Pitfalls of Climate Network Construction—A Statistical Perspective

MORITZ HAAS,[a] BEDARTHA GOSWAMI,[b] AND ULRIKE VON LUXBURG[a,c]

[a] *Department of Computer Science, University of Tübingen, Tübingen, Germany*
[b] *Machine Learning in Climate Science, University of Tübingen, Tübingen, Germany*
[c] *Tübingen AI Center, University of Tübingen, Tübingen, Germany*

(Manuscript received 21 July 2022, in final form 17 November 2022)

ABSTRACT: Network-based analyses of dynamical systems have become increasingly popular in climate science. Here, we address network construction from a statistical perspective and highlight the often-ignored fact that the calculated correlation values are only empirical estimates. To measure spurious behavior as deviation from a ground truth network, we simulate time-dependent isotropic random fields on the sphere and apply common network-construction techniques. We find several ways in which the uncertainty stemming from the estimation procedure has a major impact on network characteristics. When the data have a locally coherent correlation structure, spurious link bundle teleconnections and spurious high-degree clusters have to be expected. Anisotropic estimation variance can also induce severe biases into empirical networks. We validate our findings with ERA5 data. Moreover, we explain why commonly applied resampling procedures are inappropriate for significance evaluation and propose a statistically more meaningful ensemble construction framework. By communicating which difficulties arise in estimation from scarce data and by presenting which design decisions increase robustness, we hope to contribute to more reliable climate network construction in the future.

SIGNIFICANCE STATEMENT: Network-based approaches have gained renewed attention regarding the prediction of climate phenomena such as El Niño events, extreme regional precipitation patterns, anomalous polar vortex dynamics, and regarding understanding the Earth system. Even though climate networks are constructed from a limited amount of noisy data, they typically are not studied from a statistical perspective. However, such an approach is crucial: due to sampling uncertainty, climate networks unavoidably contain false and missing edges. We analyze how sampling artifacts impact the conclusions drawn from the networks and present both pitfalls and statistically robust procedures of network construction and evaluation. We aim to contribute to understanding the limitations and fully leveraging the potentials of network methods in climate and Earth system science.

KEYWORDS: Pattern detection; Statistics; Stochastic models; Teleconnections; Data science; Machine learning

1. Introduction

Climate networks are constructed to find complex structures such as teleconnections (Boers et al. 2019), clusters (Rheinwalt et al. 2015), hubs, regime transitions (Fan et al. 2018), and bottlenecks (Donges et al. 2009a) in the climatic system. Network-based approaches have shown considerable improvements in the prediction of several climate phenomena (Ludescher et al. 2021) such as El Niño events (Ludescher et al. 2014), extreme regional precipitation patterns (Boers et al. 2014), and anomalous polar vortex dynamics (Kretschmer et al. 2017). Typically, climate networks are constructed using a three-step procedure. First, choose a dataset of climatic variables, such as temperature or precipitation, measured on a fixed spatial grid. Then choose a notion of similarity between pairs of locations based on the corresponding time series in the dataset. Finally, construct a network with spatial locations as nodes and with edges between those pairs of locations that have the strongest similarities. Since we only have access to noisy time series of finite length, the calculated similarity values between pairs of locations will be noisy themselves: they are subject to *estimation variability*. As a consequence, any climate network that is constructed using a finite number of data might contain false edges (which should not be present) and have missing edges (which should be present). This leads us to the following important questions that have not received enough attention so far. *Which kinds of distortions are induced in climate networks due to the sampling variability of the underlying time series? Which features of climate networks can be attributed to underlying structure, and which are random artifacts due to finite-sample variation?* These are the questions we discuss in this paper from a decisively statistical point of view. First observe that a climate network is built on a large number of pairwise similarity estimates: if our grid consists of 10^4 locations, a naive procedure needs to estimate 10^8 pairwise similarities. Even extremely well-behaved estimators with a small variability will create a nonnegligible number of wrong edges in the network. But not many "wrong" or "missing" edges are necessary to distort important structural network characteristics. Even a single false long-range edge can substantially distort important network measures such as

Denotes content that is immediately available upon publication as open access.

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/JCLI-D-22-0549.s1.

Corresponding author: Moritz Haas, mo.haas@uni-tuebingen.de

DOI: 10.1175/JCLI-D-22-0549.1

shortest pathlengths, small-world properties, centrality, and betweenness measures, or the emergence of teleconnections. And through the local correlation structure that is inherent in climate data, wrong edges can propagate, leading to many wrong edges, even inducing wrong "link bundles," that is, distinct regions connected by multiple edges.

To assess the severeness of this problem, we introduce a new null model for sampling time series that shares important properties with Earth's climate system but at the same time is simple enough that we can control it, understand it, and simulate from it. To achieve this, we employ a locally correlated, isotropic data–generating process: isotropic random fields on the sphere. The key feature of this model is that the similarity of two time series only depends on the distance of the respective locations, nothing else. Locations that are close by tend to have more similar time series than locations that are far apart. Our model can thus capture important properties of real climate networks such as link-length distributions, but through its isotropic nature it is simple enough that erroneous patterns in the network can be clearly identified as statistical distortions. We introduce time dependence via a vector autoregression process [VAR(1)], which allows us to adjust the autocorrelation on each node. Consequently, the temporal autocorrelation structure can depend on the location, but the spatial correlation structure, and with it the ground truth network, remains approximately isotropic.

Sampling our null model allows us to systematically investigate the connection between noise in the similarity estimates and distortions in the network. Although the simulated data are only locally correlated, we find that complex network structures arise in the estimated networks because of imperfect estimation. For example, global spatially coherent betweenness patterns emerge (Fig. 5), which do not represent any ground truth structure. We also study the influence of choosing different similarity estimators, the influence of network sparsity on betweenness, distortions of other popular network measures, the emergence of spurious link bundles and high-degree clusters, and the biases introduced through anisotropic estimation variability. For example, we find that inappropriate estimators can result in arbitrarily wrong network estimates (Fig. 2). On the other hand, we illustrate that a conscious choice of network-construction techniques may increase robustness with respect to ground truth networks and may uncover different dynamics in the system. To filter out spurious edges, Boers et al. (2019) consider links as significant that do not appear alone but in bundled form. We show that when the data are locally highly correlated, the presence of one spurious long link makes the presence of neighboring links quite likely as well, leading to entire spurious link bundles.

In addition to our simulation results, we validate our findings with reanalysis data from the ERA5 project. We find that the tendency to form bundled connections increases with the strength of local correlations (Fig. 10). The betweenness structure in temperature-based networks highly depends on the network density and the used dataset. This raises the question of whether finding a "betweenness backbone," as in (Donges et al. 2009a), is possible and meaningful. For most climatic variables, we detect severe instability for long links. The nodes of highest degree tend to have autocorrelation (cf. Paluš et al. 2011). We conjecture that some edges from these nodes are spurious and are induced by the increased estimation variability on these nodes.

The wide range of potential empirical distortions makes a reassessment of many of the previous findings in the climate network literature desirable. However, this poses a big challenge: while our simulation study is based on a model with known ground truth, such a ground truth is not available for real-world climate networks. Yet, as our simulations show, it is extremely important to assess the reliability and robustness of findings based on empirical climate networks. Typically, researchers use approaches based on nodewise reshuffling of the time series or edgewise reshuffling of the given network. But we demonstrate that such techniques are inadequate to capture the inherent uncertainty of the network. Instead, we propose to estimate the variability in the network by computing multiple correlation estimates for each edge, while retaining the original spatial similarity structure. With this approach, we might get a statistically meaningful sense of the reliability of network patterns constructed from real, noisy time series.

Our main contributions are summarized as follows:

- We introduce a VAR(1) process of isotropic random fields on the sphere as a suitable null model for geophysical processes, for which deviations from the ground truth are easily detectable.
- We identify systematically occurring random artifacts and distortions in empirical networks and analyze why they arise.
- We show which design decisions increase the robustness of constructed networks.
- We validate our findings with ERA5 data.
- We discuss the shortcomings of common network resampling procedures for significance evaluations and propose a statistically more meaningful framework based on jointly resampling the underlying time series.

The rest of the paper is organized as follows. In section 2 we describe typical network-construction steps and introduce the isotropic data–generating process we employ in our simulations. We present intuitions about the ground truth networks and explain when spurious behavior is to be expected in the empirical networks. Section 3 demonstrates several common patterns of spurious behavior in typically constructed networks, categorized into 1) estimator selection, 2) network measures, 3) link bundles, and 4) anisotropy. Section 4 points out problematic practices in significance testing and potential improvements. Finally, section 5 provides conclusions and possibilities for future work. For readers who are unfamiliar with climate network methodology, we have assembled an introduction in the online supplemental material (section A).

2. Network construction for data from spatiotemporal random fields

To study artifacts that are introduced by estimation procedures, we need access to a "ground truth network," which is not available for real-world climate data. We therefore introduce a manageable stochastic process over Earth with known ground truth structure. We then use the model to evaluate

how estimation procedures introduce random artifacts into the network estimates depending on network-construction steps and the features of the data distribution.

a. Climate network construction

The generic procedure of constructing climate networks from spatiotemporal data is described in algorithm 1: most studies deal with univariate real-valued data at each point in time and space such as temperature, pressure, or precipitation, and so do our experiments. Given a dataset of such time series on a fixed grid, the similarity between pairs of grid points is estimated. Popular similarity measures include the Pearson correlation, mutual information (MI), and event synchronization (Quian Quiroga et al. 2002). There are several ways to construct a network based on all pairwise similarity estimates. Most often, unweighted density-threshold graphs are constructed (Tsonis and Roebber 2004; Yamasaki et al. 2008; Agarwal et al. 2019; Kittel et al. 2021), which means that an edge of weight 1 is formed between two grid locations v_i and v_j when the corresponding similarity estimate $\hat{S}_{ij}$ surpasses a certain threshold. This threshold is chosen so that a desired network density is attained. Another popular approach is edge formation based on significance tests with respect to reshuffled time series (Paluš et al. 2011; Boers et al. 2013, 2014; Deza et al. 2015). Here, the time series at both end locations of an edge are reshuffled to get a baseline distribution of how similarity estimates behave when the time series are independent. The edge is then formed if the original similarity estimate surpasses a predefined significance threshold.

Algorithm 1: *Functional network construction from spatially gridded data*

Input: **Spatiotemporal data** $\{X_{it}\}_{i\in[p],t\in[n]}$, $X_{i\cdot} = (X_{i1}, \ldots, X_{in})$ of time length n measured on p **fixed locations** $V = \{v_i | i \in [p]\}$ in some metric space $(\mathcal{X}, d)$ such as the sphere; similarity measure of interest $S : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ between two locations and estimator $\tilde{S} : \mathbb{R}^n \times \mathbb{R}^n \to [0, \infty)$ of S based on the finite time series.

1) **Estimate the similarity** between two points v_i and v_j based on the data and some estimator $\tilde{S}$ of the chosen similarity measure S: $\hat{S}_{ij} = \tilde{S}(X_{i\cdot}, X_{j\cdot})$.

2) **Construct a graph** with adjacency matrix $\hat{A}$ from the similarity estimates $\hat{S}$, parameters θ and potentially summary statistics of the data. For example, in the case of the unweighted τ-threshold graph,

$$\hat{A}_{ij} = \begin{cases} 1, & \hat{S}_{ij} \geq \tau, \\ 0, & \hat{S}_{ij} < \tau. \end{cases}$$

b. Stochastic ground truth model for spatiotemporal data

To quantify how the similarity estimation process influences the induced networks, we specify a ground truth model, using random fields over the sphere, approximating Earth's surface. Our goal is not to give an accurate model of Earth's climate, but to point out generic patterns of spurious behavior in networks constructed from a limited amount of spatiotemporal data. The simpler the data-generating model remains, the more accurately we can attribute spurious behavior to certain features of the data distribution or the employed network-construction steps. We use a data-generating process in which the correlation between data measured at different locations depends only on the distance between the locations. Such isotropic random fields are common in geostatistics (Cressie 1993; Lang and Schwab 2015) and they allow us to attribute anisotropies in the estimated networks as erroneous.

Here, we first introduce the spatial stochastic process and, in a second step, add time dependence. The mathematical process that we are going to use is an "isotropic Gaussian random field." A *random field* assigns a real value to every point of the sphere, imagine a surface temperature field. Centering (and possibly detrending and normalizing) data on each point in space yields a *zero-mean* random field, representing so-called (detrended standardized) anomalies. When evaluating a *Gaussian* random field on finitely many points, its values are jointly Gaussian distributed. For *isotropic* random fields, the covariance between two points is solely determined by the distance between the points. Hence, a zero-mean isotropic Gaussian random field is fully characterized by its covariance function k, which determines how smoothly and to what extent the random field varies across space.

Formally, a zero-mean isotropic Gaussian random field G on the sphere with covariance function $k : [0, \pi] \to \mathbb{R}$ is defined as a collection of real-valued random variables $\{G(v)\}_{v\in S^2}$ such that $\mathbb{E}[G(v)] = 0$ for all $v \in S^2$ and, given a finite grid $\{v_i\}_{i=1,\ldots,p} \subset S^2$, the random field's values on the grid points are jointly Gaussian distributed,

$$\big(G(v_1), \ldots, G(v_p)\big) \sim N(0, \Sigma),$$

with covariance $\Sigma_{ij} = k(|v_i - v_j|)$.

One popular covariance function is the Matérn covariance function (section B.2 in the supplemental material), whose smoothness parameter ν and scale parameter ℓ make it flexible as well as interpretable. It allows interpolation between the absolute exponential kernel and the Gaussian radial basis function (Stein 1999, chapter 2.10) and monotonically decreases with distance, irrespective of parameter choice. We introduce the abbreviation MIGRF(ν, ℓ) for a zero-mean isotropic Gaussian random field with Matérn covariance, smoothness ν, and length scale ℓ. Figure 1d shows realizations of an MIGRF with varying parameters, when traversing the sphere from South to North Pole. Low-smoothness ν results in abrupt changes. As expected, processes with smaller length scales ℓ contain larger fluctuations on a fixed interval. We choose $\nu \in \{0.5, 1.5\}$ and $\ell \in \{0.1, 0.2\}$ (in radians) to reflect realistic values for climatic time series (Guinness and Fuentes 2016) (section B.3 in the supplemental material), as well as to point out their influences on the estimation procedure.

We introduce time dependence via a vector autoregression VAR(1) (section B.1 in the supplemental material) that allows us to assign any desired lag-1 autocorrelation to each

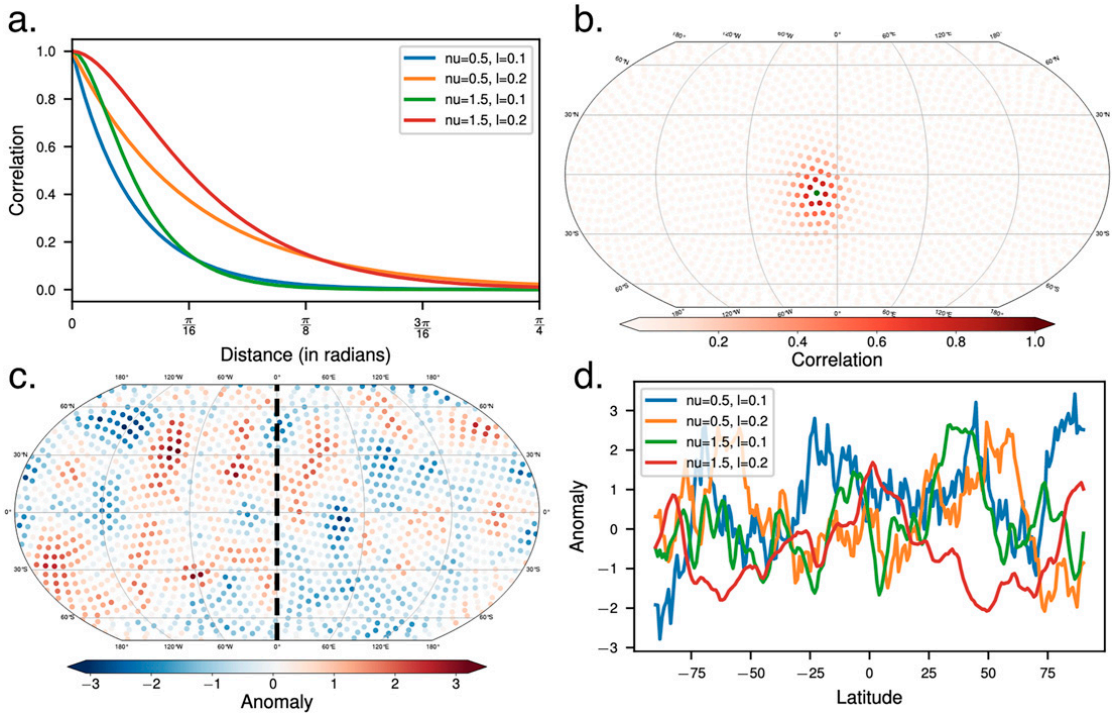

FIG. 1. Isotropic Gaussian random fields. (a) The Matérn correlation function for different parameter choices. (b) True Matérn correlation with respect to the green point for $\nu = 1.5$ and $\ell = 0.2$. (c) A realization of an MIGRF($\nu = 1.5$, $\ell = 0.2$), representing anomalies at a fixed time point. The correlation function induces smoothly varying values. The dashed black line shows the geodesic path from South Pole to North Pole used in (d). (d) Random realizations of MIGRFs with different parameters evaluated on the geodesic path shown in (c).

node of the network. Under this basic time dependence, we will be able to separate the effect of autocorrelation on the estimation procedure from other influences.

c. *Ground truth networks and imprecise estimates*

1) GROUND TRUTH NETWORKS

If we fix a grid and a network-construction method, a ground truth data distribution leads to a "true network" on this grid. Given the underlying data distribution, as in our model, we can calculate the true pairwise similarities between grid points. The network-construction procedure then determines the ground truth network based on the true similarities. For example, a ground truth density-threshold graph simply consists of the edges corresponding to the largest similarity values. How much ground truth structure of a climatic process can be captured in the ground truth network depends on the choice of climatic variable, grid, similarity measure, and network-construction scheme. Another question is then whether this ideal network can be approximated with the available empirical data and estimators.

2) ERRORS IN THE ESTIMATED NETWORKS

Given a finite amount of data, we only have access to imperfect estimates of the true similarity values. Consequently, networks constructed from data as well as their characteristics will only be estimates of the corresponding ground truth quantities and inherit *intrinsic variability*. When the chosen similarity estimator is not suitable for the estimation task, the constructed graphs can look arbitrarily wrong (Fig. 2). However, by simple inspection, it is not possible to judge whether a constructed climate network reflects "true" aspects of the physical system or whether it is dominated by random artifacts introduced through the estimation procedure. For this reason, in our simulations we mainly address the following question: *How do the estimated networks and their characteristics differ from their corresponding ground truth quantities?* The answer depends on the properties of the random field, the employed estimator, and the considered network characteristic (see section 3). To get started, let us discuss how wrong individual edges occur and then how wrong link bundles arise.

3) ERRORS IN INDIVIDUAL EDGES

Errors in the network occur because the similarity estimates between locations vary around the ground truth similarity values. False-positive edges are wrongly included in the empirical network but are not present in the ground truth network; false-negative edges appear in the ground truth network but are missing in the empirical network. Let us understand when these two cases arise in threshold graphs. Assume that the similarity estimate $\hat{S}$ over an edge with ground truth correlation S is imprecise and follows the distribution $\mathcal{N}(\mathbb{E}[\hat{S}], \sigma^2)$ (we consider the normal distribution for simplicity; other distributions lead to qualitatively similar behavior). The probability that this edge is formed in the τ-threshold graph is then given by

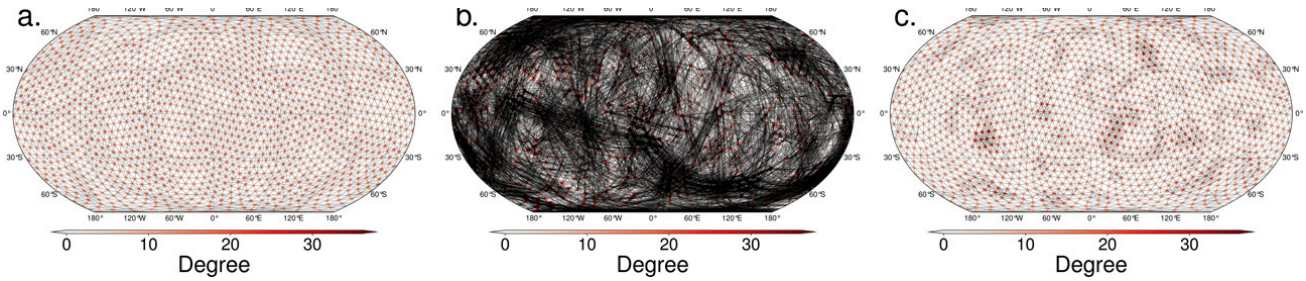

FIG. 2. (a) Ground truth network of density 0.005 for monotonically decreasing correlation structure. The shortest links possess the largest ground truth correlation values. The graph is not perfectly isotropic because an isotropic grid does not exist. (b) Empirical estimate of the left ground truth network with the *same network density* given lognormal data based on an MIGRF($\upsilon = 1.5$, $\ell = 0.2$) with variance 10 and $n = 100$ using empirical Pearson correlation [see section 3c(1)]. Many false links arise due to high estimation variance. Many long links clutter the image. Observe (strong) spurious *bundled* teleconnections. (c) Empirical estimate over the same data using Spearman correlation. No long links are formed, but we can observe spuriously dense regions.

$\Phi\left((\mathbb{E}[\hat{S}] - \tau)/\sigma\right)$, where Φ denotes the cumulative distribution function of the standard normal distribution. A false positive can only occur when the true similarity S is smaller than the threshold τ. Then the error probability is not negligible when the similarity estimates are upward biased ($\mathbb{E}[\hat{S}] \gg S$), or when the estimation variance σ^2 is large. Analogously, the probability of a false negative is not negligible when the estimate is downward biased or when the estimation variance σ^2 is large. As we will see in section 3f, the variability in the estimates grows, or in other words, their signal-to-noise ratio decreases, with data scarcity and increasing autocorrelation in the observations. A bias in the similarity estimates can be introduced by the estimator. Taken together, finding an estimator with a good bias-variance trade-off for the given similarity measure can significantly reduce the number of false edges. In particular, when the desired graph density is chosen so large that many ground truth correlation values of included and excluded edges are similarly small, the likelihood of spurious behavior increases as these edges cannot be well distinguished under the estimation variance. We see this in our experiments below when we construct dense graphs over a small-scale correlation structure (Fig. 3).

4) HOW ERRORS SPREAD LOCALLY DUE TO COVARIANCE

When the data are locally highly correlated, as is typical for climatic variables, this correlation may carry over to the joint distribution of similarity estimates. As a result, an error may propagate from one edge to edges on neighboring nodes in the following way: when the similarity estimate on one false edge is spuriously large, it is likely that the correlation estimates on edges on neighboring nodes are similarly large, so that these neighboring edges are also falsely included in the empirical network, resulting in false bundles of edges. In density-threshold graphs, this makes some regions spuriously appear denser than others. A formal argument is given in section C in the supplemental material. Combining the thoughts above, *false bundles of edges occur with high probability when measurements from close by points are highly correlated and the similarity estimates are imprecise*. Find related simulation results in section 3e.

3. Spurious behavior in networks from finite samples

In this section, we explore the effects that imprecise estimates impose on commonly constructed climate networks. We do so by simulating the isotropic Gaussian random fields introduced above.

a. *Network construction*

We construct networks following algorithm 1. To approximately remove the effects of anisotropic grids, we generate a Fekete grid (Bendito et al. 2007) after 1000 iterations with 5981 points, approximately realizing an isotropic grid of 2.5° resolution. If not stated differently, we sample an MIGRF(υ, ℓ) independently in time with $n = 100$. From the ERA5 dataset between 1979 and 2019, we consider monthly temperature of air 2 m above the surface (t2m), surface pressure (sp), total precipitation (pr), and geopotential height at 250, 500, or 850 hPa (z250, z500, z850) as well as daily t2m (dt2m). We linearly detrend all ERA5 variables and subtract the monthly climatology. Finally, real and simulated datasets are centered and normalized in each grid point to result in detrended anomalies. In some simulations, time dependence is introduced, amplifying our findings (see section 3f and section D in the supplemental material). Many studies construct correlation networks from sliding windows (Radebach et al. 2013; Hlinka et al. 2014; Fan et al. 2017; Kittel et al. 2021). Typically, these windows cover at most a year of daily observations. While more measurements in time increase the accuracy of estimated networks, our findings also hold for larger n (section E in the supplemental material).

b. *Visualizations*

For network visualizations we use a Fekete grid with 1483 points, approximately realizing a 5° resolution. In our figures, dashed lines always denote ground truth values. Uncertainty bands cover the range between the empirical 0.025 and 0.975 quantile from 30 independent repetitions. The letter x and the circle denote 95% and 99% quantiles of a distribution, respectively, and the triangles, the minimal and maximal values of a distribution.

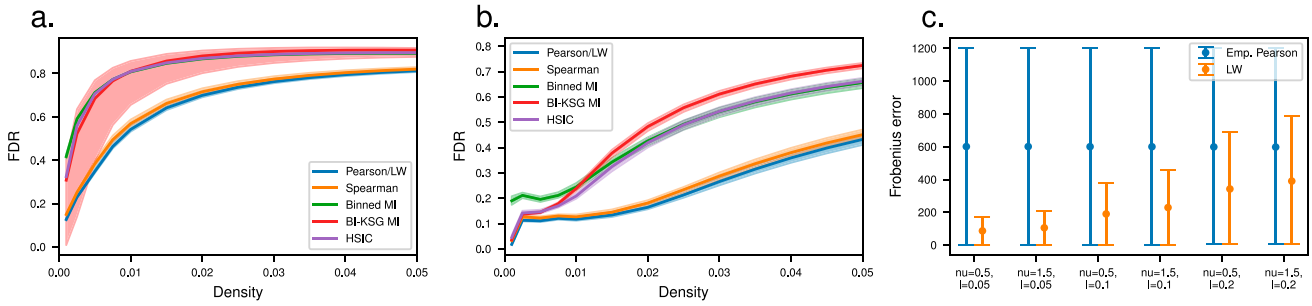

FIG. 3. Errors in empirical networks. (a) False discovery rate (FDR) for various similarity measures and different estimators of the same similarity measures for a nonsmooth MIGRF with short length scale $\nu = 0.5$ and $\ell = 0.1$. Sparse networks have a better FDR. Empirical Pearson correlation and the Ledoit–Wolf estimator coincide in unweighted density-threshold networks (see main text). (b) As in (a), but for a smooth MIGRF with long length scale $\nu = 1.5$ and $\ell = 0.2$. Estimation performance remains reasonably good up to larger network density. (c) Error of estimated correlation matrix from ground truth in Frobenius norm, which is proportional to the root-mean-square error per edge weight estimate. The number of errors in empirical networks is alarming for all hyperparameter settings. Suitable estimators, such as the Ledoit–Wolf estimator, drastically reduce the error in the edge weight estimates compared with empirical Pearson correlation.

c. *Estimation*

1) UNSUITABLE ESTIMATORS CAN INDUCE MANY WRONG EDGES

(i) Problem

When the marginal distribution on the nodes is heavy tailed, as in the case of precipitation data, commonly applied estimators become inadequate if they are sensitive to outliers. For instance, the naive correlation estimator has unusably large variance under heavy-tailed distributions; yet it has been applied to precipitation data in several studies (Scarsoglio et al. 2013; Ekhtiari et al. 2019, 2021).

(ii) Simulation results

We simulate heavy-tailed data by exponentiating the MIGRF data. Let G be a centered MIGRF with correlation function $k(\cdot)$ and variance σ^2. By setting $H(x) = \exp[G(x)]$, H defines a lognormal isotropic random field. For each point x on the sphere, we get

$$\mathrm{Cor}[H(x), H(y)] = \frac{e^{\sigma^2 k(|x-y|)} - 1}{e^{\sigma^2} - 1}.$$

Choosing σ^2 allows to continuously adjust the heaviness of the tails: while small values of σ^2 approximately recover the original correlation function k, increasing σ exponentially enhances the tail strength. For large σ^2, the correlation between grid points quickly drops to 0 with distance. We choose $\sigma^2 = 10$, which is the correct order of magnitude to fit precipitation tails on global mean (Papalexiou 2018). Note that the precipitation distribution on Earth crucially depends on the location. Here, we solely aim to illustrate the intricacy of handling heavy-tailed data through isotropic simulations. Figure 2 demonstrates that the empirical correlation fails as a correlation estimator of data sampled from H. Because the empirical covariance is an average of lognormal random variables, it will be a large variance estimator of the population covariance. The large estimation variability induces many (possibly bundled) false and missing links. For short time series, you can imagine single events dominating on each node. When these events occur at the same time for a pair of nodes, the nodes will show high empirical correlation, although the true correlation may be zero.

(iii) Consequences

Removing outliers or finding a suitable data transformation reduces this problem. By design, $\log(\cdot)$ would transform the random field back to a Gaussian random field. Alternatively, we can employ an estimator that is robust to heavy-tailed distributions (Minsker and Wei 2017). Since the Spearman correlation is invariant under monotonous transformations, it produces exactly the same results for the normal and lognormal data. An alternative to Spearman correlation with faster convergence rates is Kendall's tau (Gilpin 1993). Barber et al. (2019) consider several of the above ideas to estimate correlation in the context of hydrologic data.

2) COMPARING SIMILARITY MEASURES AS WELL AS ESTIMATORS

(i) Problem

While the empirical Pearson correlation estimator has often been equated with the corresponding similarity measure, we can strictly reduce the estimation variance in Pearson correlation networks by considering a different estimator—even for Gaussian data. Radebach et al. (2013) have shown that many characteristic network patterns are already visible in Pearson correlation networks, and historically, the Pearson correlation has been the most popular similarity measure (e.g., Tsonis and Roebber 2004; Tsonis et al. 2008; Yamasaki et al. 2008; Paluš et al. 2011; Fan et al. 2022). As estimators of mutual information need to be able to capture arbitrarily complex dependence structures, they tend to require even larger sample size to achieve reliable accuracy than do correlation estimators, resulting in more spurious behavior given the same sample size.

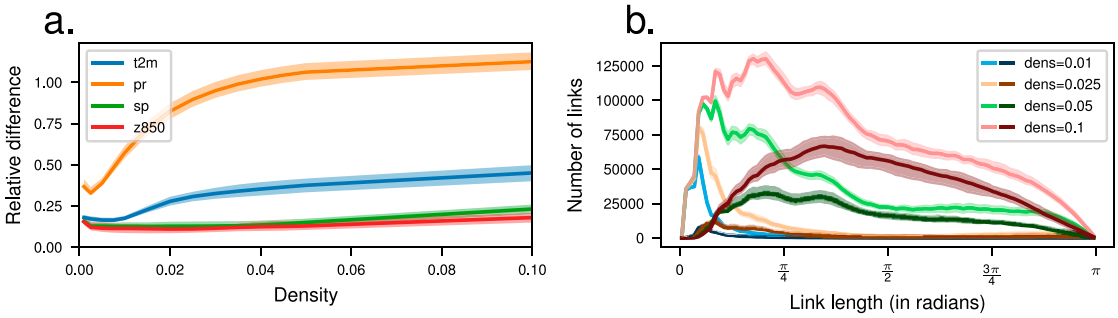

FIG. 4. Fluctuating edges in real networks. (a) Fraction of differing edges between pairs of empirical Pearson correlation networks from various climatic variables obtained by bootstrapping in time. As we divide by the number of links in one of both compared networks, the fraction of differing edges varies between 0, when no link differs, and 2, when all links differ. (b) Total link-length distribution (light) vs link-length distribution of differing edges (dark) between bootstrap networks of t2m. The link-length distribution is similar to the one from our MIGRF (cf. Fig. 7c). The short edges do not differ among bootstrap samples; long edges fluctuate heavily.

(ii) Simulation results

We consider the following similarity measures and their estimators. For MI, we use a simple binning estimator as applied in the complex network python package pyunicorn (Donges et al. 2015), where we use $\lfloor n/5 \rfloor$ bins as suggested in Cellucci et al. (2005) [a more conservative criterion than Cochran's, which was applied in Donges et al. (2009b)]. To evaluate the importance of the estimator, we also employ a bias-corrected version of the popular Kraskov, Stögbauer, and Grassberger (KSG) mutual information estimator (Gao et al. 2018; Kraskov et al. 2004) with $k = 5$. As an alternative to mutual information, we explore an estimator of the Hilbert–Schmidt independence criterion (HSIC) for random processes (Chwialkowski and Gretton 2014). For correlation, we employ the linear Ledoit–Wolf estimator (Ledoit and Wolf 2004), which counteracts the distortion of high-dimensional empirical correlation matrices by shrinking their eigenvalues.

Figure 3 shows the false discovery rate (FDR), which measures the fraction of false links, as a function of network density. Although fewer true links are available for small densities, sparse graphs are more accurate in terms of the FDR, because the correlation values of ground truth links can be empirically distinguished with high certainty from most false links under estimation variability. For random fields with long length scales, this empirical separability remains intact for longer edges. Therefore, the FDR remains low up to larger network densities (see also Fig. 9). Given the same amount of data, more complex similarity measures perform worse. For sparse graphs, the Hilbert–Schmidt independence criterion shows promising performance compared with the mutual information estimators. The bias-corrected KSG estimator is computationally expensive with fluctuating performance, and the binned MI estimator is strictly worse than HSIC. Unweighted empirical Pearson and Ledoit–Wolf density-threshold networks coincide because they produce the same ranking of edge weights. Figure 3c shows the error of the estimated correlation matrix compared with the ground truth in Frobenius norm under various hyperparameter settings. The ground truth correlations grow monotonously from left to right. Note that the empirical correlation matrix makes large estimation errors irrespective of the parameters of the random field. The linear Ledoit–Wolf estimator improves the estimation in all cases. Consequently, fixed-threshold networks, as well as weighted networks, are better approximated by the Ledoit–Wolf estimator. The less correlated the grid points are, the lower the error of the Ledoit–Wolf estimator as it shrinks the correlation estimates toward an identity matrix.

For real data, we cannot calculate the FDR as we do not know which links are false. Instead, we generate bootstrap samples of all time points and create perturbed datasets by including the measurements on the entire grid at these time points. We then construct several networks with the same density from these perturbed datasets and finally compute the fraction of differing links between pairs of sampled Pearson correlation networks (Fig. 4a). With this procedure, we approximate the network distribution induced by the dataset (see section 4b). High autocorrelation causes the need for blockwise bootstrapping to receive consistent estimates, as the network variability is increasingly underestimated with increasing autocorrelation. The results should be seen as a conservative preliminary insight into the intrinsic network variability and the number of unstable edges. A robust network-construction procedure should yield a low fraction of fluctuating links across bootstrap draws. Narrow uncertainty bands indicate that varying weighting of climatic regimes among the bootstrap samples has limited influence on the networks. Observe that in t2m and pr networks, an alarming fraction of links fluctuate (Fig. 4a), while networks from smooth variables with long length scale, such as sp and z850, fluctuate less (consistent with Fig. 9). In contrast to the synthetic data, the curves do not grow monotonically in the sparse regime. Therefore, resampled networks may be helpful in choosing a maximally robust density, minimizing the fraction of varying edges in the empirical networks. Density up to 0.01 seems to be an appropriate choice for t2m; larger densities dramatically decrease the network robustness. The differing links do not contain short edges (Fig. 4b), as the correlation values on short edges are consistently large. Longer links heavily depend on the sampled time points and are sensitive to slight perturbations of the correlation estimates. Hence, the decision

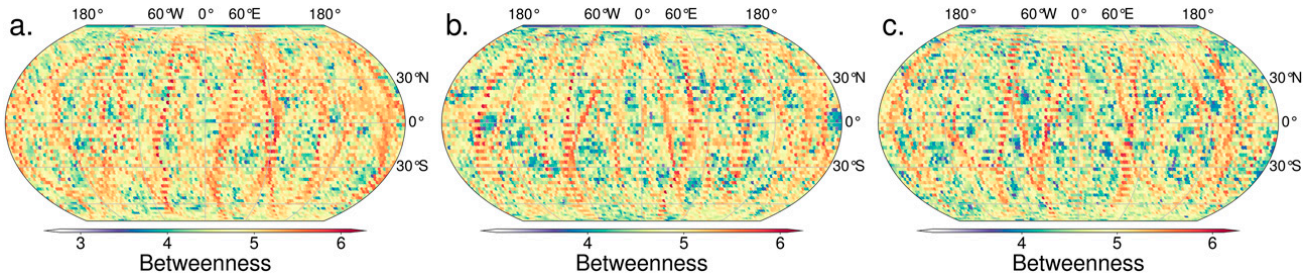

FIG. 5. Betweenness maps of simulated networks. Transformed betweenness maps [$\log_{10}(\mathrm{BC}+1)$] of empirical networks with density 0.005 on a Gaussian grid. The maps depict independent realizations of the same data-generating process MIGRF($\upsilon = 1.5$, $\ell = 0.2$). Because of the anisotropic grid, the poles are highly intraconnected. The density is chosen such that there exist few random shortest paths connecting the poles, resulting in pronounced spurious global betweenness pathways, which alter in location and extent among independent realizations of the data and do not represent ground truth structure.

as to which long links to include should not be based on a single correlation estimate. Geopotential heights behave differently and become more stable at larger densities as they have a correlation structure with an extremely large length scale.

(iii) Consequences

The selection of appropriate similarity measures depends on how much data are available. Mutual information estimators require much more data to yield reliable results than do correlation estimators. HSIC shows promising performance in our experiments. It may be worth exploring other alternatives to MI, such as Romano et al. (2018), in the future. Although not particularly well suited for random field data, the Ledoit–Wolf estimator is a uniform improvement over naive empirical cross correlation when estimating weighted Pearson correlation networks. Future work should put more focus on which estimators perform best on meteorological data.

For most climatic observables, we detect severe network variability for long links. To quantify the structural and link robustness of constructed networks, we need resampling procedures that adequately capture the intrinsic network variability (section 4b). Small network densities yield more robust networks in terms of differing/fluctuating links in resampled networks.

d. Network measures

1) Extreme betweenness values are unreliable in sparse networks

(i) Problem

The betweenness centrality of a node v_k is given by the expression $\sum_{i,j\neq k}\sigma_{i,j}(k)/\sigma_{i,j}$, where $\sigma_{i,j}$ is the total number of shortest paths from node v_i to node v_j and $\sigma_{i,j}(k)$ is the number of those paths that pass through v_k. In climate networks, high betweenness indicates that a location connects different regions. In temperature-based networks, such locations have been interpreted as key pathways of energy flow (Donges et al. 2009a). When interested in nodes of highest betweenness, it is tempting to construct sparse networks, because the most important points stand out more distinctly. However, we find that variability in betweenness also increases drastically when sparsifying the network. Donges et al. (2009a) operate exactly in this unreliable regime. Let us explain the influence of sparsity on betweenness in climate networks.

(ii) Simulation results

Figure 5 shows betweenness maps of networks, constructed as in Donges et al. (2009a), from independent draws of our locally correlated, isotropic model. Because of the standard Gaussian grid, randomly fluctuating betweenness "backbones" emerge that form global pathways but do not represent ground truth structure. The density is chosen such that there exist few pathways between the well-intraconnected poles. Which exact nodes lie on the few shortest paths between the node-dense polar regions depends on which false links are formed in the region between the poles. Phrased differently, *the supposedly important nodes with high betweenness are precisely the ones with false edges and alter between different realizations of the process.* The difference map (Fig. 2 of Donges et al. 2009a) between networks from different datasets shows strikingly similar north–south pathways. The latticelike structure makes the sparse ground truth network highly susceptible in terms of betweenness. Since the data-generating process is isotropic and the Gaussian grid is symmetric with respect to longitudinal rotations, nodes on the same latitude have equal betweenness values in the ground truth network. The empirical networks consistently show systematically different betweenness distributions. While the maximal betweenness value in the ground truth network is 1.68×10^{-3}, the empirical networks have much more pronounced extreme betweenness values of $2.36 \times 10^{-2} \pm 8.23 \times 10^{-4}$. A visualization of the heavy-tailed betweenness distribution, as well as an analysis of Forman curvature, can be found in section F.1 in the supplemental material. Even sparse ground truth networks are highly sensitive to slight perturbations of the grid and network density, as there exist few important pathways connecting different regions in a sparse, locally connected network.

Figure 6 shows betweenness maps of networks, constructed from daily t2m as in Donges et al. (2009a), but on the approximately isotropic Fekete grid. The betweenness "backbone" fluctuates more in the sparse than in the dense networks. The maps in Fig. 6 and the maps presented in the original study

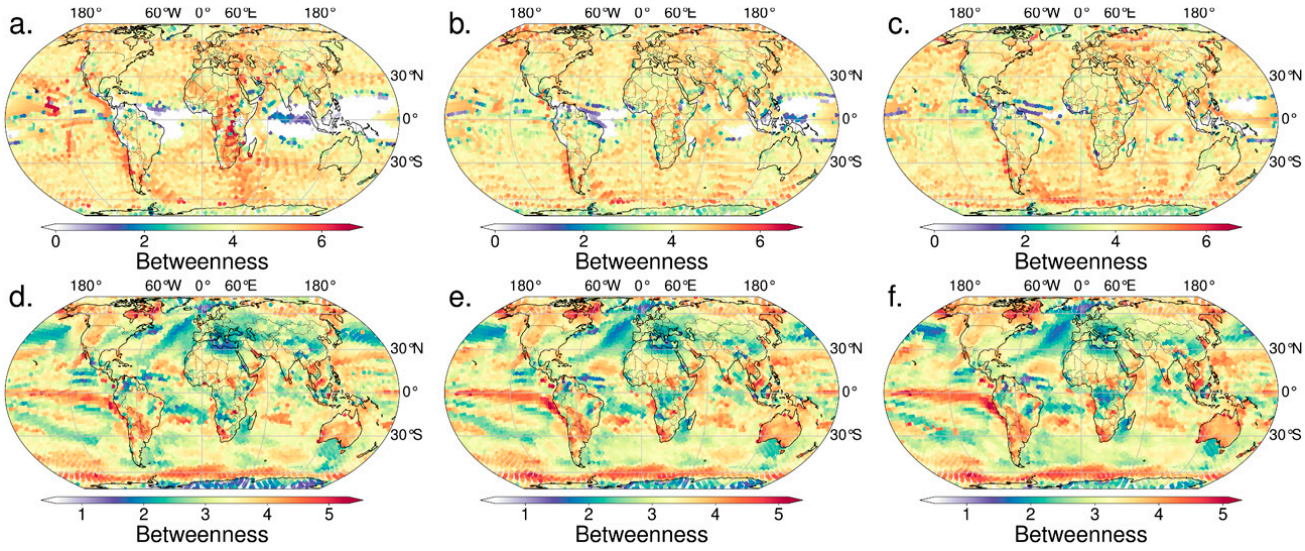

FIG. 6. Betweenness maps of temperature networks. Transformed betweenness [$\log_{10}$(BC + 1)] for daily t2m in a mutual information network (using the binning estimator) with various densities, as shown in (Donges et al. 2009a), but with (asymptotically isotropic) Fekete grid and ERA5 data between 1979 and 2019. (a)–(c) Network density = 0.004, 0.005, and 0.006, respectively. (d)–(f) Network density = 0.08, 0.1, and 0.12, respectively. Sparse networks fluctuate much more and have few pronounced extreme points. Sparse and dense networks look very different. The results in Donges et al. (2009a) also look very different from ours.

all look different because betweenness is unstable with respect to grid choice, dataset, and network density. This raises the question of which, if any, map shows a true betweenness "backbone." The dense networks (Figs. 6d–f) are more stable than the sparse networks (Figs. 6a–c) with respect to network density perturbation. But only after validating that the finite-sample network variability is also low (see section 4b) should patterns uncovered by network methods be interpreted with domain knowledge to generate novel insights. Here, a domain expert might point out stable ENSO-like patterns in the eastern Pacific in the stable dense networks (Figs. 6d–f) and get the inspiration to more closely investigate surprising patterns revealed by the betweenness map.

(iii) Consequences

Although forming sparse networks yields a better false discovery rate, some network characteristics become extremely sensitive to small perturbations of the networks and their explanatory power diminishes, even in ground truth networks. For each network measure of interest, the choice of network density constitutes a trade-off between the false discovery rate and the robustness of the measure. Here too, independent ensemble members can help to identify stable patterns (see section 4b). When a network measure fluctuates too much, as betweenness does in sparse networks, results should not be overinterpreted.

2) EMPIRICAL DISTRIBUTIONS OF NETWORK CHARACTERISTICS ARE DISTORTED

(i) Problem

Here, we present further perspectives on systematic empirical distortions of network measures. Most studies focus on the extremal nodes for any network measure, interpreting these as particularly important. Our simulation results show that, under data scarcity, *random nodes appear spuriously important in the empirical networks, not representing important nodes in the ground truth network*. Several studies have constructed Pearson correlation networks from sliding windows with 2.5° and finer resolution (Radebach et al. 2013; Hlinka et al. 2014; Fan et al. 2017, 2018, 2022). Our simulation results suggest that the naive correlation estimator and the short time scale are risk factors for false edges and distortions in global measures and extreme values of the networks.

(ii) Simulation results

Although all nodes have roughly the same degree/clustering coefficient in the ground truth graph, the observed degree/clustering coefficient distribution is more spread out in empirical networks (Fig. 7). The random distortions in empirical networks are similar in type and extent for different independent realizations. Spuriously extreme nodes in the empirical networks vary between independent realizations and do not reflect important or clustered nodes in ground truth graphs. While the average unweighted degree is consistent by construction, the weighted degree is systematically upward biased, as more links of low ground truth correlation are available that can be overestimated than links of large ground truth correlation that can be underestimated. The empirical (weighted) clustering coefficient is strongly downward biased, as spurious links connect otherwise disconnected regions and bundled connections are not formed between entire neighborhoods. Spurious teleconnections serve as shortcuts in the networks and lead to systematically smaller shortest pathlengths. Another network measure related to the clustering coefficient and shortest pathlengths is small-worldness. Since both network measures are extremely distorted in the empirical

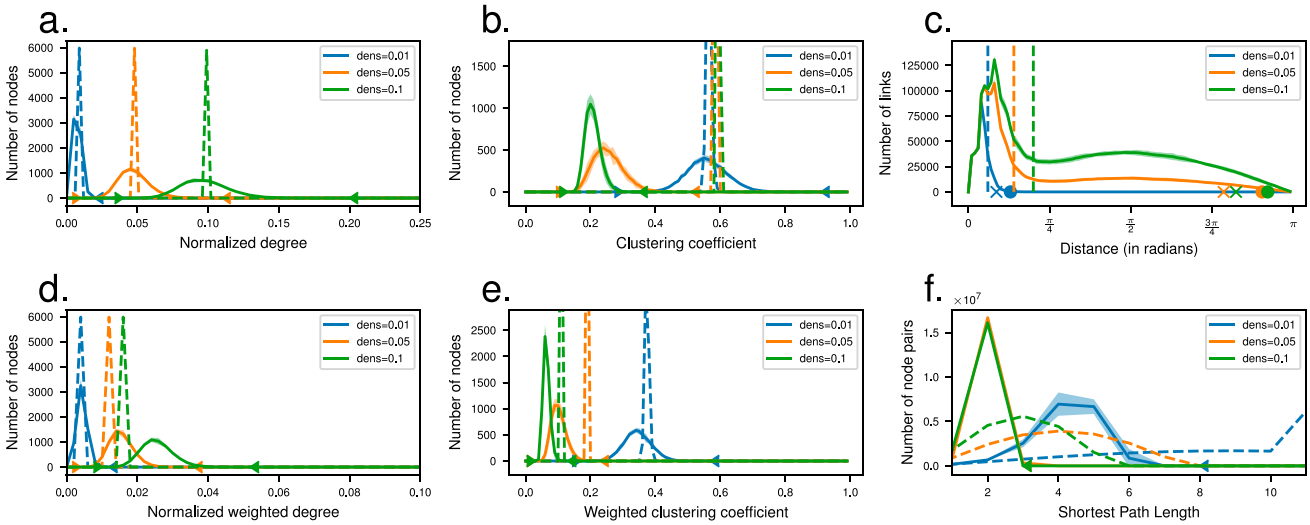

FIG. 7. Empirical vs ground truth distributions of node/edge measures. Empirical distributions of nodewise graph characteristics using Pearson correlation for an MIGRF($\upsilon = 0.5, \ell = 0.2$). (a) Normalized unweighted degree, (b) unweighted clustering coefficient, (c) unweighted link-length distribution, (d) weighted normalized degree, (e) weighted clustering coefficient (Onnela et al. 2005), and (f) unweighted shortest pathlength. Dashed lines denote the ground truth; solid lines denote the respective distribution in the empirical networks. Triangles denote the empirical extreme values; the $\times$ symbol and circle denote the 95% and 99% quantiles of a distribution averaged between independent realizations. The vertical lines in (c) denote the maximal link length in the ground truth graph; all longer links are false by design. Empirical distributions are more spread out; some measures such as the clustering coefficient or the shortest pathlength are heavily biased.

networks, a reliable conclusion about ground truth small-worldness cannot be drawn from the empirical networks in our setting. A more detailed treatment of small-worldness in spatially extended systems can be found in Bialonski et al. (2010) and Hlinka et al. (2017). The conclusions of both studies resemble ours. The length distribution of the spurious links (longer than the dashed lines in Fig. 7c) behaves as the number of available links at each distance: sinusoidal. This occurs when the corresponding true correlation values are empirically indistinguishable and has also been found in climate networks based on event synchronization (Boers et al. 2019). Under large length scales, empirical networks contain fewer erroneous edges and a more accurate link-length distribution up to higher network densities. On the other hand, empirical networks show larger spreads of degree, clustering coefficient, and shortest pathlength distributions, as false links occur in bundles (see section 3e). Under small length scales, bundling behavior is less pronounced, so that the amount of spurious links averages out, resulting in a more concentrated degree distribution, although more false links occur.

(iii) Consequences

When empirical networks are constructed with scarce data, they possess systematically different characteristics compared with the ground truth structure. In our setting, *distributions of popular node measures are more spread out as well as systematically biased*. These distortions do not become apparent by considering empirical summary statistics based on time series resampling because the empirical behavior remains consistent between independent repetitions. However, given multiple sufficiently independent network estimates, spurious and ground truth extreme nodes can be distinguished, depending on how systematically they reappear in several networks. Consequently, the distribution as well as the extreme values of network measures in empirical networks can primarily be the result of estimation errors and should not be overinterpreted. In particular, whenever the number of formed links scales with the number of available links, large estimation variability can be the cause, so researchers should make additional efforts to justify the correctness of their network when this link-length distribution arises, as in Boers et al. (2019).

e. *Local correlations give rise to spurious link bundles and high-degree clusters*

1) PROBLEM

As single false links occur with high probability in estimated networks, Boers et al. (2019) considered teleconnections in a climate network as significant only when a bundle of edges from one region to another is formed. As discussed in section 2c, this approach is unreliable when the underlying data are locally correlated, because edges tend to be formed in bundles. Spuriously dense regions in density-threshold graphs are another possible repercussion. Here, we provide empirical evidence of spurious bundling behavior.

2) SIMULATION RESULTS

Let us first define a link bundle between two locations. Intuitively, we demand that a sufficient portion of edge weight is formed between neighborhoods of both locations. Formally, let $B_\varepsilon(v) := \{u \in V | d(u, v) \leq \varepsilon\}$ be the ε-ball around location v

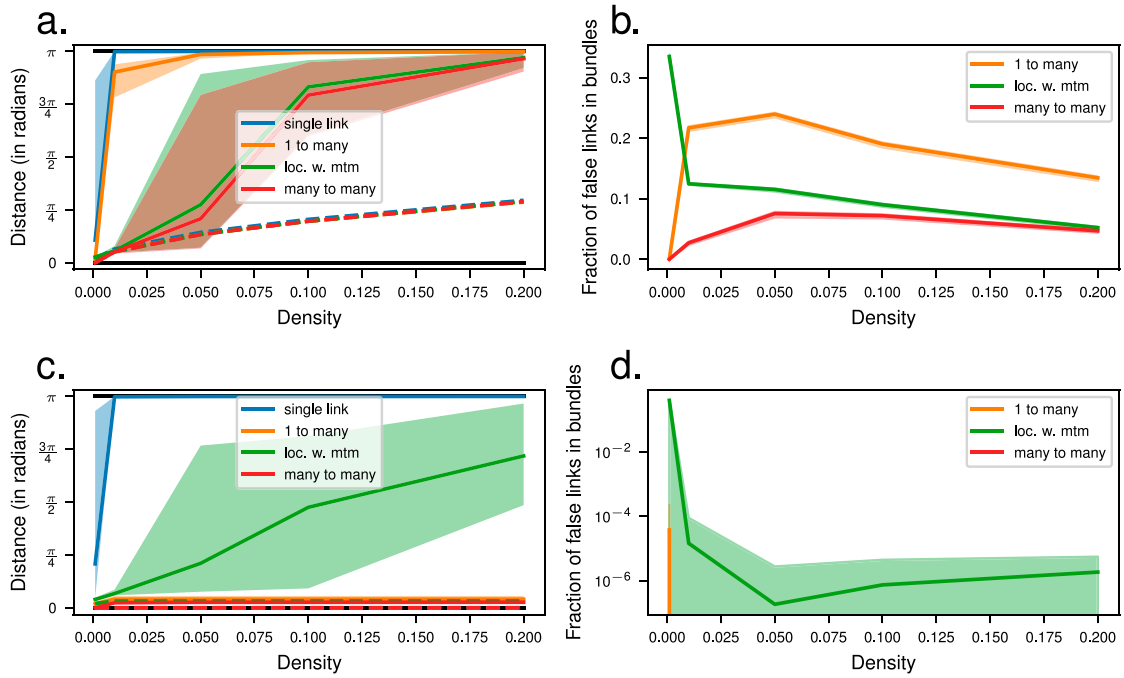

FIG. 8. Spurious link bundles frequently occur. (a),(c) Maximal length for which there exists at least one link bundle in the empirical Pearson network from an MIGRF($\upsilon = 0.5$, $\ell = 0.1$). The dashed line denotes the maximal length in the ground truth network. The definition of 1-to-many and locally weighted many-to-many (loc. w. mtm) link bundles can be found in section A.3 in the supplemental material. (a),(b) Unweighted and (c),(d) weighted link bundle notions. For weighted bundles, we choose $c = 0.5$; for unweighted 1-to-many bundles, $c = 0.9$; and for unweighted many-to-many bundles, $c = 0.8$. The radius of neighborhoods is chosen as $\varepsilon = 5°$, which corresponds to roughly 556 km. In our 2.5°-grid, the ε-balls contain 11.4 ± 1.1 nodes. (b),(d) The fraction of false links that belong to some link bundle among all false links; same setting as in (a) and (c). Many spurious long-range link bundles occur with high probability when the data are locally correlated, which is typically the case for spatiotemporal data. Strong teleconnections appear less frequently, so using edge weights can be helpful.

and let A be the graph adjacency matrix. We denote the cumulative weights between neighborhoods of v_i and v_j by

$$W_\varepsilon(v_i, v_j) := \sum_{k,\, l:\, v_k \in B_\varepsilon(v_i),\, v_l \in B_\varepsilon(v_j)} |A_{kl}|.$$

The number of edges between the regions $B_\varepsilon(v_i)$ and $B_\varepsilon(v_j)$ in the complete graph is denoted by $\rho_\varepsilon(v_i, v_j)$. We say that there is a (ε, c)-*many-to-many link bundle* between v_i and v_j in (V, E, A), if

$$\frac{W_\varepsilon(v_i, v_j)}{\rho_\varepsilon(v_i, v_j)} \geq c,$$

for some minimal connectivity $c > 0$. Of interest also might be a one-to-many and a locally weighted version of this notion (defined in section A.3 in the supplemental material).

Figure 8 shows the maximal distance of occurring link bundles (Figs. 8a,c) and the fraction of false links that belong to some bundle (Figs. 8b,d) for various notions of link bundles and for unweighted (Figs. 8a,b) and weighted (Figs. 8c,d) networks. The hyperparameters $\upsilon = 0.5$ and $\ell = 0.1$ amount to a weak local correlation structure and hence constitute an adversarial choice for bundling behavior. There is no unique definition in climate science literature of when an edge constitutes a teleconnection. Boers et al. (2019) calls an edge a teleconnection when it is longer than 2500 km or 0.12π radians; Kittel et al. (2021) sets the threshold at 5000 km or 0.25π radians. Irrespective of the exact distance, long 1-to-many link bundles already arise in unweighted networks with low density, while many-to-many link bundles consistently arise for intermediate network densities. Utilizing edge weights and tuning the minimal connectivity parameter c can reduce the number of spurious long-range link bundles by several orders of magnitude, because spuriously included links tend to lie marginally above the threshold and can therefore be distinguished from strong links. However, there is no hope of removing all spurious link bundles without removing true bundles as well. Since both true and false links naturally occur in bundles when the data are locally correlated, bundling properties cannot answer questions of significance. Results for other hyperparameters and mutual information are provided in section F in the supplemental material.

While for the FDR (Fig. 9a), large smoothness and length scale have a positive impact (because the random field varies less in total), the links that are being formed tend to occur in bundles. The essential distributional parameter for sparse networks is the smoothness of the random field, as mostly short

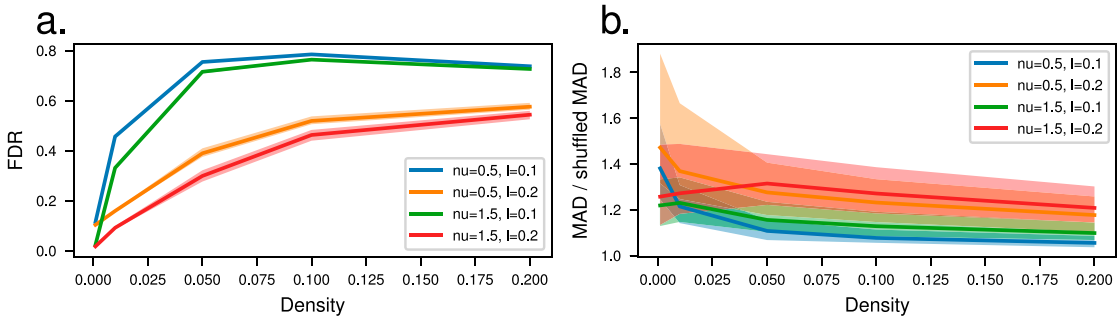

FIG. 9. Influence of random field parameters. (a) The fraction of false links (FDR) in empirical Pearson correlation networks for various hyperparameter choices of the random field. Sparse networks are more accurate in this sense (low FDR). For sparse networks, the smoothness is the essential parameter, while for denser networks, only the length scale matters. (b) Maximal average degree (MAD) in ε-balls divided by the same quantity under shuffled nodes in unweighted networks for various hyperparameter choices of the random field. Values above 1 indicate that high-degree nodes tend to be clustered. The weighted equivalent looks very similar.

links are formed. Varying the length scale has a larger impact on denser graphs as it determines the radius of spurious link bundles and the distance/network density at which ground truth correlations become empirically indistinguishable from 0. To measure whether there are regions of spuriously high degree due to the dependence between nodes, we find the ε-ball of maximal average degree (MAD) among all ε-balls $B_\varepsilon(v_i)$ (Fig. 9b). Then we compute the same quantity for randomly permuted degree values so that nodes with spuriously high degree are not spatially clustered anymore. The MAD values of the empirical networks are consistently larger than the MAD values of shuffled nodes, so a clustering of high-degree nodes occurs irrespective of the hyperparameters of the random field. The pronounced bundling behavior for larger length scales is reflected in larger MAD values.

For real data, observe a strikingly monotonous relationship between the average local correlation and the fraction of long links (longer than 5000 km or 0.25π radians) that belong to some link bundle (Fig. 10). Most differing links do not belong to a bundle, but under large local correlations the number of fluctuating long links in bundles can become nonnegligible. Also observe that our simulated data, at a given local correlation level, show a tendency to underestimate the fraction of links in bundles, indicating the existence of true bundled teleconnections in climatic variables.

3) CONSEQUENCES

We have seen that bundled connections do not necessarily represent ground truth structure but can occur spuriously when the similarity estimates are locally correlated. Even without teleconnections, random regions can appear spuriously dense. These experiments also explain the distortion of the degree distribution in Fig. 7. Using edge weights can be helpful to distinguish strong from weak connections, as spurious connections tend to lie marginally above the threshold. We conclude that when the data are locally correlated, questions of edge significance cannot be easily addressed by considering

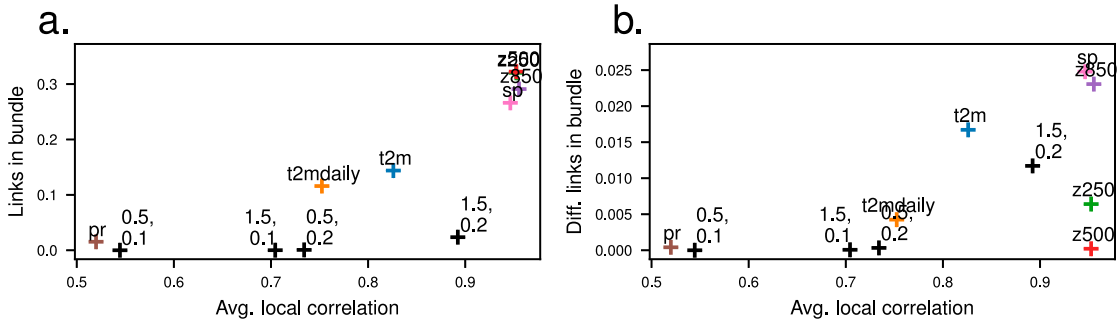

FIG. 10. Bundling behavior against local correlation. (a) Fraction of long links (longer than 5000 km or 0.25π radians) that belong to a many-to-many link bundle with $c = 0.8$ as a function of the average correlation in such a ball in Pearson correlation networks of density 0.05. A tuple a, b with a black plus symbol indicates simulated data with $v = a$ and $\ell = b$. With larger average local correlation, a larger fraction of long links is part of some bundle. Real data show even more bundling behavior than our simulated data; this indicates ground truth teleconnections. (b) Fraction of how many long links that differ between bootstrap samples belong to some bundle. Fewer differing links belong to some bundle, but their number is not negligible, especially under strong local correlation structure. Again, geopotential heights behave differently. Their correlation structure is so widespread that edge formation is highly codependent, but not very localized.

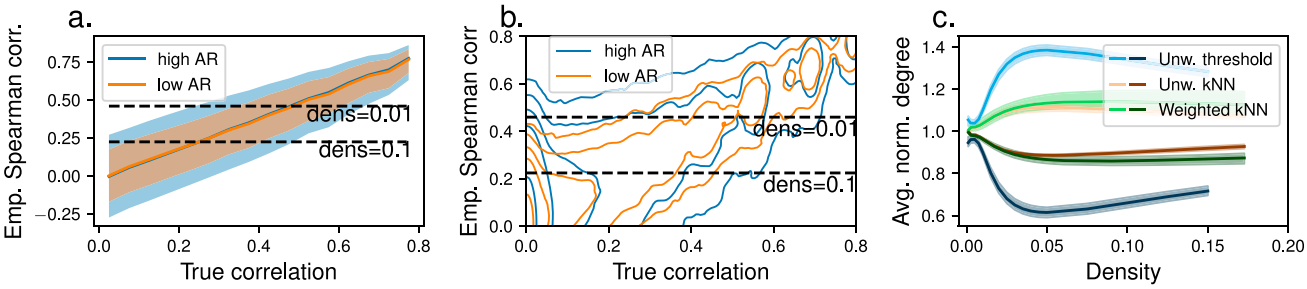

FIG. 11. Anisotropic variability of empirical estimates induces degree bias in empirical networks. We initialize a random half of the points with low lag-1 autocorrelation of 0.2 and the other half with high lag-1 autocorrelation of 0.7. Then, we analyze the distribution of Spearman correlation estimates for strongly and weakly autocorrelated nodes in an empirical network from an MIGRF($\upsilon = 1.5$, $\ell = 0.2$). (a) 5%, 50%, and 95% quantiles of the empirical Spearman correlations given the true correlation values. High autocorrelation does not introduce a bias in the correlation estimates but leads to larger variance. (b) Kernel density estimate of the edge distributions with high and low autocorrelation, respectively, at the levels 0.001, 0.01, 0.05, 0.1, and 0.5. Since most true correlation values are small, a larger variance causes a larger part of the edge distribution of the highly autocorrelated nodes to lie above the thresholds. This leads to a higher average degree for nodes with correlation estimates of higher variance. (c) Average degrees for threshold (blue), unweighted kNN (orange), and weighted kNN graphs (green) for nodes of high (light colors) and low (dark colors) autocorrelation normalized by the average degree in the network. Although kNN graphs cannot eliminate this bias, they can reduce it.

bundling behavior. Only bundling behavior that exceeds the effects of the localized correlation structure can be considered significant. Given multiple sufficiently independent empirical networks, only ground truth connections would reappear in many networks with high probability.

f. Anisotropy

1) Anisotropic autocorrelation on the nodes causes biased empirical degree

(i) Problem

Given two nodes with lag-1 autocorrelation $\alpha \in (0, 1)$ and $\beta \in (0, 1)$, the asymptotic variance $\sigma^2_{\alpha,\beta}$ of their empirical Pearson correlation scales as

$$\sigma^2_{\alpha,\beta} := 1 + 2\frac{\alpha\beta}{1 - \alpha\beta}, \tag{1}$$

which explodes for $\alpha, \beta \to 1$ (detailed explanation in section D in the supplemental material). Conversely, time length n is effectively only worth $n/\sigma^2_{\alpha,\beta}$ independent observations. The same principle applies to other estimators. Under anisotropic autocorrelation on the nodes, similarity estimates have different variability depending on the edge. While unweighted density-threshold networks are not biased by isotropic autocorrelation, anisotropic estimation variability introduces biases during the estimation procedure. Paluš et al. (2011) have already observed such biases in climate networks. Here, we explain them from a statistical perspective using our null model.

In practice, different locations have different autocorrelation patterns. Due to higher effect heat capacity, temperature over oceans has higher autocorrelation than over land (Eichner et al. 2003; Vallis 2011). Guez et al. (2014) argue that disagreement between their similarity measures is primarily caused by high autocorrelation. Our simulation results suggest that the cause of this disagreement might more fundamentally be estimation errors that vary between similarity measures.

(ii) Simulation results

We simulate anisotropic autocorrelation (Fig. 11) by employing our VAR(1) model (section B.1 in the supplemental material). We initialize a random half of the points with low lag-1 autocorrelation of 0.2 and the other half with a high lag-1 autocorrelation of 0.7. On average, empirical Spearman correlation estimates do not depend on the autocorrelation of adjacent nodes (Fig. 11a). However, the increased variance on highly autocorrelated nodes leads to both an increase of spuriously low similarity estimates for edges with high ground truth correlation, as well as more spuriously high estimates on edges with small ground truth correlation. Since most ground truth correlations are small (Fig. 11b), overall, the number of high similarity values increases. Thus, nodes of higher autocorrelation show an increased average degree in threshold graphs (Fig. 11c).

In real climate networks, the nodes of highest degree consistently have high lag-1 autocorrelation (Fig. 12). Together with our simulations, this suggests that anisotropic autocorrelation has nonnegligible spurious effects on the networks. Recalling Eq. (1), forming false links between highly autocorrelated nodes is much more likely than between nodes of small or intermediate autocorrelation. Hence, both false edges at nodes with high autocorrelation and missing edges at nodes with low autocorrelation have to be expected when some nodes attain autocorrelation values close to one, as for t2m.

(iii) Consequences

Under large isotropic autocorrelation, density-threshold networks have an increased variability but no degree bias. When the variability differs across locations, nodes with high variability receive more false edges than nodes with informative time series. State-of-the-art corrections are discussed in the next section. Using k-nearest neighbor (kNN) graphs prevents disregarding weakly autocorrelated locations. In kNN graphs, each node forms an edge to the k nodes with highest similarity. Although highly autocorrelated locations may still have more spuriously high empirical similarity values, weakly

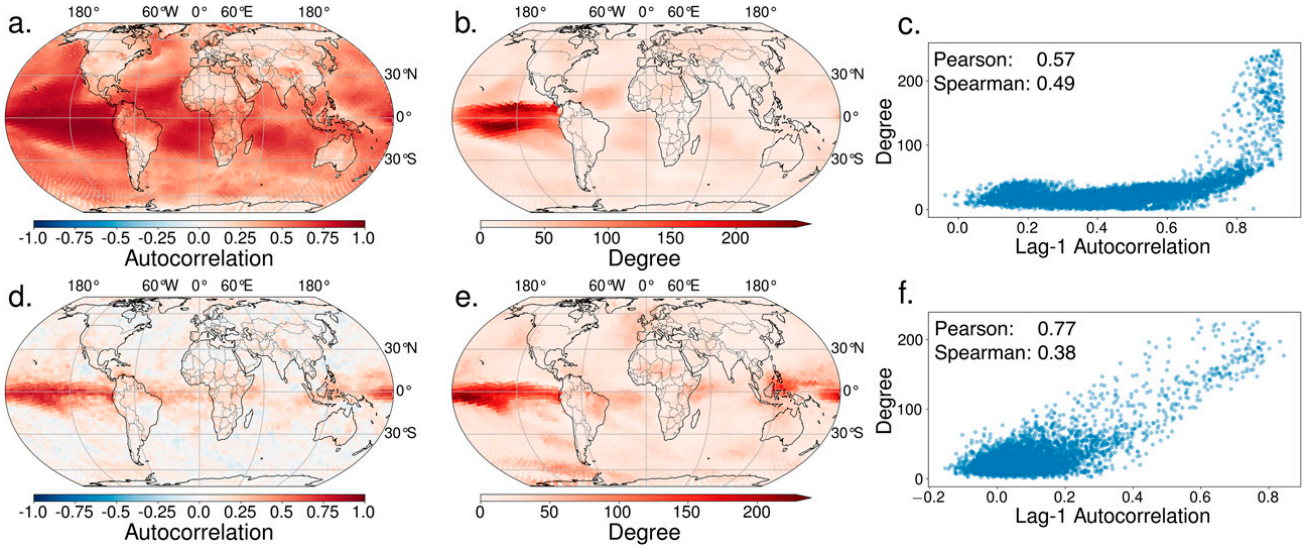

FIG. 12. Nodes of highest degree have high autocorrelation. (a),(d) The autocorrelation on the node and (b),(e) the degree in a Pearson correlation network of density 0.005 for (a), (b) monthly t2m and (d), (e) pr. (c), (f) The degree of each grid point as a function of lag-1 autocorrelation for (c) t2m and (f) pr. Observe exploding degrees for autocorrelation above 0.7 for t2m, as predicted by the variance scaling formula (1). For pr, degrees increase even earlier at autocorrelation 0.2. This suggests that the most connected nodes have many spurious edges, induced through high autocorrelation.

autocorrelated points attain more similar importance in terms of degree in unweighted as well as weighted kNN graphs.

2) FOURIER TRANSFORM-BASED RESHUFFLING REVERSES THE AUTOCORRELATION-INDUCED DEGREE BIAS IN SPARSE NETWORKS

(i) Problem

Instead of constructing density-threshold networks some, studies only include links that are significant with respect to similarity values from reshuffled data (Boers et al. 2014; Deza et al. 2015; Boers et al. 2019). For this purpose, the time series on each node is shuffled independently multiple times, and similarity values between these shuffled time series are calculated to determine the internal variability of the similarity estimates on each edge. High quantiles of this edgewise baseline distribution of similarity estimates impose restrictive thresholds above which ground truth dependence is likely. If the quantile and variance estimates are themselves noisy, yet one more source of randomness is introduced into the network estimation procedure.

(ii) Simulation results

We follow the popular approach to construct density-threshold networks not from the correlation estimates $\hat{S}_{ij}$ directly but from z scores $(\hat{S}_{ij} - \hat{\mu}^0_{ij})/\hat{\sigma}^0_{ij}$, where the edgewise mean $\hat{\mu}^0_{ij}$ and variance $\hat{\sigma}^0_{ij}$ are based on different reshuffling procedures. We compare completely random reshuffling of the time series, as in for example, Fan et al. (2022), and reshuffling with the iterative amplitude-adjusted Fourier transform (IAAFT) (Schreiber and Schmitz 1996), as proposed by Paluš et al. (2011). The IAAFT algorithm was developed by Schreiber and Schmitz (1996) to generate phase-randomized surrogate time series, which share their amplitude distribution and power spectrum with the original time series. We also construct "quantile networks" by including edges for which the empirical correlation value exceeds a high quantile of the edgewise baseline distribution. To measure the effect of autocorrelation on empirical correlation estimates, we simulate independent pairs of Gaussian time series of length $n = 100$ with varying autocorrelation and calculate the 95% quantile of empirical Pearson correlation (Fig. 13a). Naive reshuffling of the time series (blue) produces a baseline distribution of correlation estimates that does not adapt to the increased estimation variance (black line) under high autocorrelation, resulting in a too-permissive threshold and uncalibrated uncertainty. The IAAFT-based quantile estimates adapt to the increased variability but contain a large variance between independent realizations. To measure the impacts of anisotropic estimation variability on entire networks, we simulate a spatially anisotropic autocorrelation, as in the previous section. Since completely random reshuffling produces the same $\hat{\mu}^0_{ij}$ and $\hat{\sigma}^0_{ij}$ for every edge, the unweighted density-threshold network of z scores from random reshuffling exactly coincides with the unweighted density-threshold network from the original estimates $\hat{S}_{ij}$. The IAAFT surrogates remove the autocorrelation-induced degree bias for large densities (Fig. 13b). For sparse IAAFT z-score networks, the autocorrelation-induced degree bias is reversed: because the variance estimates $\hat{\sigma}^0_{ij}$ of nodes with high autocorrelation are systematically larger (section F.8 in the supplemental material), the highest z scores are formed for nodes with low autocorrelation. Since no edges are formed between nodes of high autocorrelation, the fraction of false links explodes in sparse IAAFT z-score networks (Fig. 13c). The IAAFT-based z-score networks fulfill the objective of only forming edges with small estimation variance, but then the resulting network does not represent the

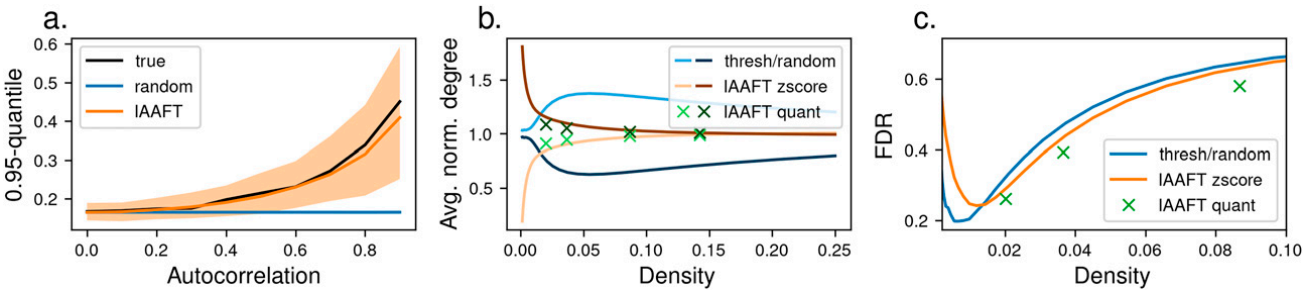

FIG. 13. Effects of (anisotropic) autocorrelation on significance-based networks. (a) Comparison of the true 0.95 quantile of empirical correlation values (black) [cf. Eq. (1)] with the 0.95 quantile obtained from naive nodewise shuffling (blue) and IAAFT nodewise shuffling (orange) on a single edge of zero ground truth correlation. We therefore calculate 10 000 shuffles of 1000 pairs of independent Gaussian time series of length $n = 100$ for each autocorrelation value. The growth of quantiles under high autocorrelation is not detected by the naive shuffling estimates. The IAAFT-based procedure detects increased variance but introduces another large source of variance between edge estimates. (b) Average normalized degrees of nodes with high (light colors) and low (dark colors) autocorrelation for unweighted threshold/z scores from random reshuffling (blue) vs networks from IAAFT-based z scores (orange) vs quantile networks from IAAFT surrogates (green) with density determined by the quantiles 0.9, 0.95, 0.99, and 0.999. The z scores from random reshuffling induce the original density-threshold network. The IAAFT surrogates correct the degree bias for large densities, but only form edges between nodes with low autocorrelation in sparse networks. (c) False discovery rates for unweighted threshold/z scores from random reshuffling (blue) vs IAAFT-based z scores (orange) vs IAAFT-based quantile networks (green). IAAFT-based z scores form many false edges in sparse networks. Quantile networks outperform z-score networks given the same density but do not reach sufficient sparsity.

spatial ground truth correlation structure. Quantile networks perform slightly better than z-score networks given the same density, but there is no reasonable significance value ≤ 0.999 that achieves a network sparsity necessary for minimal FDR.

In the following way, perfect quantile estimates allow us to determine network densities that lead to empirical networks containing few false edges. Under isotropic autocorrelation, perfect quantile estimates induce a common threshold on the entire network. Applying this threshold to the ground truth correlation matrix induces a density in the ground truth network. For our range of hyperparameters, the densities, induced by the 0.95 quantile, range from 0.007 (for $v = 0.5$, $\ell = 0.1$) to 0.032 (for $v = 1.5$, $\ell = 0.2$) without autocorrelation and from 0.001 to 0.01 given isotropic autocorrelation of 0.9. Choosing larger network densities leads to unreliable empirical networks, as ground truth correlations of longer links are not empirically distinguishable from 0 with sufficient certainty. Generally, empirical significance-based networks have a larger density than they should, because the number of spuriously high correlation estimates exceeds the number of spuriously low ones.

(iii) Consequences

In settings of low autocorrelation, completely random reshuffling yields reliable estimates of empirical correlation quantiles, resulting in a controlled false discovery rate. However, it is not able to detect anisotropic autocorrelation and can therefore not correct autocorrelation-induced degree bias. The IAAFT-based empirical networks correct this bias in dense networks, but are very biased toward edges with low estimation variance in sparse z-score networks, which results in many false links. The empirical density of significance-based networks only yields an upper bound on a desirable network density. Since *no considered network-construction technique reduces the variance in the edge estimates*, they cannot vastly improve on density-threshold networks.

3) ANISOTROPIC NOISE LEVELS ON THE NODES CAUSE NODES TO BE DISCONNECTED

(i) Problem

Observational data are generally affected by measurement errors or other sources of noise. Under isotropic additive white noise, variance in the graph construction increases (see section F.3 in the supplemental material). Even worse, anisotropic noise levels crucially distort how well nodes are connected in the graph.

A central difficulty in recovering ground truth structure is distinguishing which part of the noise is inherent to the dynamical system (aleatoric noise) and which part could be reduced through more sophisticated measurement, preprocessing, and estimation procedures (epistemic noise). While aleatoric noise affects ground truth networks and can be seen as an offset of the ground truth correlation function, everything else is an empirical distortion. Nodes over land are commonly less connected in climate networks (Donges et al. 2009b) because the underlying distributional characteristics differ across sea and different geological conditions over land. This distributional difference is at least partially aleatoric. Varying availability and reliability of measurements, on the other hand, induce epistemic noise.[1] In cases where we acknowledge that we cannot satisfactorily judge

[1] In reanalysis datasets, data over nodes in regions of high measurement density can be extrapolated with higher certainty. The density of weather stations in the United States or Europe, for example, is much higher than in parts of Africa or South America. The effort of estimating the measurement/extrapolation error in each node could alleviate the effects of an anisotropic data collection and extrapolation process. Anisotropic measurement/extrapolation noise remains to distort the constructed climate networks, and efforts should be made to gather more reliable measurements in neglected regions [cf. overview of WMO weather stations (ArcGIS 2022)].

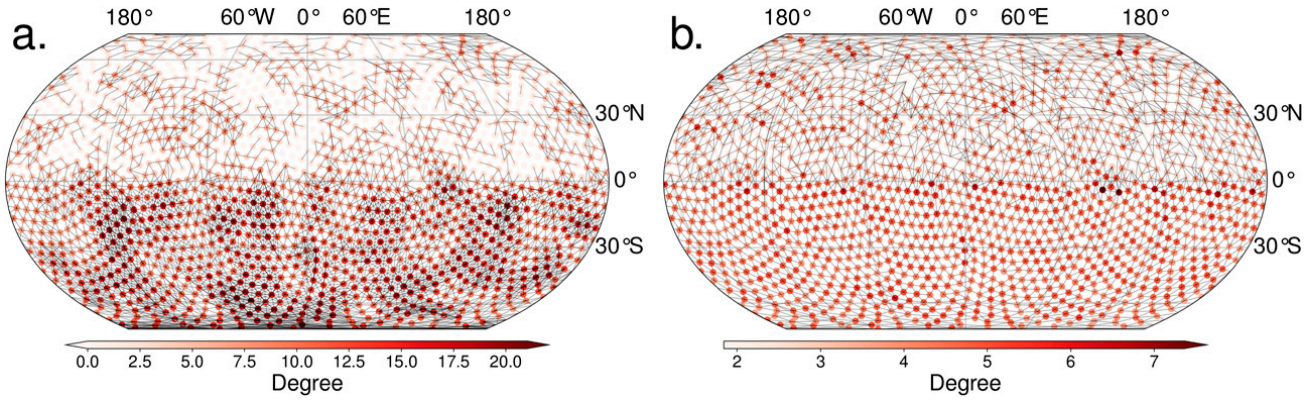

FIG. 14. Threshold vs kNN networks given anisotropic noise levels. (a) Unweighted threshold graph and (b) weighted kNN graph with approximately 0.005 of possible edges formed from an MIGRF($\nu = 1.5$, $\ell = 0.2$) and 5° resolution with additive $0.7 \times N(0, 1)$ white noise on the Northern Hemisphere. Higher measurement noise on the Northern Hemisphere leads to smaller correlation values. These nodes are less connected in the empirical as well as ground truth threshold graph. A kNN graph over the same data ensures similar total connectedness. Nodes on the Northern and Southern Hemispheres can still be differentiated in terms of weighted degree. Additionally, spurious high-degree clusters in the empirical threshold graph are not present in the weighted kNN graph.

how much our data are affected by epistemic noise, a conservative approach is to reduce the effects of anisotropy in the network construction. kNN graphs may offer a useful inductive bias in such uncertain settings.

(ii) Simulation results

By adding white noise on the Northern Hemisphere, we decrease the population correlation for these nodes. As a result, we find that (especially sparse) threshold graphs mostly form edges on the nodes with less noise (Fig. 14). To represent all nodes equally in the graph, we propose to use kNN graphs instead. By using weighted edges, the spectrum of node and link importance in terms of weighted degree does not get lost.

(iii) Consequences

When the available data are affected by epistemic noise, the connectivity structure in the network is spuriously altered. Effects of anisotropic noise on the empirical networks can be reduced by using kNN graphs. When the ground truth correlations are higher in some regions than in others (anisotropic aleatoric noise), kNN graphs can also be more informative because weakly correlated nodes are not well represented in density-threshold networks. kNN graphs pose a different inductive bias, which may be useful to detect different patterns. Although ground truth kNN graphs severely differ from ground truth density-threshold networks in anisotropic settings, given useful weights, they have shown to be useful and robust in machine-learning applications (von Luxburg 2007), while not sacrificing interpretability.

4) GROUND TRUTH NETWORKS ON ANISOTROPIC GRIDS

(i) Problem

Anisotropic grids usually introduce biases in networks that are not intentional, so that differing node connectivity does not reflect differing correlation structure in the data. Given an anisotropic grid, the nodes will have unequal characteristics in the ground truth network under isotropic correlation structure. It is well known that a regular Gaussian grid is geometrically undesirable due to its two singularities at the poles. Area weighting (Heitzig et al. 2011) becomes crucial to correct the distortions in the network. Another effect of anisotropy does not stem from anisotropic grid choice but from geographical reality. If we consider an isotropic field with a monotonically decaying correlation function on an approximately isotropic grid only defined over oceans, then the nodes in the population network will not be isotropic but will encode geometric information about the distribution of land and sea across Earth (Fig. 15). For example, sea surface temperatures are only defined over oceans.

(ii) Simulation results

The ground truth networks constructed from a monotonically decaying isotropic correlation structure simply consist of the shortest possible links. The anisotropic distribution of grid points introduces a bias to the networks that is visible in various network measures. For example, points on paths connecting different regions and points in geometric bottlenecks show higher betweenness values in sparse networks, points with large uninterrupted surroundings show higher degree, and points in inlets show larger clustering coefficient, because neighbors toward similar directions are often close to each other and thus also connected. The network density functions as a scale parameter similar to a bandwidth in kernel-density estimation, since the connection radius increases with network density.

(iii) Consequences

Estimated networks suggest misleading conclusions when false edges distort their characteristics. Even the ground truth network is a result of many design decisions that can lead to prominent behavior, readily misinterpreted when its cause is not correctly identified. For betweenness, even the ground truth values are very sensitive to small variations in network

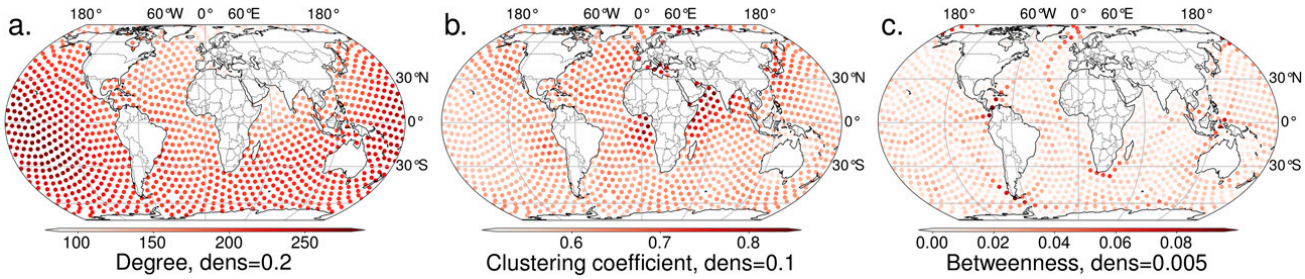

FIG. 15. Anisotropic grids induce anisotropic node characteristics. Node characteristics of ground truth networks from monotonically decaying isotropic correlation structure. (a) Degree, (b) clustering coefficient, and (c) betweenness for unweighted density-threshold graphs with density 0.2, 0.05, and 0.005, respectively.

density. Even in ground truth, conclusions are not necessarily robust. Boundary correction (Rheinwalt et al. 2012) has been proposed for networks that do not cover the entire Earth. A similar correction, using locally connected networks, could be proposed to remove the influence of the distribution of land and sea across Earth to quantify anisotropic correlation behavior.

4. Assessing significance from network ensembles

In practice, researchers are usually confronted with datasets of limited size from an unknown distribution. Based on a single network constructed with state-of-the-art climate network techniques, they cannot judge how many edges are included or excluded because of estimation errors. Given time length n and number of grid points p, the regime of "small sample size," where the observed networks significantly differ from the ground truth, can mean any order of magnitude for n and any ratio n/p, depending on the dynamics of the spatiotemporal system, measurement error, the employed estimator, and the subsequent network construction and evaluation steps. This makes general rules of thumb prohibitive, and solid uncertainty estimation based on unrestrictive assumptions crucial for the value of the study. Constructing networks with various similarity measures, datasets, resolutions, and network-construction steps (see, e.g., Radebach et al. 2013) can offer qualitative reassurance that observed patterns do not just occur under a specific setting. Significance tests offer a more quantitative approach. Here, we first discuss the shortcomings of common procedures to quantify significance in section 4a, and then offer a new probabilistic framework in section 4b that addresses these shortcomings.

a. *Resampling in current practice*

The usual approach to quantify the significance of certain findings such as hubs, pathways or teleconnections is to construct an ensemble of networks that share certain aspects of the originally constructed network while randomizing with respect to everything else through reshuffling. The effective null hypothesis of such a permutation test (also called a surrogate test) is the (limit) probability distribution over the networks that the ensemble induces. Needless to say, any permutation test can only be as meaningful as its effective null hypothesis.

All previously applied reshuffling approaches for climate networks that have been reported in the literature can be categorized into two types. Either reshuffling is directly performed on the edges to recover, for example, the original degree sequence or the degree sequence and link-length distribution (Wiedermann et al. 2015, GeoModel II), or the time series are shuffled nodewise to preserve the marginal time series dynamics with methods such as the iterative amplitude-adjusted Fourier transform (Schreiber and Schmitz 1996). In the latter case, the ensemble networks are then constructed from the shuffled dataset.

1) NODEWISE RESHUFFLING

Whenever researchers have performed permutation tests that recover marginal time series dynamics, these tests have disregarded the spatial distribution of the data completely. The nodes are assumed to be independent, so that the typical localized correlation structure, which results in a link-length distribution of predominantly short links, is replaced by a uniform one (Fig. 16a). Since the task is to construct *spatial* networks, such an ensemble is structurally unrealistic and does not induce a physically meaningful network distribution.

2) EDGE RESHUFFLING

Whenever fixing concrete network characteristics to be preserved, a preliminary question has to be addressed. Which spatial, as well as temporal, dynamics of the system or network at hand need to be preserved by the ensemble? The authors of the influential paper (Donges et al. 2009a) use a permutation test with preserved degree sequence. But what is the physical meaning of the observed degree sequence? One consistently reappearing feature of the underlying physical system is the localized correlation structure, which results in a link-length distribution of predominantly short links. As for nodewise reshuffling, this link-length distribution is destroyed and is replaced by a sinusoidal one. As a consequence of this inaccurate ensemble distribution, the authors interpret the property that the nodes of highest betweenness show degrees below average in the original network as significant behavior. In contrast, we have shown in Fig. 5 that this property is a bias that is introduced through the Gaussian grid and the latticelike connectivity behavior of the original network. The original connectivity structure gets destroyed by uniform edge

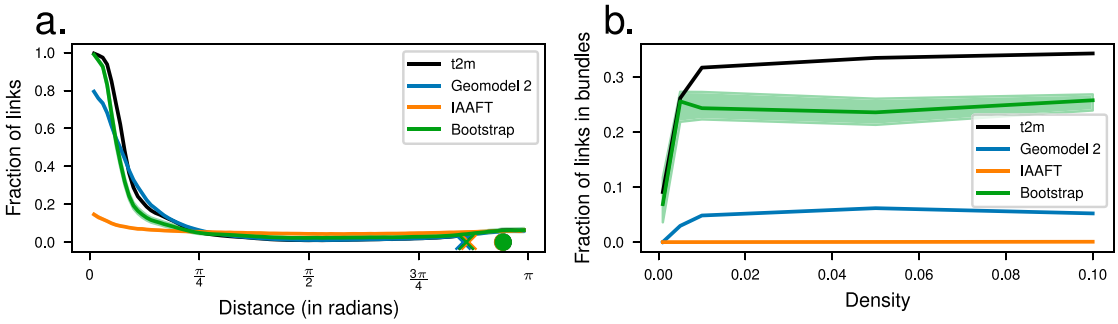

FIG. 16. Reshuffling procedures produce unrealistic network distributions. (a) The fraction of formed links among all possible links at a given distance in a Pearson correlation network for t2m. We employ Geomodel 2 with $\varepsilon = 0.05$ and min(10^7, 10 × number of edges) rewirings. Nodewise IAAFT-resampling results in approximately uniform link distribution. A naive bootstrap in time over all grid points simultaneously as well as Geomodel 2 results in a more realistic link-length distribution. (b) The fraction of links in 1-to-many link bundles. Nodewise reshuffling, as well as Geomodel 2, destroys network properties that are induced by localized correlation structure.

reshuffling; hence, the ensemble members have different betweenness properties. We have seen that betweenness is a highly unstable measure. An indication for robustness of the discovered betweenness "backbone" would be if it consistently reappeared for various subsets of the data, as well as for many network densities and similarity estimators. Since only a single instance of the climate network is presented, it is not clear if the presented backbone appears by chance.

A first improvement over random link redistribution or independence between the time series is GeoModel II (Wiedermann et al. 2015), which approximately preserves both the degree distribution and the nodewise link-length distribution. But it still does not recover the natural tendency in the network to form bundled links. Given that one link is formed, the likelihood of a neighboring link increases in a locally correlated random field. When simply recovering the total link-length distribution, as does GeoModel II, this likelihood does not increase (Fig. 16b). Furthermore, fixing the degree sequence of the nodes might not be representative of the distribution of the constructed graph. In section 3e, we have seen spurious high-degree clusters on random locations. In conclusion, no explicit resampling scheme has been proposed that recovers the joint link distribution of locally correlated random fields.

Proposing a concrete network resampling scheme always runs the risk of missing or distorting an important aspect of the underlying dynamical system and estimation procedure. The tendency to form bundled connections depends on the localized correlation structure. The expected number of spurious links depends on the (co)variability of the estimates, the spatial correlation structure of the random field, and the autocorrelation of the time series, among many other factors. All these aspects do not even cover the more complex time series dynamics that we might want to account for.

b. *Distribution-preserving ensembles*

Instead of trying to solve the impossible problem of finding which exact characteristics to preserve for the climatic question of interest, we propose to *construct network ensembles such that all members approximately reflect the network distribution that originates in the underlying physical processes.* As discussed above, state-of-the-art network resampling approaches calculate many estimates of surrogate networks that do not reflect the original distribution. Crucially, they only obtain *one* noisy estimate of the similarity value on each edge. With multiple estimates, we not only get a more robust total estimate but an approximation of the estimation variability on each edge. Instead of a single network estimate, we have access to an ensemble of equally valuable estimates that allows us to judge whether a network estimation procedure really is trustworthy and empowers us to tackle most of the issues presented in section 3. Instead of reshuffling the data nodewise, we propose to jointly subsample or resample time-windows for both end points of an edge or even for all nodes simultaneously, preserving the original dynamics in space. There is an abundance of resampling techniques for multivariate time series. One popular approach is block bootstrapping (Lahiri 2003; Shao and Tu 1995). Another is subsampling (Politis et al. 1999). Individual ensemble members should both reflect the *same data distribution* (induced by the selection of included time points), as well as be sufficiently *independent* (by representing different time-windows that are far enough apart). Such an ensemble could be constructed by the following pipeline:

(i) Decide on a network-construction procedure with unbiased edge estimates, such that the dynamics of interest behave robustly within the network distribution.
(ii) Construct many networks with the same procedure (i) by subsampling/block bootstrapping data in time on all grid points simultaneously.
(iii) Evaluate reoccuring edges and patterns such as link bundles.

If the quantity of interest can be expressed by summary statistics, the pipeline as a whole should be unbiased to yield calibrated confidence intervals. We have seen that the maximal degree or the link-length distribution of the networks can be systematically biased. Crucially, estimates of single edges or fixed neighborhoods are unbiased when the employed similarity

estimator is unbiased. In this case, given a large enough ensemble, uncertainty estimates and p values would be precise. In practice, unbiased estimators often increase estimation variance so much that the uncertainty becomes too large. In complex estimation tasks like mutual information estimation, this approach might reveal that empirically constructed networks are too incoherent due to a lack of available data. Instead of communicating false certainty, the bootstrap approach would suggest conclusions like "the amount of available data does not suffice to significantly detect this teleconnection with our small-bias mutual information estimator."

From a dynamical systems perspective, one could argue that different time points represent a different state of the dynamical system. From a statistical perspective, a distribution over networks is implicitly chosen when selecting the dataset. The remaining task is to define what characterizes the state of the dynamical system one is interested in, such as different ENSO phases. If we construct an ensemble that is independent of the state, we simply recover links and patterns that are present most of the time. Teleconnections that are only active some of the time become difficult to distinguish from noisy connections.

A single network that is more accurate than the result of a single similarity estimate per edge can be derived by selecting the edges that appear in most ensemble members. A related approach is the variable selection method *stability selection* (Meinshausen and Bühlmann 2010). In practice, stability selection often markedly improves the baseline variable selection or structure estimation algorithm. Another approach could only accept links with small estimation variability.

While most studies directly perform bootstrapping on the graph structure and not on underlying data (Chen et al. 2018; Levin and Levina 2019), a similar idea has been previously suggested (Friedman et al. 1999). To the best of our knowledge, it has not yet been applied in climate science. In practice, data scarcity, distribution shift, and varying regimes of the dynamical system complicate finding a suitable resampling or subsampling scheme that produces both sufficiently independent and identically distributed ensemble members without biases. In both bootstrap as well as subsampling techniques, a suitable choice of window size depends on the autocorrelation of the time series at hand. Highly anisotropic autocorrelation (Fig. 12) complicates designing a consistent procedure for all nodes simultaneously. With these complications and the journal's page limit in mind, we postpone proposing an explicit ensemble construction procedure to future work.

5. Conclusions

When constructing networks from data, it is not obvious that they reflect ground truth structure. Given a finite amount of data, similarity estimates contain estimation errors. Under such nonnegligible estimation variability, we find several types of spurious behavior using typical network-construction schemes:

- Not only the choice of similarity measure, but also the choice of estimators, is an influential design decision. The properties of the estimator determine how well single empirical networks approximate the population network and if the uncertainty of an ensemble is accurate.
- Global properties of finite-sample networks such as averages, variances, and maxima of network measures, or the spectrum of the adjacency matrix, can be heavily distorted.
- Links occur in bundles when the data are locally correlated and the estimator transmits the correlation structure. This leads to spurious link bundles and regions of spuriously high or low degree.
- Under anisotropic autocorrelation or marginal distributions, differing data distributions on the nodes cause anisotropic estimation variability on the edges, which in turn introduces biases in the empirical networks. Anisotropic noise levels may lead to weak representation of nodes in the network, weighted kNN graphs reduce anisotropic behavior via the inductive bias to represent all nodes equally, and differences can still be detected via edge weights.
- We find sparse networks to be more accurate in terms of false discovery rate and spurious teleconnections. Yet popular network measures such as betweenness become highly unstable in sparse networks. This constitutes a different trade-off for each estimation task. Random fields with larger length scales allow for denser networks, but also lead to more pronounced bundling behavior.

Given the variety and extremeness of possible empirical distortions, it is crucial to reliably estimate how "trustworthy" an empirical network is. State-of-the-art resampling procedures only capture particular parts of the empirical network distribution, and consequently miss other possibly relevant aspects of the dynamical system. When the implicit null hypothesis of the resampling technique does not capture all relevant properties of the dynamical system, the value of the significance test is questionable. Specifically, surrogate tests, which hypothesize independence between nodes, do not reflect a physically meaningful null hypothesis when the dynamical system is locally correlated, which the link-length distribution typically clearly reflects. Random artifacts that stem from local correlation structures will then appear significant. In the past, climate network approaches have been based on calculating a single similarity estimate on each edge; we propose to generate multiple estimates via sub- or resampling in time, in order to estimate the estimation variance on each edge. This allows us to approximate the intrinsic distribution over the constructed networks induced by the underlying data distribution and the chosen estimation procedure.

Future work

Most importantly, future network studies in climate science should estimate the underlying estimation error in each edge in order to argue about significance in a statistically meaningful way. Given scarce data, the variability of similarities cannot be precisely estimated (Fig. 13a). Further challenges inherent to climatic data have to be addressed to successfully implement our proposed framework of network ensembles, which

represent the underlying dynamics by design and are sufficiently independent in time. Adjusting the window length for block bootstrapping on each edge to the autocorrelation of the nodes could yield consistent estimates of the estimation variance on all edges. More work is needed to develop good similarity estimators, robust network-construction procedures, and resampling techniques for the climate context, respecting distribution shifts, varying regimes, anisotropy, and measurement errors in chaotic dynamical systems.

While we only construct undirected networks, directed empirical networks as well as EOFs suffer from insufficient data in analogous ways. From estimating lagged correlations to probabilistic graphical models (Koller and Friedman 2009) and causal networks (Runge et al. 2019a), a similar simulation analysis could quantify the strengths and weaknesses of different network-construction procedures. Some works have reduced the network size by clustering spatial areas of temporally coherent behavior before network-construction (cf. Rheinwalt et al. 2015; Fountalis et al. 2018; Runge et al. 2019b). Whether such an approach is statistically beneficial is task dependent and deserves a more thorough consideration in future work. On the one hand, too many nodes may induce many errors and a systematic distortion in the spectrum of empirical covariance matrices (Donoho et al. 2018; Lam 2020; Morales-Jimenez et al. 2021); on the other hand, single errors have a higher impact in smaller networks, and node estimation constitutes yet another challenging task in the network-construction pipeline. A key question in this regard is: how can we minimize the edgewise estimation variance in the downsized network? Previous work has selected representative locations for each cluster, but maybe an aggregation of regional information can boost statistical robustness further.

More complex time dynamics could introduce other kinds of spuriousness into empirical estimates we were not able to cover with our simple autoregressive model. A theoretical analysis of networks from spatiotemporal data would also be very insightful. Exploring alternatives to mutual information, such as the Hilbert–Schmidt independence criterion or the randomized information coefficient, could yield novel insights into the dynamics of the climatic system.

Networks have been constructed in several scientific fields to detect complex structures in spatiotemporal data. Geneticists try to identify connections between certain genes and the development of diseases by estimating Pearson correlation networks under the term weighted gene coexpression analysis (Horvath 2011; Niu et al. 2019). Neuroscientists (Sporns 2010) aim to understand the functional connectivity in the brain with weighted voxel coactivation network analysis (Mumford et al. 2010). While in this work we focus on the application of functional networks in climate and geoscience, our conceptual findings hold in any domain where networks are constructed from spatiotemporal data.

Acknowledgments. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (EXC 2064/1—Project 390727645). The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Moritz Haas. We thank all members of the Theory of Machine Learning and the Machine Learning in Climate Science groups in Tübingen for helpful discussions. We thank all reviewers and our editor for valuable feedback. Finally, we thank Joe Guinness for his advice concerning the use of Matérn covariance on the sphere.

Data availability statement. Hersbach et al. (2018, 2019a,b) were downloaded from the Copernicus Climate Change Service Climate Data Store. Python code for reproducing all results in this paper can be found at https://github.com/moritzhaas/climate_nets_from_random_fields/.

REFERENCES

Agarwal, A., L. Caesar, N. Marwan, R. Maheswaran, B. Merz, and J. Kurths, 2019: Network-based identification and characterization of teleconnections on different scales. *Sci. Rep.*, **9**, 8808, https://doi.org/10.1038/s41598-019-45423-5.

ArcGIS, 2022: Map of WMO weather stations. ArcGIS, accessed 24 October 2022, https://www.arcgis.com/apps/mapviewer/index.html?layers=c3cbaceff97544a1a4df93674818b012.

Barber, C., J. R. Lamontagne, and R. M. Vogel, 2019: Improved estimators of correlation and R^2 for skewed hydrologic data. *Hydrol. Sci. J.*, **65**, 87–101, https://doi.org/10.1080/02626667.2019.1686639.

Bendito, E., A. Carmona, A. M. Encinas, and J. M. Gesto, 2007: Estimation of Fekete points. *J. Comput. Phys.*, **225**, 2354–2376, https://doi.org/10.1016/j.jcp.2007.03.017.

Bialonski, S., M. T. Horstmann, and K. Lehnertz, 2010: From brain to earth and climate systems: Small-world interaction networks or not? *Chaos*, **20**, 013134, https://doi.org/10.1063/1.3360561.

Boers, N., B. Bookhagen, N. Marwan, J. Kurths, and J. Marengo, 2013: Complex networks identify spatial patterns of extreme rainfall events of the South American monsoon system. *Geophys. Res. Lett.*, **40**, 4386–4392, https://doi.org/10.1002/grl.50681.

——, ——, H. M. J. Barbosa, N. Marwan, J. Kurths, and J. A. Marengo, 2014: Prediction of extreme floods in the eastern central Andes based on a complex networks approach. *Nat. Commun.*, **5**, 5199, https://doi.org/10.1038/ncomms6199.

——, B. Goswami, A. Rheinwalt, B. Bookhagen, B. Hoskins, and J. Kurths, 2019: Complex networks reveal global pattern of extreme-rainfall teleconnections. *Nature*, **566**, 373–377, https://doi.org/10.1038/s41586-018-0872-x.

Cellucci, C. J., A. M. Albano, and P. E. Rapp, 2005: Statistical validation of mutual information calculations: Comparison of alternative numerical algorithms. *Phys. Rev. E*, **71**, 066208, https://doi.org/10.1103/PhysRevE.71.066208.

Chen, Y., Y. R. Gel, V. Lyubchich, and K. Nezafati, 2018: Snowboot: Bootstrap methods for network inference. *R J.*, **10**, 95–113, https://doi.org/10.32614/RJ-2018-056.

Chwialkowski, K., and A. Gretton, 2014: A kernel independence test for random processes. *Proc. 31st Int. Conf. on Machine Learning*, Vol. 32, Beijing, China, Association for Computing Machinery, 1422–1430, https://dl.acm.org/doi/10.5555/3044805.3045051.

Cressie, N. A. C., 1993: *Statistics for Spatial Data*. Wiley Series in Probability and Statistics, Wiley, 900 pp.

Deza, J. I., M. Barreiro, and C. Masoller, 2015: Assessing the direction of climate interactions by means of complex networks and information theoretic tools. *Chaos*, **25**, 033105, https://doi.org/10.1063/1.4914101.

Donges, J. F., Y. Zou, N. Marwan, and J. Kurths, 2009a: The backbone of the climate network. *Europhys. Lett.*, **87**, 48007, https://doi.org/10.1209/0295-5075/87/48007.

——, ——, ——, and ——, 2009b: Complex networks in climate dynamics. *Eur. Phys. J. Spec. Top.*, **174**, 157–179, https://doi.org/10.1140/epjst/e2009-01098-2.

——, and Coauthors, 2015: Unified functional network and nonlinear time series analysis for complex systems science: The pyunicorn package. *Chaos*, **25**, 113101, https://doi.org/10.1063/1.4934554.

Donoho, D., M. Gavish, and I. Johnstone, 2018: Optimal shrinkage of eigenvalues in the spiked covariance model. *Ann. Stat.*, **46**, 1742–1778.

Eichner, J. F., E. Koscielny-Bunde, A. Bunde, S. Havlin, and H. Schellnhuber, 2003: Power-law persistence and trends in the atmosphere: A detailed study of long temperature records. *Phys. Rev. E*, **68**, 046133, https://doi.org/10.1103/PhysRevE.68.046133.

Ekhtiari, N., A. Agarwal, N. Marwan, and R. V. Donner, 2019: Disentangling the multi-scale effects of sea-surface temperatures on global precipitation: A coupled networks approach. *Chaos*, **29**, 063116, https://doi.org/10.1063/1.5095565.

——, C. Ciemer, C. Kirsch, and R. V. Donner, 2021: Coupled network analysis revealing global monthly scale co-variability patterns between sea-surface temperatures and precipitation in dependence on the ENSO state. *Eur. Phys. J. Spec. Top.*, **230**, 3019–3032, https://doi.org/10.1140/epjs/s11734-021-00168-z.

Fan, J., J. Meng, Y. Ashkenazy, S. Havlin, and H. J. Schellnhuber, 2017: Network analysis reveals strongly localized impacts of El Niño. *Proc. Natl. Acad. Sci. USA*, **114**, 7543–7548, https://doi.org/10.1073/pnas.1701214114.

——, ——, ——, ——, and ——, 2018: Climate network percolation reveals the expansion and weakening of the tropical component under global warming. *Proc. Natl. Acad. Sci. USA*, **115**, E12128–E12134, https://doi.org/10.1073/pnas.1811068115.

——, ——, J. Ludescher, Z. Li, E. Surovyatkina, X. Chen, J. Kurths, and H. J. Schellnhuber, 2022: Network-based approach and climate change benefits for forecasting the amount of Indian monsoon rainfall. *J. Climate*, **35**, 1009–1020, https://doi.org/10.1175/JCLI-D-21-0063.1.

Fountalis, I., A. Bracco, B. Dilkina, C. Dovrolis, and S. Keilholz, 2018: δ-MAPS: From spatio-temporal data to a weighted and lagged network between functional domains. *Appl. Network Sci.*, **3**, 21, https://doi.org/10.1007/s41109-018-0078-z.

Friedman, N., M. Goldszmidt, and A. Wyner, 1999: Data analysis with Bayesian networks: A bootstrap approach. *Proc. 15th Conf. on Uncertainty in Artificial Intelligence*, Stockholm, Sweden, Association for Computing Machinery, 196–205, https://dl.acm.org/doi/10.5555/2073796.2073819.

Gao, W., S. Oh, and P. Viswanath, 2018: Demystifying fixed k-nearest neighbor information estimators. *IEEE Trans. Inf. Theory*, **64**, 5629–5661, https://doi.org/10.1109/TIT.2018.2807481.

Gilpin, A. R., 1993: Table for conversion of Kendall's tau to Spearman's rho within the context of measures of magnitude of effect for meta-analysis. *Educ. Psychol. Meas.*, **53**, 87–92, https://doi.org/10.1177/0013164493053001007.

Guez, O. C., A. Gozolchiani, and S. Havlin, 2014: Influence of autocorrelation on the topology of the climate network. *Phys. Rev. E*, **90**, 062814, https://doi.org/10.1103/PhysRevE.90.062814.

Guinness, J., and M. Fuentes, 2016: Isotropic covariance functions on spheres: Some properties and modeling considerations. *J. Multivar. Anal.*, **143**, 143–152, https://doi.org/10.1016/j.jmva.2015.08.018.

Heitzig, J., J. F. Donges, Y. Zou, N. Marwan, and J. Kurths, 2011: Node-weighted measures for complex networks with spatially embedded, sampled, or differently sized nodes. *Eur. Phys. J. B* **85**, 38, https://doi.org/10.1140/epjb/e2011-20678-7.

Hersbach, H., and Coauthors, 2018: ERA5 hourly data on single levels from 1959 to present. Copernicus Climate Change Service Climate Data Store, accessed 13 April 2022, https://doi.org/10.24381/cds.adbb2d47.

——, and Coauthors, 2019a: ERA5 monthly averaged data on pressure levels from 1959 to present. Copernicus Climate Change Service Climate Data Store, accessed 13 April 2022, https://doi.org/10.24381/cds.6860a573.

——, and Coauthors, 2019b: ERA5 monthly averaged data on single levels from 1959 to present. Copernicus Climate Change Service Climate Data Store, accessed 13 April 2022, https://doi.org/10.24381/cds.f17050d7.

Hlinka, J., D. Hartman, N. Jajcay, M. Vejmelka, R. Donner, N. Marwan, J. Kurths, and M. Paluš, 2014: Regional and inter-regional effects in evolving climate networks. *Nonlinear Processes Geophys.*, **21**, 451–462, https://doi.org/10.5194/npg-21-451-2014.

——, ——, ——, D. Tomeček, J. Tintěra, and M. Paluš, 2017: Small-world bias of correlation networks: From brain to climate. *Chaos*, **27**, 035812, https://doi.org/10.1063/1.4977951.

Horvath, S., 2011: *Weighted Network Analysis: Applications in Genomics and Systems Biology*. Springer, 421 pp.

Kittel, T., C. Ciemer, N. Lotfi, T. Peron, F. Rodrigues, J. Kurths, and R. V. Donner, 2021: Evolving climate network perspectives on global surface air temperature effects of ENSO and strong volcanic eruptions. *Eur. Phys. J. Spec. Top.*, **230**, 3075–3100, https://doi.org/10.1140/epjs/s11734-021-00269-9.

Koller, D., and N. Friedman, 2009: *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 1272 pp.

Kraskov, A., H. Stoegbauer, and P. Grassberger, 2004: Estimating mutual information. *Phys. Rev. E*, **69**, 066138, https://doi.org/10.1103/PhysRevE.69.066138.

Kretschmer, M., J. Runge, and D. Coumou, 2017: Early prediction of extreme stratospheric polar vortex states based on causal precursors. *Geophys. Res. Lett.*, **44**, 8592–8600, https://doi.org/10.1002/2017GL074696.

Lahiri, S. N., 2003: *Resampling Methods for Dependent Data*. Springer, 374 pp.

Lam, C., 2020: High-dimensional covariance matrix estimation. *Wiley Interdiscip. Rev.: Comput. Stat.*, **12**, e1485, https://doi.org/10.1002/wics.1485.

Lang, A., and C. Schwab, 2015: Isotropic Gaussian random fields on the sphere: Regularity, fast simulation and stochastic partial differential equations. *Ann. Appl. Probab.*, **25**, 3047–3094, https://doi.org/10.1214/14-AAP1067.

Ledoit, O., and M. Wolf, 2004: A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivar. Anal.*, **88**, 365–411, https://doi.org/10.1016/S0047-259X(03)00096-4.

Levin, K., and E. Levina, 2019: Bootstrapping networks with latent space structure. arXiv, 1907.10821v2, https://doi.org/10.48550/arXiv.1907.10821.

Ludescher, J., and Coauthors, 2021: Network-based forecasting of climate phenomena. *Proc. Natl. Acad. Sci. USA*, **118**, e1922872118, https://doi.org/10.1073/pnas.1922872118.

——, A. Gozolchiani, M. I. Bogachev, A. Bunde, S. Havlin, and H. J. Schellnhuber, 2014: Very early warning of next El Niño. *Proc. Natl. Acad. Sci. USA*, **111**, 2064–2066, https://doi.org/10.1073/pnas.1323058111.

Meinshausen, N., and P. Bühlmann, 2010: Stability selection. *J. Roy. Stat. Soc. B*, **72**, 417–473, https://doi.org/10.1111/j.1467-9868.2010.00740.x.

Minsker, S., and X. Wei, 2017: Estimation of the covariance structure of heavy-tailed distributions. *31st Conf. on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, Association for Computing Machinery, 2860–2870, https://proceedings.neurips.cc/paper/2017/file/10c272d06794d3e5785d5e7c5356e9ff-Paper.pdf.

Morales-Jimenez, D., I. M. Johnstone, M. R. McKay, and J. Yang, 2021: Asymptotics of eigenstructure of sample correlation matrices for high-dimensional spiked models. *Stat. Sin.*, **31**, 571–601, https://doi.org/10.5705/ss.202019.0052.

Mumford, J. A., S. Horvath, M. C. Oldham, P. Langfelder, D. H. Geschwind, and R. A. Poldrack, 2010: Detecting network modules in fMRI time series: A weighted network analysis approach. *Neuroimage*, **52**, 1465–1476, https://doi.org/10.1016/j.neuroimage.2010.05.047.

Niu, X., J. Zhang, L. Zhang, Y. Hou, S. Pu, A. Chu, M. Bai, and Z. Zhang, 2019: Weighted gene co-expression network analysis identifies critical genes in the development of heart failure after acute myocardial infarction. *Front. Genet.*, **10**, 1214, https://doi.org/10.3389/fgene.2019.01214.

Onnela, J.-P., J. Saramäki, J. Kertész, and K. Kaski, 2005: Intensity and coherence of motifs in weighted complex networks. *Phys. Rev. E*, **71**, 065103, https://doi.org/10.1103/PhysRevE.71.065103.

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

Papalexiou, S. M., 2018: Unified theory for stochastic modelling of hydroclimatic processes: Preserving marginal distributions, correlation structures, and intermittency. *Adv. Water Resour.*, **115**, 234–252, https://doi.org/10.1016/j.advwatres.2018.02.013.

Politis, D. N., J. P. Romano, and M. Wolf, 1999: *Subsampling*. Springer Series in Statistics, Springer, 347 pp.

Quian Quiroga, R., T. Kreuz, and P. Grassberger, 2002: Event synchronization: A simple and fast method to measure synchronicity and time delay patterns. *Phys. Rev. E*, **66**, 041904, https://doi.org/10.1103/PhysRevE.66.041904.

Radebach, A., R. V. Donner, J. Runge, J. F. Donges, and J. Kurths, 2013: Disentangling different types of El Niño episodes by evolving climate network analysis. *Phys. Rev. E*, **88**, 052807, https://doi.org/10.1103/PhysRevE.88.052807.

Rheinwalt, A., N. Marwan, J. Kurths, P. Werner, and F.-W. Gerstengarbe, 2012: Boundary effects in network measures of spatially embedded networks. *Europhys. Lett.*, **100**, 28002, https://doi.org/10.1209/0295-5075/100/28002.

——, B. Goswami, N. Boers, J. Heitzig, N. Marwan, R. Krishnan, and J. Kurths, 2015: Teleconnections in climate networks: A network-of-networks approach to investigate the influence of sea surface temperature variability on monsoon systems. *Machine Learning and Data Mining Approaches to Climate Science*, V. Lakshmanan et al., Eds., Springer, 23–33.

Romano, S., N. X. Vinh, K. Verspoor, and J. Bailey, 2018: The randomized information coefficient: Assessing dependencies in noisy data. *Mach. Learn.*, **107**, 509–549, https://doi.org/10.1007/s10994-017-5664-2.

Runge, J., P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic, 2019a: Detecting and quantifying causal associations in large nonlinear time series datasets. *Sci. Adv.*, **5**, eaau4996, https://doi.org/10.1126/sciadv.aau4996.

——, and Coauthors, 2019b: Inferring causation from time series in Earth system sciences. *Nat. Commun.*, **10**, 2553, https://doi.org/10.1038/s41467-019-10105-3.

Scarsoglio, S., F. Laio, and L. Ridolfi, 2013: Climate dynamics: A network-based approach for the analysis of global precipitation. *PLOS ONE*, **8**, e71129, https://doi.org/10.1371/journal.pone.0071129.

Schreiber, T., and A. Schmitz, 1996: Improved surrogate data for nonlinearity tests. *Phys. Rev. Lett.*, **77**, 635–638, https://doi.org/10.1103/PhysRevLett.77.635.

Shao, J., and D. Tu, 1995: *The Jackknife and Bootstrap*. Springer Series in Statistics, Springer, 517 pp.

Sporns, O., 2010: *Networks of the Brain*. The MIT Press, 424 pp.

Stein, M. L., 1999: *Interpolation of Spatial Data*. Springer Series in Statistics, Springer, 249 pp.

Tsonis, A. A., and P. J. Roebber, 2004: The architecture of the climate network. *Physica A*, **333**, 497–504, https://doi.org/10.1016/j.physa.2003.10.045.

——, K. L. Swanson, and G. Wang, 2008: On the role of atmospheric teleconnections in climate. *J. Climate*, **21**, 2990–3001, https://doi.org/10.1175/2007JCLI1907.1.

Vallis, G. K., 2011: *Climate and the Oceans*. Princeton University Press, 248 pp.

von Luxburg, U., 2007: A tutorial on spectral clustering. *Stat. Comput.*, **17**, 395–416, https://doi.org/10.1007/s11222-007-9033-z.

Wiedermann, M., J. F. Donges, J. Kurths, and R. V. Donner, 2015: Spatial network surrogates for disentangling complex system structure from spatial embedding of nodes. *Phys. Rev. E*, **93**, 042308, https://doi.org/10.1103/PhysRevE.93.042308.

Yamasaki, K., A. Gozolchiani, and S. Havlin, 2008: Climate networks around the globe are significantly affected by El Niño. *Phys. Rev. Lett.*, **100**, 228501, https://doi.org/10.1103/PhysRevLett.100.228501.

SUPPLEMENTAL MATERIAL FOR 'PITFALLS OF CLIMATE NETWORK CONSTRUCTION'

A STATISTICAL PERSPECTIVE

Moritz Haas
mo.haas@uni-tuebingen.de
University of Tübingen, Germany

Bedartha Goswami
bedartha.goswami@uni-tuebingen.de
University of Tübingen, Germany

Ulrike von Luxburg
ulrike.luxburg@uni-tuebingen.de
University of Tübingen, Germany

Contents

A Introduction to Climate Network Terminology and Methodology

Here we introduce all concepts and methods around climate networks that appear in the main text. We offer both definitions as well as brief explanations for the application in climate science.

A.1 Networks

An undirected **network** (V, E) consists of a set of **nodes/vertices** V with p elements and a set of undirected **links/edges** $\{v_i, v_j\}$ connecting pairs of nodes $v_i, v_j \in V$. We exclude self-loops $\{v_i, v_i\}$. While for undirected edges $\{v_i, v_j\}$ the ordering of the nodes does not matter, a directed edge (v_i, v_j) starts at v_i and ends at v_j, but we only consider undirected ones. In a **weighted network** each edge $\{v_i, v_j\}$ has a corresponding weight $w_{i,j}$; in an **unweighted network** all edges simply have weight $w_{i,j} = 1$. When there is no edge between v_i and v_j, we write $w_{i,j} = 0$. In this way, a network can be respresented by its **adjacency matrix** $W = (w_{i,j})_{i,j=1,\dots,p}$.

While climate networks are commonly constructed by including the edges with the largest similarity values globally, there is an alternative, which is particularly popular in machine learning (von Luxburg, 2007).

(Symmetric) k-nearest neighbor graphs (short kNN graphs) include the k edges with largest similarity for each node. This ensures that each node is connected to at least k other nodes. This is useful when some locations have much lower characteristic similarities than others, but these locations should not be ignored by the network.

An alternative network construction procedure that is popular in climate science is to include edges $\{i, j\}$ which are significant in some sense. For this purpose the time series are reshuffled several times either randomly, or preserving certain time series characteristics. This yields a baseline distribution of similarity on each edge. Now one can include an edge, when the similarity $\hat{S}_{ij}$ of the original time series exceeds a restrictive quantile of the baseline distribution like 0.99 or even 0.999. Another popular approach is to calculate the mean $\hat{\mu}^0_{ij}$ and standard deviation $\hat{\sigma}^0_{ij}$ of the baseline distribution, calculate z-scores $\frac{\hat{S}_{ij} - \hat{\mu}^0_{ij}}{\hat{\sigma}^0_{ij}}$ and construct a density-threshold network from these z-scores.

Besides eyeballing the network, there are several quantities that measure different properties of the networks and therefore potentially help to understand the network structure and discover interesting behaviour. Let us introduce the most important ones for undirected similarity networks.

The **degree** d_i of a node v_i is the cumulative weight of its connections $d_i = \sum_{i=1}^{p} w_{i,j}$. It measures the importance of the node in the sense of how well connected it is in the network. A node of exceptionally large degree is called hub and often influences many other nodes.

The **(local) clustering coefficient** C_i of a node v_i, is given by $C_i = \frac{1}{d_i(d_i-1)} \sum_{j,k=1}^{p} w_{i,j} w_{j,k} w_{k,i}$. It measures how many neighbors of node v_i are directly connected to each other by an edge. It is 0 when no neighbor of v_i has an edge to any other neighbor of v_i, and it is 1 when all neighbors of v_i are interconnected as well. A set of nodes that is highly intraconnected but less connected to other nodes is often called a 'cluster' or community of nodes. A popular approach to calculate a **global clustering coefficient** is to simply average all local clustering coefficients (Watts and Strogatz, 1998). This quantity is dominated by nodes with low degree. An alternative measure for undirecte networks which represents the global network structure more reasonably (Newman, 2018) is given by $C = \frac{\sum_{i,j,k} W_{ij} W_{jk} W_{ki}}{\sum_i d_i(d_i-1)}$.

The **weighted clustering coefficient** (Onnela et al., 2005) reflects how large triangle weights are compared to the network maximum. While there are other generalizations, this is the one implemented in the popular brain connectivity toolbox (Rubinov and Sporns, 2010). In recent work, Wang et al. (2017), Fardet and Levina (2021) point out that other definitions are more desirable for networks with highly heterogeneous weight distribution or numerous spurious edges with low weights.

The **shortest path length** from v_i to v_j quantifies the minimal number of traversed edges to move from v_i to v_j. The shortest path length at v_i is the average shortest path length from v_i to any other node. Transport networks are efficient when they have small shortest path lengths. Neuroscientists often point out small-world behaviour in functional brain networks as a justification for the brain's efficiency (Bullmore and Sporns, 2012), robustness (Levit-Binnun et al., 2013, Neal, 2017) and compositionality. Small-worldness is characterized by large clustering coefficients and low shortest path lengths. Treatments of small-worldness should be handled with care (Bialonski et al., 2010, Hlinka et al., 2017).

The **betweenness centrality** of a node v_i measures the centrality of v_i in terms of how many shortest paths in the network contain v_i. Let $\sigma_{j,k}$ denote the number of shortest paths from v_j to v_k in the network, and $\sigma_{j,k}(v_i)$ the number of these paths that also contain v_i. The betweenness centrality is then given by $\sum_{j,k \neq i} \frac{\sigma_{j,k}(v_i)}{\sigma_{j,k}}$. Similar to

shortest path length and small worldness, we find that betweenness can be heavily distorted both locally as well as globally by single false edges.

Ricci curvature of an edge describes how the connectivity of its network neighborhood differs from the connectivity of a regular grid. Positive curvature indicates a highly intraconnected neighborhood or a within cluster edge, and negative curvature indicates that an edge connects different communities (Strnad et al., 2022). Forman-Ricci curvature (Forman, 2003) and Ollivier-Ricci curvature (Ollivier, 2010) are two numerical approaches to approximate the Ricci curvature of an edge, where Forman-Ricci curvature is computationally cheaper while Ollivier-Ricci curvature is more accurate.

In climate networks, the nodes usually represent locations. Hence our network is naturally embedded in 3D space and edges have a length given by the geodesic distance between the two endpoints. In traffic networks there is a cost associated with forming edges, so that the total link length should be minimized in the network design. In similarity networks, closeby points are generally similar. This local connectivity is the baseline behaviour when analysing spatio-temporal data. Of particular interest are connected nodes that are far apart. Those indicate climatic teleconnections. An important summary statistic of the spatial properties of a network is its **link length distribution**, which subsumes the lengths of all edges in the network. Given the roughly spherical geometry of the Earth, in an infinitesimally fine grid, the number of potential links scales as $\sin(d)$ with link length d in radians. When dividing the link length distribution by $\sin(d)$ (or, for a finite grid, by the number of potential links at each distance), we see how many of the potential links are formed at a given distance. For spatiotemporal data, we expect values close to 1 for small lengths and values close to 0 for large lengths. Random link distribution results in a roughly uniform distribution across lengths, hence a sinusoidal distribution of links in the original link length distribution. Crucially, the link length distribution does not convey any information about link bundling behaviour (cf. Fig. 16). An extensive discussion of link bundles can be found in Section 3c and supplemental Section A.3.

We call a grid **isotropic** when the point density is constant over the sphere, or alternatively when the distance distribution from one grid point to all others does not depend on the considered grid point. While a perfectly isotropic grid does not exist for the sphere, there exist approximately isotropic grids, like the one we generate with the Fekete algorithm (Bendito et al., 2007) or the Fibonacci grid. In a **Gaussian grid**, the grid points are equally spaced across latitudes. In a regular Gaussian grid the number of gridpoints is constant across latitudes. While they traditionally are the most popular grid type, regular Gaussian grids have diverging densities around the poles and can therefore induce complex distortions when used in climate networks.

A.2 Similarity Measures and Estimators

Most commonly climate and neuroscientists calculate the cross-correlation between time series $\{X_t\}_{t=1,\dots,n}$ and $\{Y_t\}_{t=1,\dots,n}$ of length n, defined as $\frac{1}{n}\sum_{t=1}^{n} X_t Y_t$. Statistically, this is imprecise terminology: While they *estimate* the Pearson (cross-)correlation between the time series with the ***empirical*** **Pearson correlation estimator** $\frac{1}{n}\sum_{t=1}^{n} X_t Y_t$, there are other estimators that might result in estimates that come closer to the ground truth values we aim to approximate. For high-dimensional problems, the empirical correlation matrix is a very bad estimator: the eigenvalue spectrum of empirical correlation matrices is systematically distorted, following the Marcenko-Pastur law (Marčenko and Pastur, 1967). A long history of correlation estimation research (Fan et al., 2015, Pourahmadi, 2013) has proposed various approaches such as banding, tapering, Lasso-regularized maximum likelihood and eigenvalue shrinkage to correct these empirical distortions. One particularly popular estimator is the Ledoit-Wolf estimator (Ledoit and Wolf, 2004), which has a recent non-linear advancement (Ledoit and Wolf, 2020) claimed to accommodate dimensions up to $p = 10^4$. In Figure 3c we find that the linear Ledoit-Wolf estimator approximates the correlation matrix much better than empirical Pearson correlation. The **linear Ledoit-Wolf estimator** is defined as $\rho_1 \hat{S}_n + \rho_2 I_d$, where $\hat{S}_n$ is the empirical covariance matrix, I_d the identity matrix and the optimal choice of $\rho_1, \rho_2 \in \mathbb{R}$ can be found in Ledoit and Wolf (2004). Estimators from spatial statistics use the spatial information of the nodes and are therefore more appropriate for random field data (Li, 2009, Clifford et al., 1989). In most geostatistical models only one measurement can be made at each location; exceptions include maximum likelihood estimation of space-time variograms (Bevilacqua et al., 2012) and nearest neighbor Gaussian process models (Datta et al., 2016). Designing anisotropic spatial correlation estimators for finely gridded time series in spherical geometry remains an important direction of future work (Chen et al., 2021, Raymaekers and Rousseeuw, 2021) for climate network and EOF construction. Recently, estimation of partial correlations in Bayesian or causal network approaches (Zerenner et al., 2014, Runge et al., 2019, Gerhardus and Runge, 2020) has gained popularity, with the goal of distinguishing direct from indirect influences. The precision matrix can be estimated to recover conditional independences. The same estimation difficulties arise in this context in high dimension and suitable estimators should be carefully chosen (Lam, 2020).

Spearman's rank correlation coefficient, short **Spearman correlation**, is another similarity measure, which measures not only linear dependence structures, but all monotonous dependences. It is given by the Pearson correlation between

the time series of ranks $\{R(X_t)\}_{t=1,\dots,n}$ of $\{X_t\}_{t=1,\dots,n}$ and $\{R(Y_t)\}_{t=1,\dots,n}$ of $\{X_t\}_{t=1,\dots,n}$. It is invariant under monotonic transformations of the data. A measure of rank correlation with better finite-sample properties is given by Kendall's τ (Gilpin, 1993).

Mutual information between two random variables X and Y is given by

$$MI(X,Y) = KL(P^{X,Y}, P^X \times P^Y),$$

where KL denotes the Kullback-Leibler divergence, $P^{X,Y}$ denotes the joint distribution of X and Y, and $P^X \times P^Y$ the joint distribution of X and Y if they were independent. It measures to which extent the two random variables are dependent. Instead of only measuring linear or monotonic relationships, mutual information detects any kind of dependency. Traditionally, binning estimators of mutual information have been popular, but these are highly sensible to the number of bins and often heavily biased (Paninski, 2003). The KSG estimator (Kraskov et al., 2003) (based on a kNN approach) has been observed to be robust to the choice of its hyperparameter k (Papana and Kugiumtzis, 2009). Extensions of the KSG estimator include a bias-corrected version (Gao et al., 2018) and an adaptation to discrete-continuous mixtures (Gao et al., 2017) such as precipitation. Without distributional assumptions there exists no convergence rate that can be guaranteed in MI estimation (Antos and Kontoyiannis, 2001), so we still have to expect unprecise estimates. Especially for highly autocorrelated time series it is known that a lot of false positives (Runge et al., 2012, 2014) will arise. While evaluating the robustness of empirical MI networks, it is also worth considering other similarity measures that capture arbitrary dependences. Romano et al. (2018) compares various information theory- and correlation-based measures. The alternative we employed is the **Hilbert Schmidt Independence Criterion** (HSIC) for random processes (Chwialkowski and Gretton, 2014). It measures dependence after a mean embedding of the random variables in some reproducing kernel Hilbert space. For more details see also Gretton et al. (2012).

A.3 Link Bundle Definition

With the term 'link bundle' we want to capture the case when most links are formed between two regions. While there is a lot of work on the visualisation of graphs with focus on edge bundles (Lambert et al., 2010, Hurter et al., 2012, Selassie et al., 2011, Nocke et al., 2015), our purpose does not demand a 2D-visualisation. We are simply interested in the cumulative edge weight between two regions of interest. As we have not found a definition of link bundles in the context of spatially embedded graphs in the literature, we define our notion of link bundle over spatial graphs (V, E) with adjacency matrix A, where (V, d) is a metric space, as follows.

First define the ε-ball around v as $B_\varepsilon(v) := \{u \in V | d(u,v) \leq \varepsilon\}$. Let $\rho_\varepsilon(v_i, v_j)$ be the number of edges between $B_\varepsilon(v_i)$ and $B_\varepsilon(v_j)$ in the complete graph. It holds that $\rho_\varepsilon(v_i, v_i) = \binom{|B_\varepsilon(v_i)|}{2}$. When $B_\varepsilon(v_i) \cap B_\varepsilon(v_j) = \emptyset$, then $\rho_\varepsilon(v_i, v_j) = |B_\varepsilon(v_i)| \cdot |B_\varepsilon(v_j)|$. We denote the cumulative weights between neighborhoods of v_i and v_j by

$$W_\varepsilon(v_i, v_j) := \sum_{k,l:\, v_k \in B_\varepsilon(v_i), v_l \in B_\varepsilon(v_j)} |A_{kl}|,$$

where A denotes the adjacency matrix. Our notions of link bundles measure whether the mean weights between regions (per possible link) exceed a prespecified threshold. Whether the graph is weighted or unweighted makes no difference in the definition. Of course weights allow for a more fine-grained differentiation of link bundle strength (Fig. 8). We allow for a tolerance ($1 - c$ for unweighted graphs) of unformed edge weight as link bundles of a certain density might already be considered significant. Demanding complete link bundles might be unrealistic and sensitive to noise in practical settings.

A first idea might be a one-to-many notion, where we demand that a node v_i is connected to all nodes around a point v_j within a given radius ε. We say that there is a (ε, c)**-one-to-many link bundle** from v_i to v_j in (V, E, A), if

$$\frac{1}{|B_\varepsilon(v_i)|} \sum_{k:\, v_k \in B_\varepsilon(v_i)} |A_{kj}| \geq c.$$

We observe such one-to-many link bundles in networks dominated by extreme events and more generally when one region is locally correlated and the other point spuriously aligns with that region while differing from its surrounding. More natural might be to consider neighborhoods around both endpoints of the link bundle. We define a (ε, c)**-many-to-many link bundle** between v_i and v_j in (V, E, A), if

$$\frac{W_\varepsilon(v_i, v_j)}{\rho_\varepsilon(v_i, v_j)} \geq c.$$

In applications with anisotropic link density across space, we might want to consider teleconnections to be significant if the link density between neighborhoods is large compared to the density within the neighborhoods. There is a

(ε, c)**-locally-weighted many-to-many link bundle** between v_i and v_j in (V, E, A), if

$$\frac{W_\varepsilon(v_i, v_j)}{\rho_\varepsilon(v_i, v_j)} \geq \frac{c}{2}\left(\frac{W_\varepsilon(v_i, v_i)}{\rho_\varepsilon(v_i, v_i)} + \frac{W_\varepsilon(v_j, v_j)}{\rho_\varepsilon(v_j, v_j)}\right).$$

The notion of locally weighted link bundles is also sensible to significant teleconnections between endpoints of low link density in settings of anisotropic density. Other utilisation of intra-regional link densities is conceivable. A locally weighted version of 1-to-many link bundles can be defined analogously. In order to find link bundles in practice, we propose the distance metric between two edges (x_1, y_1) and (x_2, y_2), where $x_1, y < 1, x_2, y_2$ are nodes,

$$d_E\left((x_1, y_1), (x_2, y_2)\right) = \min\left\{d(x_1, x_2) + d(y_1, y_2), \quad d(x_1, y_2) + d(y_1, x_2)\right\} \tag{1}$$

in the undirected case. For the directed case, only take $d_E\left((x_1, y_1), (x_2, y_2)\right) = d(x_1, x_2) + d(y_1, y_2)$. In this distance two edges are close in feature space if and only if both ends of the edges are close together. Clusters found with for example hierarchical clustering over the distance matrix between edges would then constitute link bundles.

Using the weighted notions of link bundle, c has to be chosen smaller to detect link bundles, when the similarity values are typically smaller than 1. The locally weighted notion of link bundle requires less adaptation of c since it captures the edge weights characteristic to the measured random field. Analysing the edge weights in our setting indicates that the spuriously included links only lie marginally above the threshold and can therefore be distinguished from strong links. However, a ground-truth teleconnection of intermediate strength might still be difficult to distinguish from spurious bundles. While the notions of weighted link bundles offer a more fine-grained differentiation, questions of significance have to be answered differently. Although some spurious locally weighted many-to-many link bundles occur in the weighted networks, they are a lot less frequent under suitably chosen c. For large densities, the notions of locally weighted and absolute many-to-many link bundles become equivalent in unweighted networks, as the local neighborhoods become completely connected. The notion of locally weighted link bundle becomes problematic in sparse graphs, when the neighborhood connectivity structure is not sufficiently established.

B Details of the Stochastic Ground Truth Model for Spatio-Temporal Data

B.1 Time Dependence via Vector Autoregression

As the typically used data sets only resolve a fixed time resolution, we do not have to employ the computationally expensive space-time covariance functions, but opt for a vector autoregression VAR(1). As introduced in Algorithm 1, we denote the available measurements $(X_{1,t}, \ldots, X_{p,t}) \sim N(0, \Sigma)$ of the MIGRF at time t evaluated on the finite grid $\{v_i\}_{i\in[p]} \subset S^2$ with $\Sigma_{ij} = k(|v_i - v_j|)$. We generate $X_{i,t}$, for $i \in [p]$, via

$$X_{i,t} = a_i \cdot X_{i,t-1} + \varepsilon_{i,t},$$

where the innovations $(\varepsilon_{1,t}, \ldots, \varepsilon_{p,t}) \sim N(0, \Sigma^\varepsilon)$ are i.i.d. in time and a_i with $|a_i| < 1$ denotes the lag-1 autocorrelation at node v_i. What is left to do is to design an innovation covariance matrix Σ^ε that generates the desired covariance structure Σ of the MIGRF. A well-defined Σ^ε does not exist for all choices of Σ and a. Whenever it exists, it is given by $\Sigma^\varepsilon_{ij} = \Sigma_{ij}(1 - a_i \cdot a_j)$. This formula can be derived by defining $A \in \mathbb{R}^{p\times p}$ as the diagonal matrix with $A_{ii} = a_i$ and using

$$\begin{pmatrix} X_{1,t} \\ X_{2,t} \\ \vdots \\ X_{p,t} \end{pmatrix} = \sum_{k=0}^{\infty} A^k \begin{pmatrix} \varepsilon_{1,t-k} \\ \varepsilon_{2,t-k} \\ \vdots \\ \varepsilon_{p,t-k} \end{pmatrix},$$

which yields,

$$\begin{pmatrix} X_{1,t} \\ X_{2,t} \\ \vdots \\ X_{p,t} \end{pmatrix} \sim N\left(0, \sum_{k=0}^{\infty} A^k \Sigma^\varepsilon A^k\right) = N\left(0, \left(\frac{\Sigma^\varepsilon_{ij}}{1 - A_{ii} \cdot A_{jj}}\right)_{ij}\right) \overset{!}{=} N(0, \Sigma).$$

Solving for Σ^ε then leads to the desired formula.

It is easy to see that when all nodes v_i have the same autocorrelation, Σ^ε is a well-defined covariance matrix. In other cases, the above defined matrix Σ^ε might have negative eigenvalues, making it unsuitable as a covariance matrix. As a remedy, the negative eigenvalues of Σ^ε could be shifted to a small positive constant to make it positive definite. This

procedure induces a shift in the spatial covariance Σ, that would need to be quantified. The specified autocorrelation structure remains exact by design. In our experiments in Section 3d, we initialize a random half of the points with $A_{ii} = 0.2$ and the other half with $A_{ii} = 0.7$. The resulting Σ^ε is not positive definite, but we find a valid correlation matrix by shifting all negative eigenvalues of Σ^ε to 10^{-8}. This shift does not introduce a notable bias into the spatial ground truth correlations.

B.2 Matérn Covariance

The Matérn covariance function appears in many fields (Whittle, 1954, Guttorp and Gneiting, 2006). A random process on $\mathbb{R}^d$ with Matérn covariance function is a solution to the stochastic partial differential equation (Lindgren et al., 2011)

$$(\kappa^2 - \Delta)^{\alpha/2} G(x) = W(x),$$

where $W(x)$ is Gaussian white noise with unit variance, $\Delta = \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2}$ is the Laplace operator, $\kappa = \frac{\sqrt{2\nu}}{\ell}$, and $\alpha = \nu + d/2$. Therefore it regularly demonstrates its efficacy of modelling physical processes in spatial statistics (Bevilacqua et al., 2020, Stein, 2011), and can be adjusted to model non-stationary, non-isotropic, oscillating and non-separable random fields (Lindgren et al., 2011).

On $\mathbb{R}^D$, it is given by

$$C_{\nu,\ell}(d) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{d}{\ell} \right)^{\nu} K_\nu \left(\sqrt{2\nu} \frac{d}{\ell} \right),$$

where Γ is the gamma function and K_ν is the modified Bessel function of second kind. Its smoothness parameter ν and scale parameter ℓ make it flexible as well as interpretable. The scale parameter ℓ determines how far the region of high correlation extends spatially. $\nu > 0$ determines the smoothness of the random field; so how well the random field behaves locally. Large ℓ and large ν result in less fluctuation across space. A Gaussian process with Matérn covariance is $\lfloor \nu \rfloor - 1$ times differentiable in the mean-square sense. When $\nu = n + 1/2$ for some $n \in \mathbb{N}$, $C_{\nu,\ell}$ can be written as a product of an exponential and a polynomial of order n. For $\nu \to \infty$, it approximates the squared exponential function. So by varying ν, we can interpolate between the absolute exponential kernel and the Gaussian radial basis function (Stein, 1999, ch. 2.10).

Unfortunately, the Matérn is positive definite with great circle distance only if $\nu \leq 1/2$. Guinness and Fuentes (2016) studies the use of several Matérn-like covariance functions on the sphere and finds that they all fit smooth and non-smooth meteorological datasets more flexibly than alternative classes of covariance functions which respect to spherical geometry. We follow their recommendation in using the computationally efficient *chordal Matérn*. The chordal Matérn is simply the restriction of the Matérn in $\mathbb{R}^3$ to S^2. An adaptation to great circle distance, called *circular Matérn*, shows no practical gain. The authors find maximum likelihood estimates for ν close to 0.42 for ozone data and close to 1.46 for 10m height surface temperature outputs from a single run of the Community Climate System Model Version 4 (CCSM4). We employ realistic values of $\nu = 0.5$ and $\nu = 1.5$ in our simulations, being mostly interested in the effect of varying ν and ℓ.

In practice, measurements from a random field are only available on a fixed, finite grid. Instead of sampling the entire random field via the Karhunen Loeve expansion (Lang and Schwab, 2015), we can simply consider the multivariate Gaussian on the grid points, at each time step.

B.3 Choice of Smoothness ν and Length Scale ℓ

We choose ν and ℓ to reflect realistic values for climatic time series as well as to point out their influences on the estimation procedure. Our choices of $\nu \in \{0.5, 1.5\}$ and $\ell \in \{0.1, 0.2\}$ (in radians) reflect a realistic range for climatic variables (Guinness and Fuentes, 2016). Stronger smoothness only induces subtle differences. Table 1 shows network-based smoothness and length scale properties of the MIGRF as well as climatic data.

Observe that our parameter choices cover a realistic range. Given similar smoothness and length scale, the network density on simulated data is smaller as systematic teleconnections are missing.

Random field property	Simulated data				Real data				
	$\nu = 0.5$, $\ell = 0.1$	$\nu = 0.5$, $\ell = 0.2$	$\nu = 1.5$, $\ell = 0.1$	$\nu = 1.5$, $\ell = 0.2$	pr	dt2m	t2m	sp	z500
Avg. local corr.	0.544	0.734	0.704	0.892	0.520	0.753	0.826	0.946	0.953
Decorr. len., $\tau = 0.2$	0.172	0.348	0.190	0.383	0.139	0.271	0.329	0.571	0.453
Decorr. len., $\tau = 0.5$	0.076	0.152	0.104	0.216	0.059	0.141	0.183	0.346	0.305
Density, $\tau = 0.2$	0.031	0.057	0.031	0.057	0.009	0.031	0.093	0.150	0.246
Density, $\tau = 0.5$	0.001	0.005	0.002	0.010	0.002	0.006	0.013	0.047	0.117

Table 1: Distributional smoothness and length scale properties of simulated and real climatic random field data. For the *average local correlation* we compute the average correlation inside ε-balls around all vertices. We choose the radius $\varepsilon = 5^\circ$ as in Fig. 8. For the *network density* we construct unweighted Pearson correlation networks with threshold τ. Our notion of *decorrelation length* measures the length scale as a network feature. For each node, it is given by the minimal radius (in radians) at which the connectivity of the network inside an ε-ball drops below $c = 0.8$ (smaller than 1 for robustness), finally average over all nodes. The table shows average values over 30 independent realizations. For real data we construct a single network from all observations.

C Why Spurious Links Occur in Bundles

The following proposition formalizes why one spuriously large similarity estimate incurs correlated similarity estimates to be spuriously large as well, with non-negligible probability. We assume Gaussianity for simplicity. τ represents the threshold above which a link is included in the network.

Proposition C.1 (Link probability given the presence of another link). *Let $\hat{S}_1, \hat{S}_2 \sim N(\mu, \Sigma)$ be estimates of similarity values on two edges with variances $\sigma_1^2 = \Sigma_{11}$, $\sigma_2^2 = \Sigma_{22}$ and correlation $\rho = \frac{\Sigma_{12}}{\sigma_1 \sigma_2}$. Then, with $\varepsilon \in \mathbb{R}$,*

$$P\Big(\hat{S}_1 > \tau \;\Big|\; \hat{S}_2 > \tau + \varepsilon\Big) \geq \Phi\left(\frac{\frac{\sigma_1}{\sigma_2}\rho(\tau + \varepsilon - \mu_2) + \mu_1 - \tau}{\sqrt{1-\rho^2}\sigma_1}\right),$$

where Φ is the cumulative distribution function of the standard normal distribution.

The resulting probability becomes arbitrarily large when $\hat{S}_2$ becomes large (by increasing ε) and the correlation ρ between the edge estimates is large. This explains both the occurence of link bundles as well as spuriously dense regions in the empirical networks.

Proof. For any $s \in \mathbb{R}$, it holds that $\hat{S}_1 | \hat{S}_2 = s \sim N\Big(\tilde{\mu}(s),\ \tilde{\sigma}^2\Big)$ with $\tilde{\mu}(s) := \mu_1 + \frac{\sigma_1}{\sigma_2}\rho(s - \mu_2)$ and $\tilde{\sigma}^2 := (1-\rho^2)\sigma_1^2$. Then, since $\left(\frac{\hat{S}_1 - \tilde{\mu}(\tau+\varepsilon)}{\tilde{\sigma}^2} \;\middle|\; \hat{S}_2 > \tau + \varepsilon\right) \sim N(0,1)$ and with $\Phi(x) = 1 - \Phi(-x)$ for all $x \in \mathbb{R}$, we get,

$$\begin{aligned}
P\Big(\hat{S}_1 > \tau \;\Big|\; \hat{S}_2 > \tau + \varepsilon\Big) &\geq \inf_{s \geq \tau + \varepsilon} P\Big(\hat{S}_1 > \tau \;\Big|\; \hat{S}_2 > s\Big) = P\Big(\hat{S}_1 > \tau \;\Big|\; \hat{S}_2 > \tau + \varepsilon\Big) \\
&= P\Big(\frac{\hat{S}_1 - \tilde{\mu}(\tau+\varepsilon)}{\tilde{\sigma}^2} > \frac{\tau - \tilde{\mu}(\tau+\varepsilon)}{\tilde{\sigma}^2} \;\Big|\; \hat{S}_2 > \tau + \varepsilon\Big) \\
&= 1 - \Phi\Big(\frac{\tau - \tilde{\mu}(\tau+\varepsilon)}{\tilde{\sigma}^2}\Big) = \Phi\Big(\frac{\tilde{\mu}(\tau+\varepsilon) - \tau}{\tilde{\sigma}^2}\Big) \\
&= \Phi\left(\frac{\frac{\sigma_1}{\sigma_2}\rho(\tau + \varepsilon - \mu_2) + \mu_1 - \tau}{\sqrt{1-\rho^2}\sigma_1}\right).
\end{aligned}$$

□

For a more involved theoretical analysis, a realistic assumption might be that the correlation structure varies across the metric space and there can be ground truth teleconnections of intermediate strength, but data over nearby points universally shows the highest correlations. In such generality, relevant for the correctness of the estimation procedure is still only the joint distribution of edge weight estimates $\{\hat{S}_e | e \in E\}$. Future work could analyse this joint distribution more closely with respect to the resulting graph features, particularly spurious link bundles and expected false discovery rate.

D The importance of the number of effective samples

The number of time steps n has a dominating effect on errors in the network. With large enough time length $n \to \infty$, eventually all random patterns disappear, as the similarity estimates on each edge become more precise. However, the crucial quantity that determines the estimation variance is not n itself, but the *effective* time length n_{eff}, which also depends on the autocorrelation structure. Let us introduce this quantity formally and quantify the relation between n and n_{eff}.

Given two time series $\{X_t\}_{t=1,\dots,n}$ and $\{Y_t\}_{t=1,\dots,n}$ of length n, their empirical Pearson correlation (often called cross-correlation) is given by

$$\hat{C}_{xy} = \frac{1}{n}\sum_{t=1}^{n} X_t Y_t.$$

Now $\hat{C}_{xy}$ does not perfectly estimate the true underlying Pearson correlation between the time series, but has some positive variance. Hence, we define the effective time length n_{eff} as the minimal number of independent observations $\{X'_i\}$ and $\{Y'_i\}$ such that the empirical cross correlation estimator $\frac{1}{n_{\text{eff}}}\sum_{i=1}^{n_{\text{eff}}} X'_i Y'_i$ has at most the same variance as it has under the autocorrelation structure of the given time series of length n.

Below, under some assumptions, we derive the relation

$$n = \sigma_\infty^2 \cdot n_{\text{eff}}, \tag{2}$$

where σ_∞^2 denotes the asymptotic variance of empirical cross correlation in the central limit theorem. Now observe that σ_∞^2 explodes when the autocorrelation of both time series is large: For two independent AR(1)-processes with variance 1 and lag-1 autocorrelation $\alpha \in (0,1)$ and $\beta \in (0,1)$, respectively, we write $\sigma^2_{\alpha,\beta} := \sigma^2_\infty$ for which it holds that

$$\sigma^2_{\alpha,\beta} = \sum_{k=-\infty}^{\infty} Cov(X_1, X_{1+k})Cov(Y_1, Y_{1+k}) = 1 + 2\sum_{k=1}^{\infty}\alpha^k\beta^k = 1 + 2\frac{\alpha\beta}{1-\alpha\beta}.$$

Note that for $\alpha \to 1$ and $\beta \to 1$, we get $\sigma^2_{\alpha,\beta} \to \infty$. Intuitively, the given time series are essentially a single observation, when the autocorrelation approaches 1; and correlation can not be estimated with a single observation. Given that the lag-1 autocorrelation patterns of real climatic variables range from negative values to values close to 1 (Fig. 12), the effective length of climatic time series highly depends on the location. Concretely, two AR(1)-time series with lag-1 autocorrelation $\alpha = \beta = 0.9$ induce an increased estimation variance of $\sigma^2_{\alpha,\beta} = 9.53$, while $\alpha = \beta = 0.95$ already results in $\sigma^2_{\alpha,\beta} = 19.51$. Conversely, given 40 years of monthly observations $n = 480$, with the above autocorrelation patterns only yields $n_{\text{eff}} = 50.39$ and $n_{\text{eff}} = 24.60$ effective samples, respectively.

Derivation of equation (2). The central limit theorem for the empirical cross-correlation between two independent, normalized stationary time series $\{X_t\}$ and $\{Y_t\}$ (cf. (Brockwell and Davis, 1991, p. 236ff)) states that, under structural assumptions, $\sqrt{n}\left(\frac{1}{n}\sum_{t=1}^{n} X_t Y_t\right)$ is asymptotically normal with mean 0 and variance

$$\sigma_\infty = \sum_{k=-\infty}^{\infty} Cov(X_1, X_{1+k})Cov(Y_1, Y_{1+k}).$$

Because the variance of our estimates decays as $1/n$ in the central limit theorem, we need time series of length $n = \sigma_\infty \cdot n_{\text{eff}}$ to reach the same estimation variance as n_{eff} i.i.d. samples for which $\sigma_\infty = 1$.

E Our Results Retain Their Validity Under Larger Time Length n

In this section we analyse how our results evolve with increasing effective time length n.

While all estimated quantities better approximate the ground truth, the amount of errors in the networks are still troublesome and our conceptual findings still hold. In particular, the bias stemming from anisotropic autocorrelation only slightly improves with increasing n (Fig. E.6), because this bias stems from the *relative* difference in estimation variance depending on the autocorrelation strength. Also, few false shortcuts suffice to distort quantities like betweenness and shortest path lengths *globally* (Fig. E.2 and Fig. E.4). Weighted networks benefit much quicker from larger n in terms of degree estimation (Fig. E.4) and spurious link bundles (Fig. E.5).

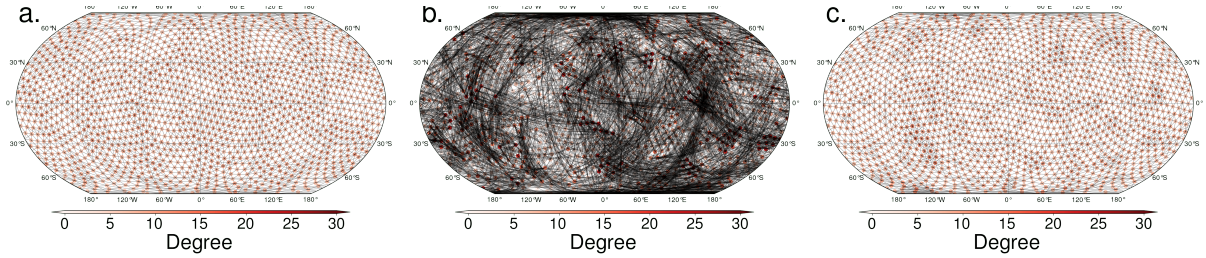

Figure E.1: Same as Fig. 2 but with $n = 1000$. The estimated networks look slightly better. But the same conceptual distortions are still clearly present: Empirical Pearson correlation fails as an estimator of heavy-tailed data. High degree clusters emerge as a result of localized correlation structure.

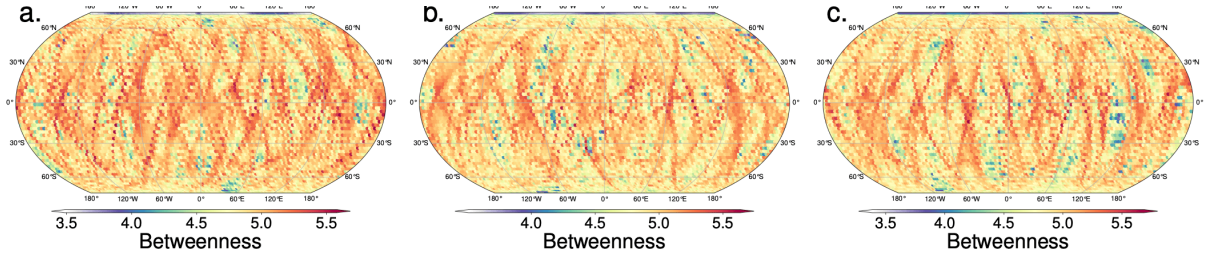

Figure E.2: Same as Fig. 5 but with $n = 500$. Few false links are enough to create spurious global betweenness structures.

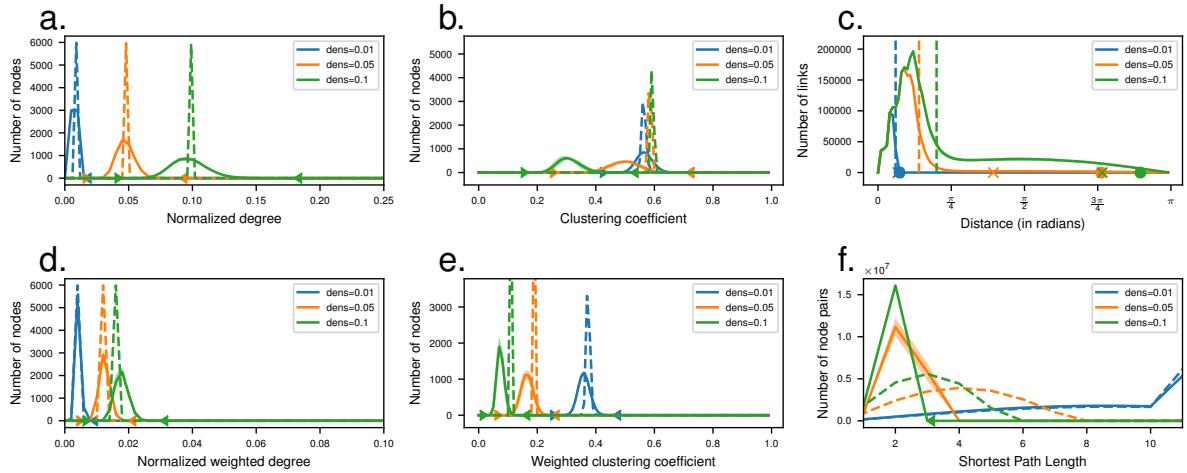

Figure E.3: Same as Fig. 7 but with $n = 500$. Compare to Fig. E.4 and observe all quantities slowly converging to their ground truth.

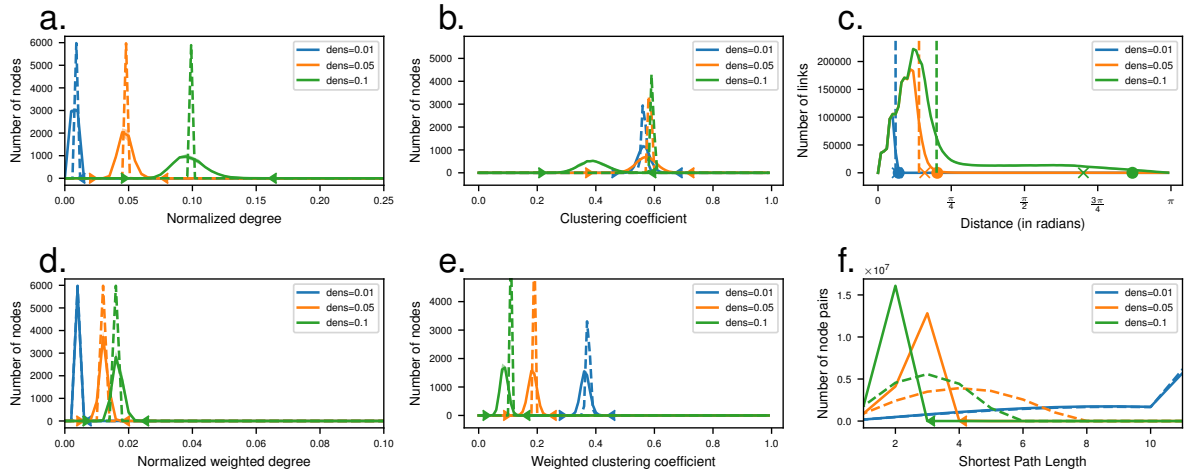

Figure E.4: Same as Fig. 7 but with $n = 1000$. Compare to Fig. E.3 and observe all quantities slowly converging to their ground truth. Weighted degree converges faster than unweighted degree. To distort shortest path length single false shortcuts suffice.

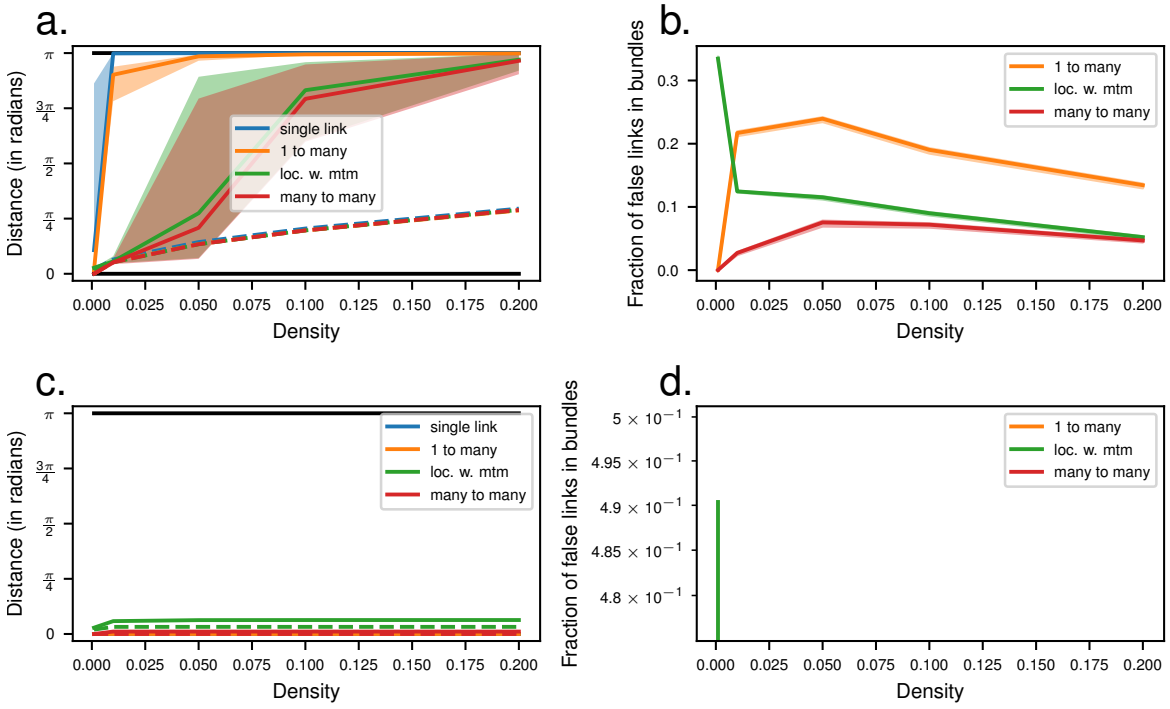

Figure E.5: Same as Fig. 8 but with $n = 500$. While the amount and length of unweighted link bundles has not decreased, the edge weights have converged faster, so that the weighted link bundles have disappeared.

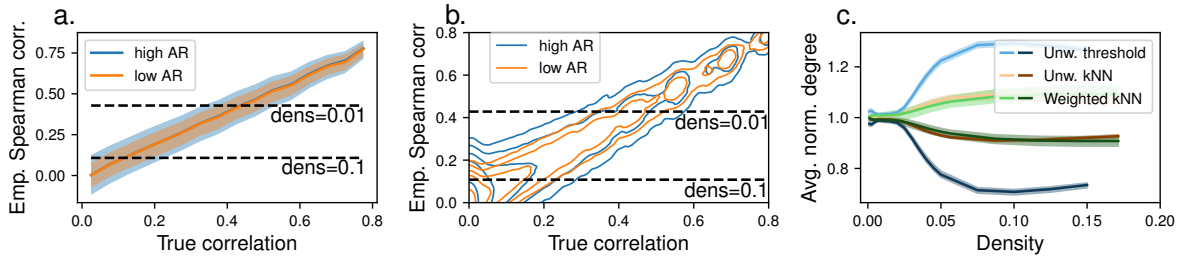

Figure E.6: Same as Fig. 11 but with $n = 500$. While the estimation variance has significantly decreased in absolute terms, the relative bias introduced through anisotropic autocorrelation has not improved.

F Further Simulation Results

F.1 Empirical Betweenness and Forman Curvature Distribution

Sparse lattice-type graphs contain nodes of spuriously high betweenness as soon as one false (tele-)connection between regions is present. The network density where the betweenness distribution is maximally distorted depends on the length scale of the random field (Fig. F.1).

An alternative betweenness measure, which has gained popularity recently, is based on the concept of curvature (Forman, 2003, Ollivier, 2010). It is and edge-based measure and does not consider shortest paths but a distance between neighborhoods of nodes. For Forman curvature we find a strong negative bias (Fig. F.1) as for the clustering coefficient in Fig. 7. Both measures are based on triangle counts.

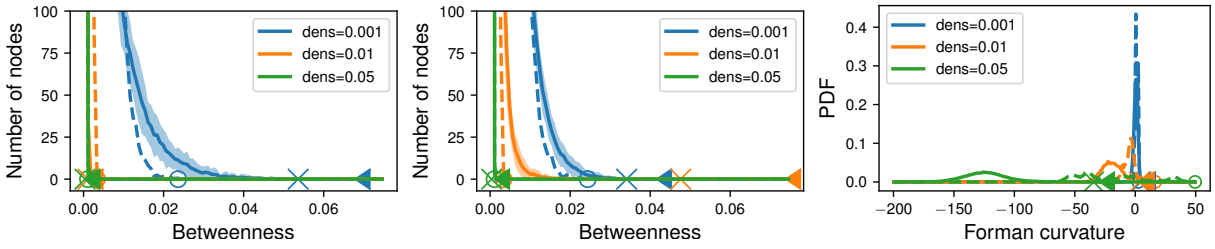

Figure F.1: **Betweenness distributions in empirical networks.** Betweenness histograms with 2σ-uncertainty bands from 30 independent realizations. 'x' and triangle denote the average and 95%-quantile of maximal betweenness values between runs, circle marks the maximal betweenness value in the ground truth graph. Denser networks contain less distortion as many noisy links average out spurious betweenness. *Left*: Density-threshold graphs using Spearman correlation from a MIGRF($\nu = 0.5, \ell = 0.1$). *Middle*: Same as left for MIGRF($\nu = 1.5, \ell = 0.2$). Note that maximal distortion occurs at higher density compared to smaller length scale, as very sparse networks here do not contain spurious long-range links. The network density, where the reliable lattice estimation transitions into more false links, is the most distorted regime in terms of betweenness. *Right*: Forman curvature distribution on the edges for networks from an MIGRF($\nu = 1.5, \ell = 0.1$). The empirical distribution is heavily downward biased, as false links are mostly 'between-cluster' links, i.e. have negative curvature. The more the lattice-like, clustered ground truth structure is disrupted, the more negative the distribution becomes.

F.2 Network Measures For Small Length Scale

Networks with large densities can not be well estimated as ground truth correlations are close to 0 already at short distances. This critical distance can be seen in the drop in the link length distribution plot that marks a phase transition from systematic short links to random false links. The estimates of network characteristics become unreliable already at small network densities, as more false links occur in all network density regimes.

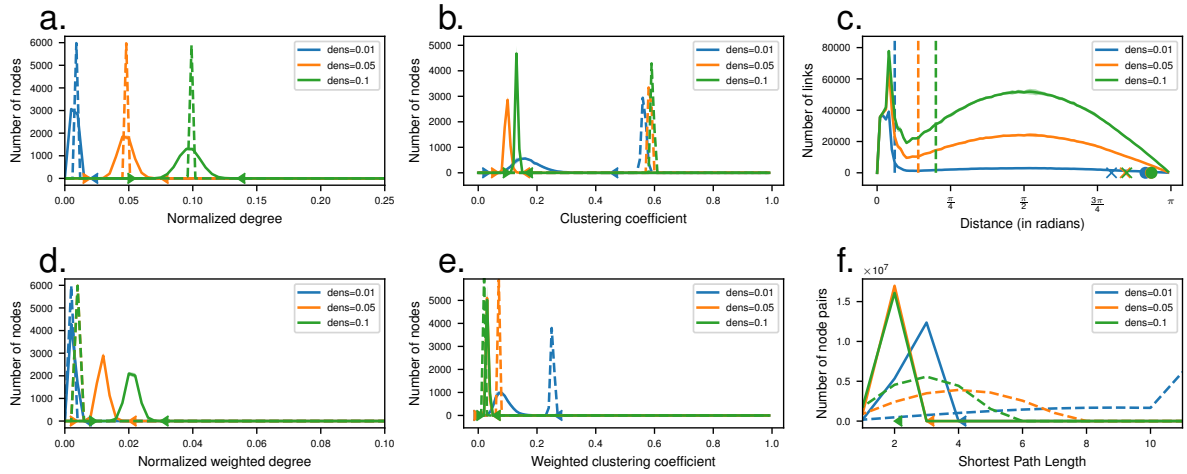

Figure F.2: **Node/edge measure distributions for small length scales.** Same as Fig. 7 but with $\nu = 0.5$ and $\ell = 0.1$. More false links in dense graphs, because more edges are empirically indistinguishable from 0. Less spread out empirical distributions, because less pronounced link bundles are formed and false links rather average out across nodes.

F.3 Networks from Noisy Measurements

Isotropic additive white noise can be understood as a modification of the correlation function to include a 'nugget effect' (Clark, 2010). Then the correlation function approaches a values smaller than 1, when the distance between points approaches 0. Figure F.3 shows the same graph measures as Fig. 7, but for data with isotropic nugget effect. Under additive noise, all population correlations decrease, hence they lie closer to each other and variance in the graph construction increases. As the correlation between similarity estimates is now bounded away from 1, the independent noise on each node will sometimes produce outliers. Under non-negligible estimation variance and for infinitesimally fine grids, this typically leads to non-vanishing fractions formed as well as unformed links between regions, no matter how small the diameter of these regions. More errors occur in the network, irrespective of distance. Even in sparse networks, under large noise, the link length distribution and shortest path length are very distorted. Otherwise the behaviour is very similar. Under fixed grid resolution, continuous but non-smooth random fields and random fields with nugget effect can resemble each other.

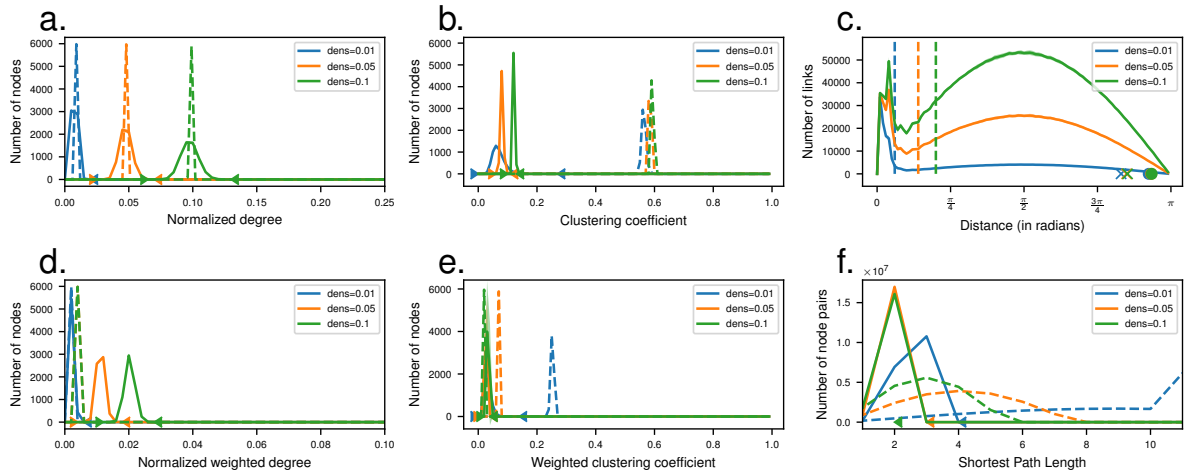

Figure F.3: Same as Fig. 7 but with $0.7N(0, 1)$-additive noise on the data, $\nu = 0.5$ and $\ell = 0.1$.

F.4 Network Measures of kNN Graphs

Figure F.4 shows the same graph measures as Figure 7, but for kNN graphs with similar (but not the same) densities. As the ground truth graphs of threshold and kNN graphs are almost identical for isotropic data, the network characteristics behave very similarly. Observe a more pronounced peak in the unweighted degree distribution. As the ground truth behaviour of kNN and threshold graphs is the same in the isotropic setting, they do not differ much. In kNN graphs the nodes become more similar so that extreme values lie closer together. In case of the clustering coefficient this does not mean that that they lie closer to the ground truth values.

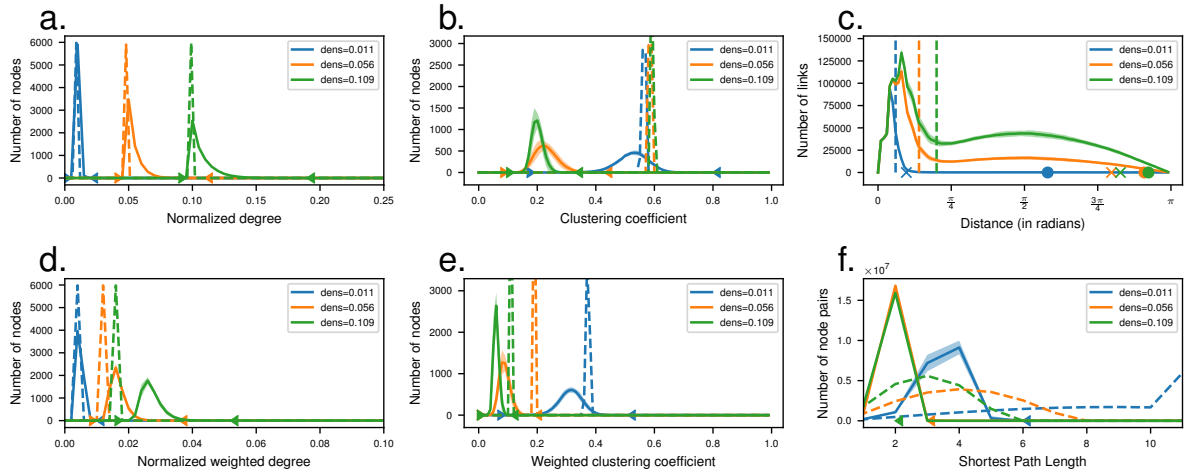

Figure F.4: Same as Fig. 7 but for kNN graphs.

F.5 Network Measures for Temperature Data

Figure F.5 shows the same graph measures for monthly t2m. Degree and clustering coefficient distributions are more stretched out, as different locations differ in distribution, but otherwise - especially the link length distribution - the plots look very similar to the empirical networks from our isotropic data. kNN graphs (Fig. F.6) augment this similarity, as they make the network more isotropic.

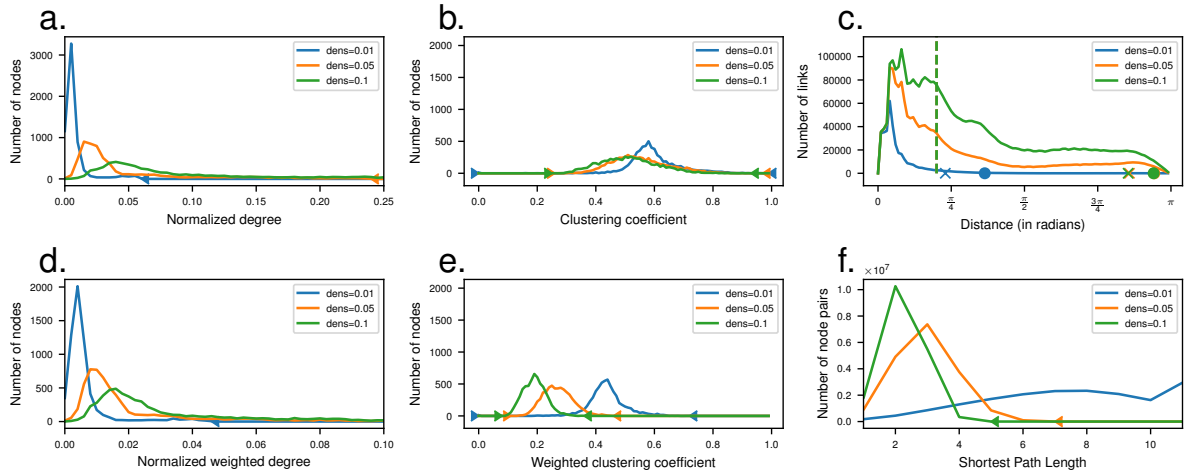

Figure F.5: Same as Fig. 7 but for t2m. Observe even more stretched out degree and clustering coefficient distributions as there are distributional differences between locations. Otherwise there is striking similarity to our isotropic data, especially in the link length distribution.

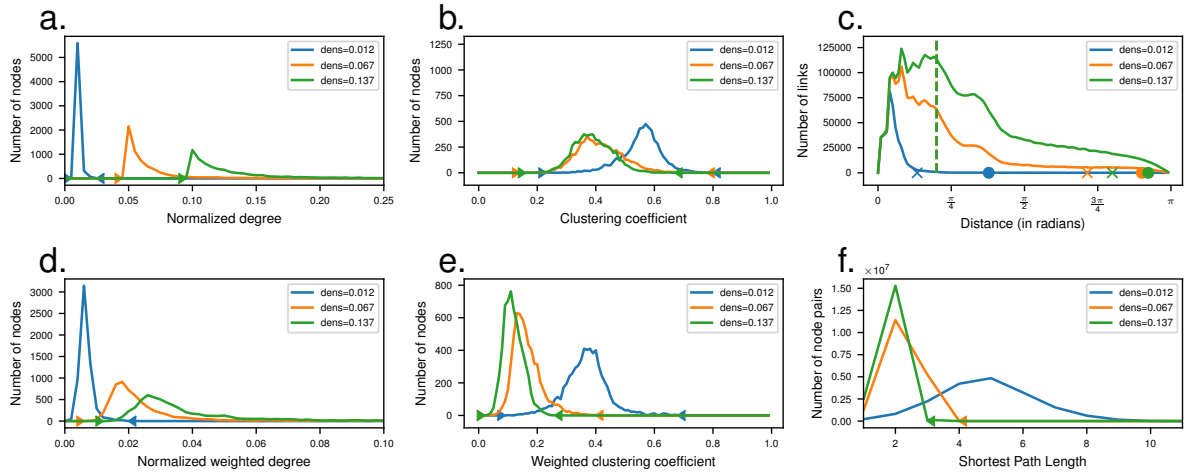

Figure F.6: Same as Fig. 7 but kNN graphs for t2m. kNN graphs increase isotropy of the network. Its characteristics look even more similar to those of our isotropic data.

F.6 Empirical Mutual Information Networks Show Less Bundling Behaviour

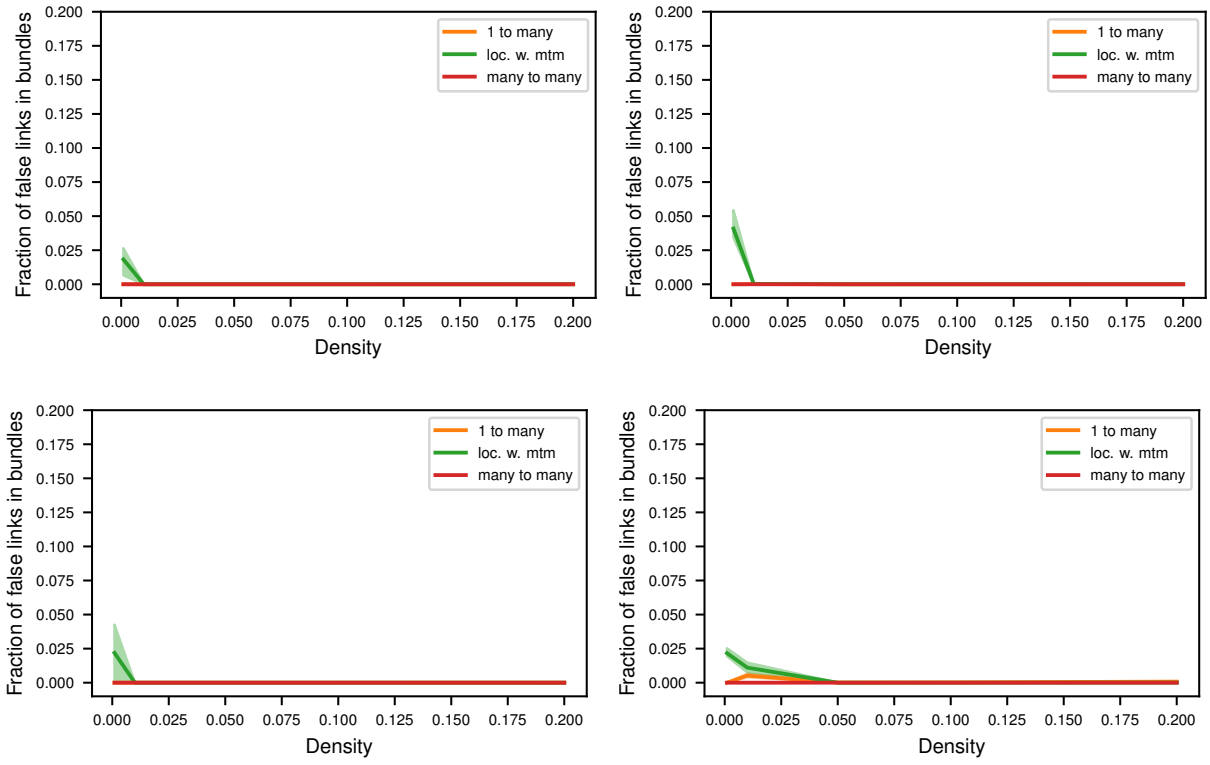

Figure F.7: **Links are much less bundled under the BIKSG estimator.** The fraction of false links in bundles as in Fig. 8 for unweighted BIKSG estimated networks for different hyperparameters of the MIGRF. The top row shows $\nu = 0.5$, the bottom row shows $\nu = 1.5$. The left column shows $\ell = 0.1$, the right column shows $\ell = 0.2$. Although the estimator makes a lot more errors, it does not transmit the localized correlation structure in the underlying data.

The BIKSG mutual information estimates are not as spatially dependent as the Pearson correlation estimates. Thus although the mutual information networks contain more false links, these are not as bundled (Fig. F.7) compared to empirical Pearson correlation in Fig. 8. Hence the estimator plays an important role in how much structural as well as spurious bundling behaviour is present in empirical networks.

F.7 Unstable Links in Climate Networks

Figure F.8 shows the link length distributions and length distributions of differing links of bootstrapped Pearson correlation networks from several climatic variables. Observe that short links are formed first and then the distribution approaches a sinusoidal form. Especially for precipitation, the network of density 0.1 has a distribution close to that of random graphs with uniform link distribution; most links differ between bootstrap samples. Geopotential heights operate on long length scales and have few differing links. This indicates that these networks are robust, especially for large densities (Fig. F.9).

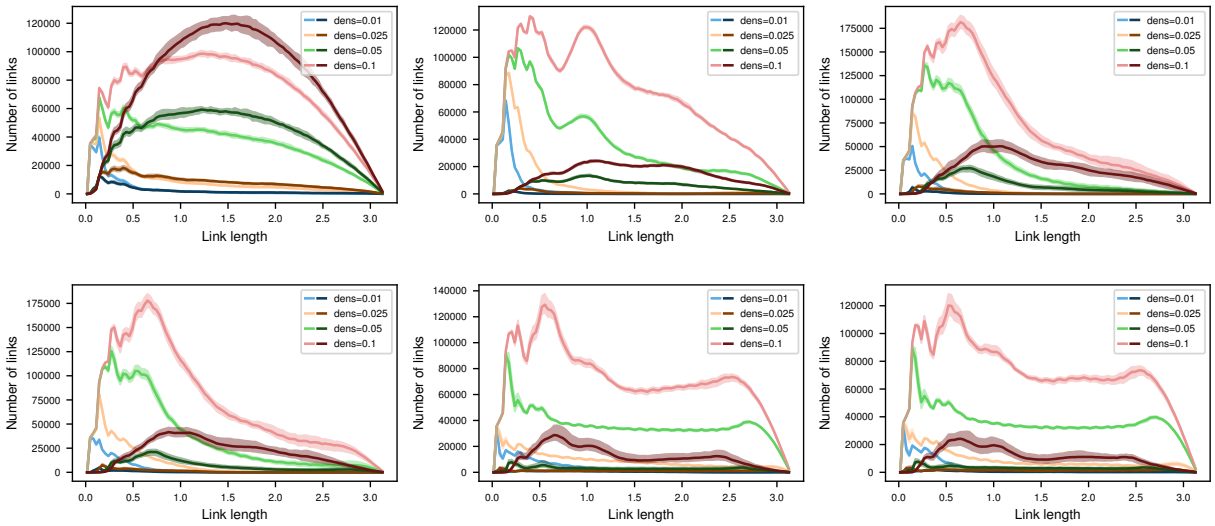

Figure F.8: Link length distributions (light) and length distribution of differing links of bootstrapped Pearson correlation networks from real data. *Top*: Precipitation, daily t2m, surface pressure. *Bottom*: Geopotential heights at 850, 500 and 250 mb.

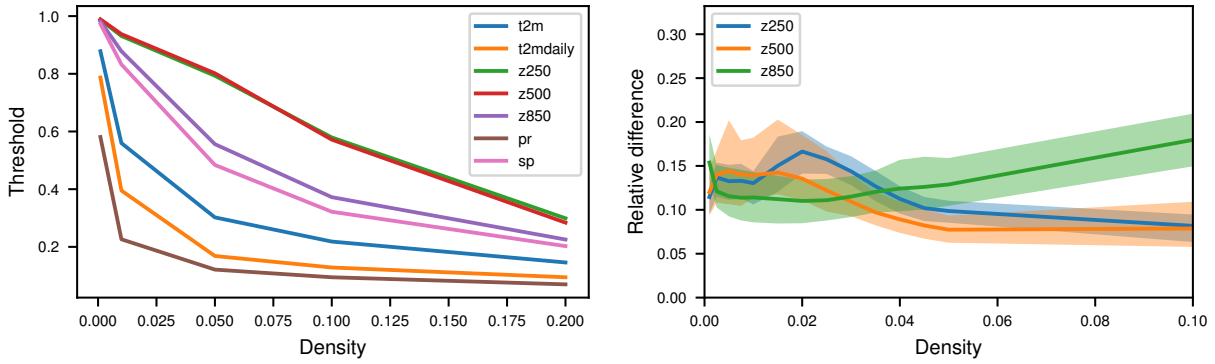

Figure F.9: *Left*: The threshold of the Pearson correlation network as a function of network density. *Right*: Same as Fig. 4 but for geopotential heights. How the relative difference between empirical networks is computed is described in Section 3a2.

F.8 Distribution of mean and variance estimates from IAAFT surrogates

Here we analyse why all edges in sparse IAAFT-based z-score networks are formed by nodes with low autocorrelation. For this purpose we compare the estimates on the edges between nodes of low autocorrelation with the estimates on the edges between nodes of high autocorrelation. Figure F.10 shows the difference between nodes of high and nodes of low autocorrelation in the distributions of mean estimates $\hat{\mu}_{ij}^0$ (left) and variance estimates $\hat{\sigma}_{ij}^0$ (right) based on IAAFT resampling. Positive values indicate more weight in the distribution of highly autocorrelated nodes and vice-versa. Naturally, the variance of the mean estimates grows with autocorrelation (left). The variance estimates for highly autocorrelated nodes are larger (right), reflecting the true increased estimation variance. Since we divide by variance estimates, which are small in absolute value, the highest z-scores are attained for nodes with low autocorrelation.

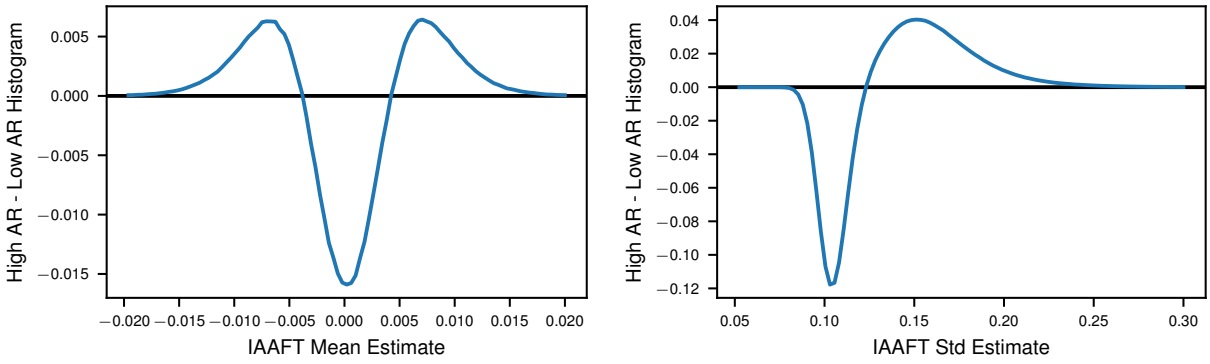

Figure F.10: Difference between nodes of high lag-1 autocorrelation 0.7 and nodes of low lag-1 autocorrelation 0.2 of the normalized histograms of mean estimates (*left*) and standard deviation estimates (*right*) based on IAAFT resampling. Positive values indicate more weight in the distribution associated with high autocorrelation and negative values indicate more weight in the distribution associated with low autocorrelation. Mean estimates have larger variance for high autocorrelation. Variance estimates are larger for high autocorrelation, as they should be.

References

A. Antos and I. Kontoyiannis. Convergence properties of functional estimates for discrete distributions. *Random Structures & Algorithms*, 19:163–193, 2001.

E. Bendito, A. Carmona, A. M. Encinas, and J. M. Gesto. Estimation of fekete points. *Journal of Computational Physics*, 225:2354–2376, 2007.

M. Bevilacqua, C. Gaetan, J. Mateu, and E. Porcu. Estimating space and space-time covariance functions for large data sets: A weighted composite likelihood approach. *Journal of the American Statistical Association*, 107:268–280, 2012.

M. Bevilacqua, C. Caamaño-Carrillo, and E. Porcu. Unifying compactly supported and Matern covariance functions in spatial statistics. *Journal of Multivariate Analysis*, 189:104949, 2020.

S. Bialonski, M. T. Horstmann, and K. Lehnertz. From brain to earth and climate systems: Small-world interaction networks or not? *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 20:013134, 2010.

P. J. Brockwell and R. A. Davis. *Time series: theory and methods*. Springer New York, 1991.

E. Bullmore and O. Sporns. The economy of brain network organization. *Nature reviews. Neuroscience*, 13:336–49, 2012.

W. Chen, M. G. Genton, and Y. Sun. Space-time covariance structures and models. *Annual Review of Statistics and Its Application*, 8:191–215, 2021.

K. Chwialkowski and A. Gretton. A kernel independence test for random processes. *Proceedings of the 31st International Conference on Machine Learning*, 32:1422–1430, 2014.

I. Clark. Statistics or geostatistics? Sampling error or nugget effect? *Journal of the Southern African Institute of Mining and Metallurgy*, 110:307 – 312, 2010.

P. Clifford, S. Richardson, and D. Hemon. Assessing the significance of the correlation between two spatial processes. *Biometrics*, 45:123, 1989.

A. Datta, S. Banerjee, A. O. Finley, and A. E. Gelfand. Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111:800–812, 2016.

J. Fan, Y. Liao, and H. Liu. An overview on the estimation of large covariance and precision matrices. *Econometrics Journal*, 19:C1–C32, 2015.

T. Fardet and A. Levina. Weighted directed clustering: Interpretations and requirements for heterogeneous, inferred, and measured networks. *Phys. Rev. Research*, 3:043124, 2021.

R. R. Forman. Bochner's method for cell complexes and combinatorial Ricci curvature. *Discrete & Computational Geometry*, 29:323–374, 2003.

W. Gao, S. Kannan, S. Oh, and P. Viswanath. Estimating mutual information for discrete-continuous mixtures. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

W. Gao, S. Oh, and P. Viswanath. Demystifying fixed k-nearest neighbor information estimators. *IEEE Transactions on Information Theory*, 2018.

A. Gerhardus and J. Runge. High-recall causal discovery for autocorrelated time series with latent confounders. *Advances in Neural Information Processing Systems*, 33:12615–12625, 2020.

A. R. Gilpin. Table for conversion of Kendall's tau to Spearman's rho within the context of measures of magnitude of effect for meta-analysis. *Educational and Psychological Measurement*, 53:87–92, 1993.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

J. Guinness and M. Fuentes. Isotropic covariance functions on spheres: Some properties and modeling considerations. *Journal of Multivariate Analysis*, 143:143–152, 2016.

P. Guttorp and T. Gneiting. Studies in the history of probability and statistics xlix on the Matérn correlation family. *Biometrika*, 93:989–995, 2006.

J. Hlinka, D. Hartman, N. Jajcay, D. Tomeĉek, J. Tintêra, and M. Paluŝ. Small-world bias of correlation networks: From brain to climate. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 27, 2017.

C. Hurter, O. Ersoy, and A. Telea. Graph bundling by kernel density estimation. *Computer Graphics Forum*, 31: 865–874, 2012.

A. Kraskov, H. Stoegbauer, and P. Grassberger. Estimating mutual information. *Physical Review E*, 69:16, 2003.

C. Lam. High-dimensional covariance matrix estimation. *Wiley Interdisciplinary Reviews: Computational Statistics*, 12:e1485, 2020.

A. Lambert, R. Bourqui, and D. Auber. 3d edge bundling for geographical data visualization. *Proceedings of the International Conference on Information Visualisation*, 2010.

A. Lang and C. Schwab. Isotropic Gaussian random fields on the sphere: Regularity, fast simulation and stochastic partial differential equations. *The Annals of Applied Probability*, 25(6), 2015.

O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88:365–411, 2004.

O. Ledoit and M. Wolf. Analytical nonlinear shrinkage of large-dimensional covariance matrices. *The Annals of Statistics*, 48:3043–3065, 2020.

N. Levit-Binnun, M. Davidovitch, and Golland Y. Sensory and motor secondary symptoms as indicators of brain vulnerability. *Journal of Neurodevelopmental Disorders*, 5, 2013.

Y. Li. *Modeling and Analysis of Spatially Correlated Data*, chapter 4, . World Scientific, 2009.

F. Lindgren, H. Rue, and J. Lindström. An explicit link between Gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73:423–498, 2011.

V. A. Marčenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1:457–483, 1967.

Z. P. Neal. How small is it? Comparing indices of small worldliness. *Network Science*, 5(1):30–44, 2017.

M. Newman. *Networks*. OUP Oxford, 2018.

T. Nocke, S. Buschmann, J. F. Donges, N. Marwan, H. J. Schulz, and C. Tominski. Review: Visual analytics of climate networks. *Nonlinear Processes in Geophysics*, 22:545–570, 2015.

Y. Ollivier. A survey of Ricci curvature for metric spaces and Markov chains. *Probabilistic Approach to Geometry*, 57, 2010.

J. P. Onnela, J. Saramäki, J. Kertész, and K. Kaski. Intensity and coherence of motifs in weighted complex networks. *Physical Review E*, 71:065103, 2005.

L. Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15:1191–1253, 2003.

A. Papana and D. Kugiumtzis. Evaluation of mutual information estimators for time series. *International Journal of Bifurcation and Chaos*, 19(12):4197–4215, 2009.

M. Pourahmadi. *High-Dimensional Covariance Estimation*. John Wiley & Sons, Inc., 2013.

J. Raymaekers and P. J. Rousseeuw. Fast robust correlation for high-dimensional data. *Technometrics*, 63:184–198, 2021.

S. Romano, N. X. Vinh, K. Verspoor, and J. Bailey. The randomized information coefficient: assessing dependencies in noisy data. *Machine Learning*, 107:509–549, 2018.

M. Rubinov and O. Sporns. Complex network measures of brain connectivity: uses and interpretations. *Neuroimage*, 52(3):1059–1069, 2010.

J. Runge, J. Heitzig, N. Marwan, and J. Kurths. Quantifying causal coupling strength: A lag-specific measure for multivariate time series related to transfer entropy. *Physical Review E, Statistical, nonlinear, and soft matter physics*, 86:061121, 2012.

J. Runge, V. Petoukhov, and J. Kurths. Quantifying the strength and delay of climatic interactions: The ambiguities of cross correlation and a novel measure based on graphical models. *Journal of Climate*, 27:720–739, 2014.

J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Glymour, M. Kretschmer, M. D. Mahecha, J. Muñoz-Marí, E. H. van Nes, J. Peters, R. Quax, M. Reichstein, M. Scheffer, B. Schölkopf, P. Spirtes, G. Sugihara, J. Sun, K. Zhang, and J. Zscheischler. Inferring causation from time series in Earth system sciences. *Nature Communications*, 10:1–13, 2019.

D. Selassie, B. Heller, and J. Heer. Divided edge bundling for directional network data. *IEEE Transactions on Visualization and Computer Graphics*, 17:2354–2363, 2011.

M. L. Stein. *Interpolation of Spatial Data*. Springer New York, 1999.

M. L. Stein. Space–time covariance functions. *Journal of the American Statistical Association*, 100:310–321, 2011.

F. Strnad, J. Schlör, C. Fröhlich, and B. Goswami. Teleconnection patterns of different El Niño types revealed by climate network curvature. *Geophysical Research Letters*, 49, 2022.

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

Y. Wang, E. Ghumare, R. Vandenberghe, and P. Dupont. Comparison of different generalizations of clustering coefficient and local efficiency for weighted undirected graphs. *Neural Computation*, 29(2):313–331, 2017.

D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature 1998 393:6684*, 393:440–442, 1998.

P. Whittle. On stationary processes in the plane. *Biometrika*, 41:434–449, 1954.

T. Zerenner, P. Friederichs, K. Lehnertz, and A. Hense. A Gaussian graphical model approach to climate networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 24:023103, 2014.