Clim Dyn (2014) 42:2675–2690
DOI 10.1007/s00382-013-1773-1

Multi-year prediction skill of Atlantic hurricane activity in CMIP5 decadal hindcasts

Louis-Philippe Caron · Colin G. Jones · Francisco Doblas-Reyes

Received: 9 January 2013 / Accepted: 15 April 2013 / Published online: 30 April 2013

Abstract Using a statistical relationship between simulated sea surface temperature and Atlantic hurricane activity, we estimate the skill of a CMIP5 multi-model ensemble at predicting multi-annual level of Atlantic hurricane activity. The series of yearly-initialized hindcasts show positive skill compared to simpler forecasts such as persistence and climatology as well as non-initialized forecasts and return anomaly correlation coefficients of $\sim$0.6 and $\sim$0.8 for five and nine year forecasts, respectively. Some skill is shown to remain in the later years and making use of those later years to create a lagged-ensemble yields, for individual models, results that approach that obtained by the multi-model ensemble. Some of the skill is shown to come from persisting rather than predicting the climate shift that occur in 1994–1995. After accounting for that shift, the anomaly correlation coefficient for five-year forecasts is estimated to drop to 0.4, but remains statistically significant up to lead years 3–7. Most of the skill is shown to come from the ability of the forecast systems at capturing change in Atlantic sea surface temperature, although the failure of most systems at reproducing the observed slow down in warming over the tropics in recent years leads to an underestimation of hurricane activity in the later period.

Keywords Decadal climate prediction · Multi-model ensemble · Forecast · Atlantic variability · Hurricane activity

L.-P. Caron (✉)
Department of Meteorology, Stockholm University, Stockholm, Sweden
e-mail: caron@misu.su.se

L.-P. Caron
Bert Bolin Centre for Climate Research, Stockholm University, Stockholm, Sweden

C. G. Jones
Rossby Centre, Swedish Meterological and Hydrological Institute, Norrköping, Sweden

F. Doblas-Reyes
Institut Català de Ciències del Clima (IC3), Barcelona, Spain

F. Doblas-Reyes
Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

1 Introduction

Near term climate prediction is a new and rapidly developing field of climate research (Smith et al. 2007; Meehl et al. 2013; Solomon et al. 2011) made possible by the increasing amount and quality of ocean surface and subsurface observations, by the computing power now available and by the significant improvement in the quality of global climate models (GCMs). Also known as decadal prediction, it aims to fill the gap in available climate information between seasonal predictions (a few months to one year) and climate change projections (>10 years), a timescale over which the influence of internal climate variability is comparable or larger to that associated with changes in radiative forcing (Hawkins and Sutton, 2009). Such forecasts are of particular interest to policy makers, due to their potential use in a range of weather-related economic activities, such as agriculture and energy planning (Mendelsohn et al. 2007; Khasnis and Nettleman 2005; Brayshaw et al. 2011).

This growing interest is also reflected in a number of new international scientific initiatives, in particular the Coupled Model Intercomparison Project Phase 5 (CMIP5),

which has a specific set of subprojects targeting decadal prediction (Taylor et al. 2012). This initiative will allow predictability and prediction skill of different forecast systems, which to date were analysed independently, to be commonly evaluated using identical criteria (Doblas-Reyes et al. 2012). For this work, we make use of the largest database of decadal hindcasts currently available to investigate and compare the ability of these hindcasts in reproducing the mean level and variability of Atlantic hurricane activity over the recent past.

Modulation of Atlantic tropical cyclone activity at the decadal timescale has been linked to slow variations in SSTs over the northern Atlantic (Zhang and Delworth 2006; Knight et al. 2006). Since forecast systems have shown some skill in predicting these SST changes (García-Serrano et al. 2012), one could reasonably expect potential predictability in associated changes in tropical cyclone activity. Already, by tracking tropical cyclone-like systems in DePreSys (Smith et al. 2007), Smith et al. (2010) found some predictability in Atlantic cyclone numbers at the multi-annual level, which they then linked to predictability in SSTs in the northern North Atlantic. Dunstone et al. (2011) provided further evidence, via data withholding experiments, for a pivotal role of the extra-tropical sub-polar gyre region in driving variability in multi-annual Atlantic tropical storms. The results of Smith et al. (2010) were supported by Vecchi et al. (2013) who inferred the number of Atlantic hurricanes in two forecast systems through an observed statistical relationship between SSTs and Atlantic storms. A multi-model approach similar to Smith et al. (2010) being computationally prohibitive, we extend the Vecchi et al. (2013) study to include all the CMIP5 forecast systems currently available to obtain the most robust estimates of predictability of multi-annual Atlantic hurricane activity.

Section 2 gives a short description of the models used in this study, the statistical relation used to infer hurricane numbers as well as the statistical tools used to evaluate the quality of the predictions. Section 3 describes the level of skill reached by the different forecast systems and how that level is modified by different factors. The paper ends with a discussion of the results in Sect. 4.

2 Methodology

2.1 Data and models

The following analysis relies on two different multi-model ensembles of simulations performed within the context of the CMIP5 project (Taylor et al. 2012). The first ensemble is comprised of five GCMs (see Table 1 for more information on the forecast systems), each of which is represented through a series of 10-year long simulations initialized using climate state from either November 1st or January 1st estimated using contemporaneous observations. Although the number of members varies from model to model, each GCM has produced multiple decadal forecasts initialized every year between 1960 and 2005 (see Table 2 for the number of hindcasts performed by each CGCM). All of these hindcasts take into account observed changes in external forcings such as greenhouse gases, solar activity, stratospheric aerosols associated with volcanic eruptions and anthropogenic aerosols until 2005, and values from the Representative Concentration Pathway (RCP) 4.5 scenario (Meinshausen et al. 2011) afterwards. This ensemble will henceforth be referred to as Ini1. It is important to point out that a true forecast could not take into account the observed changes in external forcings, but could only use a best possible estimate prior to the initialized date. Any significant unexpected changes in these forcings (e.g. large volcanic eruptions and subsequent volcanic aeorosol loadings) during the forecast period could degrade the skill of the forecast. To that effect, the skill obtained by using observed forcings is somewhat overestimated.

Table 1 CGCMs included in this study

CMIP5 I.D.	Short name	Origin	Reference
BCC-CSM1.1	BCC	China	Wu et al. (2013)
CanCM4	CanCM4	Canada	Merryfield et al. (2011)
CNRM-CM5	CNRM	France	Voldoire et al. (2012)
CMCC-CM	CMCC	Italy	Scoccimarro et al. (2011)
CCSM4	CCSM4	USA	Gent et al. (2011)
EC-Earth	EC-Earth	Europe	Du et al. (2012)
GFDL-CM2.1	GFDL	USA	Delworth et al. (2006)
UKMO-HadCM3	HadCM3	UK	Gordon et al. (2000)
IPSL-CM4	IPSL	France	Dufresne et al. (2013)
MIROC4h	MIROC4	Japan	Sakamoto et al. (2013)
MIROC5	MIROC5	Japan	Watanabe et al. (2010)
MRI-CGCM3	MRI	Japan	Yukimoto et al. (2012)

Various initialization strategies have been adopted by different model groups. A first subset of models (GFDL2.1, HadCM3, EC-Earth, CanCM4) uses what is commonly referred to as full-field initialization where the ocean fields are initialized towards the actual observed ocean state. The second subset (MIROC5, HadCM3, EC-Earth) relies on a technique referred to as anomaly initialization, where observed ocean anomalies for a given date/month are superimposed onto the model's climatology for the same date/month. This second approach aims to minimize the temporal drift of the systems towards their preferred climatology. For more information on the different

Table 2 Number of ensemble members for each model ensemble. In the instances where the same model was used to produce two series of hindcasts using two different initialization strategies, they were considered as separate members in our multi-model ensemble

Model name	INI1	INI5	NON-INI1	NON-INI5
BCC	–	3	–	1
CCSM4	–	10	–	6
CMCC	–	3	–	1
CNRM	–	10	–	1
CanCM4 (1)	10	10	10	10
CanCM4 (2)	–	10	-	10
EC-Earth (ff)	5	5	11	11
EC-Earth (an)	8	8	11	11
GFDL CM2.1	10	10	10	10
HadCM3 (ff)	10	10	10	10
HadCM3 (an)	10	10	10	10
IPSL	–	6	–	4
MIROC4	–	3	–	3
MIROC5	6	6	3	3
MRI	–	9	–	1
Ensemble	7	15	5	12

initialization techniques and model systems, we refer the interested readers to Smith et al. (2012).

Both full field and anomaly initialized hindcasts are bias corrected according to ICPO (2011). However, as pointed out in Goddard et al. (2012) and Kharin et al. (2012), these corrections do not address potential drifts caused by incorrect model response to external forcings. As such, unless the residual drift systematically improves the quality of the forecasts in all systems, it could be argued that the results presented here offer a lower bound on the skill available from decadal hindcasts at predicting Atlantic hurricane activity (from this technique).

The second ensemble is constructed using a combination of the historical (1961–2005) and the RCP 4.5 scenario (2006 onward) simulations performed using the same models as the first ensemble. As this set of runs does not use observation-based initialization, we do not expect the simulated natural variability to be in phase with the observed variability, allowing for the second ensemble to be used as a comparison basis in evaluating the added-value of initialization. This second ensemble will henceforth be referred to as NoIni1.

2.2 Hurricane index

Rather than tracking storms directly in individual simulations, frequencies of North Atlantic hurricanes are estimated using a statistical emulator shown to recover much of the observed variability in hurricane activity over the period 1982–2009 (Vecchi et al. 2011). This statistical emulator is formulated as a Poisson regression model with two predictors: mean ASO tropical SST (limited by 30°N and 30°S) and mean ASO North Atlantic tropical SST (limited by 10°N, 25°N, 80°W and 20°, referred to below as the MDR). The annual Atlantic hurricane frequency λ of this statistical emulator is given by

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

where SST_{MDR} and SST_{TROP} are, respectively, the SST anomalies of the Main Development Region (MDR) and of the tropical SSTs relative to the 1982–2005 average. According to this relation, an increase in SSTs over the MDR leads to an increase in Atlantic hurricane numbers while an increase in SSTs over the tropics at large leads to a decrease. This behaviour rests on both dynamical and thermodynamical arguments, which are discussed in more details in Vecchi et al. (2011).

We opted to estimate hurricane numbers based on this statistical relationship (as opposed to directly tracking individual storms in the simulations) because simulations performed with resolution coaser than 1°, whilst fairly good at capturing large-scale fields modulating TC activity over the Atlantic, have difficulties producing realistic hurricane activity over that same region (Caron et al. 2011). Moreover, using seasonal mean of SSTs represent a more manageable quantity of data than that required for the tracking of individual storms (when these higher frequency data are available at all).

That being said, it is important to understand the limitations of this approach. As detailed shortly, the emulator used in conjunction with observed SSTs does not succeed at capturing the full range of hurricane variability. Thus, even if a model somehow managed to perfectly reproduce the observed SSTs over the entire period of study, it would still fail to predict the exact level of hurricane activity observed during that same period. It is clear then that tracking storms directly from daily data, such as in Smith et al. (2010), offers the potential for higher skill than this statistical emulator. For this reason, we will also include, as a reference, the skill obtained when observed SSTs are used in the emulator. This can be considered the highest skill level reachable using this technique (a perfect prediction).

Moreover, despite the apparent robustness of the index in various climate conditions (Vecchi et al. 2011), it is entirely possible that the statistical relationship between relative SSTs and hurricane numbers will break down in a future climate. For example, Ranger and Niehörster (2012) demonstrate how different models with good skill at predicting Atlantic TC activity for the present climate differ significantly when applied to long-term projections. However, we believe the time horizon considered here

(<10 years) is sufficiently short to prevent the breakdown of the statistical relation.

2.3 References and statistical measures

The predicted SSTs will be compared to the Hadley Centre Global Sea Ice Coverage and Sea Surface Temperature (HadISST) database (Rayner et al. 2002) while the various hurricane forecasts will be assessed using data taken from Vecchi and Knutson (2011). This database differs slightly from the more standard IBTrACS dataset (Knapp et al. 2010) as it accounts for possible missing storms in the pre-satellite era (prior to 1966). For the pre-satellite period considered here (1961–1965), these differences are relatively small (<0.3).[1]

Two standard statistical tests will be used to evaluate the skill of our hurricane decadal forecasts. The first measure is the mean square skill score (MSSS), defined as

$$MSSS = 1 - \frac{MSE_{for}}{MSE_{ref}} \tag{2}$$

where MSE_{for} is the mean square error of our forecast and MSE_{ref} is the mean square error of a baseline forecast. MSSS = 1 shows a perfect forecast and MSSS $\leq$0, a forecast with no improvement over the baseline. It is important to keep in mind that a high MSSS does not necessarily mean a good forecast, but simply a large improvement compared to what was previously available before (the reference). We evaluate our forecasts against four different baselines:

- a forecast based on uninitialized projections (MSE_{NoIni1}).
- a forecast based on climatology (MSE_{clim}) derived from the 25 years preceding the forecast.
- a forecast based on persistence (MSE_{pers}) created by extrapolating the value of the preceding period into the next one.
- a forecast based on a linear combination of climatology and persistence (MSE_{mix}) derived from Murphy (1992), such that $f_{mix} = r \ \ f_{pers} + (1 - r) \ \ f_{clim}$, where r is the first-order correlation coefficient of the observed timeseries and f_{pers} and f_{clim} are the persistence and climatology forecasts, respectively.

A linear combination of climatological and persistence forecasts has generally been shown to be more accurate than forecasts based on climatology or persistence alone (Buell 1958). Using climatology and persistence as baselines allows comparisons with respect to simple forecasts based on past observations (which we aim to improve upon), while using the uninitialized ensemble (NoIni1) allows an evaluation of the improved skill due to initialization.

The second quality measure used in this study is the anomaly correlation coefficient (ACC). Highly auto-correlated timeseries of N elements (such as multi-annual hurricane counts) cannot be considered as timeseries of N mutually independent elements. Disregarding this time dependence would lead to an underestimation of the sampling distribution variance and to an overconfidence in the statistical significance of our results (through narrower confidence intervals than actually warranted). As such, we define an effective degree of freedom $N_{eff} < N - 2$, whereas N_{eff} is taken from Bretherton et al. (1999) and is defined as

$$N_{eff} = \frac{N}{\sum_{\tau=-(N-1)}^{N-1}(1 - |\tau|/N)\rho_{\tau}^{x}\rho_{\tau}^{y}} \tag{3}$$

where ρ_{τ}^{x}, ρ_{τ}^{y} are the autocorrelation of timeseries X, Y at lag τ and N is the number of elements in these timeseries. The confidence interval is computed after a Fisher's Z-transformation. Unless otherwise stated, the confidence intervals represent the two-sided 90 % uncertainty of the given correlation. Because we expect positive correlations between the forecasts and observation as well as an improvement from initialization, correlations can be considered one-sided and confidence intervals that do not include zero are considered significant at the 95 % level.

3 Results

We start by evaluating the skill of the initialized hindcasts using the first five and nine forecast years. We expect that evaluating the skill over periods of 5 years or more will average out the ENSO variability, which is known to strongly influence Atlantic hurricane activity (Landsea 2000). It is common practice to exclude the first forecast year of any given decadal forecast as it is considered to be within the boundary of seasonal forecasting (Goddard et al. 2012). Here however, we retain the first year in our evaluation. We feel this is justified since the systems are initialized 8 or 10 months before the official beginning of the hurricane season on August 1st. Discarding the first year would cause systems initialized on November 1st to make their first prediction 22–24 months after initialization. Furthermore, November 1st seasonal forecasts generally show no skill above that of climatology (Saunders and Lea, 2011).

Figure 1 shows the timeseries of predicted five- (top row) and nine-year (bottom row) mean hurricane numbers for both the Ini1 ensemble (left column) and NoIni1 ensemble (right column). Individual models are shown using full colored lines while the multi-model ensemble is

[1] The bias corrected hurricane count is available online at http://www.gfdl.noaa.gov/cms-filesystem-action/user_files/gav/data/lvbk09_stormcounts.txt.

represented by the dashed black line. The predictions are located over the first year of the forecasted period and each forecast is derived using only the first five or nine years of individual hindcasts. The hindcast predictions can then be contrasted with the observed activity for the corresponding five- and nine-year averaging periods (solid black line) and the index value obtained using observed SSTs (purple cross). The latter can be considered a perfect prediction of the statistical emulator.

For both five- and nine-year periods, the Ini1 ensemble exhibits better qualitative agreement with observations than does the NoIni1 ensemble. Although Ini1 fails to capture some of the observed variability, the general tendencies (decreasing activity in the first half of the record and an upward trend starting in the 90's) are well captured by the multi-model ensemble and to varying degrees by individual model ensembles. In comparison, predictions from the NoIni1 ensemble show much lower variability and fail to predict periods of low or high activity. Individual NoIni1 ensemble members do show some level of variability, but none of them come close to matching the observed activity as well as the perfect prediction.

Figure 2 shows the ACCs for both model ensembles (Ini1, black; NoIni1, blue) for both five- (first row) and nine-year horizons (bottom row). To allow for comparison, we also show the ACCs of persistence (red), and mix (green) forecasts, as well as the ACC for the "perfect prediction" (purple). For five-year forecasts, the Ini1 ensemble ACC is greater than the other three references, but, due to the large confidence interval, the null hypothesis cannot be rejected with 95 % certainty. It is worth pointing out that two models (GFDL, MIROC5) return an ACC which is statistically different from 0 (not shown). For nine-year forecasts, the Ini1 ACC increases to nearly 0.8 but does not become statistically different from 0. Interestingly, despite very small variations in the NoIni1 ensemble 9-year predictions, its ACC increases to $\sim$0.65 (and becomes statistically different from 0), suggesting that, at the decadal timescale, changes in Atlantic hurricane activity are linked, at least in part, to changes in radiative forcing, as suggested by Smith et al. (2010).

The results of the second statistical test (MSSS), which gives the improvement relative to different baseline forecasts, is shown in Fig. 2b (5 years) and 2d (nine years). Values along the y-axis return the skill score with respect to the uninitialized forecast, while values along the x-axis return the skill scores with respect to forecasts based on past observations: climatology (square), persistence (diamond) and a mix of persistence and climatology (circle). Hindcasts located in the upper-right quadrant show improvement with respect to both the uninitialized forecasts and the second baseline forecast (climatology, persistence, or mixed). We also show the MSSSs derived from using observed SSTs (purple, at y = 0), which represent an upper limit on the expected skill of the statistical emulator. For five-year forecasts, the multi-model ensemble mean returns a positive

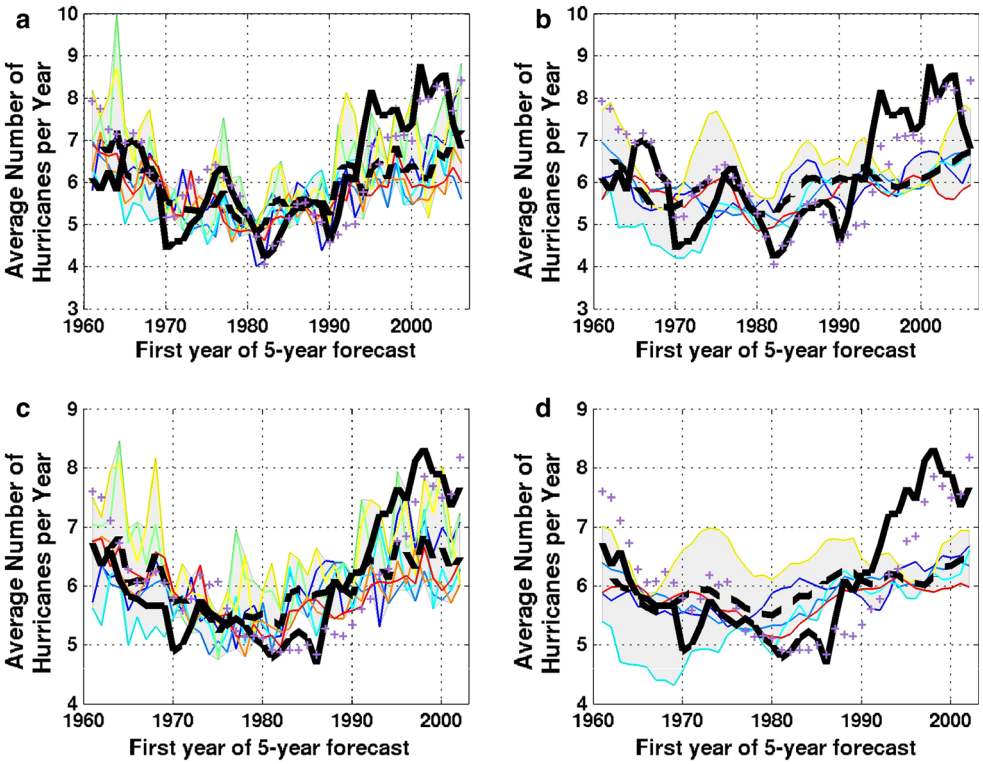


Fig. 1 *First row*: Timeseries of five-year mean hurricane numbers for **a** initialized (Ini1) and **b** non-initialized (NoIni1) forecasts. *Second row*: Timeseries of nine-year mean hurricane numbers for **c** initialized and **d** non-initialized forecasts. The *black dashed line* represents the ensemble mean, the full *color lines*, individual model means, while the full *black line* represents the observed Atlantic hurricane numbers over the corresponding five- or nine-year period. The *purple crosses* represent the predicted number of hurricanes using observed SSTs. The forecasts are aligned with respect to the first year of the forecasted period. For example, a five-year forecast covering the period 1961–1966 will be aligned with the year 1961. The color code for each GCM is shown in Fig. 2

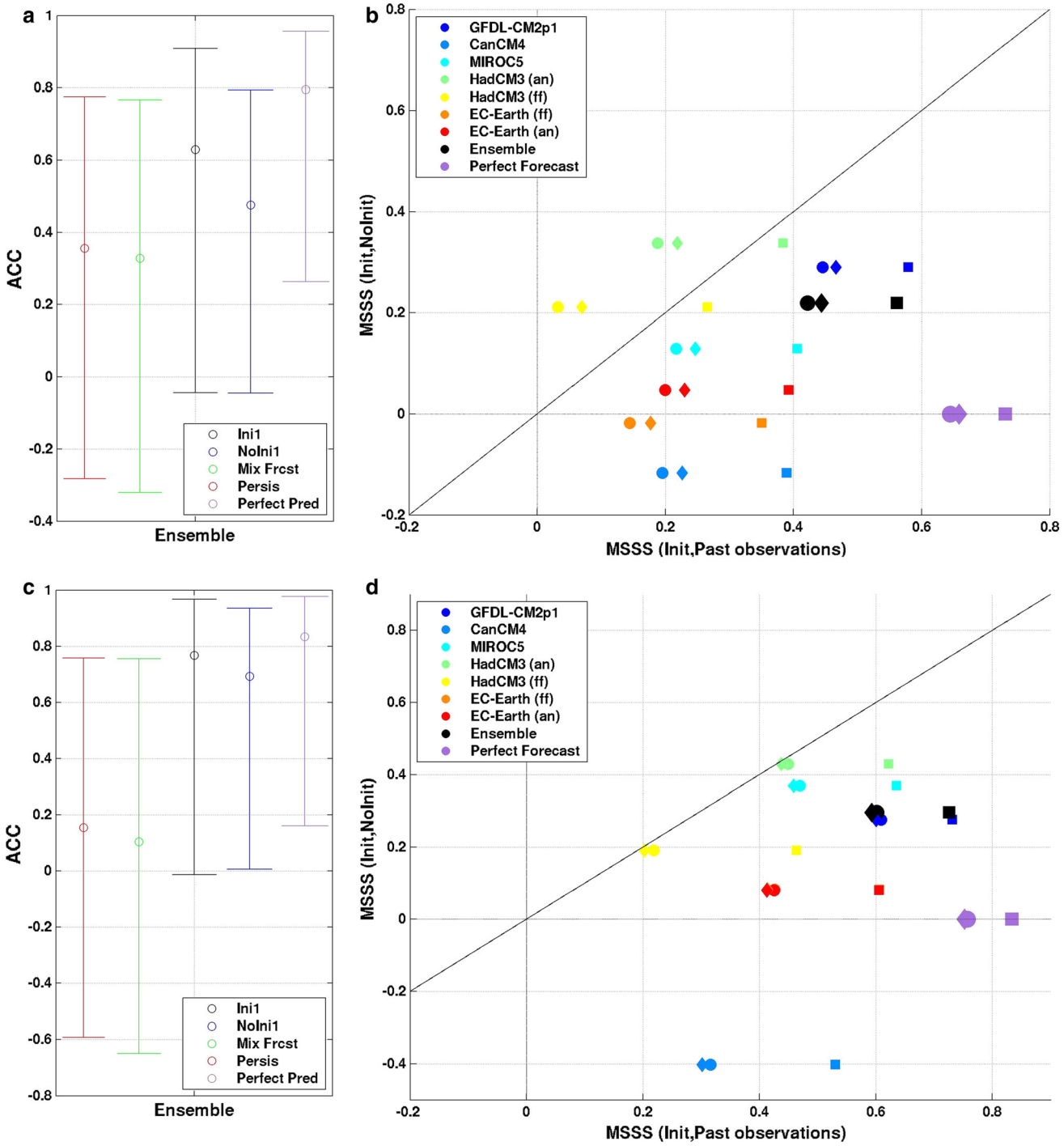


Fig. 2 *First row*: **a** ACCs of five-year mean hurricane forecasts for Ini1 (*black*) and NoIni1 (*blue*) ensembles. ACC for baseline forecasts are in *red* (persistence) and *green* (mix persistence+climatology), while the ACC derived using observed SSTs (perfect forecast) is in purple. The *bars* represent the two-sided 90 % confidence interval. **b** MSSSs of five-year mean hurricane forecasts for Ini1 ensemble (*black*) and individual members (*colors*). The skill is measured with respect to non-initialized forecasts (NoIni1) on the *vertical axis* and with respect to forecasts based on past observations in the X-axis: climatology (*square*), persistence (*diamond*) and a mix of persistence and climatology (*circle*). In both cases, the forecast is made using only the first five lead years of the hindcasts.*Second row*: **c**, **d** same as (**a**), (**b**) respectively, but for nine-year forecasts

MSSS (thus an improvement) with respect to all references, ranging from ~0.6 (w.r.t. climatology) to 0.2 (w.r.t. non-uninitiliazed forecasts). All systems outperform the climatology and persistence forecasts. Two systems (EC-Earth, CanCM4) returns a MSSS<0 (no skill) with respect to uninitialized forecasts. Given that these model's MSSSs

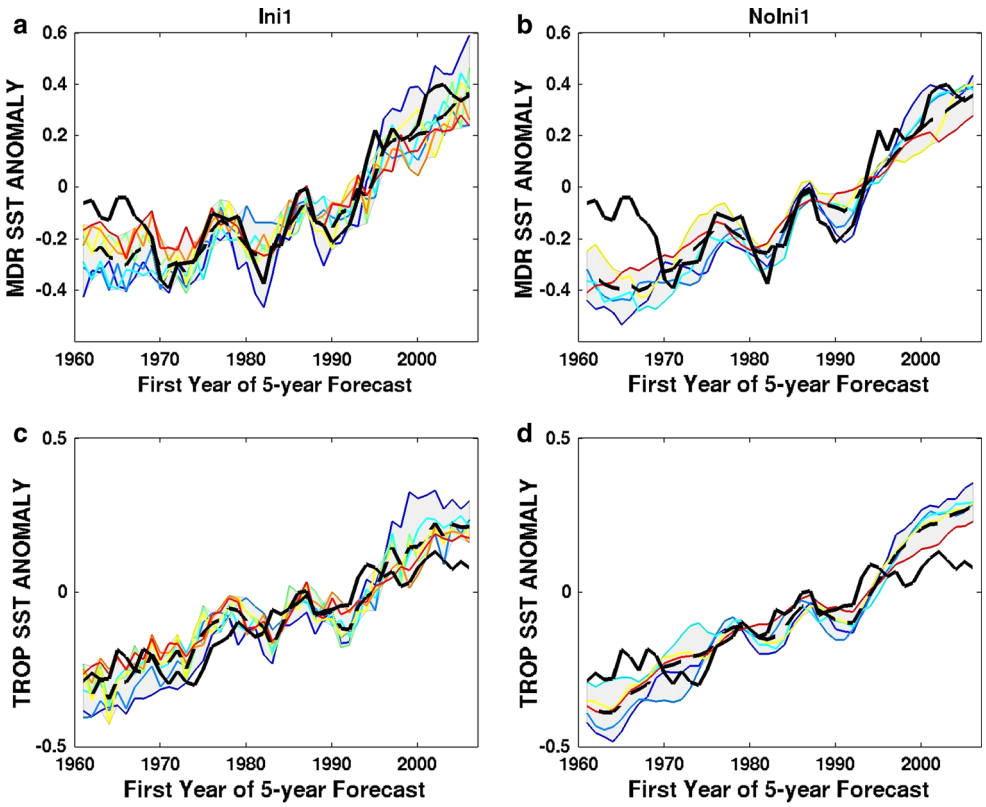


Fig. 3 First row: Timeseries of five-year mean Atlantic SST anomalies (with respect to 1982–2005 mean) for **a** initialized (Ini1) and **b** non-initialized (NoIni1) forecasts. The full *black line* represents the observed (HadISST) SSTs, the *black dashed line*, the ensemble mean while the full *color lines* represent individual model means. *Second row*: **c**, **d** same as (**a**), (**b**) respectively, but for the tropical SST anomalies. The *color code* for each GCM is shown in Fig. 2

are positive and comparable to other systems when past observations are used, this is likely more a consequence of "good" historical runs than "bad" forecasts. The reason(s) as to why initialization offers no improvement in these particular cases is beyond the scope of this study. For nine-year forecasts, the multi-model ensemble shows MSSSs ranging between $\sim$0.7 (climatology) and $\sim$0.3 (non-initialized). Over this time range, all systems offer an improvement with respect to past observations, and all but one (CanCM4) give an improvement with respect to non-initialized forecasts. Finally, it is interesting to compare systems that have adopted two different initialization strategies (EC-Earth and HadCM3). In general, the anomaly initialized predictions tend to outperform the full field predictions, the exception being EC-Earth for 9-year forecast, where both strategies return similar skill.

Having established that our initialized decadal forecasts and the statistical emulator show a certain level of skill in predicting upcoming multi-annual Atlantic hurricane activity or, at the very least, offer a definite improvement over simple multi-annual forecasts, we now investigate the origin of the detected skill.

3.1 Origin of the skill

As the statistical emulator (Eq. 1) uses only two predictors, skill can originate either from tropical North Atlantic SSTs or global tropical SSTs (or a combination of both). Figure 3 shows the mean Atlantic SST anomalies (top row) and mean tropical SST anomalies (bottom) computed using the first five years of the initialized (left column) and non-initialized (right column) forecasts. Simple qualitative comparison suggests a better improvement due to initialization for the tropical Atlantic SSTs than for the tropics at large. This is confirmed in Fig. 4, which shows the ACCs and MSSSs for Atlantic and tropical SST anomalies. The ACC of NoIni1 is about the same as that of Ini1 in the tropics while the MSSSs tend to be lower for the tropical SSTs than with Atlantic SSTs. The fact that there is much lower skill due to initialization for SST_{TROP} (Ini1 vs. NoIni1) compared to a climatology forecasts is a consequence of the signal being dominated over the last 50 years by the upward trend induced by GHGs.

Vecchi et al. (2013) identified the MDR SST anomalies as the primary source of skill and Fig. 4 supports this interpretation. The MSSS of the multi-model ensemble is $\sim$0.8 with respect to forecasts based on past observations and $\sim$0.5 with respect to NoIni1, both higher than for tropical SST anomalies. The ACC for the MDR SST anomalies is marginally larger in Ini1 than in NoIni1, but significantly different from a forecast based on persistence. All models show MSSSs greater than 0 with respect to all baselines. The CanCM4 forecast system shows almost no improvement due to initialization over the MDR, which

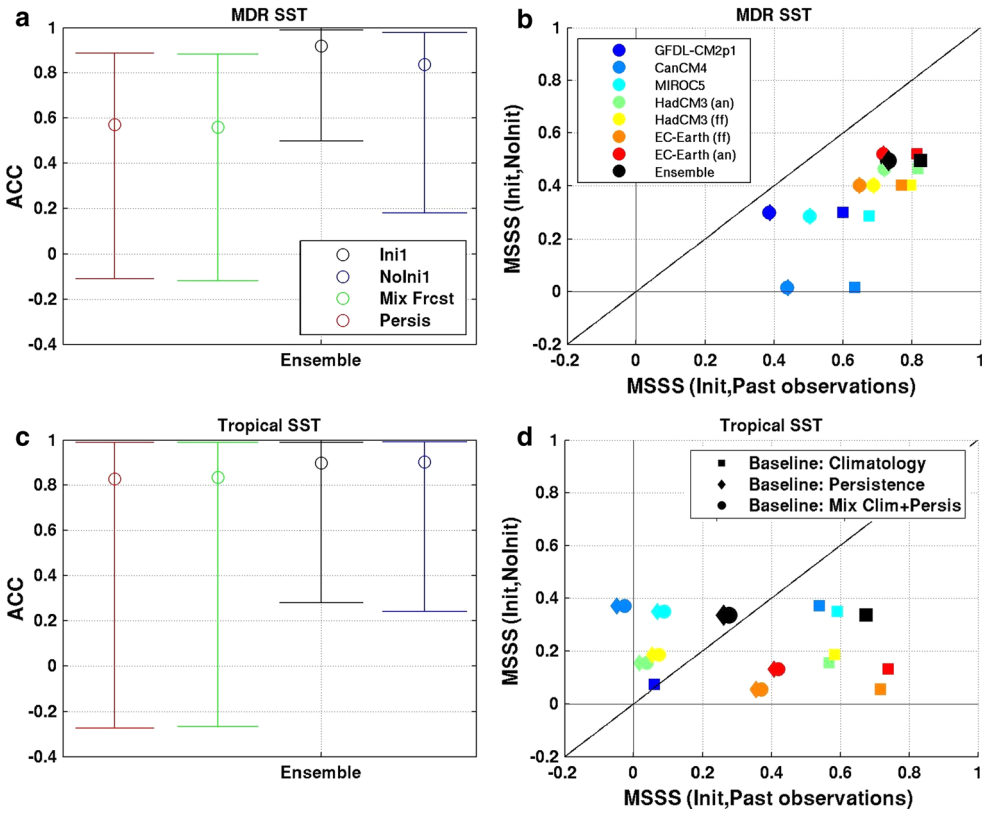


Fig. 4 *First row*: **a** ACCs of five-year mean Atlantic SST anomaly forecasts for Ini1 (*black*) and NoIni1 (*blue*) ensembles. ACC for baseline forecasts are in red (persistence) and green (mix persistence + climatology). The *bars* represent the two-sided 90 % confidence interval. **b** MSSSs of five-year mean Atlantic SST anomaly forecasts for Ini1 ensemble (*black*) and individual members (*colors*). The skill is measured with respect to non-initialized forecasts (NoIni1) on the vertical axis and with respect to forecasts based on past observations in the X-axis: climatology (*square*), persistence (*diamond*) and a mix of persistence and climatology (*circle*). In both cases the forecast is made using only the first five lead years of the hindcasts. *Second row*: **c**, **d** same as (**a**),(**b**) respectively, but for tropical SST anomaly forecasts

explains the lack of skill mentioned earlier. The key to predicting Atlantic hurricane activity with the emulator thus appears to come from accurate forecasts of tropical Atlantic SSTs.

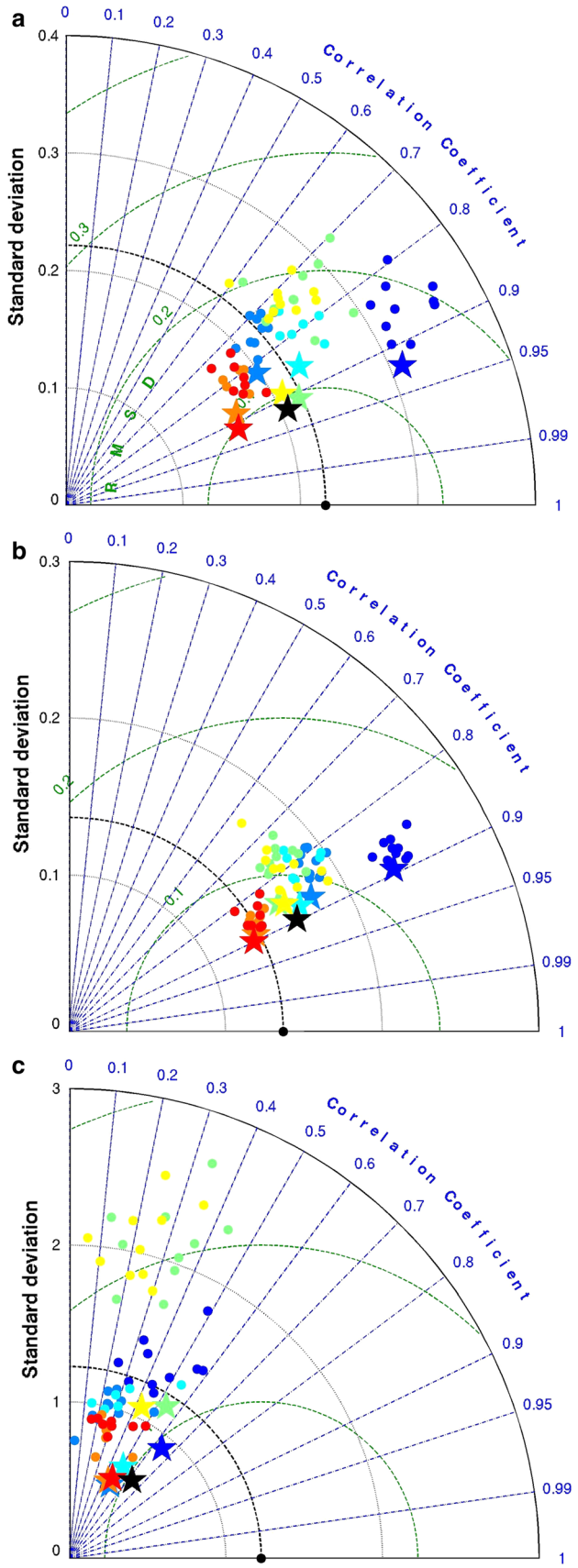
Inspection of MSSSs in Fig. 4 reveals two curious features: (1) the forecast system with some of the best score with respect to past observations for MDR SST anomalies (HadCM3) also returns the lowest MSSSs, using the same baselines, for Atlantic hurricane numbers; (2) the system with the best MSSSs for hurricanes numbers (GFDL2.1) shows the lowest skill (w.r.t. past observations) compared to other systems for MDR SSTs.

To resolve these discrepencies, we refer to the Taylor diagrams shown in Fig. 5. MDR anomalies (5a) in individual HadCM3 simulations have higher RMSDs than MDR ensemble mean anomalies in other systems. However, the HadCM3 ensemble anomalies return the smallest errors. For hurricane numbers, individual HadCM3 simulations also return poorer than average results (Fig. 5c). In this case however, computing the ensemble mean of predicted hurricane numbers does not return an equivalent improvement and HadCM3 predictions remain somewhat poorer than other systems. One could thus argue for the use of model means in Eq. 1 rather than using individual simulation SSTs, as we have done here. However, in this case, this second approach leads to a decrease in the ensemble MSSSs by $\sim$0.1 (not shown).

On the other hand, the discrepancy in the GFDL2.1 results appears to come from compensating errors in the model SST anomalies. Figure 5a, b show that both GFDL MDR and tropical SST anomalies stand apart somewhat from other model's anomaly. GFDL hindcasts show the steepest increase of all systems in SSTs over the given period due to SST anomalies that are below the ensemble mean in the 1960's and significantly above the ensemble mean after 1995. Over- or understimation of only one of the two parameters would significantly degrade the prediction, but over/underestimation of both MDR SSTs and tropical SSTs will tend to cancel each other out in Eq. 1. These compensating errors in GFDL2.1 lead to predictions of hurricane numbers in individual simulations comparable to or better than those of other forecast systems and also cause its ensemble mean to be better than other systems. That being said, GFDL2.1 is also the only model able to capture the strong dip in MDR SST in the early 80's.

3.2 Influence of forecast year and lagged-ensemble

The previous analysis of five-year long forecasts relied solely on the first to fifth year of individual initialized simulations, discarding the later years. We now evaluate the change in skill level in five-year mean forecasts as we increase the forecast lead year. Figure 6 shows the ACCs and MSSSs for the Ini1 ensemble for forecast periods



◀ **Fig. 5** Taylor diagram for five-year mean **a** Atlantic SST anomaly forecasts, **b** tropical SST anomaly forecasts and **c** Atlantic hurricane forecasts. The plots show individual simulations (*dot*) and the ensemble mean (*star*). The multi-model ensemble (Ini1) forecast is shown in *black*

ranging from lead years 0–4 to lead years 5–9. In order to avoid a different number of forecasts for different five-year periods, we perform our analysis using only year 1966 onward (e.g. there are no forecasts for which the 3rd-8th forecast years correspond to years 1961–1965). This explains why the Ini1 ACC 90 % confidence interval for years 0–4 is different from Fig. 2 and is now statistically different from 0. In fact, the ACC remains statistically different than 0, which is not the case with the persistence or NoIni1 forecasts, for a lead time of up to two years. Since some of the systems are initialized at the official end of the hurricane season on November 1st, two lead years mean, in this case, a lead time of almost three years. The ACC for lead years 5–9, although not statistically significant, is still larger than that of persistence forecasts. Figure 6 also shows the MSSSs using different skill baselines. All show their sharpest decline over the first two forecast averages (from lead year 0–2) as well as a general decline from year 0 to year 5. None shows a skill of 0 or lower, even with a five year lead time, suggesting some skill is still present in the later years of the initialized forecasts.

These results suggest that we can make use of these additional forecast years to create a lagged-ensemble, which combines hindcasts started at different start dates but which verify over the same period and has a larger size. For example, we can combine the fifth to ninth forecast years of a hindcast initialized in 1965 with the fourth to eighth year of a hindcast initialized in 1966 to obtain a forecast for the 1970–1974 period. This approach, also adopted by Smith et al. (2010) and Vecchi et al. (2013), allows us to significantly increase (factor of 10) the number of forecasts for any given year. These results are shown in Fig. 7. The ensemble mean ACC increases slightly, but so does the confidence interval, since the addition of data with longer forecast years further increases the auto-correlation of the ensemble timeseries and further reduces its degrees of freedom. This effect was also detected in Vecchi et al. (2013). Similarly, the MSSSs of the ensemble mean are not significantly impacted. On the other hand, MSSSs of individual ensemble members are generally improved by incorporating the later years into a forecast. In fact, all systems except one (GFDL2.1) show relatively comparable MSSSs with respect to baseline forecasts derived from observations. All systems display a similar behaviour: a downward trend in 5-year forecast from 1960 to 1980 and an upward trend thereafter. Generally, the systems pre-

dicting fewer storms at the beginning of the 1960's (EC-Earth, CanCM4) also predict fewer storms in the early 2000's such that the systems that offer the better forecasts in the earlier period are not the same ones that offer the better forecasts during the later period. One exception, GFDL2.1, shows results near the ensemble mean at the beginning of the period as well as predicting the strongest activity in the 2000's. Finally, part of the general increase in MSSS with respect to NoIni1 likely comes from the fact that no benefit is found from creating a lagged-ensemble for the NoIni1 forecasts: all the "forecasts" originate from the same historical run and adding later years does not effectively increase the number of forecast for a given year.

3.3 Prediction of the mid-1990's climate shift

A strong increase in Atlantic hurricane activity began occuring in 1995 and persists to this day. Vecchi et al. (2013) argue that part of the skill detected in their two forecast systems simply originated by persisting the climate shift rather than predicting it. That is, simulations initiated before 1995 were not able to predict higher activity from 1995 onward, whereas those initiated after 1995 could. They estimated that, after accounting for the ability of the systems to persist the mid-1990's climate shift, the ACC was closer to 0.4 for lead years 2–6 and to 0 for lead years 6–10. Here, we verify if their conclusion holds for the multi-model approach.

Following the method of Vecchi et al. (2013), we remove the 1995 shift by first dividing our hurricane timeseries into two periods (1961–1994, 1995–2010) and subtracting the respective climatology from each period. This procedure is applied to the observational timeseries as well as the Ini1 and NoIni1 timeseries. The resulting ACCs and MSSSs for five-year periods for lead years ranging from 0–4 years to 5–9 years are shown in Fig. 8. After removing the shift, the ensemble ACC for lead years 0–4 is much reduced compared to the one shown in Fig. 2, dropping from $\sim$0.6 to $\sim$0.4, similar to Vecchi et al. (2013) estimation. The ACC remains $\sim$0.4 (and statistically different than 0) up to lead years 3–7, and subsequently decreases to $\sim$0.2 for lead years 4–8 and 5–9. Although not statistically different than 0, this latest value is higher than the one estimated by Vecchi et al. (2013) for similar lead years. Finally, the MSSSs calculated with respect to both persistence and non-initialized forecasts similarly suggest that, even once the 1995 shift has been removed, some level of skill remains in the decadal hindcasts at least up to lead years 3–7.

3.4 Skill of the CMIP5 core set of experiments

A larger number of initialized decadal hindcasts, initialized every five years from 1960 to 2005, is also available through the CMIP5 database. This larger ensemble is part of the core set of experiments designed for the CMIP5

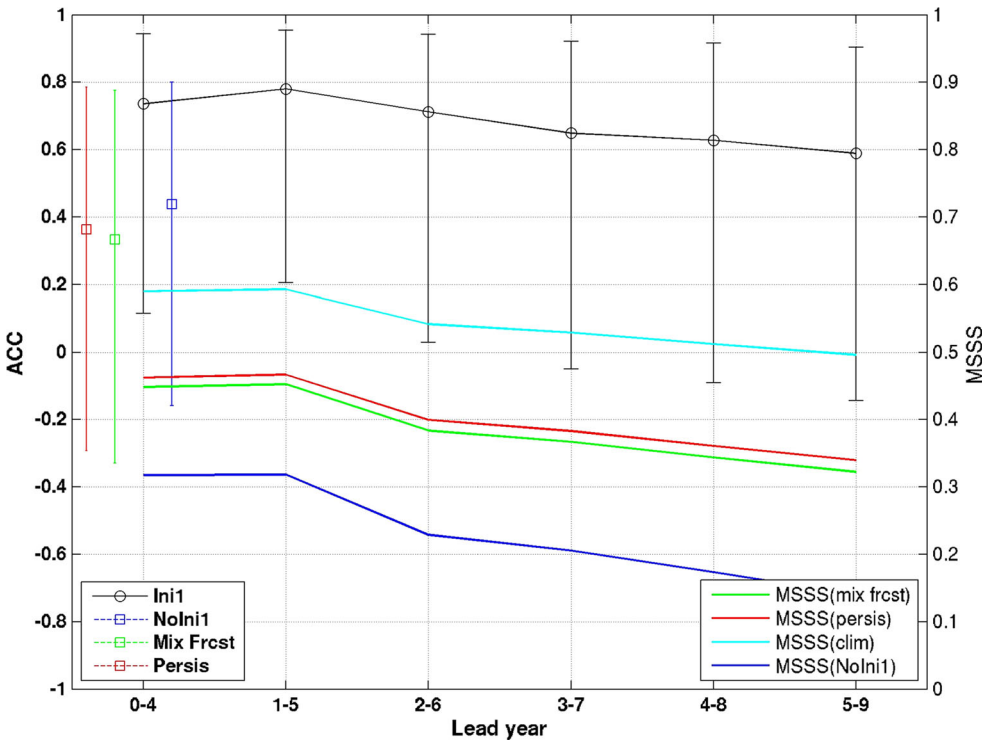


Fig. 6 *Left axis*: ACCs (and two-sided 90 % confidence intervals) for five-year mean hurricane forecasts for different lead years. The initialized ensemble is shown in *black*, while the ACC for non-initialized (Non-Ini1) forecasts, persistence forecasts and mix persistence-climatology forecasts are in *blue*, *red* and *green*, respectively. *Right axis*: MSSSs for five-year mean hurricane numbers with respect to forecasts based on climatology (*light blue*), persistence (*red*), mix climatology+persistence (*green*) and non-initialized forecasts (*blue*)

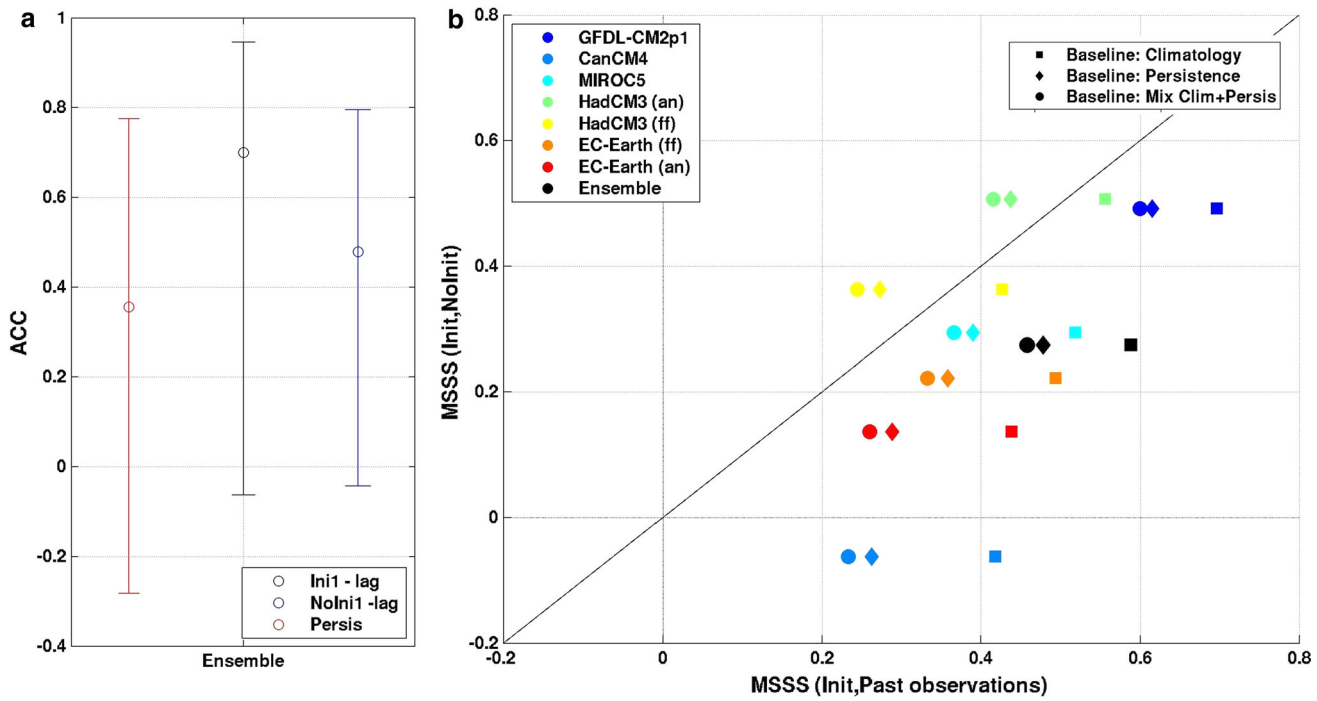


Fig. 7 Same as Fig. 2a, b, but using overlapping hindcasts years (lagged-ensemble)

project and includes more GCMs than in the Ini1 ensemble. We now investigate how the skill of this larger ensemble, which we will refer to as Ini5 (and NoIni5 for its non-initialized counterpart), compares to that of our smaller ensemble made up of yearly hindcasts. The idea here is to investigate whether skill levels similar to those obtained with Ini1 could be reached using a subset of start dates but a larger number of models.

A short description of the models comprising this larger ensemble can be found in Table 1. Figure 9 shows the timeseries for five-year mean hurricane numbers for both Ini5 (9a) and NoIni5 (9b). As with Ini1, Ini5 shows better qualitative agreement with observations, with a downward trend at the beginning of the period and an upward trend starting in the early 80's. The variance of the ensemble mean is much smaller than for observations, but this is to be expected, due to the nature of the averaging operator, which reduces the variance by a factor of N. This translates into an overestimation during the inactive periods and an underestimation during the active ones. The "predictions" of the non-initialized ensemble perform more poorly and predict a relatively constant number of storms (6), regardless of the start date.

Figure 10a shows the ACC and 90 % confidence interval of this second ensemble. Due to the smaller sample of start dates, the Ini5 ACC displays a larger confidence interval than the Ini1 ACC and is again not statistically different from 0. Figure 10b shows the MSSSs of individual simulations and the ensemble mean with respect to non-initialized hindcasts (y-axis) and climatology (x-axis). The multi-model ensemble mean returns MSSSs >0 for either of the baseline, with MSSS $\sim$0.2 in both cases.

It is interesting to compare the performance of the larger Ini5 ensemble with the performance of the smaller Ini1 ensemble broken down into five subgroups based on their start date (1960, 1965,…; 1961, 1966,…). This allows us to evaluate how the skill of the Ini5 ensemble depends upon the choice of start dates and whether a different series of start dates might have yielded different results. The ACCs and MSSSs for each of the five sub-ensembles is shown in Fig. 11. Results for the 01,06 years exclusively are poorer than results obtained using all start dates, but better than those from the larger Ini5 ensemble. It should also be noted that since the MSSS of the perfect prediction is relatively low (MSSS $\sim$0.6), the lower skill score for this ensemble of start dates should largely be attributed to shortcomings of the statistical emulator rather than to the forecast systems themselves. This contrasts with the poor performance during the 04,09 years which, in this case, seem to depend more on the forecast systems since the skill of the perfect prediction is relatively high.

Finally, the sub-group made up of years 00,05 offers a high skill level, comparable to that obtained with the entire Ini1 ensemble, with an ACC statistically different than 0. This suggests that, not only are the 01,06 start dates particularly ill-suited for the exercise at hand, but also that the skill level for systems initialized every five years is highly dependent, in this case, upon the series of start dates

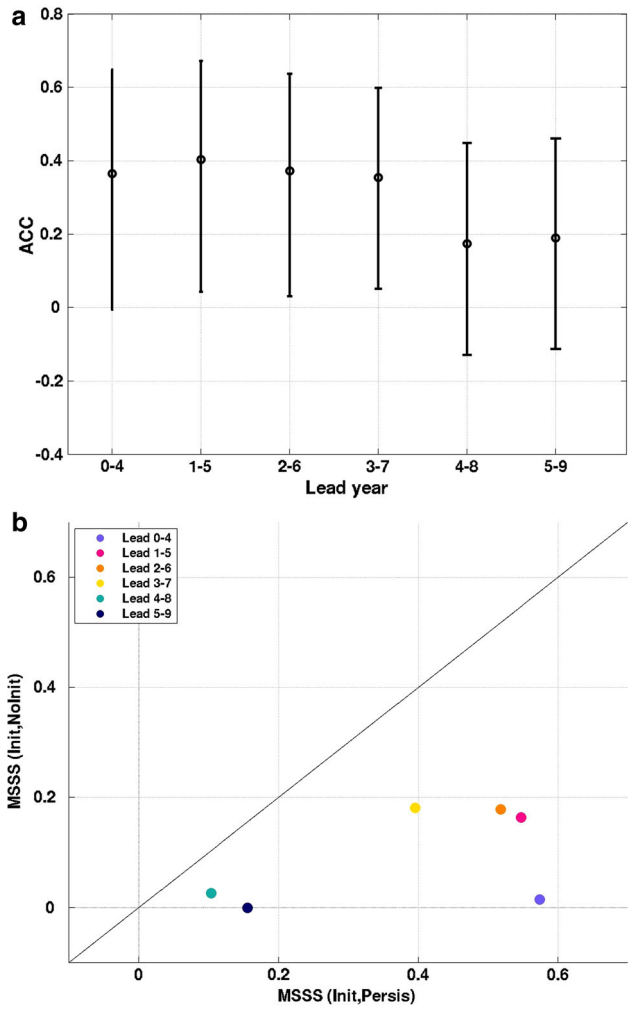


Fig. 8 **a** ACCs for five-year mean hurricane hindcasts after the removal of the 1994–1995 climate shift for different lead years of the Ini1 ensemble. The *bars* represent the two-sided 90 % confidence interval. **b** MSSSs of five-year mean hurricane hindcasts for different lead years of the Ini1 ensemble. Thee MSSS is measured with respect to persistence along the X-axis and with respect to non-initialized forecasts along the Y-axis

themselves and that higher frequency sampling is a necessary condition to obtaining a robust measure of the available skill.

4 Discussion and conclusion

Figure 2 suggests some level of skill for initialized, multi-model forecasts in predicting multi-annual levels of hurricane activity. If these results are encouraging, they should be interpreted with care. The fact that the ACC derived from HadISST (perfect prediction) is both statistically different from 0 and statistically different from the various baselines over both a five- and nine-year horizon suggests that the methodology used here offers, in theory, the possibility of predicting hurricane activity at the multi-annual level. However, over a five-year horizon, if the initialized forecasts outperform climatological, persistence and uninitialized forecasts, they do not return an ACC which is statistically different from 0 or even statistically different from either of these simpler forecasts. Morever, the skill is further reduced when the skill derived from persisting the mid-1990's climate shift is removed.

Over a nine-year horizon, the initialized forecast ensemble mean ACC increases, but does not become statistically significant. The fact that the non-initialized forecast ACCs also increase from a five- to a nine-year horizon suggests a role for radiative forcings in modulating hurricane activity, although the amplitude of the changes in the predictions are so small that it is not at all clear how these positive correlations should be interpreted. These results are consistent with those obtained by Vecchi et al. (2013) using a two-member ensemble.

Inspection of Fig. 1a reveals that there are two periods where the Ini1 ensemble forecasts are relatively poor : (1) the early 1980's and (2) the early to mid-2000's. The Ini1 ensemble predictions are also poor in the early 1960's, 1970's and mid-1990's; but given that the "perfect prediction" also fails to capture the observed hurricane activity during these three periods, this should be interpreted more as a shortcoming of the statistical emulator than a failure of the forecast systems themselves. We speculate that the failure of the ensemble prediction in the early 1980's rests on the inability of the systems to predict the strong El-Niño event that occurred during that period several years ahead of time as well as the failure of most systems to capture the full impact of the El Chichon eruption on Atlantic SSTs. This second argument is supported by Fig. 3a, which shows that only the GFDL2.1 (known to be very sensitive to aerosol forcings (Knutson et al. 2006)) system captures the full extent of MDR SST anomalies following the eruption. To a lesser extent, this argument also holds for Pinatubo's eruption in the early 1990's. In a sense, failure to capture the effective changes in radiative forcing from unpredictable volcanic eruptions may lead to a skill score more representative of the "true" skill of the forecast systems than if it had actually captured the full extent of these changes.

Based on Fig. 3a, c, failure of the index in the later period is due in large part to the inability of the systems to capture the recent slow down in global temperature increase. Starting in the early 1990's, the systems overestimate, to various degrees, the warming rate of tropical SSTs. On the other hand, MDR SSTs are captured more accurately, if somewhat underestimated in the early 2000's (except the GFDL CM2.1 system). Due to the formulation of the statistical emulator, this particular combination (negative MDR anomalies, positive tropical anomalies)

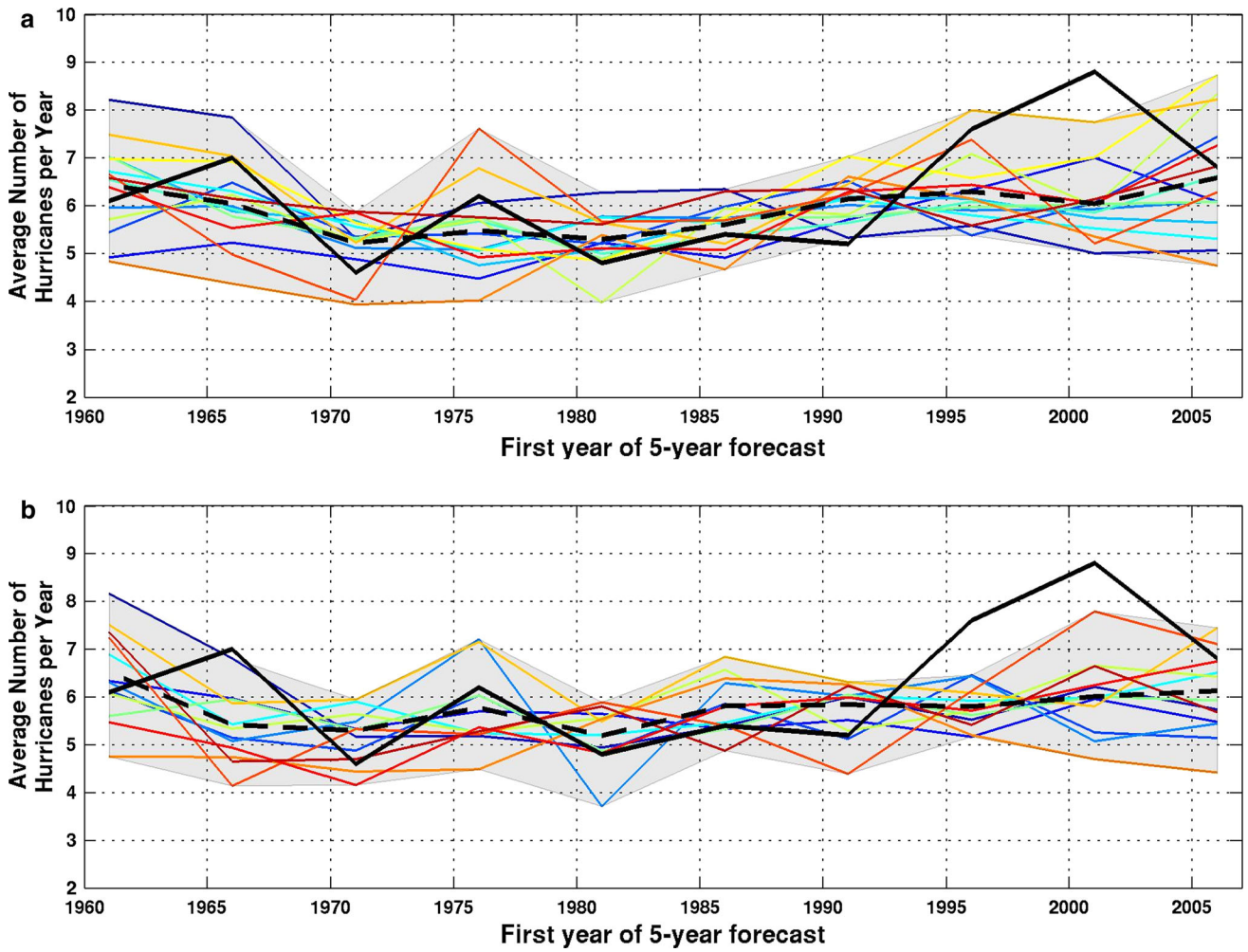


Fig. 9 Same as Fig. 1a, b, but for Ini5 ensemble (**a**) and NoIni5 ensemble (**b**)

leads to an underestimation of hurricane activity. Thus, if the skill derived from accurate prediction of MDR SST anomalies leads to relatively high ACCs and MSSSs, failure to accurately capture the warming over the tropics has severe consequences since it is directly responsible for the failure to predict the very high level of activity observed in the early to mid-2000's

Finally, we have shown that some skill remains in the later years of the hindcasts, with forecasts made by using lead time of 5–9 years still showing MSSSs>0. This suggested the creation of a lagged-ensemble for which hindcasts with different start dates but which verify over the same period could be combined to significantly increase the number of predictions for any given year. With this approach, skill of the multi-model ensemble was not significantly affected, but the skill of the individual model ensemble tends to be improved.

As mentioned previously, the results seem to confirm that the findings of Smith et al. (2010) and Vecchi et al. (2013), which suggest that Atlantic hurricane level is predictable a few years into the future, are robust. However, before reliable predictions can be made with the technique used here, some issues need to be addressed, first of which is the overestimation in tropical warming (over the last decade or so) shared by all the forecast systems (see Fig. 3c, d). The failure to capture the observed slow down in warming since ~2000 is an important (apparent) deficiency in the models. Although the reason(s) behind the slow down is not entirely clear, increase in ocean heat uptake below the mixed layer appears the likeliest candidate at this time (Loeb et al. 2012; Guemas et al. 2013). The inability of the multi-model ensemble to capture that feature causes the systems to strongly underestimate recent activity. In a forecast context (as opposed to hindcast), failure to predict the highest level of activity observed during the last ~50 years in the early to mid-2000's would be considered nothing short of a total failure. A better representation of ocean heat uptake and circulation in the climate models would likely improve predictability of TCs over the latter

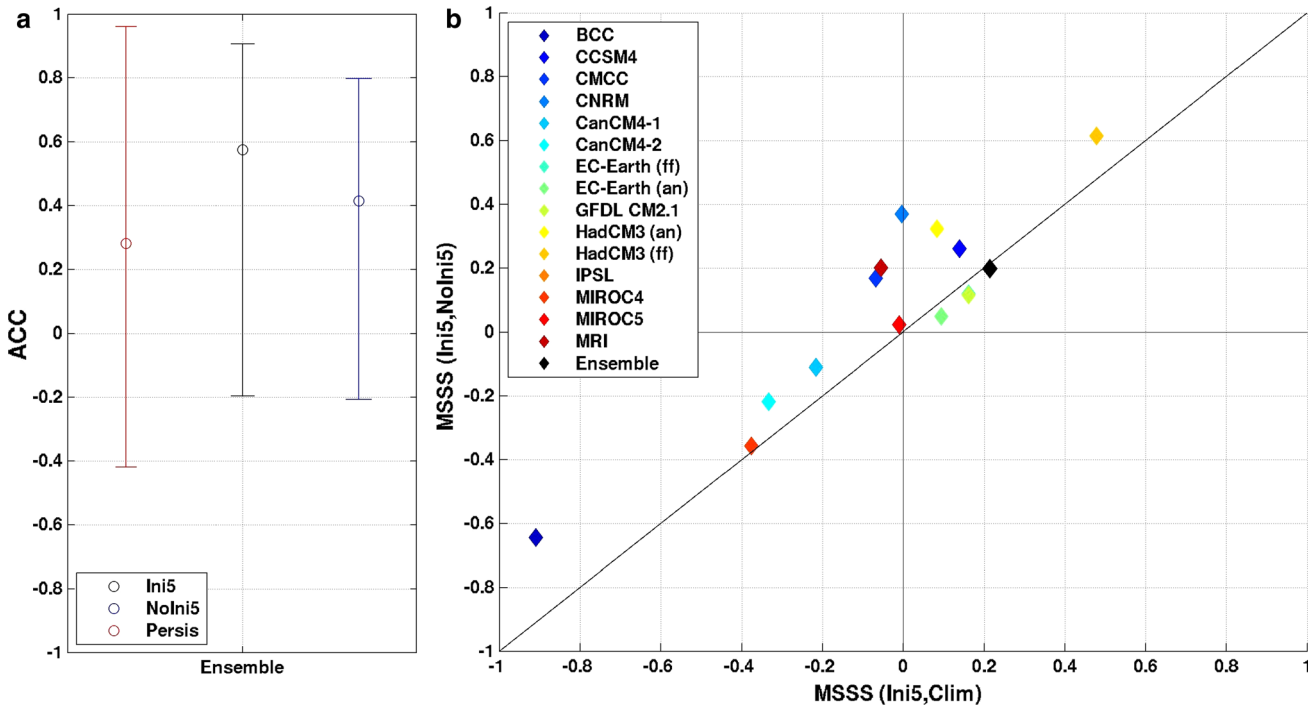


Fig. 10 **a** ACCs of five-year mean hurricane forecasts for Ini5 (*black*) and NoIni5 (*blue*) ensembles. ACC for the baseline forecast (persistence) is in *red*. The *bars* represent the two-sided 90 % confidence interval. **b** MSSSs of five-year mean hurricane forecasts for the Ini5 ensemble (*black*) and individual members (*colors*). The skill is measured with respect to climatology on the *horizontal axis* and with respect to NoIni5 on the *vertical axis*. The skill is evaluated using only the first five forecast years of the hindcasts

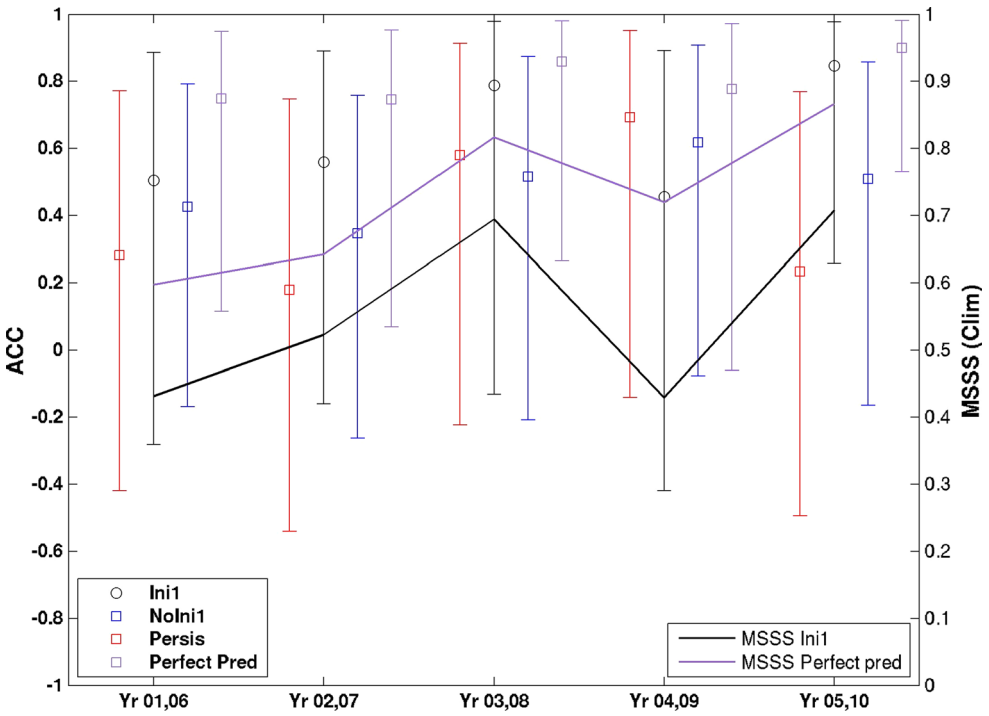


Fig. 11 *Left axis*: ACCs (and two-sided 90 % confidence intervals) for five-year mean hurricane numbers using different subsets of start years (1 every 5 years) for initialized hindcasts (*black*), non-initialized hindcasts (*blue*), and forecasts based on persistence (*red*). ACCs for perfect forecasts (made using observed SSTs) are shown in purple. *Right axis*: MSSSs of the ensemble (*full black line*) for each subset of start dates. The MSSSs of a perfect prediction is shown with the full *purple line*

part of the period and increase the skill detected here. Nonetheless, the results should still be viewed as an encouraging first step in possible reliable multi-annual prediction of Atlantic hurricane activity.

Acknowledgments We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups (listed in Table 1 of this paper) for producing and making available their model output. We are also grateful to two anonymous reviewers

whose comments led us to improve on the original work. The third author would also like to acknowledge financial support from the EU-funded SPECS project (Grant # 3038378). Finally, the first author also wishes to thank Katherine Barrett for her help in proofreading this document, Virginie Guemas for all the helpful conversations and Gabriel Vecchi for making his hurricane timeserie available and for suggesting this topic of research. For CMIP, the US Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals.

References

Bretherton CS, Widmann M, Dymnikov VP, Wallace JM, Bladé I (1999) The effective number of spatial degrees of freedom of a time-varying field. J Clim 12:1990–2009

Brayshaw DJ, Troccoli A, Fordham R, Methven J (2011) The impact of large scale atmospheric circulation patterns on wind power generation and its potential predictability: a case study over the UK. Renew Energy 36:2087–2096

Buell CE (1958) Meaning of combined climate and persistence forecast. J Meteor 15:564–565

Caron L-P, Jones CG, Winger K (2011) Impact of resolution and downscaling technique in simulating recent Atlantic tropical cyclone activity. Clim Dyn 5:869–892. doi:10.1007/s00382-010-0846-7

Delworth TL, et al. (2006) GFDL's CM2 global coupled climate models: part 1: formulation and simulation characteristics. J Clim 19(5) 643–674

Doblas-Reyes FJ, Andreu-Burillo I, Chikamoto Y, García-Serrano J, Guemas V, Kimoto M, Mochizuki T, Rodrigues LRL, van Oldenborgh GJ (2012) Near-term regional climate change prediction: impact of the initialization. Nature Comms (submitted)

Du H, Doblas-Reyes FJ, García-Serrano J, Guemas V, Soufflet Y, Wouters B (2012) Sensitivity of decadal predictions to the initial atmospheric and oceanic perturbations. Clim Dyn 39:2012–2023

Dufresne J-L, et al. (2012) Climate change projections using the IPSL-CM5 earth system model: from CMIP3 to CMIP5. Clim Dyn (Submitted)

Dunstone JJ, Smith DM, Eade R (2011) Multi-year predictability of the tropical Atlantic atmosphere driven by the high latitude North Atlantic Ocean. Geophys Res Lett 38:L1470. doi:10.1029/2011GL047949

García-Serrano J, Doblas-Reyes F-J, Coelho CAS (2012) Understanding Atlantic multi-decadal variability prediction skill. Geophys Res Lett 39:L18708. doi:10.1029/2012GL053283 (15-08-2012)

Gent PR, Danabasoglu G, Donner LJ, Holland MM, Hunke EC, et al. (2011) The community climate system model version 4. J Clim 24:4973–4991

Goddard L, Kumar A, Solomon A, Smith D, Boer G, et al. (2012) A verification framework for interannual-to-decadal predictions experiments. Clim Dyn. doi:10.1007/s00382-012-1481-1

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

Guemas V, Doblas-Reyes FJ, Andreu-Burillo I, Asif M (2012) Retrospective prediction of the global warming slowdown in the last decade.Under review for Nature Climate Change, NCLIM-12111103-T

ICPO (International CLIVAR Project Office) (2011) Data and bias correction for decadal climate predictions. International CLIVAR Project Office, CLIVAR Publication Series No. 150, p 5

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

Kharin VV, Boer GJ, Merryfield WJ, Scinocca JF, Lee W-S (2012) Statistical adjustment of decadal predictions in a changing climate. Geophys Res Lett 39:L19705. doi:10.1029/2012GL052647

Khasnis A, Nettleman MD (2005) Global warming and infectious disease. Arch Med Res 36:689–696. doi:10.1016/y.aremed.2005.03.041

Knapp KR, Kruk MC, Levinson DH, Diamond HJ, Neumann CJ (2010) The International Best Track Archive for Climate Stewardship (IBTrACS): Unifying tropical cyclone best track data. Bull Am Meteor Soc 91:363–376

Knight JR, Folland CK, Scaife AA (2006) Climate impacts of the Atlantic Multidecadal Oscillation. Geophys Res Lett 33:L17706

Knutson TR, Delworth TL, Dixon KW, Held IM, Lu J, Ramaswamy V, Schwarzkopf MD, Stenchikov G, Stouffer RJ (2006) Assessment of twentieth-century regional surface temperature trends uising the GFDL CM2 coupled models. J Clim 19:1624–1651

Landsea CW (2000) El Niño-Southern Oscillation and the seasonal predictability of tropical cyclones. In: Diaz HF, Markgraf V (eds) El Niño and the Southern Oscillation: multiscale variability and global and regional impacts. pp 149–181

Loeb NG, Lyman JM, Johnson GC, Allan RP, Doelling DR, Wong T, Soden BJ, Stephens GL (2012) Observed changes in top-of-the-atmosphere radiation and upper-ocean heating consistent within uncertainty. Nat Geosci 5:110–113. doi:10.1038/NGEO1375

Meehl GA, Goddard L, Kirtman B, Branstator G, Danabasoglu G, Hawkins E, et al. (2013) Decadal climate prediction. An update from the trenches. Bull Am Meteor Soc (Submitted)

Meinshausen M, Smith SJ, Calvin K, Daniel JS, Kainuma MLT, Lamarque JF, Matsumoto K, Montzka SA, Raper SCB, Riahi K, Thomson A, Velders GJM, van Vuuren DPP (2011) The RCP greenhouse gas concentrations and their extensions from 1765 to 2300. Clim Change 109(1–2):213–241

Mendelsohn R, Basist A, Dinar A, Kurukulasuriya P, Williams C (2007) What explains agricultural performance: climate normals or climate variance? Clim Change 81:85–99. doi:10.1007/s10584-006-9186-3

Merryfield WJ, Denis B, Fontecilla J-S, Lee W-S, Kharin V, Hodgson J, Archambault B (2011) The Canadian seasonal to interannual prediction system (CanSIPS). CMC technical report. Available online from http://collaboration.cmc.ec.gc.ca/cmc/cmoi/product_guide/docs/lib/op_systems/doc_opchanges/technote_cansips_20111124_e.pdf.

Murphy AH (1992) Climatology, persistence, and their linear combination as standard of reference in skill scores. Wea and Forecast 7:692–698

Rayner NA, Parker DE, Horton EB, Folland CK, Alexander LV, Rowell DP, Kent EC, Kaplan A, (2002) Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. J Geophys Res 108(D14):4407. doi:10.1029/2002JD002670

Ranger N, Niehörster F (2012) Deep uncertainty in long-term hurricane risk: scenario generation and implications for future climate experiments. Global Envir Change 22:703–712

Sakamoto TT, Komuro Y, Nishimura T, Ishii M, Tatebe H, Shiogame H, Hasegawa A, Toyoda T, Mori M, Suzuki T, Imada Y, Nozawa T, Takata K, Mochizuki T, Ogochi K, Emori S, Hasumi H,

Kimoto M (2013) MIROC4h: a new high-resolution atmosphere-ocean coupled genral circulation model. J Meteorol Soc Jpn. (Submitted)

Saunders M, Lea A (2011) Extended range forecast for Atlantic Hurricane activity in 2012. Tropical Storm Risk (TSR) technical report. Available online from http://tropicalstormrisk.com/docs/TSRATLForecastDec2012.pdf

Scoccimarro E, Gualdi S, Bellucci A, Sanna A, Fogli PG, Manzini E, Vichi M, Oddo P, Navarra A (2011) Effects of tropical cyclones on ocean heat transport in a high resolution coupled general circulation model. J Clim 24:4368–4384

Smith DM, Cusack S, Colman AW, Folland CK, Harris GR, Murphy JM (2007) Improved surface temperature prediction for the coming decade from a global climate model. Science 317: 796–799, doi:10.1126/science.1139540

Smith DM, Eade R, Dunstone NJ, Fereday D, Murphy JM, Pohlmann H, Scaife AA (2010) Skilful multi-year predictions of Atlantic hurricane frequency. Nat Geosci 3:846–849. doi:10.1038/ngeo1004

Smith DM, Scaife AA, Boer GJ, Caian M, Doblas-Reyes FJ, et al. (2012) Real-time multi-model decadal climate predictions. Clim Dyn. doi:10.1007/s00382-012-1600-0

Solomon A, Coauthors (2011) Distinguishing the roles of natural and anthropogenically forced decadal climate variability. Bull Am Meteor Soc 92:141–156

Taylor KE, Stouffer RJ, Meehl GA (2012) An overview of CMIP5 and the experimental design. Bull Am Meteor Soc 93. doi:10.1175/BAMS-D-11-00094.1

Vecchi GA, Msadek R, Anderson W, Chang Y-S, Delworth T, Dixon K, Gudgel R, Rosati A, Stern W, Villarini G, Wittenberg A, Yang X, Zeng F, Zhang R, Zhang S (2013) Multi-year predictions of North Atlantic hurricane frequency: promise and limitations. J Clim. doi:10.1175/JCLI-D-12-00464.1, (in press)

Vecchi GA, Zhao M, Wang H, Villarini G, Rosati A, Kumar A, Held IM, Gudgel R (2011) Statistical-dynamical predictions of seasonal North Atlantic hurricane activity. Mon Wea Rev 139(4):1070–1082

Vecchi GA, Knutson TR (2011) Estimating annual number of Atlantic hurricanes missing from the HURDAT database (1878–1965) using ship track density. J Clim 24:1736–1746

Voldoire A, Sánchez-Gómez E, Salas y Mélia D, Decharme B, Cassou C, Sénési S, Valcke S, Beau I, Alias A, Chevallier M, Déqué M, Deshayes J, Douville H, Fernandez E, Madec G, Maisonnave E, Moine M-P, Planton S, Saint-Martin D, Szopa S, Tyteca S, Alkama R, Belamari S, Braun A, Coquart L, Chauvin F (2012) The CNRM-CM5.1 global climate model: description and basic evaluation. Clim Dyn. doi:10.1007/s00382-011-1259-y

Watanabe TS, Oishi R, Komuro Y, Watanabe S, Emori S, Takemura T, Chikira M, Ogura T, Sekiguchi M, Takata K, Yamazaki D, Yokohata T, Nozawa T, Hasumi H, Tatebe H, Kimoto M (2010) Improved climate simulation by MIROC5: mean states, variability, and climate sensitivity. J Clim 23:6312–6335

Wu T, Li W, Ji J, Xin X, Li L, Wang Z, et al. (2012) Global carbon budgets simulated by the Beijing Climate Center Climate System Model for the last century. J Clim. (Submitted)

Yukimoto S, Adachi Y, Hosaka M, Sakami T, Yoshimura H, Hirabara M, Tanaka TY, Shindo E, Tsujino H, Deushi M, Mizuta R, Yabu S, Obata A, Nakano H, Koshiro T, Ose T, Kitoh A (2012) A new global climate model of Meteorological Research Institute: MRI-CGCM3 - model description and basic performance. J Meteor Soc Jpn 90A:23–64. doi:10.2151/jmsj.2012-A02

Zhang R, Delworth TL (2006) Impact of Atlantic multidecadal oscillations on India/Sahel rainfall and Atlantic hurricanes. Geophys Res Lett 33:L17712. doi:10.1029/2006GL026267