricane anomalies, we adjust the variance of the reforecasted index time series to that of the hurricane time series. Additional information on this technique can be found in Camp and Caron (2017) and Caron et al. (2015).

Combined statistical techniques. With this technique, a weighted combination of six statistical models is used to reforecast the number of hurricanes in the Atlantic basin. Model weights are based on the past performance of each model and evolve with each prediction year. Four of the statistical models use generalized linear regression models to determine the relationship between hurricane counts and either the MDR sea surface temperature or the difference in sea surface temperature between the main development region and the tropical Pacific region. Similar to the relative sea surface temperature hybrid method, a local increase in SST over the MDR leads to an increase in Atlantic hurricane numbers, while an increase in SST over the tropical Pacific leads to a decrease in hurricane activity.

The other two component models are averages of the past hurricane counts in the Atlantic basin in either active or inactive conditions—the activity state being determined using a changepoint detection technique (Jewson et al. 2009). One model includes the probability of shifting from an active to inactive state or vice versa, while the other model does not. Basin hurricane count data from 1950 to the year prior to each forecast year are used to produce a given forecast. Because the basin record is considered incomplete before the 1940s and because we require at least 30 years of data for building a reliable regression model, reforecasts cannot be made prior to 1980 with this technique. Finally, the variance of the reforecasted time series is also adjusted to match that of the hurricane time series.

DETERMINISTIC FORECASTS. Figure 1 shows that the systems capture the U shape in activity, reforecasting high activity in the 1960s (when forecasts are available), lower activity from the early 1970s to mid-1990s, and higher activity for the period that followed. There are large disagreements between the methods in the 1960s, which might be linked to the quantity and the quality of the ocean data used to initialize the climate model during those years. In terms of skill, the reforecasts generally return significant correlation coefficients for the linear (Pearson) correlation but only the AMO index technique returns a significant ranked (Kendall) correlation coefficient. The AMO index technique also returns the smallest root-mean-square error (rmse), thus suggesting an overall edge for this particular approach over the others evaluated here.

A standard technique to evaluate the reliability of ensemble forecasts consists of comparing the rmse and the spread of the ensemble. In well-calibrated forecast systems, the rmse of the ensemble mean should match the average spread of that ensemble (Fortin et al. 2014); that is, the uncertainty of the forecast should be a good measure of the error of the predictions. Here, the average model spread is defined as the square root of the time-averaged variances and the ensemble-mean spread is the square root of the sum of the average model variances, weighted according to the number of members provided by each model. In this particular case, all three systems relying on climate models are overdispersive (underconfident): the uncertainty is significantly larger than the rmse (inset, Fig. 1). Furthermore, only one observation

falls outside the prediction range with the tracking technique and none with the other two hybrid techniques (not shown). Underconfident systems will systematically give probabilities that are too low for any climate signal, thus reducing the odds that the necessary actions will be taken. It should be pointed out that the AMO index–based technique reduces the ensemble spread compared to the other techniques, both for the ensemble and for the individual models of the ensemble (not shown). In contrast, the spread of the statistical model is too small and underestimates the actual uncertainty. Such systems are said to be overconfident and underestimate the probability of extreme events.

Forecasts can also be evaluated with respect to a baseline, which in this case is a cheaper and simpler forecast, such as climatology or 10-yr persistence. The skill score (SS) is given as $1 - (\text{MAE}_{\text{forecast}}/\text{MAE}_{\text{baseline}})$, where MAE is the mean absolute error. A skill score of 1 represents a perfect reforecast and a skill score of 0 (or lower) represents no improvement over the baseline. All the techniques return a positive skill score when compared to climatology, but only the AMO index technique is significantly different from 0. The same holds when measured against a 10-yr persistence forecast, although this second baseline appears more difficult to improve upon. The better performance of the AMO index technique is likely related to the fact that the index is constructed using sea surface temperature over the northern North Atlantic, which is the region where initialization of climate models consistently returns an improvement over noninitialized climate simulations (Doblas-Reyes et al. 2013; Meehl et al. 2014), which itself has been linked to the ability of the initialized climate models to reproduce the ocean dynamics of the Atlantic meridional overturning circulation (AMOC) (Robson et al. 2012, 2014). A recent study suggests a long and robust link between the Atlantic meridional overturning circulation and the AMO (McCarthy et al. 2015). It could be argued that for the hybrid and dynamical forecast systems, much of the skill originates from the first forecast year, but as shown in the online supplemental material (https://doi.org/10.1175/BAMS-D-17-0025.2), where the results of the same analysis are repeated with forecast years 2–5 only, this does not appear to be the case.

We further evaluated whether each forecast system could accurately anticipate 5-yr periods of below-normal (lower tercile), near-normal, and above-normal (upper tercile) hurricane activity. Each such accurate prediction is identified with a colored circle at the bottom of Fig. 1. All the different techniques

have a similar success rate, in the 60%–65% range (bottom left, Fig. 1). It is worth noting that all the techniques tend to reforecast the appropriate tercile once a pattern of low or high activity has solidly been established. Around tipping points (late 1960s and mid-1990s), they tend to be less skillful, which suggests that a certain level of skill comes from persistence in the initial conditions.

WEATHER ROULETTE. To make full use of all the ensemble members and their distribution, we also adopt a probabilistic approach to reforecasting the proper terciles. And while a series of tools for probabilistic forecast evaluation exists, few are intuitive in communicating the skill to nonexperts. One
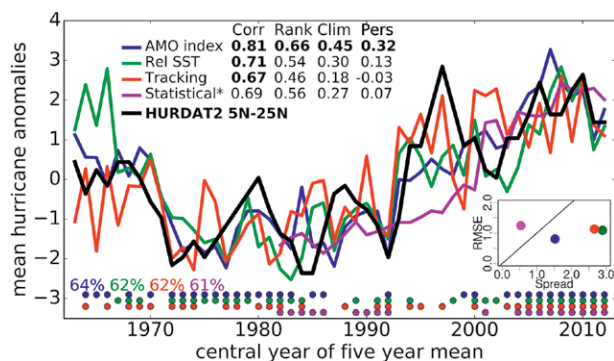
FIG. 1. Deterministic forecasts. Time series of 5-yr-mean hurricane anomalies in observations (black) and for the various forecast systems. These include forecasts made by tracking storms directly (red), forecasts based on the relative SST of the Atlantic with respect to the rest of the tropics (green), forecasts produced using a proxy for the AMO (blue), and forecasts produced using a statistical model (magenta). The 5-yr forecasts are aligned with the third year of the prediction. For observations, we consider only storms forming between 5° and 25°N. The inset table shows various measures of forecast quality: i) linear correlation index (Corr), ii) Kendall ranked correlation (Rank), and the mean absolute SS with respect to iii) a climatological forecast (Clim) and iv) a 10-yr persistence forecast (Pers). Statistically significant values for the correlations and the mean absolute SS are shown (boldface). The full circles along the x axis show the 5-yr periods for which each system's prediction landed in the right tercile, and the four numbers at the bottom left give the percentage of times that each system managed to do so. The inset plot in the bottom right compares the rmse and the spread of each technique, showing that all three forecast systems relying on climate models are underconfident and the statistical forecast system is overconfident. The asterisk denotes the statistical model skill (see inset table), which is given for the 1980–2010 period, whereas the other models are scored based on the 1961–2010 period.

diagnostic that stands out in that respect is weather roulette (Hagedorn and Smith 2009), where the skill of a forecast is quantified using an effective yearly interest rate representing the cumulative advantage obtained from using that forecast over a baseline. A game of weather roulette is illustrated in Fig. 2 and a formal description is given in the appendix.

Weather roulette is played between two opponents (a forecast and a baseline), with each player starting with the same initial capital and the roulette slots representing each of the possible terciles. The players start the first round by distributing their initial capital proportionally to the odds given for each tercile by their respective forecast. For each technique, the odds are given by the percentage of members forecasting a given tercile, while for a climatological baseline, the money is divided equally between the three categories. The money that is bet on the wrong tercile is lost (for both players), while the money that is bet on the verifying tercile is multiplied by the inverse of the probability of the baseline for that tercile [if the baseline is climatology, $1/(1/3) = 3$] and returned to each player.

The ratio of the forecast probability and the baseline probability is called the return ratio, and when the probability of the winning tercile is larger for the forecast than for the baseline, that return ratio is greater than 1 and the player betting according to the forecasts starts the next round with more money than when the round began (and vice versa). All the money is reinvested by both players in the second round (second start date), and the game is repeated for all the start dates. The skill of the forecast R is given by the geometric average of the return ratios, and the effective yearly interest rate is given by $R - 1$. A forecast that is more skillful than the baseline will return $R \geq 1$ and a positive interest rate.

Because the weather roulette requires a sufficient number of ensemble members, we can evaluate only forecast systems that rely on climate simulations (dynamical and
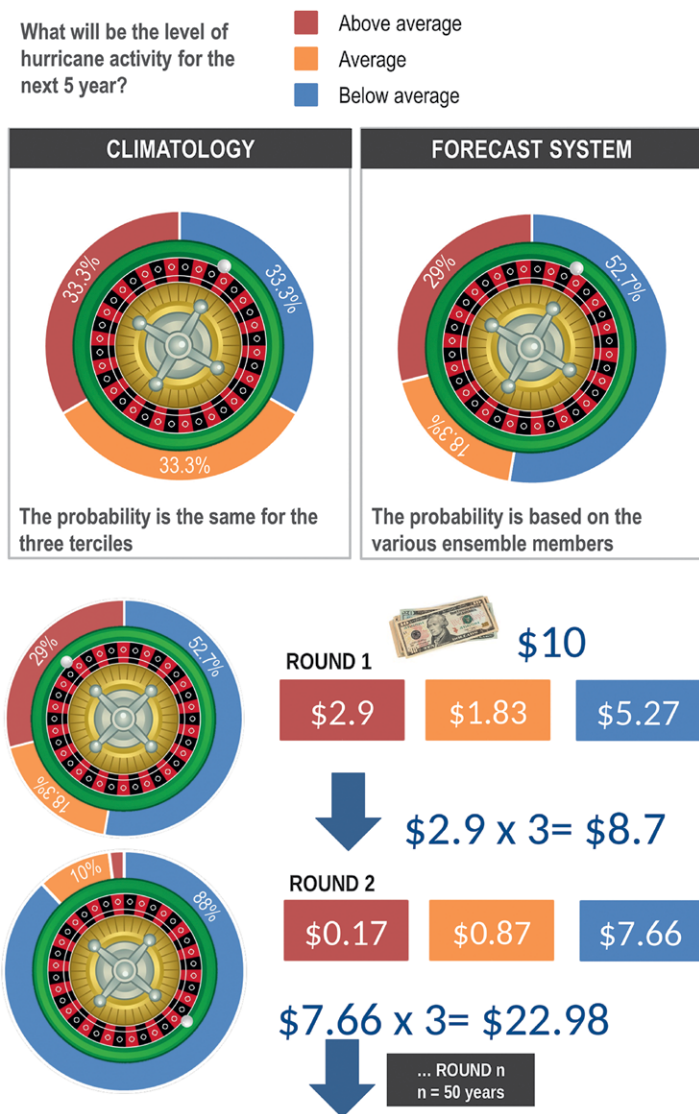

Fig. 2. An example of weather roulette. Two players bet on whether the hurricane seasons are going to be below average, near average, or above average. Both players start the game with the same amount of money (in this case $10) and spread their initial bet according to the probability given by their respective forecast. One player always bets according to climatology (top-left wheel) and always distributes 33% of the capital in each of the three categories. The second player uses a hurricane forecast system and distributes the money according to the proportion of ensemble members in each category (top-right wheel). For this player, the distribution will vary for each round. At the beginning of round 1, the player using predictions from the forecast systems puts 29% of the money on the winning category, whereas the player using the climatological forecast puts 33%. In this case, climatology gives better results and the player using a forecast system ends up with less money. This player starts round 2 with a capital of $10 × (0.29/0.33) = $8.70, whereas the other player continues with $10. In round 2, the forecast system predicts the winning category with 88% probability, thus resulting in betting 88% of the money in the right category. This player ends round 2 with $22.97 as opposed to $10 for the other player. After n rounds, the net gains associated with each strategy can be assessed.

hybrid), the three of which are measured against climatology and a combination of climatology and persistence. Figure 3 (top) compares the probability of the verifying category for each start date. The probability for the climatological forecast is always 0.33 and the verifying probability of the three forecast systems is usually greater than this value. The return ratios between the three forecast systems and the climatological forecast are given in Fig. 3 (middle). Although there is much year-to-year variation, the return ratios are usually greater than one. This is confirmed by the effective yearly interest rate, which is greater than 0 for all three systems. The skill decreases when a mix of persistence and climatology is used (Fig. 3, bottom), but again the interest rate is greater than 0 for all three forecast systems, indicating an overall better performance than the baseline. The forecast system based on the AMO index returns the highest interest rate, partly due to the high confidence in accurate forecasts calling for a higher level of activity during the later period.

CONCLUDING REMARKS. So, how skillful are the multiannual forecasts of hurricane activity originating from initialized climate models? While the current skill is still low compared to seasonal hurricane forecasts, they are better than climatological forecasts and at least as good as, but probably better than, 10-yr persistence forecasts. The constant improvement in climate models, combined with the ever-growing network of observations available to initialize them, offers hope that these forecasts will follow a path similar to that of seasonal forecasts and start providing reliable, skillful information in the not-so-distant future. Further calibration (Doblas-Reyes et al. 2005) and improvement in the correction of the climate model drift (Kharin et al. 2012) offer additional and immediate avenues by which the

current skill level can be raised. How these forecasts can be integrated into a decision-making process given the intrabasin variability (Kossin 2017) is, however, an entirely different matter.

Using a purely dynamical approach, Hermanson et al. (2014) suggested that hurricane activity will
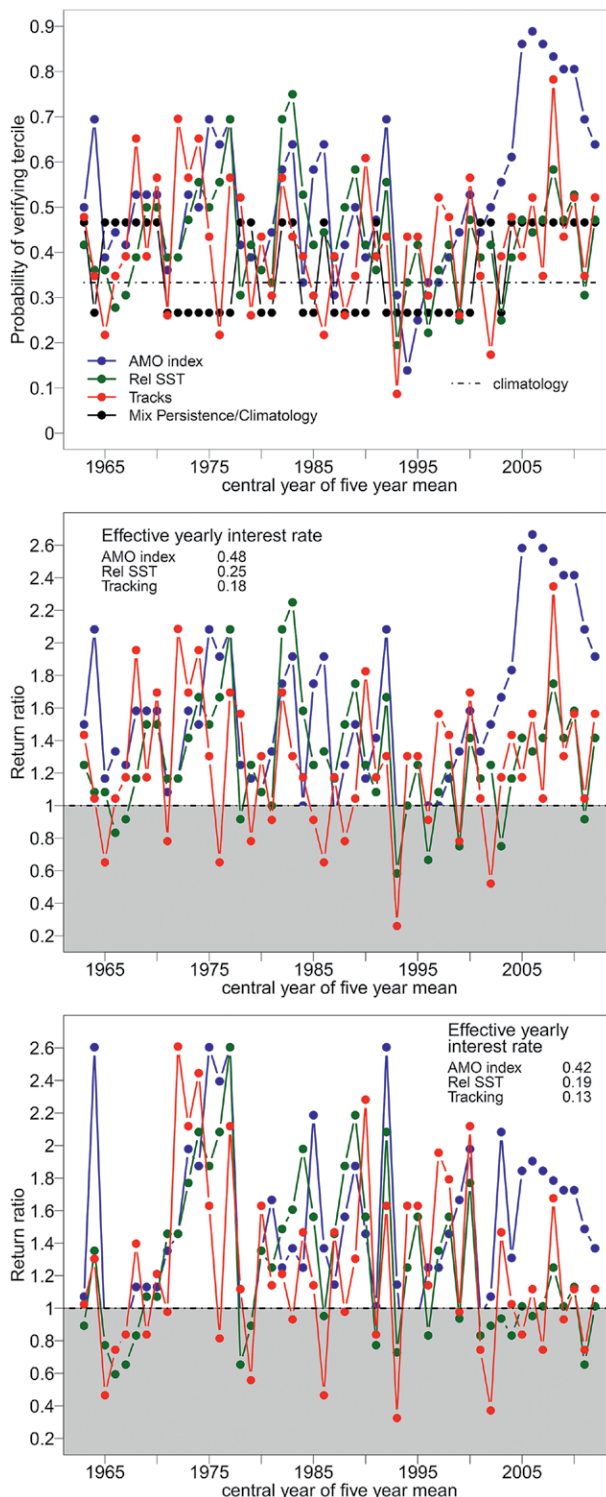
Fig. 3. Probabilistic forecast verification. (top) The probability of the verifying tercile as predicted by each of the forecast systems, the climatological forecast, and the mix climatology–persistence forecast. (middle) Return ratio for each forecast system when playing against the climatological forecast. The effective interest rate is given in the upper-left corner. (bottom) As in (middle), but when the forecast systems are measured against the mix of climatology and persistence. A return ratio >1 means that the forecast system outperformed the baseline for that year (dot sitting over the white background), while an effective yearly interest rate >0 means that the cumulative effect of using this system over the 50-yr period compared to the simpler alternative is positive.

remain low for the upcoming years. Unfortunately, most of the data available for our study originated from CMIP5, which completed in 2012. As such, the series of simulations do not cover the upcoming 5-yr period, which prevents us from using the hybrid techniques to validate that prediction. Nonetheless, there are international initiatives in the work, such as the CMIP6-endorsed Decadal Climate Prediction Project (DCPP), which will soon provide the new data required to suggest an answer to that question.

APPENDIX: STATISTICAL EVALUATION.

Anomaly correlation coefficient. Anomaly correlation coefficients (ACCs) are computed by correlating the 5-yr ensemble-mean anomalies with the observed 5-yr-mean hurricane anomalies. ACCs are computed using both standard Pearson's correlation and Kendall's rank correlation. The latter describes the ability of the forecast system to identify the relative ordering of 5-yr periods correctly and is used since we do not necessarily expect the ensemble-mean forecasts anomalies and the observed hurricane anomalies to follow a Gaussian distribution.

Autocorrelation in the time series is accounted for by considering an effective sample size n_{eff}, which approximates the number of independent data points in the time series. The effective sample size is defined such that

$$n_{eff} = \min \left\{ \frac{N}{1 + 2\sum_{\tau=1}^{N-1}[(N-\tau)/N]\rho(\tau)}, N \right\}, \quad (A1)$$

where N is the actual sample size and $\rho(\tau)$ is the autocorrelation function as a function of lag τ (von Storch and Zwiers 2001; Guemas et al. 2014). Whereas the actual sample size is the number of start dates (50), the effective sample size for the 5-yr-mean hurricane time series is much lower (10). Correlations are considered significant if the p values (shown in Table A1) are below 0.05.

Improvement over a baseline forecast. The mean absolute error skill score (SS) is used to measure improvement with respect to a baseline, taken here as either 10-yr-mean persistence forecasts or climatological forecasts. Climatology here is defined as the average from 1900 to the year prior to the forecast but using a different start point to compute the climatology does not impact the results. The mean absolute skill score is defined such that

$$SS = 1 - \frac{MAE_{forecast}}{MAE_{baseline}}, \quad (A2)$$

where $MAE_{forecast}$ and $MAE_{baseline}$ are the mean absolute errors of the forecast and the baseline, respectively. The mean absolute error is defined as

$$MAE = \frac{1}{n}\sum_{k=1}^{n}|y_k - o_k|, \quad (A3)$$

where (y_k, o_k) is the kth of n pairs of forecasts and observations.

An SS greater (less) than 0 means that the forecast offers a better (worse) performance than the reference. An SS of 1 means a perfect forecast. The confidence interval of the score was computed using the bootstrap percentile method with 10,000 replicates and a fixed block length given by $1/n_{eff}$. The SS was considered statistically significant if the confidence interval did not include 0.

Weather roulette. The weather roulette, as developed by Hagedorn and Smith (2009), is defined as a bet between two opponents—an actual forecast and a baseline—with each player betting that the odds of her/his forecast are better. The roulette slots represent each of the three possible categories (terciles). Both players start the game with the same initial capital

TABLE A1. The p values of the Pearson and Kendall correlations.

	AMO index	Relative SST	Tracking	Statistical
Pearson	0.01	0.02	0.03	0.08
Kendall	0.03	0.09	0.15	0.19

c_0 and spread all of their capital over the categories according to the probabilities given by their respective forecast.

The odds $o(i)$ of the ball falling into each of the slots (i.e., that a tercile will verify) are given by

$$o(i) = \frac{1}{p(i)}, \qquad (A4)$$

where $i = 1, 2$, and 3 and $p(i)$ is the probability of the ith outcome. The sum of probabilities over all possible outcomes is, of course, 1:

$$\sum_{i=1}^{3} p(i) = 1. \qquad (A5)$$

For each forecasting system, the odds are given by the percentage of members forecasting a given tercile. Because one model (HadCM3) has twice as many members available to produce hybrid forecasts compared to the other models, we limit the number of members for this model to 10. This will prevent this model from being overrepresented in the ensemble.

For a climatological forecast, the probabilities p_{clim} forecasted for each tercile are $1/3 = 33.3\%$. The probabilities p_{persis} of the persistence forecast are 60% for the forecasted tercile and 20% for the other two terciles. This is necessary in order to avoid the persistence forecast from going bust if an event that is not forecasted does occur. That being said, it has been shown that combinations of persistence and climatological forecasts usually perform better than either of the two standards taken individually (Buell 1958; Murphy 1992). The probabilities of the mix forecast (persistence and climatology) are constructed such that

$$p_{\text{mix}}(i) = \frac{p_{\text{persis}}(i) + p_{\text{clim}}(i)}{2}. \qquad (A6)$$

After the outcome of the first round is established, the capital that was bet on the wrong terciles is lost (for both players), while the capital c_1 that was bet on the verifying outcome v is returned to each player such that

$$c_1 = \frac{p(v)}{p_{\text{baseline}}(v)} c_0. \qquad (A7)$$

The ratio of the probabilities of the forecast over the baseline is defined as the return ratio r:

$$r = \frac{c_1}{c_0} = \frac{p(v)}{p_{\text{baseline}}(v)}. \qquad (A8)$$

When the probability of the winning tercile is larger for the forecast system than for the baseline, the return ratio is greater than 1 and that player starts the next round with more money than when the round began (and vice versa). At the end of each round, all the money is reinvested in the following round and the game is repeated until the last start date. The skill of the forecast R is given by the geometric average of the return ratios, which is given by

$$R = \sqrt[n]{\prod_{i=1}^{n} r_i}, \qquad (A9)$$

where n is the total number of rounds, which in this case is the number of forecasts produced (50). Finally, the effective yearly interest rate (IR) is given by $R - 1$. A forecast that is more skillful than the baseline will return $R \geq 1$ and a positive interest rate. Finally, it can be shown that IR is related to the ignorance score (IS), which is a proper score, by the following transformation:

$$\text{IR} = 3 \times 2^{-\text{IS}}. \qquad (A10)$$

Note that there was not a sufficiently large number of ensemble members in the statistical model to evaluate that technique with the weather roulette.

REFERENCES

AORI/NIES/JAMSTEC, 2015: MIROC5 model output prepared for CMIP5 decadals, served by ESGF. World Data Center for Climate at Deutsches Klimarechenzentrum, https://doi.org/10.1594/WDCC/CMIP5.MIM5DEC.

Bonazzi, A., A. L. Dobbin, J. K. Turner, P. S. Wilson, C. Mitas, and E. Bellone, 2014: A simulation approach for estimating hurricane risk over a 5-yr horizon. *Wea. Climate Soc.*, **6**, 77–90, https://doi.org/10.1175/WCAS-D-13-00025.1.

Buell, C. E., 1958: Meaning of combined climate and persistence forecast. *J. Meteor.*, **15**, 564–565, https://doi.org/10.1175/1520-0469(1958)015<0564:moccap>2.0.co;2.

Camargo, S. J., A. H. Sobel, A. G. Barnston, and K. A. Emanuel, 2007: Tropical cyclone genesis potential index in climate models. *Tellus*, **59A**, 428–443, https://doi.org/10.3402/tellusa.v59i4.15014.

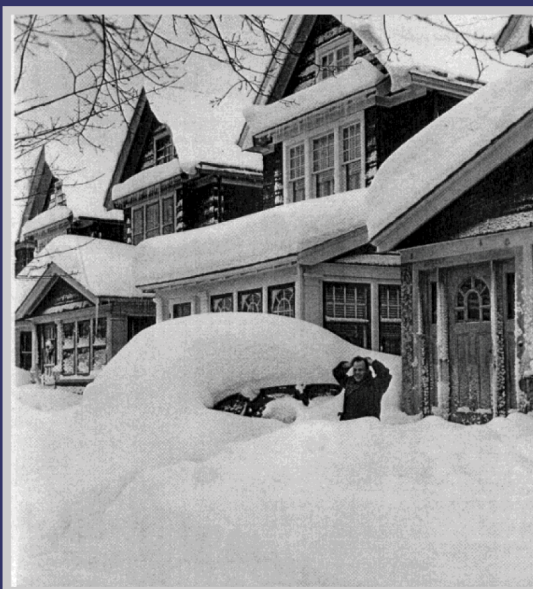Camp, J., and L.-P. Caron, 2017: Analysis of Atlantic tropical cyclone landfall forecasts in coupled GCMs

on seasonal and decadal timescales. *Hurricanes and Climate Change*, 3rd ed. J. Collins and K. Walsh, Eds., Springer, 213–241, https://doi.org/10.1007/978 -3-319-47594-3_9.

Caron, L.-P., C. G. Jones, and F. Doblas-Reyes, 2014: Multi-year prediction skill of Atlantic hurricane activity in CMIP5 decadal hindcasts. *Climate Dyn.*, **42**, 2675–2690, https://doi.org/10.1007/s00382-013 -1773-1.

——, L. Hermanson, and F. J. Doblas-Reyes, 2015: Multiannual forecasts of Atlantic U.S. tropical cyclone wind damage potential. *Geophys. Res. Lett.*, **42**, 2417–2425, https://doi.org/10.1002/2015gl063303.

Chenoweth, M., and D. Divine, 2014: Eastern Atlantic tropical cyclone frequency from 1851–1898 is comparable to satellite era frequency. *Environ. Res. Lett.*, **9**, 114023, https://doi.org/10.1088/1748 -9326/9/11/114023.

Delworth, T. L., and Coauthors, 2006: GFDL's CM2 global coupled climate models. Part I: Formulation and simulation characteristics. *J. Climate*, **19**, 643– 674, https://doi.org/10.1088/1748-9326/9/11/114023.

Doblas-Reyes, F. J., R. Hagedorn, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting—II. Calibration and combination. *Tellus*, **57A**, 234–252, https://doi .org/10.3402/tellusa.v57i3.14658.

——, and Coauthors, 2013: Initialized near-term regional climate change prediction. *Nat. Commun.*, **4**, 1715, https://doi.org/10.1038/ncomms2704.

Dunne, J. P., J. G. John, A. J. Adcroft, R. W. Hall- berg, S. M. Griffies, E. Shevliakova, R. J. Stouffer, J. P. Krasting, L. T. Sentman, P. C. D. Milly, S. L. Malyshev, W. Cooke, K. A. Dunne, M. Harrison, H. Levy, B. L. Samuels, M. J. Spelman, M. Winton, A. T. Wittenberg, P. J. Phillips, and N. Zadeh, 2014: NOAA GFDL GFDL-CM2p1, decadal experiments output for CMIP5 AR5, served by ESGF. World Data Center for Climate at Deutsches Klimarechenzentrum, http://cera-www.dkrz.de/WDCC/CMIP5/Compact .jsp?acronym=NGG2DEC.

Dunstone, N. J., D. M. Smith, and R. Eade, 2011: Multi- year predictability of the tropical Atlantic atmo- sphere driven by the high latitude North Atlantic Ocean. *Geophys. Res. Lett.*, **38**, L14701, https://doi .org/10.1029/2011gl047949.

Fortin, V., M. Abazada, F. Anctil, and R. Turcotte, 2014: Why should ensemble spread match the RMSE of the ensemble mean? *J. Hydrometeor.*, **15**, 1708–1713, https://doi.org/10.1175/jhm-d-14-0008.1.

García-Serrano, J., and F. J. Doblas-Reyes, 2012: On the assessment of near-surface global temperature and North Atlantic multi-decadal variability in the ENSEMBLES decadal hindcast. *Climate Dyn.*, **39**, 2025–2040, https://doi.org/10.1007/s00382-012-1413-1.

Goldenberg, S. B., C. W. Landsea, A. M. Mestas-Nuñez, and W. M. Gray, 2001: The recent increase in Atlantic hurricane activity: Causes and implica- tions. *Science*, **293**, 474–479, https://doi.org/10.1126 /science.1060040.

Gray, W. M., 1984: Atlantic seasonal hurricane fre- quency. Part II: Forecasting its variability. *Mon. Wea. Rev.*, **112**, 1669–1683, https://doi.org/10.1175/1520 -0493(1984)112<1669:ASHFPI>2.0.CO;2.

Guemas, V., L. Auger, and F. J. Doblas-Reyes, 2014: Hypothesis testing for autocorrelated short climate time series. *J. Appl. Meteor. Climatol.*, **53**, 637–651, https://doi.org/10.1175/JAMC-D-13-064.1.

Hagedorn, R., and L. A. Smith, 2009: Communicating the value of probabilistic forecasts with weather roulette. *Meteor. Appl.*, **16**, 143–155, https://doi.org /10.1002/met.92.

Hermanson, L., R. Eade, N. H. Robinson, N. J. Dunstone, M. B. Andrews, J. R. Knight, A. A. Scaife, and D. M. Smith, 2014: Forecast cooling of the Atlantic subpolar gyre and associated impacts. *Geophys. Res. Lett.*, **41**, 5167–5174, https://doi.org/10.1002 /2014gl060420.

Jewson, S., E. Bellone, S. Khare, T. Laepple, M. Lonfat, A. O. Shay, J. Penzer, and K. Coughlin, 2009: 5 year prediction of the number of hurricanes which make United States landfall. *Hurricanes and Climate Change*, J. B. Elsner, and T. H. Jagger, Eds., Springer, 73–99, https://doi.org/10.1007/978-0-387-09410 -6_5.

Kharin, V. V., G. J. Boer, W. J. Merryfield, J. F. Scinocca, and W.-S. Lee, 2012: Statistical adjustment of decadal predictions in a changing climate. *Geophys. Res. Lett.*, **39**, L19705, https://doi.org/10.1029/2012gl052647.

Klotzbach, P. J., and W. M. Gray, 2008: Multidecadal variability in North Atlantic tropical cyclone activ- ity. *J. Climate*, **21**, 3929–3935, https://doi.org/10.1175 /2008jcli2162.1.

——, ——, and C. Fogarty, 2015: Active Atlantic hur- ricane era at its end? *Nat. Geosci.*, **8**, 737–738, https:// doi.org/10.1038/ngeo2529.

Knight, J. R., C. K. Folland, and A. A. Scaife, 2006: Climate impacts of the Atlantic Multidecadal Oscillation. *Geophys. Res. Lett.*, **33**, L17706, https://doi .org/10.1029/2006gl026242.

——, and Coauthors, 2014: Predictions of climate sev- eral years ahead using an improved decadal predic- tion system. *J. Climate*, **27**, 7550–7567, https://doi .org/10.1175/jcli-d-14-00069.1.

Kossin, J. P., 2017: Hurricane intensification along United States coast suppressed during active

hurricane periods. *Nature*, **541**, 390–393, https://doi .org/10.1038/nature20783.

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.

Matei, D., H. Pohlmann, J. Jungclaus, W. Müller, H. Haak, and J. Marotzke, 2012: Two tales of initializing decadal climate prediction experiments with the ECHAM5/MPI-OM model. *J. Climate*, **25**, 8502–8523, https://doi.org/10.1175/jcli-d-11 -00633.1.

McCarthy, G. D., I. D. Haigh, J. J.-M. Hirschi, J. P. Grist, and D. A. Smeed, 2015: Ocean impact on decadal Atlantic climate variability revealed by sea-level observations. *Nature*, **521**, 508–510, https://doi .org/10.1038/nature14491.

Meehl, G. A., and Coauthors, 2014: Decadal climate predictions: An update from the trenches. *Bull. Amer. Meteor. Soc.*, **95**, 243–267, https://doi.org/10.1175 /bams-d-12-00241.1.

Meinshausen, M., and Coauthors, 2011: The RCP greenhouse gas concentrations and their extensions from 1765 to 2300. *Climatic Change*, **109**, 213–241, https://doi.org/10.1007/s10584-011-0156-z.

Murphy, A. H., 1992: Climatology, persistence, and their linear combination as standards of reference in skill scores. *Wea. Forecasting*, **7**, 692–698, https:// doi.org/10.1175/1520-0434(1992)007<0692:cpatlc >2.0.co;2.

Robson, J., R. Sutton, K. Lohmann, D. Smith, and M. D. Palmer, 2012: Causes of the rapid warming of the North Atlantic Ocean in the mid-1990s. *J. Climate*, **25**, 4116–4134, https://doi.org/10.1175/jcli -d-11-00443.1.

——, ——, and D. Smith, 2014: Decadal predictions of the cooling and freshening of the North Atlantic in the 1960s and the role of ocean circulation. *Climate Dyn.*, **42**, 2353–2365, https://doi.org/10.1007/s00382 -014-2115-7.

Smith, D. M., R. Eade, N. J. Dunstone, D. Fereday, J. M. Murphy, H. Pohlmann, and A. A. Scaife, 2010: Skilful multi-year predictions of Atlantic hurricane frequency. *Nat. Geosci.*, **3**, 846–849, https://doi .org/10.1038/ngeo1004.

——, H. Pohlmann, and R. Eade, 2014: HadCM3 model output prepared for CMIP5 decadal experiments, served by ESGF. World Data Center for Climate at Deutsches Klimarechenzentrum, https://cera-www.dkrz.de/WDCC/CMIP5/Compact .jsp?acronym=MOC302.

Ullrich, P. A., and C. M. Zarzycki, 2017: TempestExtremes: A framework for scale-insensitive pointwise feature tracking on unstructured grids. *Geosci. Model Dev.*, **10**, 1069–1090, https://doi.org/10.5194/gmd-10 -1069-2017.

Vecchi, G. A., and T. R. Knutson, 2011: Estimating annual numbers of Atlantic hurricanes missing from the HURDAT database (1878–1965) using ship track density. *J. Climate*, **24**, 1736–1746, https://doi .org/10.1175/2010jcli3810.1.

——, M. Zhao, H. Wang, G. Villarini, A. Rosati, A. Kumar, I. M. Held, and R. Gudgel, 2011: Statistical–dynamical predictions of seasonal North Atlantic hurricane activity. *Mon. Wea. Rev.*, **139**, 1070–1082, https://doi.org/10.1175/2010mwr3499.1.

——, and Coauthors, 2013: Multiyear predictions of North Atlantic hurricane frequency: Promise and limitations. *J. Climate*, **26**, 5337–5357, https://doi .org/10.1175/jcli-d-12-00464.1.

Vitart, F., and Coauthors, 2007: Dynamically-based seasonal forecasts of Atlantic tropical storm activity issued in June by EUROSIP. *Geophys. Res. Lett.*, **34**, L16815, https://doi.org/10.1029/2007gl030740.

von Storch, H., and F. Zwiers, 2001: *Statistical Analysis in Climate Research*. Cambridge University Press, 484 pp.

Zhang, R., and T. L. Delworth, 2006: Impact of Atlantic multidecadal oscillations on India/Sahel rainfall and Atlantic hurricanes. *Geophys. Res. Lett.*, **33**, L17712, https://doi.org/10.1029/2006gl026267.