Supporting Information for "Unravelling the community structure of the climate system by using lags and symbolic analysis"

Giulio Tirabassi[1] **and Cristina Masoller**[1,*]

[1]Departament de Fisica, Universitat Politecnica de Catalunya, Colom 11, ES-08222 Terrassa, Barcelona, Spain
[*]cristina.masoller@upc.edu

ABSTRACT

We present a detailed analysis of the communities uncovered with the symbolic method proposed in the main manuscript: we analyse the statistical features of the communities, we discuss the role of the threshold, we demonstrate the robustness of the communities found by using other algorithms, and we compare with the community structure when the network is built with correlation analysis.

Community analysis

In Fig. 1 (Fig. 3 of the main manuscript reproduced here for convenience) we can identify 4 macro-communities: extratropical continents (0) and oceans (2), tropical oceans (4) and ENSO basin (5). Then, there are also two boundary communities, 1 and 6, that are placed at the communities interfaces. Community 3, instead, includes precise areas (maritime continent, subtropical South American Monsoon system, stationary wave patterns of the North Pacific) although the connection among them is unclear. The remaining small community, 7, is clearly an artifact, and won't be examined in the following analysis.

To test the goodness of the community decomposition, we checked which is the relative intensity of the internal connections within the communities in relation with the cross-community connections. In fact, for the decomposition to be meaningful, the communities must represent well connected regions, with weaker connections among them. To investigate this feature we computed the PDFs of the weights of the internal connections of each pair of communities, and we compared them to the PDF of the cross-community connections.

As an example, in Fig. 2 we report the PDFs of the internal links of community 5 together with the cross links between community 5 and communities 2, 3, 4 and 6. We also report the internal weights of these four communities. For geographically separated communities (e.g. 5 and 2) the PDFs are clearly separated, but, the more the communities tend to be geographically close, the more the cross-links PDF overlaps with the internal ones. However there is always a certain separation among the internal and cross-links PDFs, suggesting that the decomposition is meaningful even in the case of communities 5 and 6.

We also analysed the symbolic dynamics of the nodes belonging to the same community. In Fig. 3 we report the average probabilities of symbol occurrence for the largest communities (0-5).

As it can be seen the most prominent feature of these distributions is the presence of high probabilities in the "trend" patterns, that is those ordinal patterns (OPs) in which the data values either increase or decrease for two consecutive months. This characteristic is due to the application of the running mean to the time-series at the beginning of the analysis. To understand which are the features of the symbolic dynamics that are shared by the nodes of each community, we subtract to each histogram the global symbols' distribution (that is computed from all the nodes, without classifying them in communities), obtaining the results presented in Fig. 4.

From these new histograms is evident that, while equatorial communities (4, 5) have more pronounced trends, in the extratropical ones (0, 1) V-like or Λ-like symbols are more likely to occur. These features are in good agreement with the fact that autocorrelations are higher in the tropics with respect to the extratropics.

Lastly we analyse the average transition matrix for each community. Given the abundance of trend symbols the highest transition probabilities will be among them. However, since the connections are defined by the differences between the matrix elements, this common bias is removed by the subtraction in Eq. 3 of the main manuscript. To display more clearly the differences among the average transition matrices of the different communities, we repeat the procedure done for Fig. 4 and subtract the global transition matrix, obtaining the results presented in Fig. 5.

In extratropical communities (0 and 1) there is an anomalous presence of transitions among the V-like and Λ-like patterns (labelled as 1, 2, 4 and 5) although the highest positive signal is among these patterns and trends (labeled 0 and 3). In the

extratropical community the situation is the opposite, with an anomalous prevalence of transitions among the trends. These are reflected in the histograms of Fig. 2, where the cross-links among equatorial and extratropical communities show very small weights (due to large distances between the transition matrices).

Comparison with other community detection algorithms

In the main manuscript the Infomap algorithm was used to identify the communities. Infomap was chosen for two reasons. First, it has been employed in previous climate network studies, such as[1], and second, in our study it provided a community structure that was less noisy than the other algorithms tested. As an example we report in Fig. 6 the comparison among the community structures obtained using Infomap and other three popular community detection algorithms.

In first approximation, the methods produce indeed comparable results, demonstrating the robustness of the structures obtained; however the Infomap decomposition seems to be less noisy.

Influence of the symbolic representation

The community identification algorithm uses a symbolic representation of time-series, known as ordinal analysis, by which SAT anomalies are represented by a set of ordinal patterns. In this way the information of the evolution of the SAT anomalies at the seasonal time-scale is encoded into transition probability matrices. This procedure is quite general, and can be applied to any symbolic representation of the time-series. To study the influence of the symbolic encoding method, here we use an alternative approach and digitalise the time-series with a finite set of values.

We performed this analysis using 10 equally spaced values, ranging from the maximum to the minimum value of each time-series, then we followed the same procedure as in the main text: we computed the transition matrices, and their distances, and by thresholding the distances we obtained a climate network; finally the Infomap algorithm was used to identify the communities. The results are shown in Fig. 7.

As it can be seen, the community structure obtained with this symbolic representation is much noisier than the one obtained through ordinal patterns. This is probably due to the fact that, with the ordinal patterns, each symbol encodes information about the *seasonal evolution* of SAT anomalies; in contrast, with the simpler symbolic encoding used here, the symbols only encode information about the coarse grained SAT anomaly values. We speculate that by using "blocks" composed by several coarse grained values[2], a less noisy community structure could be revealed, and this will be an interesting issue to investigate in future work.

Influence of the threshold used to construct the network

In order to construct a climate network, the weights matrix has to be pruned by using an adequate threshold W. Decreasing the threshold leads to a more connected network, while increasing it results in a sparser one. The number of communities depends on the number of connections, which in turn depends on the threshold. In order to uncover a coherent, well-defined community structure, the threshold has to be carefully chosen.

We report in Fig. 8 the number of communities and the average degree as a function of the threshold. It can be seen that there is a negative correlation between the number of communities and the average degree. The fragmentation of the network into smaller communities (as community 7 in Fig. 6a or communities 8, 9 and 10 in Fig. 6b) could be due to the removal of relevant links that keep the bigger communities together. Thus, to obtain a meaningful community structure, we selected a threshold that provided the best compromise between the need to limit the small-communities-proliferation and the need to include in the network only the relevant links.

We stress that the same qualitative behaviour is found also when using other community detection algorithms.

Comparison with networks constructed by using the Pearson correlation coefficient

In this section we contrast the community structure obtained by using the proposed methodology (based on transition probabilities computed with symbolic analysis), with that obtained with the classical approach, which measures the dynamical similarity of two time series with the Pearson correlation coefficient, r_{ij}. As in[3] we use a threshold $W = 0.5$ to prune the r_{ij} matrix.

Applying the Infomap algorithm to the obtained network results in 8604 communities, but only 20 are composed by more than 2 nodes. Figure 9 shows the largest 16 communities.

As it can be seen, only communities 0 and 1 correspond to coherent structures, namely El-Niño basin, and the tropical oceans, while the others appear to be just noise.

We also report the effect of the threshold value over the community structure in this case. As it can be seen in Fig. 10 the number of communities increases as the threshold is increased, as it occurs when using ordinal analysis (Fig. 8); however, here the change is more abrupt (note the logarithmic vertical scale), and it occurs at low values of the threshold.

Moreover, when a low threshold is used in order to limit the number of communities, the community structure obtained is meaningless, as it can be seen in Fig. 11.

References

1. Tantet, a., and H. a. Dijkstra (2014), An interaction network perspective on the relation between patterns of sea surface temperature variability and global mean surface temperature, *Earth Syst. Dynam.*, *5*(1), 1–14, doi:10.5194/esd-5-1-2014.
2. Barreiro, M., A. C. Marti, and C. Masoller (2011), Inferring long memory processes in the climate network via ordinal pattern analysis, *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *21*(1), 013,101.
3. Tsonis, A., and P. Roebber (2004), The architecture of the climate network, *Physica A: Statistical Mechanics and its Applications*, *333*, 497–504.
4. Newman, M. E. J. (2006), Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E*, *74*, 3: 036104.
5. Clauset, A., M. E. J. Newman, and C. Moore (2004), Finding community structure in very large networks. *Phys. Rev. E 70*, 066111.
6. Pons, P., and M. Latapy. (2006), Computing communities in large networks using random walks. *J. Graph Algorithms Appl. 10*, 2: 191-218.

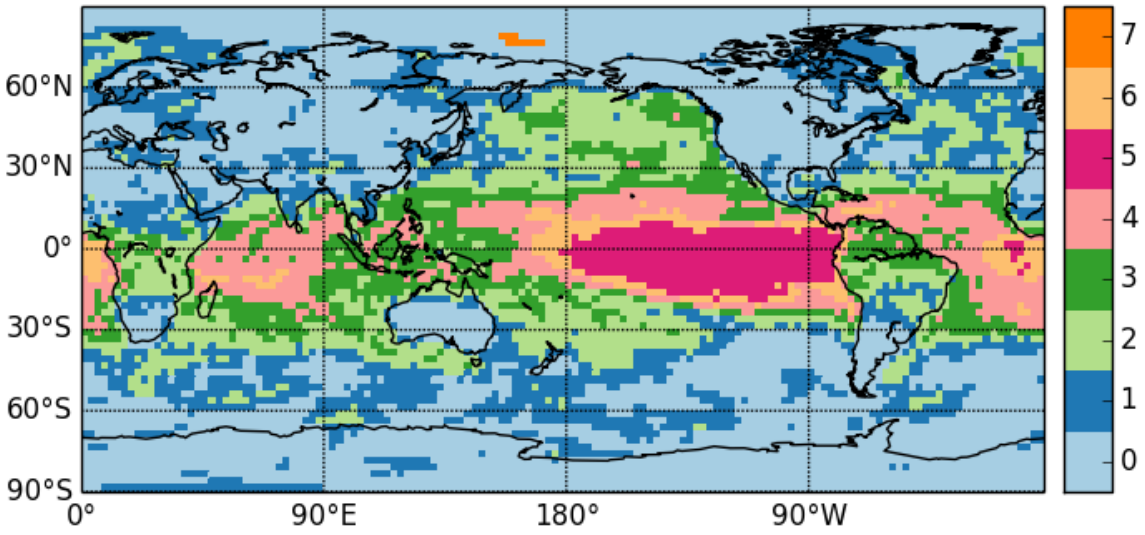

Figure 1. Figure 3 of the main manuscript: community structure from SAT anomalies detected by using symbolic time-series analysis. Matlab software (version number 7.12.0.635) was used to create this map, and all the other maps in this work.

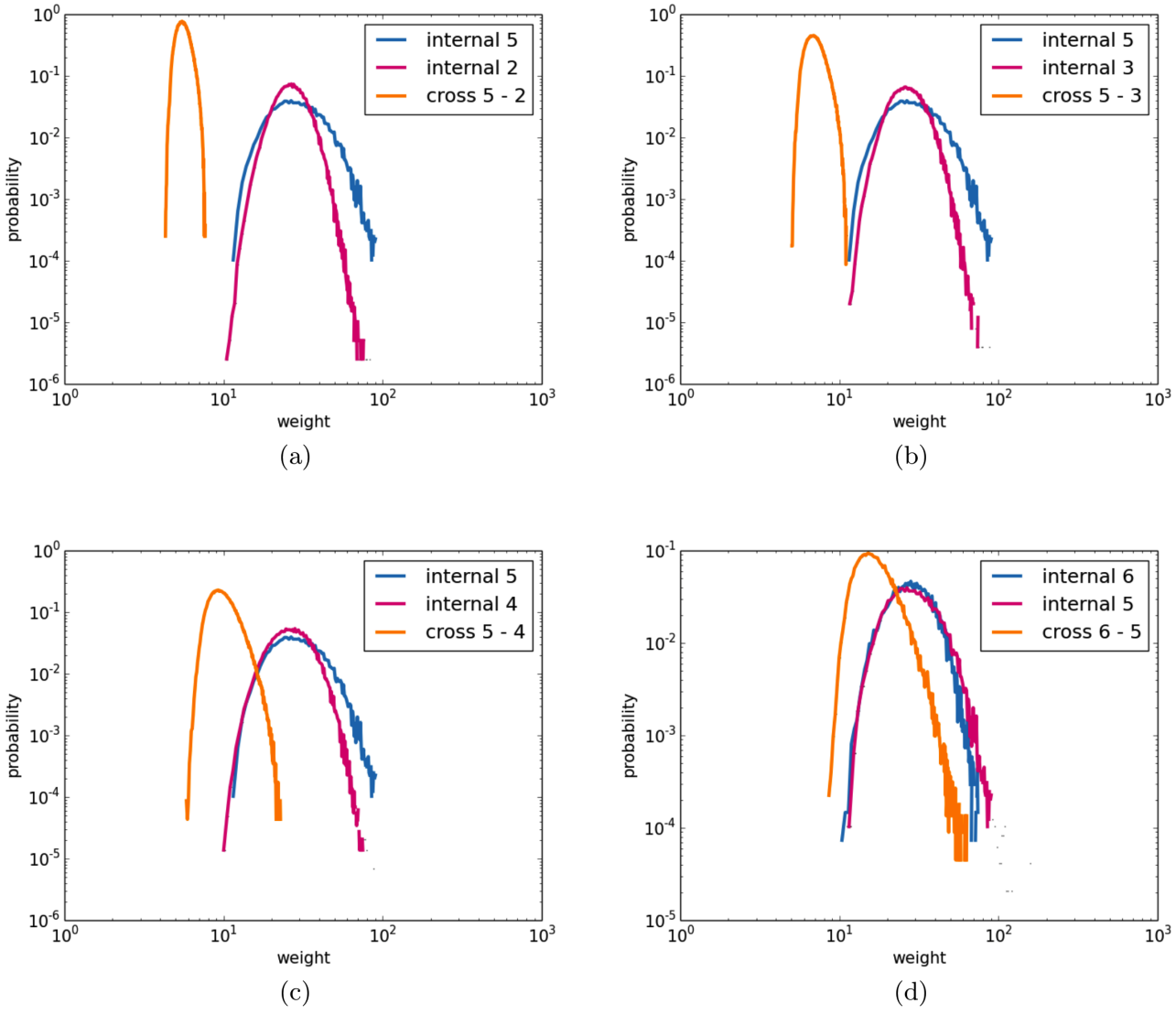

Figure 2. PDFs of the internal and cross-community weights for community 5 (El-Niño Basin) and other four communities (2, 3, 4 and 6, respectively panels **(a)**, **(b)**, **(c)** and **(d)**).

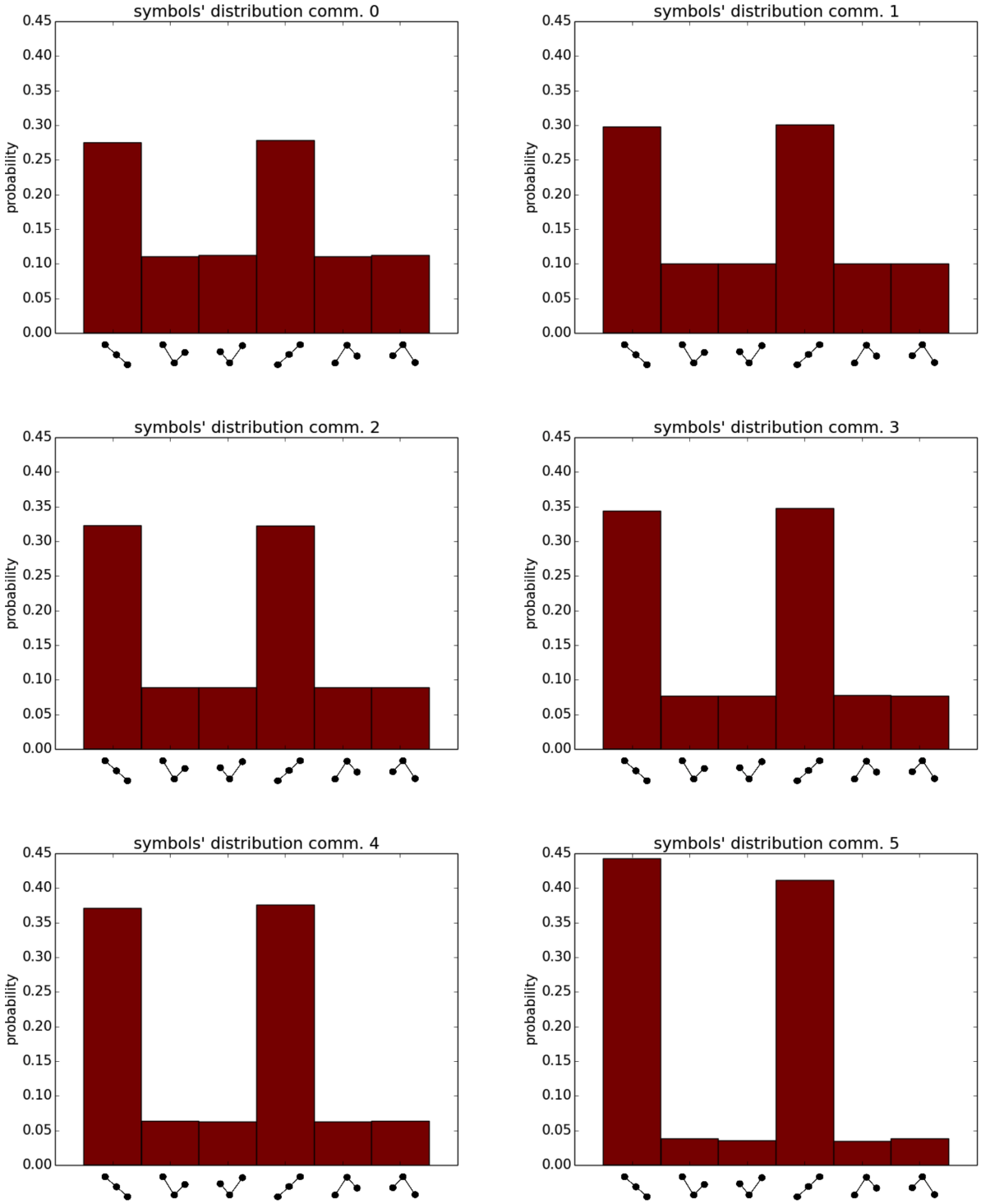

Figure 3. Average probabilities of symbol occurrence for the largest communities (0-5).

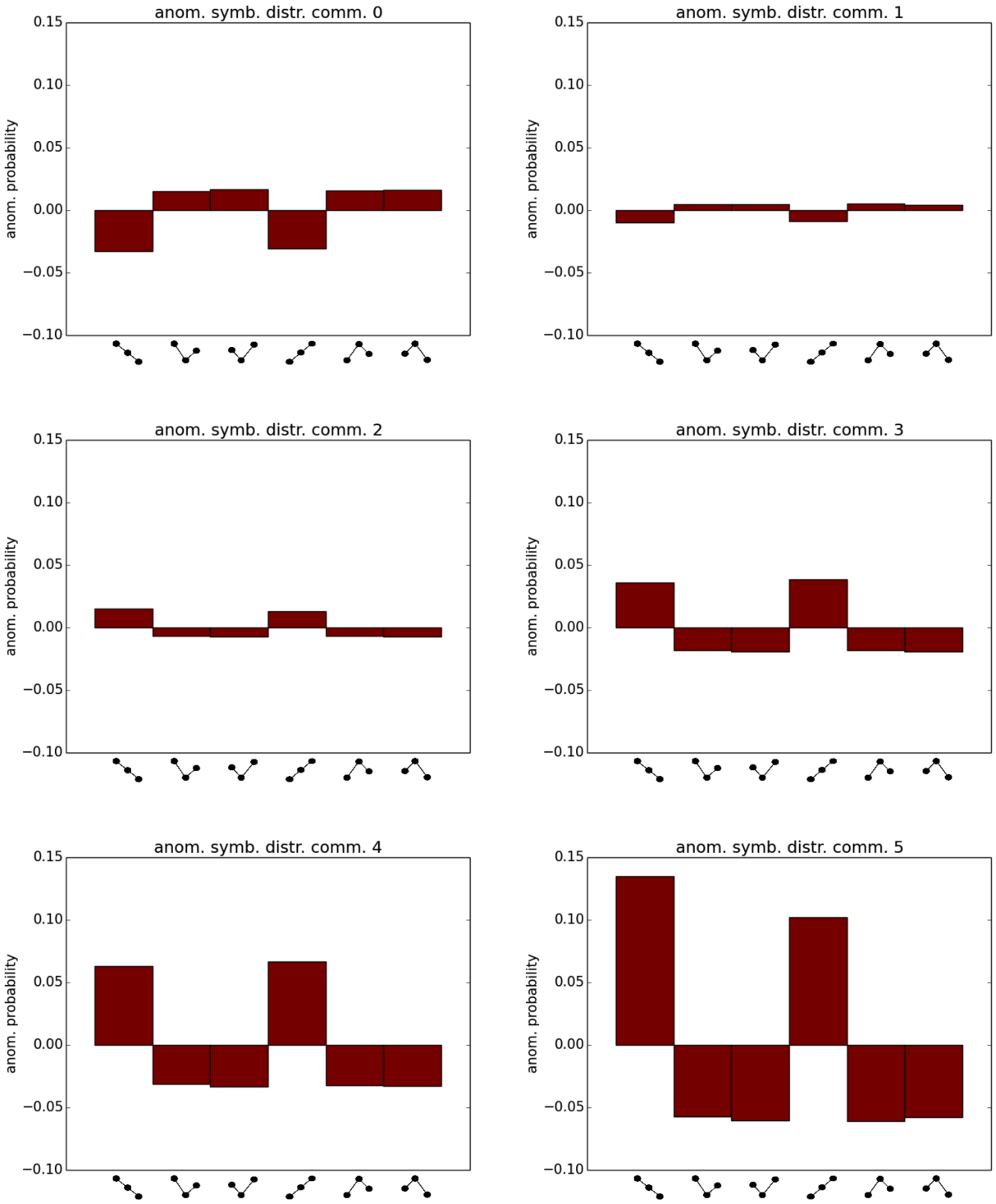

Figure 4. Anomalous symbols' distributions. The bars indicate the differences between the probabilities displayed in Fig. 3 and the global probabilities (computed from all the nodes, without classifying them in communities).

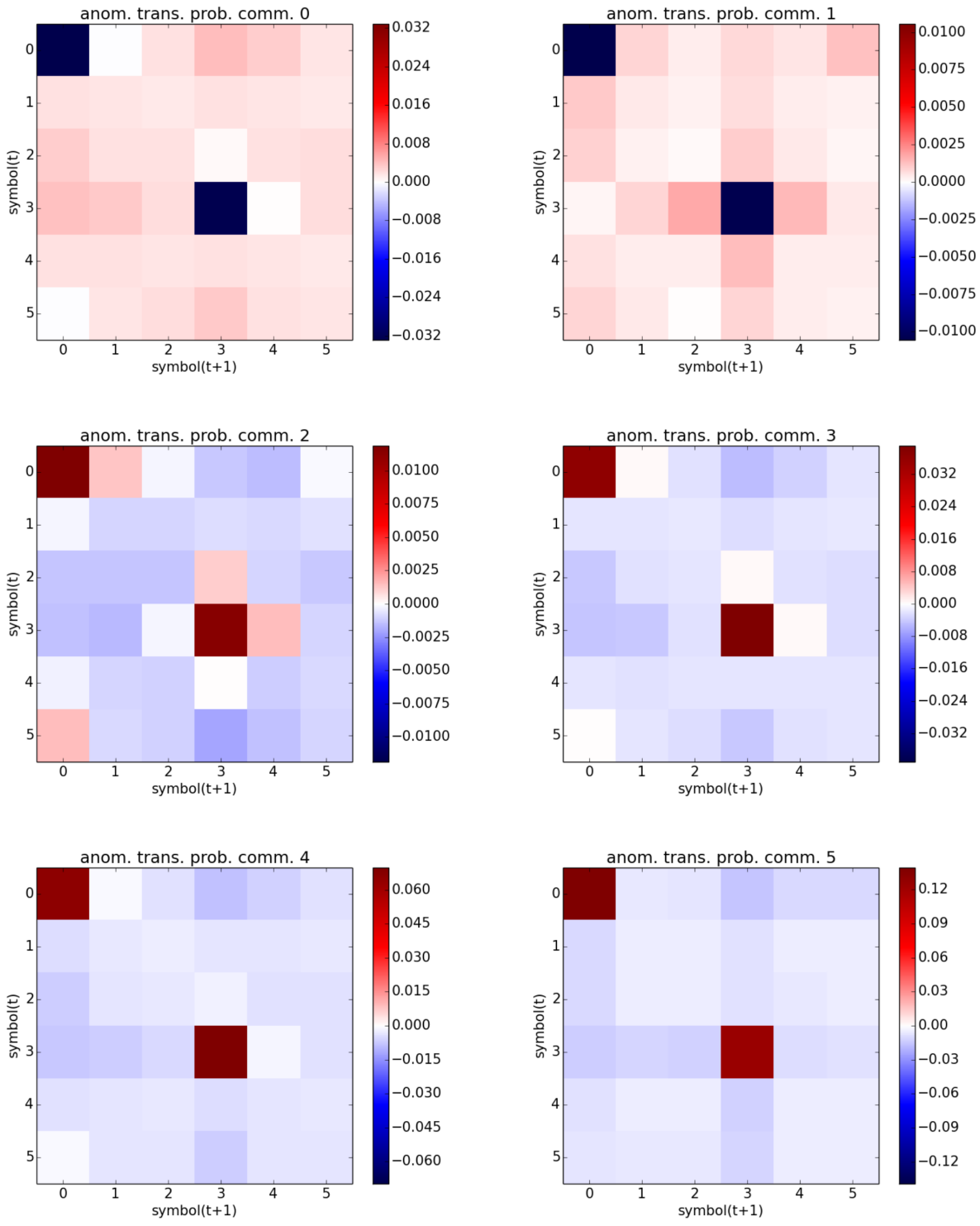

Figure 5. Anomalous transition probabilities. The color code indicates the difference between the average transition probabilities in each community and the global transition probabilities (computed from all the nodes, without classifying them in communities).

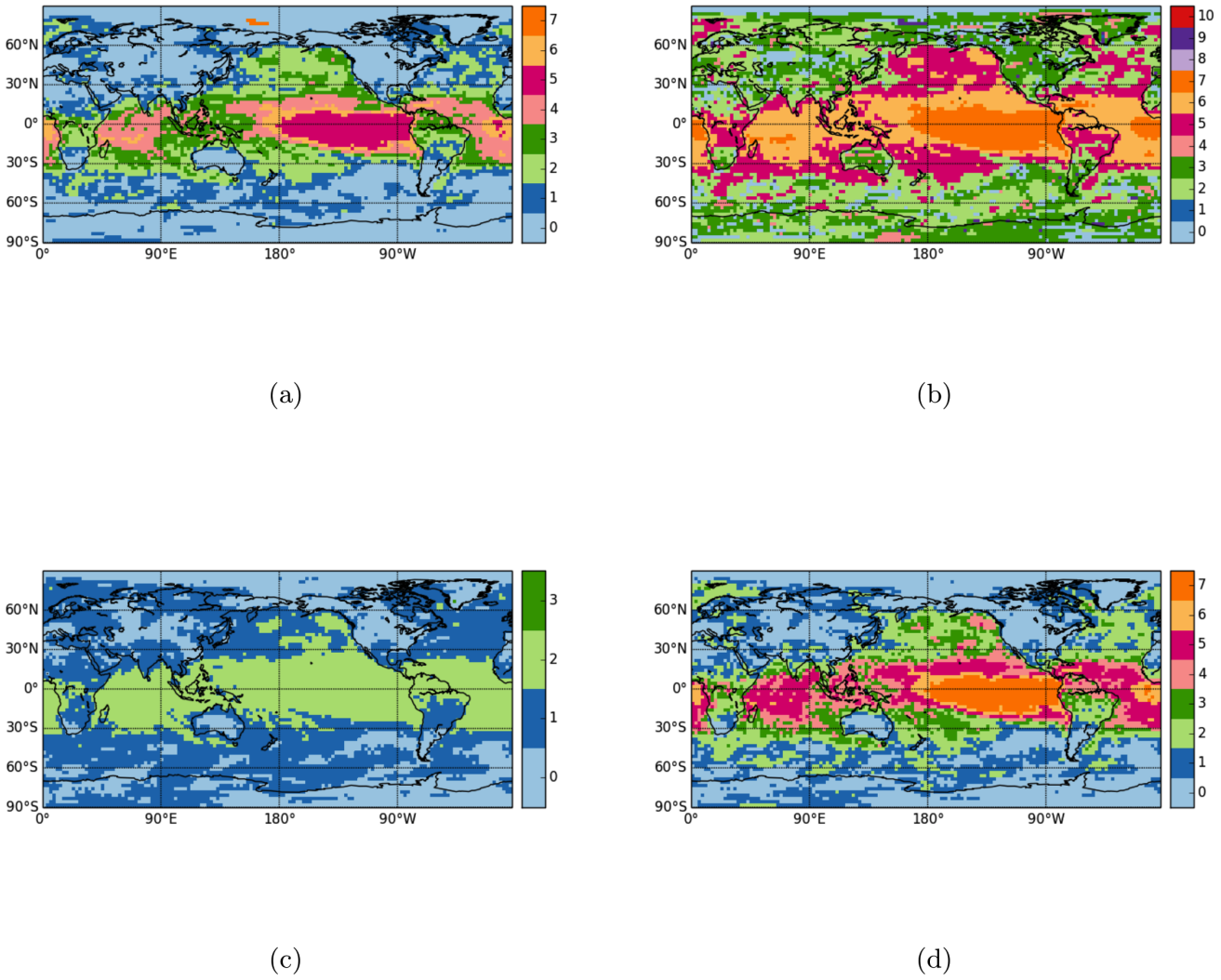

Figure 6. Community structure found by analyzing the same network with four popular community detection algorithms. **(a)** Infomap (same as Fig. 1); **(b)** Leading Eigenvector[4] (threshold used: 35); **(c)** Fast Greedy Modularity Maximization[5] (threshold used: 35); **(d)** Walktrap[6].

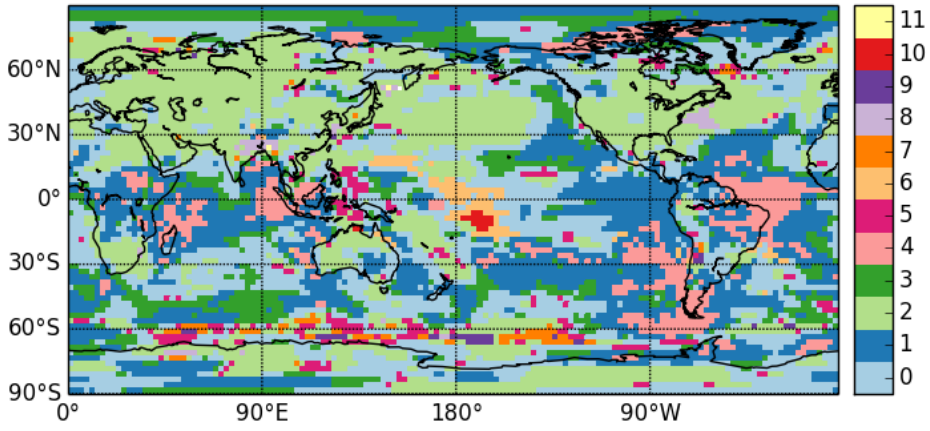

Figure 7. Community structure found with Infomap when the network is constructed by using a symbolic transformation that only encodes information about the coarse grained SAT anomaly values.

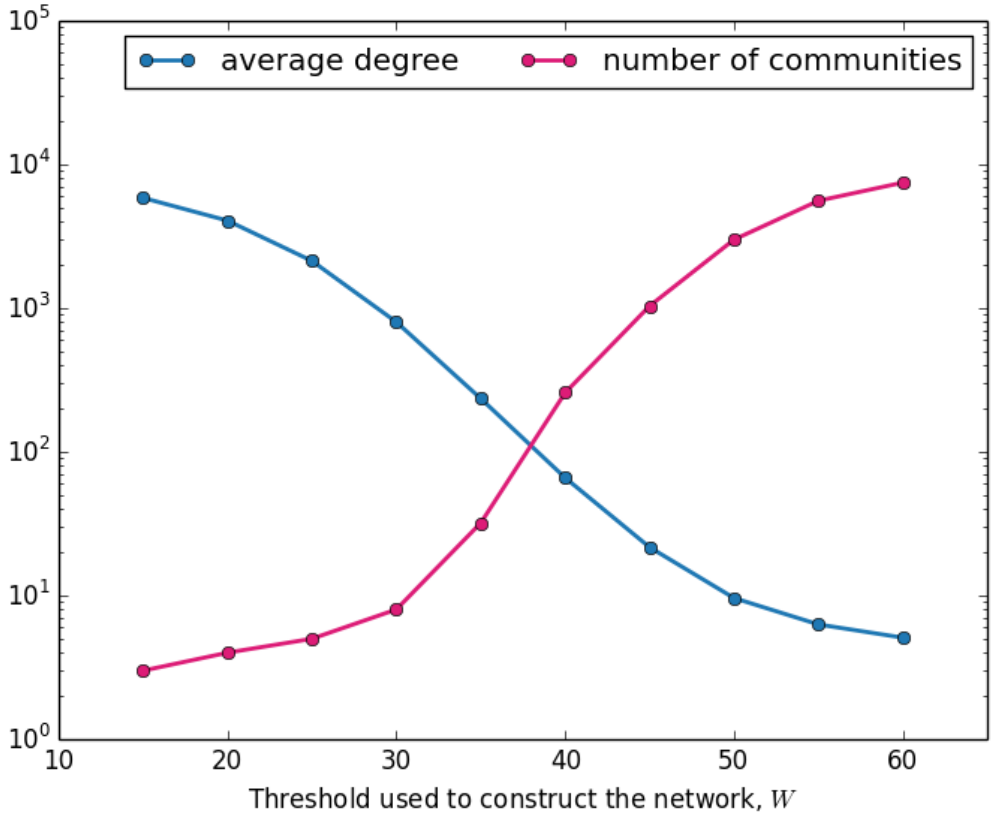

Figure 8. Dependence of the network average degree and of the number of communities with the value of the threshold used to construct the network, W. For the community structure shown in Fig. 1, the threshold used was $W = 30$.

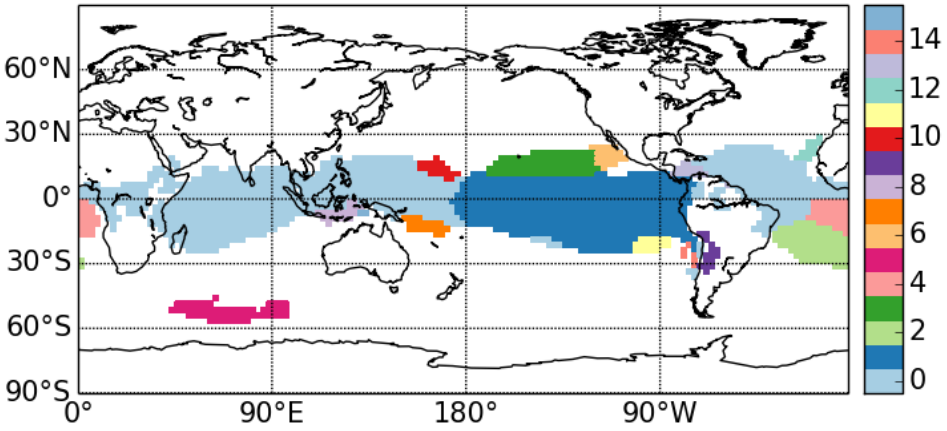

Figure 9. Community structure found with Infomap when the network is constructed by using the Pearson correlation coefficient as a measure of dynamical similarity. The value of the threshold used to construct the network is $W = 0.5$. For clarity, only the largest 16 communities are shown.

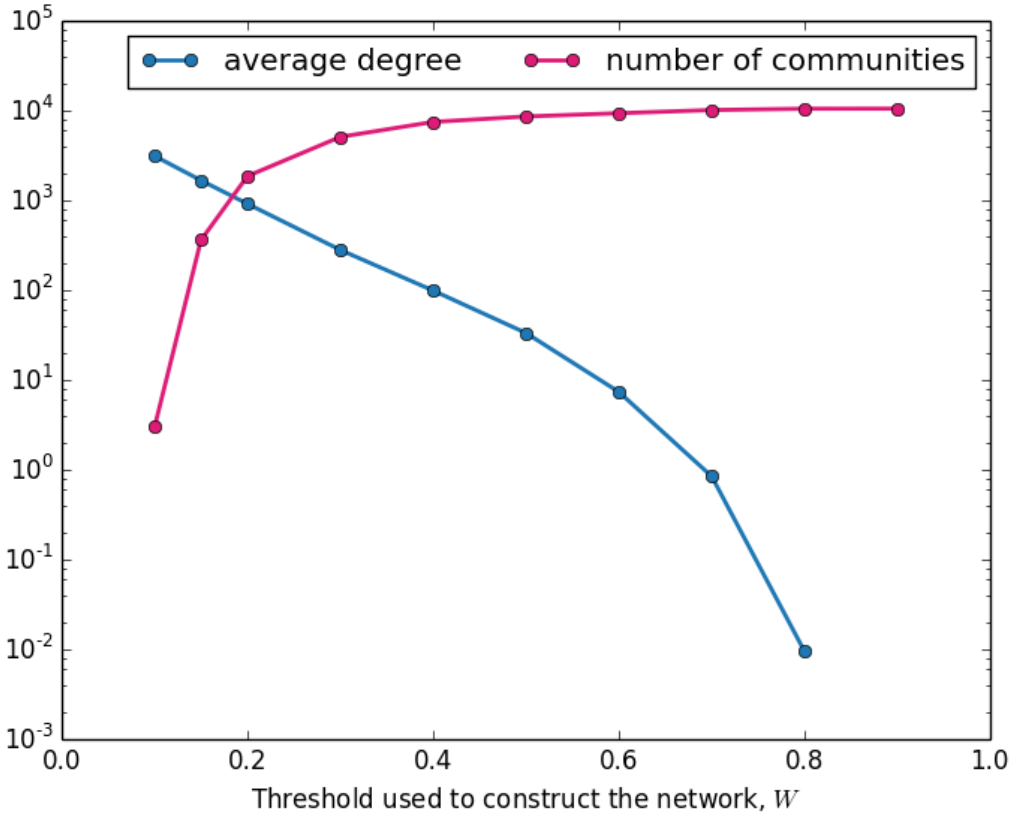

Figure 10. Dependence of the network average degree and of the number of Infomap communities from the threshold value for a network obtained thresholding the covariance matrix of the data.

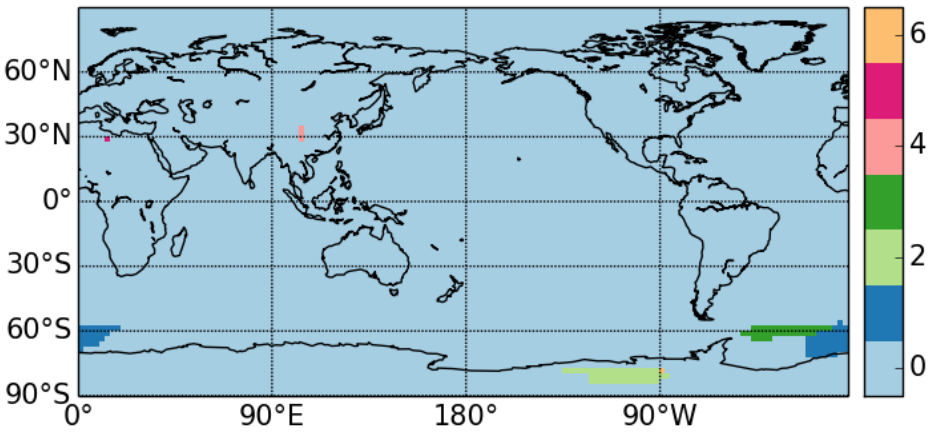

Figure 11. As Fig.9 but with $W = 0.107$.