Global and Regional Discrepancies between Early-Twentieth-Century Coastal Air and Sea Surface Temperature Detected by a Coupled Energy-Balance Analysis

DUO CHAN,[a] GEOFFREY GEBBIE,[a] AND PETER HUYBERS[b]

[a] *Department of Physical Oceanography, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts*
[b] *Department of Earth and Planetary Sciences, Harvard University, Cambridge, Massachusetts*

(Manuscript received 29 July 2022, in final form 5 November 2022)

ABSTRACT: A major uncertainty in reconstructing historical sea surface temperature (SST) before the 1990s involves correcting for systematic offsets associated with bucket and engine-room intake temperature measurements. A recent study used a linear scaling of coastal station-based air temperatures (SATs) to infer nearby SSTs, but the physics in the coupling between SATs and SSTs generally gives rise to more complex regional air–sea temperature differences. In this study, an energy-balance model (EBM) of air–sea thermal coupling is adapted for predicting near-coast SSTs from coastal SATs. The model is shown to be more skillful than linear-scaling approaches through cross-validation analyses using instrumental records after the 1960s and CMIP6 simulations between 1880 and 2020. Improved skill primarily comes from capturing features reflecting air–sea heat fluxes dominating temperature variability at high latitudes, including damping high-frequency wintertime SAT variability and reproducing the phase lag between SSTs and SATs. Inferred near-coast SSTs allow for intercalibrating coastal SAT and SST measurements at a variety of spatial scales. The 1900–40 mean offset between the latest SST estimates available from the Met Office (HadSST4) and SAT-inferred SSTs range between −1.6°C (95% confidence interval: [−1.7°, −1.4°C]) and 1.2°C ([0.8°, 1.6°C]) across 10° × 10° grids. When further averaged along the global coastline, HadSST4 is significantly colder than SAT-inferred SSTs by 0.20°C ([0.07°, 0.35°C]) over 1900–40. These results indicate that historical SATs and SSTs involve substantial inconsistencies at both regional and global scales. Major outstanding questions involve the distribution of errors between our intercalibration model and instrumental records of SAT and SST as well as the degree to which coastal intercalibrations are informative of global trends.

SIGNIFICANCE STATEMENT: To evaluate the consistency of instrumental surface temperature estimates before the 1990s, we develop a coupled energy-balance model to intercalibrate measurements of sea surface temperature (SST) and station-based air temperature (SAT) near global coasts. Our model captures geographically varying physical regimes of air–sea coupling and outperforms existing methods in inferring regional SSTs from SAT measurements. When applied to historical temperature records, the model indicates significant discrepancies between inferred and observed SSTs at both global and regional scales before the 1960s. Our findings suggest remaining data issues in historical temperature archives and opportunities for further improvements.

KEYWORDS: Sea surface temperature; Air-sea interaction; Climate change; Surface temperature; Bias

1. Introduction

Sea surface temperature (SST) estimates are crucial for a wide range of climate studies but historical estimates before the 1960s are subject to systematic uncertainties of several tenths of a degree Celsius that are comparable in magnitude to the global climate change signal. These uncertainties arise mainly from systematic offsets between bucket and engine-room intake temperatures (Kent and Taylor 2006; Kent et al. 2010; Kennedy et al. 2011; Kent et al. 2017; Hausfather et al. 2017; Kennedy et al. 2019a; Kent and Kennedy 2021; Chan 2021). Physical methods are available to estimate biases (e.g., Folland and Parker 1995), but a lack of reliable metadata necessitates assumptions regarding instrumentation and measurement protocols (Folland and Parker 1995; Kennedy et al. 2011, 2019a) that are themselves inevitably uncertain.

Other corrections seek external constraints and correct SSTs against reference temperatures. For example, nighttime marine air temperatures (NMATs) are widely used (Smith and Reynolds 2002; Huang et al. 2017), despite being biased because of increasing ship height (Kent et al. 2013) and wartime practices of reading temperatures inside ships (Folland et al. 1984). Moreover, after data collection, NMATs are often postprocessed together with SSTs, which could lead to covarying biases due to systematic data-management issues (Chan et al. 2019). Another reference is the uppermost temperatures from marine profiles (MPs; Kennedy et al. 2019a), but profile data have limited coverage before the 1940s (Meyssignac et al. 2019). Although NMATs and MPs are practical ways of correcting SSTs, their limitations make exploring alternative references or methodologies worthwhile.

A method for correcting historical SSTs was proposed by Cowtan et al. (2018) that involves referencing SSTs against air temperatures from coastal weather stations. Although station-

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/JCLI-D-22-0569.s1.

Corresponding author: Duo Chan, duo.chan@whoi.edu

DOI: 10.1175/JCLI-D-22-0569.1

based air temperatures (SATs) could be biased due to urbanization (Karl et al. 1988) or changes in instrumentation (Trewin 2010), they are generally estimated to have higher quality than early oceanic measurements because SATs have more consistent measurement techniques (Jones 2016). We expect SATs to be indicative of nearby SSTs, but the two temperatures could diverge for many reasons, including but not limited to differences in radiative absorption and emission, evaporation, heat capacity, and sampling height. Orography and patterns of atmospheric circulation could also influence differences between SATs and SSTs. To account for such differences, Cowtan et al. (2018) scaled SATs linearly using a globally uniform factor, estimated by matching global-mean trends of their corrected SSTs and HadSST3 (Kennedy et al. 2011) after the 1970s. Although such a simple scaling factor may be feasible for global mean estimates, it is likely insufficient to account for regional differences between SATs and SSTs. Moreover, different heat capacities between air and water imply a frequency dependence in the ratio of variability between SATs and SSTs (Barsugli and Battisti 1998).

These considerations suggest that an energy-balance model (EBM) could be a useful way of inferring SSTs from SATs. Barsugli and Battisti (1998) proposed a linearized model to study the power spectra, total variance, and lag covariance between MATs and SSTs in midlatitudes. This model was also extended to account for wind-driven forcing (Lee et al. 2008) and advective processes (Saravanan and McWilliams 1998). These coupled EBMs provide a simple but physical framework for intercalibrating SSTs and air temperatures. In this paper, we explore the degree to which such a physically based model of air–sea coupling improves the inference of SSTs from nearby coastal SATs. Comparing SAT-inferred SSTs with the most up-to-date SST estimates, we also explore implications for further improving the quality of historical Earth surface temperatures at global and regional scales.

2. Data and methods

a. Data

1) Observational station temperatures

Station-based land air temperatures (SATs) are from weather stations compiled within the monthly resolution Global Historical Climatology Network (GHCNm) version 4 (Menne et al. 2018a). These stations have undergone a homogenization process (Menne and Williams 2009) involving comparison of observations relative to neighboring stations (Menne et al. 2018b), but we also analyze the sensitivity of results using the unhomogenized version. We identify stations that are within 10 km of the nearest coast using the GHCN metadata compiled by Cowtan et al. (2018). Arctic stations poleward of 60°N and stations in the Baltic and Mediterranean region (28°–90°N, 0°–52°E) are excluded on account of not being representative of open ocean conditions (Cowtan et al. 2018). In total, we identify 3111 coastal stations that are, on average, 2 km inland from a coast (Fig. 1).

SAT anomalies are computed relative to a climatological period of 1982–2014, an interval coinciding with the availability of a high-resolution SST climatology (Reynolds et al. 2007). Anomalies are calculated if an SAT station has at least 16 years of data that each contains at least 6 months of data during the climatological period. In total, there are 1700 stations whose SATs satisfy these criteria, which we call tier-1 stations (Fig. 1a). Anomalies for other land-temperature stations are estimated using a technique adapted from a station homogenization protocol (Menne and Williams 2009) that is detailed in appendix A. Using this method, we calculate anomalies for 1340 more stations, which we call tier-2 stations (Fig. 1b).

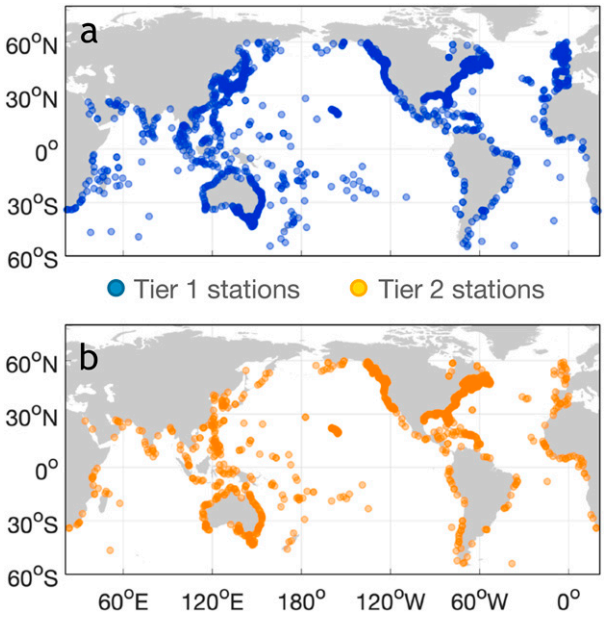

FIG. 1. Coastal stations used in this study. Coastal weather stations from GHCNmV4 are grouped into two tiers. (a) Tier-1 stations (blue) have sufficient data coverage between 1982 and 2014, and we calculate their temperature anomalies by subtracting the climatological average over 1982–2014. (b) Tier-2 stations (orange) have insufficient data coverage during the climatological period, and we evaluate their anomalies by pairing and matching temperature anomalies with nearby neighbors during overlapping months.

2) Observational SSTs

Observational SSTs are from HadSST4 (Kennedy et al. 2019b), a dataset containing estimates of monthly SST anomalies at 5° resolution. HadSST4 uses marine profile measurements to adjust SST biases after the 1940s (Kennedy et al. 2019a), although these estimates are subject to uncertainties in marine profiles (Meyssignac et al. 2019). We use the median of a 200-member HadSST4 ensemble to estimate parameters of an EMB that we then use to infer near-coast SSTs from SATs (see section 2b). Estimating EBM parameters requires matching HadSST4 with GHCNmV4 SAT data after 1960. Specifically, we sample HadSST4 at the grid box containing each SAT station. If an SST estimate is not available, we infill using the nearest neighbor SST grid or, if no adjoining grid box has available data, set SSTs as missing.

When comparing SAT-inferred SSTs with HadSST4 along global coasts (see section 4a), quantifying uncertainty in each estimate becomes crucial, especially before the 1940s when uncertainties associated with bias corrections are high. In addition to the bias correction uncertainties represented in the HadSST4 200-member ensemble, we also account for uncertainties arising from sampling and random measurement error by perturbing each member once using independently drawn samples from a zero-mean Gaussian distribution, where standard deviation estimates are provided for each grid box at monthly resolution (Kennedy et al. 2019b). Further, uncertainties associated with ship-level biases are accounted for by drawing spatially correlated uncertainties using monthly resolution covariance matrices (Kennedy et al. 2019b).

3) CLIMATE MODEL SIMULATIONS

Simulations from the Coupled Model Intercomparison phase 6 (CMIP6; Eyring et al. 2016) are used to evaluate the skill of our proposed methodology and estimate uncertainties of inferred near-coast SSTs. Specifically, we use surface air temperature (tas) and sea surface temperatures (tos) from the r1i1p1f1 member of 17 models: ACCESS-CM2, CAMS-CSM1-0, CMCC-CM2-SR5, E3SM-1-1, EC-Earth3, EC-Earth3-Veg, EC-Earth3-Veg-LR, FGOALS-f3-L, FGOALS-g3, FIO-ESM-2-0, INM-CM4-8, INM-CM5-0, MIROC6, MRI-ESM2-0, NESM3, NorESM2-LM, and NorESM2-MM. Historical all-forcing and SSP5.85 experiments are concatenated to cover 1850–2020. Simulations, whose original resolution ranges from 0.7° to 2.5° for the atmosphere and from 0.25° to 1° for the ocean, are regridded to a common 1° resolution using bilinear interpolation and sampled at locations of coastal weather stations. Sampled SATs and SSTs are further masked to have the same coverage as found in corresponding observations.

b. Coupled air–sea model

For purposes of inferring SSTs from SATs, we consider a coupled EBM framework based on Barsugli and Battisti (1998),

$$\begin{aligned} T_a' &= \beta T_s', \\ \gamma_a \frac{\partial T_a'}{\partial t} &= -\lambda_a T_a' + \kappa_a (T_o' - T_a') + F', \\ \gamma_o \frac{\partial T_o'}{\partial t} &= -\lambda_o T_o' + \kappa_o (T_a' - T_o') + kF', \end{aligned} \tag{1}$$

where T_s, T_a, and T_o denote, respectively, station air temperature, marine air temperature, and sea surface temperature; F represents stochastic heating from radiation or dynamical forcing; primed quantities, i.e., X', denote anomalies relative to a climatological state; and γ, λ, and κ denote heat capacity, damping, and thermal coupling efficiency, respectively.

This model differs from that of Barsugli and Battisti (1998) in two respects. First, to account for the possibility that temperature anomalies are amplified over land (Byrne and O'Gorman 2018), we distinguish T_s over the land from T_a over the ocean and assume that anomalous T_s and T_a variations are proportional to each other with coefficient β. Second, we append a kF' term to the SST equation, equivalent to that used by Lee et al. (2008) to represent dynamical fluxes associated with wind-induced Ekman transport. In the present context, kF' represents radiative forcing acting on both land and ocean that have different albedos and evaporative cooling, or a dynamical heat flux convergence that is correlated between the atmosphere and the ocean, though not necessarily of the same magnitude.

Combining terms and rearranging Eq. (1) gives an expression for the time rate of change of T_o',

$$\frac{\partial T_o'}{\partial t} = A \frac{\partial T_s'}{\partial t} - BT_o' + CT_s', \tag{2}$$

where $A = \beta k \gamma_a / \gamma_o$ is the relative sensitivity of stochastic external forcing between air and sea temperatures, $B = (\lambda_o + \kappa_o + k\kappa_a)/\gamma_o$ is effective thermal restoring, and $C = \beta(k\lambda_a + \kappa_o + k\kappa_a)/\gamma_o$ is effective ocean–atmosphere heat exchange.

Our overall approach is fitting Eq. (2) to recent observations and using estimated parameters A, B, and C to infer coastal SSTs from nearby station temperatures throughout the historical period. Specifically, we use data from the 1960s to fit parameters because we expect HadSST4 to be more reliable after conductivity–temperature–depth (CTD) profiles became available (Meyssignac et al. 2019). The fitting minimizes the mean squared difference between HadSST4 and predicted SSTs using Eq. (2). Details of how to integrate Eq. (2) using monthly SATs are in appendix B. Observational temperatures are binned to 10° resolution before fitting. Although fitting to individual stations is noisier because of observational error and missing values, results from individual stations give consistent estimates after fitted parameters are averaged to 10° grids.

3. Results

a. Exploring the behavior of the intercalibration model

There are two instructive limiting cases associated with Eq. (2). If A is large such that $A(\partial T_s'/\partial t)$ dominates over $-BT_o' + CT_s'$, SATs are forced by heat fluxes from the ocean and closely follow SSTs because of the air's small heat capacity. Under this limit, Eq. (2) predicts SSTs as a linear rescaling of SATs, consistent with the method used by Cowtan et al. (2018). This limiting case is more prevalent at lower latitudes, for example, in the eastern equatorial Pacific (EEP; Fig. 2a) where SATs and SSTs (black and gray) show a nearly one-to-one correspondence. The best fit of Eq. (2) to observed temperatures at this site yields an A of 0.65, whereas B is 2.5×10^{-7} s^{-1} and C is 2.4×10^{-7} s^{-1}. As expected for this regime, EBM-fitted SSTs (orange) are consistent with linearly scaled SATs (blue).

In the other limiting case, A is small, and Eq. (2) takes the form of a Hasselmann-type model (Hasselmann 1976). In our model, this regime is associated with atmospheric driving of SST anomalies, where the thermal inertia of the ocean mixed layer leads to an increasing ratio of spectral energy between SATs and SSTs with increasing frequency. This regime is more representative of high latitudes, as seen, for example, in coastal Alaska (Fig. 2c), where SATs have a variance that is 5.9 times that of SSTs. The best fit yields an A of 0.04, B of 0.8×10^{-7} s^{-1}, and C of 0.5×10^{-7} s^{-1}. EBM-fitted SSTs (orange line in Fig. 2c)

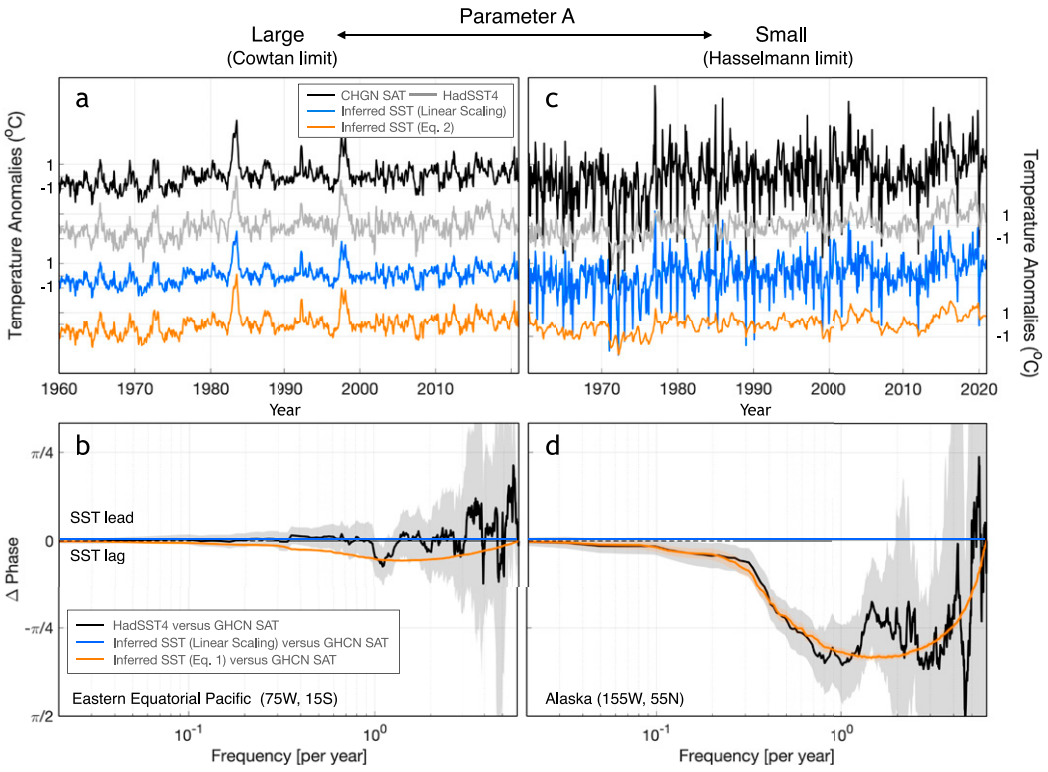

FIG. 2. Two limiting cases of SAT–SST coupling. (a) Observed coastal SAT (black, CHGNmV4) and SST (gray, HadSST4) anomalies in a 10° grid box centered on 15°S, 75°W (near the eastern equatorial Pacific). Also shown are SAT-inferred SSTs using linear scaling (blue) and our EBM [Eq. (2); orange]. Anomalies are relative to 1982–2014 climatology and offset vertically for visibility. (b) Phase lags between monthly SST and SAT anomalies. Individual curves show lags for observed SSTs (black), linearly scaled SATs (blue), and EBM-based SAT-inferred SSTs (orange). Shadings show corresponding 95% c.i. (c),(d) As in (a) and (b), respectively, but for a grid box near Alaska (centered on 55°N, 155°W).

correlate with observed SSTs (gray line) with a Pearson's r of 0.79, as compared to $r = 0.49$ for linearly scaled SATs (blue line). Fitted SSTs also capture the observed 2-month lag between SATs and SSTs ($\pi/3$ phase lag at the period of a year, Fig. 2d), in contrast to the near in-phase behavior found at the eastern equatorial Pacific site (Fig. 2b). Phase estimates are made using a multitaper coherence technique (Percival et al. 1993). This ability of our EBM to resolve phase differences is another advance over the linear scaling method. Note, however, that distinct from the original Hasselmann model where SST phase lags increase monotonically with frequency, the lag of SST behind SAT in our model peaks near a period of six months because A is not strictly zero and the component that SSTs are in phase with SATs, i.e., the small linear scaling solution, dominates the dampened Hasselmann-model solution at very high frequency.

We note a limitation of our model associated with the fact that we prescribe a single fixed ocean heat capacity, whereas ocean mixed layer depth varies, leading to differences in heat storage and air–sea exchanges (Deser et al. 2003). This limitation is somewhat mitigated, however, in that winter conditions at high latitudes show anomalies having larger variance and appear to be more important than summer conditions in terms of determining annual average anomalies. At the Alaska site, wintertime SAT variance is 10.4 times larger than the summertime counterpart (Fig. 2c). Moreover, the autocorrelation of observed SSTs (HadSST4) remains higher than 0.5 subsequent to the winter season and then decreases starting the next winter season (Fig. 3a), consistent with large wintertime SAT anomalies resetting the memory of the coupled air–sea temperatures. To further demonstrate the important role of wintertime SAT anomalies, we predict SSTs using SATs in only warm (15 March to 15 October) or cold (15 October to 15 March of the next year) months. When using cold-month SATs, we set SATs in warm months equal to their value on 14 March in the corresponding year, and vice versa. SSTs predicted using cold-month SATs (blue in Fig. 3b) have a Pearson's correlation (r) of 0.79 with the full model prediction (EBM-fitted SST; gray), a value that is significantly higher ($p < 0.05$) than the correlation of 0.51 for warm-month predictions (orange). That winter-based predictions are more skillful is also clear in individual calendar months (Figs. 3c,d).

b. *Testing model skill using CMIP6 simulations*

Before applying the EBM to infer SSTs from observed SATs, we examine its skill in the context of CMIP6

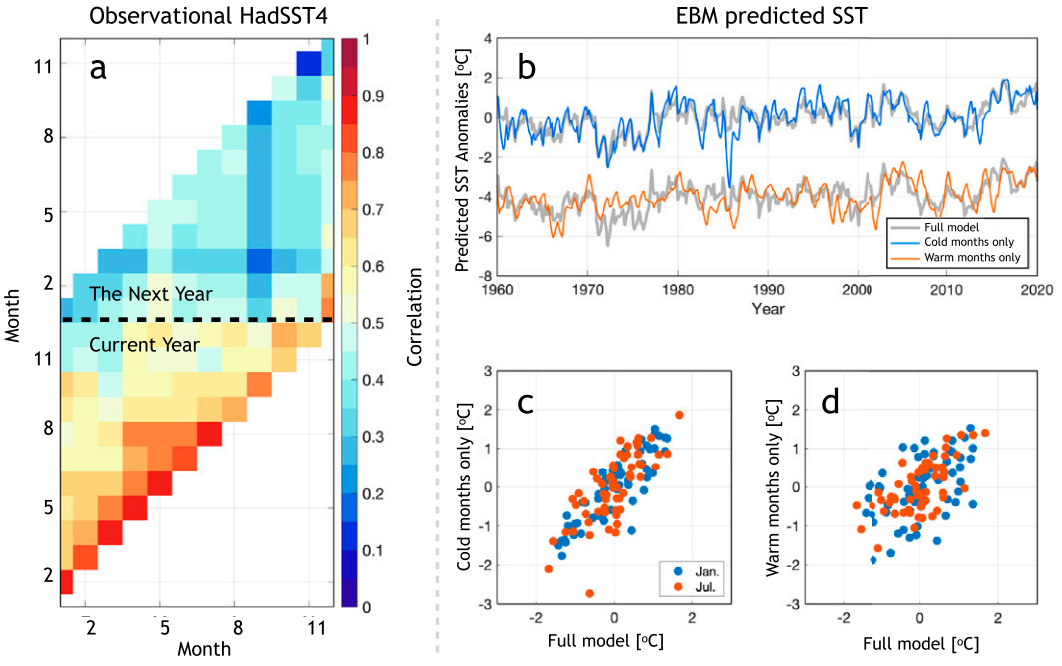

FIG. 3. Large wintertime SAT perturbations drive high-latitude SST variations. (a) The autocorrelation of SSTs (HadSST4) near the Alaska site in Fig. 2c as a function of calendar months (x axis) and lag in time (y axis). (b) SSTs predicted using cold-month SATs (blue), warm-month SATs (orange), and all-year SATs (full model, gray). (c) Scatter of predicted SSTs using cold-month SATs only (y axis) vs using all-year SATs (x axis) in January (blue) and July (red). (d) As in (c), but for using warm-month SATs only.

simulations. Fitting our model to CMIP6 simulations between 1960 and 2020, we first consider the spatial pattern of fitted parameters. Parameter A decreases from approximately 0.5 in the deep tropics to 0.1 in the extratropics (Figs. 4a,d), indicating a shift from the Cowtan et al. (2018) limit in lower latitudes to the Hasselmann limit in extratropics (Fig. 2). Parameters B and C have lower values in both the extratropics and the deep tropics (Figs. 4e,f). Such a latitudinal dependence of parameters B and C can be understood, at least partly, in the context of an air–sea bulk formula, where the product of wind speed and humidity deficit determines the efficacy of latent heat exchanges and has the highest values in subtropics.

Consider the parameter B, which equals $(\lambda_o + \kappa_o + k\kappa_a)/\gamma_o$. Air–sea interaction terms containing κ are generally an order of magnitude higher than direct radiative damping, λ_o (Barsugli and Battisti 1998), and air–sea exchange is generally dominated by latent, as opposed to longwave radiative or sensible heat fluxes. Using the bulk formula (Businger 1975), evaporation can be parameterized using a transfer coefficient (C), surface wind speed (U), and differences in the specific humidity at the air–sea interface ($q_a - q_o$),

$$E = CU(q_a - q_o). \tag{3}$$

Although the transfer coefficient C depends on background stability and surface roughness (Edson 2008), we make several simplifying assumptions. First, it is not unreasonable to assume a constant C when the monthly wind speed is below 20 m s^{-1}, as is generally the case. Second, assuming constant relative humidity allows for inferring specific humidity from temperature using the Clausius–Clapeyron equation (e.g., following the formula provided by Bolton 1980). Finally, averaging U allows representing anomalous latent heat flux E' as

$$E' \propto \overline{U}\left.\frac{\partial q^*}{\partial T}\right|_{\overline{T}}(T'_a - T'_o), \tag{4}$$

where $\overline{U}\partial q^*/\partial T|_{\overline{T}}$ is an approximation of κ, and $\partial q^*/\partial T|_{\overline{T}}$ denotes the sensitivity of saturation specific humidity to temperature evaluated at mean temperature $\overline{T}$. Whereas $\partial q^*/\partial T|_{\overline{T}}$ decreases with latitude as the mean temperature drops, $\overline{U}$ increases with latitude as a consequence of atmospheric storminess. A combination of relatively high temperature and surface wind speed maximizes E in the subtropics, whereas lower winds explain a dip in fitted B and C parameters in the deep tropics. Also note that in the Northern Hemisphere extratropics the estimated values of parameters B and C, 1–2 $\times$ 10^{-7} s^{-1} (Figs. 4b,c), are consistent with the midlatitude values diagnosed by Barsugli and Battisti (1998) from a two-layer GCM.

A cross-validation technique is used to evaluate the skill of Eq. (2) in predicting out-of-sample SSTs simulated by CMIP6 models. Specifically, for each 10° grid box, we leave out 21 years of data (1960–80 or 2000–20) and infer SSTs using parameters that are fitted on the remaining 40 years (1981–2020 or 1960–99). We quantify predictive skill using the mean squared error (MSE) and the squared Pearson's correlation (r^2)

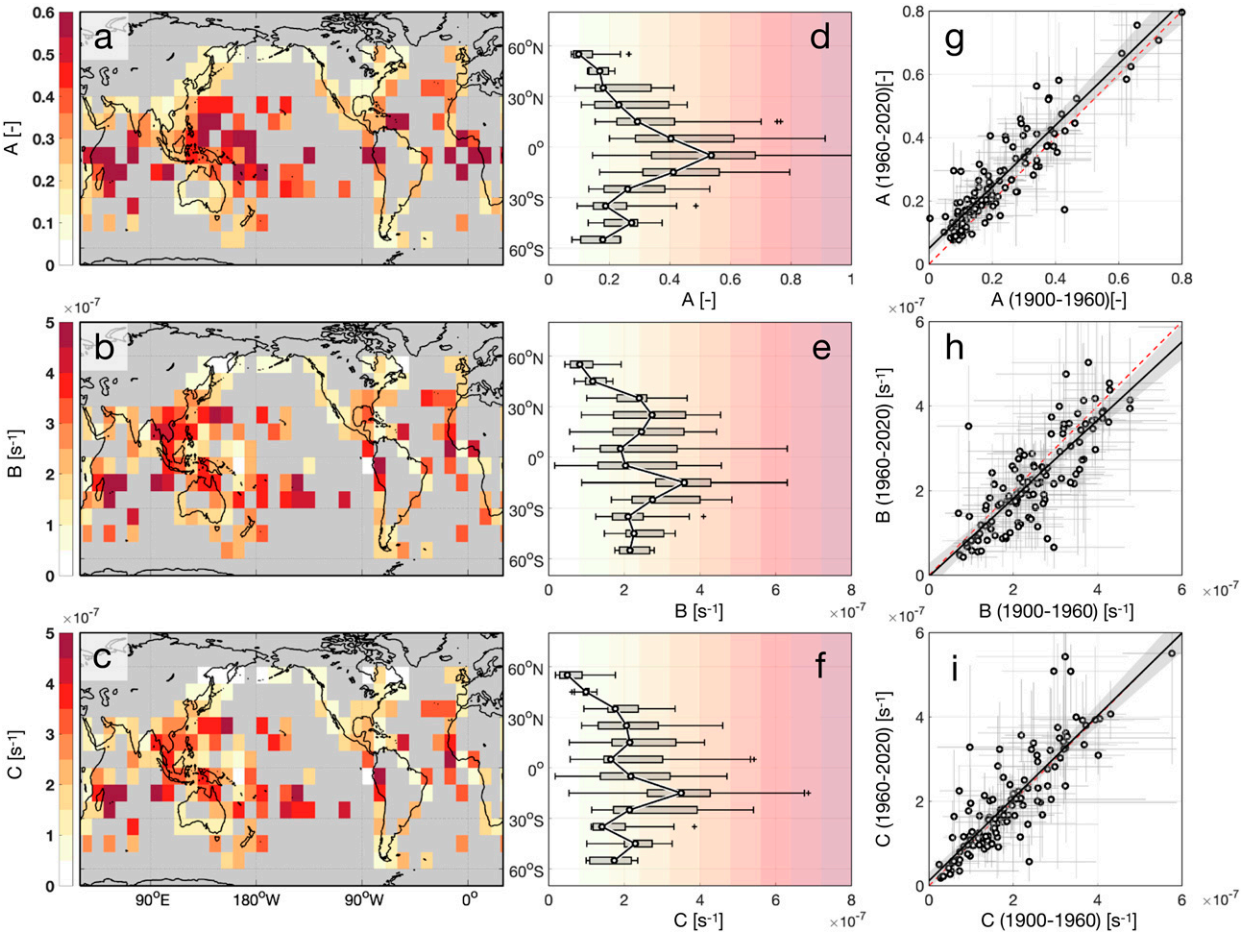

FIG. 4. CMIP6-based EBM parameters. (a)–(c) Multimodel medians of fitted parameters (a) A, (b) B, and (c) C in Eq. (2). Parameters are estimated using 1960–2020 simulations. (d)–(f) Latitudinal variations of corresponding EBM parameters in (a)–(c). Boxplots show distributions of multimodel medians across longitudes, where the box denotes interquartile range (IQR), whiskers go from the end of IQR to the furthest observation within the whisker length (1.5 × IQR), and markers denote outliers. The color shading indicates values of the x axis. (g)–(i) Parameters based on 1960–2020 simulations (y axes) vs those based on 1900–60 simulations (x axes). Markers are multimodel median values at individual 10° boxes, and error bars are interquartile ranges across CMIP6 GCMs. Solid lines are ordinary least squares fits, and shadings denote the 95% c.i. estimated by a bootstrapping technique that resamples 10° boxes 100 times with replacement. Red dashed lines denote the one-to-one relationship.

between SAT-inferred and simulated SSTs. Compared with using raw SATs, our EBM decreases MSE by an average of 49%, ranging from 18% in the deep tropics to 80% in the Northern Hemisphere extratropics (Figs. 5a,c). Increases in r^2 are largest in the Northern Hemisphere extratropics, where values increase by an average of 50% (Figs. 5b,d).

A concern is that the skill of Eq. (2) could degrade during earlier periods because parameters may change with climate. For example, processes involving ocean warming and wind stilling could affect ocean mixed layer depth and the sensitivity of SSTs to air–sea heat fluxes (Kjellsson 2015). Other changes in atmospheric circulation could also affect air–sea coupling (Thomas et al. 2008; Vautard et al. 2010). We compare parameters fitted to CMIP6 simulations between 1900–60 and 1960–2020 and find that they lie near a one-to-one relationship, with a spread that is consistent with the uncertainty estimated by bootstrapping grid boxes (Figs. 4g–i). We also use parameters fitted to 1960–2020 simulations to infer earlier SSTs (Fig. 6). The multimodel mean difference in the warming rate between coastal-mean SAT and SST is 0.07°C ([0.05°, 0.08°C]) century^{-1} from 1880 to 1960, whereas this difference is only 0.006°C ([−0.003°, 0.014°C]) century^{-1} between SAT-inferred and simulated SSTs (Figs. 6a–c). Numbers in brackets report 95% confidence intervals (c.i.) if not otherwise noted. The variance of detrended multimodel mean air–sea temperature differences is also reduced by 67% in the early twentieth century. These results support the application of parameters fitted from recent observations to infer historical SSTs.

Analyses of CMIP6 simulations also permit estimating the uncertainties in SAT-inferred SSTs relative to simulates SSTs. This evaluation allows for errors associated with the simplicity of our EBM, including components in SSTs that cannot be

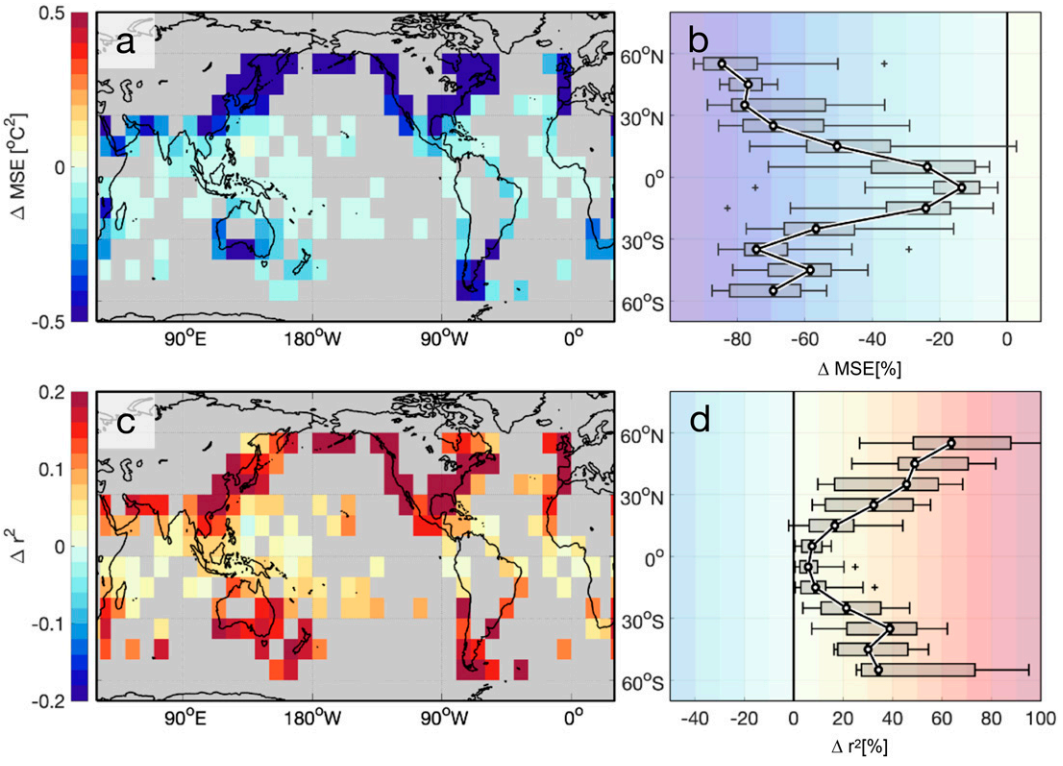

FIG. 5. Prediction skill of the EBM when applied to CMIP6. Out-of-sample predictions are performed using a cross-validation technique, and predictive skills are quantified using (a),(b) MSE and (c),(d) r^2 for monthly temperatures. Statistics are averages over the two cross-validation members and the 17 CMIP6 models. Panels show maps of (a),(c) absolute changes and (b),(d) fractional changes across latitudes. Boxplots in (b) and (d), as in Fig. 4d, show distributions across longitude bands.

fully explained by SATs such as ocean heat flux convergence. The average uncertainty of inferred SSTs, evaluated at 10° resolution, is 0.36°C (one standard deviation) at monthly resolution and 0.23°C at annual resolution. When averaged over the global coastline, the standard error at annual resolution shows clear spikes exceeding 0.1°C (one standard deviation) between 1850 and 1880 and then decreases to lower than 0.05°C after the 1880s (Fig. 6d). This error scales inversely with the square root of the number of grid boxes with observations, indicating that the interannual error structure is mainly a function of station coverage. Such an estimate indicates the intrinsic limit of using coastal and island station temperatures under historical station coverage, even given sufficiently accurate data. This result suggests that global SST estimates are potentially achievable at an accuracy of better than 0.05°C, a goal set by Kent and Berry (2008), if coastal inferences could be used to estimate SST biases over the open ocean. Before the 1880s, however, it appears that such a goal could not be met unless more historical station records are rescued. Note that such an error estimate is a lower bound because SAT records remain uncertain (to be discussed in section 4b). Moreover, propagating coastal estimates of SST biases into the ocean interior introduces additional uncertainty, which is relevant for quantifying the uncertainty of global mean temperatures, but is beyond the scope of this current paper.

c. *Observational inference of SSTs*

We next fit our model using homogenized SATs and HadSST4 observations after averaging coastal data in 10° × 10° boxes (Fig. 7). Observationally inferred parameters show consistent patterns with those obtained from CMIP6 simulations (cf. Fig. 7 to Fig. 4), but the magnitude of observational-inferred values of A and C are, on average, 26% smaller. Although this discrepancy in the magnitude of parameters could arise from observational noise and associated regression dilution effects (Fuller 2009), we favor an explanation involving the difference between local air temperature measured by stations and grid averages simulated by CMIP6 CGMs.

In Eq. (1), we assumed $T_a = \beta_{\mathrm{OBS}} T_s$, where the subscript "OBS" denotes quantities for observational data, and we expect β_{OBS} to be smaller than one on account of land–sea contrast (Byrne and O'Gorman 2018). In CMIP6 simulations, T_s is a grid average of SAT and marine air temperatures, such that the effective β is $\beta_{\mathrm{CMIP}} = \beta_{\mathrm{OBS}}/[f + (1 - f)\beta_{\mathrm{OBS}}]$, where f is the land fraction adjacent to a station. We thus expect β_{CMIP6} to be closer to one and hence fitted A and C parameters to be higher than observational estimates. The difference between β_{OBS} and β_{CMIP6} also explains the fact that the ratios of A_{OBS} to A_{CMIP6} and C_{OBS} to C_{CMIP6} increase with land fraction, f (solid lines in Fig. 8a). Using a β_{OBS} of

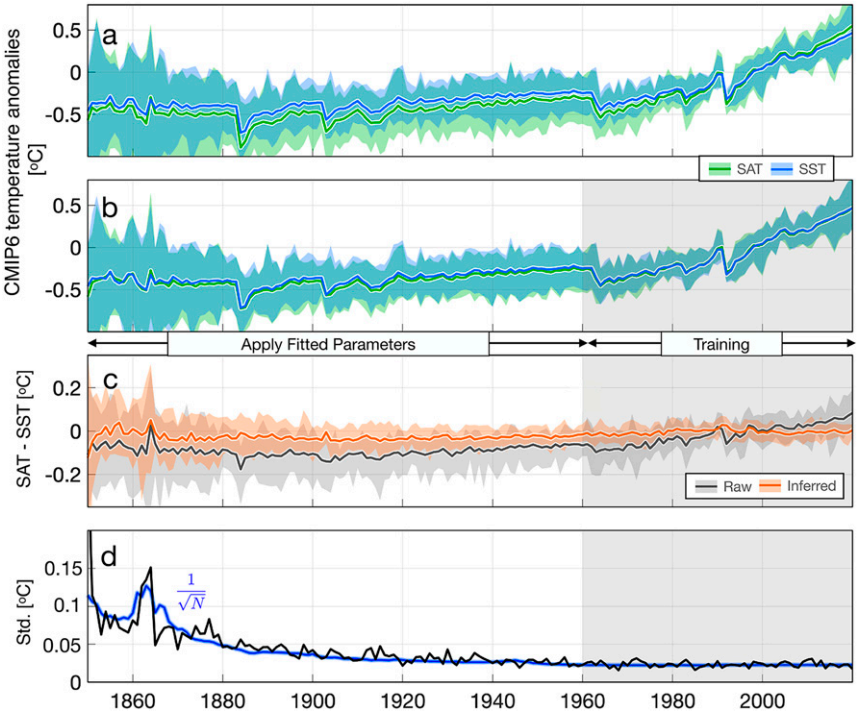

FIG. 6. Coastal-mean air and sea temperature anomalies in CMIP6 simulations. (a) Simulated SATs (green) and SSTs (blue) averaged over global coasts for multimodel means (curves) and the range across CMIP6 GCMs (shading). Anomalies are relative to 1982–2014 climatology. (b) As in (a), but for SAT-inferred SSTs (green) and simulated SSTs [blue, the same as (a)]. The inference of SSTs before the 1960s uses parameters estimated from 1960 to 2020 simulations. (c) As in (a), but for SAT minus SST differences. Our EBM method (orange) successfully removes differential air–sea warming and variability (gray). (d) The standard uncertainty (black) of temperature differences between SAT-inferred and simulated SSTs across CMIP6 GCMs. Also shown is the inverse of the square root of the number of monthly grid boxes having SAT observations in a year ($1/\sqrt{N}$; blue).

0.65 (dashed line in Fig. 8a) qualitatively reproduces the diagnosed relationship between parameter ratio and the land fraction. This explanation is also consistent with the fact that parameter B, whose definition does not involve β, is consistently inferred between observations and CMIP6 simulations.

The comparison between fitting our model to observed and simulated temperatures has implications for properly inferring SSTs from SATs. Cowtan et al. (2018) used f to calculate a linear scaling factor S for station temperatures: $S = 0.86 - 0.25f$. But, whereas the ratio of SAT and SST warming increases with f in GCM simulations (Cowtan et al. 2018), we do not expect for an f corresponding to a model-based gridding convention to apply to observations. Indeed, we find no relationship between the observed air–sea warming ratio and f in observations (Fig. 8d). That said, more detailed modeling accounting for proximity to the coast, topography, and other local meteorological factors would presumably further improve inferences of SSTs, but this level of detail is beyond the scope of the present analysis.

Applying the same cross-validation analysis as in section 3b to 1960–2020 observations gives improved MSE and r^2 relative to comparing SATs and SSTs directly (Fig. 9), and whose patterns and magnitudes are consistent with CMIP6 (Fig. 5). We also evaluate improvements relative to the linear scaling used by Cowtan et al. (2018). Linear scaling decreases MSE relative to raw SATs by an average of 24%, with a range of 11%–42% across latitude bands (Figs. 9a,d). Equation (2) further decreases MSE relative to linear scaling by an average of 24% (Figs. 9b,e). Improvements primarily occur in the Northern Hemisphere extratropics, where the Hasselmann limit of our model better applies, reducing MSE by more than 60% beyond that obtained by linear-scaled SATs. Smaller improvements in MSE of 6% are found in lower latitudes, where air–sea coupling approaches the Cowtan et al. (2018) limit (Fig. 2a).

4. Discussion

Inference of SSTs from SATs provides an opportunity to compare patterns of differences in near-coast SSTs between HadSST4 and those inferred from coastal SATs. These patterns will, however, be shown to be sensitive to the use of homogenized versus unhomogenized coastal SATs. Although

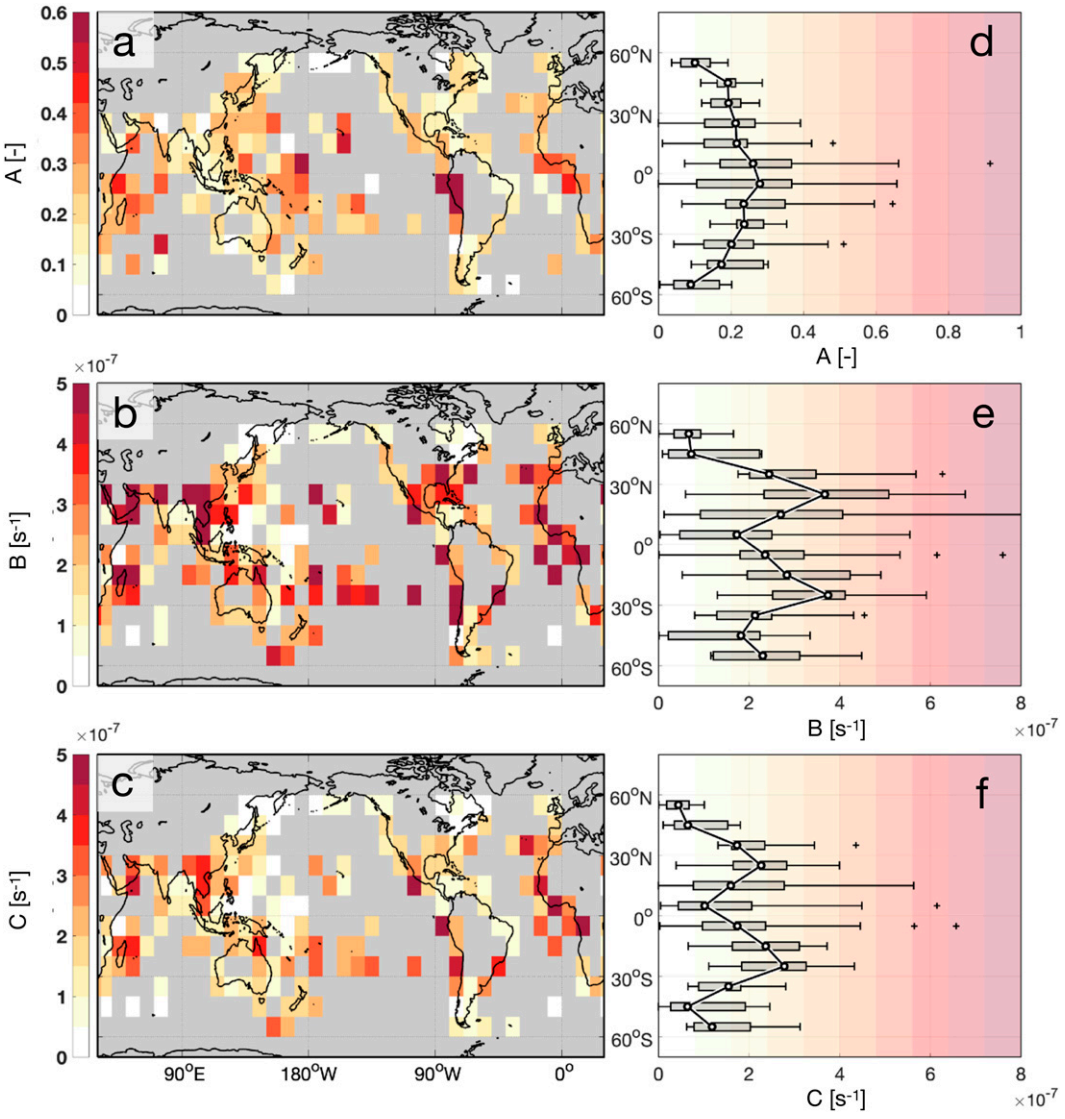

FIG. 7. Observation-based EBM parameters. (a)–(f) As in Fig. 4, but for parameters estimated from observational temperatures.

SATs and SSTs are both subject to systematic uncertainties, sources of these uncertainties are distinct, and it appears feasible to intercalibrate temperatures within and across SAT and SST archives simultaneously to provide more accurate estimates of global temperature variability.

a. Patterns of offsets between near-coast HadSST4 and SSTs inferred from homogenized SATs

Coastal average SSTs inferred from observational SATs are consistent with HadSST4 after the 1960s, but these estimates are inconsistent prior to 1960 (blue and black curves in Fig. 10). The RMSE between annual coastal-mean observational SSTs and inferred near-coast SSTs based on homogenized SAT records is 0.03°C after 1960, whereas between 1880 and 1960 the RMSE is 0.13°C (Fig. 10). During World War II (1941–45), coastal HadSST4 shows a warm anomaly of 0.21°C ([−0.04°, 0.49°C]) compared with the surrounding 10 years. Such a warm anomaly is not seen in SAT-inferred coastal SSTs and has been shown to arise from uncorrected biases associated with shifting instruments between buckets and engine-room intakes (Thompson et al. 2008; Chan and Huybers 2021). Nighttime bucket SSTs during World War II also appear overly warm, possibly on account of reading bucket water temperatures indoors to prevent detection by enemy vessels (Folland et al. 1984; Chan and Huybers 2021). Systematic offsets are also found in the early twentieth century, with SAT-inferred SSTs being 0.09°C ([0.01°, 0.17°C]) warmer than indicated by HadSST4 during 1900–40.

Coastal HadSST4 and SAT-inferred SSTs also show a variety of regional discrepancies, including the margins of the North Atlantic (NA) and equatorial Pacific (EP). Over 1900–40, coastal HadSST4 is, on average, 0.25°C ([0.14°, 0.36°C])

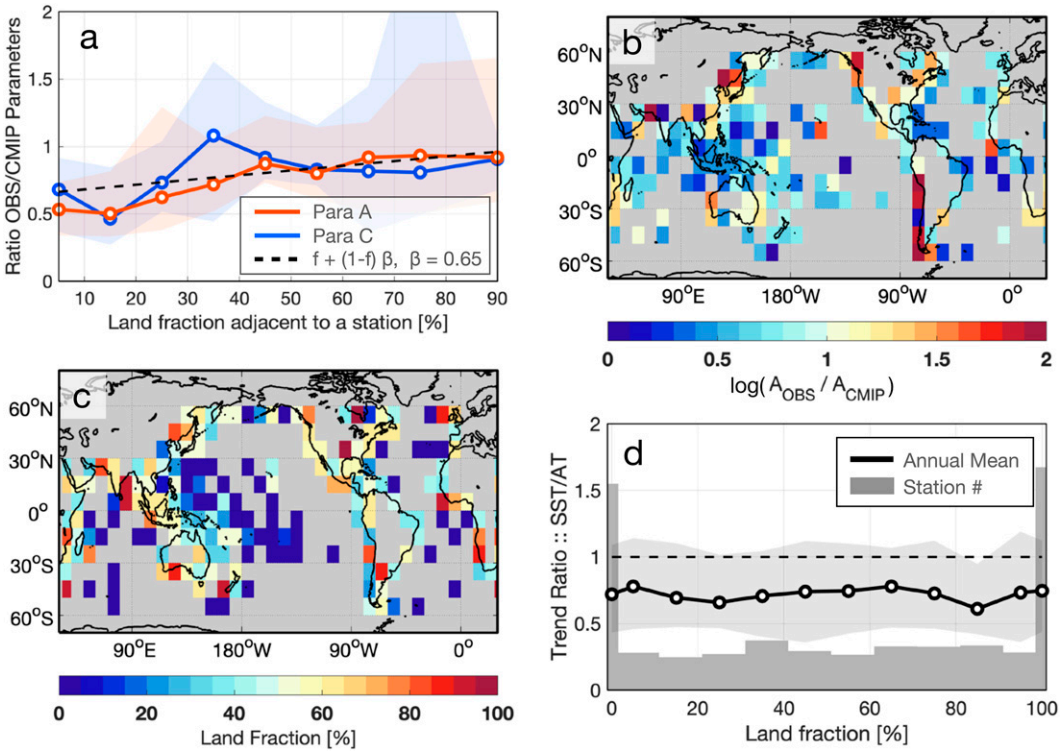

FIG. 8. Differences in fitted parameters between observations and CMIP6 simulations. (a) The ratio between observational and CMIP6-based parameter estimates (y axis) generally increases with land fraction adjacent to a station (f, x axis). Individual curves are for parameters A (orange) and C (blue), and shadings are the interquartile range across 10° grid boxes within 10% bins of f. (b) The ratio between observational and CMIP6-based estimates of parameter A (shown in log scale). (c) The mean f within each 10° grid box. (d) Warming ratio between annual SST and SAT as a function of f. Warming is the average of linear trends over 121 combinations of starting and ending years, where the starting year ranges between 1960 and 1970 and the ending year between 2010 and 2020. Shown are the median across all stations within a 10% bin of f (curve) and the associated interquartile range across stations (shading). Bars at the bottom indicate the number of stations within each bin.

warmer than inferred SSTs along NA coastlines (Fig. 11c). NA coastal SSTs in HadSST4 show consistent evolution with the so-called Atlantic multidecadal variability (AMV; Schlesinger and Ramankutty 1994). If we define a coastal AMV index as the detrended difference in the mean SST between NA coasts and global coasts, this coastal index in HadSST4 has a correlation of 0.81 with an AMV index typically defined using HadSST4 including open oceans (Trenberth and Shea 2006). For this calculation both the NA and global times series are smoothed using an 11-yr running average and the beginning and ending 5 years are excluded. Note that both the coastal and the full AMV index calculated from HadSST4 have standard deviations of 0.11°C over 1886–2015, again with 11-yr smoothing. Calculating the AMV index using SAT-inferred coastal SSTs, however, results in an index with a standard deviation of only 0.06°C, suggesting less pronounced Atlantic multidecadal SST variability. The present findings appears consistent with other SST corrections that lead to decreased warming over the North Atlantic using the early twentieth century (Chan et al. 2019). Results are also qualitatively consistent with decadal variations in North Atlantic SSTs during the past millennium being mainly forced by volcanic eruptions, suggesting a lower contribution from internal variability than previously estimated (Mann et al. 2021).

The detected differences in the EP may also have implications for trends in equatorial Pacific SST gradients (Fig. 11a). Whereas differences between inferred SSTs and HadSST4 along the western EP are indistinguishable from 0°C [−0.16°, 0.17°C] over 1900–40 (95% c.i.; Fig. 11d), inferred SSTs are 0.58°C ([0.31°, 0.87°C]) warmer along the eastern EP (Fig. 11e), suggesting a possible bias in HadSST4 toward a more La Niña–like state during the early twentieth century. Furthermore, our inferred near-coast SSTs based on homogenized SATs are better in line with CMIP6 simulations along the eastern EP (Fig. 11). Existing estimates of observed SSTs show a strengthened west-minus-east SST gradient across the equatorial Pacific throughout the twentieth century that is counter to trends in general circulation model simulations (Coats and Karnauskas 2017). Seager et al. (2019) suggested that a cold bias in the equatorial cold tongue in general circulation models could explain the discrepancy in trends. If the difference we find along the eastern EP is informative of SST biases in the Niño-3.4 region, our

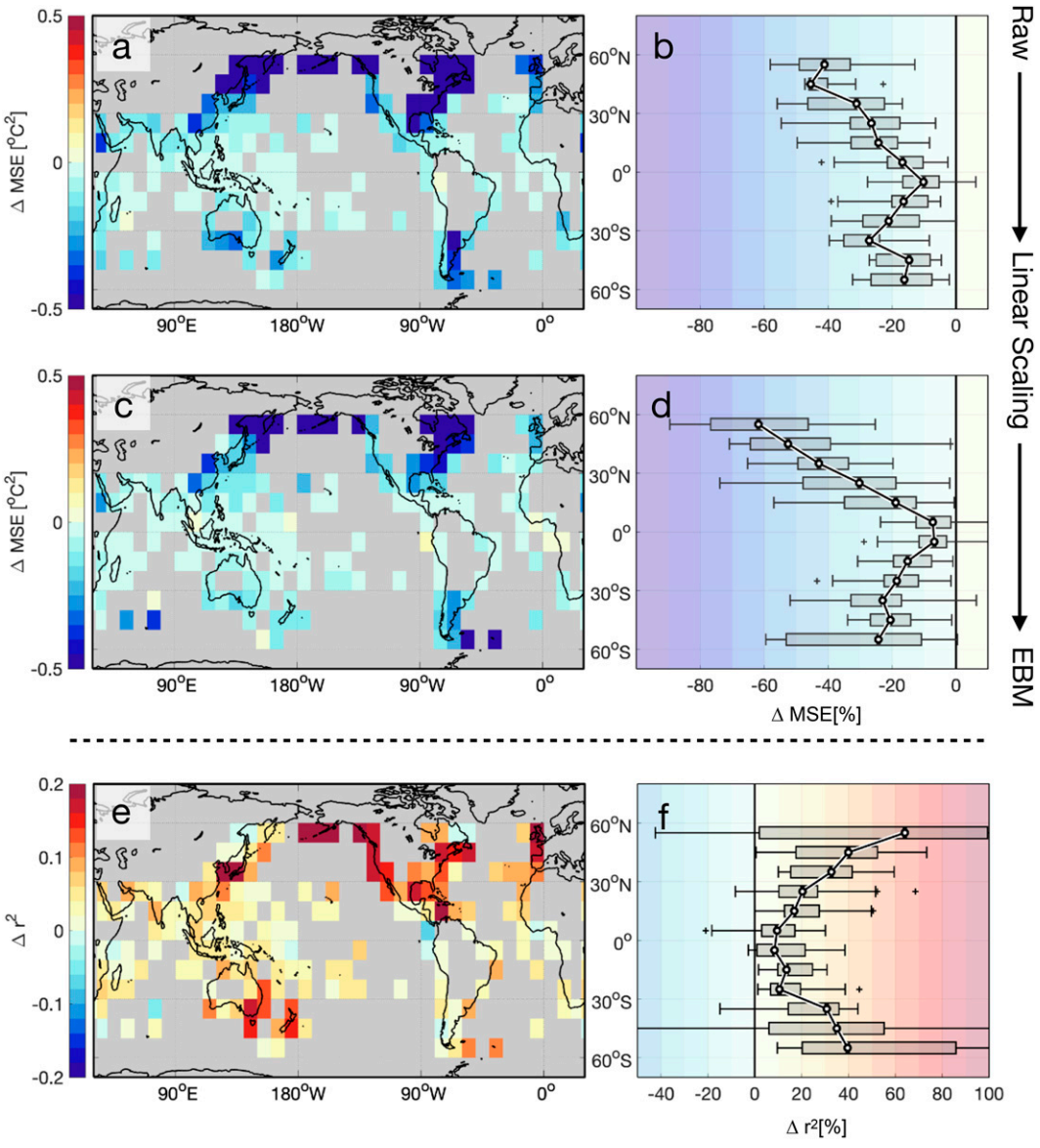

FIG. 9. Prediction skill of the EBM when applied to observational records. As in Fig. 5, except (a)–(d) MSE decreases are shown in two steps—(a),(b) changes from using raw SATs to linearly scaled SATs and (c),(d) further changes from using linearly scaled SATs to EBM-based SAT-inferred SSTs. (e),(f) Changes in r^2. Note that Δr^2 are only shown as differences between using raw SATs and EBM-based SAT-inferred SSTs because raw and linearly scaled SATs yield the same r^2.

results of a historically warmer eastern EP would suggest an even greater difference between models and observations.

The degree to which corrections to SSTs along the EP coastline have implications for SSTs in the interior EP is unclear. Even if the discrepancy between coastal temperatures reflects biases in HadSST4, boundary currents and nearshore upwelling dynamics may generate local environments that are distinct from interior ocean conditions. Moreover, SST biases also depend on bucket types and measurement protocols (Kent et al. 2010) and could be ship specific (Kennedy et al. 2012). The eastern EP is historically sparsely sampled, with the Niño 3.4 region primarily observed by ships traveling from coastal California around South America. SST observations along coastal South America are primarily from ships whose paths are more localized (Freeman et al. 2017). A detailed analysis of coastal biases and ship tracks is warranted for purposes of better understanding trends in the Niño-3.4 region.

b. *Homogenized versus unhomogenized SATs*

Our main line of analysis is sensitive to whether homogenized or unhomogenized SATs are used. Homogenized SATs, as used previously throughout this work, refer to station records that are adjusted against neighboring stations through an algorithm that

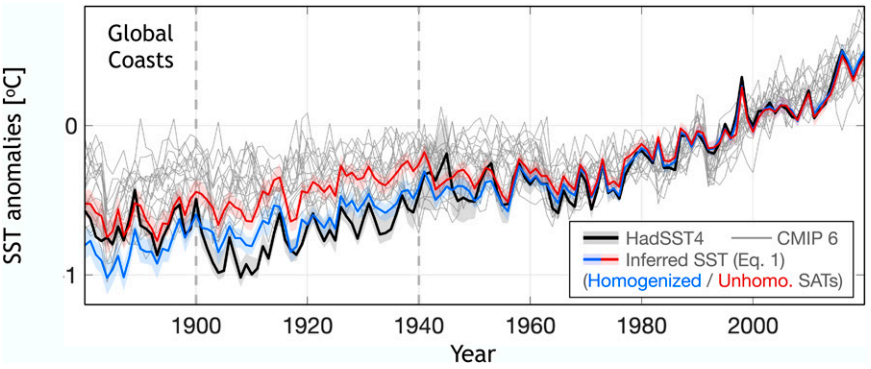

FIG. 10. Coastal mean HadSST4 and SAT-inferred SSTs. HadSST4 (gray) and EBM-based SAT-inferred SSTs (colors) generally agree after the 1960s but diverge before the 1950s. Inferred SSTs are based on homogenized (blue) or unhomogenized (red) GHCN station temperatures. Shading denotes one standard deviation errors. Also shown are 17 CMIP6 simulation (thin gray lines). Anomalies are relative to 1982–2014 climatology.

detects stepwise changes in time series (Menne and Williams 2009), whereas unhomogenized station temperature records involve error and outlier exclusion but no check for heterogeneity across stations. Adjustments are inferred to correct for the effects of urbanization, as well as changes in measurement protocols from reading temperatures at fixed hours in a day to once-per-day readings of maximum and minimum temperature during the past 24 h. Homogenized and unhomogenized records are similar since 1960, but continental mean temperature estimates from homogenized records are 0.2°–0.3°C colder during the early twentieth century than unhomogenized records (Menne et al. 2018b). Note that the larger discrepancies in the earlier interval may also related to the fact that records are normalized to zero mean between 1982 and 2014.

Cowtan et al. (2018) suggested that the SAT homogenization algorithm could artificially increase coastal trends on account of coastal regions warming less than inner continental regions. Our analysis of GHCNmV4 adjustments, however, shows no dependence between adjustment magnitudes and distance to coast (Fig. S1 in the online supplemental material). Rather, differences in temperature trends between 1900 and 2020 between homogenized and unhomogenized SATs are largest in the tropics and midlatitudes, especially in parts of the United States and China (Fig. S1). As further points of comparison, we note that the continental-mean SAT estimates from Berkeley Earth (Rohde et al. 2013) and CRUTEM5 (Osborn et al. 2021) give continental-mean SAT anomalies during 1900–40 that are, respectively, 0.05° and 0.15°C warmer than homogenized GHCNmV4 estimates and, thus, reside midway between the homogenized and unhomogenized GHCNmV4 versions. A more detailed comparison of land station temperatures, which involves reexamining characteristics of individual stations and networks, appears useful for further reducing systematic uncertainties in land temperature estimates at global and regional scales. Although such an analysis is beyond the scope of this study, systematic differences across individual continental temperature estimates of as much as 0.2°C shows that calibrating SSTs to current versions of SATs will not, of itself, substantially decrease systematic uncertainties.

It follows that SSTs inferred using homogenized or unhomogenized SATs are consistent after the 1960s, but SSTs based on the unhomogenized GHCN records are warmer by an average of 0.20°C from 1900 to 1940, implying even larger discrepancies with HadSST4 (red versus blue in Fig. 10). Some conclusions regarding SST trends are nevertheless possible regardless of calibration using homogenized or unhomogenized SATs. Both versions of inferred SSTs are warmer during the early twentieth century than HadSST4. Notable is that inferred SSTs show warming from 1880 to 1910 at a rate of 0.62°C ([0.41°, 0.89°C]) century^{-1} for homogenized and 0.23°C ([0.02°, 0.50°C]) century^{-1} for unhomogenized SATs, whereas HadSST4 shows a cooling trend of −1.03°C ([−1.18°, −0.77°C]) century^{-1}. If the SAT estimates are less biased than those directly from SSTs, they imply a smaller contribution from internal variability to global mean temperature, a conclusion in keeping with other recent assessments of historical temperature trends (Folland et al. 2018; Haustein et al. 2019).

It is also possible to make some inferences regarding patterns of SST trends. The variability in the coastal AMV index (discussed above) also decreases to 0.06°C when using unhomogenized SATs. EEP SSTs inferred from unhomogenized SATs indicate mean coastal EEP SST over 1900–30 is 0.33°C ([0.04°, 0.62°C]) warmer than 1990–2020 (Fig. 10f). On account of the overall forcing associated with increasing greenhouse gases, unhomogenized coastal EEP SATs during the early twentieth century may appear to be too warm. In CMIP6 models, coastal EEP SST anomalies over 1900–30 range from 0.30° to 0.77°C colder than 1990–2020 averages. More generally, differences relative to HadSST4 using homogenized or unhomogenized inferences of SSTs between 1900 and 1940 have a spatial correlation of 0.66 (Figs. 11a,b), indicative of the fact that using either version of SATs leads to consistent patterns of offsets.

5. Conclusions

We show that a coupled EBM framework allows for the inference of near-coast SSTs using air temperatures from

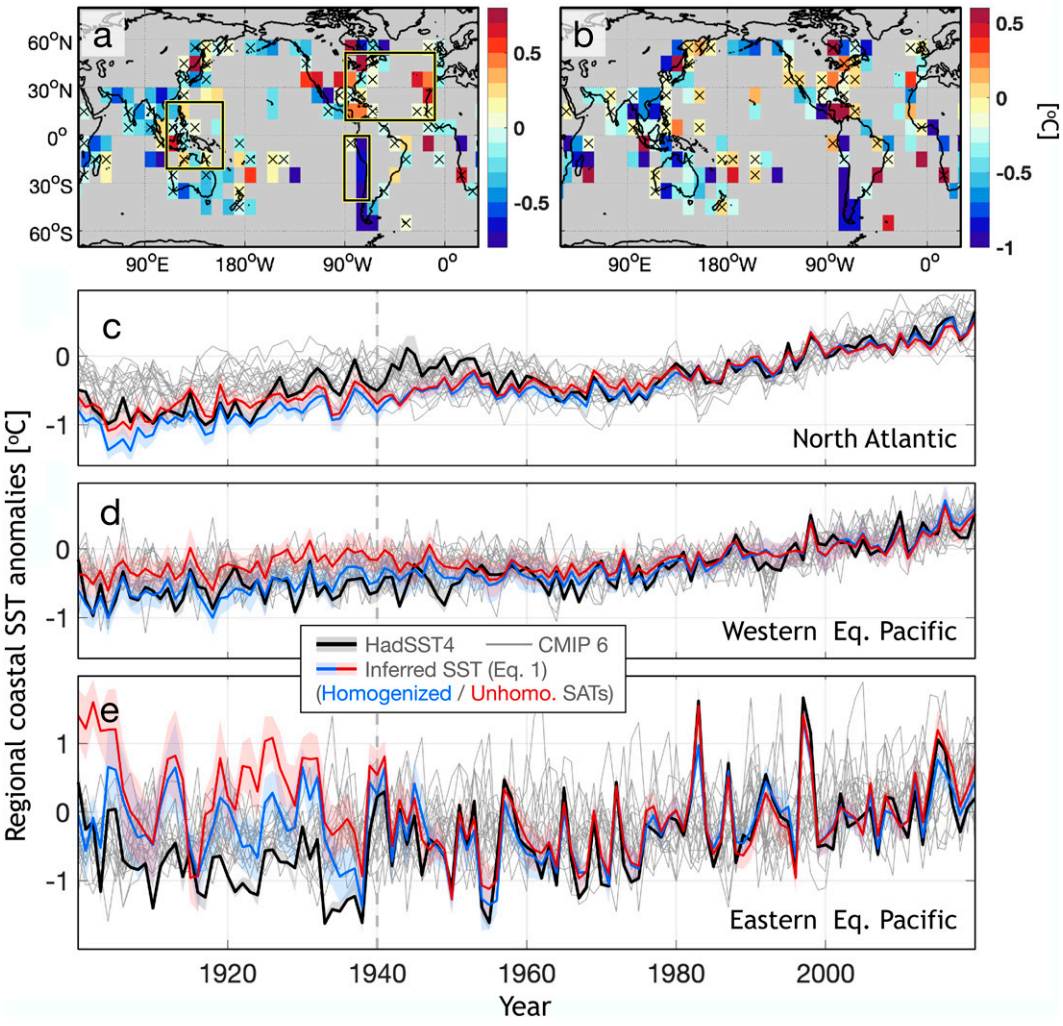

FIG. 11. Regional coastal HadSST4 and SAT-inferred SSTs. (a),(b) Maps of the differences between HadSST4 and SAT-inferred SSTs over 1900–40, based on, respectively, homogenized and unhomogenized GHCN temperatures. Differences that are insignificant from zero ($p > 0.05$) are crossed out. (c)–(e) As in Fig. 10, but for regional averages over (c) the North Atlantic, (d) the western equatorial Pacific, and (e) the eastern equatorial Pacific.

coastal weather stations. Our method captures the dynamics of regimes of air–sea temperature coupling, and is skillful when tested on CMIP6 simulations and recent observations. SAT-inferred SSTs are, however, inconsistent with near-coast HadSST4 estimates at both global and regional scales. These results provide a basis for better intercalibrating SATs and SSTs, although several caveats, discussed below, are worth highlighting.

Our EBM method has not yet accounted for seasonality. Seasonal variations in the mixed layer depth, atmospheric circulation, and local temperatures could vary parameters in Eq. (2) and hence the inferred SSTs. For example, summertime extratropical SSTs warm faster than their wintertime counterparts because of a shallower mixed layer. Although our current method brings the trend of annual-mean SATs into consistency with nearby SSTs, it does not capture the seasonality in air–sea temperature differences. In future work, it would be worthwhile to extend the EBM to account for seasonality in order to obtain accurate long-term trends in the seasonal cycle of SSTs.

The quality of the SST archive is heterogeneous because different agents measure different parts of the ocean at different times. That is, regionally varying SST biases could arise from changing nations or groups of ships that have distinct measurement characteristics (Chan and Huybers 2019). To better estimate global temperatures, it would be useful to diagnose systematic errors among individual ships and land-based weather stations. A further step in this direction would involve combining SAT-inferred SSTs with a groupwise intercomparison method (Chan and Huybers 2019, 2020). It may also be possible to use the amplitude of the diurnal cycle of SST measurements (Carella et al. 2018; Chan and Huybers 2021) and independent proxy records from corals (Pfeiffer et al. 2017) to further constrain early-twentieth-century temperature anomalies.

Finally, the systematic differences between SSTs inferred from SATs and HadSST4 along coasts, together with an apparent uncertainty associated with SAT homogenization of 0.2°C at the global scale, indicates the need for further analysis of systematic errors in SATs. Analysis and homogenization of SSTs and SATs have generally been treated as independent topics, but their proximity along the coast suggests that the homogenization of both data sources could be performed jointly in order to better estimate long-term warming rates and constrain uncertainty in these estimates.

Acknowledgments. We thank Dr. Young-Oh Kwon for discussion and comments on the project. We acknowledge three anonymous reviewers for comments that improved the readability of the paper. D. Chan is supported by the Woods Hole Oceanographic Institute Weston Howland Jr. Postdoctoral Fellowship. G. Gebbie is supported by NSF OCE-82280500. P. Huybers is supported by NSF Grant 2123295. The authors have no conflict of interests to declare.

Data availability statement. All datasets used in this study are available as follows: GHCNmV4 (https://www.ncei.noaa.gov/pub/data/ghcn/v4/; last accessed 19 May 2022); HadSST4.0.1.0, uncertainty estimates, and a 200-member ensemble (https://www.metoffice.gov.uk/hadobs/hadsst4/data/download.html; last accessed 19 May 2022). HadSST.4.0.1.0 data are © British Crown Copyright, Met Office 2022, provided under an Open Government License (http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/). Monthly CMIP6 outputs are from the ESGF portal (https://esgf-node.llnl.gov/search/cmip6/; last accessed 16 August 2021). Intermediate and final outputs of SAT-inferred near-coast SSTs, together with codes used to generate them, are archived in a Harvard Dataverse repository (https://doi.org/10.7910/DVN/Y4D8PA).

APPENDIX A

Calculating Land Station Temperature Anomalies

SAT anomalies are computed relative to the climatological period of 1982–2014. In addition to the availability of high-resolution SST climatology, as stated in the main text, this choice of climatological period also maximizes the number of tier-1 stations, for which anomalies are calculated relative to the climatological mean of the target station without referring to neighboring stations. The criteria of tier-1 stations include having at least 16 years of data that each contains at least 6 months of data during the climatological period. The numbers of identified tier-1 stations are 1659, 1648, 1700, and 1651 for the period of 1960–90, 1970–2000, 1982–2014, and 1990–2020, respectively.

Climatological anomalies for other land-temperature stations are estimated using a pairing-and-adjusting method and are referred to as tier-2 stations. Specifically, we first identify neighboring tier-1 stations within 300 km of a target tier-2 station. Monthly anomalies of the target tier-2 station are adjusted to have the same average as the mean of neighboring stations during overlapping years. Adjusted tier-2 stations can serve as neighbors for unadjusted tier-2 stations, and our algorithm iterates between pairing and adjusting, thus allowing for estimating the mean of 1340 tier-2 stations. Applying the algorithm to CMIP6 simulations, we estimate adjustment uncertainties for tier-2 stations to be 0.13°C (one standard deviation), a value similar to that for tier-1 stations (0.13°C) because of noise suppression associated with averaging across multiple neighbors for most tier-2 stations. Our results are qualitatively insensitive to the choice of threshold for identifying neighbors. Using 200 or 400 km, respectively, results in using 72 fewer or 31 more stations. Differences in estimates of monthly anomalies for tier-2 stations that are common among all analyses have a standard error of 0.11°C.

APPENDIX B

Integrating SSTs from Monthly SATs

Fitting the model requires predicting SSTs given monthly SATs. To better capture potential phase differences between SATs and SSTs, we integrate the model using a small time step. Specifically, we first linearly interpolate monthly SATs to infill missing values. The model is then integrated at a nominal 12-h time step, which requires 12-hourly SAT anomalies that are continuous in time and have the same monthly values as the original monthly SAT anomalies. Such high-resolution time series are obtained using a so-called diddling procedure, whereby a series of linear equations in the form of $T_i^* = (T_{i-1} + 6T_i + T_{i+1})/8$ are solved, where T_i^* denotes monthly average SAT in month i, and T_i denotes SATs that are linearly interpolated to 12-h resolution at the center of month i. The diddling procedure is commonly used to interpolate monthly SSTs and sea ice when specifying surface boundary conditions for atmospheric general circulation model simulations (Taylor et al. 2000). B-spline interpolation would also be possible and could have additional advantages such as guaranteeing continuous first- and second-order derivatives, although these are not required in our present context. The integration results in 12-hourly SSTs, which are then averaged back to have a monthly resolution. A final step discards months without observed SATs, yielding the final estimate of SAT-based SSTs.

REFERENCES

Barsugli, J. J., and D. S. Battisti, 1998: The basic effects of atmosphere–ocean thermal coupling on midlatitude variability. *J. Atmos. Sci.*, **55**, 477–493, https://doi.org/10.1175/1520-0469(1998)055<0477:TBEOAO>2.0.CO;2.

Bolton, D., 1980: The computation of equivalent potential temperature. *Mon. Wea. Rev.*, **108**, 1046–1053, https://doi.org/10.1175/1520-0493(1980)108<1046:TCOEPT>2.0.CO;2.

Businger, J. A., 1975: Interactions of sea and atmosphere. *Rev. Geophys.*, **13**, 720–726, https://doi.org/10.1029/RG013i003p00720.

Byrne, M. P., and P. A. O'Gorman, 2018: Trends in continental temperature and humidity directly linked to ocean warming. *Proc. Natl. Acad. Sci. USA*, **115**, 4863–4868, https://doi.org/10.1073/pnas.1722312115.

Carella, G., J. Kennedy, D. I. Berry, S. Hirahara, C. J. Merchant, S. Morak-Bozzo, and E. C. Kent, 2018: Estimating sea surface temperature measurement methods using characteristic differences in the diurnal cycle. *Geophys. Res. Lett.*, **45**, 363–371, https://doi.org/10.1002/2017GL076475.

Chan, D., 2021: Combining statistical, physical, and historical evidence to improve historical sea-surface temperature records. *Harv. Data Sci. Rev.*, **3** (1), https://doi.org/10.1162/99608f92.edcee38f.

——, and P. Huybers, 2019: Systematic differences in bucket sea surface temperature measurements among nations identified using a linear-mixed-effect method. *J. Climate*, **32**, 2569–2589, https://doi.org/10.1175/JCLI-D-18-0562.1.

——, and ——, 2020: Systematic differences in bucket sea surface temperatures caused by misclassification of engine room intake measurements. *J. Climate*, **33**, 7735–7753, https://doi.org/10.1175/JCLI-D-19-0972.1.

——, and ——, 2021: Correcting observational biases in sea surface temperature observations removes anomalous warmth during World War II. *J. Climate*, **34**, 4585–4602, https://doi.org/10.1175/JCLI-D-20-0907.1.

——, E. C. Kent, D. I. Berry, and P. Huybers, 2019: Correcting datasets leads to more homogeneous early-twentieth-century sea surface warming. *Nature*, **571**, 393–397, https://doi.org/10.1038/s41586-019-1349-2.

Coats, S., and K. B. Karnauskas, 2017: Are simulated and observed twentieth century tropical Pacific sea surface temperature trends significant relative to internal variability? *Geophys. Res. Lett.*, **44**, 9928–9937, https://doi.org/10.1002/2017GL074622.

Cowtan, K., R. Rohde, and Z. Hausfather, 2018: Evaluating biases in sea surface temperature records using coastal weather stations. *Quart. J. Roy. Meteor. Soc.*, **144**, 670–681, https://doi.org/10.1002/qj.3235.

Deser, C., M. A. Alexander, and M. S. Timlin, 2003: Understanding the persistence of sea surface temperature anomalies in midlatitudes. *J. Climate*, **16**, 57–72, https://doi.org/10.1175/1520-0442(2003)016<0057:UTPOSS>2.0.CO;2.

Edson, J. B., 2008: Review of air-sea transfer processes. *ECMWF Workshop on Ocean-Atmosphere Interactions*, Reading, United Kingdom, ECMWF, 7–24.

Eyring, V., S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor, 2016: Overview of the Coupled Model Intercomparison Project phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.*, **9**, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016.

Folland, C. K., and D. E. Parker, 1995: Correction of instrumental biases in historical sea surface temperature data. *Quart. J. Roy. Meteor. Soc.*, **121**, 319–367, https://doi.org/10.1002/qj.49712152206.

——, ——, and F. E. Kates, 1984: Worldwide marine temperature fluctuations 1856–1981. *Nature*, **310**, 670–673, https://doi.org/10.1038/310670a0.

——, O. Boucher, A. Colman, and D. E. Parker, 2018: Causes of irregularities in trends of global mean surface temperature since the late 19th century. *Sci. Adv.*, **4**, eaao5297, https://doi.org/10.1126/sciadv.aao5297.

Freeman, E., and Coauthors, 2017: ICOADS release 3.0: A major update to the historical marine climate record. *Int. J. Climatol.*, **37**, 2211–2232, https://doi.org/10.1002/joc.4775.

Fuller, W. A., 2009: *Measurement Error Models*. John Wiley and Sons, 480 pp.

Hasselmann, K., 1976: Stochastic climate models: Part I. Theory. *Tellus*, **28**, 473–485, https://doi.org/10.3402/tellusa.v28i6.11316.

Hausfather, Z., K. Cowtan, D. C. Clarke, P. Jacobs, M. Richardson, and R. Rohde, 2017: Assessing recent warming using instrumentally homogeneous sea surface temperature records. *Sci. Adv.*, **3**, e1601207, https://doi.org/10.1126/sciadv.1601207.

Haustein, K., and Coauthors, 2019: A limited role for unforced internal variability in twentieth-century warming. *J. Climate*, **32**, 4893–4917, https://doi.org/10.1175/JCLI-D-18-0555.1.

Huang, B., and Coauthors, 2017: Extended Reconstructed Sea Surface Temperature, version 5 (ERSSTv5): Upgrades, validations, and intercomparisons. *J. Climate*, **30**, 8179–8205, https://doi.org/10.1175/JCLI-D-16-0836.1.

Jones, P., 2016: The reliability of global and hemispheric surface temperature records. *Adv. Atmos. Sci.*, **33**, 269–282, https://doi.org/10.1007/s00376-015-5194-4.

Karl, T. R., H. F. Diaz, and G. Kukla, 1988: Urbanization: Its detection and effect in the United States climate record. *J. Climate*, **1**, 1099–1123, https://doi.org/10.1175/1520-0442(1988)001<1099:UIDAEI>2.0.CO;2.

Kennedy, J. J., N. A. Rayner, R. O. Smith, D. E. Parker, and M. Saunby, 2011: Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 2. Biases and homogenization. *J. Geophys. Res.*, **116**, D14104, https://doi.org/10.1029/2010JD015220.

——, R. O. Smith, and N. A. Rayner, 2012: Using AATSR data to assess the quality of in situ sea-surface temperature observations for climate studies. *Remote Sens. Environ.*, **116**, 79–92, https://doi.org/10.1016/j.rse.2010.11.021.

——, N. A. Rayner, C. P. Atkinson, and R. E. Killick, 2019a: 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, https://doi.org/10.1029/2018JD029867.

——, ——, ——, and ——, 2019b: The Met Office Hadley Centre HadSST. 4.0.1.0 data set. Met Office Hadley Centre, accessed 19 May 2022, https://www.metoffice.gov.uk/hadobs/hadsst4/data/download.html.

Kent, E. C., and P. K. Taylor, 2006: Toward estimating climatic trends in SST. Part I: Methods of measurement. *J. Atmos. Oceanic Technol.*, **23**, 464–475, https://doi.org/10.1175/JTECH1843.1.

——, and D. I. Berry, 2008: Assessment of the marine observing system (ASMOS): Final report. National Oceanography Centre Southampton Research and Consultancy Rep. 32, 55 pp.

——, and J. J. Kennedy, 2021: Historical estimates of surface marine temperatures. *Annu. Rev. Mar. Sci.*, **13**, 283–311, https://doi.org/10.1146/annurev-marine-042120-111807.

——, ——, D. I. Berry, and R. O. Smith, 2010: Effects of instrumentation changes on sea surface temperature measured in situ. *Wiley Interdiscip. Rev.: Climate Change*, **1**, 718–728, https://doi.org/10.1002/wcc.55.

——, N. A. Rayner, D. I. Berry, M. Saunby, B. I. Moat, J. J. Kennedy, and D. E. Parker, 2013: Global analysis of night marine air temperature and its uncertainty since 1880: The HadNMAT2 data set. *J. Geophys. Res. Atmos.*, **118**, 1281–1298, https://doi.org/10.1002/jgrd.50152.

——, and Coauthors, 2017: A call for new approaches to quantifying biases in observations of sea surface temperature. *Bull. Amer. Meteor. Soc.*, **98**, 1601–1616, https://doi.org/10.1175/BAMS-D-15-00251.1.

Kjellsson, J., 2015: Weakening of the global atmospheric circulation with global warming. *Climate Dyn.*, **45**, 975–988, https://doi.org/10.1007/s00382-014-2337-8.

Lee, D. E., Z. Liu, and Y. Liu, 2008: Beyond thermal interaction between ocean and atmosphere: On the extratropical climate variability due to the wind-induced SST. *J. Climate*, **21**, 2001–2018, https://doi.org/10.1175/2007JCLI1532.1.

Mann, M. E., B. A. Steinman, D. J. Brouillette, and S. K. Miller, 2021: Multidecadal climate oscillations during the past millennium driven by volcanic forcing. *Science*, **371**, 1014–1019, https://doi.org/10.1126/science.abc5810.

Menne, M. J., and C. N. Williams Jr., 2009: Homogenization of temperature series via pairwise comparisons. *J. Climate*, **22**, 1700–1717, https://doi.org/10.1175/2008JCLI2263.1.

——, ——, B. E. Gleason, J. J. Rennie, and J. H. Lawrimore, 2018a: Global historical climatology network—Monthly (GHCN-M), version 4. NOAA/National Climatic Data Center, accessed 19 May 2022, https://www.ncei.noaa.gov/products/land-based-station/global-historical-climatology-network-monthly.

——, ——, ——, ——, and ——, 2018b: The Global Historical Climatology Network monthly temperature dataset, version 4. *J. Climate*, **31**, 9835–9854, https://doi.org/10.1175/JCLI-D-18-0094.1.

Meyssignac, B., and Coauthors, 2019: Measuring global ocean heat content to estimate the Earth energy imbalance. *Front. Mar. Sci.*, **6**, 432, https://doi.org/10.3389/fmars.2019.00432.

Osborn, T. J., P. D. Jones, D. H. Lister, C. P. Morice, I. R. Simpson, J. P. Winn, E. Hogan, and I. C. Harris, 2021: Land surface air temperature variations across the globe updated to 2019: The CRUTEM5 data set. *J. Geophys. Res. Atmos.*, **126**, e2019JD032352, https://doi.org/10.1029/2019JD032352.

Percival, D. B., and A. T. Walden, 1993: *Spectral Analysis for Physical Applications*. Cambridge University Press, 583 pp.

Pfeiffer, M., J. Zinke, W.-C. Dullo, D. Garbe-Schönberg, M. Latif, and M. E. Weber, 2017: Indian Ocean corals reveal crucial role of World War II bias for twentieth century warming estimates. *Sci. Rep.*, **7**, 14434, https://doi.org/10.1038/s41598-017-14352-6.

Reynolds, R. W., T. M. Smith, C. Liu, D. B. Chelton, K. S. Casey, and M. G. Schlax, 2007: Daily high-resolution-blended analyses for sea surface temperature. *J. Climate*, **20**, 5473–5496, https://doi.org/10.1175/2007JCLI1824.1.

Rohde, R., and Coauthors, 2013: Berkeley Earth temperature averaging process. *Geoinf. Geostat.*, **1**, 20–100, https://doi.org/10.4172/2327-4581.1000103.

Saravanan, R., and J. C. McWilliams, 1998: Advective ocean–atmosphere interaction: An analytical stochastic model with implications for decadal variability. *J. Climate*, **11**, 165–188, https://doi.org/10.1175/1520-0442(1998)011<0165:AOAIAA>2.0.CO;2.

Schlesinger, M. E., and N. Ramankutty, 1994: An oscillation in the global climate system of period 65–70 years. *Nature*, **367**, 723–726, https://doi.org/10.1038/367723a0.

Seager, R., M. Cane, N. Henderson, D.-E. Lee, R. Abernathey, and H. Zhang, 2019: Strengthening tropical Pacific zonal sea surface temperature gradient consistent with rising greenhouse gases. *Nat. Climate Change*, **9**, 517–522, https://doi.org/10.1038/s41558-019-0505-x.

Smith, T. M., and R. W. Reynolds, 2002: Bias corrections for historical sea surface temperatures based on marine air temperatures. *J. Climate*, **15**, 73–87, https://doi.org/10.1175/1520-0442(2002)015<0073:BCFHSS>2.0.CO;2.

Taylor, K. E., D. Williamson, and F. Zwiers, 2000: The sea surface temperature and sea-ice concentration boundary conditions for AMIP II simulations. PCMDI Rep. 60, 28 pp.

Thomas, B. R., E. C. Kent, V. R. Swail, and D. I. Berry, 2008: Trends in ship wind speeds adjusted for observation method and height. *Int. J. Climatol.*, **28**, 747–763, https://doi.org/10.1002/joc.1570.

Thompson, D. W. J., J. J. Kennedy, J. M. Wallace, and P. D. Jones, 2008: A large discontinuity in the mid-twentieth century in observed global-mean surface temperature. *Nature*, **453**, 646–649, https://doi.org/10.1038/nature06982.

Trenberth, K. E., and D. J. Shea, 2006: Atlantic hurricanes and natural variability in 2005. *Geophys. Res. Lett.*, **33**, L12704, https://doi.org/10.1029/2006GL026894.

Trewin, B., 2010: Exposure, instrumentation, and observing practice effects on land temperature measurements. *Wiley Interdiscip. Rev.: Climate Change*, **1**, 490–506, https://doi.org/10.1002/wcc.46.

Vautard, R., J. Cattiaux, P. Yiou, J.-N. Thépaut, and P. Ciais, 2010: Northern Hemisphere atmospheric stilling partly attributed to an increase in surface roughness. *Nat. Geosci.*, **3**, 756–761, https://doi.org/10.1038/ngeo979.