δ-MAPS: From spatio-temporal data to a weighted and lagged network between functional domains

Ilias Fountalis
School of Computer Science
Georgia Tech
fountalis@gatech.edu

Annalisa Bracco
School of Earth and Atmospheric Sciences
Georgia Tech
abracco@gatech.edu

Bistra Dilkina
School of Computational Science and Engr
Georgia Tech
bdilkina@cc.gatech.edu

Constantine Dovrolis
School of Computer Science
Georgia Tech
constantine@gatech.edu

Shella Keilholz
Dept. of Biomedical Engr
Georgia Tech and Emory
shella.keilholz@bme.gatech.edu

ABSTRACT

We propose δ-MAPS, a method that analyzes spatio-temporal data to first identify the distinct spatial components of the underlying system, referred to as "domains", and second to infer the connections between them. A domain is a spatially contiguous region of highly correlated temporal activity. The core of a domain is a point or subregion at which a metric of local homogeneity is maximum across the entire domain. We compute a domain as the maximum-sized set of spatially contiguous cells that include the detected core and satisfy a homogeneity constraint, expressed in terms of the average pairwise cross-correlation across all cells in the domain. Domains may be spatially overlapping. Different domains may have correlated activity, potentially at a lag, because of direct or indirect interactions. The proposed edge inference method examines the statistical significance of each lagged cross-correlation between two domains, infers a range of lag values for each edge, and assigns a weight to each edge based on the covariance of the two domains. We illustrate the application of δ-MAPS on data from two domains: climate science and neuroscience.

1. INTRODUCTION

Spatio-temporal data become increasingly prevalent and important for both science (e.g., climate, systems neuroscience, seismology) and enterprises (e.g., the analysis of geotagged social media activity). The spatial scale of the available data is often determined by an arbitrary grid, which is typically larger than the true dimensionality of the underlying system. One major task is to identify the distinct semi-autonomous components of this system and to infer their (potentially lagged and weighted) interconnections from the available spatio-temporal data. Traditional dimensionality reduction methods, such as PCA, ICA or clustering, have been successfully used for many years but they have known limitations when the objective is to infer the functional network between all spatial components of the system.

We propose δ-MAPS, an inference method that first identifies these spatial components, referred to as "domains", and then the connections between them (§3). Informally, a *functional domain* (or simply *domain*) is a spatially contiguous region that somehow participates in the same dynamic effect or function. The exact mechanism that creates this effect or function varies across application domains; however, the key idea is that *the functional relation between the grid cells of domain results in highly correlated temporal activity.* If we accept this premise, it follows that we should be able to identify the "epicenter" or *core of a domain* as a point (or subregion) at which the local homogeneity is maximum across the entire domain. Instead of searching for the discrete boundary of a domain, which may not exist in reality, we compute a domain as the *maximum possible set* of spatially contiguous cells that include the detected core, and that satisfy a homogeneity constraint, expressed in terms of the average pairwise cross-correlation across all cells in the domain. Domains may be spatially overlapping. Also, some cells may not belong to any domain.

After we identify all domains, δ-MAPS infers a functional network between them. Different domains may have correlated activity, potentially at a lag, because of direct or indirect interactions. The proposed edge inference method examines the statistical significance of each lagged cross-correlation between two domains, applies a multiple-testing process to control the rate of false positives, infers a range of potential lag values for each edge, and assigns a weight to each edge based on the covariance of the corresponding two domains.

δ-MAPS is related to clustering, parcellation (or regionalization), network community detection, multivariate statistical methods for dimensionality reduction such as PCA and ICA, as well as functional network and lag inference methods. However, as we discuss in §2 and show with synthetic data experiments in §4, δ-MAPS is also significantly different than all these methods. δ-MAPS does not require the number of domains as an input parameter, the resulting domains are spatially contiguous and potentially overlap-

ping, and the inferred connections between domains can be lagged and positively or negatively weighted. Further, the distinction between grid cells that are correlated within the same domain and grid cells that are correlated across two distinct domains allows δ-MAPS to separate between local diffusion (or dispersion) phenomena and remote interactions that may be due to underlying structural connections (e.g., a white-matter fiber between two brain regions).

We illustrate the application of δ-MAPS on data from two domains: climate science (§5) and neuroscience (§6). First, the sea-surface temperature (SST) climate network identifies some well-known climate "tele-connections" (such as the lagged connection between the El Niño Southern Oscillation and the Indian ocean) but it also suggests the existence of some new connections that deserve further investigation by the domain experts. Second, the analysis of resting-state fMRI cortical data confirms the presence of three well-known functional brain "networks" (default-mode, occipital, and motor/somatosensory), and shows that the cortical network includes a *backbone* of relatively few regions that are densely interconnected.

2. RELATED WORK

A common approach to reduce the dimensionality of spatio-temporal data is to apply PCA (standard or rotated) or ICA techniques. For instance, in climate science, PCA (also known as Empirical Orthogonal Function (EOF) analysis) has been used to identify teleconnections between distinct climate regions [43]. The orthogonality between PCA components complicates the interpretation of the results making it difficult to identify the distinct underlying modes of variability and to separate their effects, as clearly discussed in [12]. ICA analysis is more common in the neuroscience literature, aiming to identify independent rather than orthogonal components [20]. However, ICA does not provide a relative significance for each component, and the number of independent components should be chosen based on some additional information about the underlying system.

Another broad family of spatio-temporal dimensionality reduction methods is based on clustering [5, 16, 35, 45]. These algorithms can be grouped into region-growing methods (e.g., [6, 25]), spectral (e.g., the NCUT method often applied in fMRI analysis [11, 39] – but also see a discussion of their limitations [3]), hierarchical (e.g., [7, 38]), and probabilistic (e.g., [3, 19]). These groups of algorithms are quite different but they share some common characteristics: the resulting clusters may not be spatially contiguous, they are typically non-overlapping, every grid cell needs to belong to a cluster (potentially excluding only outliers), and the number of clusters is often required as an input parameter. In particular, the lack of spatial contiguity makes it hard to distinguish between correlations due to spatial diffusion (or dispersion) phenomena from correlations that are due to remote (structural) interactions between distinct effects.

An approach of increasing popularity is to first construct a correlation-based network between individual grid cells, after prunning cross-correlations that are not statistically significant – see [23]. Then, some of these methods analyze the (binary or weighted) cell-level network directly based on various centrality metrics, k-core decomposition, spectral analysis, etc (e.g., [13, 40]) or they first apply a community detection algorithm (potentially able to detect overlapping communities, e.g., [1, 24, 28]) on the cell-level network and then analyze the resulting communities in terms of size, density, location, overlap, etc (e.g., [27, 29, 36, 37]). A community however may group together two regions that are, first, not spatially contiguous, and second, different in terms of how they are connected to other regions; an instance of this issue is illustrated in Fig. 4-C in the context of climate data analysis.

3. δ-MAPS

The input data is generated from a *spatial field* $\mathbf{X}(t)$ sampled on an arbitrary *grid* G. This grid can be modeled as a planar graph $G(V, E)$, where each vertex in V is a grid cell and each edge in E represents the spatial adjacency between two neighboring cells. A set of cells $A \subseteq V$ is *spatially contiguous*, denoted by $I_G(A)$=1, if it forms a connected component in G.

The *K-neighborhood* of a cell i, denoted by $\Gamma_K(i)$, includes i and the set of K nearest neighbors to i according to an appropriate spatial distance metric (e.g., geodesic distance for climate data, Euclidean distance for fMRI data). The K-neighborhood of a cell is always spatially contiguous.

Each grid cell i is associated with a time series $x_i(t)$ of length T ($t \in \{1, \ldots T\}$). We assume that $x_i(t)$ is sampled from a stationary signal and denote by $\tilde{\mu}_i$ and $\tilde{\sigma}_i^2$ its sample mean and variance, respectively. The similarity between the activity of two cells i and j is measured with Pearson's cross-correlation at zero-lag,

$$r_{i,j} = \frac{\sum_{t=1}^{T}(x_i(t) - \tilde{\mu}_i)(x_j(t) - \tilde{\mu}_j)}{T\,\tilde{\sigma}_i\tilde{\sigma}_j}\,. \tag{1}$$

Other similarity metrics could be used instead.

The *local homogeneity at cell i* is defined as the average pairwise cross-correlation between the $K+1$ cells in $\Gamma_K(i)$,

$$\hat{r}_K(i) = \frac{\sum_{m \neq n \in \Gamma_K(i)} r_{m,n}}{K\,(K+1)}\,. \tag{2}$$

Similarly, we define the *homogeneity of a set of cells A* as the average pairwise cross-correlation between all distinct cells in A,

$$\hat{r}(A) = \frac{\sum_{m \neq n \in A} r_{m,n}}{|A|\,(|A|-1)}\,. \tag{3}$$

3.1 Functional domains

Intuitively, a *domain A* is a spatially contiguous set of cells that somehow participate in the same dynamic effect or function. The exact mechanism that creates this effect or function varies across application domains; however, the key premise is that *the functional relation between the cells of domain A results in highly correlated temporal activity (at zero-lag), and thus high values of the homogeneity metric $\hat{r}(A)$*. A given *homogeneity threshold* δ examines if the homogeneity of A is sufficiently high, i.e., a domain A must have $\hat{r}(A) > \delta$. (the selection of δ is discussed later in this section).

If we accept this premise, it follows that we should be able to identify the "epicenter" or *core of a domain A* as a cell $i \in A$ at which the local homogeneity $\hat{r}_K(i)$ is maximum across all cells in A (and certainly larger than δ). In general, the core of a domain may not be a unique cell.

More formally now, suppose that we know that cell c is in the core of a domain. The *domain A rooted at c* has

to satisfy the following three properties: it should include cell c, be spatially contiguous, and have higher homogeneity than δ:

$$c \in A, \quad I_G(A) = 1, \quad \hat{r}(A) > \delta\,. \tag{4}$$

A domain may not have sharp spatial boundaries; instead, it may gradually "fade" into other domains or regions dominated by noise. So, instead of searching for the discrete boundary of a domain, it is more reasonable to compute a domain as the *largest possible set of cells* that satisfies the previous three constraints.

Domain identification problem: Given the field $\mathbf{X}(t)$ on the spatial grid G, a core cell c, and the threshold δ, the domain $A(c)$ is a maximum-sized set of cells that satisfies the three constraints of (4). In Appendix-1 we prove that the decision version of this problem is NP-Hard.

A given spatial field $\mathbf{X}(t)$ may include several domains. The number of identified domains, denoted by N, depends on the threshold δ. Domains may be spatially overlapping; this is the case when the cells of a region are significantly correlated with two or more distinct domain cores. Also, some cells of the grid may not belong to any domain, meaning that their signal can be thought of as mostly noise (at least for the given value of δ). Decreasing δ will typically result in a larger number of detected domain cores. Further, as δ decreases, the spatial extent of each domain will typically increase, resulting in larger overlaps between nearby domains.

δ can simply be a user-specified parameter for the minimum required average cross-correlation within a domain. Another way is to calculate δ based on a statistical test for the significance of the observed zero-lag cross-correlations. A summary of this method is given next (described in more detail in Appendix-2). We start with a random sample of pairs of grid cells. We then apply the statistical test described in §3.2 (see Equations 6 and 7) to examine if the zero-lag cross-correlation between each of these pairs passes a given significance level α (set to 10^{-2} unless specified otherwise). δ is then set to the average of the statistically significant cross-correlations in that sample. The rationale is that the average pairwise cross-correlation among cells that belong to the same domain should be higher than a sample average of statistically significant cross-correlations between cells that can be anywhere on the grid.

3.1.1 Algorithm for domain identification

Given the NP-Hardness of the previous problem, we propose a greedy algorithm that runs in two phases. In the first phase, we identify a set of cells, referred to as *seeds*; each seed is a candidate core for a domain. In the second phase, each seed is initially considered as a distinct domain. Then, an iterative and greedy algorithm attempts to identify the largest possible domains that satisfy the three constraints of (4) through a sequence of *expansion* and *merging* operations. The two phases are described next, while the complete pseudocode is presented in Appendix-3. The source code (including supporting documentation) will be available on GitHub before the final publication of this paper.

Seed selection. Recall that the core of a domain is a cell of maximum local homogeneity across all cells of that domain. So, one way to detect *potential* core cells, while the domains are still unknown, is to identify points at which the homogeneity field $\hat{r}_K(i)$ is locally maximum. Specifically, cell i is a seed if $\hat{r}_K(i) > \delta$ and $\hat{r}_K(i) \geq \hat{r}_K(j)\ \forall j \in \Gamma_K(i)$. Let S be the set of all identified seeds.

In general, a single domain may produce more than one seed because the local homogeneity field can be noisy and so it may include multiple local maxima, greater than δ. Further, additional seeds can appear in regions where domains overlap. Consequently, it is necessary to include a merging operation in which two or more seeds are eventually merged into the same domain.

Note that as K decreases, the local homogeneity field becomes more noisy and so we may detect more seeds in the same domain. On the other hand, larger values of the neighborhood size K can oversmooth the homogeneity field, removing seeds and potentially hiding entire domains. The latter is more likely if the spatial extent of a domain is smaller than $K+1$ cells. This observation implies that the spatial resolution of the given grid sets a lower bound on the size of the functional domains that can be detected.

Domain-merging operation. Two candidate domains A and B can be merged if they are spatially contiguous and if the homogeneity of their union is sufficiently high, i.e., $\hat{r}(A \cup B) > \delta$. Whenever there is more than one pair of domains that can be merged, we greedily choose the pair with the maximum union homogeneity; this greedy choice makes the merged domain more likely to expand further.

The merging operation is performed initially on the set of seeds S. It is also performed after each domain-expansion operation, whenever it is possible to do so.

Domain-expansion operation. A domain A is expanded by considering all cells that are adjacent to A, and selecting the cell i that maximizes $\hat{r}(A \cup \{i\})$; again, this greedy choice makes the expanded domain more likely to expand further.

The expansion operation is repeated in rounds. At the start of each round, domains are sorted in decreasing order of homogeneity. Then, each domain is expanded by one cell at a time, as previously described, in that order. After every expansion operation, we check whether one or more merging operations are possible. A round is complete when we have attempted to expand each domain once.

A domain can no longer expand if that would violate the homogeneity constraint δ or if there are no other adjacent cells that can be added into the domain. The domain identification algorithm terminates when no further expansion or merging operations are possible.

3.2 The domain network

Given the N identified domains $V_\delta = \{A_1, \ldots A_N\}$, the next step is to construct a network $G_\delta(V_\delta, E_\delta)$ between domains. Different domains may have correlated activity because of direct or indirect interactions. We refer to G_δ as a *functional network* to emphasize that the edges between domains are based on functional activity and correlations instead of structural or physical connections ("structural network") or causal interactions ("effective network").

We associate a *domain-level signal* $X_A(t)$ with each domain A. The definition of this signal depends on the specific application field. For instance, when we analyze climate anomaly time series, the domain-level signal is defined as the *cumulative anomaly* across all cells of that domain,

where the contribution of each signal is weighted by the relative size of that cell (it depends on the cell's latitude). For fMRI data, the domain-level signal is defined as the *average BOLD signal* across the cells of that domain.

Two different domains may be located at some distance, and so they may be correlated at a non-zero lag τ. For this reason, we examine if there is a significant cross-correlation between different domains over a range of lags ($-\tau_{max} \leq \tau \leq \tau_{max}$). The sample cross-correlation between domains A and B at a lag τ can be estimated as:

$$r_{A,B}(\tau) = \frac{\sum_{t=1}^{T-\tau}(X_A(t) - \tilde{\mu}_A)(X_B(t+\tau) - \tilde{\mu}_B)}{T\tilde{\sigma}_A\tilde{\sigma}_B}, \quad (5)$$

where $\tilde{\mu}_A$ and $\tilde{\sigma}_A$ denote sample mean and standard deviation estimates, respectively. The selection of τ_{max} should be large enough to include the typical signal propagation delays in the underlying system but at the same time it should be much lower than T. The $2\tau_{max}+1$ cross-correlations for a pair of domains can be represented with a *correlogram*; an example based on climate sea-surface temperature data (see §5) is shown in Fig. 1.

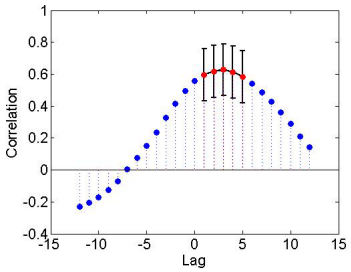


Figure 1: Correlogram between two climate time series for a lag range of ±12 months. We show the significant correlations for a false discovery rate $q = 10^{-3}$ with red. The error bars correspond to ± one standard deviation, as estimated by Eq. (6).

The next step is to examine the statistical significance of the measured cross-correlation between two domains A and B. Two uncorrelated signals can still produce a considerable sample cross-correlation if they have a strong autocorrelation structure. This is captured by Bartlett's formula [8], which is an estimator for the variance of $r_{A,B}(\tau)$ (for a fixed value of τ). Under the null-hypothesis that the domain-level signals of A and B are uncorrelated,

$$\mathrm{Var}[r_{A,B}(\tau)] = \frac{1}{T-\tau}\sum_{\tau_k=-T}^{T} r_{A,A}(\tau_k)\, r_{B,B}(\tau_k), \quad (6)$$

where $r_{A,A}(\tau_k)$ is the autocorrelation of the time series of domain A at lag τ_k.

Under the previous null-hypothesis, the expected value of $r_{A,B}(\tau)$ is zero and the following statistic approximately follows the standard normal distribution $N(0,1)$:

$$z_{A,B}(\tau) = \frac{r_{A,B}(\tau)}{\sqrt{\mathrm{Var}[r_{A,B}(\tau)]}}. \quad (7)$$

The approximation is due to the fact that $r_{A,B}(\tau)$ is bounded between $[-1,1]$. So, we can now perform hypothesis testing for every pair of domains, computing a corresponding p-value based on z.

Given that there may be several domains in G_δ, we need to control the number of false positive edges that may result from the multiple testing problem. We do so using the False Discovery Rate (FDR) method of Benjamini and Hochberg [4]. Specifically, given N domains, we need to perform $M = \frac{N(N-1)}{2}(2\tau_{max}+1)$ tests (for each potential edge and for each possible lag value), and compute the p-value for each test, based on (7). Given a False Discovery Rate q (the expected value of the fraction of tests that are false positives), the Benjamini-Hochberg procedure ranks the M p-values (p_i becomes the i'th lowest p-value) and only keeps the first $m < M$ tests (edges), where p_m is the highest p-value such that $p_m < q\,m/M$.

Lag inference and edge directionality. We infer the domain-level network G_δ as follows. Two domains $A, B \in V_\delta$ are connected if there is at least one lag value at which the cross-correlation $r_{A,B}(\tau)$ has passed the FDR test. The standard approach in *lag inference* is to consider the lag value τ^* that maximizes the absolute cross-correlation,

$$\tau^*_{A,B} = \arg\max_{\tau=-\tau_{max}\ldots\tau_{max}} \{|r_{A,B}(\tau)|\}. \quad (8)$$

The corresponding correlation is denoted as $r^*_{A,B}$. There are two problems with this approach. First, it is harder to examine the statistical significance of $|r^*_{A,B}|$ because it is the maximum of a set of random variables.[1] Second, it is often the case that there is a range of lag values that produce "almost maximum" cross-correlations, say within one standard deviation from each other. Focusing on $\tau^*_{A,B}$ and ignoring the rest of the statistically significant and almost equal cross-correlations is not well justified.

Instead, we follow a more robust approach in which an edge of the domain-level network G_δ may be associated with a range of lag values.[2] The lag range that we associate with the edge between A and B, denoted as $R_\tau(A,B)$, is defined as *the range of lags that produce significant cross-correlations, within one standard deviation from* $|r^*_{A,B}|$. If $R_\tau(A,B)$ includes τ=0, the edge is represented as *undirected*. If $R_\tau(A,B)$ includes only positive lags, the edge is directed from A to B meaning that A's signal precedes B's by the given lag range; otherwise, we associate the opposite direction with that edge. We emphasize that the directionality of the edges does *not* imply causality; it only refers to temporal ordering.

Edge weight and domain strength. How to assign a weight to each domain-level edge in G_δ? A common approach is to consider the (signed) magnitude of the cross-correlation $r^*_{A,B}$. This is reasonable if all domain signals have approximately the same signal power. In addition, we propose a new edge weight that is based on the covariance of the two domains:

$$w(A,B) = \mathrm{cov}[X_A(t), X_B(t)] = \tilde{\sigma}_A\,\tilde{\sigma}_B\, r^*_{A,B}. \quad (9)$$

[1] An analytic approach based on extreme-value statistics was proposed in [23] but it relies on several approximations. Numerical approaches based on frequency-domain bootstrapping, on the other hand, are computationally expensive [23, 26, 32].

[2] In principle, it may be a set of lag values. In practice though, significant correlations result for a continuous range of lag values.

The cross-correlation is computed at lag $\tau^*_{A,B}$ but we could use the average of all cross-correlations in $R_\tau(A,B)$ instead. The weight of an edge can be positive or negative depending on the sign of the corresponding cross-correlation.

Finally, the strength of a network node (domain) is defined as the sum of the absolute weights of all edges of that node (ignoring edge directionality).

4. ILLUSTRATION - COMPARISONS

The two objectives of this section are to illustrate how the δ-MAPS method works, and to contrast the results of the latter with commonly used methods such as PCA, ICA, spatial clustering, and overlapping community detection. We rely on synthetic data so that the ground-truth is known.

Synthetic data description. We construct five domains on a 50×70 spatial grid. Each domain i is associated with a "mother" time series $y_i(t)$, (i=1...5). To make the experiment more realistic in terms of autocorrelation structure and marginal distribution, each $y_i(t)$ is a real fMRI time series with length T=1200 (see §6). The five mother time series $y_i(t)$ are uncorrelated (absolute cross-correlation <0.05 at all lags), and they are normalized to zero-mean, unit-variance. To create correlations between domains (i.e., domain-level edges), we construct five new time series $x_i(t)$ based on linear combinations of two or more mother time series. For instance, if we set $x_i(t) = (1-\alpha)y_i(t) + \alpha y_j(t+\tau)$ with $0 < \alpha < 1$ and $x_j(t) = y_j(t)$, domains i and j become positively correlated at a lag τ; the correlation increases with α. The time series x_i are again normalized to zero-mean, unit-variance. We then scale the time series of domain i by a factor $\sqrt{s_i}$ to control the variance of each domain ($\mathrm{Var}[x_i(t)] = s_i$).

For simplicity, each domain is a circle with radius r_p. A domain has a "core region" with the same center and radius $r_c < r_p$; the core is supposed to be the epicenter of that domain. Every point in the core has the same signal $x_i(t)$ (before we add random noise). Outside the core, the signal attenuates at a distance d from the center of the domain as follows:

$$x_i(t) = \sqrt{f(d)}\, x_i(t),\ f(d) = \frac{r_p - d}{r_p - r_c},\ r_c \le d \le r_p\,. \quad (10)$$

Finally, we superimpose white Gaussian noise of zero-mean, unit-variance on the entire grid. The parameters of the five synthetic domains are shown in Table 1. The domains differ in terms of size and power (variance). The spatial extent of the domains is shown in Fig.2-A; domains 1 and 3 overlap with domain 2, while domains 4 and 5 also overlap to a smaller extent. Further, there is a strong and lagged anti-correlation between domains 1 and 3, a weaker positive correlation at zero-lag between domains 4 and 5, and an ever weaker positive correlation at zero-lag between domains 3 and 5. The edges of the domain-level network are also shown in Fig.2-A.

Table 1: Synthetic area generation parameters.

ID	r_c	r_p	s_i	$x_i(t)$
1	2	10	16	$x_1(t) = 2/3y_1(t) - 1/3y_3(t+15)$
2	4	14	11	$x_2(t) = y_2(t)$
3	2	10	16	$x_3(t) = y_3(t)$
4	0.5	5	9	$x_4(t) = 3/4y_4(t) + 1/4y_5(t)$
5	1	7	6	$x_5(t) = 4/5y_5(t) + 1/5y_3(t)$

δ-MAPS results. The parameters of δ-MAPS are set as follows: K=4 cells (up-down-left-right), and δ=0.55 (corresponds to significance level 10^{-2}). In the edge inference step, the FDR threshold is q=10% and $\tau_{max} = 20$.

Fig.2-B shows the local homogeneity field $\hat{r}_K(i)$ as well as as the identified seeds (blue dots), while Fig.2-C shows the five discovered domains. As expected, we often identify more than one seed in the core of each domain due to noise; those seeds are eventually merged into the same domain. The local homogeneity field is weaker in domains 4 and 5 (due to their lower variance) but a seed is still detected in those domains. Seeds also appear at the two overlapping regions between (1,2) and (2,3) but those seeds gradually merge with one of the domains in which they appear.

Each domain is a subset of the domain's true expanse. The reason is that some cells close to the periphery of each domain have very low signal-to-noise ratio (recall that the signal decays to zero at the periphery and so the average correlation between those cells with the rest of their domain does not exceed the δ threshold). More quantitatively, the inferred domains include about 80%-90% of the ground-truth cells in each domain. In non-overlapping regions this fraction is higher (85%-95% of the cells), while in overlapping regions it drops to 45%-80%. The extent of overlapping regions is harder to correctly identify especially when a domain (e.g., domain 2) overlaps with a stronger domain (e.g., domains 1 or 3); the stronger domain effectively masks the signal of the weaker domain. The average pairwise cross-correlation of the cells in each domain varies between 55%-70% in the ground-truth data, while the inferred domains have slightly higher average cross-correlation (65%-75%) due to their smaller expanse.

Finally, Fig. 2-C shows the inferred domain-level network. δ-MAPS identifies correctly the three edges and their polarity (positive versus negative correlations). The lag ranges always include the correct value (e.g., the edge between domains 1 and 3 has a lag range [14,15]). Also, the three edges are correctly ordered in terms of absolute cross-correlation magnitude: (1,3) followed by (4,5), followed by (3,5).

PCA/EOF results. We apply EOF analysis using Matlab's PCA toolbox. Fig. 2-E,F,G show the first three principal components, which collectively account for about 90% of the total variance. A first observation is that domains 4 and 5 are not even visible in these components – they only appear in the next two components, which account for about 5% of the variance each. This is because domains 4 and 5 are smaller and have lower variance. This is a general limitation of PCA: the variance of the analyzed field can be dominated by a small number of "modes of variability", completely masking smaller/weaker regions of interest and their connections. Second, the first three components do not provide a consistent evidence that domains 1 and 3 are strongly anti-correlated; this is due to their lagged correlation, which is missed by PCA. Third, the first component, which accounts for 40% of the total variance, can be misinterpreted to imply that domain 2 is somehow positively correlated with domains 1 and 3, even though it is actually generated by an uncorrelated signal. This is due to the overlap of domain 2 with domains 1 and 3.

ICA results. We apply ICA on the synthetic data using Matlab's FastICA toolbox. To help ICA perform better, we

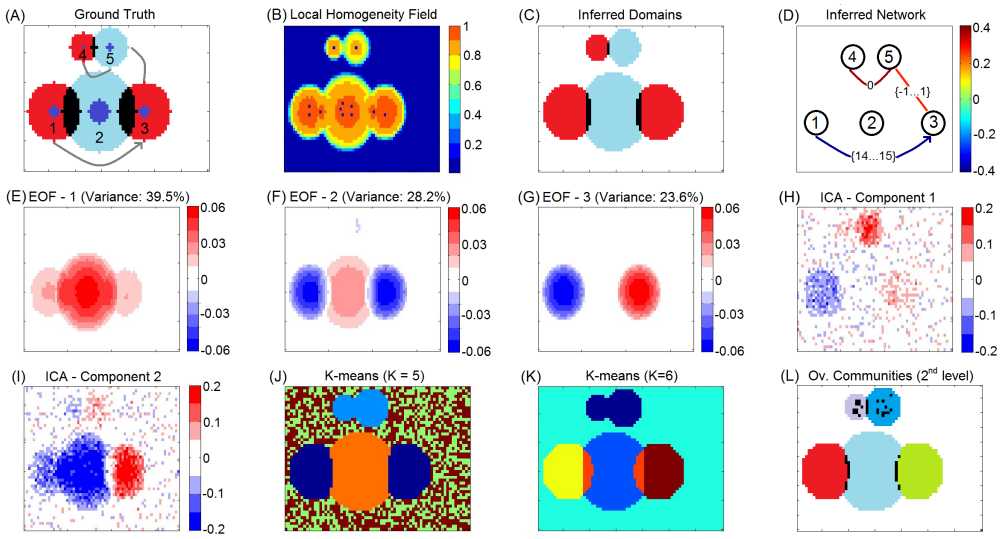


Figure 2: A: The five ground-truth domains. Adjacent domains have different colors, overlapping regions shown in black, and the core of each domain is in blue. The three constructed edges are shown in gray lines. B: The homogeneity field $\hat{r}_K(i)$ at each cell. The identified seeds are shown in blue. C: The inferred domains: adjacent domains have different colors and overlaps are shown in black. D: The inferred domain-level network: the color map refers to the edge correlation. The lag associated with each edge is also shown. E,F,G: The first three EOF (PCA) components. The variance explained by each component is shown at the top of each figure. H,I: The two ICA components. J,K: K-means clustering. L: The second hierarchical level of community structure as identified by OSLOM: each community has a distinct color and overlaps are shown in black.

actually specified the right number of independent components, which is two (domains 1,3,4,5 are indirectly correlated – domain 2 is not correlated with any other). The two independent components are shown in Fig. 2-H,I. Note that only a rough "shadow" of each domain is visible. Domains 1 and 3 appear in different colors, providing a hint that they are anti-correlated, while domains 3 and 5 appear in the same color because they are positively correlated. Overall, however, the components are quite noisy and it would be hard in practice to discover the functional structure of the underlying system if we did not know the ground-truth. The results are even harder to interpret when we request a larger number of components.

Clustering results. We apply the most well-known clustering method, *k-means*, on our synthetic data. As commonly done with correlation-based clustering, the distance between two cells i and j is determined by the maximum absolute correlation across all considered lags, as $1 - |r^*_{i,j}|$. Fig. 2-J,K shows the resulting clusters for k=5 (the number of synthetic domains) and 6, respectively. For k=5, domains 1 and 3 form a single cluster because of their strong anti-correlation; the same happens with domains 4 and 5. Further, two of the five clusters (green and brown) cover just noise. The situation changes completely when we request k=6 clusters. In that case, the overlapping regions in domain 2 form a single cluster, while domains 1 and 3 are separated in different clusters. Another clustering algorithm, resulting in spatially contiguous clusters [17], is illustrated in §5 in the context of climate data analysis (see Fig. 4-D).

Community detection results. We apply a state-of-the-art overlapping community detection method, referred to as OSLOM [24], with the default parameter values. The input to OSLOM is a positively weighted graph: each vertex is a grid cell and an edge between vertices i and j corresponds to the maximum absolute cross-correlation $|r^*_{i,j}|$ across all lags of interest. Absolute correlations less than 30% are considered insignificant and the corresponding edges are pruned.[3] As most community detection methods, OSLOM does not distinguish between positive and negative correlations. OSLOM provides a hierarchy of communities. When applied to our synthetic data, the first level of hierarchy (not shown) simply groups together domains 1,2,3 in one community (even though domain 2 is uncorrelated with domains 1 and 3), and domains 4,5 in another community. The connection between domains 3 and 5 is missed. The second level of hierarchy is shown in Fig. 2-L. Overall, OSLOM does a better job than PCA/ICA/clustering in detecting the spatial extent of each domain. A small overlap between domains (1,2) and (2,3) is discovered but to a smaller extent than δ-MAPS. However, a community in OSLOM is not constrained to be spatially contiguous. This is the reason we see some black dots in regions 4 and 5; these are non-contiguous overlaps between the communities that correspond to these two domains.

5. APPLICATION IN CLIMATE SCIENCE

We first apply δ-MAPS in the context of climate science. Climate scientists are interested in *teleconnections* between different regions, and they often rely on EOF analysis to uncover them [43]. Here, we analyze the monthly *Sea-Surface Temperature* (SST) field from the HadISST dataset [30], covering 50 years (1956-2005) at a spatial resolution

[3]We have experimented with other prunning thresholds between 20%-50% and the results are very similar at the first two hierarchy levels.

of $2.0^o \times 2.5^o$, and we focus on the latitudinal range of $[60^oS; 60^oN]$ to avoid sea-ice covered regions. Following standard practice, we pre-process the time series to form *anomalies*, i.e., remove the seasonal cycle, remove any long-term trend at each grid-point (using the Theil-Sen estimator), and transform the signal to zero-mean at each grid point.

δ-MAPS is applied as follows. We set the local neighborhood to the K=4 nearest cells so that we can identify the smallest possible domains at the given spatial resolution. Second, the homogeneity threshold δ is set to 0.37 (corresponds to a significance level of 10^{-2}). In the edge inference stage, the lag range is τ_{max}=12 months (a reasonable value for large-scale changes in atmospheric wave patterns), and the FDR threshold is set to q=3% (we identify about 30 edges and so we expect no more than one false positive).

Fig. 3-A shows the identified domains (the color code will be explained shortly). The spatial dimensionality has been reduced from about 6000 grid cells to 18 domains. 65% of the sea-covered cells belong to at least one domain; the overlapping regions are shown in black and they cover 2% of the grid cells that belong to a domain. The largest domain (domain E) corresponds to the El Niño Southern Oscillation (ENSO), which is also the most important in terms of node strength (see Fig. 3-B). Other strong nodes are domain F (part of the "horseshoe-pattern" surrounding ENSO), domain J (Indian ocean) and domain Q (sub-tropical Atlantic). The strength of the edges associated with ENSO are shown in Fig. 3-C. These observations are consistent with known facts in climate science regarding ENSO and its positive correlation with the Indian ocean and north tropical Atlantic, and negative correlations with the regions that surround it in the Pacific (horseshoe-pattern) [22].

Fig. 3-D shows the inferred domain-level network. The color code represents the (signed) cross-correlation for each edge. The lag range associated with each edge is shown in Fig. 3-E; recall that some edges are not directed because their lag range includes τ=0. The network consists of five weakly-connected components. If we analyze the largest component (which includes ENSO) as a signed network (i.e., some edges are positive and some negative) we see that it is *structurally balanced* [14]. A graph is structurally balanced if it does not contain cycles with an odd number of negative edges.[4] A structurally balanced network can be partitioned in a "dipole", so that positive edges only appear within each pole and negative edges appear only between the two poles. In Fig. 3-A, the nodes of these two poles are colored as blue and green (the smaller disconnected components are shown in other colors).

Focusing on the lag range of each edge, domain Q seems to play a unique role, as it temporally precedes all other domains in the inferred network. Specifically, its activity precedes that of domains D, E and F by about 5-10 months. The lead of south tropical Atlantic SSTs (domain Q) on ENSO has recently received significant attention in climate science [31]. Our results suggest that SST anomalies in domain Q may impact a large portion of the climate system.

Switching to lag inference, we say that a triangle is *lag-consistent* if there is at least one value in the lag range associated with each edge that would place the three nodes in a consistent temporal distance with respect to each other. For instance, in the case of the first triangle of Fig. 3-F, the triangle is lag-consistent if the edge from Q to F has a lag of 8 months and the edge between E and F has lag -2 months (meaning that the direction would be from F to E); several other values would make this triangle lag-consistent. We have verified the lag-consistency of every triangle in the climate network. One exception is the triangle between domains (C, D, G), shown at the bottom of Fig. 3-F. However, the large lag in the edge from C to G can be explained with the triangle between domains (C, E, G), which is lag-consistent. We emphasize that the temporal ordering that results from these lag relations should not be mis-interpreted as causality; we expect that several of the edges we identify are only due to indirect correlations, not associated with a causal interaction between the corresponding two nodes.

For comparison purposes, Fig. 4 shows the results of EOF analysis, community detection, and spatial clustering on the same dataset. The first EOF explains only about 19% of the variance, implying that the SST field is too complex to be understood with only one spatial component. On the other hand, the joint interpretation of multiple EOF components is problematic due to their orthogonal relation [12]. The anti-correlation between ENSO and the horseshoe-pattern regions is well captured in the first component but several other important connections, such as the negative and lagged relation between the south subtropical Atlantic and ENSO (domains Q and E, respectively), are missed.

Fig. 4-C shows the results of the overlapping community detection method OSLOM. Following [36], the input to OSLOM is a correlation-based cell-level network. Correlations less than 30% are ignored. The weight of each edge is set to the maximum absolute correlation between the corresponding two cells, across all considered lags. OSLOM identifies 22 communities. Community 6 is not spatially contiguous; it covers ENSO, the Indian ocean, a region in the north tropical Atlantic, and a region in south Pacific. This is a general problem with community detection methods: they cannot distinguish high correlations due to a remote connection from correlations due to spatial proximity. In the context of climate, the former may be due to atmospheric waves or large-scale currents while the latter may be due to local circulations.

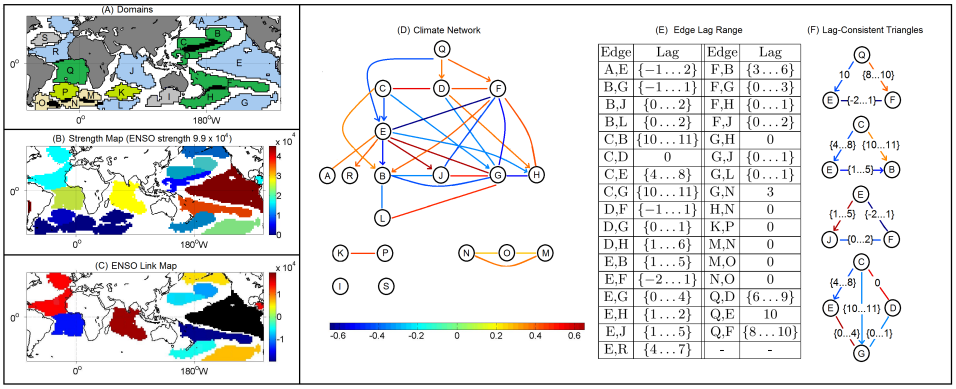
Finally, Fig. 4-D shows the results of a spatial clustering method [17], with the same homogeneity threshold δ we use in δ-MAPS. That method ensures that every cluster (referred to as "area") is spatially contiguous but it also requires that there is no overlap between areas and it attempts to assign each grid cell to an area. Consequently, it results in more areas (compared to the number of domains), some of which are just artifacts of the spatial parcellation process. Further, the spatial expanse of an area constrains the computation of subsequent areas because no overlaps are allowed.

6. APPLICATION IN FMRI DATA

Functional magnetic resonance imaging (fMRI) measures fluctuations of the blood oxygenation level dependent (BOLD) signal in the brain. The dynamics of the BOLD signal in gray matter are generally correlated with the level of neural activity. The resulting spatio-temporal field is often analyzed using ICA, clustering or network-based methods to infer *brain functional networks* [34].

Here, we illustrate δ-MAPS on cortical *resting-state* fMRI

[4]For instance, if two friends are both enemies with a third person, they form a balanced social triangle.



Edge	Lag	Edge	Lag
A,E	$\{-1\dots2\}$	F,B	$\{3\dots6\}$
B,G	$\{-1\dots1\}$	F,G	$\{0\dots3\}$
B,J	$\{0\dots2\}$	F,H	$\{0\dots1\}$
B,L	$\{0\dots2\}$	F,J	$\{0\dots2\}$
C,B	$\{10\dots11\}$	G,H	0
C,D	0	G,J	$\{0\dots1\}$
C,E	$\{4\dots8\}$	G,L	$\{0\dots1\}$
C,G	$\{10\dots11\}$	G,N	3
D,F	$\{-1\dots1\}$	H,N	0
D,G	$\{0\dots1\}$	K,P	0
D,H	$\{1\dots6\}$	M,N	0
E,B	$\{1\dots5\}$	M,O	0
E,F	$\{-2\dots1\}$	N,O	0
E,G	$\{0\dots4\}$	Q,D	$\{6\dots9\}$
E,H	$\{1\dots2\}$	Q,E	10
E,J	$\{1\dots5\}$	Q,F	$\{8\dots10\}$
E,R	$\{4\dots7\}$	-	-

Figure 3: (A) The identified domains. The color of each domain corresponds to the connected component it belongs to (the blue and green nodes belong to two different poles of the same component). (B) Color map for domain strength. The strength of ENSO (domain E) is shown at the top. (C) Edges to and from ENSO (shown in black). (D) The climate network. The color of each edge represents the corresponding cross-correlation. (E) The lag range associated with each edge. (F) Examples of lag-constistent triangles.

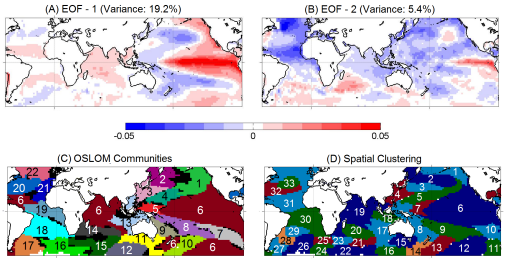


Figure 4: (A),(B) The first two components of EOF analysis. (C) Communities identified by OSLOM. Each community has a unique number and color. (D) Areas identified by spatial clustering.

data from a single subject (healthy young male adult, subject-ID: 122620) from the WU-Minn Human Connectome Project (HCP) [42]. The data acquisition parameters are described in [33]. The spatial resolution is 2mm in each voxel dimension. The pre-processing of fMRI data requires several steps; we use the "fix-extended" HCP minimal processing pipeline that includes head motion correction, registration to a structural image, masking on non-brain voxels, etc; please see [18]. MELODIC ICA and FIX are used to remove non-neuronal artifacts (e.g., physiological noise due to cardiac and respiratory cycles). We also perform bandpass filtering in the range 0.01-0.08Hz, as commonly done in resting-state fMRI.

In this paper, we analyze two scanning runs of the same subject ("scan-1" and "scan-2"). Each scan lasts about 14 minutes and results in a timeseries of length T=1200 (repetition time TR=720msec). We emphasize that major differences across different scanning sessions of the same subject are common in fMRI; studies of functional brain networks often only report group-level averages. The entire cortical volume is projected to a surface mesh (Conte69 32K) resulting in about 65K *gray-ordinate* points (as opposed to volumetric voxels) [41]. Each point of this mesh is adjacent to six other points; for this reason we set K=6. The homogeneity threshold is set to δ=0.37 (corresponds to significance level 10^{-2}). The maximum lag range τ_{max} is set to ±3, i.e., 2.2 seconds, and the FDR threshold is set to q=10^{-4} (i.e., we expect one out of 10K edges to be a false positive). The signal of a domain is defined as the average across all voxels in that domain.

The application of δ-MAPS results in a network with about 850 domains in scan-1 (1120 domains in scan-2). 80% of the domains are smaller than 30-40 voxels (depending on the scan) and 5% of the domains are larger than 250 voxels. The number of edges is 4285 in scan-1 (4200 in scan-2). The absolute value of the cross-correlation associated with each edge is typically larger than 0.5. The fraction of negative edge correlations is about 5% in scan-1 and 20% in scan-2 suggesting that the polarity of some network edges may be time-varying. The lag τ^* that corresponds to the maximum cross-correlation is 0 in 70% of the edges and ±1 in almost all other cases. 13% of the edges are directed, meaning that lag-0 does not produce a significant correlation for that pair of domains. There is a positive correlation between the degree of a domain and its physical size (the correlation coefficient between degree and $\log_{10}$(size) is 0.70 for scan-1 and 0.66 for scan-2). Further, the network is assortative meaning that domains tend to connect to other domains of similar degree (assortativity coefficient about 0.7 in both scans).

An important question is whether the δ-MAPS networks are consistent with what neuroscientists currently know about resting-state activity in the brain. During rest, certain cortical regions that are collectively referred to as the *Default-Mode Network (or DMN)* are persistently active across age and gender [44]. Other known resting-state networks are the occipital (part of the visual system) and the motor/somatosensory (associated with planning and execution of voluntary body motion). With the terminology of network theory, the previous "networks" would be referred to as *communities* within the larger functional brain network. To identify communities in the δ-MAPS network, we applied OSLOM [24]. OSLOM identifies two hierarchical levels in both scans. The first level consists of highly overlapping communities that cover almost the entire cortex. The second hierarchical level is more interesting, resulting in eight communities for scan-1 (nine for scan-2). Fig. 5 shows the three communities (C.1, C.2, C.3) for each scan that have the highest resemblance to the three previously mentioned resting-state networks: C.1 corresponds to the DMN, C.2 corresponds to the occipital resting-state network, and C.3 corresponds to the motor/somatosensory network. C.1 is quite similar across the

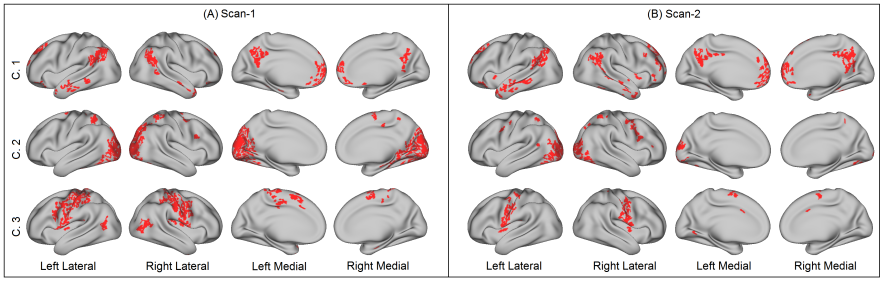


Figure 5: Three domain-level network communities for each scan. The first corresponds to the default-mode network, the second to the occipital network, and the third to the motor/somatosensory network.

two scanning sessions and it clearly captures the DMN. In C.2, the extent of the network is smaller in scan-2, which is not too surprising giving the known inter-scan variability of resting-state fMRI. C.3 is also quite similar across the two scans and consistent with the motor/somatosensory network.

To further investigate the structure of those higher degree (and typically larger) domains, we perform *k-core decomposition*.[5] The density of the remaining network, after the extraction of k=14 cores from the scan-1 network (k=16 cores in scan-2) shows a sudden increase by a factor of two. This suggests that the network includes a *densely inter-connected backbone*, also known as "rich-club". The size of this backbone is small relative to the entire network: 130 domains in scan-1 (90 in scan-2). Similar observations about the resting-state brain, but using voxel-level network analysis methods, have been previously reported [40]. Fig.6 shows the location of the backbone domains for each hemisphere and for each scan. The regions that are usually associated with the DMN dominate the backbone of both sessions. Interestingly though, scan-1 includes the regions of the motor/somatosensory network, while the backbone of scan-2 is missing those regions. One possible explanation for this discrepancy is that the subject was more relaxed during scan-2, not exerting the mental effort to stay still.

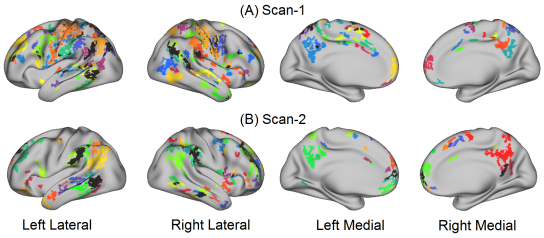


Figure 6: The domains of the backbone network for each hemisphere and scan. The color of each domain is randomly assigned (overlaps are shown in black).

7. DISCUSSION

δ-MAPS results in a correlation-based functional network. A next step could be to infer a causal, or *effective* network, leveraging the framework of probabilistic graphical models. Instead of attempting to learn the graph structure from raw data, one could use the δ-MAPS network as the underlying structure and then apply conditional independence tests to remove non-causal edges (e.g., [15]). Another direction could be to combine the inferred functional network with a structural network that shows the physical connectivity between the identified domains. This is not hard in the case of communication networks but it also becomes feasible for brain networks using diffusion-weighted MRI. The projection of the observed dynamics on the underlying structure can help to characterize the actual function and delay of each system component.

[5]A process that starts with the original network (k=0), and it removes iteratively all nodes of degree k or less in each round so that after the extraction of the k'th core all remaining nodes have degree larger than k.

8. ACKNOWLEDGEMENTS

This work was made possible by a grant from the Department of Energy, Climate and Environmental Sciences Division, SciDAC: Earth System Model Development. The work of B.D. and C.D. was also supported by a Raytheon E-Systems Faculty Fellowship award. B.D. is partially supported by the NSF grant CCF-1522054 (COMPUSTNET: Expanding Horizons of Computational Sustainability).

9. REFERENCES

[1] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466(7307):761–764, 2010.

[2] S. Arnborg, J. Lagergren, and D. Seese. Easy problems for tree-decomposable graphs. *Journal of Algorithms*, 12(2):308–340, 1991.

[3] C. Baldassano, D. M. Beck, and L. Fei-Fei. Parcellating connectivity in spatial maps. *PeerJ*, 3:e784, 2015.

[4] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[5] D. Birant and A. Kut. ST-DBSCAN: An algorithm for clustering spatial–temporal data. *Data & Knowledge Engineering*, 60(1):208–221, 2007.

[6] T. Blumensath, T. E. Behrens, and S. M. Smith. Resting-state fmri single subject cortical parcellation based on region growing. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012*, pages 188–195. Springer, 2012.

[7] T. Blumensath, S. Jbabdi, M. F. Glasser, D. C. Van Essen, K. Ugurbil, T. E. Behrens, and S. M. Smith. Spatially constrained hierarchical parcellation of the brain with resting-state fMRI. *Neuroimage*, 76:313–324, 2013.

[8] G. E. Box, G. M. Jenkins, and G. C. Reinsel. *Time series analysis: forecasting and control*, volume 734. John Wiley & Sons, 2011.

[9] X. Chen, X. Hu, and C. Wang. Finding connected dense k-subgraphs. In *Theory and Applications of Models of Computation*, pages 248–259. Springer, 2015.

[10] D. G. Corneil and Y. Perl. Clustering and domination in perfect graphs. *Discrete Applied Mathematics*, 9(1):27–39, 1984.

[11] R. C. Craddock, G. A. James, P. E. Holtzheimer, X. P. Hu, and H. S. Mayberg. A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Human brain mapping*, 33(8):1914–1928, 2012.

[12] D. Dommenget and M. Latif. A cautionary note on the interpretation of EOFs. *Journal of Climate*, 15(2):216–225, 2002.

[13] J. F. Donges, Y. Zou, N. Marwan, and J. Kurths. The backbone of the climate network. *EPL (Europhysics Letters)*, 87(4):48007, 2009.

[14] D. Easley and J. Kleinberg. *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press, 2010.

[15] I. Ebert-Uphoff and Y. Deng. Causal discovery from spatio-temporal data with applications to climate science. In *Machine Learning and Applications (ICMLA), 2014 13th International Conference on*, pages 606–613. IEEE, 2014.

[16] J. H. Faghmous and V. Kumar. Spatio-temporal data mining for climate data: Advances, challenges, and opportunities. In *Data Mining and Knowledge Discovery for Big Data*, pages 83–116. Springer, 2014.

[17] I. Fountalis, A. Bracco, and C. Dovrolis. Spatio-temporal network analysis for studying climate patterns. *Climate dynamics*, 42(3-4):879–899, 2014.

[18] M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson, T. S. Coalson, B. Fischl, J. L. Andersson, J. Xu, S. Jbabdi, M. Webster, J. R. Polimeni, et al. The minimal preprocessing pipelines for the Human Connectome Project. *Neuroimage*, 80:105–124, 2013.

[19] M. Hinne, M. Ekman, R. J. Janssen, T. Heskes, and M. A. van Gerven. Probabilistic clustering of the human connectome identifies communities and hubs. *PloS one*, 10(1):e0117179, 2015.

[20] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *Neural Networks, IEEE Transactions on*, 10(3):626–634, 1999.

[21] J. M. Keil and T. B. Brecht. The complexity of clustering in planar graphs. *J. Combinatorial Mathematics and Combinatorial Computing*, 9:155–159, 1991.

[22] S. A. Klein, B. J. Soden, and N.-C. Lau. Remote sea surface temperature variations during ENSO: Evidence for a tropical atmospheric bridge. *Journal of Climate*, 12(4):917–932, 1999.

[23] M. A. Kramer, U. T. Eden, S. S. Cash, and E. D. Kolaczyk. Network inference with confidence from multivariate time series. *Physical Review E*, 79(6):061916, 2009.

[24] A. Lancichinetti, F. Radicchi, J. J. Ramasco, S. Fortunato, et al. Finding statistically significant communities in networks. *PloS one*, 6(4):e18961, 2011.

[25] Y. Lu, T. Jiang, and Y. Zang. Region growing method for the analysis of functional MRI data. *NeuroImage*, 20(1):455–465, 2003.

[26] E. Martin and J. Davidsen. Estimating time delays for constructing dynamical networks. *Nonlinear Processes in Geophysics*, 21(5):929–937, 2014.

[27] M. P. McGuire and N. P. Nguyen. Community structure analysis in big climate data. In *Big Data (Big Data), 2014 IEEE International Conference on*, pages 38–46. IEEE, 2014.

[28] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.

[29] J. D. Power, A. L. Cohen, S. M. Nelson, G. S. Wig, K. A. Barnes, J. A. Church, A. C. Vogel, T. O. Laumann, F. M. Miezin, B. L. Schlaggar, et al. Functional network organization of the human brain. *Neuron*, 72(4):665–678, 2011.

[30] N. Rayner, D. E. Parker, E. Horton, C. Folland, L. Alexander, D. Rowell, E. Kent, and A. Kaplan. Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *Journal of Geophysical Research: Atmospheres (1984–2012)*, 108(D14), 2003.

[31] B. Rodríguez-Fonseca, I. Polo, J. García-Serrano, T. Losada, E. Mohino, C. R. Mechoso, and F. Kucharski. Are Atlantic Niños enhancing Pacific ENSO events in recent decades? *Geophysical Research Letters*, 36(20), 2009.

[32] C. Rummel, M. Müller, G. Baier, F. Amor, and K. Schindler. Analyzing spatio-temporal patterns of genuine cross-correlations. *Journal of neuroscience methods*, 191(1):94–100, 2010.

[33] S. M. Smith, C. F. Beckmann, J. Andersson, E. J. Auerbach, J. Bijsterbosch, G. Douaud, E. Duff, D. A. Feinberg, L. Griffanti, M. P. Harms, et al. Resting-state fMRI in the human connectome project. *Neuroimage*, 80:144–168, 2013.

[34] O. Sporns. *Networks of the Brain*. MIT press, 2011.

[35] M. Steinbach, P.-N. Tan, V. Kumar, S. Klooster, and C. Potter. Discovery of climate indices using clustering. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 446–455. ACM, 2003.

[36] K. Steinhaeuser, N. V. Chawla, and A. R. Ganguly. An exploration of climate data using complex networks. *ACM SIGKDD Explorations Newsletter*, 12(1):25–32, 2010.

[37] K. Steinhaeuser, N. V. Chawla, and A. R. Ganguly. Complex networks as a unified framework for descriptive analysis and predictive modeling in climate science. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 4(5):497–511, 2011.

[38] B. Thirion, G. Varoquaux, E. Dohmatob, and J.-B. Poline. Which fMRI clustering gives good brain parcellations? *Frontiers in neuroscience*, 8, 2014.

[39] M. Van Den Heuvel, R. Mandl, and H. Hulshoff Pol. Normalized cut group clustering of resting-state fMRI data. *PloS one*, 3(4):e2001, 2008.

[40] M. P. van den Heuvel and O. Sporns. Rich-club organization of the human connectome. *The Journal of neuroscience*, 31(44):15775–15786, 2011.

[41] D. C. Van Essen, M. F. Glasser, D. L. Dierker, J. Harwell, and T. Coalson. Parcellations and hemispheric asymmetries of human cerebral cortex analyzed on surface-based atlases. *Cerebral Cortex*, 22(10):2241–2262, 2012.

[42] D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, K. Ugurbil, W.-M. H. Consortium, et al. The WU-Minn human connectome project:An overview. *Neuroimage*, 80:62–79, 2013.

[43] H. Von Storch and F. W. Zwiers. *Statistical analysis in climate research*. Cambridge university press, 2001.

[44] B. T. Yeo, F. M. Krienen, J. Sepulcre, M. R. Sabuncu, D. Lashkari, M. Hollinshead, J. L. Roffman, J. W. Smoller, L. Zöllei, J. R. Polimeni, et al. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology*, 106(3):1125–1165, 2011.

[45] P. Zhang, Y. Huang, S. Shekhar, and V. Kumar. Correlation analysis of spatial time series datasets: A filter-and-refine approach. In *Advances in Knowledge Discovery and Data Mining*, pages 532–544. Springer, 2003.

Appendix I: Identifying the largest domain is NP-complete

We are given a spatio-temporal field $\mathbf{X}(t)$ on a grid G, a pairwise similarity metric between pairs of grid cells and a threshold δ. Starting from a grid cell c, the goal is to find the largest subset of grid cells that form a single spatially connected component, and whose average similarity exceeds the threshold δ. The spatial grid can be represented as a planar graph $G(V, E)$ where each grid cell is a node and

edges connect adjacent grid cells. Formally we have the following graph optimization problem:

Definition 1. Rooted Largest Connected δ-Dense Subgraph Problem (rooted LCδDS). Given a regular (grid) graph $G(V, E)$, a weight function $w : V \times V \rightarrow \mathbb{R}$ (where $w(v, v) = 0$ and symmetric), a threshold δ, and a node $c \in V$, find a maximum cardinality set of nodes $A \subseteq V$ such that $c \in A$, the induced subgraph is connected ($I_G(A) = 1$) and $\frac{\sum_{v,u \in A} w(v,u)}{|A|(|A|-1)} > \delta$ (i.e., $\hat{r}(A) > \delta$).

To show that rooted LCδDS is NP-hard we first consider a variant of the problem in which the induced subgraph A has to satisfy two conditions; it has to be a connected subgraph of G, and the average weight of the edges in A has to exceed δ. More formally:

Definition 2. Largest Connected δ-Dense Subgraph Problem (LCδDS). Given a regular (grid) graph $G(V, E)$, a weight function $w : V \times V \rightarrow \mathbb{R}$ (where $w(v, v) = 0$ and symmetric), and a threshold δ, find a maximum cardinality set of nodes $A \subseteq V$ such that $I_G(A) = 1$ and $\hat{r}(A) > \delta$.

To show that LCδDS is NP-hard we use a reduction of the densest connected k subgraph problem.

Definition 3. Densest Connected k-Subgraph Problem (DCkS). Decision version: Given a graph $G(V, E)$, and positive integers k and j, does there exist an induced subgraph on k vertices such that this subgraph has at least j edges and is connected?

DCkS (also referred to as the connected h-clustering problem) has been shown to be NP-complete on general graphs [10], as well as on planar graphs [21]. DCkS is polynomially time solvable for subclasses of planar graphs of bounded tree width [2]. Grid graphs, which are the type of graphs that arise in our application domains, are planar bipartite graphs, with non-fixed tree width, and no positive results are known for this subclass of planar graphs. The work on approximating densest/heaviest connected k-subgraphs is relatively very limited (see recent theoretical result [9]). It is easy to show that the DCkS problem can be easily reduced to an instance of the decision version of the LCδDS problem, and hence it is also NP-complete even on planar graphs.

LEMMA 1. *The decision version of the LCδDS problem is NP-complete on planar graphs.*

PROOF. This can be shown via a reduction from the DCkS. We reduce an instance $< G, k, j >$ of the DCkS to an LCδDS instance by using the same graph G, setting $w(u, v) = I(u, v) \in E$ ($w(u, v)$ is 1 if and only if the pair of nodes is connected by an edge), and $\delta = j/k(k-1)$. □

Now it is easy to show that rooted LCδDS is also NP-hard. If a poly-time algorithm existed for the rooted LCδDS, then by calling it $|V|$ times with each of the nodes of the graph, we would obtain in poly-time a solution to the NP-hard LCδDS.

Appendix II: Heuristic for the selection of δ

The threshold δ intuitively determines the minimum degree of homogeneity that the underlying field must have within each domain. The higher the threshold, the higher the required homogeneity and therefore, the smaller the size of the identified domains.

To select δ we propose the following heuristic. We start with a random sample of pairs of grid cells and for each pair i, j we compute the Pearson correlation $r_{i,j}$ at zero lag. To assess the significance of each correlation we use Bartlett's formula [8]. Under the null hypothesis of no coupling $r_{i,j}$ should have zero mean, and a reasonable estimate of its variance is given by

$$Var[r_{i,j}] = \frac{1}{T} \sum_{\tau_k=-T}^{T} r_{i,i}(\tau_k) r_{j,j}(\tau_k) \ , \qquad (11)$$

here $r_{i,i}(\tau_k)$ is the autocorrelation of the time series of grid cell i at lag τ_k. The scaled values $z_{i,j} = \frac{r_{i,j}}{\sqrt{Var[r_{i,j}]}}$ should approximately follow a standard normal distribution. To assess the significance of each correlation we perform a one sided z-test for a given level of significance α.

The threshold δ is set as the average of all significant correlations. A domain is a set of spatially contiguous grid cells, thus we require that the mean pairwise correlation for the cells belonging to the same domain to be higher than the mean pair-wise correlation of randomly picked pairs of grid cells. δ depends on the choice of the significance level α, on the autocorrelation structure of the underlying time series and on the correlation distribution of the field.

Appendix III: δ-MAPS pseudocode

```
1: Domains S = {A_1, ..., A_|S|}          ▷ The initial set of domains
2: function DomainIdentification()
3:     while True do
4:         boolean merged ← DomainMerging(S)
5:         boolean expanded ← DomainExpansion(S)
6:         if !merged&&!expanded then
7:             break                      ▷ Terminate when no further expansion or merging is possible
8:         end if
9:     end while
10: end function
```

```
function DomainExpansion(Domains S = {A_1, ..., A_|S|})
    boolean startMerging ← false
    boolean expanded ← false
    while !startMerging do ▷ Domain expansion is repeated in rounds
        expanded ← false
        sort(S) ▷ Sort domains in decreasing order of homogeneity such that r̂(A_{i-1}) > r̂(A_i) > r̂(A_{i+1})
        for i = 1 : |S| do
            Domain A_i ← S[i]
            Domain eA_i ← ExpandDomain(A_i)
            if |A_i| ≠ |eA_i| then ▷ Domain expanded
                S[i] ← eA_i
                expanded ← true
                startMerging ← CanMerge(eA_i)
                if startMerging then
                    break ▷ Exit the for loop
                end if
            end if
        end for ▷ A round of domain expansion is complete
        if !expanded then
            break ▷ Domains cannot be expanded
        end if
    end while
    return expanded
end function

function ExpandDomain(Domain A_i) ▷ Try do expand domain A_i by one cell
    Construct set Γ(A_i): all cells adjacent to A_i
    if Γ(A_i) = ∅ then
        return A_i
    else
        m ← arg max_{m∈Γ(A_i)} r̂(A_i ∪ {m}) ▷ Select the cell that maximizes r̂(A_i ∪ {m}).
        if r̂(A_i ∪ {m}) > δ then
            A_i ← A_i ∪ m
        end if
        return A_i
    end if
end function

function CanMerge(Domain A_i) ▷ Check whether one or more merging operations are possible
    boolean merge ← false
    Construct set Γ(A_i): all domains adjacent to A_i
    for j = 1 : |Γ(A_i)| do
        A_j ← Γ(A_i)[j]
        if r̂(A_i ∪ A_j) > δ then
            merge ← true
            break
        end if
    end for
    return merge
end function
```

```
function DomainMerging(Domains S = {A_1, ..., A_|S|})
    boolean merged ← false
    while True do ▷ Repeat until no pair of domains can be merged
        Domain DomainToMerge1 ← ∅
        Domain DomainToMerge2 ← ∅ ▷ Domains with the maximum union homogeneity
        maxHomogeneity ← −1
        for i = 1 : |S| do
            Domain A_i ← S[i] ▷ Get the i^th domain
            Construct set Γ(A_i)
            A_j ← arg max_{A_j∈Γ(A_i)} r̂(A_i ∪ A_j)
            if r̂(A_i ∪ A_j) > maxHomogeneity then ▷ Update the best candidates to merge
                DomainToMerge1 ← A_i
                DomainToMerge2 ← A_j
                maxHomogeneity ← r̂(A_i ∪ A_j)
            end if
        end for
        if maxHomogeneity > δ then
            S.remove(DomainToMerge1)
            S.remove(DomainToMerge2) ▷ Remove the domains that will be merged
            S ← DomainToMerge1∪DomainToMerge2
            merged ← true
        else
            break ▷ We can not merge any domains
        end if
    end while
    return merged ▷ Return true if at least one pair of domains is merged
end function
```