Improvements to the Land Surface Air Temperature Reconstruction in NOAAGlobalTemp: An Artificial Neural Network Approach

BOYIN HUANG,[a] XUNGANG YIN,[a] MATTHEW J. MENNE,[a] RUSSELL VOSE,[a] AND HUAI-MIN ZHANG[a]

[a] *NOAA/National Centers for Environmental Information, Asheville, North Carolina*

(Manuscript received 13 April 2022, in final form 5 August 2022)

ABSTRACT: NOAA global surface temperature (NOAAGlobalTemp) is NOAA's operational global surface temperature product, which has been widely used in Earth's climate assessment and monitoring. To improve the spatial interpolation of monthly land surface air temperatures (LSATs) in NOAAGlobalTemp from 1850 to 2020, a three-layer artificial neural network (ANN) system was designed. The ANN system was trained by repeatedly randomly selecting 90% of the LSATs from ERA5 (1950–2019) and validating with the remaining 10%. Validations show clear improvements of ANN over the original empirical orthogonal teleconnection (EOT) method: the global spatial correlation coefficient (SCC) increases from 65% to 80%, and the global root-mean-square difference (RMSD) decreases from 0.99° to 0.57°C during 1850–2020. The improvements of SCCs and RMSDs are larger in the Southern Hemisphere than in the Northern Hemisphere and are larger before the 1950s and where observations are sparse. The ANN system was finally fed in observed LSATs, and its output over the global land surface was compared with those from the EOT method. Comparisons demonstrate similar improvements by ANN over the EOT method: The global SCC increased from 78% to 89%, the global RMSD decreased from 0.93° to 0.68°C, and the LSAT variability quantified by the monthly standard deviation (STD) increases from 1.16° to 1.41°C during 1850–2020. While the SCC, RMSD, and STD at the monthly time scale have been improved, long-term trends remain largely unchanged because the low-frequency component of LSAT in ANN is identical to that in the EOT approach.

SIGNIFICANCE STATEMENT: The spatial interpolation method of an artificial neural network has greatly improved the accuracy of land surface air temperature reconstruction, which reduces root-mean-square error and increases spatial coherence and variabilities over the global land surface from 1850 to 2020.

KEYWORDS: Land surface; Sea surface temperature; Surface temperature; Surface observations

1. Introduction

Earth's climate change is commonly quantified by variations of global surface temperature (GST). Major GST products include NOAA global surface temperature (NOAAGlobalTemp; Vose et al. 2021; Zhang et al. 2019; Smith et al. 2008), Met Office HadCRUT (Morice et al. 2021, 2012; Brohan et al. 2006), NASA GISTEMP (Lenssen et al. 2019; Hansen et al. 2010, 1999; Hansen and Lebedeff 1987), Berkeley Earth Surface temperature (BEST; Rohde and Hausfather 2020; Rohde et al. 2013), Cowtan and Way (2014) GST, and China Merged Surface Temperature (CMST; Yun et al. 2019). These GST products employ the original land surface air temperature (LSAT) record (Rohde et al. 2013) or LSAT product such as the Global Historical Climatology Network-Monthly (GHCNm; Menne et al. 2018), and sea surface temperatures (SST) from the Extended Reconstructed SST (ERSST; Huang et al. 2017) or the Hadley Center SST (HadSST; Kennedy et al. 2011a,b, 2019).

Studies indicate that GSTs exhibit uncertainties due to limitations of observations and data processing methods, and the uncertainties are largely related to limitations in observational coverage (Huang et al. 2020a; Brohan et al. 2006), such as in the high latitudes and early in the record. Several methods have been used to best reconstruct the GST from the piecemeal historical record. Cowtan and Way (2014) used a kriging algorithm (Cressie 1990) to determine the temperature in unsampled regions according to a known covariance structure. Hansen et al. (2010) and Hansen and Lebedeff (1987) applied an interpolation algorithm that used kernel smoothing within a conical kernel of radius 1200 km. Rohde et al. (2013) adopted the kriging method to the reconstruction of LSAT and merged it with ERSST.

Smith et al. (2008) used empirical orthogonal teleconnection functions (EOTs; van den Dool et al. 2000) to interpolate LSATs in NOAAGlobalTemp. The EOTs in NOAAGlobalTemp are a set of the empirical orthogonal functions localized within a certain radius in longitude and latitude. The advantages of EOTs are that (i) they can interpolate temperature anomalies to the data-unsampled regions and (ii) they can filter out noise in observations. However, the EOTs may also damp observed signals in the data-sampled region. The filtering is desired in some sense, but the damping can be too aggressive in the regions of islands, peninsulas, or narrow lands, and in data-sparse regions such as in Antarctica and the Arctic

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/AIES-D-22-0032.s1.

Corresponding author: Boyin Huang, boyin.huang@noaa.gov

DOI: 10.1175/AIES-D-22-0032.1 e220032

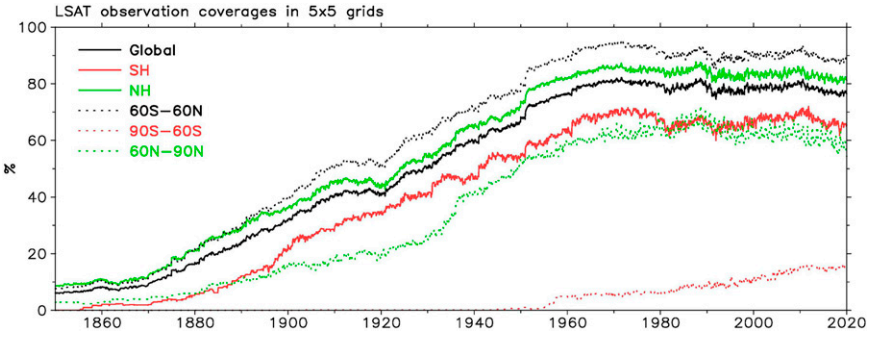

FIG. 1. Coverages (%) of LSAT observations from GHCNm in global (solid black), SH (solid red), NH (solid green), 60°S–60°N (dotted black), 90°–60°S (dotted red), and 60°–90°N (dotted green). Note that all grid boxes in 60°–90°N are counted as "land."

where LSAT observation coverage is low (Fig. 1). Since the EOTs in the data-sparse regions may not be supported by enough observations, the reconstructed temperature field was usually masked out in those regions.

Vose et al. (2021) extended the LSAT component of NOAAGlobalTemp to the Arctic region north of 65°N and applied EOTs over the Arctic region by including additional air temperature observations from ships, buoys, and drifting ice stations to the GHCNm database. The additional observation can certainly increase the confidence in the Arctic region, but still leave room for improvements in Antarctica. In particular, their study indicated that the high-frequency component of LSAT in NOAAGlobalTemp (NOAAGlobalTemp_EOT hereinafter) has very low variability over the Southern Hemisphere (SH) land surface, particularly in the data-sparse area in Antarctica, before the 1950s. This was a primary motivation for developing a new artificial neural network (ANN) interpolation method to better reconstruct the LSAT and therefore a better reconstruction of NOAAGlobalTemp (referred to as NOAAGlobalTemp_ANN hereinafter) by merging with SST.

In this paper, the LSAT reconstruction is improved using an ANN approach over the global land surface and the Arctic Ocean north of 65°N. The reason for using surface air temperature instead of SST is to better describe Earth's warming climate in NOAAGlobalTemp, since the surface air temperature warms rapidly north of 65°N while SST warms little due the existence of sea ice (Vose et al. 2021). The rest of the paper is arranged as follows: section 2 is a brief description of NOAAGlobalTemp, section 3 is a description of a three-layer ANN, section 4 is a description of ANN training and validation methods, section 5 is a comparison between the ANN and EOT methods, and section 6 provides a brief discussion and summary.

2. NOAAGlobalTemp

NOAAGlobalTemp_EOT (Huang et al. 2020a; Zhang et al. 2019; Smith et al. 2008) is a global monthly surface temperature reconstruction with 5° × 5° spatial resolution from 1850 to present. It consists of LSAT from GHCNm v4 on 5° × 5° grid boxes over the global land surface (Menne et al. 2018) and SST from ERSSTv5 on 2° × 2° grid boxes over the global ocean (Huang et al. 2017). The monthly LSAT anomaly of GHCN in a 5° × 5° grid box was generated by averaging all available station-based anomalies within the grid box. The station-based anomalies can be generated when stations have at least some observations during the base period over 1971–2000 so that the climatology can be assessed. The LSAT was reconstructed with EOTs over the global land surface, which did not include the Arctic region. Vose et al. (2021) extended the LSAT reconstruction using EOTs over the Arctic region north of 65°N, which included available air temperatures from ships, buoys, and drifting ice stations.

The monthly LSAT anomalies on 5° × 5° grid boxes over the global land surface and the Arctic Ocean north of 65°N were separated into low- and high-frequency components. The low-frequency component was derived by applying a filter of 25° in latitude and longitude and 15 years in time (Huang et al. 2017, 2015). The high-frequency component, derived initially by subtracting the low-frequency component from the original LSAT observations, was reconstructed using EOTs (Vose et al. 2021; Huang et al. 2020a; Zhang et al. 2019; Smith et al. 2008) and is now reconstructed by an ANN approach in this study. The sum of low- and high-frequency components generates the final LSAT anomalies. The LSAT anomalies are merged with SST anomalies according to the area ratio of land and ocean within a specific grid box.

3. A simple stochastic gradient descent ANN

The ANN provides a robust approach to detect the unknown based on known samples, which could be very complicated due to multiple hidden layers and numerous nodes in the hidden layers (Lippmann 1987; Goodfellow et al. 2017). Modern software for ANN implementation is available and can be implemented in Earth science (e.g., Kadow et al. 2020). To fully understand the ANN system and its internal functions and structures, we have developed a three-layer stochastic gradient descent ANN from scratch to interpolate the observed LSAT anomaly (referenced to the climatology of 1971–2000) on 5° × 5° grid boxes over the global land surface and the Arctic Ocean north of 65°N (Fig. 2). The numbers of input, hidden, and output units are represented by N_1, N_2, and N_3, respectively. Input number N_1 is determined by the number of 5° × 5° grid boxes

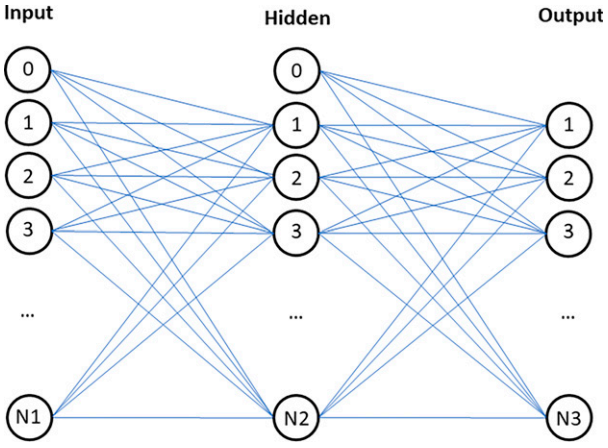

FIG. 2. A three-layer neural network: N_1, N_2, and N_3 represents the number of input, hidden, and output units, respectively; N_1 is determined by available GHCNm observations at 5° × 5° boxes ($N_1 < 1507$); N_3 is total land mask ($N_3 = 1507$); and N_2 is 200. The 0-unit in input and hidden layers represents a noise input. Each line represents the connection between units of two layers.

with valid LSAT observations. Usually, $N_1 < 1507$ because of sparse LSAT observations. $N_3 = 1507$, which represents the total number of 5° × 5° boxes over the global land surface. N_2 is set to 200 in this study. Sensitivity analyses showed that, when N_2 increases from 150, 200, to 300, the quality of interpolation increases very little while computation increases significantly. The inputs to the 0-units in input and hidden layers are a constant of 1 while their coefficients associated with the inputs are trained, which represents a noise input to damp out potential overfitting.

Algorithms of stochastic gradient descent ANN are separated into the forward propagation of inputs and backward propagation of errors (Mitchell 1997). During the forward propagation, the threshold output o_j is calculated as

$$\text{net}_j = \mathbf{w}_j \cdot \mathbf{x}, \tag{1}$$

$$o_j = \sigma(\text{net}_j) = \tanh(\text{net}_j), \quad \text{and} \tag{2a}$$

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}, \tag{2b}$$

where $\mathbf{x}$ represents input from the previous layer of the network, $\mathbf{w}_j$ represents the weighting coefficient, and o_j represents the output value of the current layer of the network. The o_j of the current layer will be used as the input for the next layer of the network. The error of the final network output is defined

$$E_d(\mathbf{w}) \equiv \frac{1}{2} \sum_{k \in \text{Output}} (t_{kd} - o_{kd})^2 \cos(\varphi_k), \tag{3}$$

where k represents the output unit ($k \in$ Output), t represents the target value, d represents training data ($d \in D$), and φ represents the latitude. For each network output unit k, the error gradient is calculated

$$\delta_k \equiv -\frac{\partial E_d}{\partial \text{net}_k} = (1 - o_k^2)(t_k - o_k)\cos(\varphi_k). \tag{4}$$

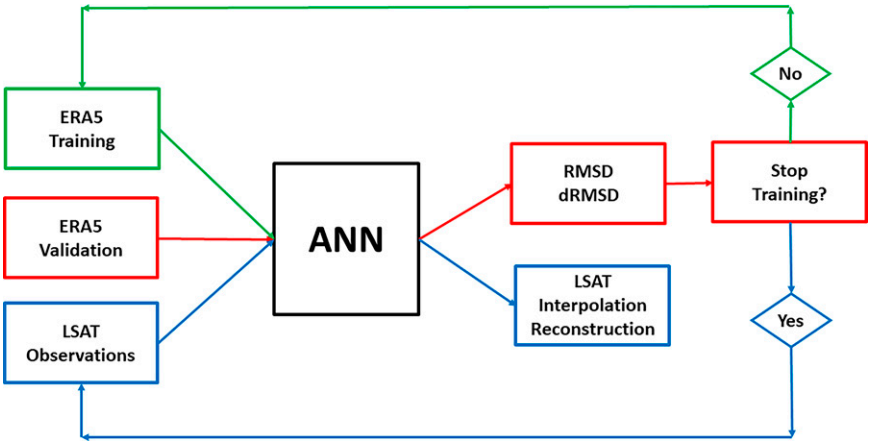

FIG. 3. Schematic processes of ANN training (green), validation (red), and reconstruction (blue): 1) Randomly draw independent ERA5 training and validation data; 2) filter the training and validation data according to GHCNm observation mask; 3) train the ANN by feeding the masked training data and update the ANN network coefficients according to the unmasked training data; 4) feed the trained ANN with the masked validation data, and calculate the RMSD between the unmasked validation data and ANN output and the RMSD difference between two consecutive trainings; 5) determine whether to stop the training; 6) if no, repeat the process in step 3; 7) if yes, feed the trained ANN with GHCNm observations; 8) save the ANN output for GHCNm reconstruction and validation statistics. The processes 1–8 are repeated for each GHCNm observation from January 1850 to December 2020.

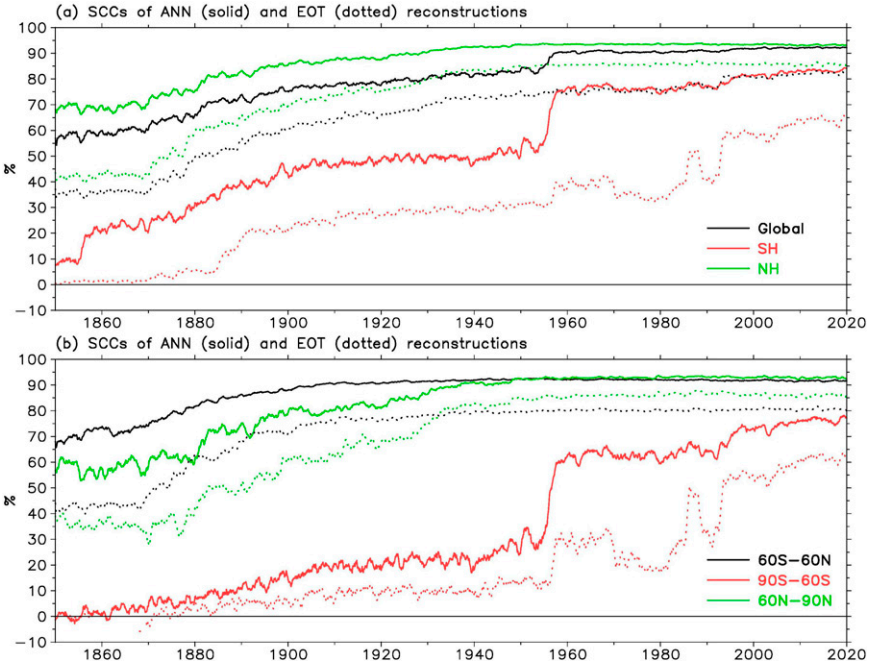

FIG. 4. SCCs (%) of ANN (solid lines) and EOT (dotted lines) reconstructions against independent ERA5 20-member ensemble LSATs over (a) the land surface in global (black), SH (red), and NH (green) and (b) 60°S–60°N (black), 90°–60°S (red), and 90°S–90°N (green). A 12-month running average is applied in plotting.

For each hidden unit j,

$$\delta_j = o_j(1 - o_j) \sum_{k \in \text{Downstream}(j)} \delta_k w_{kj} \cos(\varphi_k). \tag{5}$$

The weighting coefficient is updated according to the learning rate parameter η ($0 < \eta < 1$):

$$\Delta w_{ji} = \eta \delta_j x_{ji} \quad \text{and} \tag{6}$$

$$w_{ji} = w_{ji} + \Delta w_{ji}, \tag{7}$$

where $j = 2, 3$ and $i = j - 1$. The principle of the stochastic gradient ANN is to search for the error minima by reiterations through Eqs. (1)–(7). To avoid local minima and speed up the convergence of searching, a momentum term α ($0 < \alpha < 1$) is added to Eq. (6):

$$\Delta w_{ji}(n) = \eta \delta_j x_{ji} + \alpha \Delta w_{ji}(n - 1), \tag{8}$$

where n represents the nth reiteration. Further, to reduce the overfitting of network coefficients, a penalty term γ ($0 < \gamma < \infty$) is included upon each training reiteration by redefining the error function in Eq. (3):

$$E(\mathbf{w}) \equiv \frac{1}{2} \sum_{d \in D} \sum_{k \in \text{Output}} (t_{kd} - o_{kd})^2 \cos(\varphi) + \gamma \sum_{i,j} w_{ji}^2, \tag{9}$$

which results in updated weighting coefficient

$$w_{ji} = w_{ji}(1 - 2\gamma\eta) + \Delta w_{ji}. \tag{10}$$

The ANN coefficient w_{ji} is initialized as a random number between −0.5 and 0.5. We set $\alpha = 0.5$ and $\gamma = 0$ to optimize the interpolation skill of ANN based on a series of sensitivity tests. Specifically, these tests indicated that the ANN converges faster when momentum α increases from 0 to 0.5; likewise, the globally averaged root-mean-square difference (RMSD) between the validation and ANN output decreases as α approaches 0.5 (section 5a). In contrast, the ANN converges slower and the RMSD increases when α increases from 0.5 to 0.9. The ANN is also sensitive to the penalty γ; when $\gamma = 0.001$, the ANN converges slower and both RMSD and bias increase slightly, particularly in Antarctica.

The learning rate η is set according to the number of units of the first layer input N_1: 0.01 when $N_1 < 100$, 0.012 when $100 \le N_1 < 400$, 0.016 when $400 \le N_1 < 700$, and 0.020 when $N_1 \ge 700$. Intuitively, the larger the learning rate is, the quicker the ANN converges. However, the larger the learning rate is, the more likely it is that the ANN will be unstable. Tests indicated that a lower learning rate is needed when the number of input units is low—that is, when observational coverage is poorer, such as in the nineteenth century (Fig. 1).

4. ANN training and validation

a. *Separation of ERA5 data*

ERA5 (1950–2020; Hersbach et al. 2020) over the global land and the Arctic Ocean north of 65°N is used to train and validate the ANN. The original resolution of ERA5 is 0.25° × 0.25°, which is box averaged to 5° × 5° of the NOAAGlobalTemp for

TABLE 1. Average (1850–2020 and 20-member ensemble) SCCs (%), RMSDs (°C) and biases (°C) of ANN and EOT reconstructions using independent ERA5 LSAT in regions of global, SH, NH, 90°–60°S, 60°S–60°N, and 60°–90°N.

	Global	SH	NH	90°–60°S	60°S–60°N	60°–90°N
SCC, ANN	80	54	87	35	88	82
SCC, EOT	65	29	74	20	72	69
RMSD, ANN	0.57	0.64	0.54	1.75	0.41	0.98
RMSD, EOT	0.99	1.05	0.92	2.03	0.69	1.41
Bias, ANN	0.00	−0.00	−0.00	−0.00	−0.00	−0.00
Bias, EOT	−0.00	−0.00	−0.00	−0.01	−0.00	−0.01

the training and validation purposes. Because LSAT variations are typically larger in winter and smaller in summer, the ANN is trained separately for each month from January 1850 to December 2020.

To train the ANN for one specific month, 3 months of LSAT data from ERA5 are used. For example, the ANN for July 2000 is trained by ERA5 LSAT data of June–August over 1950–2019 (a total of 210 months of data). Furthermore, the ERA5 data is randomly separated into a training set (190 months) and an independent validation set (20 months, about 10% of the total), and LSAT data in 2020 are reserved for independent testing purposes. The use of 3 months of data increases the sample size for training.

b. *ANN training*

As required by ANNs, all LSAT anomaly data for training, validation, and reconstruction are normalized. First, the low-frequency components in all data are filtered (Huang et al. 2017, 2015); second, the maximum and minimum of the high-frequency components are found by screening through all the datasets; third, an average $\mathscr{A}$ and an amplification factor $\mathscr{F}$ are calculated as $\mathscr{A} = (\max + \min)/2$ and $\mathscr{F} = (\max - \min)/2$; and fourth, the anomalies are normalized by subtracting $\mathscr{A}$ and dividing by $\mathscr{F}$ before feeding into the ANN. The normalization can speed up the learning and therefore accelerate the convergence, and the normalized inputs between −1 and +1 are consistent with a superior sigmoid function tanh in Eq. (2b) (Stöttner 2019). The final outputs from the ANN are then retrieved by multiplying $\mathscr{F}$ and adding $\mathscr{A}$.

The ANN is trained each month from January 1850 to December 2020, since the input number N_1 varies from time to time due to changes in LSAT coverage. The trained ANN is validated in a reiterating cycle. In each cycle, the 190 training datasets randomly drawn from ERA5 are fed into the ANN (Fig. 3, boxes in green), and then the 20 validation datasets are fed into the ANN (Fig. 3, boxes in red) to determine whether the ANN is optimized. The total CPU time to train the ANN from 1850 to 2020 is approximately 540 h in Linux 3.10.0–1160.71.1.el7.x86_64, whereas the CPU time for NOAAGlobalTemp using EOTs is about 2 min.

c. *ANN validation*

After each training-validation reiteration, RMSDs between validation and ANN output are calculated for each validation month. The ensemble mean RMSD is used to quantify the performance of the ANN. The training-validation cycle continues until the following three conditions are met:

(i) the difference (dRMSD) of RMSDs between the current and the previous reiteration is less than 1×10^{-5}°C, indicating the ANN is optimized,

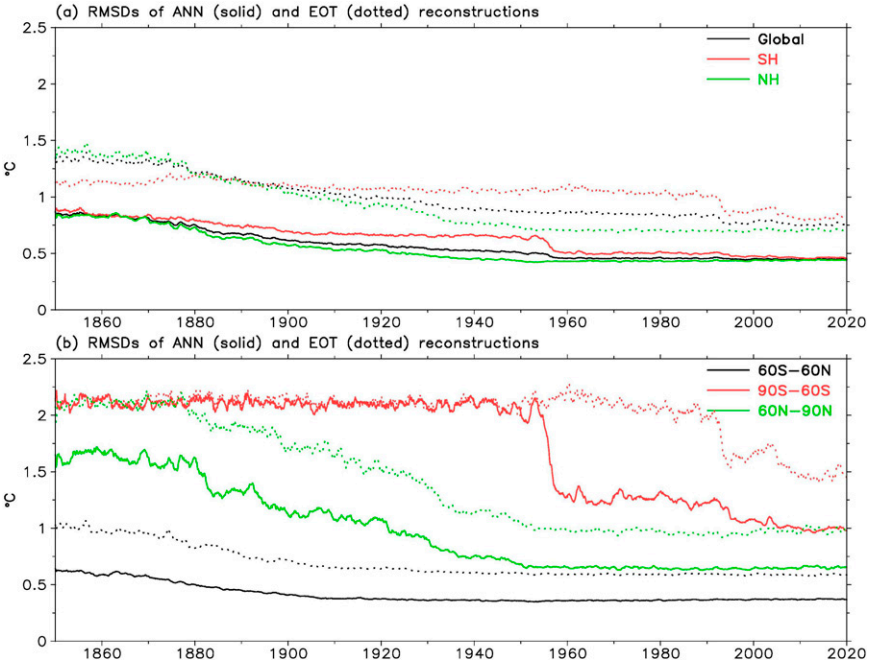

FIG. 5. As in Fig. 4, but for RMSDs (°C).

(ii) the total reiteration is larger than a minimum reiteration of 50 to ensure the performance of the ANN is stable, and
(iii) the ensemble mean RMSD is less than 1.0°C to ensure a reasonable ANN performance.

When only condition i is satisfied while conditions ii and iii are not satisfied, the training is restarted from the very beginning by reducing the learning rate η by 50% and increasing the maximum reiteration number by 50. The training is then continued until the conditions i–iii are satisfied unless one of the following two conditions happens:

(iv) the learning rate is less than its minimum of 1×10^{-4}, or
(v) the reiteration number is larger than its maximum of 300.

Conditions i–iii ensure that the ANN meets certain performance standards. In contrast, conditions iv and v are designed to train a better ANN at a high computational cost when an ideal ANN is difficult to achieve, which mostly happens in the winter months of the SH before the 1900s when LSAT observation coverage and therefore the number of the ANN input units is low.

5. Assessments

a. Methods

The quality of the ANN is quantified using the spatial correlation coefficient (SCC), RMSD, and bias (Huang et al. 2021) between validation V and ANN output O of the jth ensemble member ($1 \le j \le 20$) on 5° × 5° grid boxes:

$$\mathrm{SCC}(j) = \mathrm{cov}(O, V)/\sqrt{\mathrm{var}(O) \times \mathrm{var}(V)}, \tag{11}$$

$$\mathrm{RMSD}(j) = \left\{\frac{1}{W}\sum_{i=1}^{N_3}[O(i) - V(i)]^2 \cos(y_i)^{0.5}\right\}, \quad \text{and} \tag{12}$$

$$\mathrm{bias}(j) = \left\{\frac{1}{W}\sum_{i=1}^{N_3}[O(i) - V(i)]\cos(y_i)\right\}, \tag{13}$$

where

$$\mathrm{cov}(O, V) = \frac{1}{W}\sum_{i=1}^{N_3}[O(i) - \overline{O}][V(i) - \overline{V}]\cos(y_j), \tag{14}$$

$$\mathrm{var}(O) = \frac{1}{W}\sum_{i=1}^{N_3}[O(i) - \overline{O}]^2 \cos(y_j), \quad \text{and} \tag{15}$$

$$\mathrm{var}(V) = \frac{1}{W}\sum_{i=1}^{N_3}[V(i) - \overline{V}]^2 \cos(y_j), \tag{16}$$

where $\overline{O}$ and $\overline{V}$ represent the global average of O and V, respectively; W represents the integrated latitudinal weighting of $\cos(y_j)$; and $1 \le i \le N_3$ represents the ANN output units over the global land surface. The ensemble averages of SCC, RMSD, and bias are finally used to assess the ANN quality.

To compare the quality of the ANN and the EOT reconstruction methods, the same randomly drawn validation data on 5° × 5° grid boxes are first filtered with observation masks

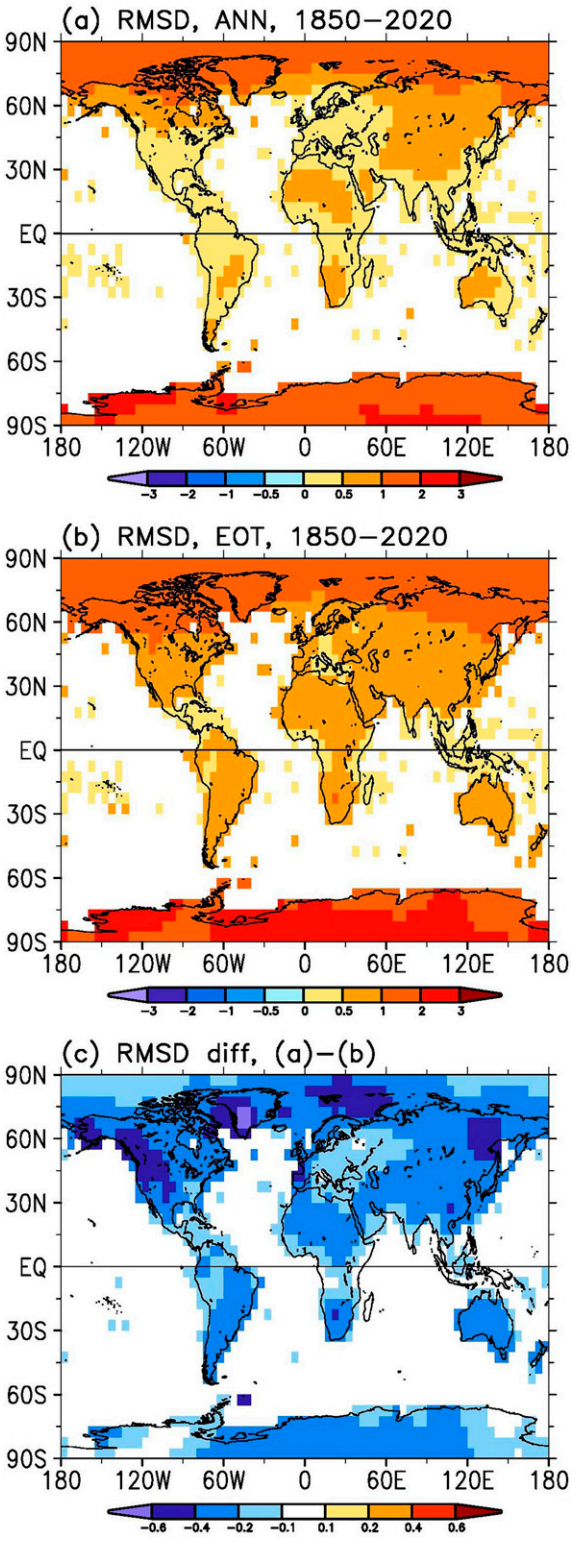

FIG. 6. RMSDs (°C) of LSAT in (a) ANN and (b) EOT methods between 1850 and 2020, and (c) the difference of RMSDs between ANN and EOT methods over the global land and the Arctic.

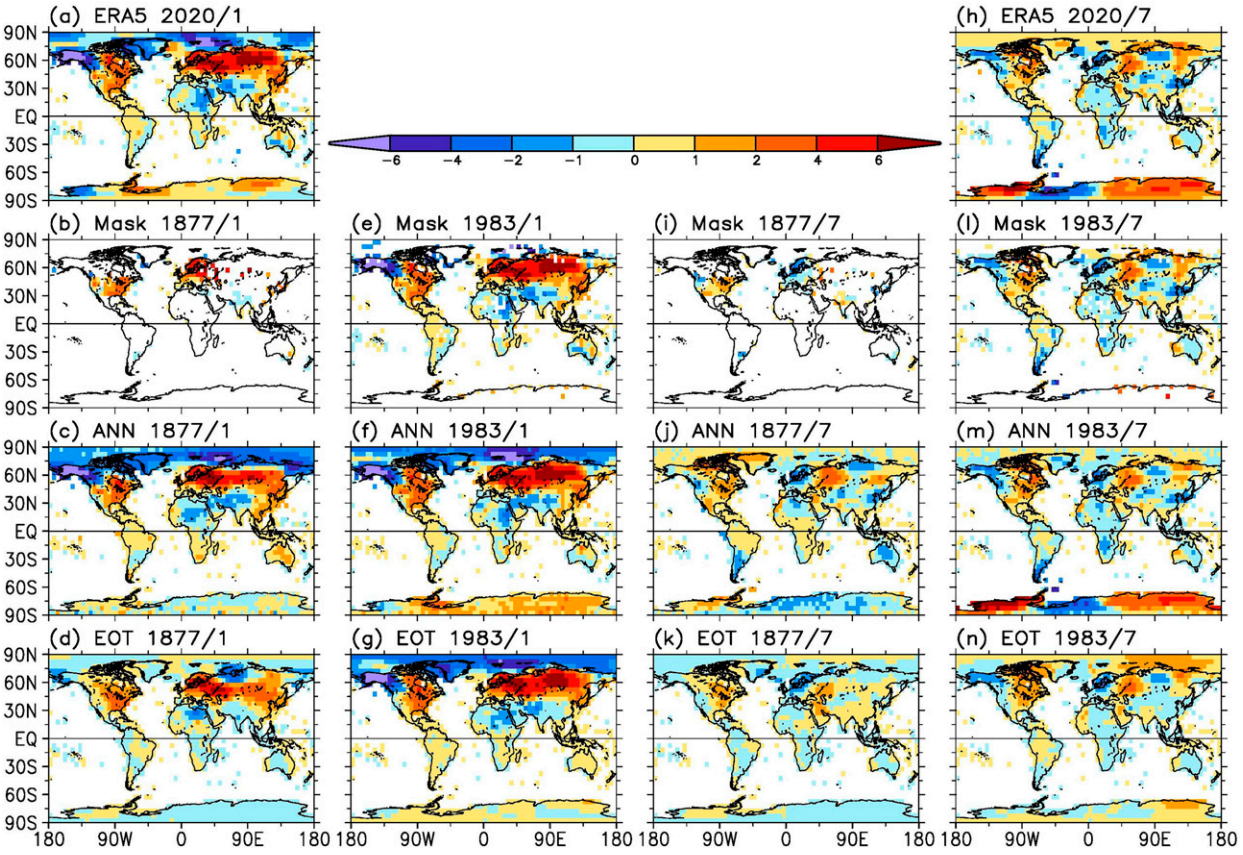

FIG. 7. LSAT anomalies (°C) in (a) ERA5 in January 2020, (b) masked by observations in January 1877, (c) ANN, and (d) EOT reconstructions and (e) masked by observations in January 1983, (f) ANN, and (g) EOT reconstructions. (h)–(n) As in (a)–(g), but using ERA5 in July 2020 and observation masks of July 1877 and 1983. SCCs between ANN and ERA5 are 79%, 96%, 35%, 91% in (c), (f), (j), and (m), respectively; and SCCs between EOT and ERA5 are 56%, 93%, 28%, and 67% in (d), (g), (k), and (n), respectively. The same figure is provided with larger size in Figs. S1 and S2 in the online supplemental material.

at a specific year and month, and then decomposed by the EOT method as in Vose et al. (2021). SCCs, RMSDs, and biases are computed as in Eqs. (11)–(16).

b. *Comparisons with independent validation data*

Figure 4 quantifies the similarity between the ANN reconstructions and 20-member ERA validations, which is quantified by the SCC in Eq. (11). Over the global land surface and the Arctic, the average SCCs of 20-member ensembles are about 55% in 1850 and gradually increase to about 92% by 2020; the overall average from 1850 to 2020 is about 80% (Table 1). The SCCs are slightly higher over the Northern Hemisphere (NH) land surface and the Arctic and increase from about 70% in 1850 to about 93% in 2020. In contrast, the SCCs are much lower over the SH land surface, rising from about 10% in 1850 to about 85% in 2020. The variations of SCCs are consistent with the variations of areal coverage of LSAT observations (Fig. 1), which are higher in NH than in SH, much higher in the Arctic than in Antarctica, and higher in the modern period after the 1950s than in the earlier period before the 1950s. The SCCs are extremely low (less than 30%) in Antarctica (90°–60°S) before the 1950s because of the lack of LSAT observations. In contrast, the SCCs are much higher over the Arctic (60°–90°N) due to a higher LSAT observation coverage. Notably, the EOT method has consistently lower SCCs than the ANN method (about 15% lower globally and about 25% lower in the SH).

Consistent with the increase in SCCs over time, RMSDs for the ANN method decreased from about 0.8°C in 1850 to 0.5°C after the 1950s in the global and both hemispheres (Fig. 5). In the Arctic (60°–90°N), RMSDs fell from about 1.6°C in 1850 to roughly 0.7°C after the 1950s. In the Antarctica (90°–60°S), RMSDs exceed 2.0°C between the 1850s and 1940s due to the lack of observations, then drop to about 1.3°C in the 1950s when more observing stations began to appear on the continent, and then gradually decrease to around 1.0°C in 2020 as the density of the station network further increased. In contrast to the ANN method, RMSDs in the EOT method are several tenths of a degree higher in all regions except the Antarctic before the 1950s, when the two methods exhibit comparable performances. Overall, the average RMSDs from 1850 to 2020 are 0.4° higher in EOT than in ANN globally and in the NH and SH (Table 1).

Figure 6 shows the spatial patterns of RMSDs over the full analysis period. In general terms, RMSDs in the ANN method

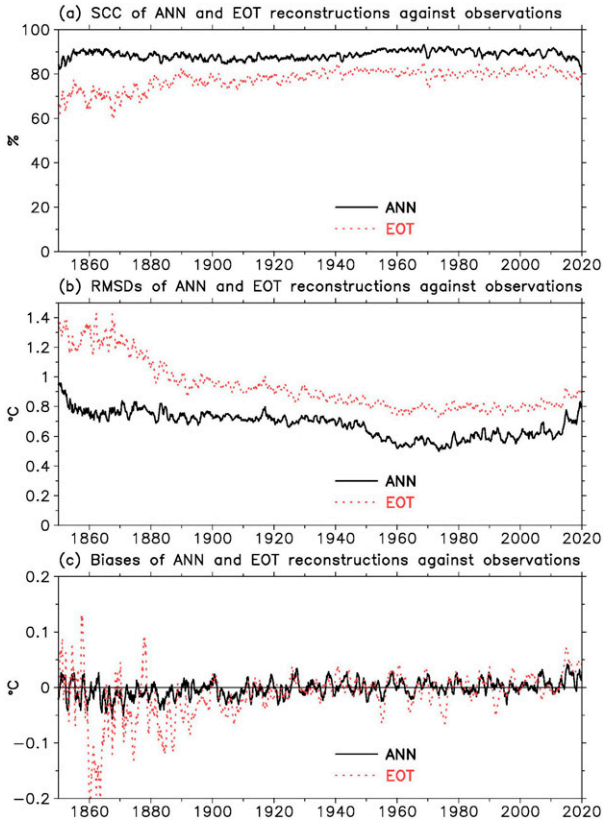

FIG. 8. (a) SCCs (%), (b) RMSDs (°C), and (c) biases (°C) between observed and reconstructed LSATs in ANN (solid black) and EOT (dotted red) over the global land surface. A 12-month running average is applied in plotting.

are about 1.0°C south of 60°S, 0.5°–1.0°C between 60°S and 60°N, and 1.0°–2.0°C north of 60°N. In contrast, RMSDs in the EOT method are about 1.0°–2.0°C south of 60°S, 1.0°C between 60°S–60°N, and 2.0°C north of 60°N (Fig. 6b). The RMSDs in ANN are about 0.1°–0.6°C lower over the global land surface (Fig. 6c), particularly in western North America, Greenland, and parts of the Arctic Ocean.

Biases in both the ANN and EOT methods are generally low (within ±0.005°C) in the global and both hemispheres (Table 1), although they are relatively higher (within ±0.015°C) in Antarctica and the Arctic. The small biases are mostly due to cancellations of errors when regional averages are calculated.

As an additional validation measure, reconstruction performance was assessed using independent ERA5 LSAT for January and July 2020, since both of which were excluded from the development of the ANN and the EOTs. It is assumed that these LSATs were observed in the strong El Niño years in 1877, 1983, and 2016 (Huang et al. 2020b) when LSAT observation coverages differ greatly. In January 2020 (Fig. 7a), the ERA5 analysis depicts warm anomalies in Antarctica, North America, and Eurasia, and cold anomalies across northern Africa, the Middle East, Indochina, and the Arctic. If these LSAT anomalies were observed in January 1877 and therefore they are filtered by observation masks in January 1877

TABLE 2. Average (1850–2020) SCCs (%), RMSDs (°C), and biases (°C) between observed and reconstructed LSATs by ANN and EOT methods in regions of global, SH, NH, 90°–60°S, 60°S–60°N, and 60°–90°N.

	Global	SH	NH	90°–60°S	60°S–60°N	60°–90°N
SCC, ANN	89	74	89	69	88	88
SCC, EOT	78	42	80	45	76	76
RMSD, ANN	0.68	0.60	0.70	0.95	0.63	0.96
RMSD, EOT	0.93	0.85	0.94	1.66	0.86	1.34
Bias, ANN	−0.00	−0.02	−0.00	−0.17	−0.00	−0.00
Bias, EOT	−0.01	−0.03	−0.01	−0.37	−0.01	−0.02

(Fig. 7b), the ANN method (Fig. 7c) reconstructs LSAT anomalies over the global land surface with a SCC of 79% against the validation data in Fig. 7a. In contrast, the EOT method only achieves a SCC of 56%; furthermore, the EOT method does not reconstruct the anomalies well in the Arctic region. If those LSAT anomalies were observed in January 1983 (Fig. 7e), both the ANN (Fig. 7f) and EOT (Fig. 7g) methods successfully reconstruct the global pattern. However, ANN also provides an accurate depiction of the warm anomaly in Antarctica whereas the EOT method does not.

In July 2020 (Fig. 7h), ERA5 analysis shows strong warm (30°E–90°W) and cold (90°W–30°E) anomalies in Antarctica and weak cold anomalies in southern South America, southern Africa. If these anomalies in 2020 were observed in July 1877 (Fig. 7i), the ANN method (Fig. 7j) can reasonably reconstruct those anomalies with a SCC of 56%, although it does not reconstruct the warm anomaly in Antarctica between 30°E and 180°. In contrast, the EOT method (Fig. 7k) only achieves a SCC of 28%. If the anomalies in 2020 were observed in July 1983 (Fig. 7l), the ANN method successfully reconstructs the anomalies in Antarctica with a SCC of 93%, while EOT fails in reconstructing anomalies in southern South America, southern Africa, and most of Antarctica with a SCC of 67%.

c. *Comparisons with observed LSATs*

Figure 8 quantifies the similarity between the reconstructions and the observed LSAT anomalies from 1850 to 2020. It should be emphasized that the SCCs, RMSDs, and biases in this section are calculated between observed and reconstructed LSATs in the regions with valid LSAT observations within a 5° × 5° grid box, which is different from those calculated between LSATs from independent validation and reconstruction over the entire land surface and the Arctic region in section 5a. Figure 8a shows that SCCs over the global land surface are generally higher (80%–95%) in the ANN method than in the EOT method (60%–85%), which suggests that the former can better reconstruct LSATs in regions with observations. Averaged over 1850–2020, the SCCs in the ANN method are 89%, 74%, and 89% for the global, SH, and NH, respectively (Table 2). SCCs in the EOT method are about 12%, 32%, and 9% lower for the global, SH, and NH, respectively.

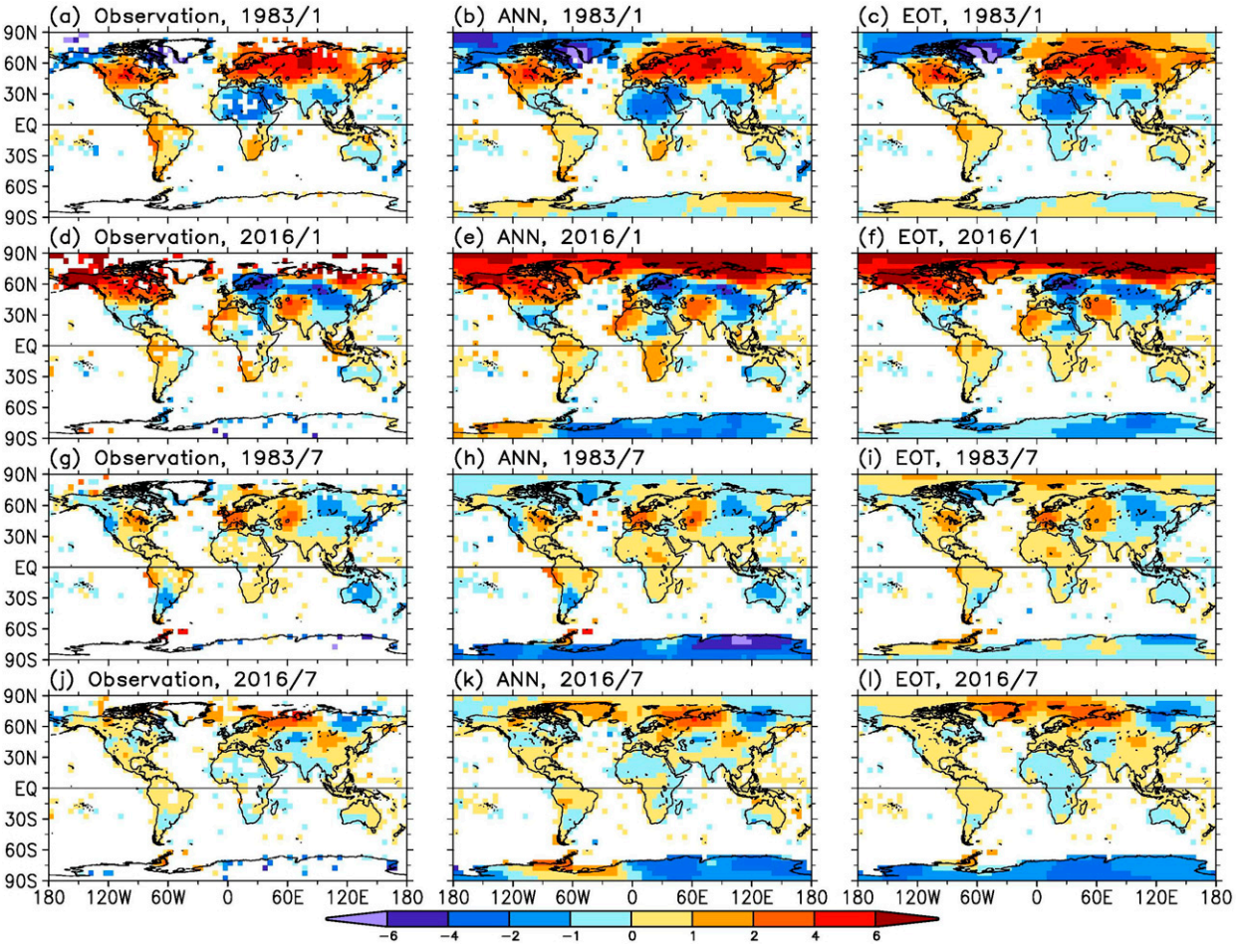

FIG. 9. LSAT anomalies (°C) in (left) observation, (center) ANN, and (right) EOT reconstructions in (a)–(c) January 1983, (d)–(f) January 2016, (g)–(i) July 1983, and (j)–(l) July 2016. SCCs between ANN and observation are 98%, 96%, 97%, and 94% in (b), (e), (h), and (k), respectively; and SCCs between EOT and observation are 96%, 92%, 80%, and 88% in (c), (f), (i), and (l), respectively.

Consistent with the higher SCCs, the RMSDs in the ANN method are lower (Fig. 8b). Specifically, ANN RMSDs are about 0.9°C in 1850, decrease to 0.5°C in the 1960s, and increase to 0.8° in 2020. In contrast, EOT RMSDs are about 1.3°C in 1950, decrease to 0.8°C in the 1960s, and increase slightly to 0.9°C in 2020. The increase in RMSDs after the 1960s may result from a slight decrease in LSAT observations (Fig. 1). Averaged over 1850–2020, RMSDs in the ANN method are 0.68°, 0.60°, and 0.70°C the global, SH, and NH, respectively (Table 2). The RMSDs in the EOT method are about 0.25°, 0.25°, and 0.24°C higher for the global, SH, and NH, respectively.

The biases over the global land surface are generally small (within ±0.05°C) in the ANN method (Fig. 8c). Biases are also small in the EOT method after the 1920s but are slightly higher (±0.20) over 1850–1920. Averaged over 1850–2020, biases are within ±0.03°C except in Antarctica, where they are about 0.2°C larger in the EOT method (Table 2). The higher biases in Antarctica may partly result from sparse observations in that region.

Some examples are given to demonstrate the capability of the ANN method in representing observed LSATs during strong El Niño years. In January 1983 (Figs. 9a–c), observations show warm anomalies in southern South America, southern Africa, and eastern Antarctica (0°–180°), and cold anomalies in Australia, New Zealand, and western Antarctica east of 30°W. These anomalies are well represented in the ANN method but are mostly not captured by the EOT. Similar findings are evident for January 2016 (Figs. 9d–f). In July 1983 (Figs. 9g–i), observations show cold anomalies in southern South America, Australia, and most of Antarctica, which are successfully represented by the ANN method but are not well depicted in the EOT method. In July 2016 (Figs. 9j–l), observations show cold anomalies in Antarctica between 0° and 90°W and warm anomalies between 90° and 0°W. These cold anomalies are well represented in both methods, but the warm anomalies are only captured by the ANN method.

d. Comparisons of LSAT variabilities

In addition to the improvements in SCCs and RMSDs, the ANN method can represent grid-scale LSAT variabilities more reasonably than the EOT method in space and time. Spatial variabilities can be quantified by standard deviations

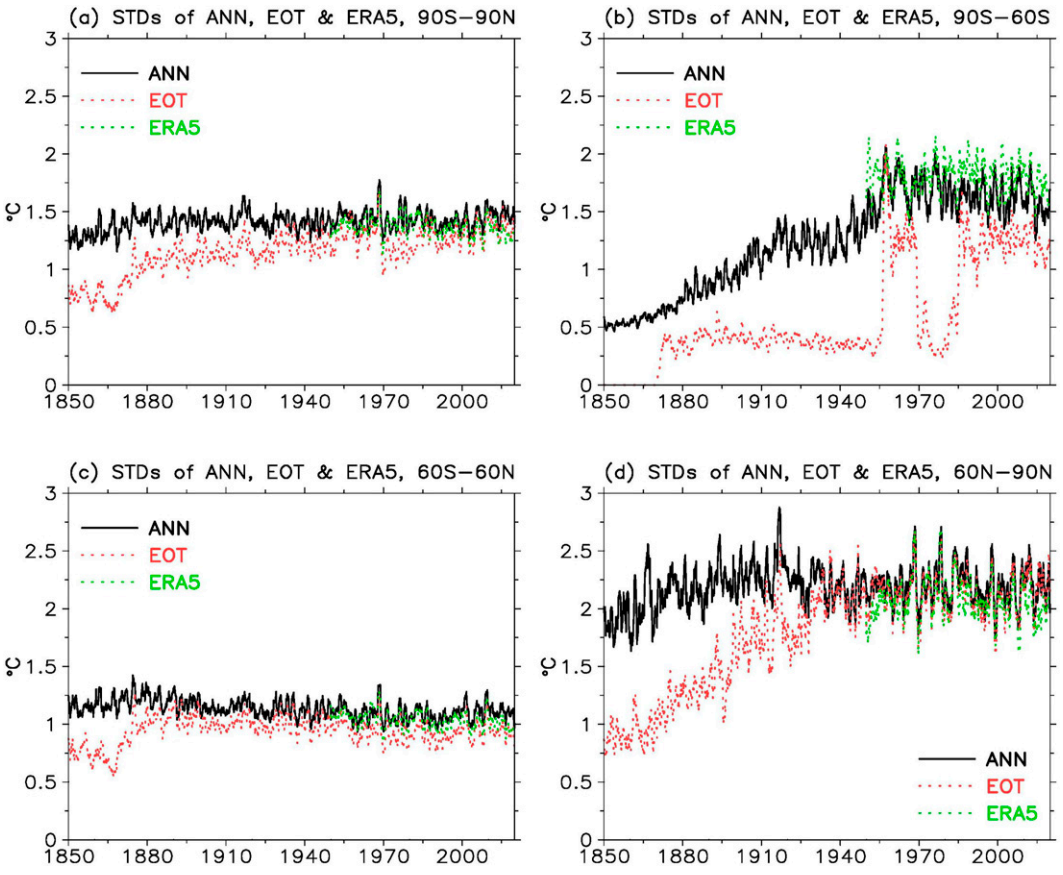

FIG. 10. STDs (°C) of LSAT by ANN (solid black) and EOT (dotted red) methods and ERA5 (dotted green): (a) global, (b) 90°–60°S, (c) 60°S–60°N, and (d) 60°–90°N from 1850 to 2020. A 12-month running filter is applied in plotting.

(STDs) for each month from 1850 to 2020, which are similar to variances in Eqs. (15)–(16) by applying a square root. Here, STDs from ERA5 (1950–2020) are used as a benchmark quantifying LSAT variability, although the reanalysis may have flaws in their variability due to dynamical model bias.

Over the global land surface (Fig. 10a; Table 3), the STDs in the ANN method are about 1.4°C from 1880 to 2020, which are very consistent with those in ERA5 over 1950–2020. In contrast, the average (1850–2020) STDs in the EOT method are slightly lower (1.2°C), which mostly appeared before the 1910s, particularly over 1850–80. Similar features are found over 60°S and 60°N (Fig. 10c). The STDs are very consistent among ANN, EOT, and ERA5 over 1950–2020 and between ANN and EOT over 1880–2020. The lower STDs in EOT in the earlier periods over 1850–80 indicate that the variabilities of LSATs are not well represented. The representation of LSAT variabilities is much worse in Antarctica and the Arctic regions (Figs. 10b,d). In Antarctica, the STDs in ANN are about 0.5°C in 1850 when observations were sparse (Fig. 1); the STDs increase to 1.5°–2.0°C after the 1950s when observations are available, which are consistent with ERA5 although they are a little smaller in ANN than in EOT. In contrast, the STDs in the EOT method are overall much smaller than ANN before the 1950s and smaller than ANN and ERA5 after the 1950s, particularly in the 1970s–80s. In the Arctic region, the STDs in the EOT method are consistent with ANN and ERA5 over 1950–2020 but are about 1°C lower than those in the ANN method before the 1930s.

TABLE 3. Average (1850–2020) STDs (°C) of LSATs by ANN and EOT methods in regions of global, SH, NH, 90°–60°S, 60°S–60°N, and 60°–90°N.

	Global	SH	NH	90°–60°S	60°S–60°N	60°–90°N
ANN	1.41	0.90	1.57	1.24	1.13	2.18
EOT	1.16	0.52	1.35	0.59	0.95	1.78

The difference in the spatial variability of LSAT may result in a difference in its correlation scale (Fig. S3 in the online supplemental material). The correlation scale in ANN (1750 km) is slightly smaller than that in EOT (2100 km), which is consistent with a higher spatial variability in ANN than in EOT. However, the correlation scale in ANN is slightly larger than that in ERA5 (1600 km). The smaller correlation scale in ERA5 indicates its higher spatial variability, which may partially result from its original high spatial resolution (0.25° × 0.25°) although the correlation is calculated using regridded data on the same

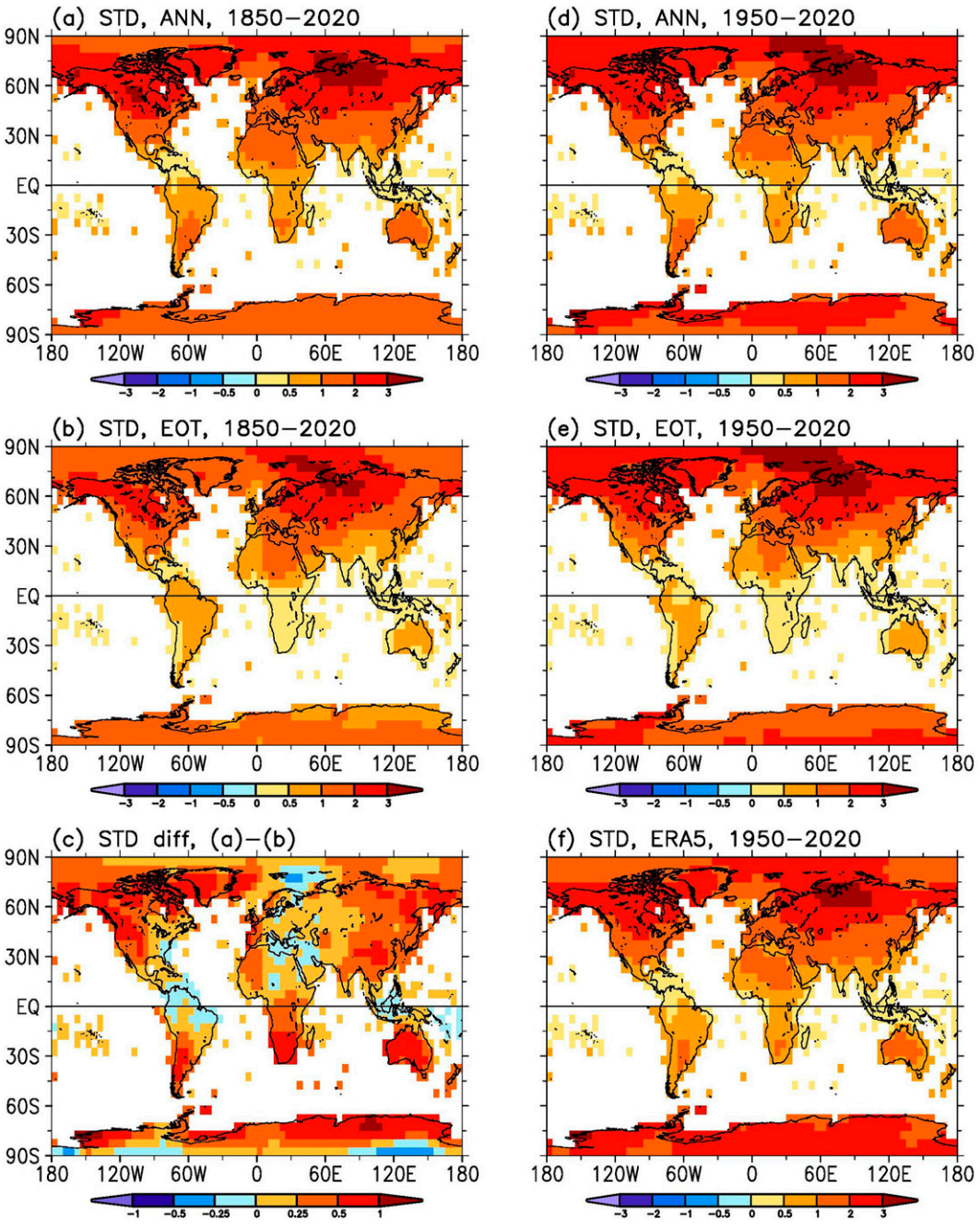

FIG. 11. STDs (°C) of LSAT in (a) ANN and (b) EOT methods between 1850 and 2020, and (c) the difference of STDs between ANN and EOT methods. Also shown are STDs in (d) ANN, (e) EOT, and (f) ERA5 between 1950 and 2020.

5° × 5° grids. However, the correlation scales in these products are larger than that (1200 km) derived from raw station observations (Hansen and Lebedeff 1987), because smoothing and filtering are commonly applied during the data reconstruction.

Similarly, temporal variabilities over 1850–2020 can be assessed by the STDs of LSAT time series in each gridbox. The STDs in the ANN method (Fig. 11a) are 1°–3°C in North America and Eurasia north of 30°N, 0.5°–1.0°C between 60°S and 30°N, and 1°–2°C south of 60°S. Relative to the ANN method, the STDs in the EOT method (Fig. 11b) are low: 1°–2°C in North America, 1°–3°C in Eurasia north of 30°N, 0.5°–1.0°C between 60°S and 30°N, and 0.5°–1.0°C south of 60°S. The differences of STDs (Fig. 11c) are about 0.5°C in North America, eastern Asia, southern South America, southern Africa, and Australia, and 0.5°–1.0°C in Antarctica, indicating that the LSAT variabilities are smaller in the EOT

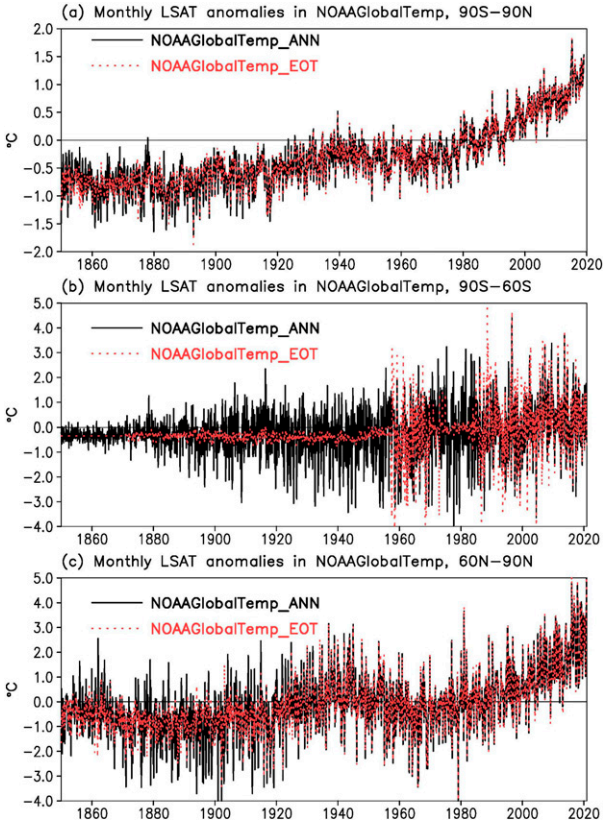

FIG. 12. Monthly LSAT anomalies (°C) in NOAAGlobalTemp_ANN (solid black) and NOAAGlobalTemp_EOT (dotted red) in (a) 90°S–90°N, (b) 90°–60°S, and (c) 60°–90°N.

method than in the ANN method. Over 1950–2020, the STDs are mostly consistent among ANN, EOT, and ERA5 (Figs. 11d–f) over the Northern Hemisphere. In the Southern Hemisphere, the STDs are comparable between ANN and ERA5. However, the STDs in EOTs are clearly underestimated in Antarctica, South America, southern Africa, and Australia.

As an additional illustration, the total LSAT anomalies were calculated by summing the common low-frequency component with the high-frequency component reconstructed by the ANN and EOT methods. Figure 12 shows the averaged LSAT anomalies in the global, Antarctica (90°–60°S), and Arctic (60°–90°N) regions. It is clear that the LSAT variabilities at monthly time scale are smaller in global average before the 1880s, smaller in the Arctic region before the 1930s, and extremely smaller in Antarctica before the 1950s and in the 1970s–80s.

6. Summary, discussion, and conclusions

This paper developed a simple three-layer ANN system to improve the spatial interpolation of LSATs over the global land surface and the Arctic Ocean. The number of input units varied with time and was determined by the LSAT observation mask, and the number of output units was constant and was determined by the land–ocean mask in NOAAGlobalTemp. The ANN system was trained by ERA5 LSAT data from 1950 to 2019; 10% of the ERA5 data were reserved for the purpose of independent validation.

The performance of the new ANN reconstruction method and its EOT predecessor was assessed by computing SCCs, RMSDs, and biases with independent validation data. At the global scale, SCCs were 80% and 65%, and RMSDs were 0.57° and 0.99°C, respectively, in the ANN and EOT methods, indicating a clear performance improvement using the ANN. The improvements were larger in the SH than in the NH and were generally larger before the 1950s. The primary exception was in Antarctica, where almost no improvements were evident until the 1950s, when observing stations became more widespread.

The trained ANN system was applied to interpolate the observed, gridded LSATs from 1850 to 2020, and the resulting reconstruction had demonstrable improvements over the EOT approach at the global scale, SCCs were 11% higher, and RMSDs were 0.25°C lower. Improvements were larger in the SH than in the NH and larger before the 1880s, which is directly associated with the availability of observations.

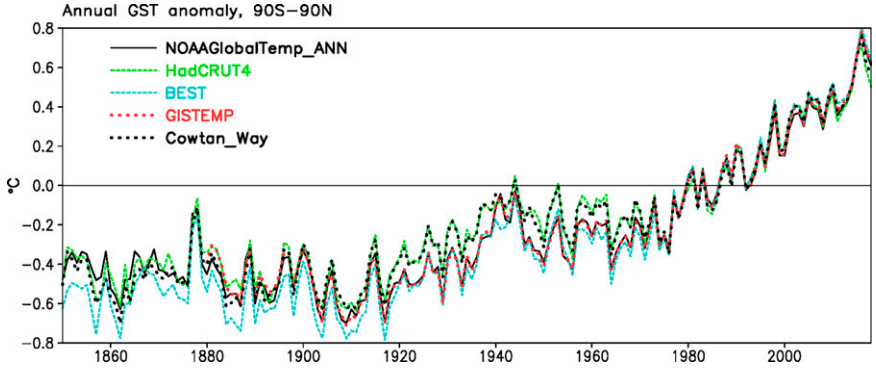

FIG. 13. Annually and globally averaged GST anomaly (°C) for NOAAGlobalTemp_ANN (solid black), the HadCRUT4 (Morice et al. 2012; dashed green), BEST (Rohde and Hausfather 2020; dashed blue), GISTEMP (Lenssen et al. 2019; dotted red), and Cowtan_Way Cowtan and Way (2014; dotted black). Data access dates for HadCRUT4, BEST, GISTEMP, and Cowtan_Way are 10 Jan 2022.

TABLE 4. Linear trends (°C decade^{-1}) and their uncertainty at the 95% confidence level of globally averaged GSTs. Confidence intervals were computed with Student's t test and an effective sampling number based on lag-1 autocorrelations (von Storch and Zwiers 1999).

	1880–2020	1950–2020	2000–20
NOAAGlobalTemp_ANN	0.077 ± 0.011	0.145 ± 0.019	0.229 ± 0.084
NOAAGlobalTemp_EOT	0.077 ± 0.011	0.144 ± 0.019	0.227 ± 0.083
NOAAGlobalTemp v5	0.075 ± 0.011	0.142 ± 0.018	0.221 ± 0.082
HadCRUT4	0.070 ± 0.010	0.125 ± 0.022	0.173 ± 0.076
BEST	0.086 ± 0.011	0.159 ± 0.020	0.244 ± 0.074
GISSTEMP	0.076 ± 0.012	0.151 ± 0.019	0.244 ± 0.075
Cowtan_Way	0.074 ± 0.010	0.130 ± 0.022	0.200 ± 0.069

It is important to note that the low-frequency component of the ANN reconstruction is identical to that of the EOT construction. Consequently, long-term trends remain the same, as do the trend differences between NOAAGlobalTemp and other GST products (Fig. 13; Table 4). The differences among GST products result largely from the differences in algorithms of SST bias correction (Huang et al. 2015, 2017). It should also be noted that this study will be followed by a new uncertainty estimation, which is an important component of NOAAGlobalTemp (Huang et al. 2020a).

The high spatial correlation between reconstructed and validation LSATs in our study is consistent with an earlier study of Kadow et al. (2020; their extended Fig. 5b), which increases from approximately 0.6 in the 1870s to 0.9 in the 2000s. It should be noted that their focus is on the global surface temperature (merged LSAT and SST). In contrast, our focus is on LSAT only, leaving the SST interpolation with the same ANN method for future development due to its high spatial resolution in 2° × 2°. Likewise, the RMSDs between reconstructed and validation LSATs in our study are consistent with their root-mean-square errors (RMSEs; their extended Fig. 3, second row), which are about 2°–3°C in Antarctica and Arctic, 0.5°–1.0°C in Eurasia and northern Africa. However, the RMSEs in tropical South America are slightly higher in their study than in our study, while the RMSEs in North America are slightly lower in their study than in our study. The difference might be associated with the average of different time periods, which are over 1870–2005 in their study and over 1850–2020 in our study, since RMSDs are higher in the earlier time because of sparse observations.

In conclusion, our study shows that the ANN method improves the LSAT interpolation over the EOT method in NOAAGlobalTemp.

Acknowledgments. The authors thank the two anonymous reviewers for providing the thoughtful and constructive comments that helped us to further improve the paper. The authors appreciate the comments from a NOAA internal reviewer, which helped a lot in improving our paper at the initial stage.

Data availability statement. NOAAGlobalTemp_ANN data are available online (https://www.ncei.noaa.gov/pub/data/cmb/ersst/v5/2022.noaaglobatemp.ann): see the NOAAGlobalTemp_ANN directory for global surface temperature data by the ANN method, the LSAT_ANN directory for LSAT data by the ANN method, the ERA5 directory for ERA5 data used for training and validation, and the CODES directory for the Fortan codes and Shell scripts used in this study. Follow the Readme files for instructions on how to read the data files.

REFERENCES

Brohan, P., J. J. Kennedy, I. Harris, S. F. B. Tett, and P. D. Jones, 2006: Uncertainty estimates in regional and global observed temperature changes: A new dataset from 1850. *J. Geophys. Res.*, **111**, D12106, https://doi.org/10.1029/2005JD006548.

Cowtan, K., and R. G. Way, 2014: Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends. *Quart. J. Roy. Meteor. Soc.*, **140**, 1935–1944, https://doi.org/10.1002/qj.2297.

Cressie, N., 1990: The origins of kriging. *Math. Geol.*, **22**, 239–252, https://doi.org/10.1007/BF00889887.

Goodfellow, I., Y. Bengio, and A. Courville, 2017: Convolutional networks. *Deep Learning*, Adaptive Computation and Machine Learning Series, MIT Press, 321–362.

Hansen, J., and S. Lebedeff, 1987: Global trends of measured surface air temperature. *J. Geophy. Res.*, **92**, 13 345–13 372, https://doi.org/10.1029/JD092iD11p13345.

——, R. Ruedy, J. Glascoe, and M. Sato, 1999: GISS analysis of surface temperature change. *J. Geophys. Res.*, **104**, 30 997–31 022, https://doi.org/10.1029/1999JD900835.

——, ——, M. Sato, and K. Lo, 2010: Global surface temperature change. *Rev. Geophys.*, **48**, RG4004, https://doi.org/10.1029/2010RG000345.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quart. J. Roy. Meteor. Soc.*, **146**, 1999–2049, https://doi.org/10.1002/qj.3803.

Huang, B., and Coauthors, 2015: Extended reconstructed sea surface temperature version 4 (ERSST.v4). Part I: Upgrades and intercomparisons. *J. Climate*, **28**, 911–930, https://doi.org/10.1175/JCLI-D-14-00006.1.

——, 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.

——, and Coauthors, 2020a: Uncertainty estimates for sea surface temperature and land surface air temperature in NOAAGlobalTemp version 5. *J. Climate*, **33**, 1351–1379, https://doi.org/10.1175/JCLI-D-19-0395.1.

——, M. L'Heureux, Z.-Z. Hu, X. Yin, and H.-M. Zhang, 2020b: How significant was the 1877/78 El Niño? *J. Climate*, **33**, 4853–4869, https://doi.org/10.1175/JCLI-D-19-0650.1.

——, C. Liu, E. Freeman, G. Graham, T. Smith, and H.-M. Zhang, 2021: Assessment and intercomparison of NOAA Daily Optimum Interpolation Sea Surface Temperature (DOISST) version 2.1. *J. Climate*, **34**, 7421–7441, https://doi.org/10.1175/JCLI-D-21-0001.1.

Kadow, C., D. M. Hall, and U. Ulbrich, 2020: Artificial intelligence reconstructs missing climate information. *Nat. Geosci.*, **13**, 408–413, https://doi.org/10.1038/s41561-020-0582-5.

Kennedy, J. J., N. A. Rayner, R. O. Smith, D. E. Parker, and M. Saunby, 2011a: Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 1. Measurement and sampling uncertainties. *J. Geophys. Res.*, **116**, D14103, https://doi.org/10.1029/2010JD015218.

——, ——, ——, ——, and ——, 2011b: 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.

——, ——, C. P. Atkinson, and R. E. Killick, 2019: 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.*, **124**, 7719–7763, https://doi.org/10.1029/2018JD029867.

Lenssen, N. J. L., G. A. Schmidt, J. E. Hansen, M. J. Menne, A. Persin, R. Ruedy, and D. Zyss, 2019: Improvements in the GISTEMP uncertainty model. *J. Geophys. Res. Atmos.*, **124**, 6307–6326, https://doi.org/10.1029/2018JD029522.

Lippmann, R., 1987: An introduction to computing with neural nets. *IEEE ASSP Mag.*, **4**, 4–22, https://doi.org/10.1109/MASSP.1987.1165576.

Menne, M. J., C. N. Williams, B. E. Gleason, J. J. Rennie, and J. H. Lawrimore, 2018: 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.

Mitchell, T. M., 1997: *Machine Learning*. McGraw-Hill, 414 pp.

Morice, C. P., J. J. Kennedy, N. A. Rayner, and P. D. Jones, 2012: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. *J. Geophys. Res.*, **117**, D08101, https://doi.org/10.1029/2011JD017187.

——, and Coauthors, 2021: An updated assessment of near-surface temperature change from 1850: The HadCRUT5 data set. *J. Geophys. Res. Atmos.*, **126**, e2019JD032361, https://doi.org/10.1029/2019JD032361.

Rohde, R. A., and Z. Hausfather, 2020: The Berkeley Earth land/ocean temperature record. *Earth Syst. Sci. Data*, **12**, 3469–3479, https://doi.org/10.5194/essd-12-3469-2020.

——, and Coauthors, 2013: Berkeley Earth temperature averaging process. *Geoinfor. Geostat.*, **1** (2), 1–13, https://doi.org/10.4172/2327-4581.1000103.

Smith, T. M., R. W. Reynolds, T. C. Peterson, and J. Lawrimore, 2008: Improvements to NOAA's historical merged land–ocean surface temperatures analysis (1880–2006). *J. Climate*, **21**, 2283–2296, https://doi.org/10.1175/2007JCLI2100.1.

Stöttner, T., 2019: Why data should be normalized before training a neural network. Towards Data Science, accessed 25 July 2020, https://towardsdatascience.com/why-data-should-be-normalized-beforetraining-a-neural-network-c626b7f66c7d.

van den Dool, H. M., S. Saha, and A. Johansson, 2000: Empirical orthogonal teleconnections. *J. Climate*, **13**, 1421–1435, https://doi.org/10.1175/1520-0442(2000)013<1421:EOT>2.0.CO;2.

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

Vose, R. S., and Coauthors, 2021: Implementing full spatial coverage in NOAA's global temperature analysis. *Geophys. Res. Lett.*, **48**, e2020GL090873, https://doi.org/10.1029/2020GL090873.

Yun, X., B. Huang, J. Cheng, W. Xu, S. Qiao, and Q. Li, 2019: A new merge of global surface temperature datasets since the start of the 20th century. *Earth Syst. Sci. Data*, **11**, 1629–1643, https://doi.org/10.5194/essd-11-1629-2019.

Zhang, H.-M., and Coauthors, 2019: Updated temperature data give a sharper view of climate trends. *Eos*, **100**, https://doi.org/10.1029/2019EO128229.