Check for updates

ATMOSPHERIC SCIENCE

Reconstructing historical climate fields with deep learning

Nils Bochow[1,2,3]*, Anna Poltronieri[1], Martin Rypdal[1], Niklas Boers[3,4,5]

Copyright © 2025 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. Distributed under a Creative Commons Attribution License 4.0 (CC BY).

Historical records of climate fields are often sparse because of missing measurements, especially before the introduction of large-scale satellite missions. Several statistical and model-based methods have been introduced to fill gaps and reconstruct historical records. Here, we use a recently introduced deep learning approach based on Fourier convolutions, trained on numerical climate model output, to reconstruct historical climate fields. Using this approach, we are able to realistically reconstruct large and irregular areas of missing data and to reproduce known historical events, such as strong El Niño or La Niña events, with very little given information. Our method outperforms the widely used statistical kriging method, as well as other recent machine learning approaches. The model generalizes to higher resolutions than the ones it was trained on and can be used on a variety of climate fields. Moreover, it allows inpainting of masks never seen before during the model training.

INTRODUCTION

Observational climate data are typically sparse before systematic observations such as buoys, ship measurements, or satellite measurements were introduced. In general, the further back in time we go, the fewer observations are available (*1*). Temperature and precipitation records are the best-observed climate fields in the recent past and reach back until the 19th century, but measurements are still sparse and rely heavily on interpolation, especially for earlier parts of the records (*2*, *3*). Even more severely, for many important climate variables, such as sea-ice thickness or vegetation indices, no measurements exist at all before the introduction of large-scale satellite missions. The corresponding time series often span a few decades or even only years (*4*, *5*). The low spatial and temporal resolution introduces large uncertainties and limits our understanding of important climatic processes (*1*, *2*, *6*).

Several approaches and methods to produce historical climate fields based on the available observations have been developed in the past. One approach is to run state-of-the-art weather models with observations and past weather forecasts to produce reanalysis products that provide a complete picture of the past weather and climate for the past decades (*7*–*10*). While reanalyses are successful in providing spatiotemporally continuous and consistent data, they often struggle with specific regions and variables and inherit biases the used numerical models suffer from (*9*, *11*).

An alternative approach is to use statistical methods to reconstruct missing information. In this regard, kriging or Gaussian process regression is widely used in the geosciences (*12*–*14*). However, statistical methods typically do not include knowledge of the temporal and spatial patterns of the underlying climatic fields and therefore fail to reconstruct these patterns, especially for large missing areas.

In recent years, machine learning (ML) has become widely used in geoscience and climate science, with the promise of better performance than statistical methods while still providing easy usability and, to some extent, knowledge of the underlying physical processes (*15*, *16*). The applications of ML in climate science are vast and range from classical time series forecasting (*17*–*19*), downscaling, and postprocessing of numerical models (*20*, *21*) to time series reconstruction (*15*, *22*). Furthermore, there is a substantial ongoing effort to combine traditional numerical Earth system models with ML methods to leverage the advantages of both approaches (*16*, *23*–*30*).

In this study, we consider the reconstruction of spatial climate fields as an image inpainting problem. Inpainting images based on given information is a classical problem in computer vision, and many approaches have been proposed in recent years (*31*–*33*). We apply the recently introduced state-of-the-art deep learning approach resolution-robust large mask inpainting with Fourier convolutions (LaMa) (*34*) to reconstruct different climate fields with a focus on surface temperature records. We train our model on numerical climate model output from the Coupled Model Intercomparison Project to reconstruct the missing measurements in observational data. Our method is able to reconstruct climate fields with very sparse information and highly irregular missing data. We show that our approach outperforms kriging and other ML methods. Moreover, it is able to inpaint different datasets than the ones it was trained on and can be used on a variety of structurally different climatic fields at varying resolutions.

The surface temperature is one of the most important climate variables as a direct measure of climate change. Global instrumental temperature records reach back to the mid-19th century (*2*) with local observations reaching back as far as the mid-17th century (*35*). However, on average, less than 30% of Earth's surface before the year 1900 AD has measurements in the state-of-the-art observational dataset HadCRUT4 (Fig. 1A). This is similar for other widely used long-term temperature data. Therefore, surface temperature records serve as perfect proof-of-concept applications for the image inpainting task in climate science.

RESULTS

To inpaint the temperature records, we train our model on the historical surface temperatures from the Coupled Model Intercomparison Project 5 (CMIP5) ensemble (1850 to 2012 AD); see Methods.

[1]Department of Mathematics and Statistics, Faculty of Science and Technology, UiT The Arctic University of Norway, Tromsø, Norway. [2]Physics of Ice, Climate and Earth, Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark. [3]Potsdam Institute for Climate Impact Research, Potsdam, Germany. [4]Earth System Modelling, School of Engineering and Design, Technical University of Munich, Munich, Germany. [5]Department of Mathematics and Global Systems Institute, University of Exeter, Exeter, UK.
*Corresponding author. Email: nils.bochow@uit.no

Downloaded from https://www.science.org on August 14, 2025

Fig. 1. Spatial map and temporal time series of missing observations in HadCRUT4. (**A**) Time series of the missing value ratio on the grid cell level in HadCRUT4 for the whole Earth. There is a steady increase in the observational temperature coverage with some exceptions such as the two world wars. (**B**) Spatial missing ratio in HadCRUT4 over the whole time span from 1850 to 2022. The polar regions show the lowest coverage of temperature records.

We follow a previously introduced mask generation approach (*15*) and mask the training data with the missing masks derived from the observational gaps in the temperature dataset HadCRUT4 (*2*) during training.

First, we evaluate the model on the same held-out CMIP5 member as in (*15*) to directly compare with their inpainting approach, which is based on partial convolutions (PConv) and trained on the same CMIP5 training set. In a second step, we evaluate the trained model on each HadCRUT4 mask for 2251 randomly held-out months of the CMIP5 ensemble. This gives a total of 4,641,750 combinations of images and masks. As a baseline comparison, we compare our results with statistical kriging. Subsequently, we reconstruct the HadCRUT4 data and show examples of other applications.

Comparison with related work

In order for our method to be directly comparable to the previously introduced PConv method (*15*), we evaluate LaMa on the same held-out CMIP5 ensemble member. We mask the held-out 145th CMIP5 member with the corresponding HadCRUT4 masks for each month over the time span of 1870 to 2005 AD to have the same temporal range as Kadow *et al.* (*15*) (Fig. 2, A and B). We reconstruct the held-out CMIP5 member over the whole time span with LaMa and kriging and compare it with the PConv approach (Fig. 2). We use the spatial root-mean-square error (RMSE) and site-wise RMSE as main evaluation metrics, where the main difference between the metrics is the order of averaging in time and space (see Methods). We refer to the square root of the spatially weighted average of the squared differences between the ground truth and the inpainted image as the spatial RMSE. In addition, we define the site-wise RMSE as the RMSE at each site (i.e., grid cell), averaged over the time dimension. The mean site-wise RMSE is the spatially weighted average of the site-wise RMSE in all grid cells. In each time step, we exclude the grid cells that have known values for the calculations of the RMSE.

Our model is able to realistically reconstruct the spatial patterns and amplitude of the surface temperature. An example spatial reconstruction for February 1870 for all methods, i.e. LaMa, kriging, and PConv (*15*), is shown in Fig. 2. There is high agreement between the reconstructed temperature fields and the ground truth. While the tropical and subtropical regions show strong similarities between the ground truth and reconstruction, the polar regions show the strongest deviation from the ground truth for all methods (Fig. 3 and Table 1).

All ML methods show an improvement compared to kriging in terms of the mean site-wise and mean spatial RMSE (Fig. 3 and Table 1). LaMa outperforms all other methods with a 15% lower mean site-wise RMSE than the reference PConv method (Fig. 3F and Table 1). LaMa shows the largest improvement in the Northern Hemisphere, especially in North America and Asia, but also in the subtropical Southern Hemisphere and in West Antarctica compared to kriging (green regions in Fig. 3D). Note that eastern Antarctica is in the west in the geographic projection used in these maps here. LaMa and the PConv method show a lower site-wise RMSE in 79 and 71% of the grid cells compared to kriging, respectively. For all methods, the temporally averaged site-wise RMSE is the largest in the Arctic and Antarctic regions. The PConv approach shows a slightly worse performance in the northern and southern tropical Pacific than kriging in terms of the site-wise RMSE (Fig. 3E), while LaMa shows better performance in this region. Both ML methods show considerably worse performance in East Antarctica and slightly worse performance in the Indian Ocean than kriging. The weaker performance in the aforementioned regions is likely a result of the high-temperature variation in these regions. Kriging, by definition, produces a smooth temperature field that might be closer to the ground truth for highly temporally variable regions. Especially for Antarctica, the reconstructed temperature fields via kriging are very homogeneous because of the distance of known measurements, while the ML approaches inpaint a highly variable Antarctica as learned from the CMIP5 data (Fig. 2). This suggests that the available temperature information fed into the networks is not sufficient to successfully infer the temperature patterns in Antarctica.

In 78% of the grid cells, LaMa shows a lower site-wise RMSE than the reconstruction based on PConv (Fig. 3F). LaMa is also able to reconstruct the global mean temperature (GMT) time series reasonably and closely follows the ground truth (fig. S1). LaMa shows comparable performance to the PConv method (*15*) with slightly worse RMSE of the yearly GMT but lower spatial RMSE and higher correlation between the yearly GMT time series than PConv and kriging (fig. S1). The lower spatial RMSE of LaMa's reconstruction implies that LaMa is generally better in reconstructing the spatial temperature patterns. On the other hand, the slightly greater RMSE of the GMT time series implies that LaMa shows slightly worse performance in reconstructing the temporal variability on a global scale.

Downloaded from https://www.science.org on August 14, 2025

Fig. 2. Example reconstructed CMIP5 temperatures for all methods. (**A**) Masked ground truth for February 1870 derived from a held-out CMIP5 member, masked with the corresponding HadCRUT4 mask for this date. White areas denote masked regions. (**B**) Ground truth without masking. (**C**) Infilled temperatures via LaMa. The spatial patterns are very similar to the ground truth. (**D** and **E**) Same as (C) but for PConv (*15*) and kriging, respectively. While the spatial patterns reconstructed by the deep learning methods are very similar to the ground truth, some regions show the opposite trend in the temperature, e.g., northern South America.

Evaluation on CMIP5

In addition to the evaluation on the single held-out CMIP5 ensemble member, we evaluate the model on 2251 held-out months of the entire CMIP5 ensemble against all 2064 HadCRUT4 masks. We calculate the spatial and site-wise ensemble RMSE of the infilled evaluation temperature data and compare it with kriging (Fig. 4). Here, ensemble mean refers to the mean over all months of the 2251 CMIP5 ensemble members, and temporal mean refers to a mean over the masks, which corresponds to the temporal dimension of the HadCRUT4 records.

LaMa has a temporal ensemble mean of the spatial RMSE of 1.02 K, and the reconstructed images via kriging have a mean spatial RMSE of 1.23 K (Fig. 4D). The ensemble mean of the site-wise RMSE shows a similar behavior (Fig. 4, A to C). LaMa shows a more than 20% smaller mean site-wise RMSE than kriging (Fig. 4, A to C) but shows a higher spatial RMSE than kriging for very large masks (>80%). Otherwise, LaMa consistently outperforms kriging for all masks.

The spatial RMSE of the infilled images via kriging and LaMa depends on the ratio of missing values and shows a decrease around mask numbers 1200 to 1300, which correspond to years 1950 to 1960 AD (Fig. 4D). This is due to the introduction of large-scale observational instruments and, therefore, greater coverage of temperature observations. Both methods show a seasonal dependency of the spatial RMSE (Fig. 4D) because of a difference in the seasonal global temperature coverage, mostly in the polar regions. The summer months in the polar regions have a greater coverage than the winter months. This leads to a higher spatial RMSE for the austral winter than in the boreal winter when observations in Antarctica are sparse, because Antarctica is the region with the largest uncertainty.

The spatial patterns of the site-wise RMSE are similar to the site-wise RMSE of the single-member test set, with the maximum RMSE close to the poles and a minimum in the tropical and subtropical regions (Fig. 4, A and B). In particular, the RMSE in the Northern Hemisphere is notably smaller for LaMa than for kriging.

It is important to note that LaMa does generally not have any information about the temporal order of the training set. That is, even when randomly permuting the fields of the training set, the results would remain the same. However, when randomly selecting months for evaluation from the training and validation sets, there might be hidden temporal information in the previous or subsequent months because of the temporal correlation of individual months. To rule out this possibility, we first show that the local

Downloaded from https://www.science.org on August 14, 2025

Fig. 3. Average site-wise RMSE and comparison between methods for a single held-out CMIP5 member. (**A**) Temporally averaged RMSE at each site for an inpainted CMIP5 held-out member using LaMa fixed. The white areas denote the regions with available temperature records for the whole period 1870 to 2005 AD. The spatially weighted average of the site-wise RMSE is 0.64°C. The polar regions, especially Antarctica, show the greatest RMSE, while the tropical and subtropical regions show the smallest RMSE. (**B** and **C**) Same as (A) but for PConv (*15*) and kriging. Both ML methods show a smaller mean RMSE than kriging. In particular, in the Northern Hemisphere, the RMSE is smaller. (**D**) Difference between temporally averaged RMSEs at each site for LaMa and kriging. Green areas denote regions where the RMSE of LaMa is smaller than that of the baseline kriging method. Purple areas denote greater RMSE than for kriging. LaMa fixed shows a lower RMSE than kriging in 79% of the grid cells. (**E**) Same as (D) but for PConv (*15*). PConv shows in 71% of the grid cells a smaller RMSE than kriging. (**F**) Comparison between LaMa and PConv. Green areas denote regions where the RMSE of LaMa is smaller than the RMSE of the PConv method. LaMa shows a lower RMSE than PConv in 78% of all grid cells.

Table 1. Comparison between all methods in terms of the RMSE. Spatially weighted average of the site-wise RMSE and average spatial RMSE for both LaMa methods, PConv, and kriging for the single CMIP5 ensemble member. The improvement compared to kriging is denoted in the parentheses. LaMa shows considerable improvement in comparison with the other methods. LaMa random outperforms kriging and has a similar performance to PConv.

Model	Spatially averaged site-wise RMSE (°C)	Temporally averaged spatial RMSE (°C)
LaMa	0.64 (21.0%)	0.99 (16.8%)
LaMa random	0.74 (8.6%)	1.10 (7.6%)
PConv	0.74 (8.6%)	1.08 (9.2%)
Kriging	0.81	1.19

Downloaded from https://www.science.org on August 14, 2025

Fig. 4. Error statistics on the held-out CMIP5 members for each HadCRUT4 mask. (**A**) Average site-wise RMSE for kriging on the randomly held-out 2251 months from the CMIP5 ensemble. We calculate the site-wise RMSE for every combination of HadCRUT4 masks from 1850 to 2021 AD and month. The spatially weighted average of the site-wise RMSE is 0.85°C. The polar regions show the largest RMSE, while the tropics and subtropical oceans show the lowest RMSE. White grid cells denote regions where temperature observations are available for the whole time span. (**B**) Same as (A) but for LaMa. The site-wise RMSE is lower in most grid cells than for kriging. In particular, in the Northern Hemisphere, there is a strong improvement compared to kriging. LaMa outperforms kriging in terms of the spatially averaged site-wise RMSE. (**C**) Difference between temporally averaged RMSE at each site for LaMa and kriging. Green areas denote regions where the RMSE of LaMa is smaller than that of the baseline kriging method. Purple areas denote greater RMSE than for kriging. (**D**) Spatial RMSE for both methods and all HadCRUT4 masks, which are ordered in time; note that, in general, the size of the masks in terms of the number of missing data declines over time. LaMa outperforms kriging almost consistently for all masks. LaMa shows a higher RMSE than kriging for very large masked areas (≥ 80%) but outperforms kriging otherwise. Especially for small masks, the average spatial RMSE is substantially lower for LaMa than for kriging.

similarity of consecutive months of the same CMIP5 model is generally low. We demonstrate this using several similarity metrics, including the correlation between two consecutive months (fig. S2, B to E). We also show that when filling in the missing grid cell values of a given month by the previous, following, or same month of the following year, the RMSE is on average substantially larger than when infilling via LaMa (fig. S2A). However, this test of local similarity does not consider the known correlation on long spatial scales between adjacent months. To show that our test set does not contain any temporal information and is thus strictly independent from the training set, we additionally train our model on a different train-validation-test split, where we hold out 17 random years from the training set instead of every 9th month (compare Methods). We repeat our analysis on the 145th held-out CMIP5 member as well as on the alternative training set. In both cases, the error metrics are very similar (fig. S3). The model trained on the alternative training-validation-test split partially even outperforms the model shown in the main text. These tests together show that our approach for reconstructing the field of a given month does not rely on simply using similar months from the training data.

HadCRUT4

After demonstrating that LaMa is able to reconstruct spatial and temporal patterns of the CMIP5 temperatures, we apply the trained network to the HadCRUT4 observational data. We show that our method is able to accurately reconstruct the spatial and temporal patterns of the HadCRUT4 dataset. As there are no control data for the reconstructed temperature observations, we first compare with reconstructed HadCRUT4 temperatures via kriging (*6*) and also analyze spatial patterns in the reconstruction, focusing on known historical events.

We take a spatially weighted mean to obtain the temporal time series of the GMTs for all methods. The reconstructed yearly GMTs show a strong correlation (Pearson correlation coefficient $r > 0.99$) and the same trend as the masked HadCRUT4 temperature time series for all methods (Fig. 5). LaMa shows a lower GMT for the mid-19th century relative to the masked mean and the temperature reconstructed via kriging or PConv. The main contributor is a slightly colder Antarctica in LaMa's temperature reconstruction compared to the other methods (fig. S4, B and D). There is no a priori reason to believe that the global mean time series reconstructed by LaMa is unreasonable. However, given another validation on the HadCRUT5 dataset that we carry out (figs. S5 and S6) and given the underestimation of temperatures in Antarctica on the validation CMIP5 member (Fig. 3), it is likely that the GMT in the mid-19th century, as reconstructed by LaMa, is underestimated.

Because of the nature of the data, there is no ground truth that we can compare our reconstructions to. Hence, we compare the reconstructed temperature fields to well-known historical events such as strong El Niño episodes. The El Niño in the year 1877/1878 AD is known to have been extraordinarily strong and is linked to famines around the globe (*36*, *37*). However, historical temperature records for these years are sparse. LaMa is able to reproduce the warm Pacific Ocean on the basis of the sparse records, whereas kriging is not able to reconstruct the spatial patterns of the temperature anomaly (fig. S7). The reconstruction based on PConv (fig. S7E) also shows a warm

Downloaded from https://www.science.org on August 14, 2025

Fig. 5. Reconstructed HadCRUT4 GMT time series. (**A**) Monthly HadCRUT4 time series from 1850 to 2022 with anomalies relative to the years 1960 to 1990. The reconstructions from LaMa and kriging as well as the mean of the masked HadCRUT4 records are shown. The dashed black curve is the spatially averaged GMT derived from the incomplete HadCRUT4 observations. (**B**) Same as (A) but for the yearly averaged GMT. In addition, we show the reconstruction based on the PConv method. The Pearson correlation coefficients between the masked time series, i.e., yearly GMT, are $r_{\text{LaMa}} = 0.99$ for LaMa, $r_{\text{kriging}} = 0.99$ for kriging, and $r_{\text{PConv}} = 0.99$ for PConv. (**C** and **D**) Same as (A) and (B) but for the difference between the reconstructions and the masked HadCRUT4 time series.

Pacific but with a smaller spatial extent. We also show an opposite example of a strong La Niña year with a cold Pacific for February 1917 AD (fig. S8) (*38*, *39*). Kriging does not reconstruct the same spatial extent of the cold Pacific compared to LaMa, which shows a strongly anomalous cold Pacific (fig. S8). Statistical interpolation tends to inpaint large missing areas with values close to zero (e.g., fig. S7). However, even for these anomalous historical events, it is still hard to verify the validity of the reconstructed temperature anomalies. We compare our example reconstructions visually with the 20CRv3.SI reanalysis (figs. S7F and S8F) (*8*, *40*). The ML reconstructions for the two example months show a strong similarity to the reanalysis, while the kriging reconstruction does not show the same spatial patterns. This suggests that LaMa is able to capture the dynamics underlying the global temperature fields. However, the temperature anomalies in Antarctica, in particular, show different spatial patterns across the different reconstructions and temperature products. It has to be noted that the 20CRv3.SI reanalysis should not be seen as a ground truth but rather as an independent temperature dataset to compare with. While the 20CRv3.SI reanalysis has been shown to perform reasonably even in the 19th century, there are known biases in the 20CRv3.SI product that should be treated with caution (*11*).

We also evaluate our reconstruction method on the more recent HadCRUT5 (*41*) dataset, which has further improved coverage in recent years compared to earlier versions. We choose a late month of the HadCRUT5 data that has no gaps, here January 2021, apply all HadCRUT4-derived masks, and subsequently reconstruct the artificially masked temperature field (fig. S5). In general, LaMa is able to reconstruct the temperature fields well compared to the ground truth. Even for the mask with the largest RMSE, we find good agreement of the spatial temperature patterns with the ground truth (fig. S5B). LaMa is able to reconstruct the spatial patterns in most parts of the Pacific, America, and Eurasia. However, there is a notable difference between the ground truth and reconstructions in the polar regions, with considerably underestimated Arctic temperatures in the reconstructions. For the mask that leads to the lowest RMSE, most masked grid cells are reconstructed very well, including the polar regions (fig. S5C). On average, the site-wise RMSE shows the largest differences in the polar regions, northern Asia, and northern North America (fig. S5D). We calculate the difference between the ground truth GMT and the reconstructed GMT for each mask (fig. S6). The reconstructed GMT shows the largest difference for masks with a coverage of more than 70%, with a maximum difference of −0.27°C. For the early masks, the reconstructed GMT is generally lower than the ground truth GMT. For later masks, the reconstructed GMT is close to the ground truth GMT. The average difference between the ground truth GMT and the reconstructed GMT over all masks is close to 0°C (fig. S6).

Downloaded from https://www.science.org on August 14, 2025

Beyond HadCRUT

LaMa is able to generalize to higher resolutions than the ones it is trained on and is not restricted to temperature fields. We inpaint the Berkeley Earth Surface Temperatures (BEST; 90 by 90 pixels) (*42*) using the LaMa model trained on the CMIP5 images (72 by 72 pixels; fig. S9). We also show an example reconstruction of sea ice concentration to show the application to a structurally different climatic field (fig. S11).

We do not modify the trained model before evaluating on the BEST temperature records. We transform the BEST temperature records to images (90 by 90 pixels) with the same procedure as before, which corresponds to a 156.25% higher resolution than the images we trained the model on. LaMa, trained solely on the HadCRUT4 masks, shows visible artifacts in the inpainted spatial fields, especially at the edges of the gaps (fig. S10), and is therefore not suitable for reconstructions on unseen masks. This problem can be easily facilitated by using a different mask generation algorithm during training. By generating random masks during training on the fly, using a previously proposed mask generation algorithm (*34*), LaMa generalizes to different masks than the ones seen during training. In the following, we call LaMa with randomly generated masks during the training LaMa random. We show that LaMa random is able to inpaint the missing areas in the BEST record without any strong artifacts (fig. S9). LaMa random is therefore better suited for generalization tasks when the final mask shapes for inference are not known during the training.

It should be noted that the very low resolution we use for the training images limits the application to higher resolutions. However, it has been shown that LaMa can generalize to resolutions up to four times higher than those it is trained on (*34*). Hence, a substantially stronger upsampling than from CMIP5 (72 by 72 pixels) to BEST (90 by 90 pixels) should be possible.

To show the applicability to a different climate field, we train LaMa on the daily sea ice concentration from 1979 to 2022, taken from the ERA5 reanalysis (*9*, *43*) with a resolution of 180 by 1440 pixels (Northern Hemisphere). This gives a total of 15,450 monthly fields, where we hold out 1054 random months for evaluation and 1043 random months for validation. Even for this relatively small training sample, LaMa is able to reconstruct the spatial extent and concentration of the sea ice reasonably well (fig. S11). LaMa learns the continent distribution during training and correctly predicts the extent of the sea ice, given very little information from the unmasked areas (fig. S11, A and D). LaMa correctly reconstructs the seasonality of the sea ice concentration with a maximum in winter (fig. S11a) and a minimum in summer (fig. S11D). The largest deviations from the ground truth are generally at the edges of the sea ice, while the central Arctic shows the lowest error (fig. S11C). This example case of the sea ice concentration shows that LaMa can be applied to a variety of structurally different climate fields.

DISCUSSION

Reconstruction of historical observations is an active and important research field in climate science with vast implications for the present climate, short- and long-term future projections, and climate change attribution. Previously used methods often struggle with large irregular gaps in climate fields or with resolving spatial patterns. We show that LaMa is able to realistically reconstruct global temperature records across different datasets and resolutions. LaMa clearly outperforms the widely used kriging with a 21.0% smaller spatially averaged site-wise RMSE (Table 1). Furthermore, our method outperforms a previously proposed deep learning method based on PConv (*15*). In terms of the spatially averaged site-wise RMSE, LaMa outperforms PConv by 13.5% (Table 1), and 78% of the grid cells show a lower site-wise RMSE on the test set (Fig. 3F). We note that approaches using principal components analysis (PCA) have been successfully used to reconstruct temperature records (*44*, *45*). In particular, PCA has been shown to be able to reconstruct the El Niño-Southern Oscillation remarkably well (*46*). While, in general, PCA shows a better performance than kriging in reconstructing spatial temperature patterns, PCA shows overall worse performance than PConv (*15*). More specifically, a thorough comparison of the PConv method with a PCA-based reconstruction shows that PCA introduces additional biases when applying cross-validation and is generally outperformed by PConv (*15*). Because, as we have shown, our approach generally outperforms the PConv method, it is thus clear that it also gives better reconstructions than PCA.

A substantial advantage of LaMa over PConv and other statistical methods such as kriging is the global context provided by fast Fourier convolutions (FFCs), in contrast to the local context of PConv. For climate fields with global teleconnections, such as temperature fields, the FFC blocks in LaMa allow for a global context, promising more accurate spatial reconstructions. We train our model on rather low-resolution images (72 pixels by 72 pixels), which makes it difficult to resolve global teleconnections. Nonetheless, our model is able to realistically reconstruct spatial temperature patterns on a global level. Training on high-resolution images is limited by the available graphics processing unit (GPU) infrastructure. However, because of the ability of LaMa to train on lower-resolution images than the ones it is evaluated on, this problem can be mitigated. While the training of the model can take several wall time days, depending on the training size, the evaluation is done in the order of minutes. It is hence much faster than gap filling with dynamic models and still faster than statistical methods.

While LaMa can generalize to data different from its training set, the masks derived from BEST appear to be too dissimilar from the HadCRUT4 masks, which LaMa was trained on, to yield sensible reconstructions. We find that randomly generated masks during training ensure applicability on masks never seen during training, but on the HadCRUT4-derived masks, LaMa almost consistently outperforms LaMa random in terms of the error metrics (fig. S12). We note, however, that an improved random mask generation during training could potentially further improve LaMa random. Furthermore, we show that LaMa can be applied to a variety of structurally very different climate fields (fig. S11).

The global time series reconstructed by LaMa shows slightly lower GMT until the year 1880 AD. With LaMa, we find a GMT of 1.09°C above the preindustrial level (1850 to 1900) for the period 2010 to 2020, while the best estimate based on the masked HadCRUT4 dataset gives a warming level of 1.00°C for the same period. This is mostly due to a colder Southern Hemisphere reconstructed by LaMa, especially in the Antarctic region. Because of the sparse observations at the poles, it is difficult to validate the plausibility of the reconstructed temperatures in these areas. While the colder Antarctica reconstructed by LaMa is a priori not implausible, LaMa shows worse performance in larger parts of Antarctica than PConv and kriging (Fig. 3, D and F). In addition, the evaluation on the HadCRUT5 month shows that LaMa underestimates the GMT in the first 300 to 400 masks, suggesting that our approach generally

Downloaded from https://www.science.org on August 14, 2025

underestimates the GMT in the beginning of the reconstruction. We attribute a possible underestimation of Antarctic temperatures to two reasons. First, the variability of the surface temperature in Antarctica across the CMIP5 ensemble is large (*47*). For a similar global surface temperature distribution, the temperatures in Antarctica might differ vastly between the single ensemble members. This makes it hard for the ML model to learn useful spatial connections that lead to reasonable Antarctic temperature predictions. Second, the inpainting problem turns into an outpainting problem for the polar regions, which is inherently harder. There are almost no known measurements for any time step in close proximity to Antarctica. While LaMa is able to extrapolate the polar regions reasonably well, the performance is worse than in the nonpolar regions (Fig. 3). However, the left and right edges of the images, corresponding to the prime meridian, do not necessarily show a higher RMSE than the other regions (Fig. 3), which suggests that this is not the main reason.

For a masking ratio of more than 80%, LaMa shows a higher mean site-wise RMSE on the testing set than the other methods (Fig. 4D). This is due to artifacts with unusually high temperatures in western Antarctica (bottom right corner of the image) for some of the CMIP5 ensemble members. Similarly to Kadow *et al.* (*15*), we attribute this to effects at the edges of the images, as mentioned above. LaMa random does not exhibit these artifacts, which suggests that a more sophisticated mask generation during training could resolve these issues. We do not observe any artifacts in the other reconstructed datasets.

Convolutional neural networks (CNNs) are generally not able to capture the spherical geometry of Earth well, which can lead to the aforementioned artifacts at the edges and worse performance in the polar regions. Graph neural networks or spherical convolutions could facilitate this problem (*15*, *48*, *49*). First attempts of applying CNNs with spherical harmonics as basis functions to weather forecasting show promising results (*50*). However, spherical CNNs remain to be computationally much more expensive than regular convolutional networks (*50*). Alternatively, one could train several different models for different, smaller regions of the Earth surface using corresponding geographical projections and merge the resulting output. This would ameliorate the problem of the model not being aware of the changing grid cell size and, hence, correlation toward the poles in a regular projection. However, besides the increased computational demand, this is only feasible if the training data are available in appropriate projections or if the raw observational/model data are available so that the different projections can be carried out. Otherwise, interpolation artifacts from reprojecting would lead to additional errors. The rising popularity of generative models makes it a promising alternative to CNN-based models for the reconstruction of climate fields. In particular, recently introduced diffusion models (*51*, *52*) show promising performance on image inpainting tasks. Furthermore, by using video-inpainting techniques (*53*) rather than image inpainting, the temporal dimension of the data could directly be taken into account. However, little work has been done so far in that direction and the physical plausibility of such models remains uncertain.

Reconstructions via deep learning can aid in understanding past and present changes in the Earth system. By learning the spatiotemporal patterns of the underlying climate fields, LaMa is able to realistically reconstruct a variety of observables with varying resolutions. Our easy-to-use deep learning model clearly outperforms previous methods and therefore serves as an alternative to conventional methods used in the geosciences.

METHODS

Resolution-robust large mask inpainting with Fourier convolutions (LaMa)

We use the recently introduced LaMa model (*34*) that builds on FFCs for reconstructing missing image regions. LaMa is a feed-forward ResNet-like inpainting network with a multicomponent loss. LaMa has been shown to outperform other ML methods such as aggregated contextual transformations for high-resolution image inpainting (AOT GAN) (*54*), image inpainting with learnable feature imputation (*55*), or latent diffusion models (*34*, *56*, *57*) and is able to inpaint large missing areas with a high receptive field.

LaMa is designed to inpaint masked images. Given an image x and a binary mask m, the input is prepared by stacking the masked image with the mask itself, $x' = \text{stack}(x \odot m, m)$, where $\odot$ denotes element-wise multiplication. This results in a four-channel input tensor. The core of LaMa is the FFC block, which uses channel-wise fast Fourier transform. FFC's primary advantage is its ability to capture global context in early network layers, unlike conventional convolutions with limited spatial receptive fields (*58*). An FFC block consists of two interconnected branches:

1) A local branch in the spatial domain using conventional convolutions.

2) A global branch in the spectral domain using a real fast Fourier transform.

The input tensor x is split along feature channels into local (x^l) and global (x^g) components:

1) $x^l \in \mathbb{R}^{H \times W \times (1-\alpha)C}$ captures small-scale information.

2) $x^g \in \mathbb{R}^{H \times W \times \alpha C}$ captures large-scale context.

Here, $H \times W$ represents spatial dimensions, C is the number of channels, and $\alpha \in [0, 1]$ is the hyperparameter determining the channel split ratio between global and local branches. The FFC block's output $y = \{ y^l, y^g \}$ maintains the same dimensions and split ratio as the input. The block's operations can be described as

$$y^l = y^{l \to l} + y^{g \to l} = f_l\left(x^l\right) + f_{g \to l}(x^g) \tag{1}$$

$$y^g = y^{g \to g} + y^{l \to g} = f_g(x^g) + f_{l \to g}\left(x^l\right) \tag{2}$$

where f_l is the local operation via conventional convolutions, f_g is the global spectral transformer, and $f_{g \to l}$ and $f_{l \to g}$ denote the inter-path transitions via conventional convolutions. The spectral transform f_g in the global path ensures global context in early layers. While the global branch can suffice for realistic inpainting, the local branch adds network stability (*59*). FFC's image-wide receptive field allows for global context in early network layers. Conventional convolutions would require substantially more layers to achieve a similar receptive field, increasing computational demands.

The multicomponent loss of LaMa is the second crucial part for achieving realistic inpainting results. The full loss function $\mathscr{L}$ is given by

$$\mathscr{L} = \kappa L_{\text{adv}} + \alpha \mathscr{L}_{\text{HRFPL}} + \beta \mathscr{L}_{\text{DiscPL}} + \gamma R_1 \tag{3}$$

with an adversarial loss L_{adv}, the high receptive field perceptual loss L_{HRFPL}, a discriminator-based perceptual loss or feature-matching loss $\mathscr{L}_{\text{DiscPL}}$, and a regularization term R_1. The adversarial loss L_{adv} ensures that the inpainting network $f_\Theta\left(x'\right)$ generates natural-looking local details. The high receptive field perceptual loss L_{HRFPL} evaluates the similarity between the target and predicted image

Downloaded from https://www.science.org on August 14, 2025

using a pretrained network, here, a pretrained ResNet50 network with dilated convolutions. The discriminator-based perceptual loss increases stability and performance (*34*). The regularization term prevents overfitting and allows better generalization. The weights are given by $\kappa = 10$, $\alpha = 30$, $\beta = 100$, and $\gamma = 0.001$.

LaMa is able to inpaint high-resolution images even if trained on lower-resolution images. We extend and modify the model to allow drawing of pregenerated masks from Hierarchical Data Format version 5 files during training and evaluation, as well as to enable training and evaluation on rectangular images. LaMa outperforms the method based on PConv (*15*) in terms of spatial metrics and shows comparable performance on a temporal mean global scale. For the full description of the network architecture, we refer to the original paper (*34*).

Training procedure, preprocessing, and validation

Training strategy

We train the model on the monthly surface temperature (tas) of 239,616 CMIP5 ensemble members following (*15*). For the training and evaluation, we transform the temperature records into grayscale png images (72 by 72 pixels) with three identical RGB channels. We normalize the images with respect to the maximum and minimum values in the full CMIP5 set such that the maximum temperature corresponds to 255 and the minimum value to 0. Therefore, the maximum resolution of the reconstruction is given by $\frac{|\mathbf{u}_{\max}| + |\mathbf{u}_{\min}|}{255}$ with $\mathbf{u}$ as the climatic field of interest. For our monthly mean temperature reconstruction, this leads to an effective maximum resolution of ~0.19°C. We convert the floats to integers during the transformation of the temperature records to images by truncation. In the following, we take the transformed temperature records as ground truth when we refer to the metrics of our ML model. For simplicity, we only plot the nontransformed HadCRUT4 temperature time series and add the difference between the HadCRUT4 masked temperature records and the transformed HadCRUT4 temperature records when we plot the reconstructed LaMa temperature time series. We hold out 27,702 images for validation and 2250 images for evaluation. Furthermore, we hold out one CMIP5 ensemble member from training for comparison with the approach in (*15*). We use two different mask-generation methods during the training. For the first approach, we train the model on randomly drawn masks derived from the HadCRUT4 (*2*) missing masks. In the following, we name this model LaMa. Our second approach generates random masks following the approach in (*34*) during training, hereafter LaMa random.

To show the independence of our test set from the training set, we retrain our model on an alternative training-validation-test split and repeat the error analysis. We withhold 17 random years from the training set, instead of withholding every ninth month from the training set. By withholding a year from the training set, we mean that the whole year for all 144 CMIP5 ensemble members is removed from the training set. We use 15 of these 17 years for validation (corresponding to 25,670 monthly fields) and 2 years for testing (corresponding to 3456 monthly fields). We train the model for 55 epochs and otherwise follow the same training strategy as mentioned above.

Random mask generation

The random mask generation algorithm works as follows. With a 77% probability, several line segments are drawn. A random starting point on the image is chosen where multiple line segments (we choose a minimum of four and a maximum of five) are drawn, which correspond to the masked area. Each of the line segments has a random angle, length (max. 40 pixels), and width (max. 35 pixels). With a 23% probability, random rectangle boxes with a minimum size of 10 pixels and a maximum size of 70 pixels (minimum of one box and maximum of five boxes) are drawn instead of line segments. Examples of some random masks generated during the training are shown in fig. S13.

Hardware

We train both models on two NVidia Tesla V100 GPUs or alternatively on one NVIDIA H100 Tensor Core GPU with a maximum of 60 epochs with a batch size of 100. We choose the training checkpoint with the lowest error metrics on the evaluation dataset for each model. Therefore, we use the 57th checkpoint for LaMa random and the 60th checkpoint for LaMa fixed. For the full training parameters, we refer to the configuration files in the GitHub repository.

Evaluation

As main evaluation metrics, we use the spatial RMSE and the site-wise RMSE. We refer to the square root of the spatially weighted average of the squared differences between the ground truth and the inpainted image as the spatial RMSE

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} w_i \left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n} w_i}} \tag{4}$$

with the inpainted grid cell y_i and the ground truth $\hat{y}_i$ at the given time steps, the number of grid cells n, and the weights $w_i = \cos(\phi_i)$ that are given by the latitude ϕ_i of the corresponding point y_i, which corresponds to the area of the grid cell for a regular rectangular grid. In addition, we define the site-wise RMSE as the RMSE at each site (i.e., grid cell), averaged over the time dimension. The mean site-wise RMSE is the spatially weighted average of the site-wise RMSE in all grid cells. Hence, the main difference between the two main evaluation metrics that we focus on is the order of averaging in time and space. In each time step, we exclude the grid cells that have known values for the calculations of the RMSE. It is explicitly stated if we show a nonweighted RMSE.

In addition to the described validation procedures, we compare our reconstructions visually with the 20CRv3.SI reanalysis (*8*). 20CRv3.SI is a reanalysis, or four-dimensional weather reconstitution, spanning the years 1836 to 2015. The reanalysis product only assimilates surface observations of synoptic pressure into the National Oceanic and Atmospheric Administration's (NOAA) Global Forecast System and prescribes sea surface temperature and sea ice distribution to estimate other climate variables.

Supplementary Materials

This PDF file includes:

Figs. S1 to S13

References

REFERENCES AND NOTES

1. M. Ben-Yami, V. Skiba, S. Bathiany, N. Boers, Uncertainties in critical slowing down indicators of observation-based fingerprints of the Atlantic Overturning Circulation. *Nat. Commun.* **14**, 8344 (2023).

Downloaded from https://www.science.org on August 14, 2025

2. C. P. Morice, J. J. Kennedy, N. A. Rayner, P. D. Jones, Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. *J. Geophys. Res. Atmos.* **117**, D08101 (2012).
3. I. Harris, T. J. Osborn, P. Jones, D. Lister, Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. *Sci. Data* **7**, 109 (2020).
4. E. Vermote, NOAA CDR Program, NOAA climate data record (CDR) of AVHRR normalized difference vegetation index, version 5 (NOAA National Centers for Environmental Information, 2018); www.ncei.noaa.gov/metadata/geoportal/rest/metadata/item/gov.noaa.ncdc:C01558/html.
5. J. C. Landy, G. J. Dawson, M. Tsamados, M. Bushuk, J. C. Stroeve, S. E. L. Howell, T. Krumpen, D. G. Babb, A. S. Komarov, H. D. B. S. Heorton, H. J. Belter, Y. Aksenov, A year-round satellite sea-ice thickness record from CryoSat-2. *Nature* **609**, 517–522 (2022).
6. K. Cowtan, R. G. Way, Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends. *Q. J. Roy. Meteorol. Soc.* **140**, 1935–1944 (2014).
7. E. Kalnay, M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, A. Leetmaa, R. Reynolds, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K. C. Mo, C. Ropelewski, J. Wang, R. Jenne, D. Joseph, The NCEP/NCAR 40-year reanalysis project. *Bull. Am. Meteorol. Soc.* **77**, 437–471 (1996).
8. G. P. Compo, J. S. Whitaker, P. D. Sardeshmukh, N. Matsui, R. J. Allan, X. Yin, B. E. Gleason, R. S. Vose, G. Rutledge, P. Bessemoulin, S. Brönnimann, M. Brunet, R. I. Crouthamel, A. N. Grant, P. Y. Groisman, P. D. Jones, M. C. Kruk, A. C. Kruger, G. J. Marshall, M. Maugeri, H. Y. Mok, Ø. Nordli, T. F. Ross, R. M. Trigo, X. L. Wang, S. D. Woodruff, S. J. Worley, The twentieth century reanalysis project. *Q. J. Roy. Meteorol. Soc.* **137**, 1–28 (2011).
9. B. Bell, H. Hersbach, A. Simmons, P. Berrisford, P. Dahlgren, A. Horányi, J. Muñoz-Sabater, J. Nicolas, R. Radu, D. Schepers, C. Soci, S. Villaume, J.-R. Bidlot, L. Haimberger, J. Woollen, C. Buontempo, J.-N. Thépaut, The ERA5 global reanalysis: Preliminary extension to 1950. *Q. J. Roy. Meteorol. Soc.* **147**, 4186–4227 (2021).
10. C. Soci, H. Hersbach, A. Simmons, P. Poli, The ERA5 global reanalysis from 1940 to 2022. *Q. J. Roy. Meteorol. Soc.* **150**, 4014–4048 (2024).
11. L. C. Slivinski, G. P. Compo, P. D. Sardeshmukh, J. S. Whitaker, C. McColl, R. J. Allan, P. Brohan, X. Yin, C. A. Smith, L. J. Spencer, R. S. Vose, M. Rohrer, R. P. Conroy, D. C. Schuster, J. J. Kennedy, L. Ashcroft, S. Brönnimann, M. Brunet, D. Camuffo, R. Cornes, T. A. Cram, F. Domínguez-Castro, J. E. Freeman, J. Gergis, E. Hawkins, P. D. Jones, H. Kubota, T. C. Lee, A. M. Lorrey, J. Luterbacher, C. J. Mock, R. K. Przybylak, C. Pudmenzky, V. C. Slonosky, B. Tinz, B. Trewin, X. L. Wang, C. Wilkinson, K. Wood, P. Wyszyński, An evaluation of the performance of the twentieth century reanalysis version 3. *J. Clim.* **34**, 1417–1438 (2021).
12. T. Berezowski, M. Szcześniak, I. Kardel, R. Michałowski, T. Okruszko, A. Mezghani, M. Piniewski, CPLFD-GDPT5: High-resolution gridded daily precipitation and temperature data set for two largest Polish river basins. *Earth Syst. Sci. Data* **8**, 127–139 (2016).
13. A. Sekulić, M. Kilibarda, D. Protić, M. P. Tadić, B. Bajat, Spatio-temporal regression kriging model of mean daily temperature for Croatia. *Theor. Appl. Climatol.* **140**, 101–114 (2020).
14. L. Belkhiri, A. Tiri, L. Mouni, Spatial distribution of the groundwater quality using kriging and Co-kriging interpolations. *Groundw. Sustain. Dev.* **11**, 100473 (2020).
15. C. Kadow, D. M. Hall, U. Ulbrich, Artificial intelligence reconstructs missing climate information. *Nat. Geosci.* **13**, 408–413 (2020).
16. C. Irrgang, N. Boers, M. Sonnewald, E. A. Barnes, C. Kadow, J. Staneva, J. Saynisch-Wagner, Towards neural Earth system modelling by integrating artificial intelligence in Earth system science. *Nat. Mach. Intell.* **3**, 667–674 (2021).
17. T. Mitsui, N. Boers, Seasonal prediction of Indian summer monsoon onset with echo state networks. *Environ. Res. Lett.* **16**, 074024 (2021).
18. B. Lim, S. Zohren, Time-series forecasting with deep learning: A survey. *Philos. Trans. R. Soc. A* **379**, 20200209 (2021).
19. R. Lam, A. Sanchez-Gonzalez, M. Willson, P. Wirnsberger, M. Fortunato, F. Alet, S. Ravuri, T. Ewalds, Z. Eaton-Rosen, W. Hu, A. Merose, S. Hoyer, G. Holland, O. Vinyals, J. Stott, A. Pritzel, S. Mohamed, P. Battaglia, Learning skillful medium-range global weather forecasting. *Science* **382**, eadi2336 (2023).
20. D. Grattarola, P. Vandergheynst, Generalised implicit neural representations. arXiv:2205.15674 [cs, eess] (2022).
21. P. Hess, M. Drüke, S. Petri, F. M. Strnad, N. Boers, Physically constrained generative adversarial networks for improving precipitation fields from Earth system models. *Nat. Mach. Intell.* **4**, 828–839 (2022).
22. Y. Huang, L. Yang, Z. Fu, Reconstructing coupled time series in climate systems using three kinds of machine-learning methods. *Earth Syst. Dyn.* **11**, 835–853 (2020).
23. C. Monteleoni, G. A. Schmidt, S. McQuade, Climate informatics: Accelerating discovering in climate science with machine learning. *Comput. Sci. Eng.* **15**, 32–40 (2013).
24. M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, Deep learning and process understanding for data-driven Earth system science. *Nature* **566**, 195–204 (2019).
25. J. Yuval, P. A. O'Gorman, C. N. Hill, Use of neural networks for stable, accurate and physically consistent parameterization of subgrid atmospheric processes with good performance at reduced precision. *Geophys. Res. Lett.* **48**, e2020GL091363 (2021).
26. Y. Zhu, R. H. Zhang, J. N. Moum, F. Wang, X. Li, D. Li, Physics-informed deep-learning parameterization of ocean vertical mixing improves climate simulations. *Natl. Sci. Rev.* **9**, nwac044 (2022).
27. M. Gelbrecht, A. White, S. Bathiany, N. Boers, Differentiable programming for Earth system modeling. *Geosci. Model Dev.* **16**, 3123–3135 (2023).
28. T. Schneider, S. Behera, G. Boccaletti, C. Deser, K. Emanuel, R. Ferrari, L. R. Leung, N. Lin, T. Müller, A. Navarra, O. Ndiaye, A. Stuart, J. Tribbia, T. Yamagata, Harnessing AI and computing to advance climate modelling and prediction. *Nat. Clim. Change* **13**, 887–889 (2023).
29. C. O. de Burgh-Day, T. Leeuwenburg, Machine learning for numerical weather and climate modelling: A review. *Geosci. Model Dev.* **16**, 6433–6477 (2023).
30. D. Kochkov, J. Yuval, I. Langmore, P. Norgaard, J. Smith, G. Mooers, M. Klöwer, J. Lottes, S. Rasp, P. Düben, S. Hatfield, P. Battaglia, A. Sanchez-Gonzalez, M. Willson, M. P. Brenner, S. Hoyer, Neural general circulation models for weather and climate. *Nature* **632**, 1060–1066 (2024).
31. Z. Qin, Q. Zeng, Y. Zong, F. Xu, Image inpainting based on deep learning: A review. *Displays* **69**, 102028 (2021).
32. X. Zhang, X. Wang, C. Shi, Z. Yan, X. Li, B. Kong, S. Lyu, B. Zhu, J. Lv, Y. Yin, Q. Song, X. Wu, I. Mumtaz, DE-GAN: Domain embedded GAN for high quality face image inpainting. *Pattern Recognit.* **124**, 108415 (2022).
33. Y. Chen, R. Xia, K. Zou, K. Yang, RNON: Image inpainting via repair network and optimization network. *Int. J. Mach. Learn Cybern.* **14**, 2945–2961 (2023).
34. R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, V. Lempitsky, Resolution-robust large mask inpainting with Fourier convolutions. arXiv:2109.07161 [cs, eess] (2021).
35. D. E. Parker, T. P. Legg, C. K. Folland, A new daily central England temperature series, 1772–1991. *Int. J. Climatol.* **12**, 317–342 (1992).
36. M. Davis, *Late Victorian Holocausts: El Niño Famines and the Making of the Third World* (Verso Books, 2002).
37. B. Huang, M. L'Heureux, Z.-Z. Hu, X. Yin, H.-M. Zhang, How significant was the 1877/78 El Niño? *J. Clim.* **33**, 4853–4869 (2020).
38. B. S. Giese, S. Ray, El Niño variability in simple ocean data assimilation (SODA), 1871–2008. *J. Geophys. Res.* **116**, C02024 (2011).
39. E. N. Voskresenskaya, O. V. Marchukova, Qualitative classification of the La Niña events. *Phys. Oceanogr.* 14–24 (2015).
40. L. C. Slivinski, G. P. Compo, J. S. Whitaker, P. D. Sardeshmukh, B. S. Giese, C. McColl, R. Allan, X. Yin, R. Vose, H. Titchner, J. Kennedy, L. J. Spencer, L. Ashcroft, S. Brönnimann, M. Brunet, D. Camuffo, R. Cornes, T. A. Cram, R. Crouthamel, F. Domínguez-Castro, J. E. Freeman, J. Gergis, E. Hawkins, P. D. Jones, S. Jourdain, A. Kaplan, H. Kubota, F. Le Blancq, T.-C. Lee, A. Lorrey, J. Luterbacher, M. Maugeri, C. J. Mock, G. W. K. Moore, R. Przybylak, C. Pudmenzky, C. Reason, V. C. Slonosky, C. A. Smith, B. Tinz, B. Trewin, M. A. Valente, X. L. Wang, C. Wilkinson, K. Wood, P. Wyszyński, Towards a more reliable historical reanalysis: Improvements for version 3 of the twentieth century reanalysis system. *Q. J. Roy. Meteorol. Soc.* **145**, 2876–2908 (2019).
41. C. P. Morice, J. J. Kennedy, N. A. Rayner, J. P. Winn, E. Hogan, R. E. Killick, R. J. H. Dunn, T. J. Osborn, P. D. Jones, I. R. Simpson, An updated assessment of near-surface temperature change from 1850: The HadCRUT5 data set. *J. Geophys. Res. Atmos.* **126**, e2019JD032361 (2021).
42. R. A. Rohde, Z. Hausfather, The berkeley earth land/ocean temperature record. *Earth Syst. Sci. Data.* **12**, 3469–3479 (2020).
43. Copernicus Climate Change Service (C3S), Sea ice concentration daily gridded data from 1979 to present derived from satellite observations (2020); https://cds.climate.copernicus.eu/cdsapp#!/dataset/satellite-sea-ice-concentration?tab=overview.
44. J. M. Beckers, M. Rixen, EOF calculations and data filling from incomplete oceanographic datasets. *J. Atmos. Oceanic Tech.* **20**, 1839–1856 (2003).
45. K. Wang, G. D. Clow, Reconstructed global monthly land air temperature dataset (1880–2017). *Geosci. Data J.* **7**, 4–12 (2020).
46. T. M. Smith, R. W. Reynolds, R. E. Livezey, D. C. Stokes, Reconstruction of historical sea surface temperatures using empirical orthogonal functions. *J. Climate* **9**, 1403–1420 (1996).
47. M. S. Y. Tang, S. N. Chenoli, A. A. Samah, O. S. Hai, An assessment of historical Antarctic precipitation and temperature trend using CMIP5 models and reanalysis datasets. *Polar Sci.* **15**, 1–12 (2018).
48. R. Keisler, Forecasting global weather with graph neural networks. arXiv:2202.07575 [physics] (2022).
49. S. Scher, G. Messori, Physics-inspired adaptions to low-parameter neural network weather forecasts systems. arXiv:2008.13524 [physics] (2023).
50. C. Esteves, J.-J. Slotine, A. Makadia, "Scaling spherical CNNs," in *Proceedings of the 40th International Conference on Machine Learning* (MLResearchPress, 2023), vol. 202, pp. 9396–9411.
51. A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, L. Van Gool, RePaint: Inpainting using denoising diffusion probabilistic models. arXiv:2201.09865 [cs] (2022).

Downloaded from https://www.science.org on August 14, 2025

52. C. Wei, K. Mangalam, P.-Y. Huang, Y. Li, H. Fan, H. Xu, C. Xie, A. Yuille, C. Feichtenhofer, Diffusion models as masked autoencoders. arXiv:2304.03283 [cs] (2023).
53. T. Höppe, A. Mehrjou, S. Bauer, D. Nielsen, A. Dittadi, Diffusion models for video prediction and infilling. arXiv:2206.07696 [cs, stat] (2022).
54. Y. Zeng, J. Fu, H. Chao, B. Guo, Aggregated contextual transformations for high-resolution image inpainting. *IEEE Trans. Vis. Comput. Graph* **29**, 3266–3280 (2021).
55. H. Hukkelås, F. Lindseth, R. Mester, Image inpainting with learnable feature imputation. arXiv:2011.01077 [cs, eess] (2020).
56. R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models. arXiv:2112.10752 [cs] (2022).
57. P. Kulshreshtha, B. Pugh, S. Jiddi, Feature refinement to improve high resolution image inpainting. arXiv:2206.13644 [cs, eess] (2022).
58. L. Chi, B. Jiang, Y. Mu, "Fast fourier convolution," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin, Eds. (Curran Associates, Inc., 2020), vol. 33, pp. 4479–4488.
59. Y. Kilcher, Resolution-robust large mask inpainting with Fourier convolutions (w/ author interview) (2021); www.youtube.com/watch?v=Lg97gWXsiQ4.
60. F. Crameri, G. E. Shephard, P. J. Heron, The misuse of colour in science communication. *Nat. Commun*. **11**, 5444 (2020).
61. B. Murphy, R. Yurchak, S. Müller, GeoStat-Framework/PyKrige: v1.7.0 (2022); https://zenodo.org/record/7008206.
62. Z. Wang, A. C. Bovik, Mean squared error: Love it or leave it? A new look at Signal FidelityMeasures. *IEEE Signal Process. Mag*. **26**, 98–117 (2009). https://ieeexplore.ieee.org/document/4775883.

Acknowledgments: This is ClimTip contribution no. 19; the ClimTip project has received funding from the European Union's Horizon Europe research and innovation program under grant agreement no. 101137601. Parts of the computations were performed on resources provided by Sigma2–the National Infrastructure for High Performance Computing and Data Storage in Norway under the project nn8008k. Support for the Twentieth Century Reanalysis Project version 3 dataset is provided by the US Department of Energy, Office of Science Biological and Environmental Research (BER), by the National Oceanic and Atmospheric Administration Climate Program Office, and by the NOAA Earth System Research Laboratory Physical Sciences Laboratory. We gratefully acknowledge the Ministry of Research, Science and Culture (MFWK) of Land Brandenburg for supporting this project by providing resources on the high-performance computer system at the Potsdam Institute for Climate Impact Research. Parts of the publication charges for this article have been funded by a grant from the publication fund of UiT The Arctic University of Norway. Some of the figures (Figs. 1B and 2 to 4 and figs. S3, S5, and S7 to S12) are made following the guidance for color maps by Crameri *et al.* (*60*). We thank R. Suvorov for answering technical questions regarding the LaMa model, J. Meuer for providing the CMIP5 training and validation data, and C. Kadow for providing the output of the model. **Funding:** N.Boe. acknowledges funding from ClimTip, which has received funding from the European Union's Horizon Europe research and innovation program under grant agreement no. 101137601. N.Boe. acknowledges further funding by Volkswagen Foundation and the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement no. 956170, as well as from the Federal Ministry of Education and Research under grant no. 01LS2001A. M.R., A.P., and N.Bow. were supported by the UiT Aurora Centre Program, UiT The Arctic University of Norway (2020), and the Research Council of Norway (project number 314570). **Author contributions:** N.Bow. and N.Boe. conceived and designed the study. N.Bow. carried out the computations and analyzed the results with assistance from A.P. N.Boe. and M.R. supervised the project. All authors discussed the results. N.Bow. wrote the paper with contributions from all authors. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. The HadCRUT4 data (*2*) are available at www.metoffice.gov.uk/hadobs/hadcrut4/. The HadCRUT5 data (*41*) are available at www.metoffice.gov.uk/hadobs/hadcrut5/data/HadCRUT.5.0.2.0/download.html. The NOAA/CIRES/DOE 20th Century Reanalysis (version 3) data are provided by the NOAA PSL, Boulder, Colorado, US, and are available at their website at https://psl.noaa.gov/data/gridded/data.20thC_ReanV3.html (*40*). The BEST dataset is freely available at https://berkeleyearth.org/data/ (*42*). The sea ice concentration data are available at https://cds.climate.copernicus.eu/cdsapp##!/dataset/satellite-sea-ice-concentration. The model check points as well as the reconstructed fields shown are available on Zenodo at www.doi.org/10.5281/zenodo.10512175. The original LaMa model is available on GitHub at https://github.com/advimman/lama. Our modified version is available at https://github.com/NilsBochow/lama_reconstruction. Both LaMa versions are also archived on Zenodo at www.doi.org/10.5281/zenodo.10512175. We use the software package PyKrige (*61*) for reconstructing the temperature records via kriging.

Submitted 6 March 2024
Accepted 26 February 2025
Published 2 April 2025
10.1126/sciadv.adp0558

Downloaded from https://www.science.org on August 14, 2025