ATMOSPHERIC SCIENCE

Improved simulation of 19th- and 20th-century North Atlantic hurricane frequency after correcting historical sea surface temperatures

Duo Chan[1]*, Gabriel A. Vecchi[2], Wenchang Yang[2], Peter Huybers[1]

Confidence in dynamical and statistical hurricane prediction is rooted in the skillful reproduction of hurricane frequency using sea surface temperature (SST) patterns, but an ensemble of high-resolution atmospheric simulation extending to the 1880s indicates model-data disagreements that exceed those expected from documented uncertainties. We apply recently developed corrections for biases in historical SSTs that lead to revisions in tropical to subtropical SST gradients by ±0.1°C. Revised atmospheric simulations have 20% adjustments in the decadal variations of hurricane frequency and become more consistent with observations. The improved simulation skill from revised SST estimates not only supports the utility of high-resolution atmospheric models for hurricane projections but also highlights the need for accurate estimates of past and future patterns of SST changes.

INTRODUCTION

Changes in Atlantic hurricane activity as a consequence of anthropogenic climate variations remain uncertain (1, 2) but could have major societal implications (3, 4). Historic records show substantial multidecadal variations in Atlantic hurricane activity (5) that covary with sea surface temperature (SST) differences between the Atlantic main development region and the remainder of the tropics (5, 6). Both statistical (5, 7, 8) and dynamical models (9, 10) are skillful in reproducing variations in observational estimates of hurricane frequency over recent decades. This covariation supports an interpretation that SST variations are a proxy for variations in the thermodynamic potential for hurricane genesis associated with the temperature difference between the surface and tropical tropopause, as well as large-scale circulation changes influencing hurricane activity (11–13).

When extended to cover the late 19th century and the full 20th century with commonly used reconstructions of SSTs, however, models fail to capture the amplitude of multidecadal variations in reconstructed hurricane counts. For example, statistical models (5) based on tropical SST differences in HadISST1 (14) predict hurricane activity that is 17% weaker over 1885–1899 and 16% stronger over 1930–1955 than observational estimates (5, 8).

Discrepancies in the long-term relationship between reconstructed and modeled Atlantic hurricane counts may arise for a variety of reasons. These discrepancies could reflect errors in historical hurricane reconstructions. For example, before the satellite era, hurricane reconstructions must be corrected for missed events, a process that is inevitably uncertain (5, 7). The classification of hurricanes can be uncertain on account of errors in maximum wind speed estimates (15). Model-data discrepancies may also reflect an inadequacy of using SST variations alone to recover past hurricane activity. For example, upper-level atmospheric conditions have the potential to evolve independently of SSTs (8, 16). Recent simulations also indicate that global hurricane frequency decreases with increasing CO_2 independent of an SST influence (17–19).

An additional possibility is that errors in SST estimates corrupt past simulation skill. All widely used estimates of historical SST variability depend on in situ observations compiled under the International Comprehensive Ocean-Atmosphere Data Set (ICOADS) (20–22). These data require corrections to account for temporal and spatial inhomogeneity in available measurements (23–25). Before the 1980s, data come largely from measurements made using buckets, comprising 40% of observations between 1942 and 1981 and 95% of observations before 1942 (26). Bucket temperatures are estimated to be, on average, biased cool by 0.4°C over the early 20th century (26, 27), foremost because of cooling from wind-induced evaporation (23). Other biases are also present, however, such as those associated with heating from solar absorption (27, 28). The degree to which cooling and heating influence temperature measurements variously depends on the design of a bucket and measurement protocols.

Lack of metadata by which to make specific corrections has necessitated simplifying assumptions regarding the spatial and temporal structure of bucket biases. For example, HadISST1 uses globally uniform and linear weights to represent a transition from wooden buckets to less insulated canvas buckets (14). Correctly diagnosing the spatial and temporal evolution of these biases, however, is important because hurricanes and other climate phenomena are sensitive to patterns of tropical SST changes (13, 29–31).

A recently developed method allows for identifying systematic offsets among groups of ships and improved corrections of regional biases resulting from their uneven spatial and temporal sampling of SST (32). This method involves pairing nearby measurements from distinct groups and estimating systematic offsets using a linear mixed-effect (LME) model. Application of the method to version 3.0 of ICOADS (22) results in the detection of highly significant offsets (P<0.05) between nations and data collecting groups (32). Physical and historical evidence supports the validity of estimated groupwise offsets. Groupwise SST offsets have variously been found to arise as a consequence of truncation of observations during digitization, misclassifying engine room intake measurements as bucket SSTs, or as a result of buckets having different thermal insulation properties or remaining longer on ship deck before measurement (33, 34). Correcting for groupwise offsets leads to SST adjustments ranging between ±0.5°C for monthly 5° × 5° ocean grids and results in more

[1]Department of Earth and Planetary Sciences, Harvard University, Cambridge, MA, USA. [2]Department of Geosciences, Princeton University, Princeton, NJ, USA.
*Corresponding author. Email: duochan@g.harvard.edu

homogeneous patterns of warming over the early 20th century that are in better agreement with surface air temperature data from coastal weather stations (*33*). Moreover, groupwise adjustments also remove a significant but otherwise unexplained warm anomaly during World War II (*35*), bringing instrumental SST estimates into consistency with coastal and island station records (*36*) as well as marine proxies of SST (*37*).

Here, we explore whether biases in historical SST estimates are a major limiting factor in the ability of models to reproduce historical North Atlantic hurricane frequency and whether model-data discrepancies in SST-based predictions of historical hurricanes indicate a need to account for additional processes in models. To answer these questions, we perform a series of SST-forced atmospheric model simulations using the National Oceanic and Atmospheric Administration–Geophysical Fluid Dynamics Laboratory (NOAA-GFDL) High-Resolution Atmospheric Model (HiRAM) (*9*) and a 25-km version of the NOAA-GFDL Atmospheric Model version 2.5 (AM2.5) (*38*). These models skillfully simulate many aspects of the climatology of tropical cyclones (TCs) (*9, 39*) and are also skillful in reproducing the interannual variability of North Atlantic hurricane frequency in the satellite era, when high-frequency SST variability is well observed (fig. S1). We first use HadISST1 (*14*) to obtain an eight-member simulation ensemble (five from HiRAM and three from AM2.5) that is used as a control. We then obtain a second ensemble of simulations using a version of HadISST1 that is revised to account for groupwise bucket corrections and that is referred to as HadISST1b. See Materials and Methods for more details of groupwise SST adjustments and high-resolution atmospheric simulations.

RESULTS

Simulations prescribed with HadISST1 show only weak trends in North Atlantic hurricanes over the full interval (Fig. 1A and fig. S2A) but have clear decadal variations that are generally in phase with observational hurricane reconstructions (Fig. 1A). Similar to findings using statistical models (*5, 8*), simulated hurricane counts are lower in the late 19th century and higher in the middle 20th century (Fig. 1A), resulting in a significant model-data discrepancy. For example, during 1885–1899, simulations using HadISST1 yield, on average, 7.0±0.4 (2 SDs) hurricanes per year, a value that is 17% less ($P = 0.06$) than the value of 8.4±1.4 from the observational estimates (Fig. 1D). On the other hand, during 1930–1959, the simulated count averages 8.4±0.3 hurricanes per year, which is 22% higher ($P < 0.01$) than the observational value of 6.9±0.9. Uncertainties are reported as two standard errors if not otherwise specified. The error estimate for HadISST1-based simulations only includes atmospheric intrinsic variability, ϵ_i, uncertainties due to perturbing model's initial conditions. The error of observational hurricane frequency estimates accounts for both atmospheric intrinsic variability, ϵ_i, and reconstruction error, ϵ_o, summed in quadrature. It is necessary to consider ϵ_i in both simulation and observation because they are independent. Significance is evaluated using a two-sided z test and reported using P values; values smaller than 0.01 are reported as $P < 0.01$. See Materials and Methods for further details regarding uncertainty estimates.

Groupwise bucket SST adjustments lead to changes in patterns of SSTs that are sustained over decades (Fig. 2). Here, we focus on a relative pattern of SST defined as the difference between SST averaged over the North Atlantic main development region (20° to 80°W, 10° to 25°N) and the entire tropical ocean, referred to as relative
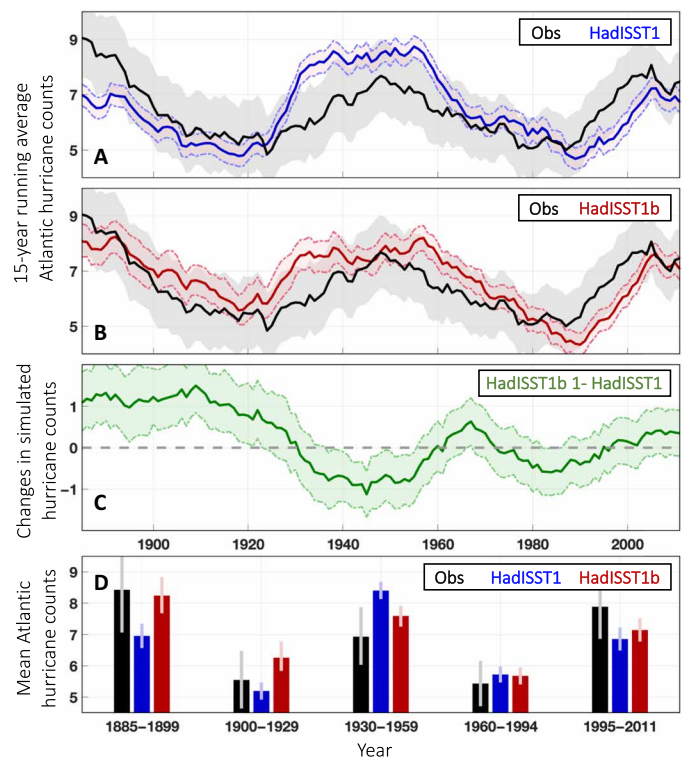

Fig. 1. Observed and simulated Atlantic hurricane counts. (**A**) Simulations using HadISST1 give lower hurricane counts in the late 19th century than observational estimates and higher counts in the mid-20th century. (**B**) Simulated and observed hurricane counts become consistent using HadISST1b, which includes groupwise bucket SST adjustments. (**C**) Difference in hurricane counts between simulations using HadISST1 and HadISST1b [(B) minus (A)]. Uncertainties are for atmospheric intrinsic variability and uncertainties in hurricane adjustments (added in quadrature, gray shading, 95% CI), atmospheric intrinsic variability (blue shading, 95% CI), and atmospheric intrinsic variability and uncertainties arising from uncertain SST adjustments (added in quadrature, red shading, 95% CI). Curves in (A) to (C) are 15-year running averages, with the initial (1878–1884) and final (2012–2018) 7 years truncated. (**D**) Mean hurricane counts over active and inactive periods where uncertainties (vertical bars, 95% CI) correspond to those in (A) and (B). Shown results combine five members of HiRAM runs and three members of AM2.5 runs, but an improvement in the skill of hurricane simulations is also found using the ensemble from either model (fig. S3).

SST (RSST) (*8*). RSST covaries with North Atlantic hurricane frequency (*8, 13*). RSST also covaries with the Atlantic Multidecadal Oscillation (AMO) index (*40*), although we focus on RSST because simulated hurricanes have a higher correlation with RSST than the AMO (figs. S4 and S5). Over 1885–1920, adjustments of SST data coming from Germany, the Netherlands, and a group of data referred to as deck number 156 because its nationality is unknown make the main development region 0.10°±0.05°C warmer relative to the rest of the tropics (Fig. 2A). Between 1930 and 1960, British and German SST adjustments result in colder SSTs in the Atlantic main development region, whereas Japanese and Dutch SST adjustments give warmer SSTs in the tropical Indian Ocean and the western Pacific, leading to a decrease in RSST by −0.05°±0.03°C (Fig. 2B). Bucket SST adjustments after the 1960s have a smaller magnitude because bucket measurements make up a smaller fraction of total measurements.
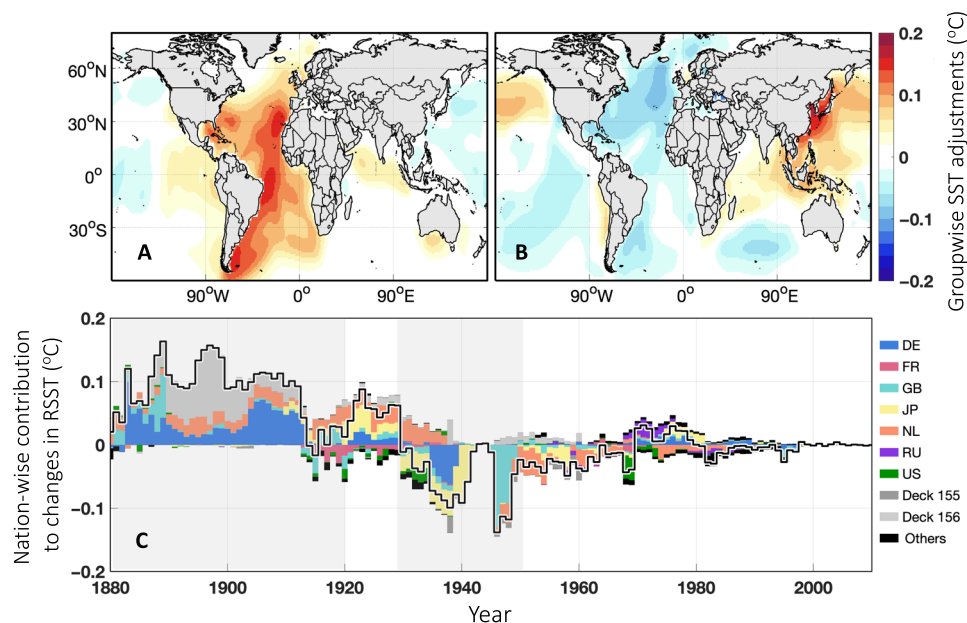
Fig. 2. Adjustments in the relative SST index. (**A** and **B**) Mean groupwise bucket SST adjustments incorporated in HadISST1b over 1885–1920 (A) and 1930–1949 (B). (**C**) Contributions from individual nations (stacked bars) to changes in RSST (black line). Nation abbreviations are for Germany (DE), France (FR), Great Britain (GB), Japan (JP), the Netherlands (NL), Russia (RU), and the United States (US). Groups without nation information are combined using deck information (such as deck 155 and 156), where deck is an indicator of marine data collectors in ICOADS (*22*). Note that changes in RSSTs tend toward a lower magnitude with time because groupwise bucket adjustments incorporated in HadISST1b are scaled by the fraction of bucket versus other measurements in individual grid boxes.

Changes in RSST associated with groupwise bucket adjustments are of comparable magnitude with internal and forced variations. Internal multidecadal variability in RSST simulated by 14 Coupled Model Intercomparison Project Phase 5 (CMIP5) models (*41*) has an SD of 0.11°C. Furthermore, changes in RSST during 2081–2100 relative to 1986–2005 climatology average only −0.02°C in the RCP4.5 ensemble, and changes in individual CMIP5 models used in an ensemble range from −0.34° to 0.40°C. The fact that biases and uncertainties in historical large-scale tropical SST patterns are of a magnitude similar to internal and forced variations has not previously been recognized.

Revising SSTs from HadISST1 to HadISST1b increases simulated Atlantic hurricane activity in the late 19th century and decreases activity in the middle 20th century (Fig. 1C). Changes in simulated hurricane count are consistent with changes in RSSTs and bring simulated hurricane activity into better alignment with historical observations (Figs. 1B and 3). Note that the uncertainty of HadISST1b-based simulations (Fig. 1B) contains not only atmospheric intrinsic variability but also errors associated with uncertain groupwise SST adjustments. Because of limited computing power, we estimate uncertainties associated with SST errors by converting perturbations in the RSST index to expected variations in hurricane counts. Specifically, we take an ensemble of random groupwise SST adjustments and scale each according to the sensitivity of hurricane counts to RSSTs. The sensitivity is estimated by regressing simulated hurricane counts against RSSTs at annual resolution (fig. S4). For the active period during 1885–1899, the simulated count increases from 7.0 ± 0.4 hurricanes per year when using HadISST1 to 8.2 ± 0.6 when using HadISST1b (Fig. 1D). Such an increase (*P*<0.01) is equivalent to an 18% fractional change relative to the climatological value of 6.6 hurricanes per year. HadISST1b-based hurricane count is statistically consistent with the observational estimate of 8.4 ± 1.4 hurricanes per year (Fig. 1D). During the next active period, between 1930 and 1959, the simulated count decreases from 8.4 ± 0.3 hurricanes per year when using HadISST1 to 7.6 ± 0.3 when using HadISST1b, where the latter is, again, consistent with the observed value of 6.9 ± 0.9 hurricanes per year.

Model skill in reproducing the decadal variability of historical Atlantic hurricane counts increases significantly when using HadISST1b. Model misfit is quantified using the root mean square error (RMSE) between 15-year running average observed and simulated hurricane counts. RMSE decreases from 1.06 hurricanes per year when using HadISST1 to 0.82 when using HadISST1b (Table 1). This decrease in RMSE is significant (*P* = 0.03) as assessed using a one-sided test against a null hypothesis that SST adjustments have no skill (Fig. 4A). The effect of unskillful SST adjustments is represented using a null distribution constructed by randomly permuting the mean difference between HadISST1- and HadISST1b-based simulations 10,000 times. The squared Pearson's correlation (r^2) between observed and simulated counts increases from 0.31 when using HadISST1 to 0.50 when using HadISST1b, an improvement that is also significant (*P* = 0.04; Fig. 4B). Results are robust to a variety of plausible alterations (Table 1), including examining ensemble members from individual models instead of combining them, using a 25-year window for running averages instead of 15 years or calibrating models by multiplying simulated hurricane counts instead of shifting the threshold level.

DISCUSSION
Our results reconcile the model-data discrepancy in Atlantic hurricane frequency at decadal time scales and do not indicate a need for
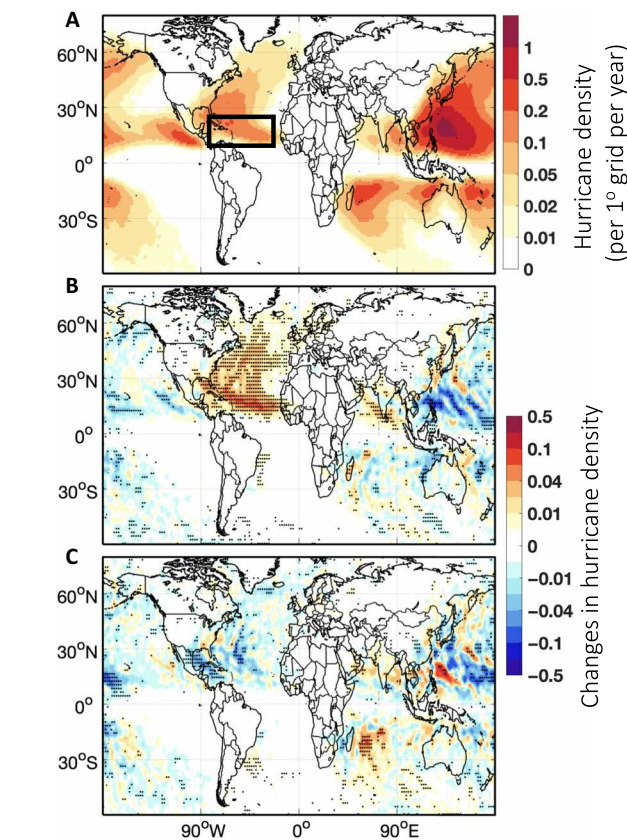
Fig. 3. Maps of hurricane track density. (**A**) Climatological hurricane track density averaged over 1885–1920 and eight members based on HadISST1. The Atlantic main development region is highlighted by a black box. (**B**) Ensemble-mean changes in simulated hurricane track density (HadISST1b- minus HadISST1-based simulations). Accounting for groupwise SST offsets significantly increases hurricane density in the North Atlantic over 1885–1920 (dots, $P<0.05$). (**C**) as (B) but for changes over 1930–1949. The pattern of changes in hurricane density is similar using either HiRAM or AM2.5 runs (fig. S6). For visualization purposes, hurricane track density on 1° gridding is smoothed using a nine-grid two-dimensional (2D) convolutional smoother.

additional processes to be represented within models. This consistency can be demonstrated using a simple model and error budget

$$H = \mathbf{F}(T) + \epsilon_i + \epsilon_o + \epsilon_T \qquad (1)$$

Hurricane count, H, not only is represented as a process that maps SSTs to expected hurricane count, $\mathbf{F}(T)$, but also is subject to additional stochastic and systematic terms, depending on whether counts are from observations or simulations. Observational estimates, H_{obs}, contain errors associated with atmospheric intrinsic variability, ϵ_i, and corrections for missed hurricanes, ϵ_o. Simulated hurricane counts, H_{sim}, are subject to atmospheric intrinsic variability, ϵ_i, and SST errors, ϵ_T.

We first construct a null distribution of the RMSE between simulated and observed hurricane counts under the assumption that $\mathbf{F}(T)$ is equivalent between the observations and simulations and that ϵ_T is negligible. In this scenario, the RMSE expected for HadISST1 is constructed only accounting for the ϵ_i and ϵ_o terms in $H_{obs} - H_{sim}$. ϵ_i is realized by drawing random time series 10,000 times from a zero-centered Gaussian distribution whose SD is the cross-member spread of hurricane counts within simulation ensembles. ϵ_o is also realized 10,000 times by randomly perturbing parameters in the hurricane adjustment algorithm (5). The null distribution of RMSE for HadISST1 gives a 95% confidence interval (CI) ranging from 0.43 to 0.93 hurricanes per year (Fig. 5A), whereas the actual difference between simulated and observed counts using HadISST1 has an RMSE of 1.06 hurricanes per year. Such a result gives the appearance that improvements in climate models or a reevaluation of errors in historical hurricane counts are required.

A revised null distribution based on HadISST1b reconciles the model-data discrepancy in two ways. First, the null distribution for HadISST1b additionally includes SST uncertainty, ϵ_T, in recognition of the fact that groupwise SST adjustments are correlated across space and time and only partially cancel under averaging. Accounting for ϵ_T widens the 95% CI of the null distribution to 0.44 to 0.96 hurricanes per year. Second, systematic errors in simulated hurricanes associated with SST biases are reduced, with the RMSE between observations and HadISST1b-based simulations decreasing to 0.82 hurricanes per year (Fig. 5B). As a result, the HadISST1b RMSE estimate is consistent with known error sources, supporting the accuracy of the current generation of models with respect to predicting changes in hurricane activity.

Table 1. Model skill in reproducing historical North Atlantic hurricane counts. Shown statistics are squared Pearson's correlation coefficient r^2 and RMSE between observational and ensemble-mean of simulated hurricane counts. We explore the sensitivity of results by investigating models of different resolutions (50 km HiRAM or 25 km AM2.5), turning off SST adjustments in the satellite era (splice), tracking simulated hurricanes using a threshold of 33 m/s and then calibrating by multiplying 1.2 (×1.2), and using smoothing windows of a different length (25 years). Statistics are for 1878–2018, but we omit an interval equal to half of the smoothing window from the beginning and end. Numbers in parentheses are P values of incorrectly rejecting the null hypothesis that increases in r^2 or decreases in RMSE arise from unskillful SST adjustments.

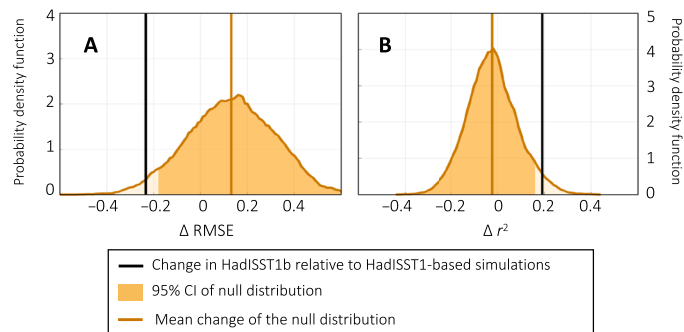|	r^2(HadISST1)	r^2(HadISST1b)	RMSE(HadISST1)	RMSE(HadISST1b)
Standard	0.31	0.50 ($P = 0.04$)	1.06	0.82 ($P = 0.03$)
50 km	0.23	0.42 ($P = 0.04$)	1.14	0.89 ($P = 0.04$)
25 km	0.42	0.57 ($P = 0.07$)	0.97	0.78 ($P = 0.05$)
Splice	0.31	0.47 ($P = 0.05$)	1.06	0.82 ($P = 0.03$)
×1.2	0.30	0.52 ($P = 0.02$)	1.18	0.97 ($P = 0.03$)
25 years	0.30	0.45 ($P = 0.08$)	0.91	0.72 ($P = 0.05$)

Fig. 4. Significant improvements in hurricane simulation skill. (**A**) Compared with HadISST1-based simulations, HadISST1b-based simulations show significantly lower RMSE ($P = 0.03$) and (**B**) higher correlation with observed hurricane counts (r^2; $P = 0.04$). Differences in statistics between HadISST1b- minus HadISST1-based simulations (black lines) are shown. RMSE and r^2 are computed between 1885 and 2011 after smoothing counts using a 15-year running average. Null distributions (gold shading) are constructed, assuming that groupwise SST adjustments to HadISST1 contain no skill (see Materials and Methods). Unskillful adjustment is expected to increase the RMSE between observed and simulated hurricanes and decrease r^2 (gold vertical lines).
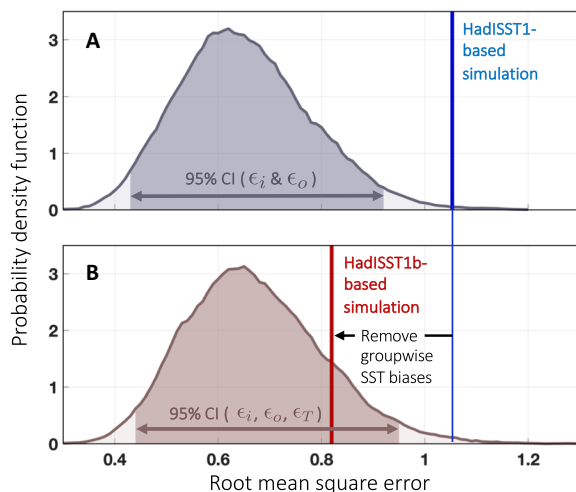

Fig. 5. Groupwise SST adjustments reconcile the model-data discrepancy in Atlantic hurricane frequency. (**A**) Atmospheric intrinsic variability (ϵ_i) and missed hurricane corrections (ϵ_o) are insufficient to explain the discrepancy between observational and HadISST1-based simulations (blue line). Here, the model-data discrepancy is quantified using RMSE and calculated using 15-year running average hurricane counts between 1885 and 2011. The null distribution (shading) is reconstructed using a Monte Carlo method by randomly realizing ϵ_i and ϵ_o for 10,000 times (see Materials and Methods). (**B**) as (A) but for the discrepancy between observational and HadISST1b-based simulations (red line). Accounting for groupwise SST offsets decreases biases in simulated hurricanes. Meanwhile, accounting for additional uncertainty arising from errors in groupwise SST adjustments (ϵ_T) widens the null distribution (shading).

Beyond the consistency of the HadISST1b scenario, however, there appears substantial scope for further reducing discrepancies between observed and model-reproduced hurricane counts. One line of future research is to continue to improve historical SST estimates to better evaluate model-data consistency. For example, engine room

intake measurements of SST, which are more prevalent in the second half of the 20th century, are potentially subject to systematic warm biases of several tenths of degrees Celsius associated with changes in sampling depth, engine room design, and conversion to hull-mounted sensors (24). Groupwise offsets and associated adjustments have, however, not yet been developed for engine room intake measurements, and mismatches over more recent decades might be reduced by adjusting engine room intake measurements. In addition, SST biases associated with individual ships may also contribute substantial uncertainty to regional SST patterns (42, 43). Offsets have been estimated and adjusted in HadISST1b after averaging ships coming from the same nation and data-collecting groups (33), but ships within the same group may have distinct SST biases that depend on sampling characteristics or ship design. Improvements in reconstructions of historical SSTs and hurricane counts are also possible as more historical ship logs are rescued (7, 22, 24). Removing biases in historical SSTs and hurricane simulations may also have implications for the detection and attribution of hurricane changes that are still limited by the relative small signal-to-noise ratio (1).

Improvements of SSTs could also arise from improving mapping techniques. For example, patterns of SST variability may vary with time and, thereby, differ from stationary SST patterns used in most mapping methods. Moreover, artificial biases may be introduced over data-sparse regions through mapping when remaining SST biases project onto large-scale SST patterns. Understanding the influence of mapping algorithms and possible interactions with bias corrections also appears an important goal for future work.

Further improvements in hurricane simulations are, of course, also possible. Climate models could be further improved through better resolving the structure of hurricanes and more fully incorporating relevant physical processes and environmental factors (1, 19). Furthermore, processes such as mid-tropospheric humidity (44) or global and regional increases in TC intensity (45) could alter in a warming climate in ways that are not sampled in the historical record. Inability to demonstrate that variations in historical hurricane counts demand improvement does not preclude other physical lines of evidence for where opportunities exist to improve prediction, but does imply that observationally testing for improved skill may be difficult.

Accurate projections of evolving SST patterns are known to be critical for predicting basin-scale changes in hurricane frequency (1). Our major finding is that biases in historical SST patterns were a dominant limiting factor in the ability of models to reproduce historical Atlantic hurricane counts at multidecadal time scales. Corrections to SST patterns significantly improves the model's reproduction skill. The remaining model-data mismatch could have arisen from atmospheric intrinsic variability and errors in hurricane reconstruction. Continued improvement of historical SST and hurricane estimates will facilitate more accurate tests of the skill of hurricane simulations.

MATERIALS AND METHODS
Observed North Atlantic hurricane counts
North Atlantic hurricane observations covering 1878–2018 come from the Best Track Data (HURDAT2) (46). We identify hurricanes as tropical storms in the North Atlantic that have a maximum sustained wind speed higher than 33 m/s. Annual counts of hurricanes in HURDAT2 are adjusted according to an estimate of missed hurricanes

before 1965 (5), which involves adding a correction factor to observed hurricane counts based on sampling satellite observations using early ship tracks in the ICOADS (21).

The uncertainty of the hurricane correction (ϵ_o) takes into account the year of satellite data used, the size of hurricanes, and the day in a year a storm was paired with observations, which yields an ensemble of 27,950 adjustment time series. Because we are interested in decadal variability of hurricane frequency, time series are 15-year running average, with the uncertainty estimated by drawing random samples from the adjustment ensemble. Specifically, for each year, 10,000 samples are randomly drawn from 27,950 possible values without replacement and under the assumption that years are independent. After smoothing the 10,000 random realizations of possible adjustments, ϵ_o is estimated to have an SD of 0.37 hurricanes per year between 1885 and 1964. Because of increasing numbers of ship tracks, ϵ_o decreases with time, from 0.44 hurricanes per year in the late 19th century to 0.23 hurricanes per year in the early 1960s. Note that although the number of missed hurricanes in individual years are integers and, therefore, follow a Poisson distribution, in practice, 15-year or longer averages are well approximated by Gaussian distributions on account of the central limit theorem.

Historical SST estimates

We use HadISST1 (14) as the baseline SST estimate. To explore the implication of groupwise bucket adjustments for the simulation of historical hurricanes, we merge groupwise bucket adjustments with HadISST1 to obtain a new estimate, which we call HadISST1b. Groupwise adjustments are estimated in (32, 33) and account for systematic offsets among groups of bucket SSTs. In short, offsets are estimated by applying the LME model to pairs of nearby measurements from distinct nation and data-collecting groups, where pairs are identified as the closest two measurements that are within 300 km and 2 days of one another. Offsets are estimated relative to the mean of all paired SST measurements, and regional and temporal variations in offsets for individual groups are simultaneously estimated. These offsets are then removed from individual SST measurements according to group, location, and year.

To merge groupwise bucket adjustments to HadISST1, these adjustments are averaged within $2° \times 2°$ grid boxes that contain bucket measurements. Because HadISST1 uses SST measurements from a variety of methods, not only buckets, groupwise bucket adjustments are multiplied by the ratio of bucket to all SST measurements in individual grids for each month. In the North Atlantic, adjustments are multiplied by a fraction that averages 97% before the 1940s but decreases to 16% after the 1980s because of the increasing prevalence of engine room intake and drifter measurements. Scaled adjustment fields are smoothed in space using a two-dimensional (2D) convolutional smoother with a spatial scale of five grid boxes, and fields are interpolated to global coverage using biharmonic spline interpolation, as encoded by Matlab's griddata function using the V4 method. Last, adjustments in individual boxes are tapered to zero according to an exponential decay with an 1100-km length scale or 10° at the equator.

Note that HadISST1 made use of satellite infrared observations after 1982 (14). When calculating the ratio of bucket measurements to scale groupwise adjustments, we assume that the mass of satellite observations is five times of that from simultaneous buoy and drifter measurements. To assess the influence of this assumption, we turn off groupwise bucket SST adjustments after 1982 and still find robust improvements in the skill of HiRAM and AM2.5 (Table 1).

Simulating North Atlantic hurricanes using prescribed SSTs

We explore a series of SST-forced atmospheric model simulations using the NOAA-GFDL HiRAM and the NOAA-GFDL AM2.5 model. HiRAM has the finite volume cubed-sphere dynamical core at a global 50-km resolution (180×180 grid points on each of the cube faces, or C180) at 32 vertical levels (9). AM2.5 has the finite volume cubed-sphere dynamical core at a global 25-km resolution (360 × 360 grid points on each of the cube faces, or C360) at 32 vertical levels (38). These models skillfully simulate many aspects of the climatology of TCs (9, 39) and are widely used for process-level studies of cyclone dynamics (17, 47–49). Using these models, we obtain two ensembles of time-varying SST-forced experiments from 1871 to 2018, one prescribed with HadISST1 and the other with HadISST1b. Each ensemble consists of eight members, with five from HiRAM and three from AM2.5. Individual members have small perturbations in their initial condition. Radiative forcing changes are prescribed from the CMIP5 historical scenario for 1871–2004 and from the CMIP5 RCP4.5 scenario for 2005–2018.

The tracking algorithm of tropical storms follows (50). If using the observational threshold of 33 m/s to identify simulated hurricanes, models average 5.8 hurricanes per year in the North Atlantic, whereas HURDAT2-based reconstructions have an average value of 6.6 hurricanes per year. We, therefore, relax the wind speed threshold to 31.7 m/s such that the long-term climatological hurricane counts in simulations equal observational reconstructions.

The spread in simulated hurricane counts across ensemble members is used to estimate contributions from atmospheric intrinsic variability, ϵ_i. The SD of ϵ_i is calculated for each eight-member ensemble using the departures from the respect ensemble mean, and the resulting two values are averaged, giving a value of 2.19 hurricanes per year. ϵ_i is effectively independent across years, having a lag-1 Pearson's r^2 of less than 0.01. It follows that the SD of ϵ_i for observed 15-year running average counts becomes $2.19/\sqrt{15}$, or 0.57 hurricanes per year. Averaging over an eight-member ensemble mean further decreases the SD of ϵ_i to 0.20 hurricanes per year. Although hurricanes are quantized, the average is again well approximated as Gaussian.

Simulated hurricane counts are also subject to uncertainties in SST. Uncertainties in regional SST patterns are almost an order of magnitude larger than previously recognized (33). Because estimating the sensitivity of hurricane counts to arbitrary regional SST patterns using HiRAM or AM2.5 would be computationally prohibitive, we focus on the pattern associated with the June to November mean RSST index. RSST is defined as $1.707 + 1.388 \, T'_{MDR} - 1.521 \, T'_{Trop}$ (8), where T'_{MDR} is the average SST anomalies over the North Atlantic main development region (20° to 80°W, 10° to 25°N) relative to the 1982–2005 climatology, and T'_{Trop} is the average SST anomaly over tropical ocean (30°S to 30°N). Note that the AMO index is largely collinear with the decadal variability of RSST, but RSST has a stronger linear covariance with simulated North Atlantic hurricanes (see figs. S4 and S5 for a comparison between RSST and AMO).

We use an ensemble of HadISST1b realizations to estimate uncertainty in RSSTs and the two atmospheric model ensembles to estimate the sensitivity of hurricane counts to RSSTs. An SD error associated with ϵ_{RSST} of 0.02°C is estimated from a 20-member ensemble of HadISST1b realizations that are randomly perturbed according to uncertainties in groupwise bucket adjustments. Ensemble-averaged hurricane counts in simulations with HadISST1b and HadISST1 are regressed against their respective variations in RSST,

giving a sensitivity of 7.30 ± 0.35 hurricanes per year per degree Celsius (2 SDs; see fig. S4). Our best estimate of the error in hurricane counts arising from uncertain SSTs, $7.30 \times \epsilon_{RSST} = \epsilon_T$, therefore, implies SDs between 0.23 hurricanes per year in the late 19th century and 0.07 hurricanes per year in the 1980s. Note that the persistence of groupwise SST errors makes their contribution to hurricane count uncertainty comparable in magnitude to atmospheric effects at decadal time scales. Sampling and random SST errors (27), on the other hand, cancel under regional and temporal averaging, and their contribution to 15-year smoothed RSST errors is only 0.01°C in SD. We, therefore, only account for uncertainties associated with groupwise SST adjustments.

Comparison with CMIP5 simulations

We compare the magnitude of historical SST adjustments with internal variability and the range of projected changes in CMIP5 simulation. Specifically, we use 8674 years of preindustrial simulations from 14 models to estimate the range of internal RSST variability at decadal time scales. For each pi-control simulation, RSSTs are first calculated from detrended SSTs ("tos" in CMIP5 outputs) averaged over June to November, and then smoothed using a 15-year running window. An SD of decadal variability (0.11°C) in CMIP5 RSSTs is calculated after concatenating 15-year smoothed RSSTs across CMIP5 models. We also use the "r1i1p1" member of historical and RCP4.5 runs from 17 CMIP5 models to estimate the distribution of projected changes in RSSTs. We quantify changes as the difference in unsmoothed RSSTs between 2081–2100 and 1986–2005. The 17 CMIP5 models are chosen to be consistent with (51) and are ACCESS1.0, ACCESS1.3, CanESM2, CCSM4, CMCC-CM, CSIRO Mk3.6.0, GFDL-CM3, GFDL-ESM2G, GFDL-ESM2M, GISS-E2-H*, GISS-E2-R*, HadGEM2-CC, HadGEM2-ES, MIROC-ESM, MIROC-ESM-CHEM*, MPI-ESM-LR, and NorESM1-M. Marker "*" indicates that the tos output from preindustrial runs is not available for these models. If we include all 37 CMIP5 models available, the SD of internal variability remains similar (0.10°C); changes in the RCP4.5 scenario have a wider range of −1.24° to 0.40°C compared with the −0.35° to 0.40°C range when using 17 models.

Significance test of improvements in model's skill

The significance of increases in model's reproduction skill, as measured by RMSE or squared cross-correlation (r^2), is assessed using a one-sided test against a null distribution, assuming that groupwise SST adjustments have no skill. The null distribution is realized using a Monte Carlo technique whereby the ensemble-mean difference between HadISST1- and HadISST1b-based simulations is permuted using 10-year blocks and then smoothed to mimic the effect of randomized SST adjustments. Changes in RMSE and r^2 obtained when introducing randomized SST adjustments are calculated from a total of 10,000 random realizations to construct the null hypothesis. The expected change is positive for RMSE and negative for r^2 because introducing perturbations having no skill will generally increase noise in reconstructions.

Model-data mismatches arising from known sources of uncertainties

We use a Monte Carlo method to obtain a null distribution of RMSE that could have arisen from atmospheric intrinsic variability, $\epsilon_{i,\,sim}$ and $\epsilon_{i,\,obs}$, hurricane adjustment errors, ϵ_o, and errors in SST patterns, ϵ_T (Fig. 5). The null distribution is calculated from 10,000 randomly realized time series of respective errors. Specifically, $\epsilon_{i,\,sim}$ and $\epsilon_{i,\,obs}$ are realized by first drawing time series of independent and identically distributed samples from a Gaussian distribution, i.e., $N(0, 2.19^2)$. These random time series are then smoothed and averaged to account for temporal and ensemble averaging. ϵ_o is realized by bootstrapping randomized hurricane corrections (see the "Observed North Atlantic hurricane counts" section). ϵ_T is realized using RSSTs in an ensemble of HadISST1b and then multiplied by the sensitivity factor 7.30 ± 0.35 (see the "Simulating North Atlantic hurricanes using prescribed SSTs" section).

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/7/26/eabg6931/DC1

REFERENCES AND NOTES

1. T. Knutson, S. J. Camargo, J. C. Chan, K. Emanuel, C.-H. Ho, J. Kossin, M. Mohapatra, M. Satoh, M. Sugi, K. Walsh, L. Wu, Tropical cyclones and climate change assessment: Part II: Projected response to anthropogenic warming. *Bull. Am. Meteorol. Soc.* **101**, E303–E322 (2020).
2. E. J. Cha, T. R. Knutson, T.-C. Lee, M. Ying, T. Nakaegawa, Third assessment on impacts of climate change on tropical cyclones in the typhoon committee region—Part II: Future projections. *Trop. Cyclone Res. Rev.* **9**, 75–86 (2020).
3. R. A. Pielke Jr., J. Gratz, C. W. Landsea, D. Collins, M. A. Saunders, R. Musulin, Normalized hurricane damage in the United States: 1900–2005. *Nat. Hazards Rev.* **9**, 29–42 (2008).
4. K. Emanuel, Global warming effects on U.S. hurricane damage. *Weather Clim. Soc.* **3**, 261–268 (2011).
5. G. A. Vecchi, T. R. Knutson, Estimating annual numbers of Atlantic hurricanes missing from the HURDAT database (1878–1965) using ship track density. *J. Clim.* **24**, 1736–1746 (2011).
6. L.-P. Caron, L. Hermanson, A. Dobbin, J. Imbers, L. Lledó, G. A. Vecchi, How skillful are the multiannual forecasts of Atlantic hurricane activity? *Bull. Am. Meteorol. Soc.* **99**, 403–413 (2018).
7. G. A. Vecchi, T. R. Knutson, On estimates of historical North Atlantic tropical cyclone activity. *J. Clim.* **21**, 3580–3600 (2008).
8. G. A. Vecchi, R. Msadek, W. Anderson, Y.-S. Chang, T. Delworth, K. Dixon, R. Gudgel, A. Rosati, B. Stern, G. Villarini, A. Wittenberg, X. Yang, F. Zeng, R. Zhang, S. Zhang, Multiyear predictions of North Atlantic hurricane frequency: Promise and limitations. *J. Clim.* **26**, 5337–5357 (2013).
9. M. Zhao, I. M. Held, S.-J. Lin, G. A. Vecchi, Simulations of global hurricane climatology, interannual variability, and response to global warming using a 50-km resolution GCM. *J. Clim.* **22**, 6653–6678 (2009).
10. T. Knutson, J. L. McBride, J. Chan, K. Emanuel, G. Holland, C. Landsea, I. Held, J. P. Kossin, A. Srivastava, M. Sugi, Tropical cyclones and climate change. *Nat. Geosci.* **3**, 157–163 (2010).
11. K. Emanuel, The maximum intensity of hurricanes. *J. Atmos. Sci.* **45**, 1143–1155 (1988).
12. A. H. Sobel, J. Nilsson, L. M. Polvani, The weak temperature gradient approximation and balanced tropical moisture waves. *J. Atmos. Sci.* **58**, 3650–3665 (2001).
13. G. A. Vecchi, B. J. Soden, Effect of remote sea surface temperature change on tropical cyclone potential intensity. *Nature* **450**, 1066–1070 (2007).
14. N. A. Rayner, D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell, E. C. Kent, A. Kaplan, Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res. Atmos.* **108**, D14 (2003).
15. C. W. Landsea, S. Feuer, A. Hagen, D. A. Glenn, J. Sims, R. Perez, M. Chenoweth, N. Anderson, A reanalysis of the 1921–30 Atlantic hurricane database. *J. Clim.* **25**, 865–885 (2012).
16. K. Emanuel, S. Solomon, D. Folini, S. Davis, C. Cagnazzo, Influence of tropical tropopause layer cooling on Atlantic hurricane activity. *J. Clim.* **26**, 2288–2301 (2013).
17. I. M. Held, M. Zhao, The response of tropical cyclone statistics to an increase in CO2 with fixed sea surface temperatures. *J. Clim.* **24**, 5353–5364 (2011).
18. M. Sugi, J. Yoshimura, Decreasing trend of tropical cyclone frequency in 228-year high-resolution AGCM simulations. *Geophys. Res. Lett.* **39**, L19805 (2012).
19. G. A. Vecchi, T. L. Delworth, H. Murakami, S. D. Underwood, A. T. Wittenberg, F. Zeng, W. Zhang, J. W. Baldwin, K. T. Bhatia, W. Cooke, J. He, S. B. Kapnick, T. R. Knutson, G. Villarini, K. van der Wiel, W. Anderson, V. Balaji, J. Chen, K. W. Dixon, R. Gudgel, L. M. Harris, L. Jia, N. C. Johnson, S.-J. Lin, M. Liu, C. H. J. Ng, A. Rosati, J. A. Smith, X. Yang,

Tropical cyclone sensitivities to CO2 doubling: Roles of atmospheric resolution, synoptic variability and background climate changes. *Clim. Dyn.* **53**, 5999–6033 (2019).

20. S. Woodruff, H. Diaz, J. Elms, S. Worley, COADS Release 2 data and metadata enhancements for improvements of marine surface flux fields. *Phys. Chem. Earth* **23**, 517–526 (1998).

21. S. J. Worley, S. D. Woodruff, R. W. Reynolds, S. J. Lubker, N. Lott, ICOADS release 2.1 data and products. *Int. J. Climatol.* **25**, 823–842 (2005).

22. E. Freeman, S. D. Woodruff, S. J. Worley, S. J. Lubker, E. C. Kent, W. E. Angel, D. I. Berry, P. Brohan, R. Eastman, L. Gates, W. Gloeden, Z. Ji, J. Lawrimore, N. A. Rayner, G. Rosenhagen, S. R. Smith, ICOADS Release 3.0: A major update to the historical marine climate record. *Int. J. Climatol.* **37**, 2211–2232 (2017).

23. C. Folland, D. Parker, Correction of instrumental biases in historical sea surface temperature data. *Q. J. R. Meteorol. Soc.* **121**, 319–367 (1995).

24. E. C. Kent, J. J. Kennedy, T. M. Smith, S. Hirahara, B. Huang, A. Kaplan, D. E. Parker, C. P. Atkinson, D. I. Berry, G. Carella, Y. Fukuda, M. Ishii, P. D. Jones, F. Lindgren, C. J. Merchant, S. Morak-Bozzo, N. A. Rayner, V. Venema, S. Yasui, H.-M. Zhang, A call for new approaches to quantifying biases in observations of sea surface temperature. *Bull. Am. Meteorol. Soc.* **98**, 1601–1616 (2017).

25. E. C. Kent, N. A. Rayner, D. I. Berry, R. Eastman, V. Grigorieva, B. Huang, J. J. Kennedy, S. R. Smith, K. M. Willett, Observing requirements for long-term climate records at the ocean surface. *Front. Mar. Sci.* **6**, 441 (2019).

26. J. J. Kennedy, N. A. Rayner, R. O. Smith, D. E. Parker, M. Saunby, Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 2. biases and homogenization. *J. Geophys. Res. Atmos.* **116**, D14104 (2011).

27. J. J. Kennedy, N. A. Rayner, C. P. Atkinson, R. E. Killick, An ensemble data set of sea surface temperature change from 1850: The Met Office Hadley Centre HadSST. 4.0.0.0 data set. *J. Geophys. Res. Atmos.* **124**, 7719–7763 (2019).

28. G. Carella, J. Kennedy, D. Berry, S. Hirahara, C. J. Merchant, S. Morak-Bozzo, E. Kent, Estimating sea surface temperature measurement methods using characteristic differences in the diurnal cycle. *Geophys. Res. Lett.* **45**, 363–371 (2018).

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

30. G. A. Vecchi, M. Zhao, H. Wang, G. Villarini, A. Rosati, A. Kumar, I. M. Held, R. Gudgel, Statistical–dynamical predictions of seasonal North Atlantic hurricane activity. *Mon. Weather Rev.* **139**, 1070–1082 (2011).

31. G. Villarini, G. A. Vecchi, Twenty-first-century projections of North Atlantic tropical storms from CMIP5 models. *Nat. Clim. Chang.* **2**, 604–607 (2012).

32. D. Chan, P. Huybers, Systematic differences in bucket sea surface temperature measurements among nations identified using a linear-mixed-effect method. *J. Clim.* **32**, 2569–2589 (2019).

33. D. Chan, E. C. Kent, D. I. Berry, P. Huybers, Correcting datasets leads to more homogeneous early-twentieth-century sea surface warming. *Nature* **571**, 393–397 (2019).

34. D. Chan, P. Huybers, Systematic differences in bucket sea surface temperatures caused by misclassification of engine room intake measurements. *J. Clim.* **33**, 7735–7753 (2020).

35. D. Chan, P. Huybers, Correcting observational biases in sea-surface temperature observations removes anomalous warmth during World War II. *J. Clim.* **1**, 44 (2021).

36. K. Cowtan, R. Rohde, Z. Hausfather, Evaluating biases in sea surface temperature records using coastal weather stations. *Q. J. R. Meteorol. Soc.* **144**, 670–681 (2018).

37. M. Pfeiffer, J. Zinke, W.-C. Dullo, D. Garbe-Schönberg, M. Latif, M. E. Weber, Indian Ocean corals reveal crucial role of World War II bias for twentieth century warming estimates. *Sci. Rep.* **7**, 14434 (2017).

38. H. Murakami, G. A. Vecchi, S. Underwood, T. L. Delworth, A. T. Wittenberg, W. G. Anderson, J.-H. Chen, R. G. Gudgel, L. M. Harris, S.-J. Lin, F. Zeng, Simulation and prediction of category 4 and 5 hurricanes in the high-resolution GFDL HiFLOR coupled climate model. *J. Clim.* **28**, 9058–9079 (2015).

39. M. Zhao, I. M. Held, G. A. Vecchi, Retrospective forecasts of the hurricane season using a global atmospheric model assuming persistence of SST anomalies. *Mon. Weather Rev.* **138**, 3858–3868 (2010).

40. M. E. Schlesinger, N. Ramankutty, An oscillation in the global climate system of period 65–70 years. *Nature* **367**, 723–726 (1994).

41. K. E. Taylor, R. J. Stouffer, G. A. Meehl, An overview of CMIP5 and the experiment design. *Bull. Am. Meteorol. Soc.* **93**, 485–498 (2012).

42. E. Kent, D. Berry, Assessment of the marine observing system (ASMOS): Final report (2008).

43. J. Kennedy, R. Smith, N. Rayner, Using AATSR data to assess the quality of in situ sea-surface temperature observations for climate studies. *Remote Sens. Environ.* **116**, 79–92 (2012).

44. C.-Y. Lee, S. J. Camargo, A. H. Sobel, M. K. Tippett, Statistical–dynamical downscaling projections of tropical cyclone activity in a warming climate: Two diverging genesis scenarios. *J. Clim.* **33**, 4815–4834 (2020).

45. G. Villarini, G. A. Vecchi, Projected increases in North Atlantic tropical cyclone intensity from CMIP5 models. *J. Clim.* **26**, 3231–3240 (2013).

46. C. W. Landsea, J. L. Franklin, Atlantic hurricane database uncertainty and presentation of a new database format. *Mon. Weather Rev.* **141**, 3576–3592 (2013).

47. M. Zhao, I. M. Held, S.-J. Lin, Some counterintuitive dependencies of tropical cyclone frequency on parameters in a GCM. *J. Atmos. Sci.* **69**, 2272–2283 (2012).

48. T. M. Merlis, M. Zhao, I. M. Held, The sensitivity of hurricane frequency to ITCZ changes and radiatively forced warming in aquaplanet simulations. *Geophys. Res. Lett.* **40**, 4109–4114 (2013).

49. T. Knutson, J. Sirutis, M. Zhao, R. E. Tuleya, M. Bender, G. A. Vecchi, G. Villarini, D. Chavas, Global projections of intense tropical cyclone activity for the late twenty-first century from dynamical downscaling of CMIP5/RCP4.5 scenarios. *J. Clim.* **28**, 7203–7224 (2015).

50. L. M. Harris, S.-J. Lin, C. Tu, High-resolution climate simulations using GFDL HiRAM with a stretched global grid. *J. Clim.* **29**, 4293–4314 (2016).

51. K. Bhatia, G. Vecchi, H. Murakami, S. Underwood, J. Kossin, Projected response of tropical cyclone intensity and intensification in a global climate model. *J. Clim.* **31**, 8281–8303 (2018).

52. K. E. Trenberth, D. J. Shea, Atlantic hurricanes and natural variability in 2005. *Geophys. Res. Lett.* **33**, (2006).

Citation: D. Chan, G. A. Vecchi, W. Yang, P. Huybers, Improved simulation of 19th- and 20th-century North Atlantic hurricane frequency after correcting historical sea surface temperatures. *Sci. Adv.* **7**, eabg6931 (2021).

Science Advances

Improved simulation of 19th- and 20th-century North Atlantic hurricane frequency after correcting historical sea surface temperatures

Duo Chan, Gabriel A. Vecchi, Wenchang Yang and Peter Huybers