1 | INTRODUCTION

Land air temperatures (1.5–2 m above the land surface) are one of the most fundamental climate variables for documenting and understanding rapid environmental changes (Stocker *et al.*, 2013). However, there is increasing evidence that global air temperature datasets are biased due to the sparseness of observations prior to the 1950s, particularly in the Arctic where enhanced climate changes are expected (Cowtan and Way, 2014; Huang *et al.*, 2017; Wang *et al.*, 2017b). The coverage bias is evident in the global dataset assembled by Karl *et al.* (2015), in the HadCRUT4 dataset (Cowtan and Way, 2014), the gridded Berkeley Earth temperature dataset (Way *et al.*, 2017), and others. This bias has led to underestimates in both the Arctic and global temperature trends during the last several decades (Cowtan and Way, 2014; Huang *et al.*, 2017).

There are several approaches for addressing data sparseness issues. Adding more stations to the datasets is the most effective approach for improving the spatial coverage (Wang *et al.*, 2017b). Although it may be possible to fill in missing historical data with auxiliary in-situ measurements in some cases, this approach is unlikely to be successful everywhere, especially during the early historical period (prior to the 1950s). Thus, several mathematical and statistical methods have been developed to interpolate or reconstruct missing data values in historical temperature records. For example, Cowtan and Way (2014) used kriging to spatially interpolate the HadCRUT4 global temperature dataset, Huang *et al.* (2017) used Data INterpolating Empirical Orthogonal Functions (DINEOF) to improve coverage in the Arctic, while Wang *et al.* (2017a) used the Biased Sentinel Hospitals Areal Disease Estimation (BSHADE) method to improve global coverage over land.

Similar methodologies are currently used to produce the most widely used global temperature datasets such as HadCRU and Berkeley Earth. Briefly, the first step is to construct a baseline or background temperature map which is generally created from data over a 30 year period. Temperature anomalies are then produced by subtracting the baseline temperatures and the anomalies spatially interpolated to produce a map product. A variety of interpolation methods have been used and these are continually evolving. For the HadCRU products, triangulated linear interpolation (Harris *et al.*, 2014) was used for the third version while Angular Distance Weighting has been used since version 4 (URL: https://crudata.uea.ac.uk/cru/data/hrg/cru_ts_4.03/Release_Notes_CRU_TS4.03.txt). Meanwhile, Berkeley Earth used kriging to implement spatial interpolation (Rohde *et al.*, 2013). An important issue with these interpolation methods is that they rely on a distance relationship, which is problematic for the early period because of the sparse station coverage in the Arctic and elsewhere (examples will be shown in Figure 6).

For global air temperatures, spatial and temporal coverage issues are not independent, an issue that the DINEOF method (Beckers and Rixen, 2003) is designed to cope with. It determines the number of statistically significant empirical orthogonal functions (EOF) by a cross-validation procedure for incomplete datasets, as well as quantification of the noise level and interpolation errors. This method does not need prior information about the error covariance structure so it is self-consistent and parameter-free (Beckers and Rixen, 2003). By estimating both the spatial and temporal EOFs, it considers the spatiotemporal features in a set of available records. It has been applied successfully to incomplete spatial images, e.g., mapping changes in sea surface temperatures (Alvera-Azcárate *et al.*, 2005) and salinity (Alvera-Azcárate *et al.*, 2016). Huang *et al.* (2017) recently applied DINEOF to reconstruct missing data gaps in annual air temperature records for the Arctic.

In this study, we use the DINEOF method to produce a reconstructed mean monthly air temperature dataset for global land areas over the period 1880–2017. The raw dataset used in this study is the Global Historical Climatology Network (GHCN)-monthly temperature dataset Version 4 (GHCNm V4), which includes many more data sources than previous versions (Peterson and Vose, 1997; Lawrimore *et al.*, 2011). The recent version (V4) provides mean monthly air temperature records from more than 27,000 stations (Menne *et al.*, 2018). Two different validation methods are used to evaluate the accuracy of our reconstructed temperature data. Finally, the temporally extended station data significantly improve the spatial coverage during the period prior to the 1950s. This product is expected to be useful for global and regional climate analyses and provides a new perspective for improving global temperature products in the future.

2 | DATA PRODUCTION METHODS

A work-flow procedure was designed to prepare, reconstruct, and post-process station data from the GHCNm V4 network (Figure 1).

2.1 | Step 1: Input file preparation

In this step, raw GHCNm V4 data were filtered and converted to the format required by DINEOF software: (a) Stations with missing latitude, longitude, or elevation information were deleted. (b) To avoid short records, stations with <10 years of data for any month were removed. (c) Any records beyond the 1880–2017 time period were discarded. (d) Bad or questionable data were eliminated based on the quality flags in the GHCNm V4 raw files. (e) To improve the robustness of the reconstruction, a 5σ principle was used to detect and remove

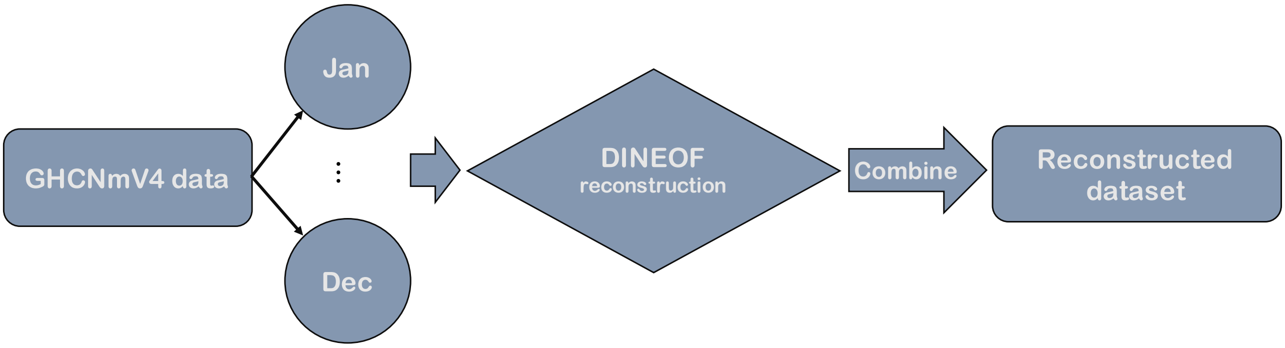
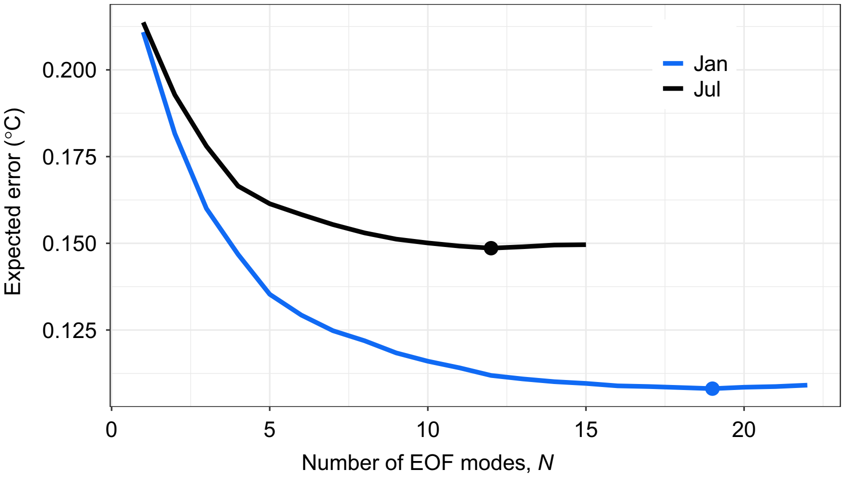


FIGURE 1 Schematic overview of the workflow used to process and reconstruct the station data

FIGURE 2 Cross-validation error convergence with increasing number of EOFs (N) used for the reconstruction of January and July monthly air temperatures. The minimum error (filled circles) determines the optimal number of EOF modes for each month. Note that the expected errors are for the normalized temperature records rather than the true temperature data



any remaining outliers, i.e., any data points beyond ±5 standard deviations of the long-term mean were removed. (f) The preprocessed data were split into 12 monthly (binary) files to facilitate separate analysis with the DINEOF software. In the end, a total of 26,253 stations were used in this study, i.e., ~5% of the GHCNm V4 sites were removed because of insufficient record length or missing station location information.

2.2 | Step 2: Reconstruction of mean monthly temperature data using DINEOF

The reconstruction procedure follows that described by Beckers and Rixen (2003), although here we treat each month as an independent time series. Briefly, the procedure was: (i) Initial input monthly data were normalized by the temporal mean and variance at each station and the missing data values set to zero. This created a normalized spatiotemporal temperature matrix **X**. (ii) Missing data values were replaced by estimates obtained from the EOF analysis. This must be done iteratively because the calculated EOFs depend to some extent on the missing data values. In addition, the missing data estimates depend on the total number N of EOFs retained in the analysis. In our analysis, we let N range from 1 to 50 where the EOFs had been ordered according to their significance as measured by the magnitude of their associated singular values. For each N, a Singular Value Decomposition (SVD) was performed to decompose the temperature matrix **X** into spatial and temporal EOFs. Utilizing the first N EOFs, the missing data values were estimated and the EOFs reevaluated. This step was repeated until convergence occurred. Once convergence was obtained, the missing data/EOF evaluation was repeated using $N + 1$ EOFs. (iii) Once the spatial and temporal EOFs were known for each N, the cross-validation technique described by Beckers and Rixen (2003) was applied to determine the optimal number of EOFs to retain in the final reconstruction; the optimal number of EOFs occurs when the cross-validation error is minimized (Figure 2). For the GHCNm V4 dataset, we found that the

TABLE 1 Optimal number of EOF modes for each month

Month	N	Month	N
January	19	July	12
February	18	August	13
March	18	September	17
April	18	October	16
May	14	November	16
June	11	December	18

optimal number of EOF modes varied from 11 to 19, depending on month (Table 1); the optimal number of modes was generally higher during the northern hemisphere's winter months. (iv) The first two steps (*i* and *ii*) were then repeated using only the optimal EOF modes. (v) A new normalized spatiotemporal temperature matrix **X2** was created using the optimal spatial and temporal EOF modes for each month. Missing data values in the original matrix **X** are replaced by optimal EOF estimates in **X2**. (vi) The temporal mean and variance were then added back to the reconstructed matrix **X2** to recover the actual temperature.

2.3 | Step 3: Output file preparation

The 12 reconstructed monthly temperature matrices **X2** were merged into a single file. Details concerning the file format are described in Section 4.

3 | VALIDATION

To validate the reconstructed dataset, we first randomly set ~10% of the total available data to missing data values and used the same method as Section 2 to reconstruct these artificial missing points. Comparing the reconstructed artificial missing data with the true observations, the mean absolute error

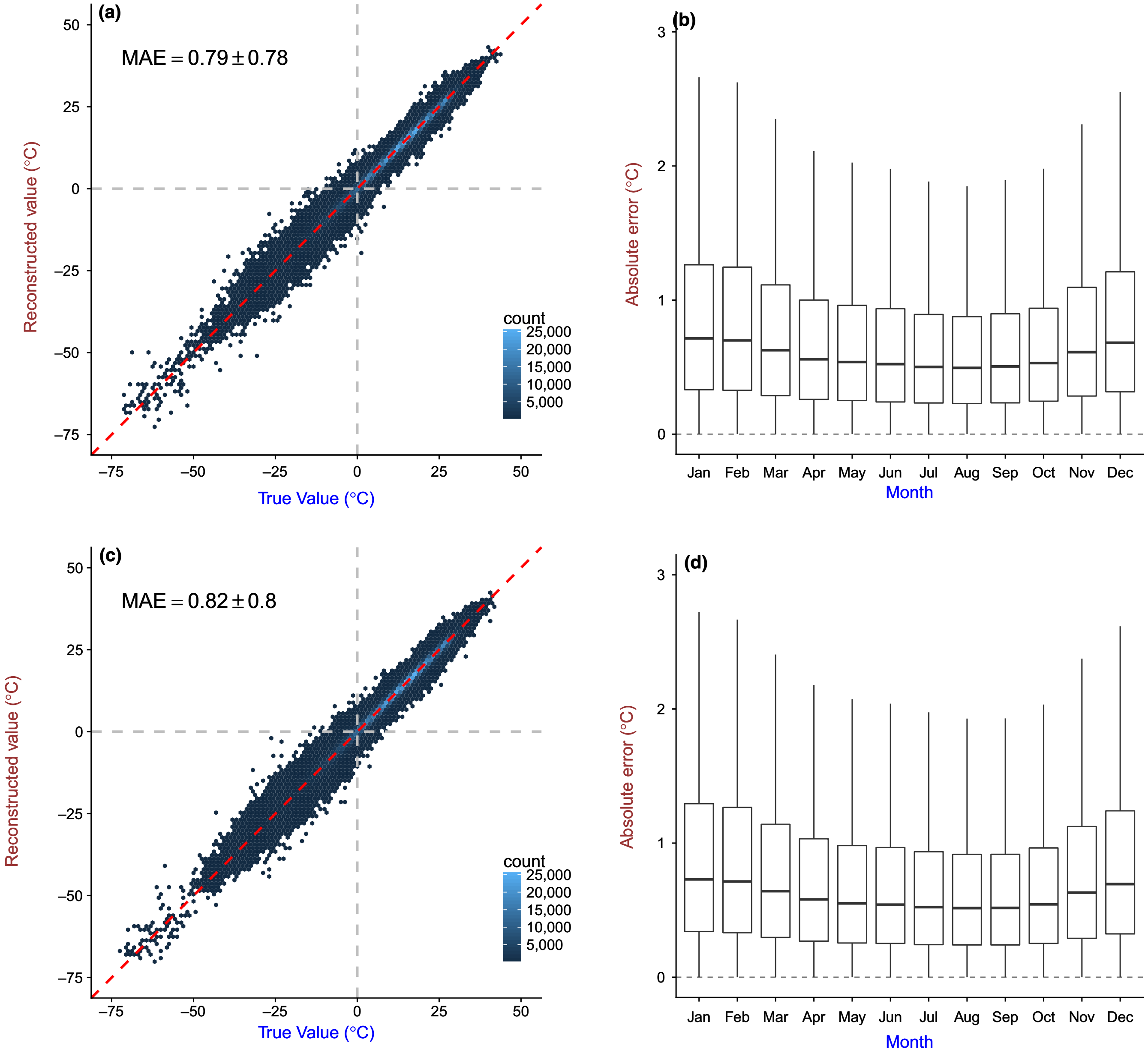


FIGURE 3 Overall accuracy (a) and monthly biases (b) for pure random missing value experiments; (c) and (d) are same with (a) and (b), but for the block missing value experiments. Red dashed lines in (a) and (c) are 1:1 lines. MAE is the mean absolute error

TABLE 2 Monthly mean absolute errors (unit: °C) in reconstructed temperature records

	Random			Block		
Month	**Mean**	**STDEV**[a]	***N***[b]	**Mean**	**STDEV**	***N***
Jan	1.02	0.99	104,313	1.05	1.02	106,424
Feb	0.99	0.95	106,236	1.02	0.99	108,056
Mar	0.86	0.82	105,395	0.88	0.85	106,399
Apr	0.76	0.74	106,057	0.79	0.76	107,870
May	0.72	0.68	106,257	0.75	0.71	107,918
Jun	0.69	0.64	105,964	0.72	0.66	109,150
Jul	0.67	0.63	106,316	0.69	0.65	107,734
Aug	0.65	0.59	106,471	0.67	0.61	108,400
Sep	0.67	0.62	106,230	0.68	0.61	107,402
Oct	0.71	0.66	105,815	0.73	0.69	109,588
Nov	0.84	0.80	106,472	0.87	0.84	107,829
Dec	0.96	0.92	105,366	0.98	0.94	107,990

[a] STDEV: standard deviation of MAEs across grids.

[b] N is number of monthly temperature pairs between reconstructed and true record.

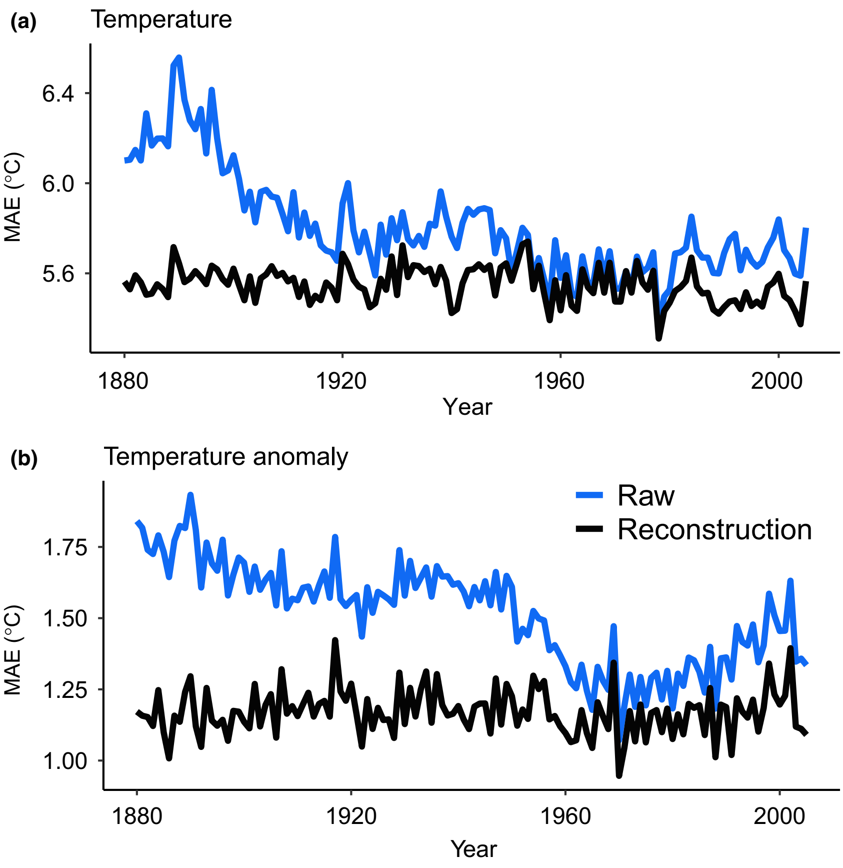


FIGURE 4 Differences in mean monthly air temperature between the gridded raw (reconstructed) GHCNm V4 dataset and CMIP5 climate model output (a); Differences in temperature anomalies between the raw (reconstructed) GHCNm V4 data and CMIP5 output (b). Differences (absolute and anomaly) are represented by the mean absolute error (MAE), which is derived from raw GHCNm V4 dataset (blue curve) and reconstructed dataset (black curve)

(MAE) of the reconstruction is found to be ~0.79 ± 0.78°C (Figure 3a). The estimated errors vary by month with the largest error occurring in January (~1.02 ± 0.99°C) and the smallest occurring in August (~0.65 ± 0.59°C) (Figure 3b, Table 2).

Second, we used block sampling to validate the case where continuous blocks of data 13 years long (about 10% out of the total 138 years) are missing. The first step was to randomly sample ~10% of the stations. For each of these stations we then artificially set a contiguous block of 13 years to missing data values. Reconstructing the artificial missing data as before and comparing with the true observations, we find the overall accuracy of the reconstruction (i.e., the average of MAEs between true observations and reconstructions)

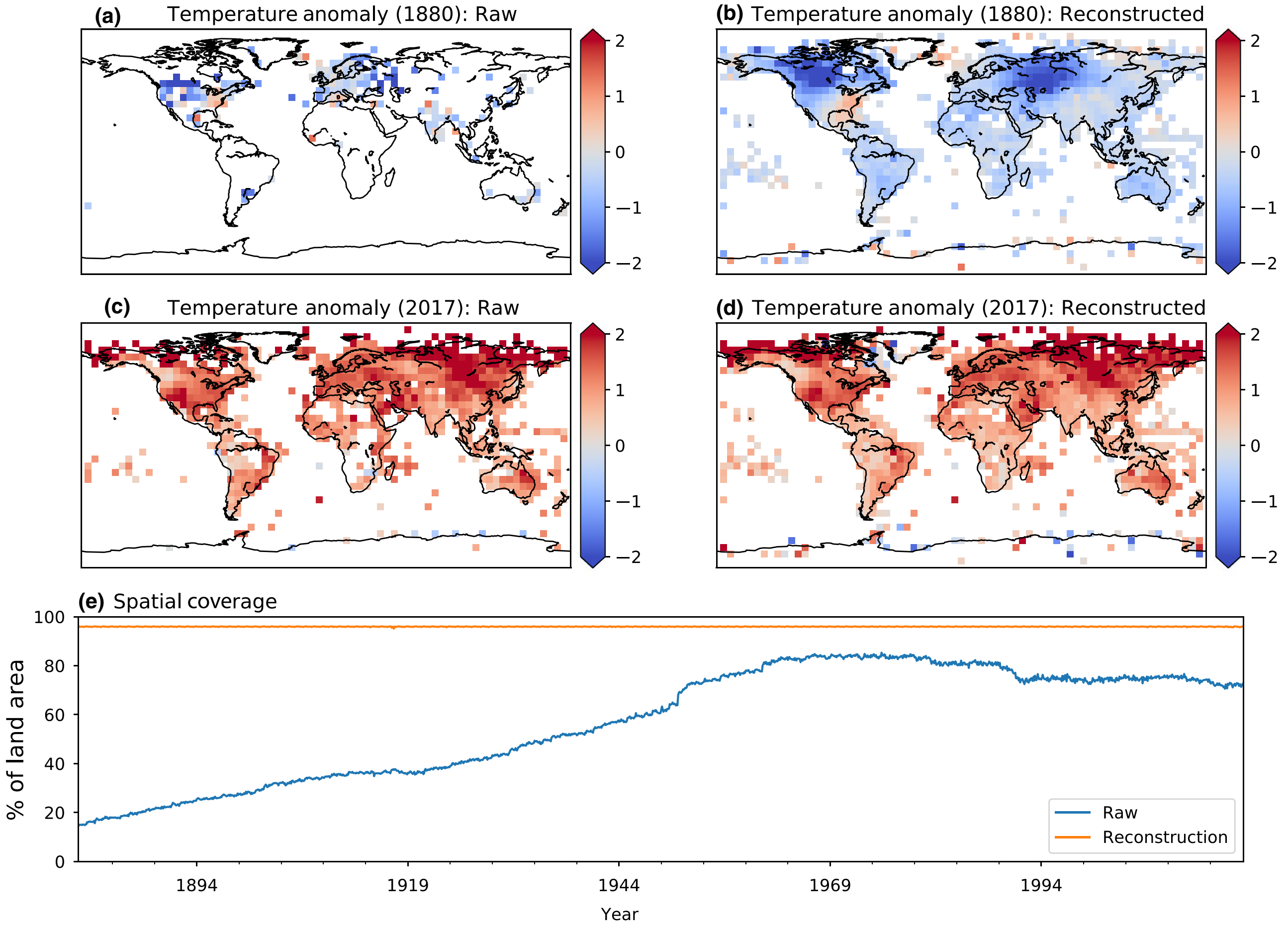

(a) Temperature anomaly (1880): Raw
(b) Temperature anomaly (1880): Reconstructed
(c) Temperature anomaly (2017): Raw
(d) Temperature anomaly (2017): Reconstructed
2
1
0
−1
−2
(e) Spatial coverage
% of land area
100
80
60
40
20
0
1894
1919
1944
1969
1994
Year
Raw
Reconstruction

FIGURE 5 Panels (a) and (c) are gridded raw GHCNm V4 temperature anomalies during 1880 and 2017, respectively. Panels (b) and (d) are the same as (a) and (c) but for the reconstructed data. Spatial resolution is a regular 5° × 5° latitude-longitude grid. Panel (e) shows the spatial coverage of the raw and reconstructed datasets

is ~0.82 ± 0.80°C (Figure 3c), which is slightly larger than for the purely random sampling experiments. As before, MAEs vary by month with the maximum error occurring in January (~1.05 ± 1.02°C) while the minimum occurs during August (~0.67 ± 0.61°C) (Figure 3d, Table 2).

In summary, validation tests indicate the DINEOF reconstruction is reliable. The accuracy of the reconstruction during May – September is slightly better than during the other months. This is because the majority of the land stations are located in the Northern Hemisphere where larger mean absolute errors occur during the cold seasons.

An auxiliary way to evaluate the reconstructed dataset quality is to compare it with climate model output. Ordinarily, observations are used to evaluate climate model performance. However, climate model output can also be used to assess observational data quality as pointed out by Massonnet *et al.* (2016). The basic assumption is that observational data and climate model output play symmetrical roles so it is possible to use one to infer how close the other is from the true state, and vice versa. If observational data quality is consistent with time, the difference between the observations and the model outputs should remain relatively constant.

To reevaluate the observational data quality, we compared both the raw and reconstructed GHCNm V4 mean monthly air temperatures in each 5° × 5° grid box with CMIP5 ensemble climate model output; CMIP5 ensemble outputs were obtained from http://climexp.knmi.nl/selectfield_cmip5.cgi. The algorithm used to grid the GHCNm V4 temperatures was from Osborn and Jones (2014). We also implemented a similar comparison using temperature anomalies where the anomalies were calculated relative to the 1961–1990 climatological means. Mean absolute errors between the GHCNm V4 datasets and CMIP5 output were calculated for each 5° × 5° grid cell and then averaged over the entire spatial domain. As indicated by the MAE trends, differences between the gridded raw GHCNm V4 temperatures and CMIP5 output varies substantially with time (Figure 4a) with the differences being significantly larger prior to the 1920s. The maximum bias during the early part of the record exceeded 6.50°C. In contrast, the differences generally were around 5.60°C in recent

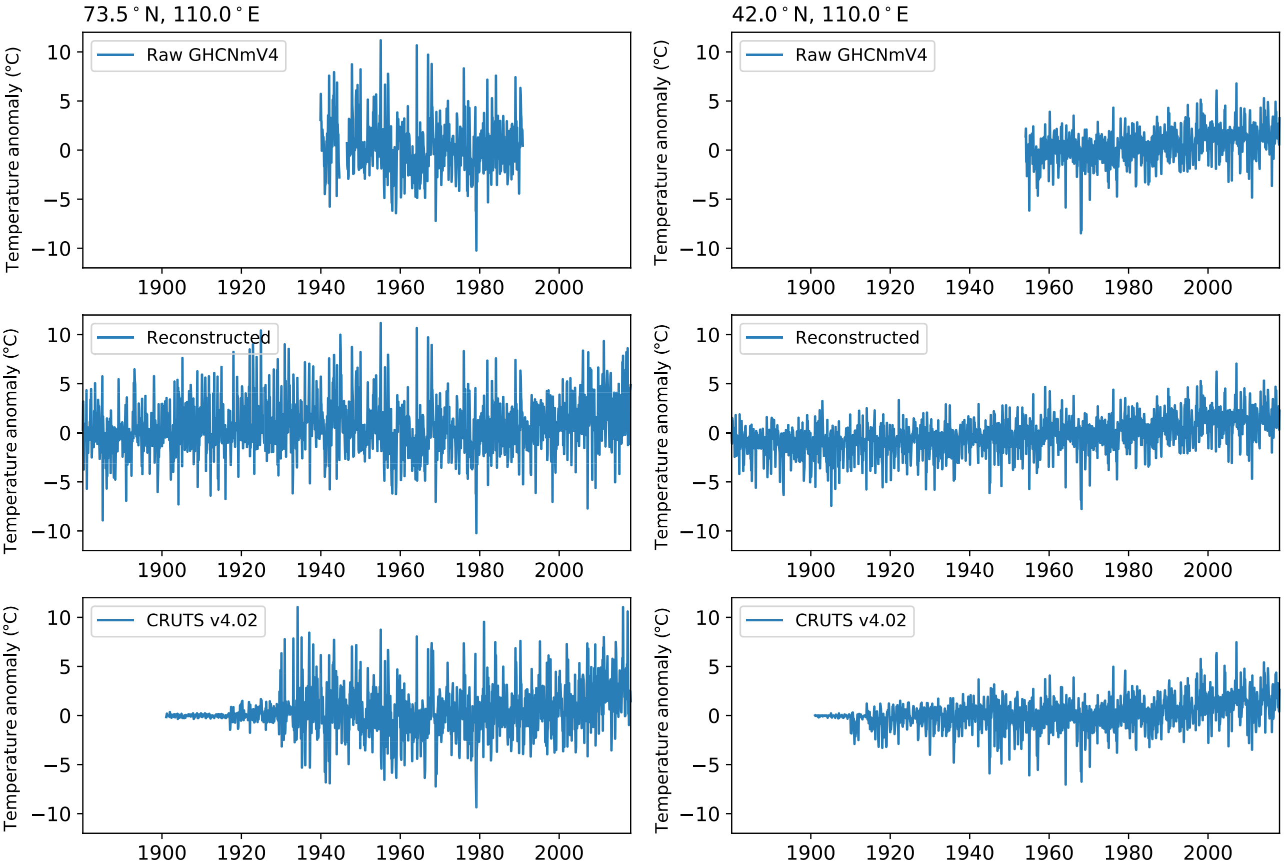


FIGURE 6 Left panels (top to bottom) are the raw GHCNm V4 air temperature anomalies, reconstructed GHCNm V4 air temperature anomalies, and CRU TS4.02 temperature anomalies at 73.5°N, 110°E. Right panels are same as on the left but for the area around 42.0°N, 110°E

decades (since the 1950s). Similarly, temperature anomalies show a declining bias over time (Figure 4b).

In comparison with the raw GHCNm V4 data, differences between the reconstructed GHCNm V4 air temperatures and CMIP5 output were much more homogeneous over the entire period (Figure 4), i.e., differences between the reconstructed temperatures and the model output remain ~5.60°C over the entire period while differences in the temperature anomalies generally remain less than 1.25°C. This implies that the reconstructed data during the early period are as reliable as that in recent decades. Thus, the reconstructed dataset effectively overcomes the time-dependent coverage bias inherent in the raw dataset (Figure 5e).

Figure 5 provides a quick visualization of the gridded temperature anomalies from the raw GHCNm V4 and our reconstructed datasets. Considering a regular 5° × 5° latitude-longitude grid (as it is widely used in existing global temperature products), the spatial coverage has increased to ~95%. The reconstruction improves the spatial coverage by about 80% during the earliest decades (1880–1900) and by 10%–20% since the 1950s.

To illustrate the improvement between our reconstruction and CRU TS4.02 (a widely used spatial interpolation product by Climate Research Units), the temperature anomalies for two locations (73.5°N, 110°E and 42.0°N, 110°E) are shown in Figure 6. As reported by Macias-Fauria *et al.* (2014), the CRU TS4.02 product has insufficient temporal variability during the early period (lower panels, Figure 6). This defect is likely due to how the CRU spatial interpolation method behaves with sparse data coverage. Our reconstructed temperature product significantly improves this aspect (middle panels, Figure 6).

4 | DATASET LOCATION AND FORMAT

The global monthly air temperature dataset is available at *figshare*, https://doi.org/10.6084/m9.figshare.7961120. Three files are available at this site:

1. List.dat: This file provides the GHCNm V4 identification information for each climate station, including: alphanumerical station ID (column 1), numerical station ID code (column 2), longitude (column 3), latitude (column 4), and elevation (column 5).
2. Raw_GHCN4_Data.dat: This file contains all the raw data we used for the reconstruction after applying the quality

control and reformatting procedures (Step 1, Section 2). The file consists of 14 columns: station ID code consistent with *List.dat* (column 1), year (column 2), and monthly air temperatures for January through December at 0.01°C resolution (columns 3–14). Missing data are flagged as −9999, a convention that is maintained throughout this study.
3. Reconstructed_GHCN4_Data.dat: This file provides all the reconstructed data. It has exactly same data structure as *Raw_GHCN4_Data.dat*.

5 | DATASET USE AND REUSE

The reconstructed global land air temperature dataset produced by this study can facilitate a variety of studies relevant to climatic and environmental research at regional, national, or continental scales. We caution that the reconstructed dataset may have several potential limitations.

1. Some studies using the previous version of the GHCNm (version 3) found the homogenized data may be significantly biased in regions where the station spatial coverage is sparse such as in the Arctic (e.g., Wang *et al.*, 2017b). In GHCNm V4, the station data at high-latitudes (≥65°N, ≥65°S) has been separated during the data homogenization process to reduce biasing in these regions (Lawrimore, 2018). However, there may still be unknown issues for stations located in high-altitude regions such as the Qinghai-Tibet Plateau.
2. The DINEOF procedure used here fills gaps for all available stations, extending the limited data back to the late 19th century. The resulting reconstructed dataset substantially increases the spatial coverage, particularly during the early period. However, the procedure does not perform a spatial interpolation. Thus, the reconstructed dataset still contains blank areas where no station data exists.
3. The DINEOF method depends partly on the length and quality of input time-series. Thus, stations with records that are too short (<10 years) have been removed from this study. As data records are added in the future, these stations may be included in the analysis once the record duration criterion is satisfied. Although quality controls of raw station records were implemented by both this study and GHCNm V4, there may still be questionable values at a few sites.

ACKNOWLEDGEMENTS

We thank the GeoHydrodynamics and Environment Research group at the University of Liège for freely providing the DINEOF software (http://modb.oce.ulg.ac.be/mediawiki/index.php/DINEOF) and NOAA for compiling and making the world-wide station data publicly accessible. This study was supported by the U.S. National Science Foundation (Award No. 1503559). Pre- and post-processing codes in MATLAB are available from the authors upon request. The gridded data files with NetCDF format are available upon request. Publication of this article was funded by the University of Colorado Boulder Libraries Open Access Fund.

CONFLICT OF INTEREST

The authors declare that there are no conflicts of interest.

ORCID

Kang Wang https://orcid.org/0000-0003-3416-572X

REFERENCES

Alvera-Azcárate, A., Barth, A., Rixen, M. and Beckers, J.M. (2005) Reconstruction of incomplete oceanographic data sets using empirical orthogonal functions: application to the Adriatic Sea surface temperature. *Ocean Modelling*, 9(4), 325–346. Available at: https://doi.org/10.1016/j.ocemod.2004.08.001

Alvera-Azcárate, A., Barth, A., Parard, G. and Beckers, J.-M. (2016) Analysis of SMOS sea surface salinity data using DINEOF. *Remote Sensing of Environment*, 180, 137–145. Available at: https://doi.org/10.1016/j.rse.2016.02.044

Beckers, J.M. and Rixen, M. (2003) EOF calculations and data filling from incomplete oceanographic datasets. *Journal of Atmospheric and Oceanic Technology*, 20(12), 1839–1856. Available at: https://doi.org/10.1175/1520-0426(2003)020<1839:ECADFF>2.0.CO;2

Cowtan, K. and Way, R.G. (2014) Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends. *Quarterly Journal of the Royal Meteorological Society*, 140(683), 1935–1944. Available at: https://doi.org/10.1002/qj.2297

Harris, I., Jones, P.D., Osborn, T.J. and Lister, D.H. (2014) Updated high-resolution grids of monthly climatic observations – the CRU TS3.10 dataset. *International Journal of Climatology*, 34(3), 623–642. Available at: https://doi.org/10.1002/joc.3711

Huang, J., Zhang, X., Zhang, Q., Lin, Y., Hao, M., Luo, Y. *et al.* (2017) Recently amplified arctic warming has contributed to a continual global warming trend . *Nature Climate Change*, 7(12), 875–879. Available at: https://doi.org/10.1038/s41558-017-0009-5

Karl, T. R., Arguez, A., Huang, B., Lawrimore, J. H., McMahon, J. R., Menne, M. J. *et al.* (2015) Possible artifacts of data biases in the recent global surface warming hiatus. *Science*, 348(6242), 1469–1472. Available at: https://doi.org/10.1126/science.aaa5632

Lawrimore, J. (2018) *Climate algorithm theoretical basis document (CATBD), Global Historical Climatology Network-Monthly (GHCN-M) Mean Temperature (Version 4)*. Natl. Oceanic and Atmos. Admin., U.S. Dep. of Commerc. Available at: https://www1.ncdc.noaa.gov/pub/data/ghcn/v4/documentation/CDRP-ATBD-0859%20Rev%201%20GHCN-M%20Mean%20Temperature-v4.pdf [Accessed 08 September 2019].

Lawrimore, J. H., Menne, M. J., Gleason, B. E., Williams, C. N., Wuertz, D. B., Vose, R. S. *et al.* (2011) An overview of the global historical climatology network monthly mean temperature data set, version 3. *Journal of Geophysical Research: Atmospheres*, 116(D19). Available at: https://doi.org/10.1029/2011JD016187

Macias-Fauria, M., Seddon, A. W. R., Benz, D., Long, P. R. and Willis, K. (2014) Spatiotemporal patterns of warming. *Nature Climate Change*, 4(10), 845–846. Available at: https://doi.org/10.1038/nclimate2372

Massonnet, F., Bellprat, O., Guemas, V. and Doblas-Reyes, F.J. (2016) Using climate models to estimate the quality of global observational data sets. *Science*, 354(6311), 452–455. Available at: https://doi.org/10.1126/science.aaf6369

Menne, M.J., Williams, C.N., Gleason, B.E., Jared Rennie, J. and Lawrimore, J.H. (2018) The global historical climatology network monthly temperature dataset, version 4. *Journal of Climate*, 31(24), 9835–9854. Available at: https://doi.org/10.1175/JCLI-D-18-0094.1

Muller, R.A., Rohde, R., Jacobsen, R., Muller, E. and Wickham, C. (2013) A new estimate of the average earth surface land temperature spanning 1753 to 2011. *Geoinformatics and Geostatistics: An Overview*, 1, 1–7. Available at: https://doi.org/10.4172/2327-4581.1000101

Osborn, T. J. and Jones, P. (2014) The CRUTEM4 land-surface air temperature data set: Construction, previous versions and dissemination via Google Earth. *Earth System Science Data*, 6(1), 61–68. Available at: https://doi.org/10.5194/essd-6-61-2014

Peterson, T. C. and Vose, R. S. (1997) An overview of the global historical climatology network temperature database. *Bulletin of the American Meteorological Society*, 78(12), 2837–2849. Available at: https://doi.org/10.1175/1520-0477(1997)078<2837:AOOTGH>2.0.CO;2

Stocker, T. F., Qin, D., Plattner, G. K., Tignor, M. M. B., Allen, S. K., Boschung, J. *et al.* (2013) *Climate Change 2013: The Physical Science Basis*. Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge Univ. Press.

Wang, J., Chengdong, X., Maogui, H., Li, Q., Yan, Z. and Jones, P. (2017a) Global land surface air temperature dynamics since 1880. *International Journal of Climatology*, 38, e466–e474. Available at: https://doi.org/10.1002/joc.5384

Wang, K., Zhang, T., Zhang, X., Clow, G. D., Jafarov, E. E., Overeem, I. *et al.* (2017b) Continuously amplified warming in the Alaskan Arctic: Implications for estimating global warming hiatus. *Geophysical Research Letters*, 44(17), 9029–9038. Available at: https://doi.org/10.1002/2017GL074232

Way, R. G., Oliva, F. and Viau, A. E. (2017) Underestimated warming of northern Canada in the Berkeley Earth temperature product. *International Journal of Climatology*, 37(4), 1746–1757. Available at: https://doi.org/10.1002/joc.4808

How to cite this article: Wang K, Clow GD. Reconstructed global monthly land air temperature dataset (1880–2017). *Geosci Data J*. 2019;00:1–9. https://doi.org/10.1002/gdj3.84