Climate field completion via Markov random fields – Application to the HadCRUT4.6 temperature dataset.

Adam Vaccaro,[1] Julien Emile-Geay,[1] Dominique Guillot,[2] Resherle Verna,[1]

Colin Morice,[3] John Kennedy,[3] and Bala Rajaratnam[4]

Corresponding author: Adam Vaccaro (avaccaro@usc.edu)

[1]Department of Earth Sciences,

University of Southern California, Los

Angeles, CA, USA

[2]Department of Mathematical Sciences,

University of Delaware, Newark, DE, USA

[3]Met Office Hadley Centre, Fitzroy Road,

Exeter, EX1 3PB, UK

[4]Department of Statistics, University of

California, Davis, CA, USA

Abstract. Surface temperature is a vital metric of Earth's climate state, but is incompletely observed in both space and time. Here, we leverage GraphEM, a recently developed imputation method, to construct a spatially complete estimate of the HadCRUT4.6 global surface temperature dataset back to AD 1850. GraphEM leverages Gaussian Markov random fields (aka Gaussian graphical models) to better estimate covariance relationships within a climate field. An investigation of the resulting network reveals that GraphEM is able to detect anisotropic features, such as land/ocean contrasts, orography, ocean currents and wave-propagation pathways – all leading to improved estimates of missing values. This interpolated analysis of HadCRUT4.6 data is available as a 100-member ensemble, propagating information about sampling variability available from the original HadCRUT4.6 dataset. A comparison of NINO3.4 and global mean monthly temperature series with published datasets reveals similarities and differences that can be directly traced to the spatial interpolation method. Notably, the GraphEM-completed HadCRUT4.6 global temperature displays no "hiatus" in the early twenty-first century, consistent with recent analyses using interpolated fields. Known events like the 1877/1878 El Niño are recovered with greater fidelity than a competing method (kriging), and result in different assessments of changes in ENSO variability through time. Gaussian Markov random fields provide a more geophysically-motivated way to impute missing values in climate fields, and the associated graph provides a tool to analyze teleconnections patterns. We close with a discussion of wider applications of Markov random fields in climate science.

1. Introduction

The surface temperature record, comprised of land surface air temperature (LSAT) measurements and sea-surface temperature (SST) readings from around the world, is a critical resource in quantifying and understanding climate change over the industrial era. By offering a glimpse into climate conditions of the past, it plays a vital role in climate science by helping frame recent trends in the context of past variability [*Stocker et al.*, 2013]. Since they are based on in-situ temperature measurements, instrumental surface datasets are also fundamental to the validation of remote sensing retrieval algorithms [*Karl et al.*, 2006], and also serve as key training datasets for paleoclimatic reconstructions [*Tingley et al.*, 2012; *Hakim et al.*, 2016]. They also play a critical role in the evaluation of climate models [*Simmons et al.*, 2004; *Flato et al.*, 2013].

However, since instrumental datasets draw on a variety of observational platforms, the inhomogeneities among the various types of measurements (such as different instrumentation, measurement type, and time of day), can manifest as bias in metrics calculated from the measurements if necessary corrections are not applied [*Kennedy et al.*, 2011; *Williams et al.*, 2012; *Hausfather et al.*, 2013; *Kennedy*, 2013; *Kent et al.*, 2016]. Recent studies suggest that sources of inhomogeneity in the instrumental data that were previously unaccounted for are at least partially responsible for the so-called global warming "hiatus" (a slowdown of temperature rise between about 1998 and 2013) apparent in some global temperature estimates [*Karl et al.*, 2015]. A separate problem is spatial coverage, which is highly nonstationary, with regions of the globe that lack observational coverage of any kind at various points in time (Fig. 1). A spatially complete, homogeneous global

temperature dataset is preferred by some researchers, but such datasets do not exist for long periods of time (satellite-based estimates are only available since 1979, and they too suffer from biases [e.g. *Mears and Wentz*, 2005; *Thorne et al.*, 2010]). It would therefore be desirable to fill in the gaps in the surface temperature record in a way that is maximally consistent with existing observations. To this end, let us consider the HadCRUT4 dataset.

The U.K. Met Office Hadley Centre and the University of East Anglia's Climatic Research Unit's global surface temperature dataset, HadCRUT [*Brohan et al.*, 2006; *Morice et al.*, 2012], is among the most widely used global instrumental temperature datasets available [*Flato et al.*, 2013; *Hausfather et al.*, 2013]. It is derived from near-surface air temperature data from the Hadley Centre and Climatic Research Unit's land-based CRUTEM4 [*Jones et al.*, 2012] combined with sea surface temperature data from the Hadley Centre's HadSST3 dataset [*Kennedy et al.*, 2011a; *Kennedy et al.*, 2011b]; the result is a global gridded surface temperature dataset dating back to 1850. Indeed, it is the longest of the three major independently produced global surface temperature datasets, which include NOAA/NCDC's Merged Land-Ocean Surface Temperature Analysis [*Vose et al.*, 2012] and NASA's GISTEMP [*Hansen et al.*, 2010]. The coverage is fragmentary, however: observations were not available for slightly more than half of the grid cells, most of them at the poles, over Africa, and generally more so earlier in the record (Fig. 1).

This uneven distribution of missing observations introduces a bias into the estimation of the global mean temperature. Under-sampling of the fastest warming parts of the globe (the poles) leads to an underestimation of the global mean temperature trend [*Simmons et al.*, 2010; *Folland et al.*, 2013]. This coverage bias is of particular concern over recent

decades due to the different rates of warming exhibited between high and low latitudes and between land and ocean points [*Hansen et al.*, 2006].

A previous study [*Cowtan and Way*, 2014, hereafter CW14] identified the uneven distribution of unsampled regions around the globe as a source of bias in global temperature trends calculated using the HadCRUT4 dataset. In particular, CW14 found that incomplete sampling of high-latitude polar regions lead to an underestimation of the global warming trend over recent decades (i.e., the so-called "hiatus") found in other global temperature reconstructions [*Easterling and Wehner*, 2009; *Hansen et al.*, 2010; *Karl et al.*, 2015]. After applying a kriging-based interpolation method [*Cressie*, 1990] to the raw HadCRUT4 data, the authors found that the warming trend had been restored, suggesting that an optimal imputation scheme is necessary for a more accurate portrayal of temperature variations in this and other surface temperature datasets. Other studies have replicated CW14's results, and shown that interpolation over Arctic regions eliminates the supposed hiatus [*Huang et al.*, 2017].

Spatial estimation methods have a long history in this field, which we recount more fully in section 2. For now we note that, whether Bayesian or frequentist in essence, these approaches belong to two broad categories: (1) *local* methods, which parametrically model spatial covariances to infer missing data as a function of distance only, usually within a short radius [e.g. *Hansen and Lebedeff*, 1987; *Reynolds and Smith*, 1994; *Smith et al.*, 1996; *Reynolds et al.*, 2002; *Tingley and Huybers*, 2010; *Rohde et al.*, 2013; *Cowtan and Way*, 2014]; (2) *global* methods, which leverage long-range correlations, but must regularize the covariance estimation problem either via truncation [*Kaplan et al.*, 1997, 1998, 2003] or model selection [*Guillot et al.*, 2015]. Very few methods bridge this dichotomy; *Karspeck*

et al. [2012] offer one example, constructing a non-stationary Matérn model to estimate the North Atlantic sea surface temperature field in a Bayesian framework. Multiresolution lattice kriging *Nychka et al.* [2015] offers another, recently applied to surface temperature [*Ilyas et al.*, 2017].

The objective of this paper is to apply the theory of Markov random fields (MRFs, aka Graphical models) to this problem. Recently, *Guillot et al.* [2015, hereafter, G15] used MRFs to flexibly model temperature covariance in the context of paleoclimate reconstructions, a closely related problem [*Tingley et al.*, 2012]. Here we apply G15's algorithm (GraphEM) to produce a 100-member ensemble of spatially complete realizations of the latest HadCRUT dataset (HadCRUT4.6). We show that the imputation results in more realistic reconstructions of past climate events, like the 1877/78 El Niño, and stronger historical warming trends than in uninterpolated data.

The rest of this paper is structured as follows: we first describe the data, as well as the methods used to infill the gaps herein (Section 2). Results are presented in section 3, followed in section 4 by an analysis of the climate network afforded by the graphical approach. Section 5 discusses the benefits and limitations of our approach, and offers conclusions.

2. Data and Methods

2.1. Imputation of missing values

Following *Schneider* [2001], we consider an incomplete climate dataset with n samples of p variables (say, monthly surface temperature over the past 167 years on a grid with $p = 2592$ points, corresponding to HadCRUT4's $5° \times 5°$ grid). Let us denote by X the $n \times p$ data matrix. For a given row x of X, let x_a and x_m denote the portions of the values in x

that are available and missing, respectively. Let μ and Σ denote the mean and covariance matrix of the dataset, similarly partitioned between available and missing values for a given row. Our goal is to fill in the blanks of the dataset in a way that is consistent with the available data. This is generally accomplished (for each row of the data matrix) by a linear regression of the form

$$\mathsf{x}_m - \boldsymbol{\mu}_m = (\mathsf{x}_a - \boldsymbol{\mu}_a)\mathsf{B} + \mathsf{e} \tag{1}$$

where e is the imputation error, of zero mean and covariance matrix C. Many techniques have been applied to estimate the matrix of regression coefficients B but the infilled values $\hat{\mathsf{x}}_m$ are generally expectation values based on the regression model Eq(1), that is, linear combinations of available observations with weights given by the regression coefficients B. The regression coefficients B depend on the $\hat{\mu}$ and $\hat{\Sigma}$ statistics of the data. If they can be estimated, the missing values can be imputed. Conversely, the statistics of the data depend on the missing values. So estimating statistics of incomplete data and imputing missing values are parts of an estimation problem that is generally nonlinear [*Schneider*, 2001]. For normal data, the mean and covariance matrix are sufficient statistics and determine the parameters in the regression model (1). Given an estimate $\hat{\Sigma}$ of the covariance matrix of the data, partitioned into the submatrices describing relationships between observed and missing values,

$$\hat{\Sigma}_{aa} \equiv \widehat{\mathrm{Cov}}(\mathsf{x}_a, \mathsf{x}_a), \quad \hat{\Sigma}_{am} \equiv \widehat{\mathrm{Cov}}(\mathsf{x}_a, \mathsf{x}_m), \quad \hat{\Sigma}_{mm} \equiv \widehat{\mathrm{Cov}}(\mathsf{x}_m, \mathsf{x}_m), \tag{2}$$

the ordinary least-squares estimate of the regression coefficients can be expressed as

$$\hat{\mathsf{B}} = \hat{\Sigma}_{aa}^{-1}\hat{\Sigma}_{am}. \tag{3}$$

(Hats denote estimated quantities.) Under normality, then, the mean and covariance matrix are sufficient statistics that determine the expected values of the missing data; the estimate (3) is the conditional maximum likelihood estimate of the regression coefficients given a covariance matrix estimate of full rank. The estimate (3) of the regression coefficients implies two central challenges in carrying out an imputation of missing values based on (1): (*i*) the covariance matrices must be reliably estimated; (*ii*) some submatrices ($\hat{\Sigma}_{aa}$) must be inverted. Although these problems are obviously linked, they are somewhat distinct in practice, and both can be problematic. The standard sample covariance matrix typically used for $\hat{\Sigma}$ may not be a good estimator, as we will discuss further below. Moreover, since the number of variables in climate datasets is typically much larger than the number of samples, the sample covariance matrix will be rank-deficient. Hence, it must be *regularized* to obtain a stable inverse and estimate of the regression coefficients. The central mathematical problems at hand are therefore those of covariance estimation and regularization. Mean vectors $\boldsymbol{\mu}$ must also be estimated, but this step is less problematic than the estimation of covariance matrices from incomplete data [*Little and Rubin*, 2002].

The imputation problem therefore boils down to the accurate estimation of a covariance matrix and its subsequent inversion, two tasks that pose a number of problems in high dimensions. Indeed, for HadCRUT4, the number of variables p exceeds the number of monthly observations ($n \sim 2000$). This "large p, small n problem" is well-known to result in sample covariance matrices that are not only rank-deficient, but also very imprecise [*Johnstone*, 2001] This prohibits their inversion, thus imposing the need for some form of regularization to impute the missing data. Note that this need arises universally in spatial estimation problems, regardless of the framework.

In climate studies, covariance regularization has been accomplished in three main ways: (i) explicit modeling of the spatial dependence of covariances [e.g., *Hansen and Lebedeff*, 1987; *Jones and Moberg*, 2003; *Brohan et al.*, 2006; *Tingley and Huybers*, 2010]; (ii) expansion of covariance matrix estimates in a truncated principal component basis to achieve dimension reduction [*Reynolds and Smith*, 1994; *Smith et al.*, 1996, 2008; *Kaplan et al.*, 1997, 1998, 2000, 2003]; (iii) filtering of trailing principal components of covariance matrix estimates through Tikhonov regularization/ridge regression [*Schneider*, 2001]. In the first case, regularization is achieved by choosing spatial covariance functions that drop sharply with radial distance (e.g., Matérn kernels). This choice is subjective and so far has not allowed for anisotropy or inhomogeneities that may be associated with surface features like coastlines or mountain ranges. In the second case, regularization is achieved through the choice of a truncation parameter (the number of principal components to be retained). Heuristics have typically been used for this choice, representing a limitation of this approach as practiced in climate applications. In the third case, regularization is achieved through the choice of a continuous regularization parameter ("ridge parameter"), which measures the strength of the filtering of trailing principal components [*Hoerl and Kennard*, 1970a, b; *Tikhonov and Arsenin*, 1977]. Ridge regression with determination of a regularization parameter by generalized cross-validation (GCV; *Golub et al.* 1979; *Craven and Wahba* 1979) has been used in the regularized expectation maximization (EM) algorithm [*Schneider*, 2001] and was shown to lead to more accurate SST estimates for recent decades than the widely used interpolation method of *Reynolds and Smith* [1994].

A novel approach consists of modeling temperature as a Markov random field [*Guillot et al.*, 2015]. This involves representing conditional independence relations between locales

with the aid of graph $G = (V, E)$, where V are the vertices (temperature grid points) and E the edges of the graph (connections between grid points). Because this approach will be unfamiliar to most readers, we now provide a brief overview.

2.2. Graphical covariance modeling

An attractive feature of graph-based methods is their demonstrated link to dynamical features of fluid flow (land-ocean boundaries, orography, or wave propagation pathways, aka teleconnections), which traditional covariance modeling methods do not capture [*Tupikina et al.*, 2016]. This makes such graphs applicable to any geophysical field that may be modeled by a multivariate normal distribution, e.g. including those representing passive tracers. This includes monthly surface temperature, which may be regarded as a passive tracer of atmospheric flow on such scales [*Tupikina et al.*, 2016]. Care must be taken, however, in the definition of such graphs: *Zerenner et al.* [2014] found that working with covariance or concentration graphs (i.e. looking at marginal vs conditional independence relations) lead to radically diffent structures. This is not surprising, as a zero in the concentration matrix ($\Omega = \Sigma^{-1}$) means that the nodes are conditionally independent given values of the field at all other nodes, while zeros in Σ indicate marginal independence [*Dempster*, 1972; *Lauritzen*, 1996]. In practice, fields that show long-range correlations will show very few true zero entries in Σ, though they could be quite abundant in Ω. Concentration graph models leverage these zeros to reduce the dimensionality of the estimation problem, resulting in a well-conditioned covariance matrix Σ.

Recently, G15 embedded such concentration graphs within an Expectation-Maximization (EM) algorithm [*Dempster et al.*, 1977]. The resulting method, GraphEM, is a generalization of the RegEM algorithm [*Schneider*, 2001], and was successfully used to

estimate annual temperature over the Common Era [*Wang et al.*, 2015] with demonstrably higher spatial fidelity than competing methods [*Wang et al.*, 2014; *Guillot et al.*, 2015], including RegEM variants. The main difficulty in this approach is to reliably estimate the concentration graph G from a finite dataset in the presence of observational noise.

G15 explored two approaches to graph estimation: (1) modeling G as a neighborhood graph – that is, a graph obtained by declaring all points with a radius R of each vertex to be a neighbor; (2) using the graphical lasso [GLASSO, *Friedman et al.*, 2008], an ℓ_1 penalized likelihood method, to discover the conditional independence relations hiding in the data. Both approaches involve a tuning parameter (the radius R or the penalty parameter ρ [see *Guillot et al.*, 2015, Eq (2.1)]), and both parameters may be optimized via cross-validation. G15 found that for suitable choices of ρ, the GLASSO approach yielded better estimates of the temperature field than the neighborhood graph method. The choice of ρ is a tradeoff between including too few neighbors (ρ large), which in the limit makes the graph strictly local; or including too many neighbors (ρ small), some of which are spurious and could raise imputation error. Once a graph is identified, GraphEM uses it to obtain well-conditioned estimates of Σ and its inverse Ω, which are then used to estimate the missing values via regression (Fig. 2).

In the following, we apply the GLASSO variant of GraphEM to the HadCRUT4.6 dataset, which we now describe. The computational steps are outlined in Section 2.4.

2.3. Source Data

The original source of historical temperature data used in this study is HadCRUT4 (version 4.6.0.0, hereinafter HadCRUT4.6). HadCRUT4 is a monthly-resolved global surface temperature anomaly dataset placed on a 5° x 5° grid. The anomalies are calculated with

respect to a 1961-1990 reference period and are available from January 1850 to September 2018. The HadCRUT4 dataset is nearly ubiquitous in the climate science community, and indeed, is one of the most widely used temperature datasets [*Brohan et al.*, 2006; *Morice et al.*, 2012; *Cowtan and Way*, 2014]. However, like any surface temperature dataset, HadCRUT4 suffers from incomplete coverage: historical observations are not available for slightly more than half of the grid cells (∼53%) over the entire time period (1850-2017), but HadCRUT4 shows improved coverage (∼70%) since 1979 (Fig.1).

CW14 identified such coverage gaps as a major source of bias in global average temperature derived from the HadCRUT4 dataset. In particular, the authors found that incomplete sampling of high-latitude polar regions leads to an underestimation of the global warming trend over recent decades found in other preexisting global temperature reconstructions. The authors performed three reconstructions: null, kriging, and hybrid. The null reconstruction involves setting the missing values to the global mean temperature, whereas kriging allows for the estimation of unknown values by making use of the weighted average of neighboring sampled data. The hybrid reconstruction features the combination of both the kriging method and the University of Alabama-Huntsville's satellite dataset [*Spencer*, 1990; *Christy et al.*, 2007].

The authors found that in all cases, the kriging and the hybrid approaches outperformed the null reconstruction. The authors also found that the hybrid approach outperformed the kriging approach in a few cases, notably at high latitudes. However, due to the hybrid approach spanning over a short period time, we use the kriging solution as a starting point for our work, since it covers the period from 1850 to 2017.

We build on CW14 by using the global temperature series that they obtained using their kriging reconstruction to update the HadCRUT4 data before infilling the remaining missing values. HadCRUT4 is organized as a 100 ensemble member dataset, which samples the distribution of likely surface temperature anomalies [*Morice et al.*, 2012]. Structuring the dataset in this manner allows for better representation of complex temporal and spatial interdependencies of measurement and bias uncertainties, which allows correlated uncertainties to be considered [*Morice et al.*, 2012].

Additionally, we make use of a hybrid method which consists of both the kriging method and Cowtan and Way's reconstruction version 2, which features the MERRA short reconstruction. The MERRA short reconstruction consists of data spanning the 1979 – 2017 period and offers near global coverage [*Bosilovich et al.*, 2015]. Our results will be compared against a comprehensive set of surface temperature datasets (Table 1).

2.4. Workflow

Here, we produce a spatially-complete global gridded temperature dataset extending back until 1850. We start off by estimating both the conditional independence graph and the covariance matrix of the gridded temperature field using information from the HadCRUT4$CW(MERRA)$ reconstruction. We use two versions of Cowtan and Way's HadCRUT4 dataset: one featuring only the kriging (HadCRUT4$_{CW}$) reconstruction that we use as an initial guess in GraphEM and the other using a hybrid approach that features both kriging and satellite data (HadCRUT4$_{CW(Merra)}$) [*Cowtan and Way*, 2014; *Bosilovich et al.*, 2015] over the 1979–2017 period. The latter dataset was used to estimate the graph, which assumes that this graph is constant through the entire 1850–2017 period. The graph is then obtained via the ℓ_1-penalized maximum likelihood method described

in G15 (Section 2.1). More specifically, the graphical lasso is used to obtain a sparse estimate of the concentration matrix of the field for a given target sparsity level. Once this concentration matrix is obtained, the graph can be estimated from the pattern of zeros in the concentration matrix. The optimal target sparsity parameter for glasso is chosen using 10-fold cross-validation performed over available grid points. In this context, cross-validation is performed over both time and space: for each fold, a partition representing 10% of the entire time axis is randomly selected for each grid cell. Cross-validation scores are shown in Figure 4 Testing with higher sparsities showed that despite yielding lower expected prediction error over grid cells that were available in the raw data, the reconstructions tended to show unrealistically large magnitudes of temperature further back in time when observations are especially scarce. These anomalous values are likely the result of overfitting the regularization parameter.

To identify a sparsity level that minimized the number of extreme temperature anomalies, we examined the temperature frequency distribution at varying sparsity levels. Kernel density estimates of temperature anomaly probabilities are displayed in Figure 5. Of the sparsity levels tested, 0.6% sparsity showed the least amount of distortion compared to the raw HadCRUT4.6 and yielded the lowest frequency of extreme anomalies. As such, 0.6% sparsity was selected as the optimal sparsity level.

Because GraphEM requires an initial guess for the field before interpolating missing values, we used the kriging solution from CW14. The results are not sensitive to this choice, though it greatly accelerated computations by providing a "warm" start.

3. Results

Here we examine the results of the GraphEM imputation against published HadCRUT4 variants, as well as other datasets produced independently of HadCRUT4 (Table 1). Because the full dataset is large ($2016 \times 2592 \times 100$), we focus for this comparison on two indices: the global mean temperature (GMT) and the NINO3.4 index, a common metric of El Niño activity [*Trenberth and Stepaniak*, 2001]. Spatial context will be given as appropriate.

3.1. Global variability

We first compare GMT from the GraphEM-infilled dataset to series computed from the raw HadCRUT4.6 median data and from HadCRUT4.6$_{CW}$. This will illustrate the effect of the GraphEM infilling and quantify the difference between it and the kriging method used by CW14.

Figure 6 shows GMT and NINO3.4 monthly anomalies from the GraphEM solution HadCRUT4.6$_G$ to the same series obtained from the raw HadCRUT4.6 median (green) and the CW14 median (HadCRUT4.6$_{CW}$, red). Unsurprisingly, the datasets are very close over recent times (post 1950), but diverge markedly back in time. The monthly GMT series from HadCRUT4.6$_G$ exhibits much more variability than the raw data pre-1950 and shows a secular warming trend similar to HadCRUT4.6$_{CW}$. Hints of an enhanced seasonal cycle are also detectable in the two interpolated version (Fig 6, bottom left). These differences are entirely driven by data availability: series using the raw HadCRUT4.6 shrink to the mean, while their interpolated counterparts make use of what little observations are available, and propagate them in space in accordance to their covariance model.

Figure 7 gathers such warming trends calculated over various intervals, expressed in °C/century as a basis for comparison. It shows that the trend calculated over 1998-2013

in HadCRUT4.6$_G$ (0.08°C/decade) is larger than in the raw median (0.05°C/decade), but quite a bit smaller than that from HadCRUT4.6$_{CW}$ (0.1°C/decade). This confirms CW14's finding that spatial incompleteness is a major contributor to the global warming "hiatus" found in previous analyses using the raw HadCRUT4 dataset. Looking at the longest common period (1880–2017), HadCRUT4.6$_{CW}$ displays the strongest trends ($_{0.686}0.709$°C/century$_{0.731}$, 95% confidence interval) of any HadCRUT4 variant. So while it is true that trends in interpolated datasets are greater than in the raw HadCRUT4.6 median over all periods analyzed, the type of interpolation is seen to be quite consequential in these estimates.

Figure 8 shows the monthly-resolved GMT series from HadCRUT4$_G$, NASA's GISTEMP and NOAAGlobalTemp plotted over the entire interval (top) and over a recent period (1985-present, bottom). The overall shape of all three series look similar, especially the long term trends, but HadCRUT4$_G$'s GMT is sometimes colder by 0.5–1°C in the 1880's – a large difference indeed, resulting in stronger secular trends for HadCRUT4$_G$ prior to the mid twentieth century. However, HadCRUT4$_G$ (like its other HadCRUT4 variants) is colder over recent decades, so the late twentieth century trend is larger in GISTEMP and GlobalTemp ($_{1.052}1.136$°C/century$_{1.221}$ and $_{1.040}1.118$°C/century$_{1.195}$, respectively). Differences become more apparent at higher resolution (Fig 8, bottom).

A notable point from Fig 7 is that interpolation methods result in notably different trends between HadCRUT4 variants and external datasets (NASA's GISTEMP, NOAAGlobalTemp). This underlies the importance of interpolation and motivates using the most appropriate methods. Another source of differences between the HadCRUT4 variants and GISTEMP/NOAAGlobalTemp is the bias adjustments applied to SSTs, which

affect trends across all the period analyzed, particularly the post 1998 "hiatus" and the period from 1950 on. *Kent et al.* [2016] covers the longer periods, *Hausfather et al.* [2017], the more recent.

To gain insight into GMT behavior, the zonally-averaged temperature evolution from HadCRUT4$_G$ is charted in Fig 9. It shows the largest variations over polar regions, particularly the Arctic, illustrating the well-known "polar amplification" phenomenon. This reinforces the result of CW14 that missing values over this region were key to the underestimation of recent trends in the raw HadCRUT4 dataset. It also clearly shows that the onset of twentieth century warming occurred earlier in high northern latitudes.

3.2. Regional variability: the case of ENSO

With a comprehensive depiction of spatio-temporal temperature variability clearly beyond the score of this paper, we focus here on El Niño-Southern Oscillation (ENSO) variability. ENSO is the leading mode of global interannual variability, influencing climate and weather over much of the globe [*Sarachik and Cane*, 2010; *Trenberth et al.*, 1998]. ENSO can be described by many metrics [*Trenberth and Stepaniak*, 2001], the NINO3.4 index (average SST over [170°–120°W, 5°S– 5°N]) being a common one [*Trenberth*, 1997]. One motivation for assessing the impact of imputation on this index is that *Emile-Geay et al.* [2013a, b] found that instrumental trends in NINO3.4 had a leading-order influence on the amplitude of reconstructed NINO3.4 variability over the past millennium.

NINO3.4 indices computed from each of the HadCRUT4 variants are plotted in Figure 6 (right). Overall, the three datasets show similar trends over recent decades (1960-present), but show divergence pre-1960. HadCRUT4$_{CW}$ shows the least amount of variance further back in time, while HadCRUT4$_G$ shows the most. This is again a consequence of

interpolation: when no interpolation is performed, the index is dominated by the few observations available at that time, which mostly follow shipping tracks between Australia and the United States (Fig 1, *Bunge and Clarke* [2009]). The CW14 kriging method, which assumes a relatively narrow decorrelation radius, smooths these anomalies to a field mean estimated by generalized least squares away from shiptracks (i.e. over much of the NINO3.4 box). In contrast, the GraphEM solution leverages large-scale teleconnections to infer NINO3.4 temperature (section 4), resulting in stronger anomalies. This is particularly clear for the year 1917, which stands out in differences to the raw HadCRUT4 of up to 5K (Fig. 6, bottom right). The large negative anomaly arises from isolated ship observations. Where observations are isolated in this way, the automated quality control is less reliable and occasional large outliers can pass the checks. Reliable interpolation methods could be used to improve the background fields against which quality control checks are made, thereby improving data quality.

It is instructive to look at spatial anomalies to understand what is at work. Figure 10 illustrates this for the well-known 1877/78 El Niño, the biggest documented event prior to 1982 [*Quinn*, 1992], which caused widespread famines in China and India, estimated to have caused the premature death of about 20 million people [*Davis*, 2001]. Figure 10 shows a comparison between the raw (uninterpolated) HadCRUT4 dataset, HadCRUT4$_{CW}$, and HadCRUT4$_{G}$ over various seasonal windows straddling 1877/1878. The raw HadCRUT4 dataset is shown to be missing vast swathes of the Pacific. Yet, subtle hints of the equatorial warming associated with the El Niño event can be detected. After applying a kriging-based interpolation method to the dataset (CW, bottom row), the presence of the El Niño event becomes clearer, but the pattern is patchy, with reduced amplitude away

from sampled grid points – a natural consequence of distance-based kriging that produces unphysical features. In contrast, GraphEM can recover the full structure of this known El Niño event, including its far field effects over Eurasia, either with a neighborhood graph or a glasso graph.

Thus, the two covariance modeling methods yield very different estimates of past climate variability: HadCRUT4$_{CW}$ seems more conservative with outliers like the 1917 anomaly, but distorts the spatial patterns and likely overdamps anomalies away from observed locales. On the other hand, HadCRUT4$_G$ appears to retain more faithful spatial patterns, but can be vulnerable to outliers. This is a reminder that interpolation methods are only as good as the data that go into them.

Figure 11 compares the HadCRUT4$_G$ NINO3.4 to analogous series derived from global SST datasets (ERSSTv5 [*Huang et al.*, 2015; *Liu et al.*, 2015], COBE SST [*Folland and Parker*, 1995; *Ishii et al.*, 2005]), as well as stand-alone NINO3.4 products [*Bunge and Clarke*, 2009; *Kaplan et al.*, 1998]. See Table 1 for details. As before, differences are most pronounced before 1875, and around 1917, though differences of up to 1.5K are seen through the 2000's for some months (Bunge & Clarke, pink line). These differences are a testament to the difficulty of characterizing ENSO state even with modern observing platforms [*Huang and Kumar*, 2013], and need to be kept in mind when assessing changes in ENSO variability – variance among datasets cannot be neglected.

The comparison is summarized in Table 2, which shows the correlations calculated between HadCRUT4$_G$ and the other series. Most correlations are above 0.9, with the largest differences observed with Kaplan EXTENDED at decadal and longer scales.

4. Analysis of the Graph

To better understand the reasons underlying the better performance of GraphEM in preserving climate features like ENSO teleconnections, we now explore the characteristics of the GLASSO-estimated graph, G. Figure 12 shows the mean degree (number of neighbors) of every vertex in the graph, a basic measure of network connectivity (top). The degree ranges from 5 in parts of the southern ocean – where mesoscale eddies control the dynamics – to 60 in the deep tropics – which act as a Rossby wave source [*Sardeshmukh and Hoskins*, 1988] able to project influence at long distances [*Horel and Wallace*, 1981; *Simmons et al.*, 1983].

Figure 12 (middle) displays the average great circle distance to neighbors. It is seen to be equally variable as the mean degree, ranging from a minimum of 150 km in polar and extratropical land regions, to over 10,000 km in equatorial land regions. To some extent, this is a consequence of gridding in 5° boxes, which packs many neighbors in small areas near the poles. Nonetheless, the pattern shows marked deviations from zonal symmetry, including marked land/ocean contrasts.

Consistent with past work on climate networks [*Tsonis et al.*, 2006; *Zerenner et al.*, 2014], our analysis also finds tropical nodes to be highly connected to the rest of the globe (many neighbors, with large average distances over equatorial forests). However, this picture is highly granular, implying that there is a lot more to this connectivity than latitude.

One way to probe the structure of G is to contrast it to a neighborhood graph G_R. The rationale for this is twofold: (1) by definition, a neighborhood is radially isotropic

and geographically confined, and (2) neighboorhood graphs have been found to provide a reasonable first guess in estimating Σ_{aa} (G15).

We thus define G_R for various values of R. To measure similarity between G and G_R at each location l, we restricted G to the points within R km of l, then computed the percentage of edges that are common to that subgraph of G and G_R. From this we obtain the fraction of common edges (f). We find its mode to be maximized for $R = 1000$ km, and henceforth focus on this value of R.

Figure 12 (bottom), which shows f over the sphere, reveals some similarity between G and G_{1000}, in the sense that the glasso algorithm tends to identify geographical neighbors as climate neighbors. This is particularly true in the tropics, where $f \geq 70\%$ (Fig 12, bottom and Fig. 13, top left). On the other hand, a more detailed comparison (Fig. 13) shows important differences: while the average distance to neighbors is clustered around 1,000 km in G_{1000} (by design), this metric shows tremendous spread in the glasso graph G (Fig. 13, top right), implying very local as well as long-range connections.

This has profound consequences for other measures of network topology. The local clustering coefficient C, which measures local interconnectedness [*Bollobás*, 1998; *Zerenner et al.*, 2014], is also very different between G and G_{1000} (Fig. 13, bottom left): while it is tightly clustered around 0.45 in G_{1000}, it ranges from 0.1 to 0.7 in G, with an average of 0.3. In both cases these measures are much larger than the local clustering coefficient of a random graph with identical degree (dashed lines). Similarly, the average shortest path length L [*Bollobás*, 1998], which measures how many steps are needed on average to get from one network node to another randomly picked node [*Zerenner et al.*, 2014], is much larger for G_{1000} than G (18 vs 7). Both are much larger than their random counterparts

(around 2 in both cases). Taken together, these large C and L metrics are indicative of small-world networks [*Watts and Strogatz*, 1998], but clearly G is able to capture much more long-range connections than G_{1000}.

This is illustrated in Figure 14. The top left panel shows the neighbors of a point in the Weddell Sea, its neighbors tightly clustered around it. Compare this graph to the top right panel, which shows the neighbors of a point in the eastern Pacific Cold Tongue. There the neighbors are seen to spread along the equatorial waveguide, as well as Baja California, connected by coastal Kelvin wave dynamics [e.g. *Cane*, 1984]. The bottom left panel shows the neighbors of a point in the California Current, closely hugging the western coast of the United States. The neighbors of a the point centered in Western Europe (bottom right) is very close to a distance-based neighborhood graph.

Finally, we investigate whether G appears to contain meaningful clusters, that is, a collection of regions where the vertices are well-connected within a region, but not among regions. We use spectral clustering [see e.g. *Von Luxburg*, 2007] to produce the clustering, and compare it to the same neighborhood graph as above (Fig. 15). Clusters were obtained using the spectral clustering implementation in scikit-learn [*Pedregosa et al.*, 2011] with default parameters, and 8 clusters. Because of its isotropy, the neighborhood graph looks the same from any vantage point, so spectral clustering returns a symmetric partition of the sphere. On the other hand, GLASSO delineates very different regions. In particular, it contrasts a North American cluster (light blue) with a Eurasian/North African cluster (cyan). In the southern hemisphere, the spectral clustering of G identifies 3 Antarctic clusters, one located over the East Antarctic ice sheet, one adjacent to the Ross sea, and one adjacent to the Weddell sea. Most world oceans form a single cluster (deep

blue), which also encompasses tropical landmasses; this likely reflects efficient spreading of climate information by Kelvin and Rossby waves within the tropical waveguide. It is remarkable that clustering the GLASSO graph extracts such climatically-relevant features from the data alone, without any other source of information than the number of clusters to retrieve. This suggests that graphs may be used to probe relationships in climate fields.

5. Discussion

We have applied Gaussian Markov random fields to the imputation of missing values in a leading surface temperature dataset, HadCRUT4.6. Gaussian graphical models allow flexible modeling of the covariance structure of climate fields, characterized by strong anisotropy resulting from wind or ocean currents, wave propagation patterns, land-ocean contrasts, or orography. This results in estimates of surface temperature that better capture the known structure of climate patterns (e.g. ENSO, Fig 10). They do so by encoding the conditional independence structure of the field, identifying which neighbors are most essential to infer the value of the field at one point (Markov property). By ignoring the vast majority (here, over 99%) of the others, they greatly reduce the number of parameters to estimate, thereby resulting in a well-conditioned covariance matrix, thus a well-posed estimation problem.

It is important to note that, while graph neighbors might also be geographical neighbors, this process can carry information along vast distances. An analogy might be useful: the celebrated autoregressive process of order 1 (AR(1)) is an example of a Markov process, where, by definition, the value of a series x_t is conditionally independent of all previous values but x_{t-1}. Yet, for large enough values of the autocorrelation parameter (which quantifies x_t's dependence on x_{t-1}), the process can take many time steps to "forget" a

past excursion. Similarly, Markov random fields may capture long-range teleconnections even for purely local graphs (e.g. G_{1000}). The principal advantage of discovering the concentration graph via GLASSO, instead of specifying it via distance-based measures, is therefore in allowing non-isotopic patterns to be pulled out of the data (Fig 14), something that distance-based methods at the heart of many kriging approaches cannot allow. At the same time, because concentration graphs tend to favor local information, these graphs better prevent the spread of errors across the field that global methods that rely on an eigendecomposition of the covariance matrix, where the leading modes (by definition, the most energetic ones) are also the most global, and are thus more vulnerable to observational errors. Methods based on concentration graphs, like GraphEM, therefore offer a useful middle-ground between local and global methods.

This raises another important point: graph-based methods need not be the purview of an EM-approach, and in fact, should not be. As recounted by *Schneider* [2006], the EM algorithm draws from the center of a distribution (the M in EM corresponds to maximizing the likelihood, which selects the mode), which is known to lower the variance of the dataset [*Little and Rubin*, 2002]. Unless explicit steps are taken to utilize the estimates of imputation error output by GraphEM, using only the central estimate will underestimate the true imputation error (this is, of course, in addition to all other sources of uncertainty, [*Kennedy*, 2013]). There would therefore be value in other inference frameworks that leverage graph-based covariance estimation. Bayesian hierarchical models [*Gelman et al.*, 2013, chapter 5] are one mechanism to accomplish this, and have the potential to more fully represent uncertainties [*Karspeck et al.*, 2012], including known biases and imputation error.

The framework of the Berkeley Earth Surface Temperature dataset [*Rohde et al.*, 2013], which currently uses some form of kriging, would be a natural candidate for this. Another realm is data assimilation [*Kalnay*, 1996; *Compo et al.*, 2011; *Hakim et al.*, 2016], where the need for regularizing rank-deficient covariance matrices also arises [*Gaspari and Cohn*, 2006]. In addition to atmosphere-ocean applications, another potential domain of application are geophysical inverse problems like seismic tomography [*Dębski*, 2009].

Beyond their use as regularizing tools, concentration graph models also appear useful in characterizing teleconnections within a geophysical field (section 4). This echoes previous work [*Tsonis et al.*, 2006; *Paluš et al.*, 2011; *Tupikina et al.*, 2016] and suggests that concentration graphs should be considered a *bona fide* analytical tool to be archived alongside the climate fields they served to complete, as done herein (see link to data repository below). One potential application of such graphs would be to characterize the similarity between climate networks simulated by general circulation models and those derived from observational products such as HadCRUT4. This may provide yet another benchmark for model evaluation, one focused on relationships within a field or between fields (e.g. temperature, pressure and precipitation).

Finally, the results shown here (Section 3) testify once again to the large differences in basic climate metrics (e.g. GMT trends) arising from various interpolation methods, or lack thereof. It is therefore particularly important that centers that generate major surface temperature datasets utilize the best available statistical methods to obtain complete climate fields, which tends to be what users want. The present work offers one such approach, and we hope it will stimulate further research and applications in this area.

Acknowledgments. J.E.G acknowledges support from grants AGS 1003818 and DMS 1025465 from the U.S National Science Foundation. J.K. and C.M. were supported by the Met Office Hadley Centre Climate Programme funded by BEIS and Defra. Code is freely available at github.com/advaccaro/hadcrut4.6-graphem. Data will be made available at ... upon publication.

References

Bollobás, B. (1998), Random graphs, in *Modern graph theory*, pp. 215–252, Springer.

Bosilovich, M., R. Lucchesi, and M. Suarez (2015), Merra-2: File specification.

Brohan, P., J. J. Kennedy, I. Harris, S. F. Tett, and P. D. Jones (2006), Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850, *Journal of Geophysical Research: Atmospheres (1984–2012)*, *111*(D12).

Bunge, L., and A. J. Clarke (2009), A verified estimation of the el niño index niño-3.4 since 1877, *Journal of Climate*, *22*(14), 3979–3992.

Cane, M. A. (1984), Modeling Sea-Level During El Niño, *J. Phys. Oceanogr.*, *14*(12), 1864–1874.

Christy, J. R., W. B. Norris, R. W. Spencer, and J. J. Hnilo (2007), Tropospheric temperature change since 1979 from tropical radiosonde and satellite measurements, *Journal of Geophysical Research: Atmospheres (1984–2012)*, *112*(D6).

Compo, G. P., 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.

532 Woodruff, and S. J. Worley (2011), The twentieth century reanalysis project, *Quarterly*
533 *Journal of the Royal Meteorological Society*, *137*(654), 1–28, doi:10.1002/qj.776.
534 Cowtan, K., and R. G. Way (2014), Coverage bias in the hadcrut4 temperature series and
535 its impact on recent temperature trends, *Quarterly Journal of the Royal Meteorological*
536 *Society*, *140*(683), 1935–1944.
537 Craven, P., and G. Wahba (1979), Smoothing noisy data with spline functions: esti-
538 mating the correct degree of smoothing by the method of generalized cross-validation,
539 *Numerische Mathematik*, *31*, 377–403.
540 Cressie, N. (1990), The origins of kriging, *Mathematical Geology*, *22*(3), 239–252, doi:
541 10.1007/BF00889887.
542 Davis, M. (2001), *Late Victorian Holocausts: El Niño Famines and the Making of the*
543 *Third World*, 464 pp., Verso, New York.
544 Dębski, W. (2009), Seismic tomography by monte carlo sampling, *Pure and Applied Geo-*
545 *physics*.
546 Dempster, A. P. (1972), Covariance selection, *Biometrics*, pp. 157–175.
547 Dempster, A. P., N. M. Laird, and D. B. Rubin (1977), Maximum likelihood estimation
548 from incomplete data via the EM algorithm (with discussion), *J. Roy. Stat. Soc. B.*, *39*,
549 1–38.
550 Easterling, D. R., and M. F. Wehner (2009), Is the climate warming or cooling?, *Geo-*
551 *physical Research Letters*, *36*(8).
552 Emile-Geay, J., K. Cobb, M. Mann, and A. T. Wittenberg (2013a), Estimating Central
553 Equatorial Pacific SST variability over the Past Millennium. Part 1: Methodology and
554 Validation, *J. Clim.*, *26*, 2302–2328, doi:10.1175/JCLI-D-11-00510.1.

555 Emile-Geay, J., K. Cobb, M. Mann, and A. T. Wittenberg (2013b), Estimating Central
556 Equatorial Pacific SST variability over the Past Millennium. Part 2: Reconstructions
557 and Implications, *J. Clim.*, *26*, 2329–2352, doi:10.1175/JCLI-D-11-00511.1.

558 Flato, G., J. Marotzke, B. Abiodun, P. Braconnot, S. Chou, W. Collins, P. Cox, F. Dri-
559 ouech, S. Emori, V. Eyring, C. Forest, P. Gleckler, E. Guilyardi, C. Jakob, V. Kattsov,
560 C. Reason, and M. Rummukainen (2013), Evaluation of Climate Models, in *Climate*
561 *Change 2013: The Physical Science Basis. Contribution of Working Group I to the*
562 *Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by
563 T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. Allen, J. Boschung, A. Nauels,
564 Y. Xia, V. Bex, and P. Midgley, pp. 741–866, Cambridge University Press, Cambridge,
565 United Kingdom and New York, NY, USA, doi:10.1017/CBO9781107415324.020.

566 Folland, C., and D. Parker (1995), Correction of instrumental biases in historical sea sur-
567 face temperature data, *Quarterly Journal of the Royal Meteorological Society*, *121*(522),
568 319–367.

569 Folland, C. K., A. W. Colman, D. M. Smith, O. Boucher, D. E. Parker, and J.-P. Vernier
570 (2013), High predictive skill of global surface temperature a year ahead, *Geophysical*
571 *Research Letters*, *40*(4), 761–767.

572 Friedman, J., T. Hastie, and R. Tibshirani (2008), Sparse inverse covariance estimation
573 with the graphical lasso, *Biostatistics*, *9*(3), 432–441.

574 Gaspari, G., and S. E. Cohn (2006), Construction of correlation functions in two and three
575 dimensions, *Quarterly Journal of the Royal Meteorological Society*, *125*(554), 723–757,
576 doi:10.1002/qj.49712555417.

577 Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2013), *Bayesian Data Analysis*,
578 2nd ed., 675 pp., Chapman and Hall, New York, NY.

579 Golub, G. H., M. Heath, and G. Wahba (1979), Generalized cross-validation as a method
580 for choosing a good ridge parameter, *Technometrics*, *21*(2), 215–223.

581 Guillot, D., B. Rajaratnam, J. Emile-Geay, et al. (2015), Statistical paleoclimate recon-
582 structions via markov random fields, *The Annals of Applied Statistics*, *9*(1), 324–352.

583 Hakim, G. J., J. Emile-Geay, E. J. Steig, D. Noone, D. M. Anderson, R. Tardif, N. Steiger,
584 and W. A. Perkins (2016), The last millennium climate reanalysis project: Framework
585 and first results, *Journal of Geophysical Research: Atmospheres*, *121*, 6745 – 6764,
586 doi:10.1002/2016JD024751.

587 Hansen, J., and S. Lebedeff (1987), Global trends of measured surface air temperature,
588 *Journal of Geophysical Research*, *92*, 13,345–13,372, doi:10.1029/JD092iD11p13345.

589 Hansen, J., M. Sato, R. Ruedy, K. Lo, D. W. Lea, and M. Medina-Elizade (2006), Global
590 temperature change, *Proceedings of the National Academy of Sciences*, *103*(39), 14,288–
591 14,293.

592 Hansen, J., R. Ruedy, M. Sato, and K. Lo (2010), Global surface temperature change,
593 *Reviews of Geophysics*, *48*(4).

594 Hausfather, Z., M. J. Menne, C. N. Williams, T. Masters, R. Broberg, and D. Jones (2013),
595 Quantifying the effect of urbanization on us historical climatology network temperature
596 records, *Journal of Geophysical Research: Atmospheres*, *118*(2), 481–494.

597 Hausfather, Z., K. Cowtan, D. C. Clarke, P. Jacobs, M. Richardson, and R. Rohde (2017),
598 Assessing recent warming using instrumentally homogeneous sea surface temperature
599 records, *Science Advances*, *3*(1), doi:10.1126/sciadv.1601207.

Hoerl, A. E., and R. W. Kennard (1970a), Ridge regression: Biased estimation for non-orthogonal problems, *Technometrics*, *12*, 55–67.

Hoerl, A. E., and R. W. Kennard (1970b), Ridge regression: Applications to non-orthogonal problems, *Technometrics*, *12*, 69–82, correction, **12**, 723.

Horel, J., and J. Wallace (1981), Planetary-scale atmospheric phenomena associated with the Southern Oscillation, *Mon. Weather Rev.*, *109*, 814–829.

Huang, B., V. F. Banzon, E. Freeman, J. Lawrimore, W. Liu, T. C. Peterson, T. M. Smith, P. W. Thorne, S. D. Woodruff, and H.-M. Zhang (2015), Extended Reconstructed Sea Surface Temperature Version 4 (ERSST. v4). Part I: Upgrades and Intercomparisons, *Journal of Climate*, *28*(3), 911–930.

Huang, J., X. Zhang, Q. Zhang, Y. Lin, M. Hao, Y. Luo, Z. Zhao, Y. Yao, X. Chen, L. Wang, S. Nie, Y. Yin, Y. Xu, and J. Zhang (2017), Recently amplified arctic warming has contributed to a continual global warming trend, *Nature Climate Change*, *7*(12), 875–879, doi:10.1038/s41558-017-0009-5.

Huang, M. L. J. L. C. L. H. Z. V. B. Z. H., B., and A. Kumar (2013), Why did large differences arise in the sea surface temperature datasets across the tropical pacific during 2012?, *J. Atmos. Oceanic. Technol.*, *30*, 2944–2953, doi: https://doi.org/10.1175/JTECH-D-13-00034.1.

Ilyas, M., C. M. Brierley, and S. Guillas (2017), Uncertainty in regional temperatures inferred from sparse global observations: Application to a probabilistic classification of El Niño, *Geophysical Research Letters*, *44*(17), 9068–9074, doi:10.1002/2017GL074596.

Ishii, M., A. Shouji, S. Sugimoto, and T. Matsumoto (2005), Objective analyses of sea-surface temperature and marine meteorological variables for the 20th century using

623 icoads and the kobe collection, *International Journal of Climatology*, *25*(7), 865–879.

624 Johnstone, I. (2001), On the distribution of the largest eigenvalue in principal components
625 analysis, *Annals of Statistics*, *29*, 295–327.

626 Jones, P. D., and A. Moberg (2003), Hemispheric and Large-Scale Surface Air Temper-
627 ature Variations: An Extensive Revision and an Update to 2001., *Journal of Climate*,
628 *16*, 206–223, doi:10.1175/1520-0442(2003)016<0206:HALSSA>2.0.CO;2.

629 Jones, P. D., D. H. Lister, T. J. Osborn, C. Harpham, M. Salmon, and C. P. Morice
630 (2012), Hemispheric and large-scale land-surface air temperature variations: An exten-
631 sive revision and an update to 2010, *Journal of Geophysical Research: Atmospheres*,
632 *117*(D5).

633 Kalnay, E. & coauthors. (1996), The NCEP/NCAR 40-year Reanalysis Project, *Bull.*
634 *Amer. Meteor. Soc.*, *77*, 437–471.

635 Kaplan, A., Y. Kushnir, M. A. Cane, and M. B. Blumenthal (1997), Reduced space
636 optimal analysis for historical data sets: 136 years of Atlantic sea surface temperatures,
637 *J. Geophys. Res. - Oceans*, *102*(C13), 27,835–27,860.

638 Kaplan, A., M. A. Cane, Y. Kushnir, A. C. Clement, M. B. Blumenthal, and B. Ra-
639 jagopalan (1998), Analyses of global sea surface temperature 1856-1991, *Journal of*
640 *Geophysical Research*, *103*(18), 567–18.

641 Kaplan, A., Y. Kushnir, and M. Cane (2000), Reduced space optimal interpolation of
642 historical marine sea level pressure: 1854-1992, *J. Climate*, *13*(16).

643 Kaplan, A., M. A. Cane, and Y. Kushnir (2003), *Reduced space approach to the optimal*
644 *analysis interpolation of historical marine observations: Accomplishments, difficulties,*
645 *and prospects*, pp. 199–216, World Meteorological Organization, Geneva, Switzerland.

Karl, T., S. Hassol, C. Miller, and W. Murray (2006), Temperature trends in the lower atmosphere. steps for understanding and reconciling differences, in *A Report by the US Climate Change Science Program and the Subcommittee on Global Change Research*, National Oceanic and Atmospheric Administration.

Karl, T. R., A. Arguez, B. Huang, J. H. Lawrimore, J. R. McMahon, M. J. Menne, T. C. Peterson, R. S. Vose, and H.-M. Zhang (2015), Possible artifacts of data biases in the recent global surface warming hiatus, *Science*, *348*(6242), 1469–1472.

Karspeck, A. R., A. Kaplan, and S. R. Sain (2012), Bayesian modelling and ensemble reconstruction of mid-scale spatial variability in north atlantic sea-surface temperatures for 1850–2008, *Quarterly Journal of the Royal Meteorological Society*, *138*(662), 234–248, doi:10.1002/qj.900.

Kennedy, J., N. Rayner, R. Smith, D. Parker, and M. Saunby (2011), Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 2. biases and homogenization, *Journal of Geophysical Research: Atmospheres (1984–2012)*, *116*(D14).

Kennedy, J. J. (2013), A review of uncertainty in in situ measurements and data sets of sea surface temperature, *Reviews of Geophysics*, *52*(1), 1–32, doi:10.1002/2013RG000434.

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*(D14), doi:10.1029/2010JD015218.

Kennedy, J. J., N. A. Rayner, R. O. Smith, D. E. Parker, and M. Saunby (2011b), Reassessing biases and other uncertainties in sea surface temperature observations mea-

669 sured in situ since 1850: 2. Biases and homogenization, *Journal of Geophysical Research*
670 *(Atmospheres)*, *116*, D14104, doi:10.1029/2010JD015220.

671 Kent, E. C., J. J. Kennedy, T. M. Smith, S. Hirahara, B. Huang, A. Kaplan, D. E. Parker,
672 C. P. Atkinson, D. I. Berry, G. Carella, Y. Fukuda, M. Ishii, P. D. Jones, F. Lindgren,
673 C. J. Merchant, S. Morak-Bozzo, N. A. Rayner, V. Venema, S. Yasui, and H.-M. Zhang
674 (2016), A call for new approaches to quantifying biases in observations of sea surface
675 temperature, *Bulletin of the American Meteorological Society*, *98*(8), 1601–1616, doi:
676 10.1175/BAMS-D-15-00251.1.

677 Lauritzen, S. L. (1996), *Graphical Models*, Oxford: Clarendon Press.

678 Little, R. J. A., and D. B. Rubin (2002), *Statistical analysis with missing data*, Wiley
679 series in probability and statistics, New York, NY.

680 Liu, W., B. Huang, P. W. Thorne, V. F. Banzon, H.-M. Zhang, E. Freeman, J. Lawrimore,
681 T. C. Peterson, T. M. Smith, and S. D. Woodruff (2015), Extended reconstructed sea
682 surface temperature version 4 (ersst. v4): Part ii. parametric and structural uncertainty
683 estimations, *Journal of Climate*, *28*(3), 931–951.

684 Mears, C. A., and F. J. Wentz (2005), The effect of diurnal correction on
685 satellite-derived lower tropospheric temperature, *Science*, *309*(5740), 1548–1551, doi:
686 10.1126/science.1114772.

687 Morice, C. P., J. J. Kennedy, N. A. Rayner, and P. D. Jones (2012), Quantifying uncer-
688 tainties in global and regional temperature change using an ensemble of observational
689 estimates: The HadCRUT4 data set, *Journal of Geophysical Research: Atmospheres*
690 *(1984–2012)*, *117*(D8).

Nychka, D., S. Bandyopadhyay, D. Hammerling, F. Lindgren, and S. Sain (2015), A multiresolution gaussian process model for the analysis of large spatial datasets, *Journal of Computational and Graphical Statistics*, *24*(2), 579–599.

Paluš, M., D. Hartman, J. Hlinka, and M. Vejmelka (2011), Discerning connectivity from dynamics in climate networks, *Nonlinear Processes in Geophysics*, *18*(5), 751–763, doi: 10.5194/npg-18-751-2011.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011), Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research*, *12*, 2825–2830.

Quinn, W. H. (1992), Large-scale ENSO events, the El Niño and other important regional features, in *Registro del fenómeno El Niño y de eventos ENSO en América del Sur*, vol. 22, edited by L. Macharé, José; Ortlieb, pp. 13–22, Institut Fran cais d'Etudes Andines, Lima.

Reynolds, R. W., and T. M. Smith (1994), Improved Global Sea Surface Temperature Analyses Using Optimum Interpolation., *J. Climate*, *7*, 929–948.

Reynolds, R. W., N. A. Rayner, T. M. Smith, D. C. Stokes, and W. Wang (2002), An improved in situ and satellite sst analysis for climate, *Journal of climate*, *15*(13), 1609–1625.

Rohde, R., R. Muller, R. Jacobsen, S. Perlmutter, A. Rosenfeld, J. Wurtele, J. Curry, C. Wickham, and S. Mosher (2013), Berkeley earth temperature averaging process, *Geoinfor. Geostat.: An Overview*, *1*(2), 1–13.

Sarachik, E. S., and M. A. Cane (2010), *The El Niño-Southern Oscillation Phenomenon*, 384 pp., Cambridge University Press, Cambridge, UK.

Sardeshmukh, P., and B. Hoskins (1988), The generation of global rotational flow by steady idealized tropical divergence, *J. Atmos. Sci.*, *45*, 122–1251.

Schneider, T. (2001), Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values, *Journal of Climate*, *14*(5), 853–871.

Schneider, T. (2006), Analysis of incomplete data: Readings from the statistics literature, *Bulletin of the American Meteorological Society*, *87*(1410–1411).

Simmons, A., P. Jones, V. da Costa Bechtold, A. Beljaars, P. Kållberg, S. Saarinen, S. Uppala, P. Viterbo, and N. Wedi (2004), Comparison of trends and low-frequency variability in cru, era-40, and ncep/ncar analyses of surface air temperature, *Journal of Geophysical Research: Atmospheres (1984–2012)*, *109*(D24).

Simmons, A., K. Willett, P. Jones, P. Thorne, and D. Dee (2010), Low-frequency variations in surface atmospheric humidity, temperature, and precipitation: Inferences from reanalyses and monthly gridded observational data sets, *Journal of Geophysical Research: Atmospheres (1984–2012)*, *115*(D1).

Simmons, A. J., J. M. Wallace, and G. W. Branstator (1983), Barotropic Wave Propagation and Instability, and Atmospheric Teleconnection Patterns., *J. Atmos. Sci.*, *40*, 1363–1392.

Smith, T., R. Reynolds, T. C. Peterson, and J. Lawrimore (2008), Improvements to NOAA's historical merged land-ocean surface temperature analysis (1880-2006), *J. Clim.*, *21*, 2283–2296.

Smith, T. M., R. W. Reynolds, R. E. Livezey, and D. C. Stokes (1996), Reconstruction of historical sea surface temperatures using empirical orthogonal functions, *Journal of Climate*, *9*(6), 1403–1420.

Spencer, R. W. (1990), Precise monitoring of global temperature trends, *Science*, *247*, 1558–1558.

Stocker, T., D. Qin, G.-K. Plattner, L. Alexander, S. Allen, N. Bindoff, F.-M. Bréon, J. Church, U. Cubasch, S. Emori, P. Forster, P. Friedlingstein, N. Gillett, J. Gregory, D. Hartmann, E. Jansen, B. Kirtman, R. Knutti, K. Krishna Kumar, P. Lemke, J. Marotzke, V. Masson-Delmotte, G. Meehl, I. Mokhov, S. Piao, V. Ramaswamy, D. Randall, M. Rhein, M. Rojas, C. Sabine, D. Shindell, L. Talley, D. Vaughan, and S.-P. Xie (2013), *Technical Summary*, book section TS, p. 33âĂŞ115, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, doi: 10.1017/CBO9781107415324.005.

Thorne, P. W., J. R. Lanzante, T. C. Peterson, D. J. Seidel, and K. P. Shine (2010), Tropospheric temperature trends: history of an ongoing controversy, *Wiley Interdisciplinary Reviews: Climate Change*, *2*(1), 66–88, doi:10.1002/wcc.80.

Tikhonov, A. N., and V. Y. Arsenin (1977), *Solution of Ill-Posed Problems*, Scripta Series in Mathematics, 258 pp., V. H. Winston and Sons, Washington.

Tingley, M. P., and P. Huybers (2010), A Bayesian Algorithm for Reconstructing Climate Anomalies in Space and Time. Part 1: Development and applications to paleoclimate reconstruction problems, *J. Clim.*, *23*, 2759–2781, doi:10.1175/2009JCLI3016.1.

Tingley, M. P., P. F. Craigmile, M. Haran, B. Li, E. Mannshardt, and B. Rajaratnam (2012), Piecing together the past: statistical insights into paleoclimatic reconstructions,

Quaternary Science Reviews, *35*(0), 1 – 22, doi:10.1016/j.quascirev.2012.01.012.

Trenberth, K. E. (1997), The Definition of El Niño, *Bull. Amer. Met. Soc.*, *78(12)*, 2771–2777.

Trenberth, K. E., and D. P. Stepaniak (2001), Indices of El Niño Evolution, *Journal of Climate*, *14*(8), 1697–1701, doi:10.1175/1520-0442(2001)014<1697:LIOENO>2.0.CO;2.

Trenberth, K. E., G. W. Branstator, D. Karoly, A. Kumar, N.-C. Lau, and C. Ropelewski (1998), Progress during TOGA in understanding and modeling global teleconnections associated with tropical sea surface temperatures, *J. Geophys. Res.*, *103*, 14,291–14,324, doi:10.1029/97JC01444.

Tsonis, A. A., K. L. Swanson, and P. J. Roebber (2006), What do networks have to do with climate?, *Bulletin of the American Meteorological Society*, *87*(5), 585–596, doi:10.1175/BAMS-87-5-585.

Tupikina, L., N. Molkenthin, C. López, E. Hernández-García, N. Marwan, and J. Kurths (2016), Correlation networks from flows. the case of forced and time-dependent advection-diffusion dynamics, *PLOS ONE*, *11*(4), e0153,703–.

Von Luxburg, U. (2007), A tutorial on spectral clustering, *Statistics and computing*, *17*(4), 395–416.

Vose, R. S., D. Arndt, V. F. Banzon, D. R. Easterling, B. Gleason, B. Huang, E. Kearns, J. H. Lawrimore, M. J. Menne, T. C. Peterson, et al. (2012), Noaa's merged land-ocean surface temperature analysis, *Bulletin of the American Meteorological Society*, *93*(11), 1677–1685.

Wang, J., J. Emile-Geay, J. E. Smerdon, D. Guillot, and B. Rajaratnam (2014), Evaluating climate field reconstruction techniques using improved emulations of real-world

conditions, *Climate of the Past*, *10*(1), 1–19.

Wang, J., J. Emile-Geay, D. Guillot, N. P. McKay, and B. Rajaratnam (2015), Fragility of reconstructed temperature patterns over the common era: Implications for model evaluation., *Geophysical Research Letters*, *42*, 7162–7170, doi:10.1002/2015GL065265.

Watts, D. J., and S. H. Strogatz (1998), Collective dynamics of "small-world" networks, *Nature*, *393*, 440 EP –, doi:10.1038/30918.

Williams, C. N., M. J. Menne, and P. W. Thorne (2012), Benchmarking the performance of pairwise homogenization of surface temperatures in the united states, *Journal of Geophysical Research: Atmospheres (1984–2012)*, *117*(D5).

Zerenner, T., P. Friederichs, K. Lehnertz, and A. Hense (2014), A gaussian graphical model approach to climate networks, *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *24*(2), 023,103, doi:10.1063/1.4870402.

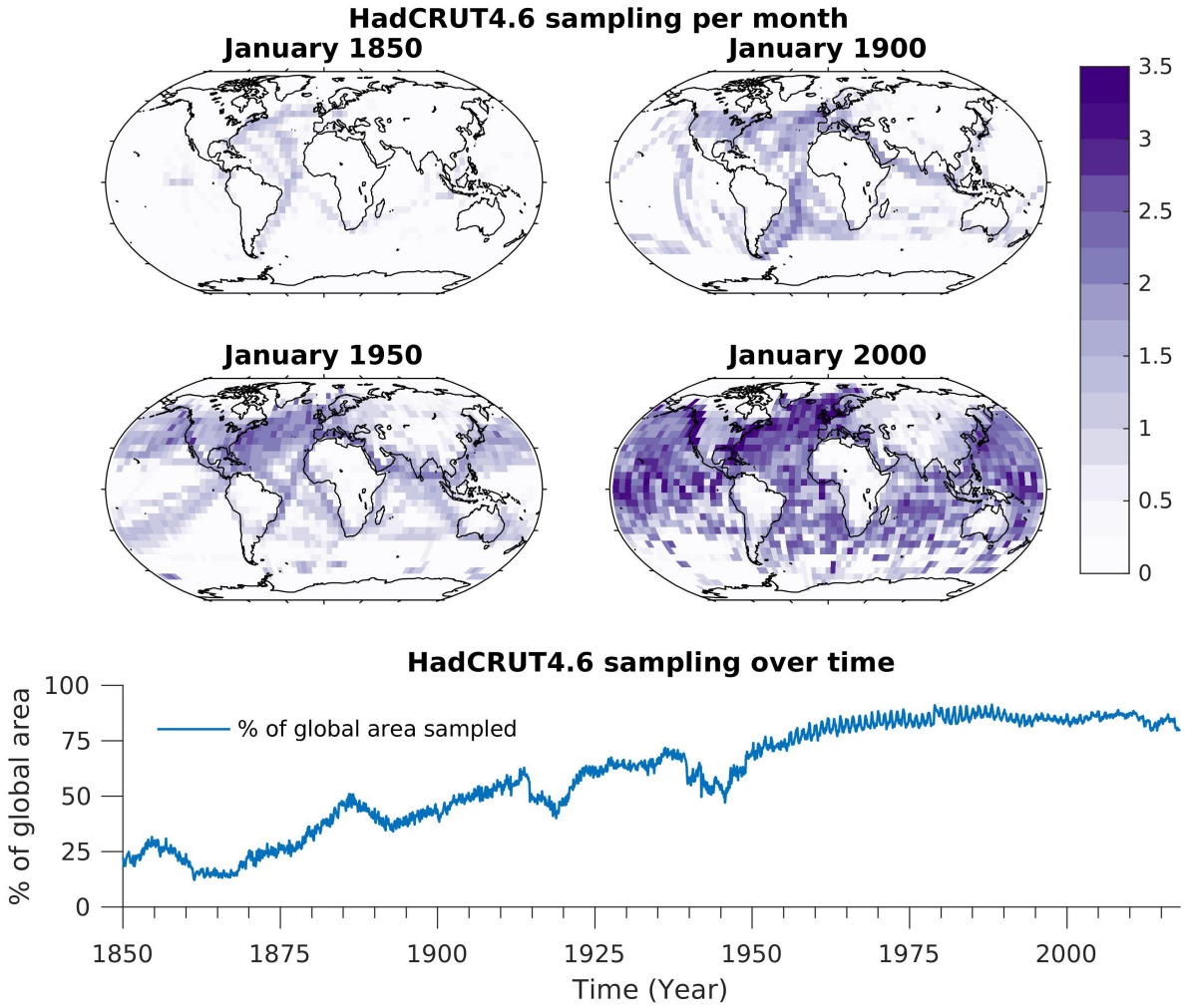

Figure 1. **(Top)** Spatial distribution of the number of available monthly observations over marine grid boxes and the number of weather station reports over land cells for January 1850, 1900, 1950, and 2000. Colors are based on a logarithmic scale (base 10). **(Bottom)** Total number of monthly observations available across all grid cells over time. From about 1950 to 1980, a clear seasonal cycle is observed.

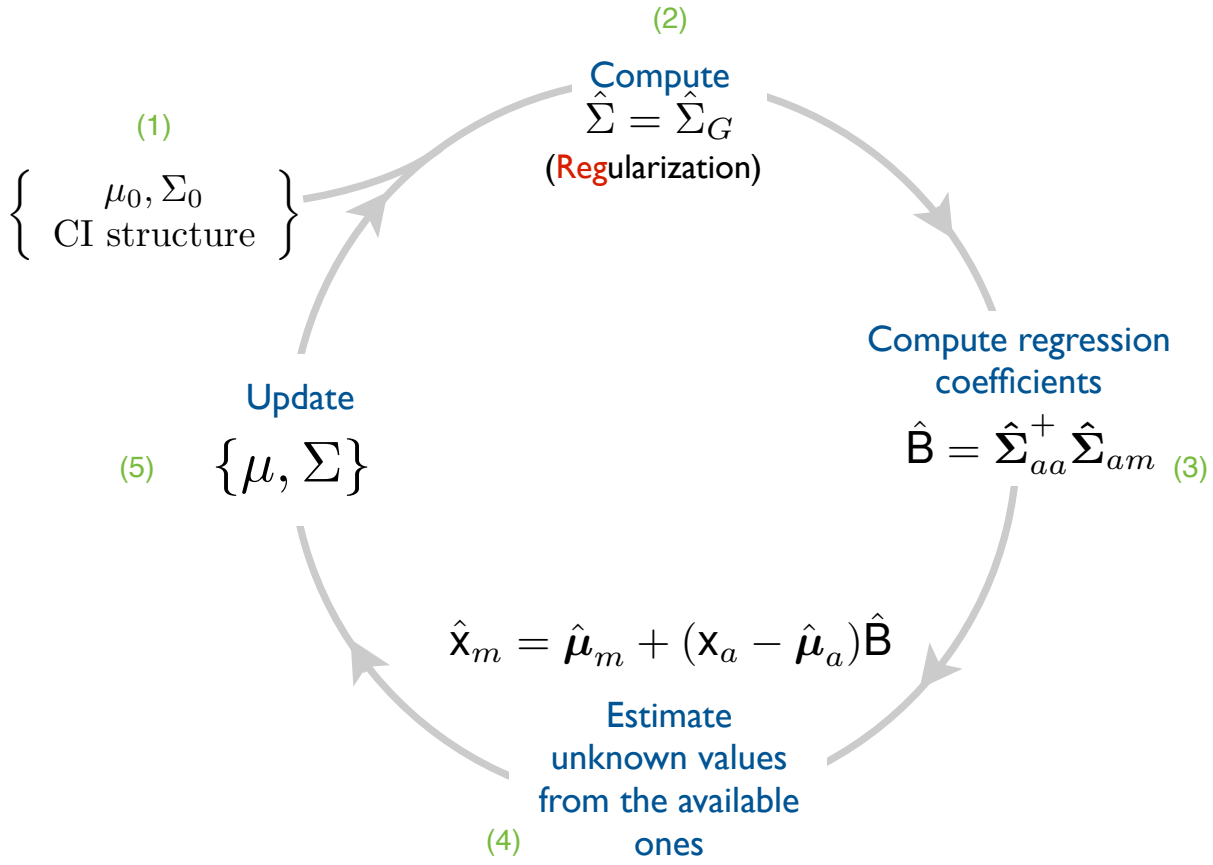

Figure 2. Schematic of the GraphEM algorithm. Starting from an initial guess of the mean and covariance, as well as some information about the conditional independence structure (i.e. the concentration graph G), the algorithm first computes a graphical estimate of the covariance matrix, uses it to compute regression coefficients, impute the missing values, then uses the latter to update the mean and covariance of the dataset. The algorithm proceeds until convergence, which is always guaranteed for a fixed graph.

Table 1. Surface temperature datasets used in this study.

L+O: land and ocean. O: ocean only.

Dataset	Reference	Period	Spatial coverage	Resolution
HadCRUT4.6	*Morice et al.* [2012]	Jan 1850 – Dec 2017	global L+O	5° x 5°
HadCRUT4.6$_{CW}$	*Cowtan and Way* [2014]	Jan 1850 – Dec 2017	global L+O	5° x 5°
HadCRUT4.6$_{G}$	this study	Jan 1850 – Dec 2017	global L+O	5° x 5°
GISTEMP	*Hansen et al.* [2010]	Jan 1880 – Feb 2018	global L+O	2° x 2°
NOAAGlobalTemp	*Vose et al.* [2012]	Jan 1880 – Apr 2018	global L+O	5° x 5°
COBE SST	*Ishii et al.* [2005]	Jan 1850 –Dec 2016	global O	1° x 1°
ERSSTv5	*Huang et al.* [2017]	Mar 1854 – Feb 2018	global O	2° x 2°
Kaplan extended	*Kaplan et al.* [1998]	Jan 1856 – Aug 2015	NINO3.4 region	n/a
Bunge & Clarke	*Bunge and Clarke* [2009]	Jan 1873 – Mar 2008	NINO3.4 region	n/a

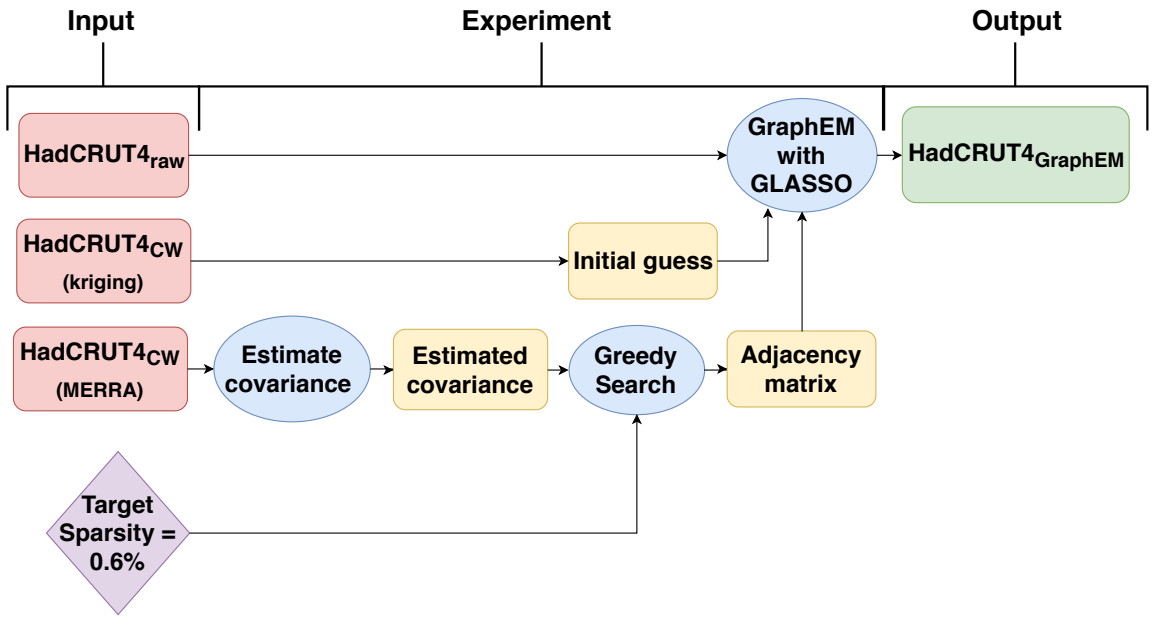

Figure 3. Schematic showing the interpolation workflow used in this study. First, the HadCRUT4$_{CW(MERRA)}$ data are used to estimate the covariance matrix of the temperature field. Then, using a target sparsity of 0.6%, the adjacency matrix, or graph, of the field is obtained from a greedy search algorithm. The adjacency matrix is then incorporated into GraphEM with GLASSO algorithm to infill the missing values in the raw HadCRUT4 data, using HadCRUT4$_{CW(kriging)}$ as an initial guess.

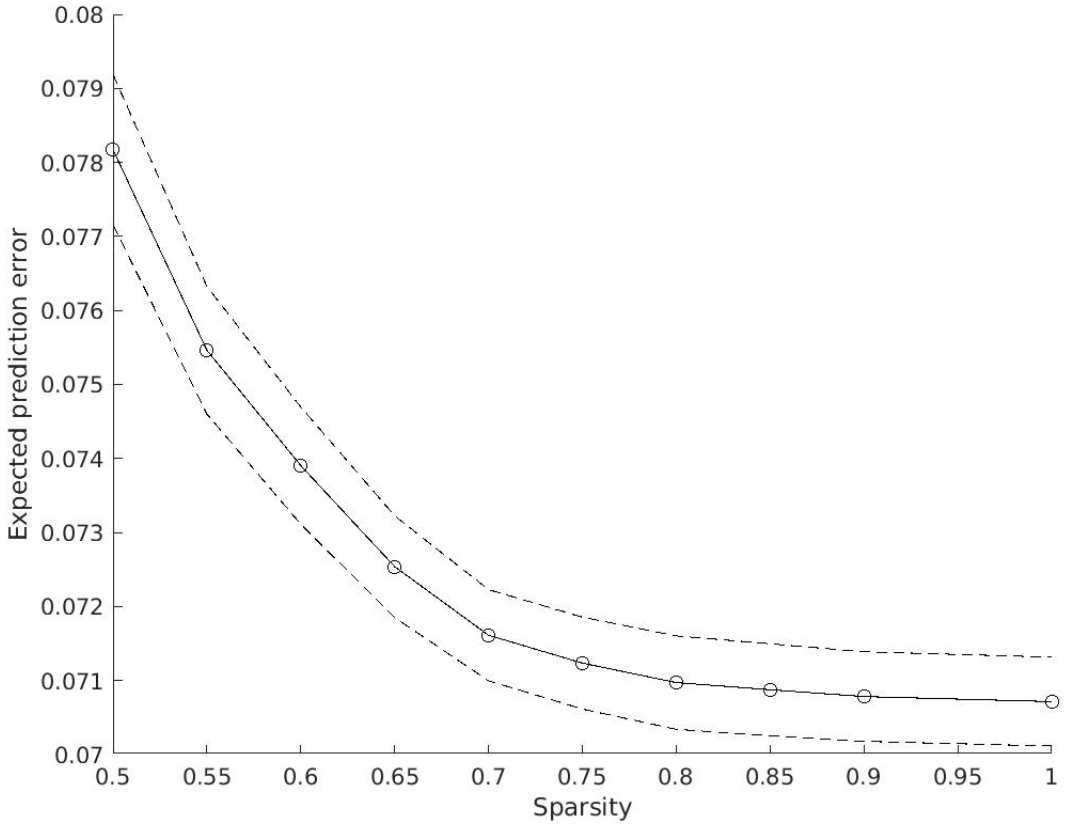

Figure 4. Cross-validation scores for different target sparsities when infilling the HadCRUT4.6 median.

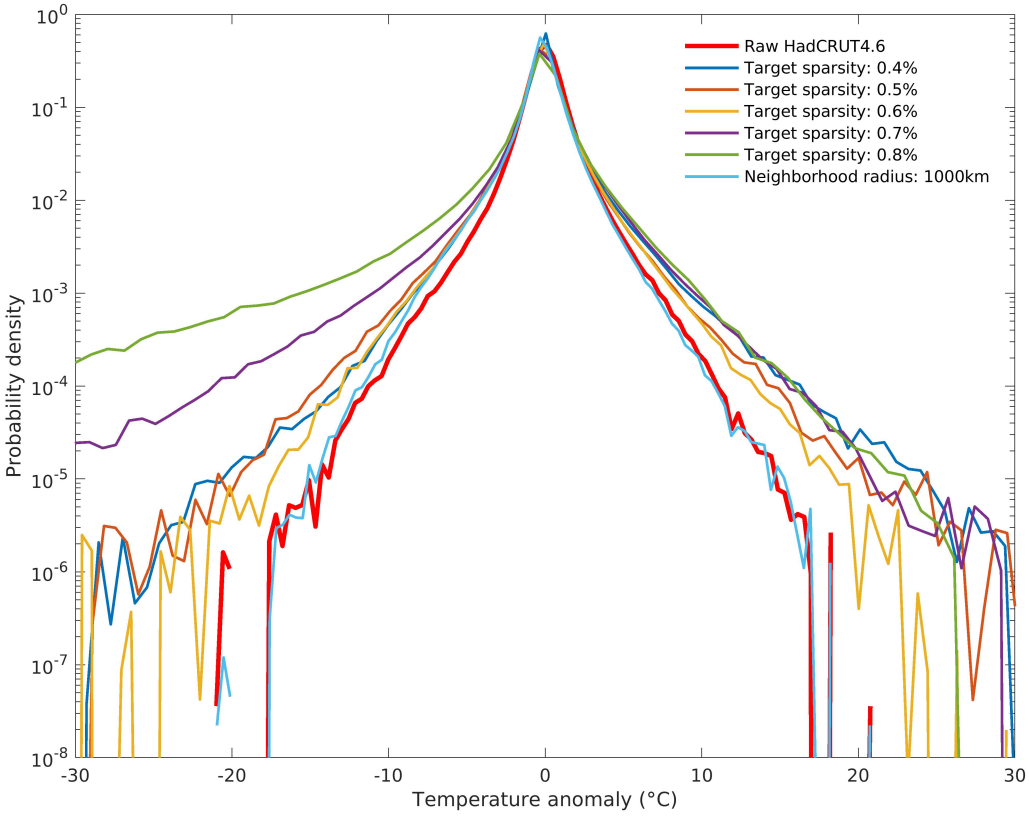

Figure 5. Kernel density estimates of probability density plotted on a logarithmic scale for the raw HadCRUT4.6 median, GraphEM with GLASSO products at various target sparsities, and GraphEM neighborhood graph with 1000 km radius.

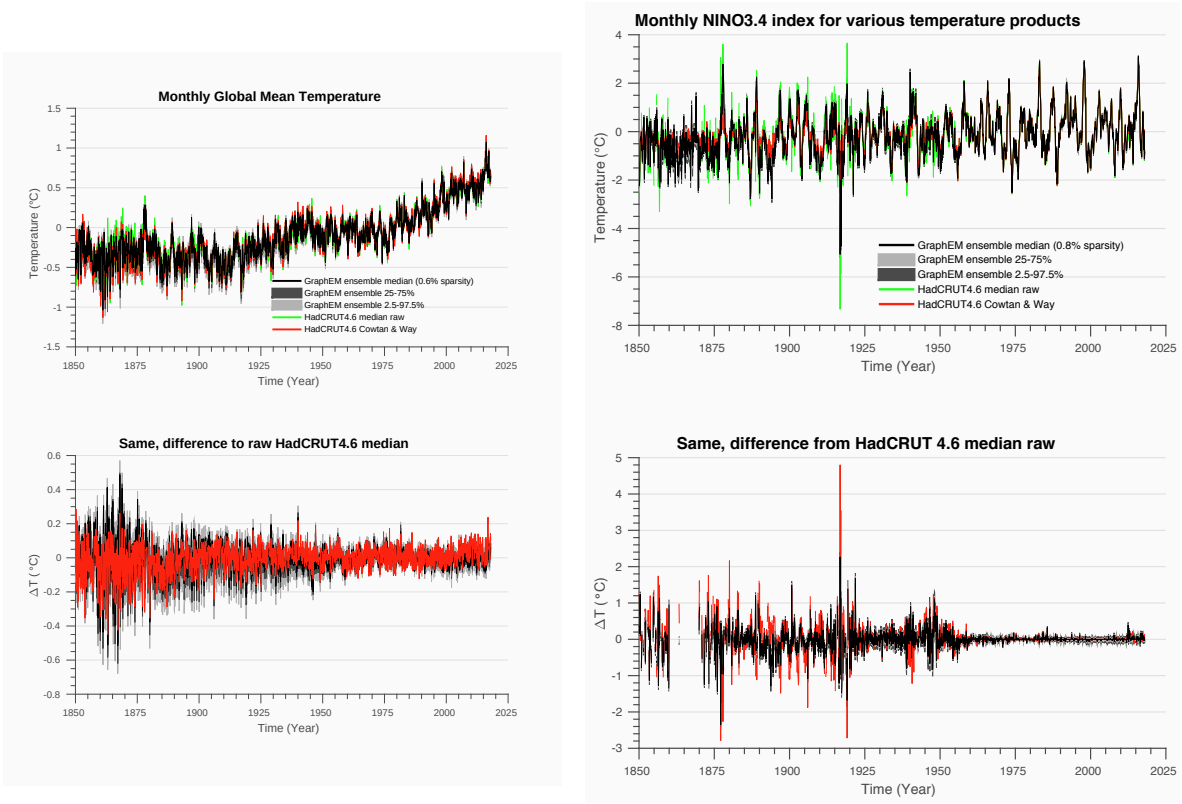

Figure 6. GMT (left) and NINO3.4 (right) in HadCRUT4 variants. Central and extreme quantiles of the: (Top) raw HadCRUT4.6 median (green), CW14 HadCRUT4.6 (red), and HadCRUT4$_G$ (black). (Bottom) Same, plotted as difference to the raw HadCRUT4.6 median.

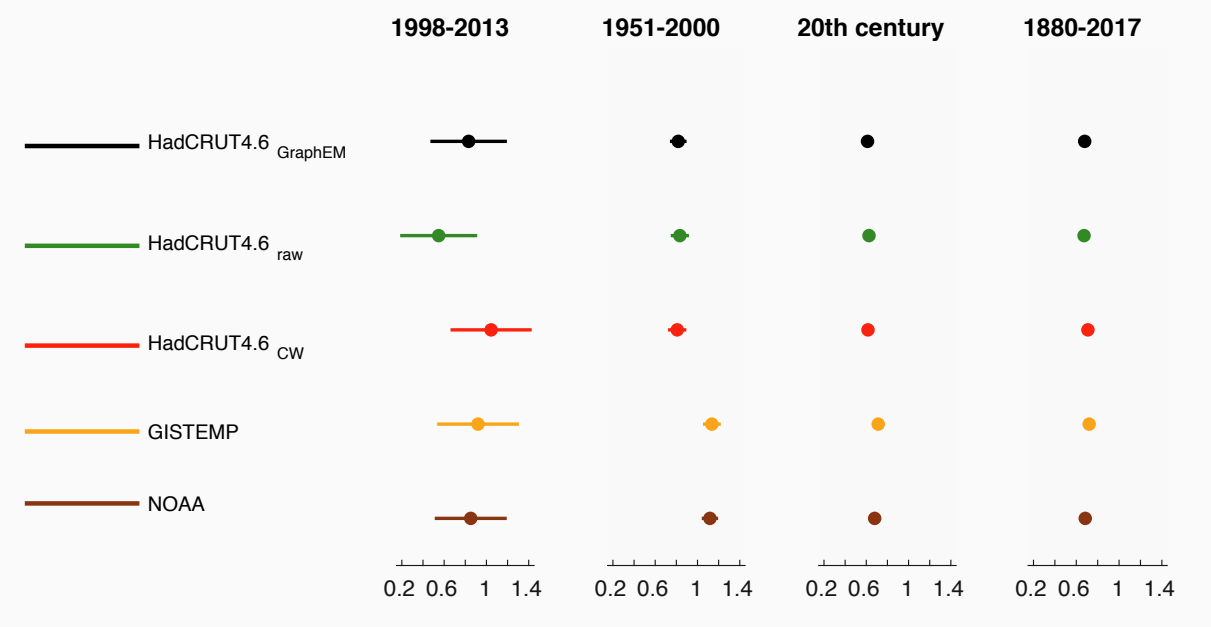

Figure 7. Global mean temperature trends (°C/century). Linear trends are calculated as ordinary least squares fits over the median of each dataset over 4 intervals: (1) the 1998-2013 "hiatus" period, (2) the late twentieth century (1951-2000), (3) the twentieth century, and (4) the longest common period (1880-2017). A 95% confidence interval is represented by lines extending outward from the central dots. In the two rightmost columns, these intervals are narrower than the central dot, hence not visible.

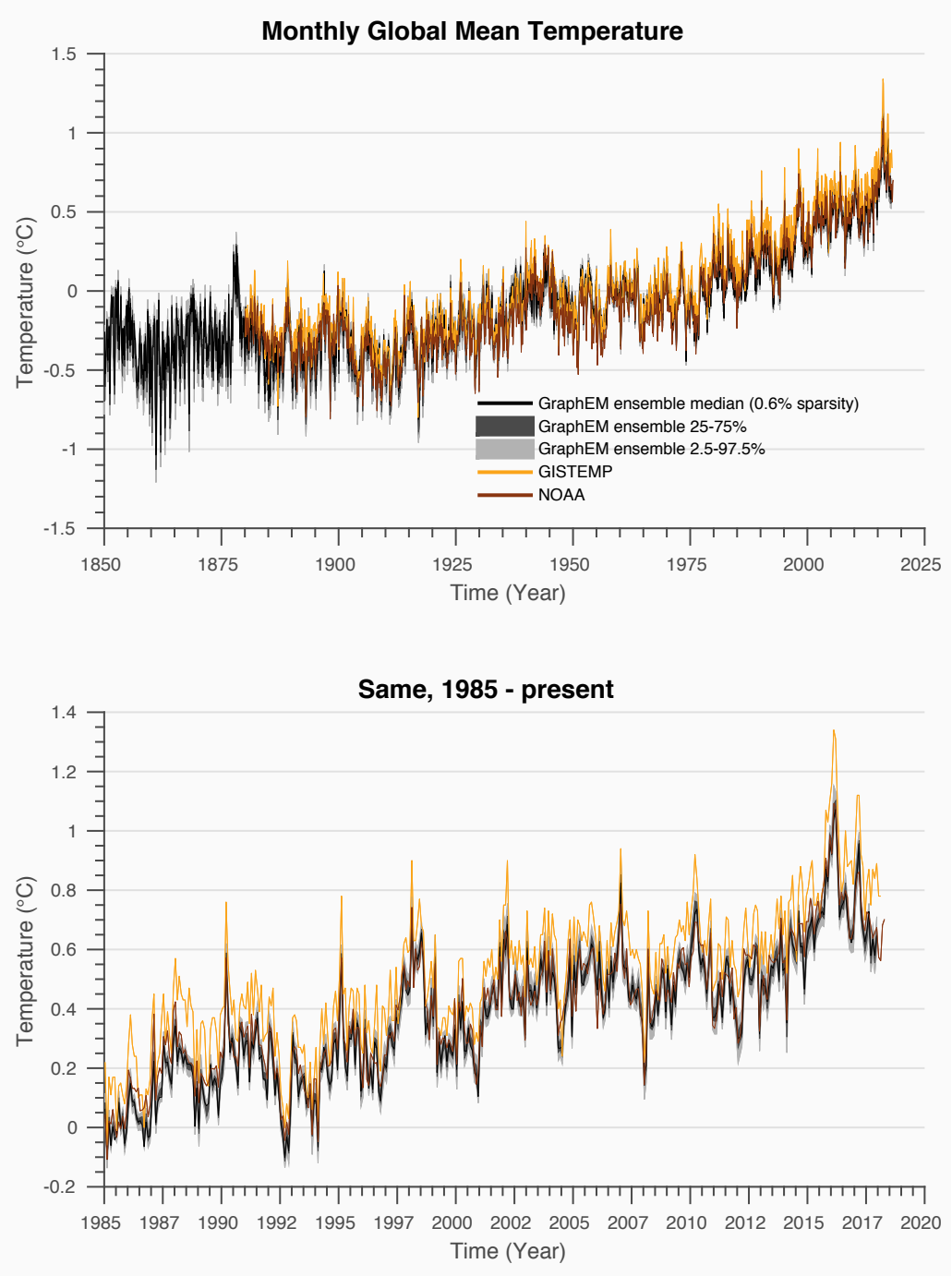

Figure 8. **(Top)** Monthly global mean temperature series plotted over their entire time interval for various global surface temperature datasets: GraphEM-infilled HadCRUT4.6 (black), NASA's GISTEMP (yellow), and NOAA's Global Surface Temperature analysis (brown). **(Bottom)** Same, except the series are plotted over the 1985-2018 period: GraphEM-infilled HadCRUT4.6 (light blue), NASA's GISTEMP (yellow), and NOAA's Global Surface Temperature analysis (brown).

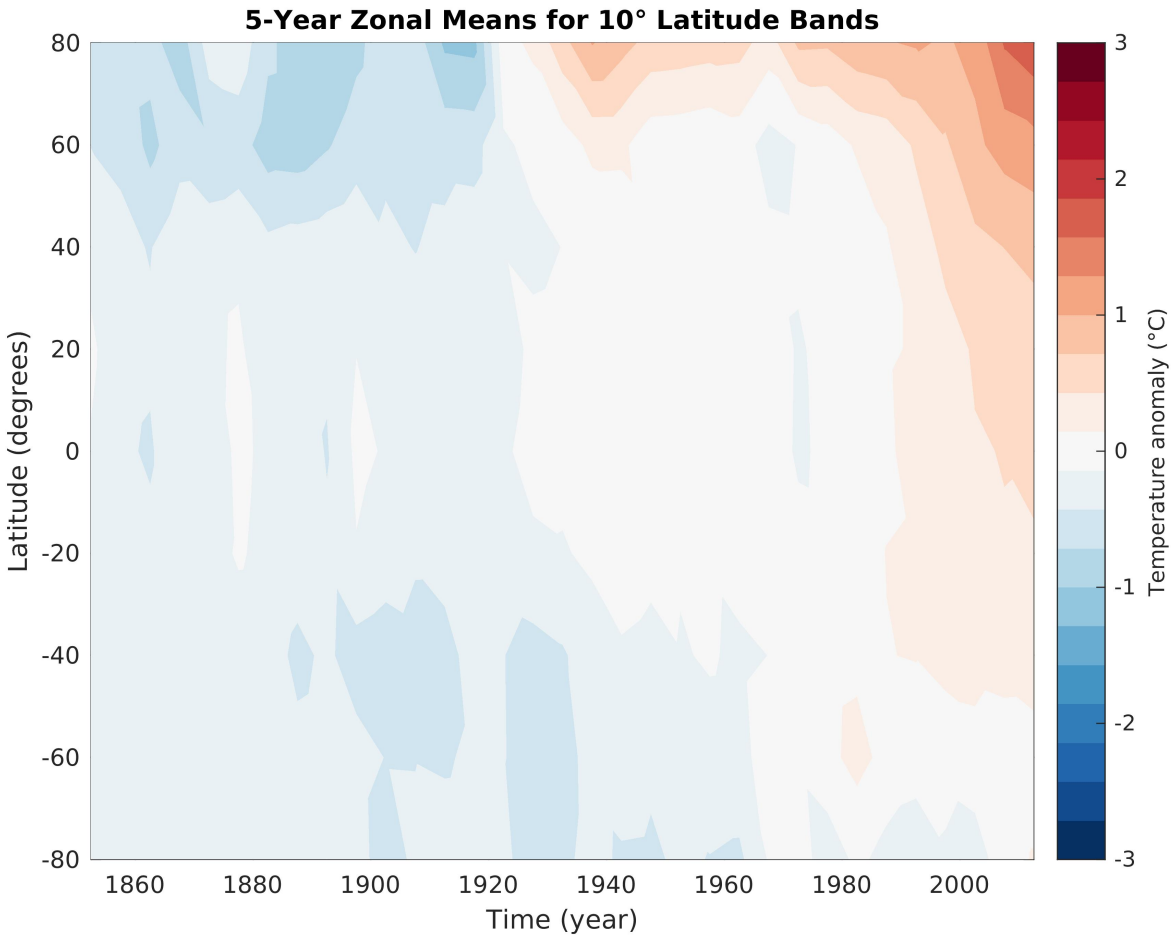

Figure 9. The HadCRUT4$_G$ temperature field zonally-averaged over 10° latitude bands and 5 year intervals. The interpolated solution shows strong polar amplification and an increased rate of warming over Arctic regions.

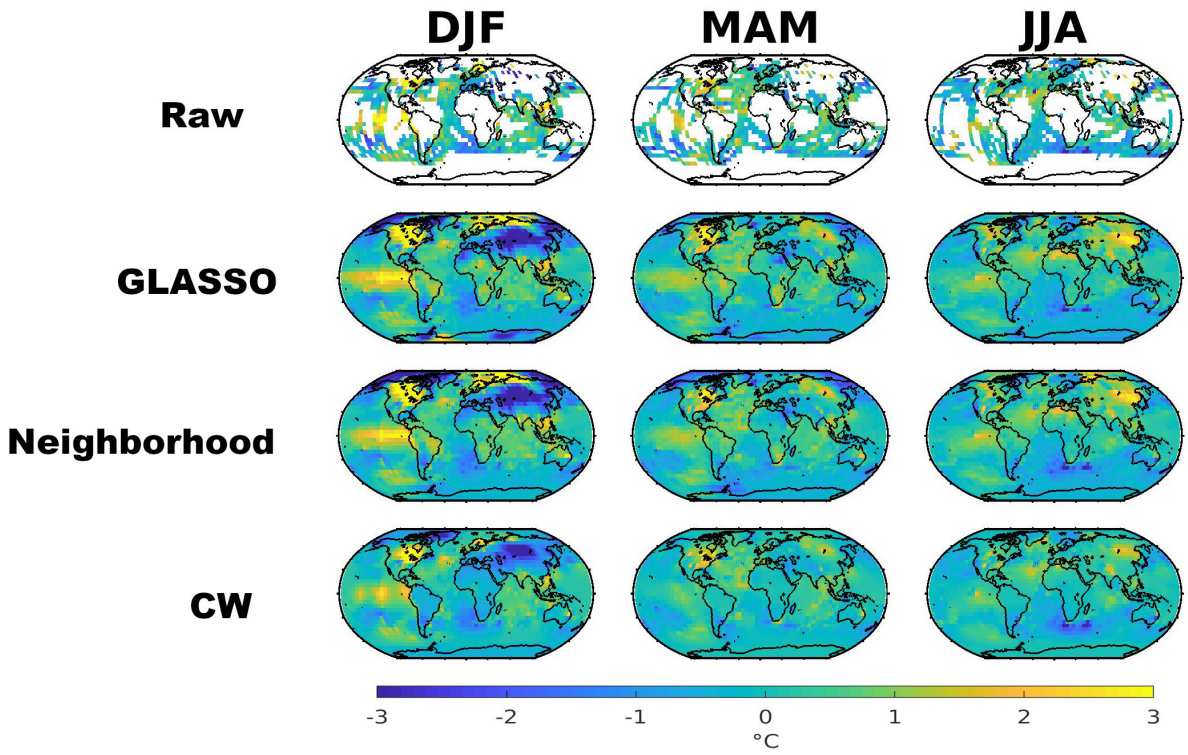

Figure 10. Temperature field averaged across each grid cell over DJF, MAM, and JJA seasonal periods during the 1877/1878 El Niño event. **(First row)** HadCRUT4.6$_{raw}$ **(Second row)** HadCRUT4.6$_G$ with graph estimated using GLASSO and 0.6% target sparsity **(Third row)** HadCRUT4.6$_G$ using a neighborhood graph with a 1000km cutoff radius **(Fourth row)** HadCRUT4.6$_{CW}$

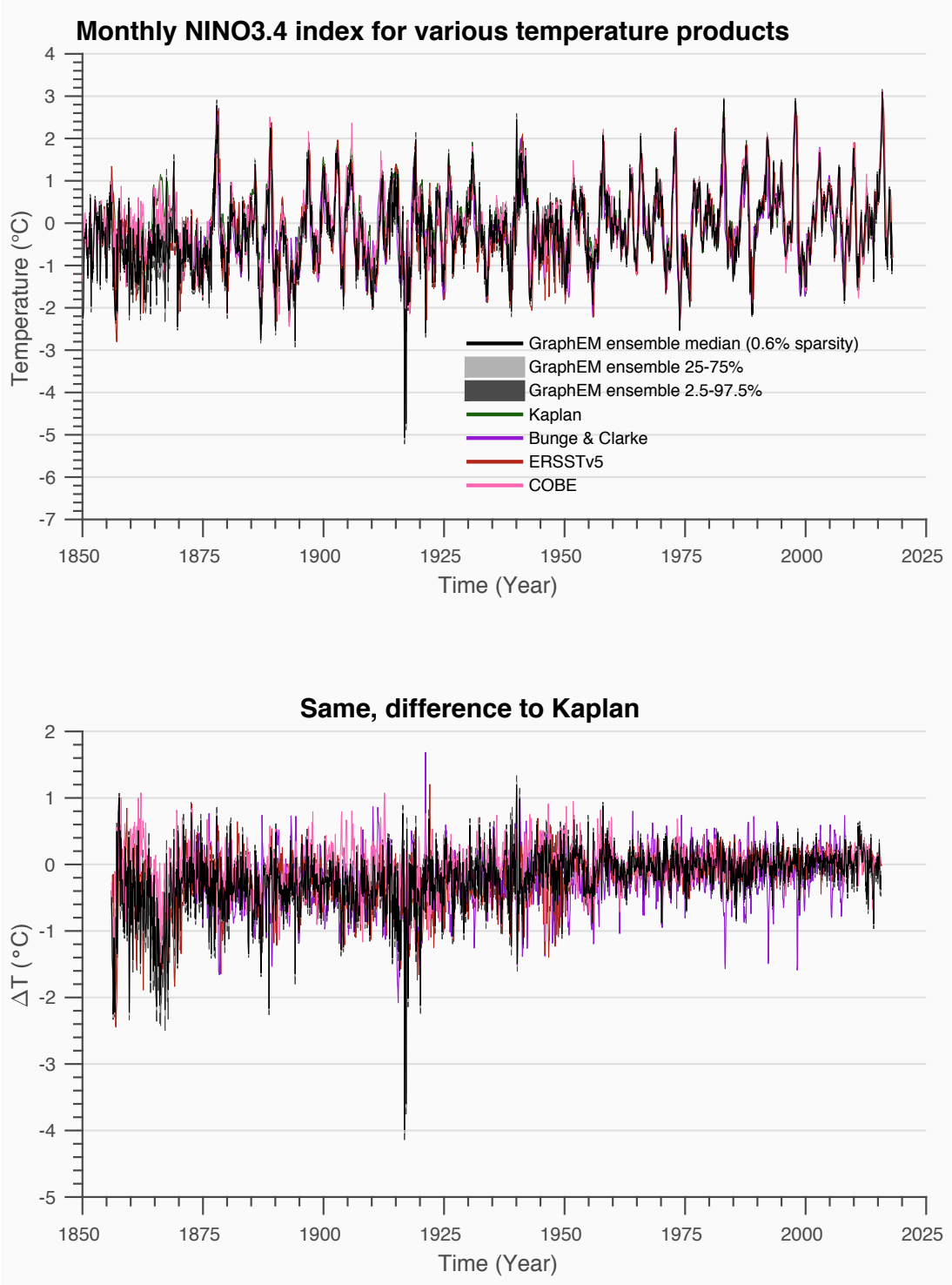

Figure 11. **(Top)** Monthly-resolved NINO3.4 series for Kaplan (dark green), Bunge & Clarke (purple), ERSSTv5 (brown), COBE SST (salmon), and HadCRUT4$_G$ (blue). **(Bottom)** Same, except series are plotted as differences to the Kaplan NINO3.4 series.

Dataset	Period	$\rho(N3.4m)$	$\rho(N3.4a_{yr})$	$\rho(N3.4d_{yr})$	$\rho(N3.4a_{djf})$	$\rho(N3.4d_{djf})$
$HadCRUT4.6_{raw}$	1/1850–12/2017	**+0.95**	**+0.94**	**+0.96**	**+0.98**	**+0.93**
$HadCRUT4.6_{CW}$	1/1850–12/2017	**+0.93**	**+0.95**	**+0.93**	**+0.95**	**+0.92**
COBE SST	1/1850–12/2016	**+0.87**	**+0.94**	**+0.89**	**+0.93**	**+0.87**
Bunge & Clark	1/1873–3/2008	**+0.85**	**+0.94**	**+0.93**	**+0.94**	**+0.90**
ERSSTv5	3/1854–2/2018	**+0.90**	**+0.96**	**+0.93**	**+0.92**	**+0.86**
Kaplan extended	1/1856–8/2015	**+0.88**	**+0.91**	**+0.79**	**+0.95**	**+0.76**

Table 2. Comparison of NINO3.4 indices. Correlations were computed for the monthly-resolved NINO3.4 series (N3.4m), an annualized series calculated using a 12-month mean ($N3.4a_{yr}$), a decadally smoothed series obtained by applying a low-pass filter with a cutoff frequency of 1/120 months to the monthly resolved series ($N3.4d_{yr}$), an annualized series calculated using only DJF months ($N3.4a_{djf}$), and a decadally smoothed series computed by applying a zero-phase low-pass filter with a cutoff period 10 years to the annual series calculated using DJF months only ($N3.4d_{djf}$)

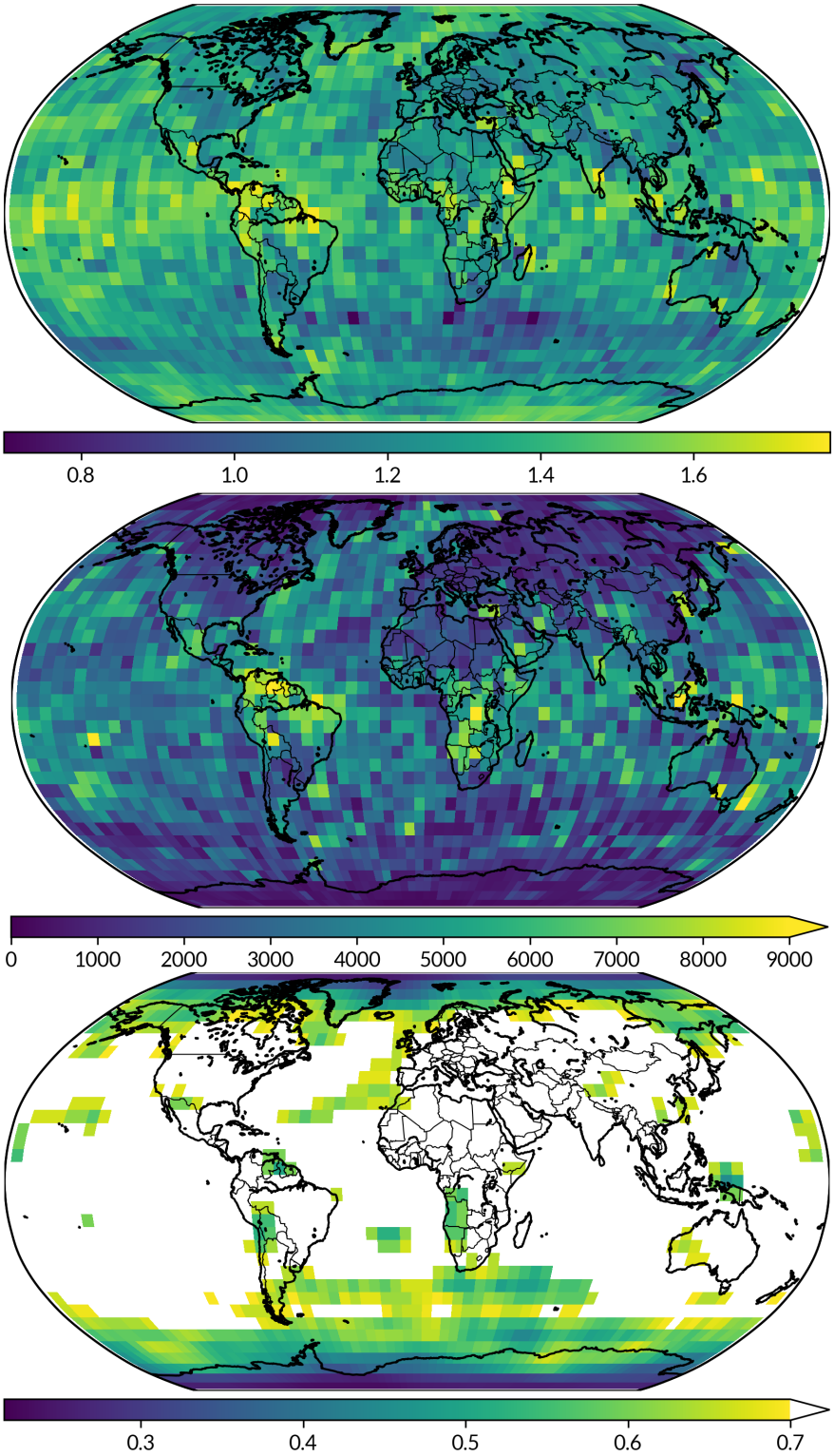

Figure 12. (top) Degree of vertices of G on a $\log_{10}$ scale; (middle) Average distance to neighbors in G (in km); (bottom) Fraction of common edges between G and G_{1000}. To highlight equator/pole contrasts, only regions where this number falls below 70% are shown.

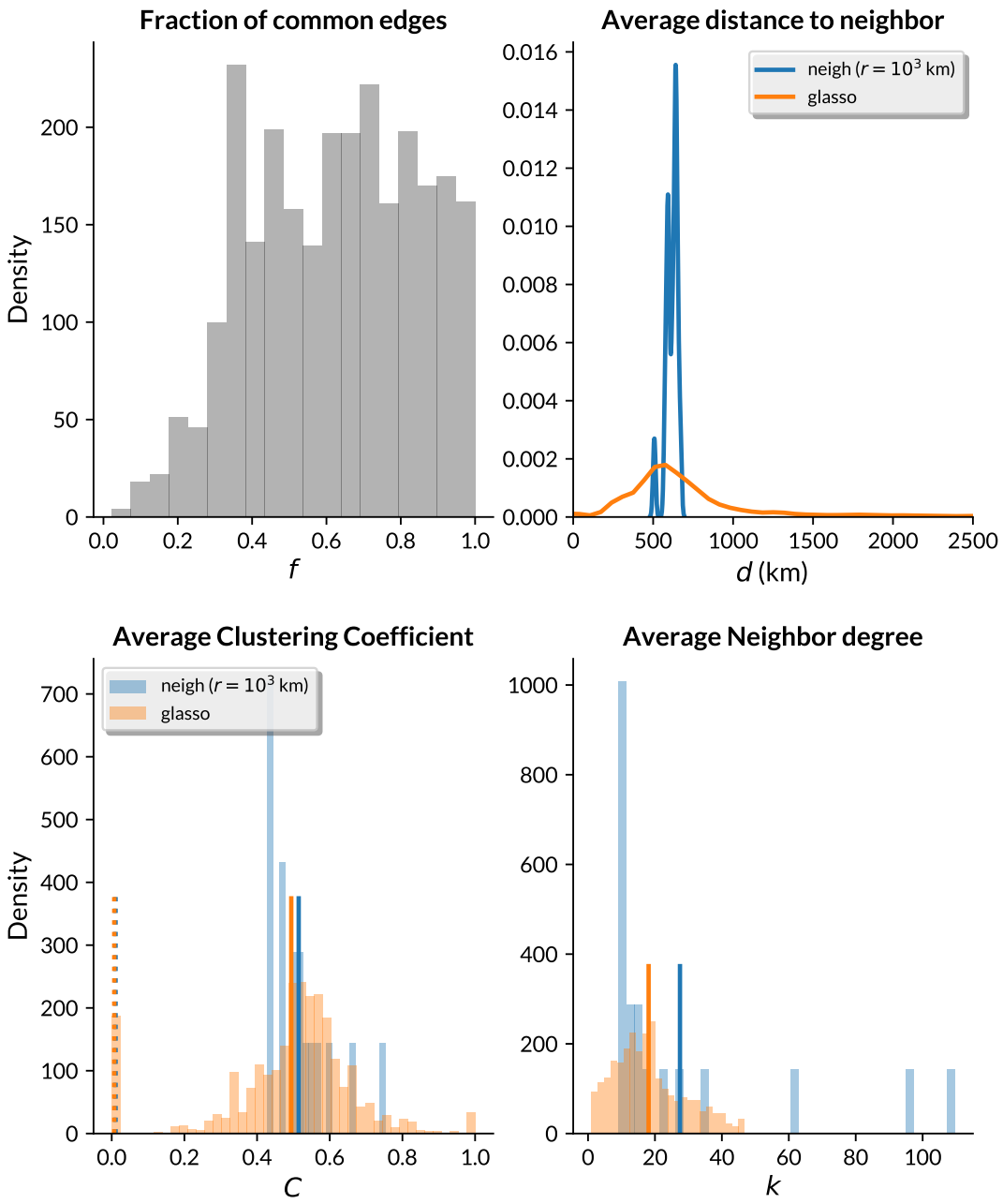

Figure 13. Network connectivity in G (graphical lasso) vs G_{1000} (neighborhood graph). Fraction of common edges between graphs, as raw histogram (f) (top left). Average distance to neighbor in both graphs, as kernel density estimate (top right). Local clustering coefficient C (bottom left) and average network degree k (bottom right).

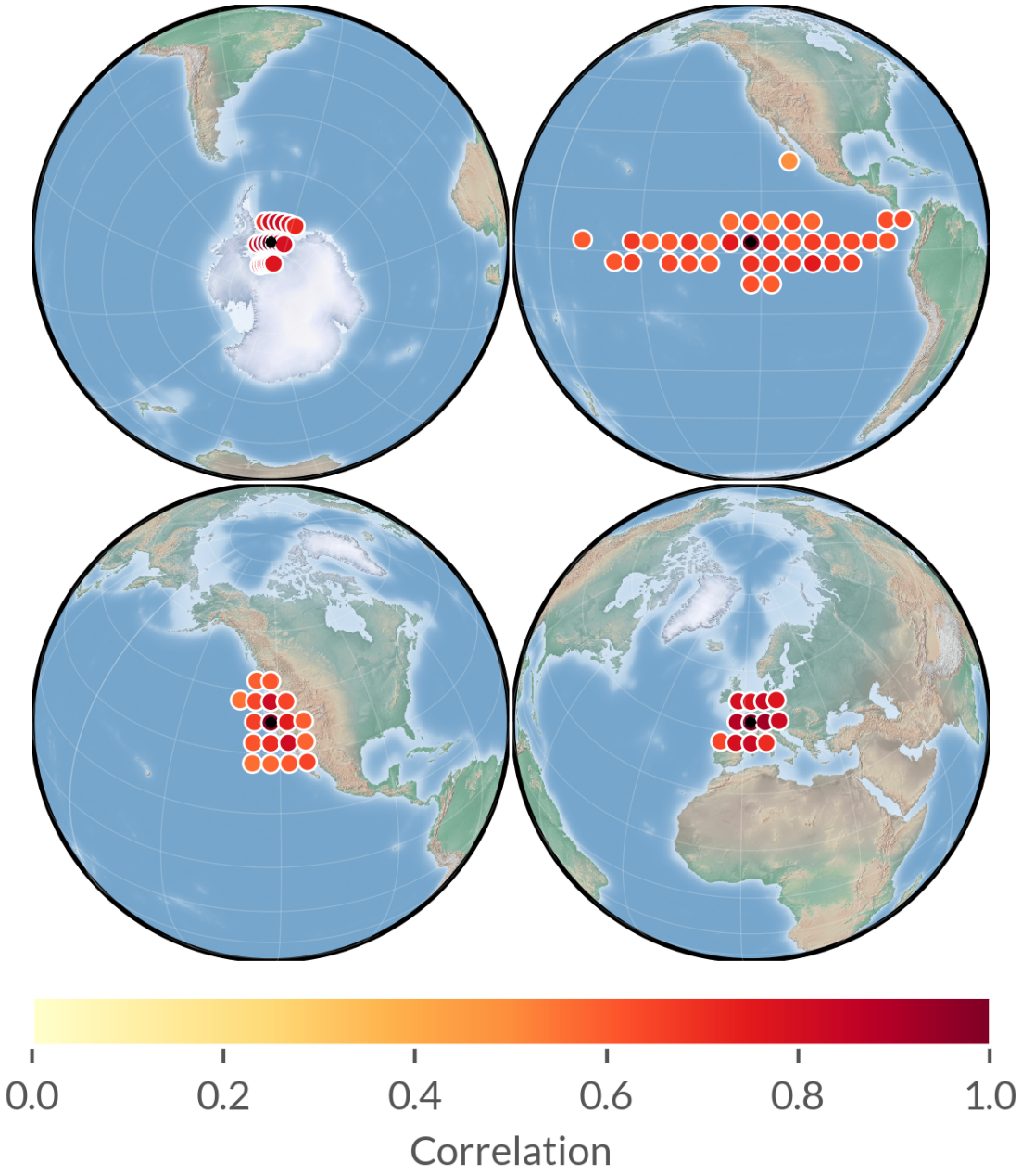

Figure 14. Illustration of the GLASSO graph. Each panel displays the neighbors of a subjectively chosen vertex (star) in the GLASSO graph, colored by the values of their correlation to the node. This illustrates the unequal weight of each node's neighbors in the estimation problem.

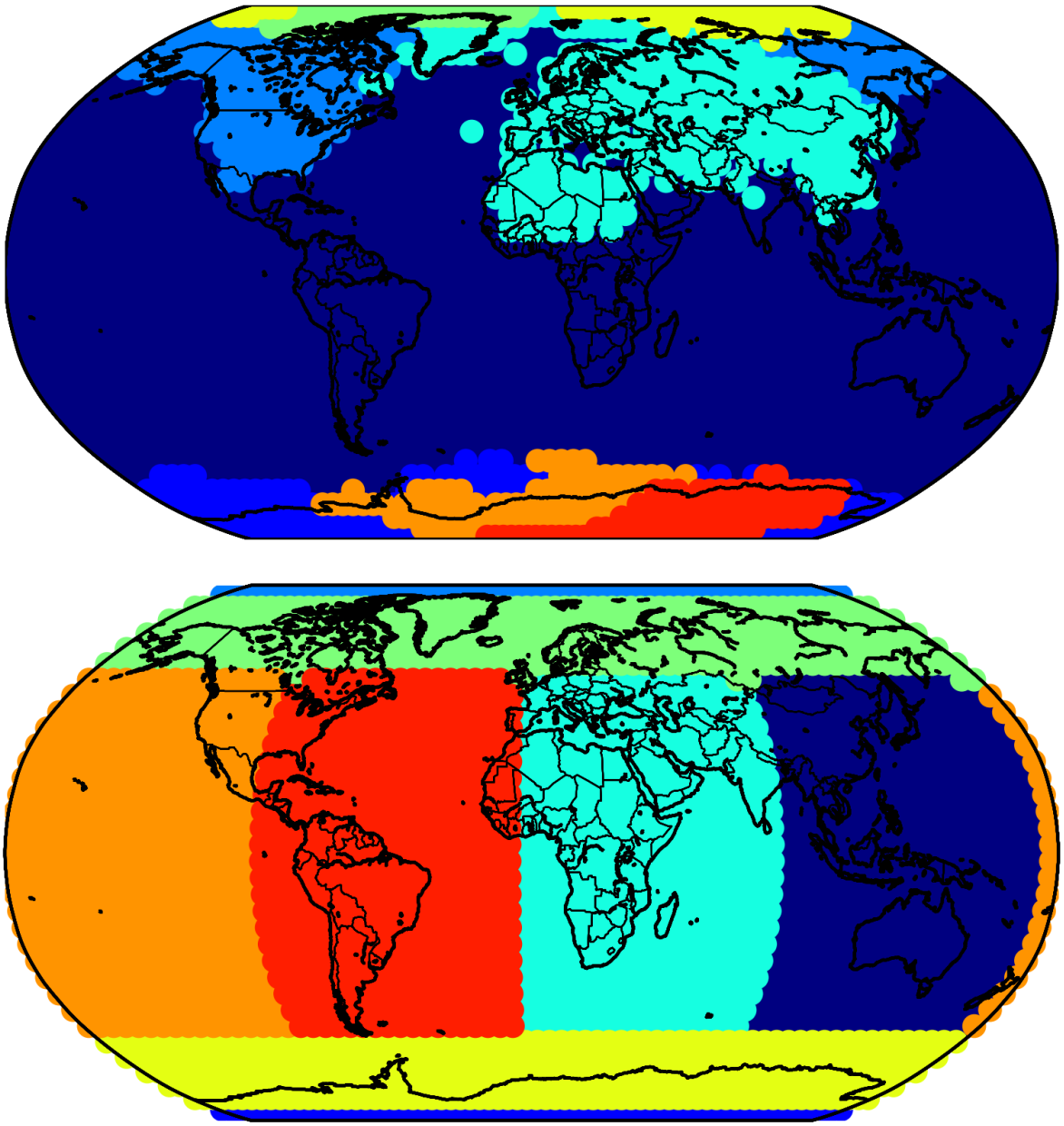

Figure 15. Spectral clustering of (top) the glasso graph G; (bottom) a 1000 km neighborhood graph, G_{1000}. Both use 8 clusters.