Multiyear Predictions of North Atlantic Hurricane Frequency: Promise and Limitations

GABRIEL A. VECCHI, RYM MSADEK, WHIT ANDERSON, YOU-SOON CHANG, THOMAS DELWORTH, KEITH DIXON, RICH GUDGEL, ANTHONY ROSATI, AND BILL STERN

Geophysical Fluid Dynamics Laboratory, NOAA, Princeton, New Jersey

GABRIELE VILLARINI

IIHR–Hydroscience & Engineering, The University of Iowa, Iowa City, Iowa

ANDREW WITTENBERG, XIASONG YANG, FANRONG ZENG, RONG ZHANG, AND SHAOQING ZHANG

Geophysical Fluid Dynamics Laboratory, NOAA, Princeton, New Jersey

(Manuscript received 19 July 2012, in final form 30 January 2013)

ABSTRACT

Retrospective predictions of multiyear North Atlantic Ocean hurricane frequency are explored by applying a hybrid statistical–dynamical forecast system to initialized and noninitialized multiyear forecasts of tropical Atlantic and tropical-mean sea surface temperatures (SSTs) from two global climate model forecast systems. By accounting for impacts of initialization and radiative forcing, retrospective predictions of 5- and 9-yr mean tropical Atlantic hurricane frequency show significant correlations relative to a null hypothesis of zero correlation. The retrospective correlations are increased in a two-model average forecast and by using a lagged-ensemble approach, with the two-model ensemble decadal forecasts of hurricane frequency over 1961–2011 yielding correlation coefficients that approach 0.9. These encouraging retrospective multiyear hurricane predictions, however, should be interpreted with care: although initialized forecasts have higher nominal skill than uninitialized ones, the relatively short record and large autocorrelation of the time series limits confidence in distinguishing between the skill caused by external forcing and that added by initialization. The nominal increase in correlation in the initialized forecasts relative to the uninitialized experiments is caused by improved representation of the multiyear tropical Atlantic SST anomalies. The skill in the initialized forecasts comes in large part from the persistence of a mid-1990s shift by the initialized forecasts, rather than from predicting its evolution. Predicting shifts like that observed in 1994/95 remains a critical issue for the success of multiyear forecasts of Atlantic hurricane frequency. The retrospective forecasts highlight the possibility that changes in observing system impact forecast performance.

I. Introduction

Predicting and projecting future North Atlantic Ocean hurricane activity is a topic of scientific interest (e.g., Gray 1984; Knutson and Tuleya 2004; Emanuel 2005; Camargo et al. 2007a; Vecchi et al. 2008; Smith et al. 2010, hereafter S10; Knutson et al. 2010; Vecchi et al. 2011; Villarini et al. 2011b; Villarini and Vecchi 2012b, 2013a,b) and high societal significance (Pielke et al. 2008; Mendelsohn et al. 2012; Peduzzi et al. 2012). The seasonal basinwide frequency of North Atlantic hurricanes has exhibited variability on a variety of time scales, from interannual to multidecadal, although it remains unclear whether there has been any century-scale trend in Atlantic hurricane frequency (e.g., Mann and Emanuel 2006; Vecchi and Knutson 2008, 2011; Landsea et al. 2010; Villarini et al. 2011a).

The scientific basis for predictions of seasonal hurricane activity at leads of one to three seasons has been developed (e.g., Gray 1984; Elsner and Jagger 2006; Vitart 2006; Camargo et al. 2007a,b; Vitart et al. 2007;

Corresponding author address: Gabriel A. Vecchi, Geophysical Fluid Dynamics Laboratory, NOAA, U.S. Route 1, Forrestal Campus, Princeton, NJ 08542.
E-mail: gabriel.a.vecchi@noaa.gov

DOI: 10.1175/JCLI-D-12-00464.1

Klotzbach and Gray 2009; Wang et al. 2009; Kim and Webster 2010; LaRow et al. 2010; Zhao et al. 2010; Alessandri et al. 2011; Chen and Lin 2011; Vecchi et al. 2011; Villarini and Vecchi 2013b), leading to the identification of different potential sources of skill, both local and remote.

Decadal to centennial projections of seasonal hurricane activity in response to changes in external forcing (greenhouse gases, aerosols, volcanoes, and solar) have been made (e.g., Oouchi et al. 2006; Knutson et al. 2008; Emanuel et al. 2008; Gualdi et al. 2008; Vecchi et al. 2008; Sugi et al. 2009, 2012; Zhao et al. 2009; Bender et al. 2010; Knutson et al. 2010; Villarini et al. 2011b; Zhao and Held 2011; Villarini and Vecchi 2012b, 2013a). The basis for these projections is the possibility that radiatively forced climate change could influence the climatic conditions to which hurricanes are sensitive, such as large-scale circulation, wind shear, ocean temperatures, potential intensity, and humidity (e.g., Emanuel 1987, 2007; Broccoli and Manabe 1990; Shen et al. 2000; Knutson and Tuleya 2004; Camargo et al. 2007b; Vecchi and Soden 2007a,b). Recent model results span a relatively wide range of possibilities for North Atlantic hurricane frequency (including increases or decreases) under enhanced CO_2-induced warming, while there is a wider tendency for hurricane intensity to increase in these studies (e.g., Knutson and Tuleya 2004; Knutson et al. 2008; Emanuel et al. 2008; Gualdi et al. 2008; Knutson et al. 2008; Vecchi et al. 2008; Sugi et al. 2009, 2012; Zhao et al. 2009; Bender et al. 2010; Knutson et al. 2010; Villarini et al. 2011b; Villarini and Vecchi 2012b, 2013a). There are indications that changes in atmospheric aerosols could influence past and projected hurricane activity, with increases (decreases) in Atlantic aerosol loading driving decreases (increases) in Atlantic hurricane activity (Mann and Emanuel 2006; Evan et al. 2009; Villarini and Vecchi 2012b, 2013a).

Assessing hurricane predictability at intermediate time scales, between seasonal predictions and multidecadal projections, is an emerging field of research. In addition to potential influences caused by changes in radiative forcing, internal variations of the climate system could play a large role in changes of hurricane frequency on time scales of decades (e.g., Goldenberg et al. 2001; Zhang and Delworth 2006, 2009; Knight et al. 2005; Latif et al. 2007; Dunstone et al. 2011; Villarini et al. 2011b; Villarini and Vecchi 2012b). There are physical reasons to expect coherent multiyear hurricane variations to be tied to ocean changes (e.g., Goldenberg et al. 2001; Zhang and Delworth 2005, 2006, 2009; Knight et al. 2005; Latif et al. 2007; Dunstone et al. 2011). There are also indications that some of the relevant ocean changes may be potentially predictable on decadal time scales (e.g., Griffies and Bryan 1997a,b; Pohlmann et al. 2004; Collins et al. 2006; Pohlmann et al. 2009; Msadek et al. 2010; S10; Teng et al. 2011; Chikamoto et al. 2013; van Oldenborgh et al. 2012; A. Rosati et al. 2012, unpublished manuscript; Yang et al. 2012; Yeager et al. 2012). As decadal variability and the associated predictability can result from both internally and externally forced fluctuations (e.g., Rotstayn and Lohmann 2002; Hawkins and Sutton 2009; C. Chang et al. 2011; Villarini et al. 2011b; Booth et al. 2012; Zhang et al. 2013; Villarini and Vecchi 2012b), one has to consider skill arising from both external factors and internal variability on multiyear time scales. A number of modeling groups are now following the same framework for phase 5 of the Coupled Model Intercomparison Project (CMIP5; Taylor et al. 2012) to be assessed as part of the Fifth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC AR5), by performing decadal predictions initialized with estimates of the observed state of the climate system (Taylor et al. 2012; Meehl et al. 2013). While for sea surface temperatures (SSTs) most of the skill on multiyear time scales arises from predicting the warming trend associated with radiative forcing changes (e.g., van Oldenborgh et al. 2012; A. Rosati et al. 2012, unpublished manuscript), there is at least one study suggesting that initialization can increase the skill in multiyear hurricane forecasts (S10). In this paper, we explore the ability of a hybrid statistical–dynamical hurricane forecasting system to retrospectively predict multiyear hurricane activity in the Atlantic using two different coupled climate models, including the one used by S10. We explore the skill of North Atlantic hurricane frequency resulting from changing radiative forcing and from natural variability. We assess the improvement in skill caused by initialization and discuss the source of this improved skill and its implications for future multiyear forecasts of North Atlantic hurricane frequency.

2. Data and methods

a. Statistical hurricane emulator

We use a hybrid statistical–dynamical North Atlantic hurricane frequency prediction framework to explore the predictability of multiyear hurricane activity. This framework has been shown to exhibit retrospective skill in seasonal hurricane forecasts from as early as boreal winter prior to the hurricane season (Vecchi et al. 2011). It combines a statistical emulator of a high-resolution dynamical atmospheric model (Zhao et al. 2009, 2010) and initialized forecasts of SST. The statistical emulator is formulated as a Poisson regression model with two

predictors: Tropical Atlantic SST and tropical-mean SST, each averaged over the August–October season.

The choice of these two predictors is motivated by dynamical considerations, observed relationships between hurricane activity and SST and the sensitivity of dynamical models to SST perturbations. Observational analyses have highlighted correlations between SST changes in the tropical Atlantic and hurricane activity indices (e.g., Elsner and Jagger 2006; Emanuel 2005). However, observational correlations as high or higher have been found between hurricane activity and the weighted difference between Atlantic and tropical-mean SSTs (the SST changes in the Atlantic relative to the tropics, or ''relative SST'') by other studies (e.g., Swanson 2007, 2008; Vecchi et al. 2008; Villarini et al. 2010, 2011b, 2012; Villarini and Vecchi 2012a). The physical basis for exploring relative SST as a predictor of hurricane activity is based on the tendency of free tropospheric temperature changes to follow those of tropical-mean SST (Sobel et al. 2002) or SSTs in the Indo-Pacific Oceans region where the bulk of tropical convection resides (Tang and Neelin 2004) as described by the weak temperature gradient approximation (Sobel and Bretherton 2000). An Atlantic SST warming that is larger than that of the tropical average, with a tropospheric warming in the Atlantic that follows tropical-mean SST, would lead to a large-scale destabilization of the atmosphere in the Atlantic, to changes in the large-scale vorticity, shear, and atmospheric humidity, and to increases in tropical cyclone (TC) potential intensity (e.g., Latif et al. 2007; Vecchi and Soden 2007a; Gualdi et al. 2008; Sugi et al. 2009, 2012; Zhao et al. 2009; Xie et al. 2010; Zhao and Held 2011; Ramsay and Sobel 2011; Camargo et al. 2013; Vecchi et al. 2013). Supporting the notion of relative SST as a predictor for Atlantic hurricane activity, dynamical modeling studies have found that the threshold for TC genesis under projected climate changes over the twenty-first century increases along with the overall tropical warming (e.g., Knutson et al. 2008). The interannual, decadal, and climate change response of North Atlantic TC frequency simulated across a range of dynamical frameworks is also well explained by relative SST (e.g., Vecchi et al. 2008; Sugi et al. 2009, 2012; Zhao et al. 2009, 2010; Vecchi et al. 2011; Villarini et al. 2011b; Knutson et al. 2013; Zhao and Held 2011), although strong departures caused by moist adiabatic warming can complicate relative SST models of hurricane frequency (e.g., Vecchi et al. 2013).

Following Vecchi et al. (2011), we model the rate of occurrence λ (the expected value of the aggregate seasonal number) of North Atlantic hurricane frequency using a Poisson regression model as follows:

$$\lambda = e^{1.707+1.388\mathrm{SST}_{\mathrm{MDR}}-1.521\mathrm{SST}_{\mathrm{TROP}}}, \quad (1)$$

where $\mathrm{SST}_{\mathrm{MDR}}$ and $\mathrm{SST}_{\mathrm{TROP}}$ are anomalies in the regional SST indices relative to the 1982–2005 average, as described in section 2c. Note that $\mathrm{SST}_{\mathrm{MDR}}$ is the average over the hurricane main development region (10°–25°N, 80°–20°W), and $\mathrm{SST}_{\mathrm{TROP}}$ is the global, 30°S–30°N average of SST. As discussed in Vecchi et al. (2011), this statistical emulator of the sensitivity of hurricane frequency to SST changes in the Zhao et al. (2009, 2010) high-resolution atmospheric model was trained across a broad range of climate states, including multiple realizations of the historical period and various projections of twenty-first-century SST change. This statistical model's performance against the observed record satisfies a necessary condition for its application to interannual to decadal predictions (Vecchi et al. 2011). The parameters in this statistical emulator, built on the output of a high-resolution AGCM, are very similar to those that arise from modeling-adjusted hurricane frequency over the 1878–2008 period (Villarini et al. 2012). This statistical emulator is able to reproduce much of the observed variability in hurricane activity ($r^2 = 0.58$; Vecchi et al. 2011), and its ability to recover changes in hurricane frequency compares well with hindcasts and projections from high-resolution dynamical models (e.g., Zhao et al. 2009, 2010; Villarini et al. 2011b; Knutson et al. 2013). The low computational cost of the statistical emulator allows us to efficiently perform a variety of retrospective forecasts using multiple input datasets, described below.

b. Global climate model predictions

The statistical emulator (described above) is applied to predictions of SST from two global climate models: the National Oceanic and Atmospheric Administration (NOAA) Geophysical Fluid Dynamics Laboratory (GFDL) Climate Model, version 2.1 (CM2.1), and the Met Office (UKMO) Decadal Prediction System (DePreSys) Perturbed Physics Ensemble (PPE), referred to as GFDL-DecPre and UKMO-DePreSys, respectively. The forecast system specifications are summarized in Table 1. These models are just two of those that will be part of the CMIP5 decadal prediction experiments, although the CMIP5 version of UKMO-DePreSys is slightly different from the one used here. Exploration of those models allows us to compare the behavior of a prediction system that has shown skill in interannual hurricane predictions using the hybrid statistical–dynamical framework (i.e., GFDL-DecPre; Vecchi et al. 2011) and also to apply the hybrid framework to a model system that has shown high multiyear correlations using an alternative approach

TABLE 1. Summary of the two dynamical multiyear experimental forecast systems explored in this manuscript.

Forecast system	Underlying GCM	Initialization procedure	Ensemble type	Initialization date	Treatment of volcanoes
GFDL CM2.1 DecPre (A. Rosati et al. 2012, unpublished manuscript; Yang et al. 2012)	GFDL CM2.1 (Delworth et al. 2006)	Fully coupled ensemble Kalman filter (Zhang et al. 2007), full variable assimilation	Ten ensemble members from ensemble Kalman filter (EnKF) assimilation	1 Jan (each year) 1960–2011	Forcing from future volcanoes included
UKMO-DepPreSys PPE (Smith et al. 2007; S10)	HadCM3 (Gordon et al. 2000)	Atmospheric and oceanic conditions relaxed to observations. Ocean anomaly initialization (Smith and Murphy 2007)	Nine ensemble member PPE	1 Nov (each year) 1960–2005	Forcing from past volcanoes included

(i.e., UKMO-DePreSys; S10). Additionally, these two models generated a full ensemble of initialized predictions each year, rather than every five years as in many other CMIP5 experiments (Meehl et al. 2013), allowing us to more fully explore past performance.

The GFDL decadal climate hindcasts (i.e., GFDL-DecPre) are carried out over the period 1961–2011 using the GFDL CM2.1 coupled system (Delworth et al. 2006), in which both the atmosphere and the ocean are initialized through a full-field assimilation to bring the state of the coupled model close to observations. The initial conditions are produced with the GFDL fully coupled reanalysis version 3.1 of the GFDL Ensemble Coupled Data Assimilation System (ECDA3.1), which is based on an ensemble Kalman filter (Zhang et al. 2007; Zhang and Rosati 2010; Y. Chang et al. 2011b) and has been shown to produce a realistic ocean-mean state and variability (Chang et al. 2013). Ensembles of 10 members are produced starting on 1 January every year from 1961 to 2011 and run for 10 years. Historical radiative forcing is used for the 1961–2005 period and the representative concentration pathway (RCP) 4.5 scenario for the predictions starting after 2005. A 10-member ensemble of uninitialized runs with the same forcings has also been produced to investigate the impact of initialization. This forecast suite is further discussed in A. Rosati et al. (2012, unpublished manuscript) and its retrospective skill in predicting Atlantic multidecadal oscillation (AMO)-like variability is described in Yang et al. (2012).

The DePreSys (Smith et al. 2007) is based on the Hadley Centre Coupled Model, version 3 (HadCM3; Gordon et al. 2000). The UKMO-DePreSys PPE (S10) is an updated version that uses a 9-member ensemble of model variants that aims to sample model uncertainties through perturbations to poorly constrained atmospheric and surface parameters. Initial conditions are created by relaxing the model's components toward atmospheric [European Centre for Medium-Range Weather Forecasts (ECMWF) analysis and reanalysis] and oceanic (Smith and Murphy 2007) analysis, with values assimilated as anomalies with respect to the model climate. The purpose of anomaly assimilation is to minimize climate drift after the assimilation is switched off, but this does not totally suppress the bias as discussed in Robson (2011). The 10-yr-long decadal retrospective forecasts consist of 9-member ensembles starting on 1 November every year from 1960 to 2005. A parallel set of nine uninitialized experiments using the DePreSys PPE is also used, and is referred to as the UKMO-DePreSys uninitialized forecast runs. The DePreSys experiments do not include future volcanic information in them, only volcanic aerosols from eruptions prior to the initialization; thus, each initial year has a unique suite of uninitialized experiments. We use the UKMO-DePreSys PPE data rather than the CMIP5 UKMO-DePreSys output in order to make a comparison with the results of S10.

We also perform a two-model average prediction by first running the statistical emulator on the output from each model, and then averaging the predictions of the two models. Previous experience with interannual hurricane forecasts indicates that a two-model average can have advantages over each individual model (Vecchi et al. 2011). Further work with the full suite of CMIP5 models is underway (Caron et al. 2013).

c. *Lead-dependent climatology*

The statistical hurricane emulator is defined in terms of SST anomalies (SSTAs) with respect to the 1982–2005 climatology (Vecchi et al. 2011). The initialized and uninitialized model forecasts have their own climatology, which—for initialized forecasts using both models and for uninitialized forecasts using UKMO-DePreSys PPE—can depend on the lead time of the forecast. The uninitialized forecasts of DePreSys PPE have a lead-dependent climatology because the history of radiative forcing seen by forecasts verifying on the same year can

depend on the initialization year, because no ''future'' volcanic information is included in these uninitialized experiments. Therefore, we define a different climatology for each experiment (initialized and uninitialized) for each model (GFDL-DecPre and UKMO-DePreSys PPE). For the initialized model experiments we build a climatology that depends on lead time by averaging, for each lead time between 1 and 10 years, the forecasts that verify the years 1982–2005. We choose this as our reference period for two principal reasons: 1) the statistical model of Vecchi et al. (2011) was trained referenced to 1982–2005, and 2) the trade-off between trying to train over a period in which the observing system used to initialize the forecasts was relatively stable and there was the desire to have a long record to faithfully define the model drift. Using other reference periods does not alter the principal results of this manuscript. To compute the model climatology we average all 10 ensemble members for GFDL-DecPre, but because UKMO-DePreSys PPE is a ''perturbed physics ensemble'' a different climatology is defined for each of its 9 ensemble members. Note that a key impact of subtracting the lead-dependent climatology is to remove a systematic bias that arises in the forecasts as the models drift toward their own mean state when initialized with observations (Stockdale 1997; ICPO 2011). The drift of the models used here is toward each model's free running climatology, though even after 10 years there are regions where the initialized experiments have not yet settled at the free running climatology—these regions tend to roughly coincide with the regions where a potentially predictable decadal signal has been identified in the literature (e.g., Yang et al. 2012). A key assumption is that the systematic drift of the models does not depend on the initialization period—that is, that the systematic drift does not depend on the changes to the climate observing system that have occurred in the last 50 years. The stationary drift assumption has been shown to be problematic in interannual predictions, where change in observing system can modify the drift, and a suggested solution is to use different lead-dependent climatologies across major changes in observing system (e.g., Kumar et al. 2012). The assumption that the drift is stationary will be further discussed in section 4.

d. Skill measures

We explore two statistical measures to quantitatively assess retrospective performance: the anomaly correlation coefficient (ACC) and the mean squared skill score (MSSS). These statistics are not independent, but offer slightly different views of the forecast model skill. The ACC is the sample correlation coefficient as a function of lead time t (or an average of lead times) between a set of forecast anomalies F'_j and observed anomalies O'_j over $j = 1, \ldots, n$ years after removing the mean of each:

$$\mathrm{ACC}(t) = \frac{\sum_{j=1}^{n} [F'_j(t) \times O'_j(t)]}{\sqrt{\sum_{j=1}^{n} F'_j(t)^2 \sum_{j=1}^{n} O'_j(t)^2}}, \tag{2}$$

where $F'_j = F_j - \overline{F}, O'_j = O_j - \overline{O}$, and the overbar denotes the time mean over the climatological period 1982–2005, which is a function of lead time t. The ACC values can range from -1 to 1, and they measure the degree to which large positive and negative excursions from the mean co-occur in the forecast and verification.

The root-mean-square error (RMSE) is often used as a measure of accuracy of the forecasts. It is defined as the square root of the mean square error (MSE):

$$\mathrm{RMSE}(t) = \sqrt{\mathrm{MSE}(t)} = \sqrt{\frac{1}{n} \sum_{j=1}^{n} [F'_j(t) - O'_j(t)]^2}. \tag{3}$$

We use here a related statistical measure, the mean squared skill score (Murphy 1988) following recommendations by Goddard et al. (2013). The MSSS is based on the MSE between the forecast and the observed climatology and represents the improvement in accuracy of the forecast over climatology:

$$\mathrm{MSSS}(t) = 1 - \frac{\mathrm{MSE}_F(t)}{\mathrm{MSE}_{\overline{X}}(t)}. \tag{4}$$

The highest MSSS value of 1 is reached when $\mathrm{MSE}_F = 0$ and $\mathrm{MSE}_{\overline{X}} \neq 0$.

Instead of using climatology as reference forecast one can use the MSE of the uninitialized projections (i.e., MSE_P) to evaluate the improved skill from initialization:

$$\mathrm{MSSS}(t) = 1 - \frac{\mathrm{MSE}_F(t)}{\mathrm{MSE}_P(t)}, \tag{5}$$

where a positive MSSS indicates that the initialized forecasts outperform the uninitialized ones. MSSS can be expressed as a function of correlation and conditional bias (Goddard et al. 2013), which is useful when interpreting an improvement of skill from initialization.

e. Assessment of statistical significance

We explored three different estimates to assess statistical significance of the correlation results against a null of zero correlation, and to compute the confidence intervals of the retrospective correlations. For the estimates of statistical significance the effective number of

degrees of freedom N_{eff} of the correlation of two time series X and Y was computed using the methodology described in Bretherton et al. (1999), using the biased estimates of autocorrelation spectrum of the various time series:

$$N_{\text{eff}} = \frac{N}{\sum_{\tau=0}^{N-1} [(1 - |\tau|)/N] r_\tau^X r_\tau^Y}, \tag{6}$$

where N is the number of samples in each time series, and r_τ^X and r_τ^Y are the estimates of autocorrelation of each time series at lag τ. Because of the large autocorrelation of the time-smoothed predicted and observed hurricane time series at even long lags, the effective degrees of freedom can be considerably smaller than the number of years in the time series. Typically, when compared with observations, the 5-yr mean initialized forecasts tend to have between 6 and 8 effective degrees of freedom and the uninitialized forecasts tend to have between 10 and 12 effective degrees of freedom—even though there are around 50 years of data that are compared. Without accounting for the strong autocorrelation in these time series, one would estimate much narrower confidence intervals and a smaller p value for the null hypothesis; failure to account for the diminished degrees of freedom can lead to a substantial overestimation of forecast skill.

Although hurricane frequency is not normally distributed, we are exploring multiyear averages of hurricane frequency, which allows us to approximate the distribution as normal. To compute confidence intervals of a correlation we use a two-sided test (because it is possible that initialization could lead to degradation in performance), and we use a one-sided test against the null hypothesis of zero correlation (because a significantly negative correlation would be a failure of the forecast system), and we have compared the results from three methods:

1) Fisher's z transformation: The sample estimate of the correlation coefficient $r_{X,Y}$ between two time series X and Y is transformed using

$$Z_{X,Y} = 0.5 \ln[(1 + r_{X,Y})/(1 - r_{X,Y})]. \tag{7}$$

The new quantity $Z_{X,Y}$ follows a z distribution with an N_{eff}-of 3 degrees of freedom (Fisher 1915, 1924; von Storch and Zwiers 1999). Using standard z-statistic tables one can estimate the confidence intervals on the mean and test against a null hypotheses of zero mean from the sample estimate $Z_{X,Y}$. To transform the confidence interval estimates of the z statistic back to correlation space, we employ the inverse Fisher's z transformation:

$$r^*_{X,Y} = \frac{e^{2Z^*_{X,Y}} - 1}{e^{2Z^*_{X,Y}} + 1}, \tag{8}$$

where $Z^*_{X,Y}$ is the estimate of the upper or lower bound on the confidence interval of the z statistic and $r^*_{X,Y}$ is the estimate of the upper or lower bound on the confidence interval of the correlation coefficient.

2) Full distribution of the correlation coefficient: Johnson et al. (1995) provide the distribution of the sample correlation coefficient R when the population correlation coefficient ρ is equal to zero:

$$p_R(r) = \frac{\Gamma[(n-1)/2]}{\Gamma(1/2)\Gamma[(n-2)/2]}(1 - r^2)^{(n-4)/2}, \quad \text{for} \quad -1 < r < 1, \tag{9}$$

where $\Gamma(\cdot)$ is the gamma function and n is the sample size. This distribution is symmetric around the zero. By using p_R, we can test the null hypothesis of no correlation at a given significance level α, by checking whether the sample correlation coefficient lies within or outside the rejection or critical region.

3) Monte Carlo estimate: For sample sizes ranging between 2 and 100, we build 100 000 estimates of the distribution of the sample correlation coefficient between two normally distributed time series of length N_{eff} and an underlying correlation ρ. We sample underlying correlation coefficients between -1 and 1, at intervals of 0.01. From this Monte Carlo estimate of the probability density function (PDF) of the sample correlation coefficient, we estimate significance against a null of zero correlation as the probability of a correlation as large as or larger than a particular sample correlation given an underlying correlation of zero. In an analogous manner, we also compute the confidence intervals on the sample correlation given an underlying correlation.

We have compared the three estimates of the confidence intervals on the correlation coefficient and null test against a correlation of zero for the retrospective forecast correlations, and have found that they are consistent with each other. For simplicity, in the manuscript we only show the estimates from the Fisher's z transformation.

3. Results

a. Retrospective hurricane forecasts

Figure 1 shows the 5-and 9-yr mean (centered on the midpoint of each interval) initialized and uninitialized

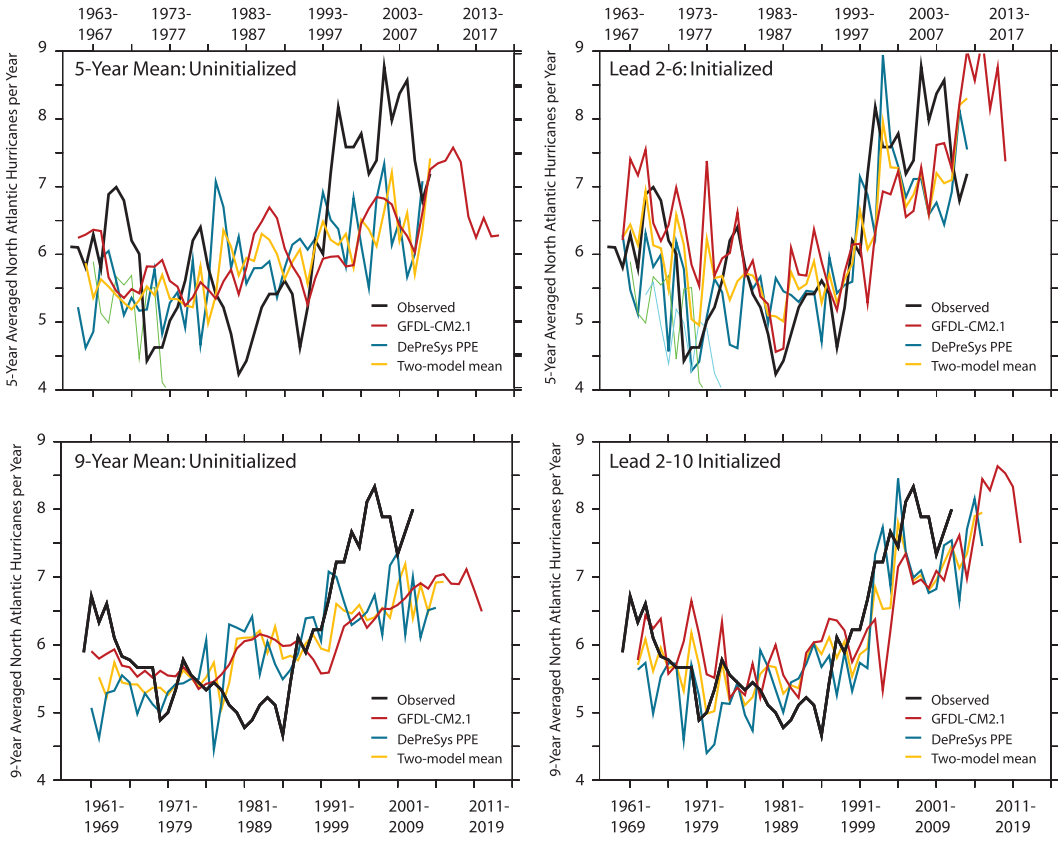

FIG. 1. Retrospective and future forecasts of hurricane frequency: (top) retrospective forecasts for 5-yr-running hurricane frequency and (bottom) 9-yr-running forecasts, showing results from (left) uninitialized and (right) initialized experiments. Black lines show the observed 5- and 9-yr hurricane counts from the NOAA Hurricane Database (HURDAT; Jarvinen et al. 1984, MacAdie et al. 2009), which includes an adjustment for observing inhomogeneity prior to 1966 described in Vecchi and Knutson (2011). For the retrospective forecasts, the red line shows the forecasts from the GFDL CM2.1 system, the blue line shows the UKMO-DePreSys PPE system, and the yellow line shows the two-system ensemble mean.

forecasts of North Atlantic hurricane frequency in GFDL-DecPre and UKMO-DePreSys PPE compared with observations. The observed record of 5-yr mean hurricane frequency is largely characterized by two distinct states with low values (~5–6 hurricanes per year) in the first half of the record and a shift in the mid-1990s (e.g., Elsner et al. 2004; Li and Lund 2012) toward a more active state (~8 hurricanes per year). The uninitialized predictions capture a tendency for an increase in hurricane frequency over the late twentieth century, indicating that part of the recent increase in Atlantic hurricane frequency was caused by changes in radiative forcing, consistent with other recent findings (e.g., S10; Villarini and Vecchi 2012b, 2013a). However, the uninitialized experiments fail to capture the abrupt shift in the mid-1990s. The initialized retrospective forecasts show better qualitative agreement to observations than do the initialized runs, suggesting an improvement from initialization.

Despite the time averaging, both observations and the model predictions have year-to-year variability in 5-yr North Atlantic hurricane frequency, which complicates detection of decadal changes (Fig. 1). The year-to-year variations in the multiyear initialized forecasts are larger than that in observations, even though the forecasts are ensemble averages. This result is particularly striking given that the statistical emulator should only recover a fraction of the observed variance, and suggests that the initialized forecasts have too much internal variability. An alternative interpretation, which is discussed further in section 3c below, is that the initial conditions for each year's initialization persist too strongly, so that each initialization year's climate reflects the average of multiple subsequent years.

The anomaly correlation between the observed hurricane counts and the model predictions for both initialized and uninitialized experiments is shown in Fig. 2 for 5- and 9-yr means. A persistence forecast is given as

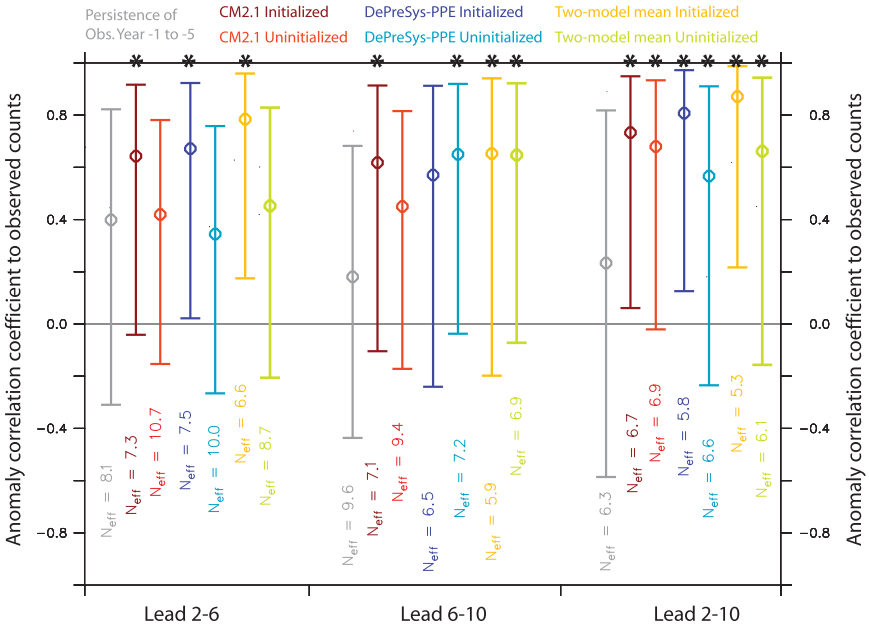

FIG. 2. Correlation for retrospective multiyear forecasts of North Atlantic hurricane frequency, with 90% uncertainty estimates. Each cluster of bars shows the retrospective correlation of multiyear hurricane frequency forecasts for leads of (left) 2–6, (middle) 6–10, and (right) 2–10 yr. Gray symbols indicate the correlation of the persistence of the 5-yr average count preceding the initialization of the model. Red symbols are for the GFDL-DecPre system, blue are for UKMO-DePreSys PPE, and yellow is for the two-model average. The initialized and uninitialized versions of each model are distinguished by different coloring. The sample correlation estimate is shown by the circle; the bars show the two-sided 90% uncertainty of a correlation given an underlying correlation with the value shown by the corresponding circle. An asterisk on top of a bar shows correlations that are significantly different from a null hypothesis of an underlying correlation of zero at $p = 0.1$, single-sided, with the effective degrees of freedom estimated as in Bretherton et al. (1999).

a reference test forecast, where the 5-yr (9-yr) mean persistence is defined as the observed average over the 5 (9) years that precede the model's initialization (persisting the SSTA indices does not improve the performance of the persistence null model, with correlations ranging between 0.16 and 0.4 depending on the SST dataset used). So, for example, the persistence forecast for the lead 2–6 forecast centered in 1992 (e.g., initialized in 1989) is the observed hurricane count averaged over 1984–88. Consistent with Fig. 1, at lead 2–6 the initialized retrospective predictions show higher correlations than the uninitialized ones, for both models. The values are significantly different from zero and exceed the values given by persistence, which is not the case for the uninitialized predictions. Comparable skill is found between the two models, slightly higher in UKMO-DePreSys; these retrospective correlations are comparable to those reported in S10 using an alternative methodology applied to DePreSys PPE. Computing the two-model mean increases the signal-to-noise ratio, leading to higher correlations than in either individual model. At lead 2–10, all the predictions outperform the persistence forecast. The decadal correlations are nominally higher in the initialized retrospective predictions than in the uninitialized, with the largest values, exceeding 0.8, occurring when taking the two-model mean. This decadal skill does not come only from the first few years because the correlations at lead 6–10 are also large (Fig. 2), although it is ambiguous if there is improvement from initialization. At lead 6–10, GFDL-DecPre shows larger correlations for the initialized predictions but UKMO-DePreSys indicates higher values for the uninitialized runs, yielding undistinguishable values between the initialized and the noninitialized experiments for the two-model mean.

These results suggest that coupled GCMs that account for both changes in initial state and radiative forcings can lead to skillful multiyear retrospective predictions of hurricane frequency. The nominal improvement from initialization should, however, be interpreted with care given the large confidence intervals associated with the point estimates of the correlations (Fig. 2). As discussed

above in section 2e, although the observed record is 50 years long, because of the large autocorrelation of the time series each year is not independent from those nearby. Hence, the effective number of degrees of freedom is largely reduced to less than 10 for most lead times, as indicated on Fig. 2, based on Bretherton et al. (1999). Therefore, even if the initialized predictions give a correlation that is statistically different from climatology and is nominally higher than in the uninitialized predictions, the large confidence intervals indicate that the retrospective correlation of the initialized forecasts is not different from persistence or the uninitialized experiments at $p = 0.1$. Some of the correlations of the initialized forecasts are significantly larger than the noninitialized experiments at $p = 0.2$.

The nonsignificance of the difference between the initialized and noninitialized correlations does not depend strongly on the effective sample size, as long as some level of autocorrelation is assumed. We recomputed the confidence intervals on the sample correlations using an unrealistic assumption that two years were needed for each new degree of freedom, and the initialized to uninitialized correlation differences were still not significantly different at $p = 0.1$. If we assume, even more unrealistically, that a new degree of freedom is achieved every 1.5 years, then the differences between the initialized and uninitialized experiments are significant at $p = 0.1$. However, we wish to stress that these perturbation experiments yield an extremely unrealistically high estimate of the number of degrees of freedom, considering we are exploring 5-yr running averages of quantities with a pronounced trend and interdecadal variation. The record is too short, and the difference between initialized and uninitialized correlations too small, to yield a statistically significant difference.

Improvement from initialization on the two-model mean lead 2–6 forecast is close to being significant even at $p = 0.1$, suggesting potentially higher confidence in multimodel ensembles. For the lead 2–6 and 2–10 forecasts, for both model systems there is a consistent nominal improvement of retrospective correlation from initialization relative to the uninitialized experiments. Because of this, and because of the small sample size, we speculate that the lack of significance at $p = 0.1$ may reflect a ''lack of power'' by the significance test, rather than a ''lack of effect'' from initializing (Johnson 1999). For the lead 6–10 forecast, however, the nominal difference between the initialized and noninitialized forecasts changes sign (there is nominal indication of improvement in GFDL-DecPre, but a nominal degradation in UKMO-DePreSys PPE), so we interpret the lack of significance in this case as indicating a lack of effect from initialization. Therefore, it appears that the nominal improvement in the lead 2–10 forecast arises in the first part of the decade and represents potential multiyear forecast skill rather than decadal skill.

A lagged-ensemble approach, in which past forecasts are used to augment the effective ensemble size of more recent forecasts (e.g., by creating a forecast where the current year's lead 1–5 and the previous year's lead 2–6 forecasts are averaged), can increase forecast performance [e.g., Vecchi et al. (2011) showed improvement in interannual hurricane forecasts from lagged ensembles]. We explored the impact of lagged ensembles in the retrospective hurricane forecasts (not shown) at lags of up to three years (i.e., averaging lead 1–5, 2–6, and 3–7 verifying the same years together), resulting in nominal improvements in the correlation coefficient (on the order of 0.02–0.05). However, the smoothing induced by the lagged ensemble led to a further reduction of degrees of freedom. Because the uncertainty in a correlation estimate increases with decreasing correlation or sample size, the uncertainty estimates on the correlation coefficient did not show substantial change: even after lagged-ensemble averaging, the retrospective correlation of the uninitialized and initialized forecasts were in each other's confidence intervals.

As a complement to the skill estimate using ACC, we show in Fig. 3 the MSSS for various 5- and 9-yr mean leads. Both the improvement relative to climatology [Eq. (4)] and that from initialization [Eq. (5)] are indicated on the x and y axes, respectively. None of the retrospective initialized forecasts has a negative MSSS on the x axis, which indicates at least a nominal improvement relative to climatology. An improvement from initialization is also suggested at all leads in GFDL-DecPre, and at most leads except for 5–9 and 6–10 in UKMO-DePreSys PPE, leading to a smaller MSSS at those lead times for the two-model mean. Both models indicate an improved skill at decadal scale from initialization, with the highest values in UKMO-DePreSys. As shown in Goddard et al. (2013), the MSSS is a function of both the correlation and the conditional bias, and the higher MSSS from initialization is mainly caused by a reduction of the conditional bias that is large in the uninitialized predictions.

b. SST source of hurricane forecast skill

Our hurricane frequency index is based on SST averaged over the tropical Atlantic and over the global tropics [Eq. (1)], so both quantities are potential sources for the better predictability in the initialized forecasts. We can explore retrospective forecasts and skill measures of these two indices with the hope of finding the role each had in recovering the past history of hurricane activity (Fig. 4). Overall, there is no indication that

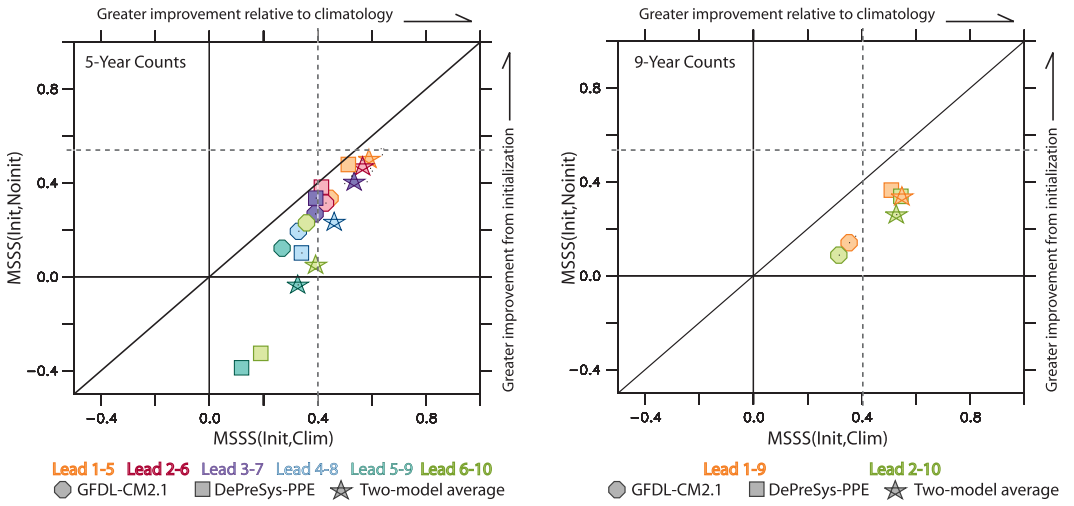

FIG. 3. The MSSS of retrospective initialized multiyear hurricane frequency forecasts for various leads and models, for (left) 5- and (right) 9-yr running-mean forecasts. The *x* axis shows the MSSS against climatology. The *y* axis shows the MSSS against the uninitialized forecasts. The diagonal line indicates the one-to-one line. Circles show the values for the GFDL-DecPre system, squares for UKMO-DePreSys PPE, and stars for the two-model ensemble mean. Different colors indicate different forecast leads.

retrospective forecasts of tropical-mean SST are improved by initializing the coupled GCMs (top, Fig. 4), with the relatively monotonic warming of the tropics dominating the observed and modeled signals. The dominance of the long-term trend in both SST indices cuts the effective degrees of freedom severely, to the point where for tropical-mean SST the interpretation of correlation as a skill metric is likely too ambiguous to be useful. The GFDL-DecPre system has marginally higher retrospective correlation in both SST indices than does UKMO-DePreSys, likely because of the inclusion of future volcanic information in its radiative forcing (Table 1). However, this nominally larger skill in GFDL-DecPre for the two SST indices does not translate into even a nominal increase of the hurricane forecasts (Fig. 2) because the volcanic signals are primarily spatially uniform. Across both model systems there is a consistent nominal improvement of retrospective correlation of Atlantic main development region (MDR) SST predictions from initialization, but the effect is small relative to the number of degrees of freedom. Only in the GFDL-DecPre does the initialized forecast of MDR SST approach a significant improvement over a persistence forecast. Because of the dominance of a quasi-monotonic trend, for tropical-mean SST all the forecast methods (initialized and uninitialized GCM forecasts and persistence) yield comparable results. For both SST indices all of the forecast methodologies lead to statistically significant retrospective correlations against a null of zero correlation, again largely because of the dominance of a trend.

The results in Fig. 4 suggest that the nominal improvement in retrospective correlation from initialization came from improvements to the forecast of Atlantic MDR SST. However, because the time series of each SST index includes a substantial component that is coherent across both indices, and because the hurricane frequency emulator is based on the difference between the two indices, interpreting the source of hurricane predictability from each index is not necessarily straightforward, as was noted in Vecchi et al. (2011). An alternative approach to assessing the influence of each index on the role of initialization on forecast skill is to use values of one index from the initialized experiments and the other from the uninitialized experiments. For example, taking values for SST_{MDR} from the initialized experiment, but keeping the SST_{TROP} from the uninitialized one, yields comparable hurricane retrospective forecast results (Fig. 5a) to when both indices are taken from the initialized experiments (Fig. 2). The impact of initialization on SST_{MDR} yields 5-yr mean fluctuations of this hurricane frequency index that show rather good agreement with observations for both models with a correlation of 0.70 and 0.59 in GFDL-DecPre and UKMO-DePreSys, respectively (both significantly different from zero correlation at $p < 0.05$) at lead 2–6. Using values for SST_{MDR} from the uninitialized experiments but those of SST_{TROP} from the initialized experiments leads to very different results (Fig. 5b). The correlation drops to 0.21 in GFDL-DecPre and to 0.43 in UKMO-DePreSys, with neither correlation significantly different from $\rho = 0$ (even at $p < 0.2$) and

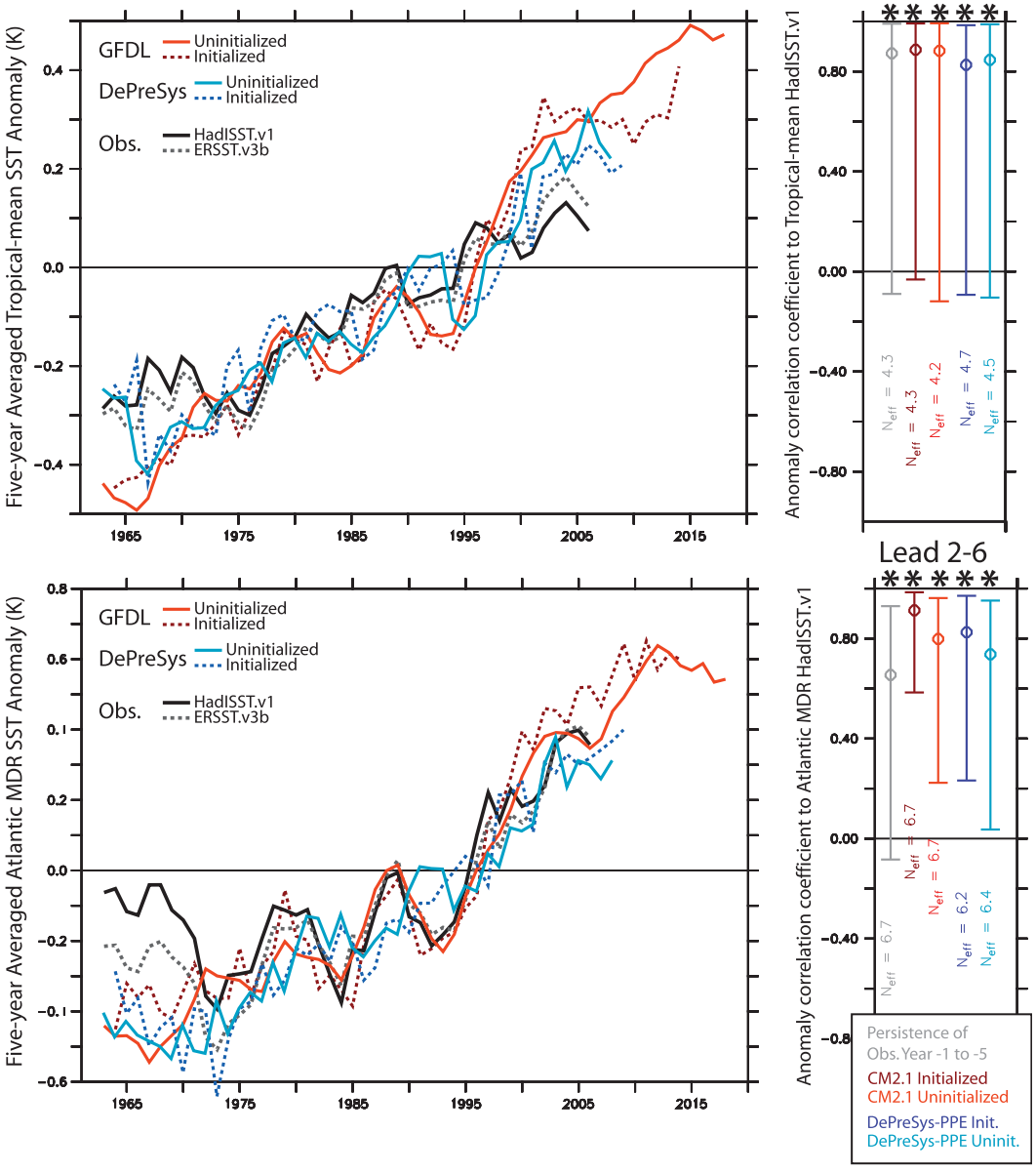

FIG. 4. Retrospective and future forecasts of the SST indices used for the hurricane emulator. (left) Time series of the 5-yr mean SSTAs averaged over the (top) global tropics and (bottom) Atlantic hurricane MDR, at lead 2–6. Black lines show observational estimates from the Hadley Centre Sea Ice and SST dataset, version 1 (HadISST.v1; Rayner et al. 2003; solid), and Extended Reconstructed SST, version 3b (ERSST.v3b; Smith et al. 2008; dotted). Colored lines show initialized (dashed) and uninitialized (solid) experiments from GFDL-DecPre (red) and UKMO-DePreSys PPE (blue). (right) Retrospective correlations of the forecasts at lead 2–6 against the HadISST.v1 SST product. In the right panels, an asterisk on top of a bar shows correlations that are significantly different from a null hypothesis of an underlying correlation of zero at p 5 0.1, single sided, with the effective degrees of freedom estimated as in Bretherton et al. (1999).

neither model able to reproduce the observed sharp increase in the mid-1990s. This indicates that the nominal improvement in correlation in the initialized multiyear predictions results from a better representation of the Atlantic main development region when initializing the coupled models, with little beneficial impact from initialized predictions of the global-mean tropical SST.

For the GFDL-DecPre system, the difference in retrospective correlation when swapping initialized/uninitialized SST_{MDR} and SST_{TROP} is significant at $p < 0.1$. Note in Fig. 5b that there is a large increase in hurricane frequency around 2005 in GFDL-DecPre, as in Fig. 1a. This increase, which we currently consider to be spurious, is a large contributor to the reduction in

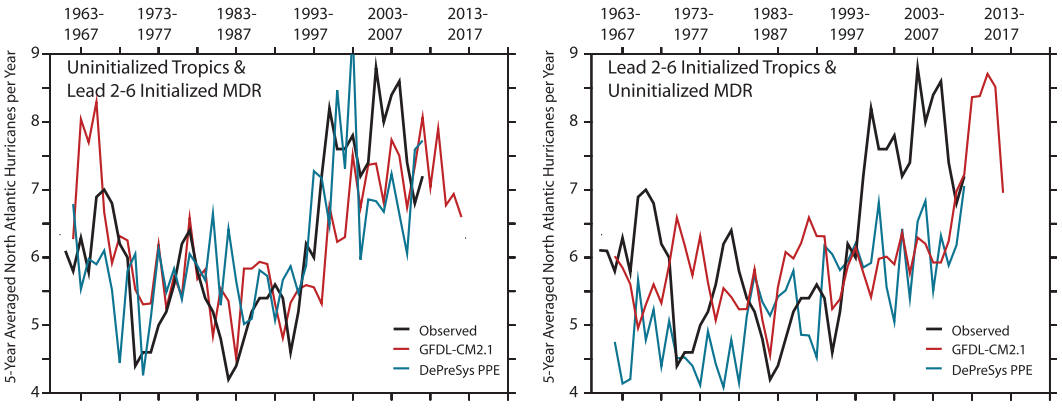

FIG. 5. Retrospective forecasts exploring the source of the initialized vs uninitialized components: (left) Atlantic MDR SST from initialized experiments and tropical-mean SST from uninitialized; (right) tropical-mean SST from initialized experiments and Atlantic MDR SST from uninitialized experiments. The skill comes from the improvement of tropical Atlantic SST in the initialized experiments.

correlation from the impact of initialization on tropical-mean SST in the GFDL model. There is a coincidence between the global implementation of the Array for Real-Time Geostrophic Oceanography (Argo) drifting float profiles in 2003 and the spurious shift of 9-yr forecasts centered around 2005/06, suggesting that enhanced observational sampling after 2003 may have led to a change in the lead-dependent climatology. Experiments are underway to test this possibility. The lack of such a spurious increase in UKMO-DePreSys could arise from different initialization processes, or from the fact that the last initialized forecast in UKMO-DePreSys begins in 2006—so the late spike would not be evident. Were the introduction of Argo found to be the driver of this spurious increase, in addition to developing methods to minimize the impact of observing system changes, the impact of other large changes to the observing system would also have to be explored (e.g., the introduction of altimetry in the early 1990s and the completion of the TAO array in the mid-1990s).

c. Role of the mid-1990s climate shift

The nominal improvement in skill from initialization should be interpreted with care. Even if the initialized retrospective predictions outperform climatology at almost all lead times (Fig. 3), the skill could still come from persistence—just persistence that cannot be captured with our observationally based persistence model. Figures 6a and 6b compare the retrospective predictions of hurricane frequency for 5-yr means ranging between leads 1–5 and 6–10. The forecasts at each lead show a tendency to have a systematic 1-yr shift with respect to the preceding lead, with the mid-1990s shift in each model trailing in time for longer leads rather than capturing the observed 1995 shift (e.g., Elsner et al. 2004; Li and Lund 2012) at the right time. By performing changepoint analysis (Pettitt test) on the models' retrospective predictions, we find a shift in forecasts initialized in 1991 in UKMO-DePreSys and forecasts initialized in 1995 in GFDL-DecPre. This tendency for forecasts to lock across the shift can be seen more clearly when the same time series are plotted as a function of initialization year instead of verification time (Figs. 6c,d): forecasts initialized the same year are very similar to each other, independent of when they verify. Notice that the mid-1990s shift for each model appears at the same initialization year for all lead times, as does the potentially spurious mid-2000s shift in GFDL-DecPre.

Up to now we have been largely comparing the results of forecasts initialized at different years at the same lead, without focusing on the evolution of hurricane counts of each forecast as the lead increases. A correct forecast of the mid-1990s climate shift would have indicated at some point prior to the shift that there was an increased probability of hurricane frequency increasing in time. For example, if a forecast initialized in early 1991 showed counts averaged in 1992–96 that were larger than those in 1991, or an increased number of ensemble members with large increases, one would have evidence for a future shift. Do these two forecast systems produce such a shift? Figure 7 shows that in the observational record, reflecting the rapid increase in frequency in 1995, the difference in hurricane counts averaged over the five years following the years 1991–94 exceeded the counts over each of those years by an unusually large amount, relative to the distribution over the 1961–2006 period. However, neither forecast system (colored lines in Fig. 7) shows a tendency for their forecasts to increase in time relative to the first forecast year when initialized in the early 1990s. In fact, there is

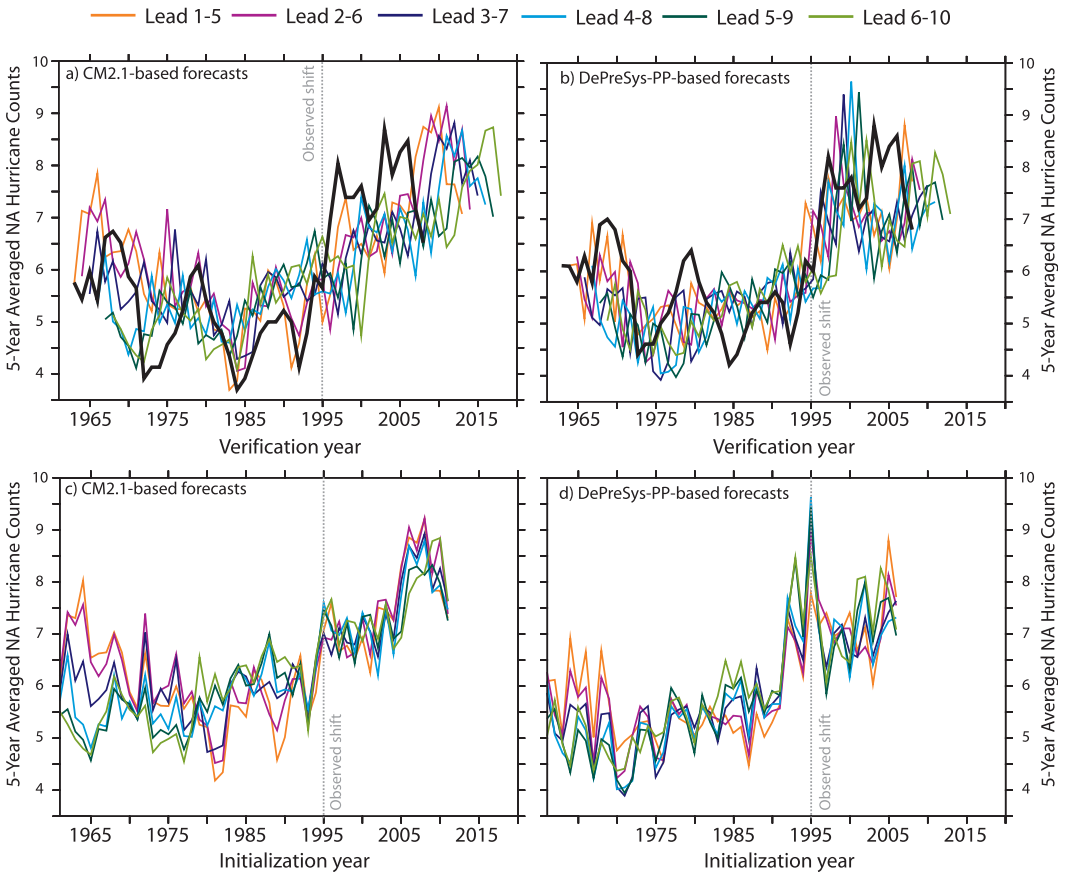

FIG. 6. Retrospective forecasts arranged by verification and initialization date. (a),(b) Retrospective forecasts of 5-yr running hurricane averages for various leads, arranged so that each point on the time axis corresponds to the midpoint of the 5-yr interval over which the average is computed (e.g., 1992 corresponds to the midpoint of the 1990–94 average). (c),(d) Retrospective 5-yr forecasts for various leads arranged so that each point on the time axis corresponds to the date in which the model was initialized. Data are from (left) GFDL CM2.1 forecasts and (right) the UKMO-DePreSys PPE system. Dark line in the top panels shows the observed 5-yr running counts.

a nominal tendency for these forecasts to decrease in time from the first forecast year, relative to the distribution of tendencies across all initialization dates, 1961–2006. That is, the models did not forecast a tendency toward higher frequency in the mid-1990s (Fig. 7), even though the sequence of forecast values exhibits a climate shift in the mid-1990s (Figs. 1 and 6).

To further highlight the influence of the mid-1990s shift on the retrospective skill estimation, we explore forecast performance after removing the mid-1990s shift from both the forecasts and the observations. The shift is "removed" by simply referencing each period before and after the 1994/95 shift to its own climatology; for instance, the time-mean hurricane count preceding 1995 is removed from all years before 1995, and the time-mean hurricane count following 1995 is removed from all years after 1995. We note that using each model's changepoint instead of 1995 does not affect the character of the results. Figures 8 and 9 indicate that removing the shift leads to a substantial reduction of correlation in the initialized predictions at lead 2–6 (particularly for UKMO-DePreSys PPE), and no indication of skill beyond that lead time, further confirming that the decadal signal is dominated by the trend that arises from the existence of the mid-1990s changepoint. Therefore, future real (as opposed to the retrospective forecasts explored here) multiyear and decadal predictions of hurricane frequency should not be expected to show the same skill as over the 1961–2011 period unless there are changepoints of similar character to the mid-1990s shift. Our results are encouraging for the feasibility of multiyear forecasts of hurricane frequency with the current prediction systems. However, this analysis highlights that substantial challenges remain—or, viewed more optimistically, that it is possible to improve the performance of the system beyond its current capability.

An interesting side effect of removing the mid-1990s shift is to increase the effective degrees of freedom,

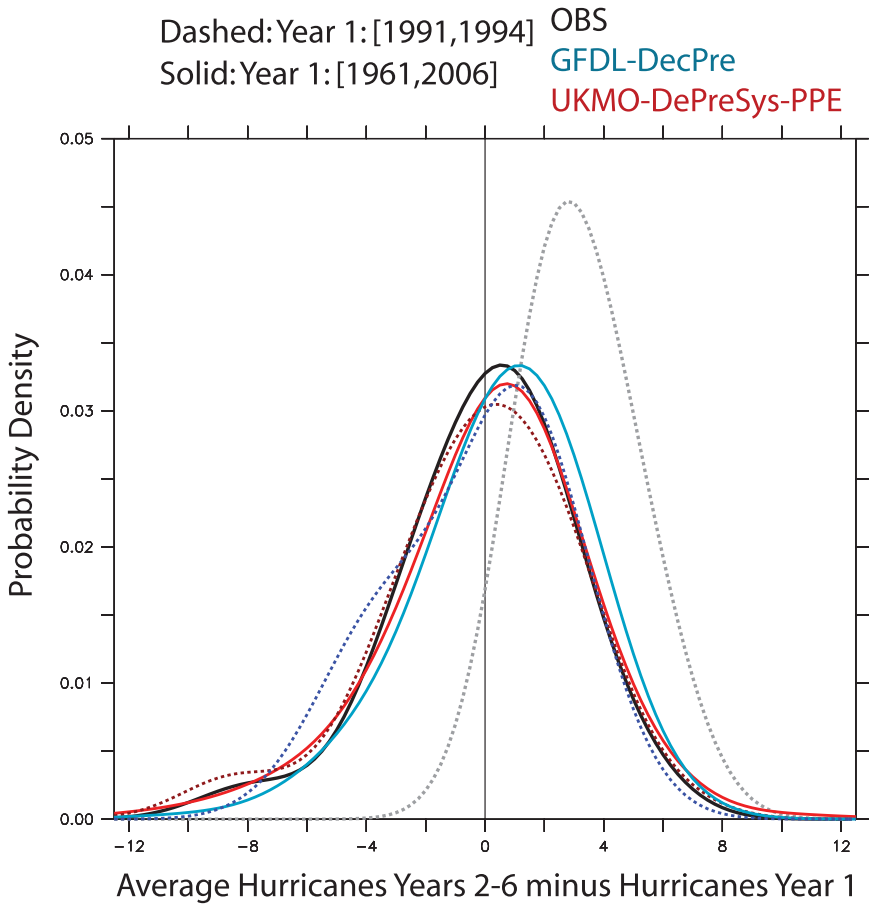

FIG. 7. Empirical PDF estimates for the change in seasonal hurricane counts over the entire record and over the four years that preceded the 1994/95 climate shift. The quantity explored is the difference in hurricane counts averaged over the five years following a given year with the counts of that year (e.g., for 1991 it is the difference of hurricane counts averaged 1992–96 with those in 1991); PDFs are estimated through Gaussian convolution with an *e*-folding scale of 2.5 hurricanes per year. Black lines are based on observations, blue lines on the forecasts with GFDL-DecPre, and red lines on the forecasts using UKMO-DePreSys; solid lines are computed over the 1961–2006 period and dashed lines over 1991–94. The separation of the solid and dashed black lines is a reflection of the increase in storm counts that occurred in 1995. Notice that there is no tendency for forecasts initialized in the early 1990s toward intensification through the early years of the forecast: the forecast systems do not dynamically predict the occurrence of the 1994/95 shift.

narrowing the confidence intervals associated with the point estimates of the correlation coefficient (cf. Figs. 2 and 9). In addition, the retrospective correlation in the uninitialized forecasts without changepoint disappeared—because it largely arose from the projection of the observed shift onto the models' forced trend over this period. In this modified context, there are now indications that for the GFDL model and the two-model ensemble the correlations (although lower than in the case including the shift; Fig. 2) are significantly higher than those of the uninitialized versions of the model at lead 2–6. That is, there is significant (at $p < 0.1$) indication that GFDL-DecPre and the two-model ensemble may be able to predict the types of variations in hurricane frequency that occurred in the early 1980s and early 1990s better than the uninitialized experiments. In Fig. 2, the nominal improvement from initialization in the correlation of the lead 2–6 and 6–10 mean hurricane counts in GFDL CM2.1 was larger than that for the lead 2–10 forecasts; this may reflect the ability of GFDL CM2.1 to retrospectively forecast some multiyear variations beyond the 1994/95 climate shift—which is the dominant signal in the 9-yr running counts. This further highlights the limitations of a data record that is short relative to the dominant time scales in order to assess the impact of multiyear forecast skill. While it is entirely possible that some of the non-significant differences between the initialized and uninitialized models shown in Figs. 2 and 3 could become

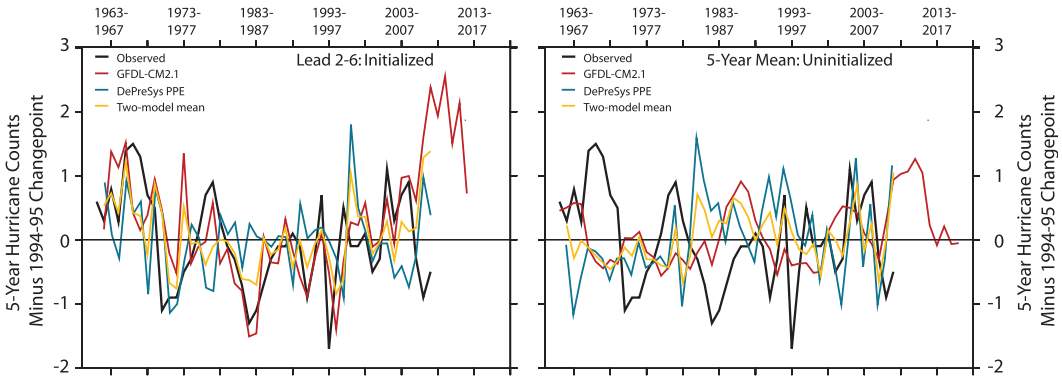

FIG. 8. Retrospective forecasts of North Atlantic hurricane frequency after removing the1994/95 shift in the mean from forecasts and verification (see section 3a): (left) initialized forecasts at lead 2–6 and (right) uninitialized experiments. Black line shows the observed counts, the red line is from the GFDL-DecPre system, the blue line is from UKMO-DePreSys PPE, and the yellow line is the two-system average, all after removing the 1994/95 shift in the mean.

significant from a longer record, it is also possible that the impact of initialization could also decrease and remain nonsignificant in a longer record.

4. Summary and discussion

Predictions of North Atlantic hurricane frequency were investigated in two global coupled models initialized toward estimates of the observed climate state. We find statistically significant retrospective correlation of multiyear to decadal initialized hurricane frequency forecasts by accounting for both initialization and radiative forcing changes. The two systems explored, GFDL-DecPre and UKMO-DePreSys PPE, show comparable skill. The two-model mean had the best skill, encouraging the pursuit of broader multimodel studies (e.g., Caron et al. 2013); lagged averages lead to nominal correlation increases. The retrospective correlations from initialized multiyear hurricane forecasts are comparable to those reported in S10 using an alternative methodology.

Taken together, our results and those of S10 indicate that initializing a climate model and accounting for radiative forcing changes, together, can lead to significant retrospective skill in multiyear hurricane forecasts (relative to climatological forecasts). The performance of the initialized forecasts was nominally better than that of uninitialized forecasts, both in correlation and in MSSS (Goddard et al. 2013). However, because of the short observational record and the persistent character of the time series, the confidence intervals associated with all the forecasts are large, and the difference between initialized and uninitialized forecasts is not statistically significant at $p = 0.1$ (although some are at $p = 0.2$). Because of the consistency of correlations across studies and the visual improvement, we hypothesize that lack of significant improvement from initialization may indicate of lack of ''power'' (i.e., the probability that the test will correctly reject the null hypothesis) by the statistical test (arising from too few degrees of freedom and a relatively strong correlation arising from radiative forcing alone) rather than a lack of effect of initialization (e.g., Johnson 1999). Additional years could lead to enhancement of our confidence; however, the large autocorrelation of the time series indicates that we require about seven years of data to gain a degree of freedom—so many years will be required to improve our confidence, even if we include the past 50 years in future estimates of forecast skill.

The observed time series of North Atlantic hurricane frequency is dominated by a strong and abrupt rise in 1995 leading to a trend over the 1961–2011 period. The high correlations of the retrospective predictions of North Atlantic hurricane frequency depend on the presence of this shift. While predictions from both models are for more hurricanes after the mid-1990s than before, the increase is not actually predicted by the evolution of the models, but is present in the initial state (i.e., forecasts initialized after the shift exhibited by each model remain high, but those initialized prior do not show the shift; Figs. 6 and 7). That is, the large retrospective skill estimates (Figs. 2 and 3) do not come from predicting the dynamical evolution of the climate system resulting in the hurricane frequency shift, but from ''recognizing'' that a climate shift has occurred and persisting that shift. This behavior mirrors experience in seasonal forecasts of El Niño, where transitions from climatological conditions to a warm ENSO state can be problematic to predict (e.g., Landsea and Knaff 2000; Vecchi et al. 2006), and successful forecasts often reflect the continued updating of subsurface conditions. This reduces our

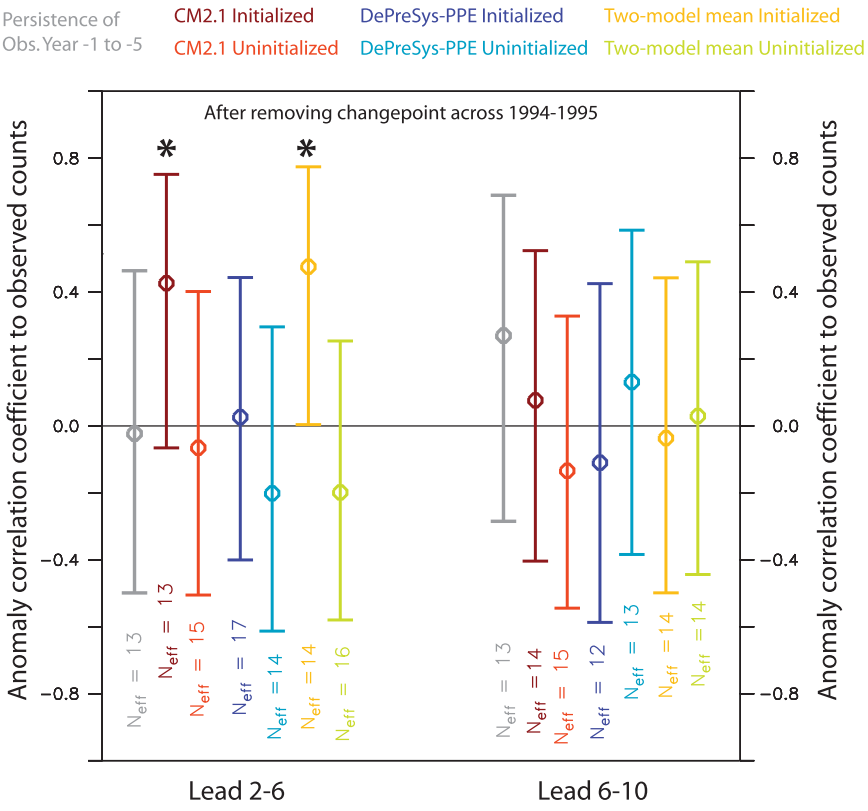

FIG. 9. Retrospective correlations of forecasts after removing 1994/95 shift in the mean from forecasts and verification. The gray symbol is the correlation of the persistence of the 5-yr average count preceding the initialization of the model. Red symbols are for the GFDL-DecPre system, blue are for UKMO-DePreSys PPE, and yellow are for the two-system average. The initialized and uninitialized versions of each model are distinguished by different coloring. The sample correlation estimate is shown by the circle; the bars show the two-sided 90% uncertainty of a correlation given an underlying correlation with the value shown by the corresponding circle. An asterisk on top of a bar shows correlations that are significantly different from a null hypothesis of an underlying correlation of zero at $p = 0.1$, single-sided, with the effective degrees of freedom estimated as in Bretherton et al. (1999).

confidence that the onset of a similar shift in the near future could be successfully predicted with current prediction systems. It also highlights the need to better understand the origin of the changepoint in the observations and to assess whether the modeled mechanisms are consistent with those in the real world (e.g., Robson et al. 2012).

Despite high correlation values, the mean retrospective skill of these forecasts may provide a poor and even misleading guide to the future performance. In the absence of a major climate shift, like the 1994/95 shift, the long-term estimates of correlation (e.g., 0.6–0.9) are not representative, and the lower retrospective correlations assessed after removing the shift (e.g., 0–0.4; Figs. 8 and 9) may be closer to those one should expect.

Neither model system successfully predicts the highest values of observed 5-yr hurricane frequency that appear in the mid-2000s. GFDL-DecPre shows a comparable rise but 5–10 years later than observed, whereas UKMO-DePreSys shows a more modest increase with a several-year delay as well. Forecasts with GFDL-DecPre that extend past the present suggest an increase in hurricane frequency through the mid-2010s (Fig. 1). However, observations have been tending in the opposite direction, with recent years being less active than those in the mid-2000s. This period coincides with a fundamental change in the ocean observing system, with the global introduction of Argo floats after 2003 bringing considerably better coverage of the surface and subsurface ocean. Changes in observing systems have previously impacted the behavior of initialized forecasts, in part by changing the character of the initialized model's drift (e.g., Kumar et al. 2012); therefore the introduction of Argo could impact the lead-dependent climatology.

Thus, we hypothesize that this increase predicted by with GFDL-DecPre is spurious and reflects the impact

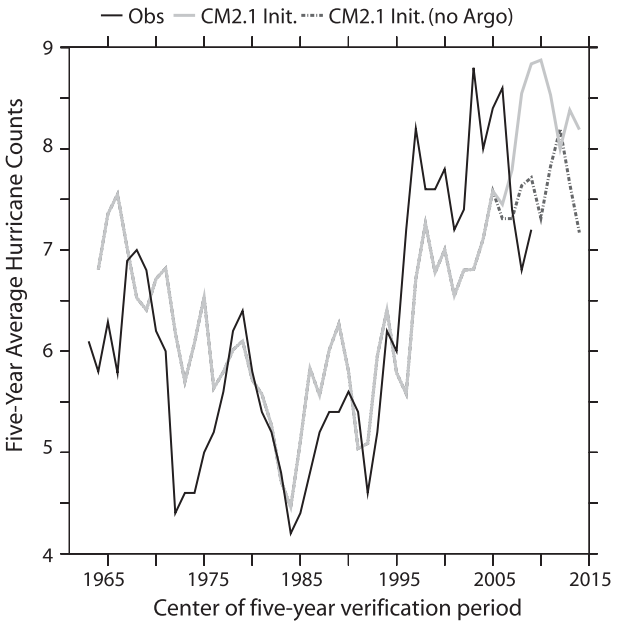

FIG. 10. Impact of Argo on retrospective and future forecasts of hurricane frequency using GFDL-DecPre. Lagged-ensemble (leads 1–5 and 2–6) forecasts of 5-yr Atlantic hurricane frequency based on the standard GFDL-DecPre system (gray line), and from a perturbation experiment in which forecasts initialized 2004 and later do not include data from Argo floats in the initialization (dashed line); black line shows observed 5-yr counts. A change in the drift of the initialized forecasts after the introduction of Argo leads to an increase in the predicted number of hurricanes after 2004.

of Argo data on the GFDL-DecPre drift. To test this hypothesis a set of experiments was performed in which Argo data were withheld from the initialization scheme of GFDL-DecPre after 2004. The predicted abrupt increase after 2004 is severely reduced when Argo is removed (Fig. 10), largely because of changes to model drift in regions that were poorly observed prior to Argo. These experiments support our hypothesis, so a more plausible prediction for the coming years is that shown in the left panel of Fig. 5, in which there is a tendency for relative stability to a reduction of hurricane frequency in coming years. Changes in drift (lead-dependent climatology) arising from the introduction of Argo impact the character of predictions of tropical- and global-mean temperature in the GFDL-DecPre system, leading to spuriously cold predictions of both if a single lead-dependent climatology is used to analyze the pre- and post-Argo period. We speculate that related errors may arise in other prediction systems from observing system changes. Methodologies to deal with the impact of observing system changes on drift must be developed in order to fully realize the potential of multiyear predictions; as the post-Argo record lengthens, motivated by Kumar et al. (2012), a potential solution is to use different lead-dependent climatologies for the pre- and post-Argo period. In addition, the impact of other observing system changes bear exploration, such as the introduction of the Pacific Tropical Atmosphere–Ocean moored buoy array in the early 1990s (McPhaden 1993) and expendable bathythermographs in the late 1960s. Interpretation of forecasts needs to be keenly constrained by our knowledge of changing observing practices both in the predictands (e.g., Vecchi and Knutson 2008, 2011; Landsea et al. 2010; Villarini et al. 2011a) and in the observations used to initialize the climate model (e.g., Zhang et al. 2007; Kumar et al. 2012).

Identifying the source of skill in retrospective predictions is key to the success of future forecasts. Recent studies (Mann and Emanuel 2006; Evan et al. 2009; S10; Villarini and Vecchi 2012b, 2013a) have argued that the recent (since the 1980s) increase of Atlantic hurricane activity was not caused by internal variability alone but also included an externally forced component driven largely by changing aerosol concentrations. Our results partially support this interpretation, indicating high correlations (significantly lead 2–10) in the uninitialized forecasts. Yet the sharp mid-1990s increase in Atlantic hurricane frequency is not retrospectively predicted in the uninitialized experiments. Its better representation in the initialized predictions could be interpreted as an indication of a key role for internal variability in the mid-1990s shift, supporting various studies (e.g., Zhang and Delworth 2005, 2006, 2009; Robson et al. 2012; Yeager et al. 2012; R. Msadek et al. 2013, unpublished manuscript). However, the nominal improvement from initialization could also reflect a failure in the radiative forcing/response in these models that is corrected when they are constrained with observations.

Our results indicate that the impact of initialization on forecasts of the Atlantic MDR relative to the tropics was key to the higher skill in the initialized forecasts (Figs. 4 and 5). Zhang and Delworth (2006) suggested that multiyear changes in hurricane activity could be driven by changes to the heat transport over the entire North Atlantic. S10 and Dunstone et al. (2011) further suggested that the subpolar North Atlantic was the main source of multiyear predictability of Atlantic hurricane frequency. The North Atlantic also stands out as the region where initialized forecasts outperform uninitialized ones in the GFDL model (A. Rosati et al. 2012, unpublished manuscript; Yang et al. 2012; R. Msadek et al. 2013, unpublished manuscript), suggesting a potential link between North Atlantic variability and Atlantic hurricane predictability in GFDL-DecPre. Further, Kang et al. (2008) showed that changes in the North Atlantic could lead to changes in atmospheric circulation over the tropical Atlantic in GFDL CM2.1. However, in our

retrospective forecasts of hurricane activity, the relevant source of skill must have been present in tropical Atlantic SST—so any role for extratropical forcing must involve a subsequent change to tropical Atlantic SST. Thus, improved representation of processes controlling tropical Atlantic climate (e.g., Doi et al. 2012) is key to enhanced skill in forecasts of hurricane activity by systems like those used here.

Acknowledgments. We are grateful to Doug Smith (UK Met Office) for making the UKMO-DePreSys PPE data available. We thank Ming Zhao and Tom Knutson for comments and suggestions.

REFERENCES

Alessandri, A., A. Borrelli, S. Gualdi, E. Scoccimarro, and S. Masina, 2011: Tropical cyclone count forecasting using a dynamical seasonal prediction system: Sensitivity to improved ocean initialization. *J. Climate,* **24,** 2963–2982.

Bender, M. A., T. R. Knutson, R. E. Tuleya, J. J. Sirutis, G. A. Vecchi, S. T. Garner, and I. M. Held, 2010: Modeled impact of anthropogenic warming on the frequency of intense Atlantic hurricanes. *Science,* **327,** 454–458.

Booth, B. B., N. J. Dunstone, P. R. Halloran, T. Andrews, and N. Bellouin, 2012: Aerosols implicated as a prime driver of twentieth-century North Atlantic climate variability. *Nature,* **484,** 228–232.

Bretherton, C. S., M. Widmann, V. P. Dymnikov, J. M. Wallace, and I. Bladé, 1999: The effective number of spatial degrees of freedom of a time-varying field. *J. Climate,* **12,** 1990–2009.

Broccoli, A. J., and S. Manabe, 1990: Can existing climate models be used to study anthropogenic changes in tropical cyclone climate? *Geophys. Res. Lett.,* **17,** 1917–1920.

Camargo, S. J., A. G. Barnston, P. Klotzbach, and C. W. Landsea, 2007a: Seasonal tropical cyclone forecasts. *WMO Bull.,* **56,** 297–309.

——, K. A. Emanuel, and A. H. Sobel, 2007b: Use of a genesis potential index to diagnose ENSO effects on tropical cyclone genesis. *J. Climate,* **20,** 4819–4834.

——, M. Ting, and Y. Kushnir, 2013: Influence of local and remote SST on North Atlantic tropical cyclone potential intensity. *Climate Dyn.,* **40** (5–6), 1515–1529, doi:10.1007/s00382-012-1536-4.

Caron, L. P., C. G. Jones, and F. Coblas-Reyes, 2013: Multi-year prediction skill of Atlantic hurricane activity in CMIP5 decadal hindcasts. *Climate Dyn.,* doi:10.1007/s00382-013-1773-1, in press.

Chang, C.-Y., J. C. H. Chiang, M. F. Wehner, A. Friedman, and R. Ruedy, 2011: Sulfate aerosol control of tropical Atlantic climate over the 20th century. *J. Climate,* **24,** 2540–2555.

Chang, Y.-S., S. Zhang, and A. Rosati, 2011: Improvement of salinity representation in an ensemble coupled data assimilation system using pseudo salinity profiles. *Geophys. Res. Lett.,* **38,** L13609, doi:10.1029/2011GL048064.

——, ——, ——, T. Delworth, and W. F. Stern, 2013: An assessment of oceanic variability for 1960–2010 from the GFDL ensemble coupled data assimilation. *Climate Dyn.,* **40,** 775–803.

Chen, J.-H., and S.-J. Lin, 2011: The remarkable predictability of inter-annual variability of Atlantic hurricanes during the past decade. *Geophys. Res. Lett.,* **38,** L11804, doi:10.1029/2011GL047629.

Chikamoto, Y., and Coauthors, 2013: An overview of decadal climate predictability in a multi-model ensemble by climate model MIROC. *Climate Dyn.,* **40,** 1201–1222, doi:10.1007/s00382-012-1351-y.

Collins, M., and Coauthors, 2006: Interannual to decadal climate predictability in the North Atlantic: A multimodel-ensemble study. *J. Climate,* **19,** 1195–1203.

Delworth, T. L., and Coauthors, 2006: GFDL's CM2 global coupled climate models. Part I: Formulation and simulation characteristics. *J. Climate,* **19,** 643–674.

Doi, T., G. A. Vecchi, A. J. Rosati, and T. L. Delworth, 2012: Biases in the Atlantic ITCZ in seasonal–interannual variations for a coarse- and a high-resolution coupled climate model. *J. Climate,* **25,** 5494–5511.

Dunstone, N. J., D. M. Smith, and R. Eade, 2011: Multi-year predictability of the tropical Atlantic atmosphere driven by the high latitude North Atlantic Ocean. *Geophys. Res. Lett.,* **38,** L14701, doi:10.1029/2011GL047949.

Elsner, J. B., and T. H. Jagger, 2006: Prediction models for annual U.S. hurricane counts. *J. Climate,* **19,** 2935–2952.

——, X. Niu, and T. H. Jagger, 2004: Detecting shifts in hurricane rates using a Markov chain Monte Carlo approach. *J. Climate,* **17,** 2652–2666.

Emanuel, K. A., 1987: The dependence of hurricane intensity on climate. *Nature,* **326,** 483–485.

——, 2005: Increasing destructiveness of tropical cyclones over the past 30 years. *Nature,* **436,** 686–688.

——, 2007: Environmental factors affecting tropical cyclone power dissipation. *J. Climate,* **20,** 5497–5509.

——, R. Sundararajan, and J. Williams, 2008: Hurricanes and global warming—Results from downscaling IPCC AR4 simulations. *Bull. Amer. Meteor. Soc.,* **89,** 347–367.

Evan, A. T., D. J. Vimont, A. K. Heidinger, J. P. Kossin, and R. Bennartz, 2009: The role of aerosols in the evolution of tropical North Atlantic Ocean temperature anomalies. *Science,* **324,** 778–781.

Fisher, R. A., 1915: Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika,* **10,** 507–521.

——, 1924: The distribution of the partial correlation coefficient. *Metron,* **3,** 329–332.

Goddard, L., and Coauthors, 2013: A verification framework for interannual-to-decadal predictions experiments. *Climate Dyn.,* **40,** 245–272.

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

Gordon, C., C. Cooper, C. Senior, H. Banks, J. Gregory, T. Johns, J. Mitchell, and R. Wood, 2000: The simulation of SST, sea ice extents and ocean heat transports in a version of the Hadley Centre coupled model without flux adjustments. *Climate Dyn.,* **16,** 147–168.

Gray, W. M., 1984: Atlantic seasonal hurricane frequency. Part I: El Niño and 30 mb quasi-biennial oscillation influences. *Mon. Wea. Rev.,* **112,** 1649–1668.

Griffies, S. M., and K. Bryan, 1997a: Predictability of North Atlantic multidecadal climate variability. *Science,* **275,** 181–184.

——, and ——, 1997b: A predictability study of simulated North Atlantic multidecadal variability. *Climate Dyn.,* **13,** 459–487.

Gualdi, S., E. Scoccimarro, and A. Navarra, 2008: Changes in tropical cyclone activity due to global warming: Results from a high-resolution coupled general circulation model. *J. Climate,* **21,** 5204–5228.

Hawkins, E., and R. Sutton, 2009: The potential to narrow uncertainty in regional climate predictions. *Bull. Amer. Meteor. Soc.,* **90,** 1095–1107.

ICPO, 2011: Data and bias correction for decadal climate predictions. International CLIVAR Project Office (ICPO), CLIVAR Publication Series No. 150, 6 pp. [Available online at http://www.wcrp-climate.org/decadal/index.shtml.]

Jarvinen, B. R., C. J. Neumann, and M. A. S. Davis, 1984: A tropical cyclone data tape for the North Atlantic Basin, 1886–1983: Contents, limitations, and uses. NOAA Tech. Memo. NWS NHC 22, 24 pp.

Johnson, D. H., 1999: The insignificance of significance testing. *J. Wildl. Manage.,* **63,** 763–772.

Johnson, N. L., S. Kotz, and N. Balakrishnan, 1995: *Continuous Univariate Distributions.* Vol. 2. Wiley, 752 pp.

Kang, S. M., I. M. Held, D. M. W. Frierson, and M. Zhao, 2008: The response of the ITCZ to extratropical thermal forcing: Idealized slab-ocean experiments with a GCM. *J. Climate,* **21,** 3521–3532.

Kim, H.-M., and P. J. Webster, 2010: Extended-range seasonal hurricane forecasts for the North Atlantic with a hybrid dynamical-statistical model. *Geophys. Res. Lett.,* **37,** L21705, doi:10.1029/2010GL044792.

Klotzbach, P. J., and W. M. Gray, 2009: Twenty-five years of Atlantic basin seasonal hurricane forecasts. *Geophys. Res. Lett.,* **36,** L09711, doi:10.1029/2009GL037580.

Knight, J. R., R. J. Allan, C. K. Folland, M. Vellinga, and M. E. Mann, 2005: A signature of persistent natural thermohaline circulation cycles in observed climate. *Geophys. Res. Lett.,* **32,** L20708, doi:10.1029/2005GL024233.

Knutson, T. R., and R. E. Tuleya, 2004: Impact of CO_2-induced warming on simulated hurricane intensity and precipitation: Sensitivity to the choice of climate model and convective parameterization. *J. Climate,* **17,** 3477–3495.

——, J. J. Sirutis, S. T. Garner, G. A. Vecchi, and I. Held, 2008: Simulated reduction in Atlantic hurricane frequency under twenty-first-century warming conditions. *Nat. Geosci.,* **1,** 359–364; Corrigendum, **1,** 479.

——, and Coauthors, 2010: Tropical cyclones and climate change. *Nat. Geosci.,* **3,** 157–163.

——, and Coauthors, 2013: Dynamical downscaling projections of late twenty-first century Atlantic hurricane activity: CMIP3 and CMIP5 model-based scenarios. *J. Climate,* in press.

Kumar, A., M. Chen, L. Zhang, W. Wang, Y. Xue, C. Wen, L. Marx, and B. Huang, 2012: An analysis of the nonstationarity in the bias of sea surface temperature forecasts for the NCEP Climate Forecast System (CFS) version 2. *Mon. Wea. Rev.,* **140,** 3003–3016.

Landsea, C. W., and J. A. Knaff, 2000: How much skill was there in forecasting the very strong 1997–98 El Niño? *Bull. Amer. Meteor. Soc.,* **81,** 2107–2119.

——, G. A. Vecchi, L. Bengtsson, and T. R. Knutson, 2010: Impact of duration thresholds on Atlantic tropical cyclone counts. *J. Climate,* **23,** 2508–2519.

LaRow, T. E., L. Stefanova, D. W. Shin, and S. Cocke, 2010: Seasonal Atlantic tropical cyclone hindcasting/forecasting using two sea surface temperature datasets. *Geophys. Res. Lett.,* **37,** L02804, doi:10.1029/2009GL041459.

Latif, M., N. Keenlyside, and J. Bader, 2007: Tropical sea surface temperature, vertical wind shear, and hurricane development. *Geophys. Res. Lett.,* **34,** L01710, doi:10.1029/2006GL027969.

Li, S., and R. Lund, 2012: Multiple changepoint detection via genetic algorithms. *J. Climate,* **25,** 674–686.

MacAdie, C. J., C. W. Landsea, C. J. Neumann, J. E. David, E. Blake, and G. R. Hammer, 2009: Tropical cyclones of the North Atlantic Ocean, 1851–2006. NCDC Tech. Memo., 238 pp. [Available online at http://www.nhc.noaa.gov/abouttrackbooks.shtml and from the National Climatic Data Center, 151 Patton Ave., Room 120, Asheville, NC 28801-5001.]

Mann, M. E., and K. A. Emanuel, 2006: Atlantic hurricane trends linked to climate change. *Eos, Trans. Amer. Geophys. Union,* **87,** 233–241, doi:10.1029/2006EO240001.

McPhaden, M. J., 1993: TOGA–TAO and the 1991–93 El Niño–Southern Oscillation event. *Oceanography,* **6,** 36–44.

Meehl, G., and Coauthors, 2013: Decadal climate prediction: An update from the trenches. *Bull. Amer. Meteor. Soc.,* in press.

Mendelsohn, R., K. Emanuel, S. Chonabayashi, and L. Bakkensen, 2012: The impact of climate change on global tropical cyclone damage. *Nat. Climate Change,* **2,** 205–209.

Msadek, R., K. W. Dixon, T. L. Delworth, and W. Hurlin, 2010: Assessing the predictability of the Atlantic meridional overturning circulation and associated fingerprints. *Geophys. Res. Lett.,* **37,** L19608, doi:10.1029/2010GL044517.

Murphy, A. H., 1988: Skill scores based on the mean squared error and their relationships to the correlation coefficient. *Mon. Wea. Rev.,* **116,** 2417–2424.

Oouchi, K., J. Yoshimura, H. Yoshimura, R. Mizuta, S. Kusumoki, and A. Noda, 2006: Tropical cyclone climatology in a global warming climate as simulated in a 20-km-mesh global atmospheric model: Frequency and wind intensity analysis. *J. Meteor. Soc. Japan,* **84,** 259–276.

Peduzzi, P., B. Chatenoux, H. Dao, A. De Bono, C. Herold, J. Kossin, F. Mouton, and O. Nordbeck, 2012: Global trends in tropical cyclone risk. *Nat. Climate Change,* **2,** 289–294.

Pielke, R. A., Jr., J. Gratz, C. W. Landsea, D. Collins, M. A. Saunders, and R. Musulin, 2008: Normalized hurricane damages in the United States: 1900–2005. *Nat. Hazards Rev.,* **9,** 29–42.

Pohlmann, H., M. Botzet, M. Latif, A. Roesch, M. Wild, and P. Tschuck, 2004: Estimating the decadal predictability of a coupled AOGCM. *J. Climate,* **17,** 4463–4472.

——, J. H. Jungclaus, A. Köhl, D. Stammer, and J. Marotzke, 2009: Initializing decadal climate predictions with the GECCO oceanic synthesis: Effects on the North Atlantic. *J. Climate,* **22,** 3926–3938.

Ramsay, H. A., and A. H. Sobel, 2011: Effects of relative and absolute sea surface temperature on tropical cyclone potential intensity using a single-column model. *J. Climate,* **24,** 183–193.

Rayner, N. A., D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell, E. C. Kent, and A. Kaplan, 2003: Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res.,* **108,** 4407, doi:10.1029/2002JD002670.

Robson, J., 2011: Understanding the performance of a decadal prediction system. Ph.D. thesis, University of Reading, 233 pp. [Available online at http://www.met.reading.ac.uk/~swr06jir/thesis/JIR_thesis.pdf.]

——, R. Sutton, K. Lohmann, D. Smith, and M. Palmer, 2012: Causes of the rapid warming of the North Atlantic Ocean in the mid-1990s. *J. Climate,* **25,** 4116–4134.

Rotstayn, L. D., and U. Lohmann, 2002: Tropical rainfall trends and the indirect aerosol effect. *J. Climate,* **15,** 2103–2116.

Shen, W., R. E. Tuleya, and I. Ginis, 2000: A sensitivity study of the thermodynamic environment on GFDL model hurricane intensity: Implications for global warming. *J. Climate,* **13,** 109–121.

Smith, D. M., and J. Murphy, 2007: An objective ocean temperature and salinity analysis using covariances from a global climate model. *J. Geophys. Res.,* **112,** C02022, doi:10.1029/2005JC003172.

——, S. Cusack, A. W. Colman, C. K. Folland, G. R. Harris, and J. M. Murphy, 2007: Improved surface temperature prediction for the coming decade from a global climate model. *Science,* **317,** 796–799.

——, R. Eade, N. J. Dunstone, D. Fereday, J. M. Murphy, H. Pohlmann, and A. A. Scaife, 2010: Skillful multi-year predictions of Atlantic hurricane frequency. *Nat. Geosci.,* **3,** 846–849.

Smith, T. M., R. W. Reynolds, T. C. Peterson, and J. Lawrimore, 2008: Improvement to NOAA's historical merged land–ocean surface temperature analysis (1880–2006). *J. Climate,* **21,** 2283–2296.

Sobel, A. H., and C. S. Bretherton, 2000: Modeling tropical precipitation in a single column. *J. Climate,* **13,** 4378–4392.

——, I. M. Held, and C. S. Bretherton, 2002: The ENSO signal in tropical tropospheric temperature. *J. Climate,* **15,** 2702–2706.

Stockdale, T. N., 1997: Coupled ocean–atmosphere forecasts in the presence of climate drift. *Mon. Wea. Rev.,* **125,** 809–818.

Sugi, M., H. Murakami, and J. Yoshimura, 2009: A reduction in global tropical cyclone frequency due to global warming. *SOLA,* **5,** 164–167.

——, ——, and ——, 2012: On the mechanism of tropical cyclone frequency changes due to global warming. *J. Meteor. Soc. Japan,* **90A,** 397–408.

Swanson, K. L., 2007: Impact of scaling behavior on tropical cyclone intensities. *Geophys. Res. Lett.,* **34,** L18815, doi:10.1029/2007GL030851.

——, 2008: Nonlocality of Atlantic tropical cyclone intensities. *Geochem. Geophys. Geosyst.,* **9,** Q04V01, doi:10.1029/2007GC001844.

Tang, B. H., and J. D. Neelin, 2004: ENSO influence on Atlantic hurricanes via tropospheric warming. *Geophys. Res. Lett.,* **31,** L24204, doi:10.1029/2004GL021072.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2012: An overview of CMIP5 and the experiment design. *Bull. Amer. Meteor. Soc.,* **93,** 485–498.

Teng, H., G. Branstator, and G. A. Meehl, 2011: Predictability of the Atlantic overturning circulation and associated surface patterns in two CCSM3 climate change ensemble experiments. *J. Climate,* **24,** 6054–6076.

van Oldenborgh, G. J., F. J. Doblas-Reyes, B. Wouters, and W. Hazeleger, 2012: Decadal prediction skill in a multi-model ensemble. *Climate Dyn.,* **38,** 1263–1280.

Vecchi, G. A., and B. J. Soden, 2007a: Effect of remote sea surface temperature change on tropical cyclone potential intensity. *Nature,* **450,** 1066–1071.

——, and ——, 2007b: Global warming and the weakening of the tropical circulation. *J. Climate,* **20,** 4316–4340.

——, and T. R. Knutson, 2008: On estimates of historical North Atlantic tropical cyclone activity. *J. Climate,* **21,** 3580–3600.

——, and ——, 2011: Estimating annual numbers of Atlantic hurricanes missing from the HURDAT database (1878–1965) using ship track density. *J. Climate,* **24,** 1736–1746.

——, A. T. Wittenberg, and A. Rosati, 2006: Reassessing the role of stochastic forcing in the 1997–1998 El Niño. *Geophys. Res. Lett.,* **33,** L01706, doi:10.1029/2005GL024738.

——, K. L. Swanson, and B. J. Soden, 2008: Whither hurricane activity? *Science,* **322,** 687–689.

——, 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.

——, S. Fueglistaler, I. M. Held, T. R. Knutson, and M. Zhao, 2013: Impacts of atmospheric temperature trends on tropical cyclone activity. *J. Climate,* **26,** 3877–3891.

Villarini, G., and G. A. Vecchi, 2012a: North Atlantic Power Dissipation Index (PDI) and accumulated cyclone energy (ACE): Statistical modeling and sensitivity to sea surface temperature changes. *J. Climate,* **25,** 625–637.

——, and ——, 2012b: Twenty-first-century projections of North Atlantic tropical storms from CMIP5 models. *Nat. Climate Change,* **2,** 604–607, doi:10.1038/nclimate1530.

——, and ——, 2013a: Projected increases in North Atlantic tropical cyclone intensity from CMIP5 models. *J. Climate,* **26,** 3231–3240.

——, and ——, 2013b: Multi-season lead forecast of the North Atlantic Power Dissipation Index (PDI) and accumulated cyclone energy (ACE). *J. Climate,* **26,** 3631–3643.

——, ——, and J. A. Smith, 2010: Modeling the dependence of tropical storm counts in the North Atlantic basin on climate indices. *Mon. Wea. Rev.,* **138,** 2681–2705.

——, ——, T. R. Knutson, and J. A. Smith, 2011a: Is the recorded increase in short duration North Atlantic tropical storms spurious? *J. Geophys. Res.,* **116,** D10114, doi:10.1029/2010JD015493.

——, ——, ——, M. Zhao, and J. A. Smith, 2011b: North Atlantic tropical storm frequency response to anthropogenic forcing: Projections and sources of uncertainty. *J. Climate,* **24,** 3224–3238.

——, ——, and J. A. Smith, 2012: U.S. landfalling and North Atlantic hurricanes: Statistical modeling of their frequencies and ratios. *Mon. Wea. Rev.,* **140,** 44–65.

Vitart, F., 2006: Seasonal forecasting of tropical storm frequency using a multi-model ensemble. *Quart. J. Roy. Meteor. Soc.,* **132,** 647–666.

——, M. Huddleston, D. Deque, T. Palmer, T. Stockdale, M. Davey, S. Ineson, and A. Weisheimer, 2007: Dynamically-based seasonal forecasts of Atlantic tropical storm activity issued in June by EUROSIP. *Geophys. Res. Lett.,* **34,** L16815, doi:10.1029/2007GL030740.

von Storch, H., and F. W. Zwiers, 1999: *Statistical Analysis in Climate Research.* Cambridge University Press, 484 pp.

Wang, H., J. K. E. Schemm, A. Kumar, W. Wang, L. Long, M. Chelliah, G. D. Bell, and P. Peng, 2009: A statistical forecast model for Atlantic seasonal hurricane activity based on the NCEP dynamical seasonal forecast. *J. Climate,* **22,** 4481–4500.

Xie, S. P., C. Deser, G. A. Vecchi, J. Ma, H. Teng, and A. T. Wittenberg, 2010: Global warming pattern formation: Sea surface temperature and rainfall. *J. Climate,* **23,** 966–986.

Yang, X., and Coauthors, 2013: A predictable AMO-like pattern in GFDL's fully-coupled ensemble initialization and decadal forecasting system. *J. Climate,* **26,** 650–661.

Yeager, S., A. Karspeck, G. Danabasoglu, J. Tribbia, and H. Teng, 2012: A decadal prediction case study: Late twentieth-century North Atlantic Ocean heat content. *J. Climate,* **25,** 5173–5189.

Zhang, R., and T. L. Delworth, 2005: Simulated tropical response to a substantial weakening of the Atlantic thermohaline circulation. *J. Climate,* **18,** 1853–1860.

——, and ——, 2006: Impact of Atlantic multidecadal oscillations on India/Sahel rainfall and Atlantic hurricanes. *Geophys. Res. Lett.,* **33,** L17712, doi:10.1029/2006GL026267.

——, and ——, 2009: A new method for attributing climate variations over the Atlantic hurricane basin's main development region. *Geophys. Res. Lett.,* **36,** L06701, doi:10.1029/2009GL037260.

——, and Coauthors, 2013: Have aerosols caused the observed Atlantic multidecadal variability? *J. Atmos. Sci.,* **70,** 1135–1144.

Zhang, S., and A. Rosati, 2010: An inflated ensemble filter for ocean data assimilation with a biased coupled GCM. *Mon. Wea. Rev.,* **138,** 3905–3931.

——, M. J. Harrison, A. Rosati, and A. T. Wittenberg, 2007: System design and evaluation of coupled ensemble data assimilation for global oceanic climate studies. *Mon. Wea. Rev.,* **135,** 3541–3564.

Zhao, M., and I. M. Held, 2011: The response of tropical cyclone statistics to an increase in CO_2 with fixed sea surface temperatures. *J. Climate,* **24,** 5353–5364.

——, ——, S.-J. Lin, and G. A. Vecchi, 2009: Simulations of global hurricane climatology, interannual variability, and response to global warming using a 50-km resolution GCM. *J. Climate,* **22,** 6653–6678.

——, ——, and G. A. Vecchi, 2010: Retrospective forecasts of the hurricane season using a global atmospheric model assuming persistence of SST anomalies. *Mon. Wea. Rev.,* **138,** 3858–3868.