Manuscript prepared for Nonlin. Processes Geophys.
with version 5.0 of the LaTeX class copernicus.cls.
Date: 5 February 2014

On the influence of spatial sampling on climate networks

Nora Molkenthin[1,2], **Kira Rehfeld**[1,3], **Veronika Stolbova**[1,2], **Liubov Tupikina**[1,2], **and Jürgen Kurths**[1,2]

[1]PIK Potsdam Institute of Climate Impact Research, P.O.Box 601203, 14412 Potsdam, Germany
[2]Department of Physics, Humboldt-Universität zu Berlin, Newtonstr. 15, 12489 Berlin, Germany
[3]Alfred-Wegner Institute for Polar and Marine Research, Telegrafenberg A43, 14473 Potsdam, Germany

Correspondence to: Nora Molkenthin
(Molkenthin@pik-potsdam.de)

Abstract. Climate networks are constructed from climate time series data using correlation measures. It is widely accepted that the geographical proximity as well as other geographical features such as ocean and atmospheric currents have a large impact on the observable time-series similarity. Therefore it is to be expected that the spatial sampling will influence the reconstructed network. Here we investigate this by comparing analytical flow networks, networks generated with the START model and networks from temperature data from the Asian monsoon domain. We evaluate them on a regular grid, a grid with added random jittering and two variations of clustered sampling. We find that the impact of the spatial sampling on most network measures only distorts the plots if the node distribution is significantly inhomogeneous. As a simple diagnostic measure for the detection of inhomogeneous sampling we suggest the Voronoi cell size distribution.

1 Introduction

Complex networks are used in many areas of physics to describe systems as different as the brain, the internet and social interactions (Boccaletti et al., 2006; Barthélemy, 2011). While the first applications were discrete by nature, complex networks have recently also found use in continuous systems like climate (Tsonis et al., 2006; Donges et al., 2009). In those applications, the system's spatial description is reduced to a number of discrete points, between which correlations are evaluated to construct a network (Malik et al., 2011; Rehfeld et al., 2012). This reveals a variety of interesting structures, many of which can be explained with geographical features. Spatial proximity often leads to large correlations, but mountains for example, can act as a barrier and prohibit exchange. Regions with a dominant wind or ocean flow can exhibit longer links. With all this an effect of the sampling locations has to be expected.

Aliasing effects can occur, if the sampling steps are larger than the half-width of investigated phenomena and can distort the actual signal substantially (Unser, 2000; Dippé and Wold, 1985). Spatial sampling effects have been discussed (Heitzig et al., 2012), where they proposed weighted measures to deal with differently sized nodes. In climate networks the related issue of temporal sampling and irregularity are treated in (Rehfeld et al., 2011) and the role of boundary effects was investigated in (Rheinwalt et al., 2012). While previous studies have mostly used grids for the locations of the measurements (Yamasaki et al., 2008; Tsonis et al., 2010), in this paper we study the effects of the spatial sampling itself on two typical model example systems and one set of observational data. In contrast to previous studies, we look at the sampling effects compared to the underlying system, rather than just differently sized regions. The first example is a model, that generates networks directly from flows, introduced in (Molkenthin et al., 2013). The second example is the START model, introduced in (Rehfeld et al., 2013), in which artificial autocorrelated time series are transported by a flow field. In these models the node locations can be chosen freely, making them ideal for studying the effects of spatial sampling. The third example is the analysis of temperature data from the Asian Monsoon area.

We compute networks of a regular grid, a grid with a uniformly random jitter and clustered node distributions and compare their degree, betweenness and link length distribution to see how the spatial sampling affects common network measures.

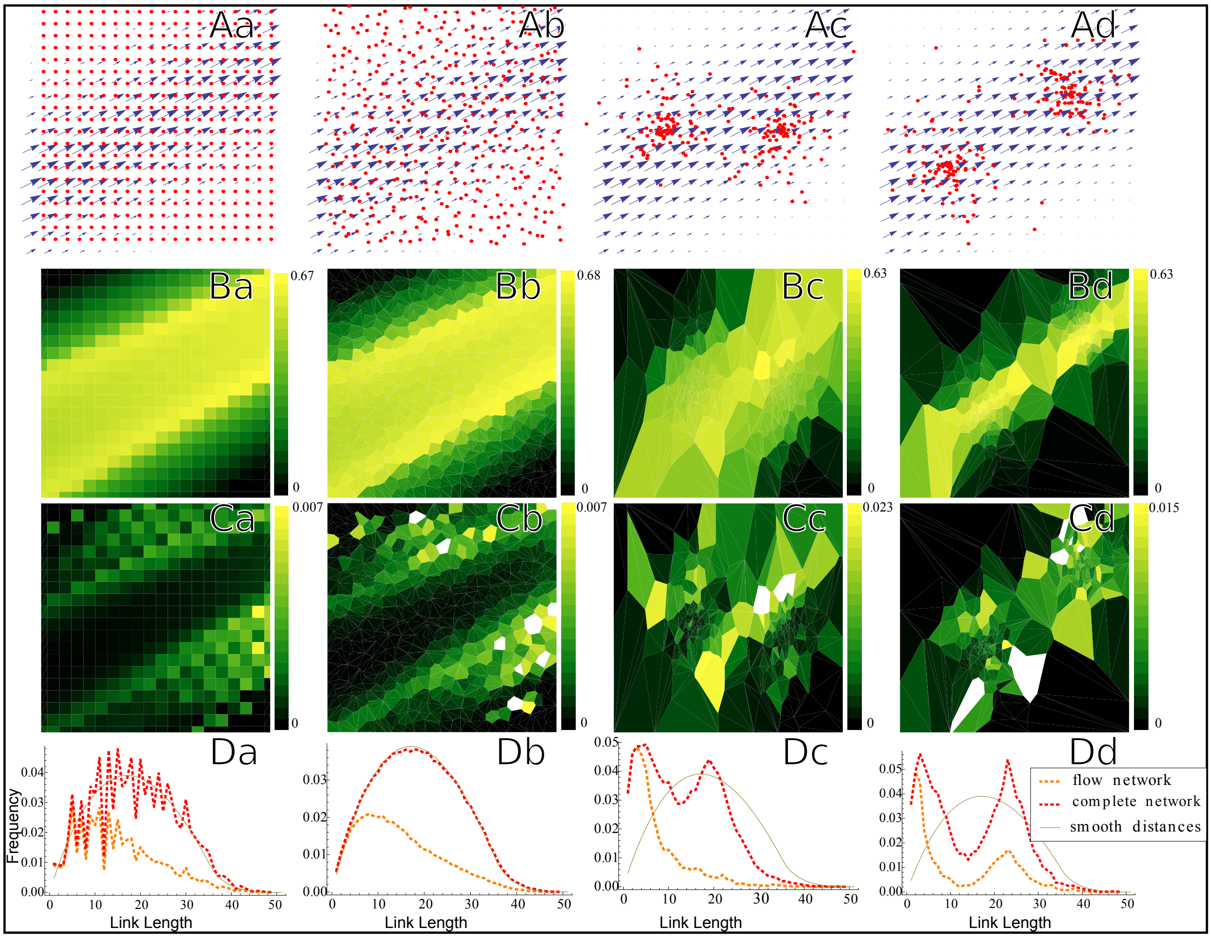

Fig. 1. A diagonal flow is sampled with Aa) a grid, Ab) a jittered grid, Ac) two clusters on opposite sides of the flow, Ad) two clusters inside the flow. The networks are constructed with a link density of 40 percent. Row B) shows the degree, row C) the betweenness and row D) the link length distribution.

2 Spatial effects in flow networks

Flow networks as introduced in (Molkenthin et al., 2013) are constructed using analytical solutions of the advection-diffusion-equation (ADE) instead of time series in the definition of the correlation. The correlation measure is based on the scalar product between the temperature profile of a single δ-peak's evolution due to advection and diffusion, evaluated at each pair of nodes. We construct these networks and analyze the network measures degree, betweenness and link length distribution of the resulting networks and compare the influence of the spatial sampling on the result. Flow networks for a diagonal flow pattern, are constructed for the node distributions shown in row A of Fig.1, on a regular 20×20 grid, a jittered 20×20 grid and for two versions of Gaussian clustering. The jittered sampling pattern was generated by uniformly drawing a point from each grid cell. For the clustering, a Gaussian distribution around two central points on opposite sides of the flow (Ac) and along the flow (Ad) with 200 nodes. We plotted the node distribution in the flow, node degrees, node betweenness and the link length distribution for each of the sampling patterns.

Degree: We observe that the degree (shown in row B in Fig. 1) is higher in the middle stripe, where the flow velocity is highest. The degree is not very sensitive to the spatial sampling. In all four plots the degree is clearly highest where the absolute velocity is highest. Clustered sampling can lead to a skewed shape of the stripe, or underestimate its width if the outer region is poorly sampled.

Betweenness: The betweenness is shown in row C in Fig. 1. Grid and grid plus jitter show similar patterns: the betweenness is highest in the transition zone between the fast and slow flowing areas. This structure is barely visible in the clustered sampling plots. In the clusters on opposite sides, two stripes are visible, but the lower one of them is located in the center of the flow, rather than at its side, in the middle between the two clusters.

This indicates that rather than showing a transition zone of the flow, it emphasizes a region, that is poorly sampled but central in the flow. In the plot of the two clusters in the flow, the outer regions are poorly sampled, therefore the transition is not visible.

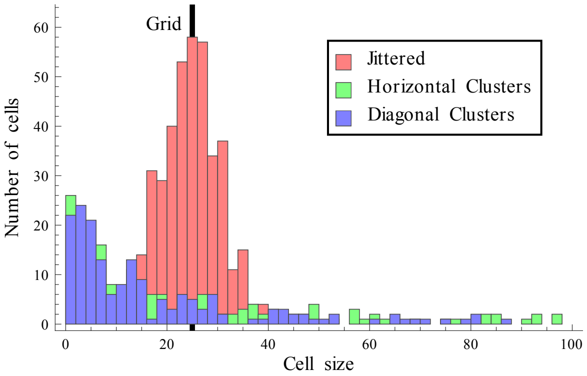

Fig. 2. Tile size distribution for jittered and clustered sampling. The jittered sampling is peaked around the grid cell size, clustered sampling results in many small and few very large tiles. Note that grid and jittered grid have twice as many nodes as the clustered sampling.

Voronoi tesselation analysis: A Voronoi tesselation assigns a cell to each node, such that every point in the cell is closer to that node than to any other. If the nodes are uniformly distributed in space the resulting cell sizes will have a clear peak around A_{sam}/N_{nod}, where A_{sam} is the size of the total sampled area and N_{nod} is the number of nodes. The differences in the sampling can be quantified using the area size distributions of their Voronoi tesselations (Barthélemy, 2011) as shown in Fig. 2.

In the grid all tiles are exactly the same size, so the histogram would be one very sharp peak. In the case of grid with jitter the peak is broadened. The two clustered sampling peaks around much smaller values.

Link length distribution: The link length distribution, shown in row D in Fig. 1 is spiky for the grid and considerably smoother for the jittered node distribution. In clustered sampling, the link length distributions show two peaks. The beige line shows the distance distribution of pairs of points in a continuous square for comparison.

The regular arrangement of points in the grid leads to an overrepresentation of some specific linklength i.e. multiples of the grid constant, while other node-node distances are excluded by the node distribution. The distance distribution of pairs of points in a continuous square is the continuous analogue to the link length distribution of the fully connected graph of nodes in the discretized network description. Since climate is a continuous phenomenon, an appropriate discretization should well approximate that distance distribution. The distance distribution is obtained by integrating the distances between all pairs of points x and y, which is equivalent to integrating over the area of the square and all circle segments, that lie in the square of circles of radius d. Normalizing the resulting function of d one gets the distance distribution. It coincides well with the link length distribution of the fully connected nodes of the jittered sampling. The fully connected grids link length distribution is more spiky but varies around the same curve, while the clustered node sampling leads to a biased link length distribution with one maximum for links within each cluster and another maximum of links connecting the clusters.

Degree distribution: Furthermore we find (results not shown) that the spatial sampling has little impact on the degree distribution. For all sampling types the degree distribution was constant with one peak, the position of which depended on the value of the threshold.

Link number with threshold: The total number of links is a decreasing function of the threshold (results not shown). For both, grid and grid with jitter, this function is decreasing smoothly rather than in steps.

3 Spatial sampling in networks from the START-model

Next we consider gridded, jittered and clustered networks using the START (Stream Transported Auto Regressive Temperature) model described in (Rehfeld et al., 2013).

It is used to simulate time series at each node and construct the network by computing the correlations for each pair. The 20 percent strongest correlations are used to set the links in the network. The START-model assumes three sources of random information, that is transported by three independent flows. The ADE is used to compute variance factors $f_X(\mathbf{p}, F)$, that approximate the influence of source X on position $\mathbf{p}$ at forcing F. At each point, the signal is computed as the sum of the contributions from the three sources, scaled with the corresponding variance factor. A local noise contribution is also added:

$$R_i = f_X(i,F)R_X + f_Y(i,F)R_Y + f_Z(i,F)R_Z + R_{noise}. \tag{1}$$

Here we use a forcing where only one of the flows is active for better comparison to the diagonal flow in the flow networks.

Despite also being constructed according to advection and diffusion, START is significantly different from the flow networks. Firstly, instead of averaging over all possible initial peak positions, there is a particular source location. This results in a decay of signal strength with distance from the source.

Fig. 5 shows the node degrees for the resulting networks, which drop off with distance to the source, unlike the flow networks, where the degree only depended on the velocity. We can see that, while the jittered node distribution shows the same structure as the grid, using clustered node locations leads to a distortion of the pattern.

Additionally we can see that due to the addition of random fluctuations at each point, the degree in the stagnant and far away regions go to zero.

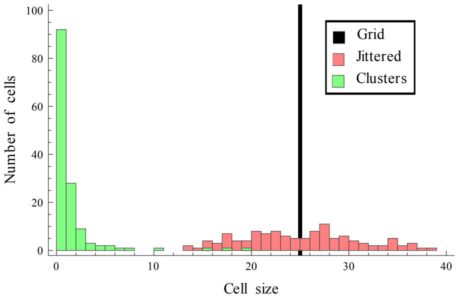

Fig. 3. Tile size distribution for the three node distributions in START.

The Link length distribution is similar to the link length distributions for the flow networks, see Fig. 4.

The betweenness (results not shown) shows the same tongue like structure as the degree. This can be explained with the random fluctuations reducing the dynamic areas to only one, which is the region inside the flow, while everything else is essentially unconnected.

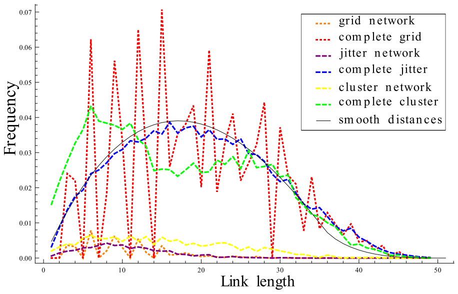

Fig. 4. Link length distribution for the three node distributions in START.

4 Spatial sampling effects in data from the Asian monsoon domain

The data analysis was done on NCEP/NCAR (Kalnay and Kanamitsu, 1996; NCEP/NCAR) reanalysis daily temperature anomalies from January to December 2011 with a resolution of 2.5×2.5 degrees, which results in a grid constant $a = 280$ km. The network was constructed from the anomaly time series using Pearson correlation. We threshold the correlation matrix using a link density of the 5 percent strongest correlations and obtain degree and betweenness as shown in Fig. 7. The geographical coordinates of each node are marked by a point at the center of each cell. To investigate sampling effects, we added a jitter in longitude and latitude at each grid point with a uniformly random number between $\pm a/2$. The tile size distributions are shown in Fig. 6

The new time series are constructed as Gaussian weighted averages of all of the original time series on the grid. The weights depend on the euclidean distance between the grid points and the new location.

Fig. 7 shows a) the annual average of the absolute values of the surface wind speeds, b) the link length distribution. The degree is presented for grid and jittered sampling in the second row (c,d). The bottom row shows the betweenness for grid and jittered sampling (e,f).

The node degree is highest in the bottom left corner of Fig.7. This coincides with the Indian ocean surface current. We conclude that the more persistent ocean currents have a larger influence on the degree than the atmospheric currents, which are subject to more change. In contrast to the correlation between high velocities and high degrees, which was established earlier, fast average velocities in the Tibetian plateau Fig.7a), correspond to some of the lowest degrees in the Asian monsoon domain, as shown in Fig.7c and d).

The plateau is relatively secluded from the rest of the network due to its high altitude, this means that even if the link density inside the secluded area would be very high, the degree would be low, as connections to other regions are less likely.

Other regions of relatively high degree can be seen in 35-40° north, which coincides with the influence of the westerlies, and along the Chinese coastline. Comparing Fig.7 c and d, we find that the effect of spatial sampling on the degree patterns is reasonably small.

The shortest path betweenness (e,f) varies more spatially than the degree. Most nodes have medium betweenness and those who have higher or lower values do not persist when the sampling is changed. Take, for example, the betweenness at 25°N, 90°E, which in the gridded sampling is one of the lowest values but in the jittered sampling is average. It is unclear if this effect is due to aliasing or a physical effect, however, it could imply that the betweenness measure is too sensitive to draw robust conclusions on the systems dynamics in this case.

5 Discussion and Conclusions

We have analyzed the influence of the spatial distribution of nodes on the network topology for three examples: Flow networks, the START model and surface temperature correlation networks. We find that the degree is robust, while the betweenness reacts more sensitively to the spatial sampling. All these examples are related to climate, yet the results can be generalized to all networks coming from discretizations of any continuous system. For a smooth linklength distribution the node to node separation has to cover all possible distances. A grid does not fulfill that as all distances are of the form: $d(x_1, x_2) = a\sqrt{n^2 + m^2}$, where a is the grid constant and n and m are integer numbers. Distances, that can not be represented in this way, are not present in a grid.

While the effects of the area sizes themselves are discussed in detail in Heitzig et al. (2012) and can be removed using their proposed consistently weighted network measures, this

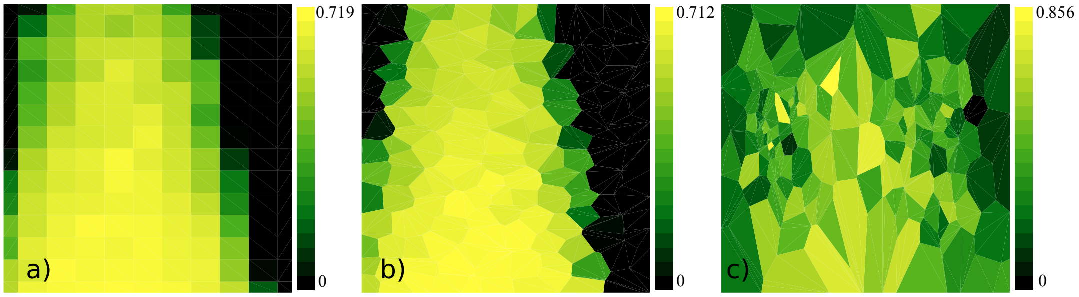

Fig. 5. Degree of the network of the flow generated with START for grid (a), grid with jitter (b) and two clusters (c)

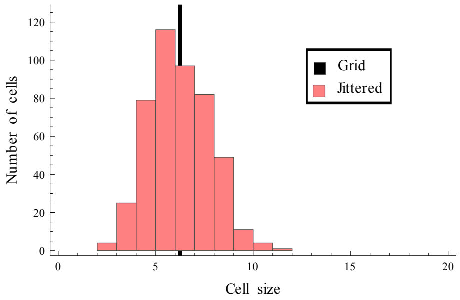

Fig. 6. Tile size distribution for the three node distributions in START.

study shows that the problem goes further, when you also consider the underlying physical system. While the node size distribution in the two clustered sampling versions discussed in section 2 is very similar, their relative position to the underlying flow is not and they have very different distorting effects on the network structure.

In future work it might be interesting to compare different jitter realizations quantitatively, for example using the common component evolution function (CCEF) as introduced in (Tupikina et al., 2013).

So in summary, we found that, as long as the node distribution in space is sufficiently homogeneous, the exact spatial sampling chosen has little impact on the topology of the network in all cases we analyzed. Only in cases of significantly inhomogeneous sampling, distortion and misleading structures arose. It is therefore important to discuss the sampling and its impact when analyzing spatially sampled data. As a simple test of the spatial sampling we suggest to look at the Voronoi size distribution. In a homogeneous sampling this will have a clear peak around A_{sam}/N_{nod}, where A_{sam} is the size of the total sampled area and N_{nod} is the number of nodes. If the peak is shifted or very spread out, that suggests poor spatial sampling.

Acknowledgements. The authors acknowledge funding by the German Science Foundation (DFG graduate school 1539), the German Federal Ministry for Education and Research (BMBF project PROGRESS, support code 03IS2191B) and the support of the LINC project (no. 289447) funded by EC's Marie-Curie ITN program (FP7-PEOPLE-2011-ITN). We would like to thank Norbert Marwan for discussion.

References

Barthélemy, M.: Spatial networks, Physics Reports, 499, 1–101, doi:10.1016/j.physrep.2010.11.002, 2011.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., and Hwang, D.: Complex networks: Structure and dynamics, Physics Reports, 424, 175–308, doi:10.1016/j.physrep.2005.10.009, 2006.

Dippé, M. and Wold, E.: Antialiasing through stochastic sampling, ACM Siggraph Computer Graphics, 19, 69–78, 1985.

Donges, J. F., Zou, Y., Marwan, N., and Kurths, J.: Complex networks in climate dynamics, The European Physical Journal Special Topics, 174, 157–179, doi:10.1140/epjst/e2009-01098-2, 2009.

Heitzig, J., Donges, J., Zou, Y., Marwan, N., and Kurths, J: Node-weighted measures for complex networks with spatially embedded, sampled, or differently sized nodes, The European Physical Journal B 85, 38, doi:10.1140/epjb/e2011-20678-7, 2012.

Kalnay, E. and Kanamitsu, M.: The NCEP/NCAR 40-year reanalysis project, Bull. Amer. Meteor. Soc., 1996.

Malik, N., Bookhagen, B., Marwan, N., and Kurths, J.: Analysis of spatial and temporal extreme monsoonal rainfall over South Asia using complex networks, Climate Dynamics, 39, 971–987, doi:10.1007/s00382-011-1156-4, 2011.

Molkenthin, N., Rehfeld, K., Marwan, N., and Kurths, J.: Networks from flows - from dynamics to topology, Submitted to Sci. Reports, 2013.

NCEP/NCAR: NCEP/NCAR, http://www.erls.noaa.gov/psd.

Rehfeld, K., Marwan, N., Heitzig, J., and Kurths, J.: Comparison of correlation analysis techniques for irregularly sampled time series, Nonlinear Processes in Geophysics, 18, 389–404, doi:10.5194/npg-18-389-2011, http://www.nonlin-processes-geophys.net/18/389/2011/, 2011.

Rehfeld, K., Marwan, N., Breitenbach, S. F. M., and Kurths, J.: Late Holocene Asian summer monsoon dynamics from small but complex networks of paleoclimate data, Climate Dynamics, 41, 3–19, doi:10.1007/s00382-012-1448-3, http://link.springer.com/10.1007/s00382-012-1448-3, 2012.

Rehfeld, K., Molkenthin, N., and Kurths, J.: Spatio-temporal climate transitions from paleoclimate networks, Submitted to Non-

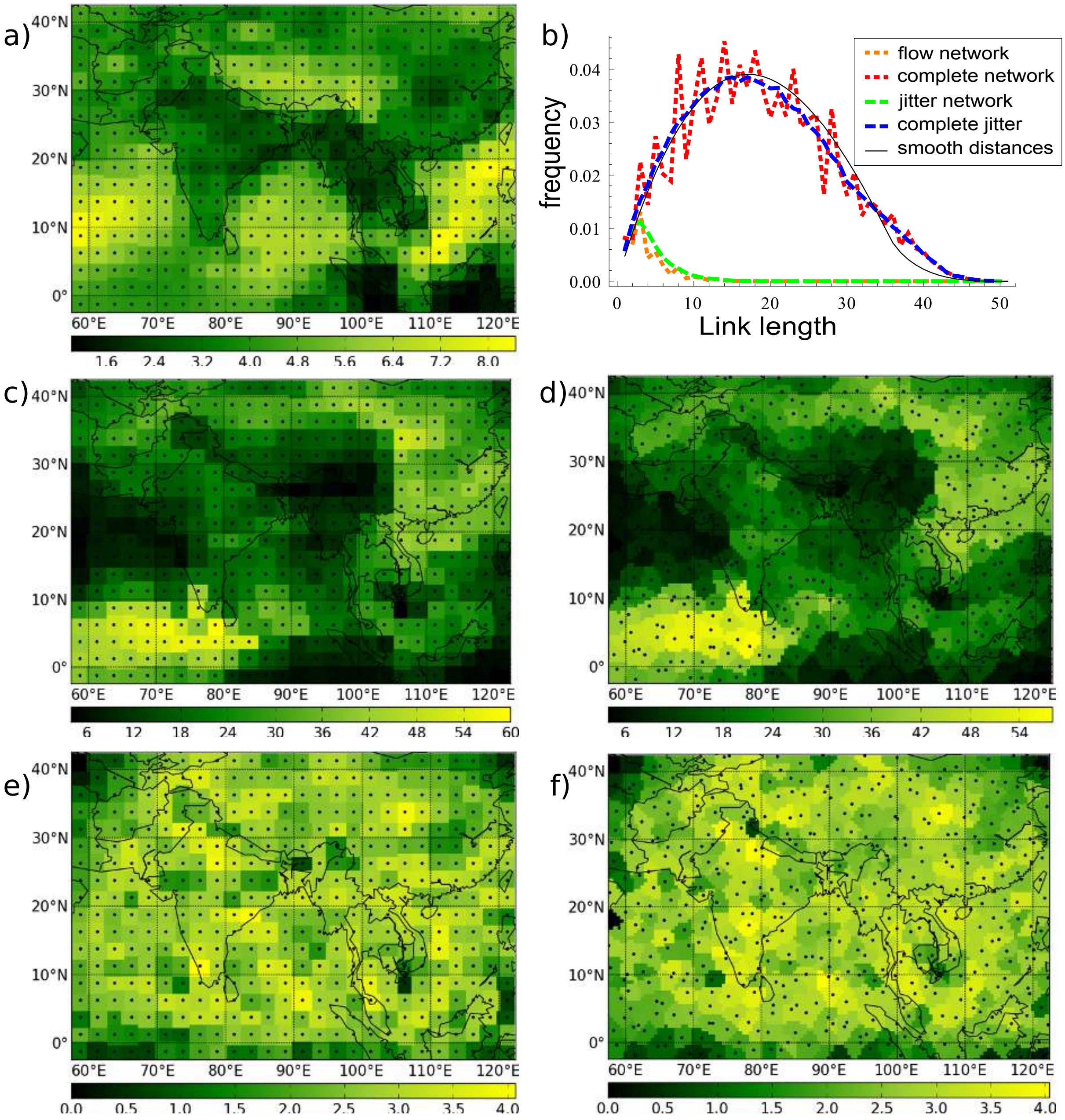

Fig. 7. The correlation network of temperature data from NCEP/NCAR on a grid and a jittered grid. a) absolute mean wind velocity, b) link length distributions, c) degree of the grid network, d) degree of the jittered network, e) betweenness of the grid network, f) betweenness of the jittered network.

lin. Proc. Geophys., 2013.

Rheinwalt, A., Marwan, N., Kurths, J., Werner, P., and Gerstengarbe, F.-W.: Boundary effects in network measures of spatially embedded networks, EPL (Europhysics Letters), 100, 28 002, doi:10.1209/0295-5075/100/28002, http://stacks.iop.org/0295-5075/100/i=2/a=28002?key=crossref.173d39051c8ecb38e3e08eaeebe86b35, 2012.

Tsonis, A. A., Swanson, K. L., and Roebber, P. J.: What Do Networks Have to Do with Climate?, Bulletin of the American Meteorological Society, 87, 585–595, doi:10.1175/BAMS-87-5-585, 2006.

Tsonis, A. A., Wang, G., Swanson, K. L., Rodrigues, F. A., and Costa, L. D. F.: Community structure and dynamics in climate networks, Climate Dynamics, 37, 933–940, doi:10.1007/s00382-010-0874-3, 2010.

Tupikina, L., Rehfeld, K., Molkenthin, N., Stolbova, V., Marwan, N., and Kurths, J.: Detecting evolution of networks using spatial-temporal autocorrelation function, Submitted to Nonlin. Proc. Geophys., 2013.

Unser, M.: Sampling-50 years after Shannon, Proceedings of the IEEE, 88, 569–587, doi:10.1109/5.843002, 2000.

Yamasaki, K., Gozolchiani, A., and Havlin, S.: Climate Networks around the Globe are Significantly Affected by El Niño, Physical Review Letters, 100, 228 501, doi:10.1103/PhysRevLett.100.228501, 2008.