# Unravelling the community structure of the climate system by using lags and symbolic time-series 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

Many natural systems can be represented by complex networks of dynamical units with modular structure in the form of communities of densely interconnected nodes. Unraveling this community structure from observed data requires the development of appropriate tools, particularly when the nodes are embedded in a regular space grid and the datasets are short and noisy. Here we propose two methods to identify communities, and validate them with the analysis of climate datasets recorded at a regular grid of geographical locations covering the Earth surface. By identifying mutual lags among time-series recorded at different grid points, and by applying symbolic time-series analysis, we are able to extract meaningful regional communities, which can be interpreted in terms of large-scale climate phenomena. The methods proposed here are valuable tools for the study of other systems represented by networks of dynamical units, allowing the identification of communities, through time-series analysis of the observed output signals.

## Introduction

Many real-world complex systems can be represented in terms of networks of interacting nodes embedded in space. Examples include power grids, fiber-optic networks, road networks, flight connections, etc[1–3]. Such networks are usually organized in modules or communities of densely interconnected nodes[4–11]. The spatial embedding of the network can hidden the underlying community structure, rendering the identification of communities a challenging a task[12–14]. The effects of space in the topology of the network are particularly important when the network is built with correlation analysis of output signals which are recorded at a regular grid of observation points. Examples of this situation include brain functional networks[15–17] and climate networks[18–21].

Here we focus on climate networks, which provide relevant insight into global climate phenomena[22–29]. Climate communities reveal coherent subsystems[30], can be used for model inter-comparisons[31], and can advance climate predictability[32]. For example, communities obtained from the analysis of sea surface temperature (SST) reveal information about long-term SST variability[33].

In climate networks, detecting communities representing geographical regions with similar climate is challenging because the links are defined via correlation analysis (for example, by using the Pearson coefficient). Thus, in the resulting network, the nodes are linked mainly to neighboring nodes (because of the high correlation that results from physical proximity), and long distance links are scarce. This spatial effect can hide, for example, the fact that distant extratropical land masses (in the two hemispheres) are likely to have similar climate. This similarity is not captured because, when the network is built with correlation analysis, the northern and southern hemispheres are indirectly or only weakly connected.

Here propose and validate two methodologies to overcome this problem. From time-series recorded at a regular grid of points covering the Earth's surface, the methods extract different and relevant properties of our climate. With the first method, which is based in computing mutual lags between time-series, we are able to infer communities defined by regions in which the oscillations of a climate variable (the surface air temperature, or the geopotential height) are in-phase; with the second method, which is based in symbolic time-series analysis, we group together regions that share similar properties of the symbolic dynamics. We validate these methods by uncovering meaningful communities, which can be related to known properties of the climate system.

## Data

We analyze monthly-averaged surface air temperature (SAT) and geopotential height (GH) reanalysis data from NCEP/NCAR (state-of-the-art model simulation with data assimilation using past observed data where and when is available,[34]). The data

covers the period from January 1948 to May 2012 ($T = 773$) data points and has a spatial grid resolution of 2.5 degrees ($N = 10226$ nodes). The data can be freely downloaded from the NCEP/NCAR reanalysis project webpage:

http://www.esrl.noaa.gov/psd/data/reanalysis/reanalysis.shtml

## Time lags method

The first method proposed for community identification unveils geographical regions in which the oscillations of a climate variable are in-phase, revealing similar response to annual solar forcing. To identify such regions, for each time-series we first compute the annual cycle, and then compare the mutual lags among all pairs of time-series. Thus, for each $x_i^y(t)$, where $x$ indicates either SAT or GH, $i$ indicates the geographical location, $y$ indicates the year and $t$ indicates the month within that year, we first compute the seasonal cycle as $x_i(t) = (1/Y)\sum_y x_i^y(t)$ where $Y$ is the number of years (64 or 65 depending on the month). Then, for each pair of time-series, $i$ and $j$, we compute the lagged cross-correlation of the seasonal cycles, $C_{ij}(\tau) = (1/12)\sum_t x_i(t)x_j(t+\tau)$, and determine their mutual lag, $\ell_{ij}$, as the value of $\tau$ that maximizes $C_{ij}(\tau)$. The seasonal cycle is by definition periodic, therefore, we search for a maximum in $\tau \in [0, 11]$[35,36]. With $\ell_{ij}$, we calculate $\ell_{ji}$ as: $\ell_{ji} = 12 - \ell_{ij}$ if $\ell_{ij} \neq 0$, else $\ell_{ji} = 0$.

If the mutual lags among any three regions ($i$, $j$, $k$) are well defined, they should satisfy:

$$\ell_{ij} \quad = \quad (\ell_{ik} + \ell_{kj}) \bmod 12. \tag{1}$$

To fix the ideas, let us consider that $i$ is a region in continental Europe, $j$ is in the tropical eastern Pacific Ocean and $k$ is in southern South America. If the lag between $i$ and $j$ is 8 months, and the lag between $i$ and $k$ is 6 months, then, the lag between $j$ and $k$ should be 2 months.

Therefore, one vector containing the lags between a region, $k$, and any other region, $i$, $\vec{\ell}_k = \{\ell_{ik}\}$, contains in fact all the information needed for computing the lag between any two regions $i$ and $j$: if we know $\ell_{ik}$ and $\ell_{jk}$, $\ell_{ij}$ can be calculated from Eq. (1).

However, because we consider monthly-averaged data, $\ell_{ij}$, $\ell_{ik}$ and $\ell_{kj}$ are integer numbers of months, and thus, because of round-off errors (the real lags are not necessarily integer numbers) Eq. (1) will not hold for all the triples $(i, j, k)$.

In order to identify the regions that have well-defined lags among them, we chose a reference node $i$, and, for each other node $j$, we test Eq. (1) for all the possible $k$s. If the relation is satisfied in more than 50% of the cases, we consider $\ell_{ij}$ to be a well defined lag, otherwise no value is assigned. This is in fact a simple work-around solution to a complex optimisation problem: how to remove the minimum number of $\ell_{ij}$ values, so that Eq.(1) is valid for all the remaining ones.

Then, the information about all mutual lags, $\{\ell_{ij} \forall i, j\}$, can be summarized in just one map, which displays the lags between a region, $k$, and any other region $i$ (i.e., displays the vector $\vec{\ell}_k$), because, from this map, any lag $\ell_{ij}$ can be calculated using Eq.(1). For SAT time-series, the resulting map is plotted in Fig. 1a for a reference region in continental Europe and in Fig. 1b for one in southern South America. In these plots, all the areas sharing the same color present a seasonal cycle in phase, and the white areas indicate regions in which the lag with the reference point is not well-defined.

The two panels are very similar; the white areas are a little fraction of the total area, and they are located at the boundaries of well defined regions, thus confirming a coherent community decomposition. We can see that, in spite of the fact that the annual solar forcing is zonally symmetric, the maps of lag times are heterogeneous. In particular, wide ocean areas have a one-month delay with respect to the landmasses. In the eastern boundaries of the oceans this delay reaches two months and even three months in El Niño region. While the one-month delay can be expected due to thermal inertia of the water respect to the land, the longer delays have no straightforward explanation and require further investigation.

By applying this methodology to the geopotential height at 500 hPa, we uncover a very different community structure, displayed Fig. 2. In this case, due to the fact that the seasonal cycle is highly non-linear and heterogeneous, the white areas with not well-defined lags increase with respect to the SAT lag map. In particular, a wide part of the equatorial belt, as well as the polar region, have undefined lags. Also, in the northern hemisphere, two regions with undefined lags are consistent with the North Atlantic Oscillation pattern, which on long time-scales can act as a source of noise for the lag determination. Nevertheless, several consistent features can be seen, including the six-month symmetry between the two hemispheres.

## Symbolic method

The second method proposed for community identification allows to uncover regions that share similar symbolic patterns of climate variability. To rule out similarities which are due to the periodicity induced by the solar annual cycle, the analysis is now performed on *anomaly* time-series, $y_i(t)$, computed by subtracting the seasonal cycle to the raw data[37].

In order to construct a network in which regions with similar climate are connected, we first use symbolic analysis and transform each time-series, $y_i(t)$, in a symbolic sequence, $s_i(t)$. Next, for each symbolic sequence we calculate the transition probabilities, $M_i(\alpha, \beta)$, among all possible pairs of symbols, $\alpha$ and $\beta$. Specifically, we compute the number of times $\beta$ occurs

after $\alpha$, over the total number of transitions. The transition probabilities (TPs) describe the statistics of the symbolic sequence. In order that two regions, $i$ and $j$, with similar (dissimilar) TPs, are strongly (weakly) linked, we define the weight of the link between $i$ and $j$ as

$$w_{ij} = \left( \sum_{\alpha,\,\beta} (M_i(\alpha,\beta) - M_j(\alpha,\beta))^2 \right)^{-1} . \tag{2}$$

Next, we construct a network by considering only the strongest links, *i.e.,* we threshold $\{w_{ij}\}$ and obtain the adjacency matrix, $A_{ij} = \mathrm{H}(w_{ij} - W)$, where $H$ is the Heaviside step-function and $W$ is a threshold chosen such that the network is not too sparse or too connected[38]. Then, we apply the Infomap algorithm of community identification[7,33]. To summarize, in this second method, the symbolic information obtained from $N$ time-series is encoded in $N$ TP matrices, and then we identify the regions which have similar TPs.

There are many ways of perform the symbolic data reduction. Here we use the method of *ordinal analysis*[39–41] because it has been proven useful to construct climate networks[22,25,35] (a comparison with other symbolic methods is presented in the *Supplementary Information*, SI). In this approach, each time series is divided into non-overlapping segments of length $Q$, and each segment is assigned a symbol, $s$, (known as ordinal pattern) according to the ranking of the values inside the segment. For example, with $Q = 3$, if $y_i(t) < y_i(t+1) < y_i(t+2)$, $s_i(t)$ is "012", if $y_i(t) > y_i(t+1) > y_i(t+2)$, $s_i(t)$ is "210", and so forth. Thus, the symbols take into account the *relative temporal ordering* of the values and not the values themselves. In this way, each symbol encodes information about the evolution of the time-series during $Q$ months. In order to estimate the TPs with good statistics, the length of the time-series must be much longer than the number of possible transitions, *i.e.,* $T \gg Q!$[2]. Thus, with $T = 773$ months, we use $Q = 3$.

The community structure inferred from SAT anomalies is presented in Fig. 3. As it can be seen, the algorithm divides the world in 8 areas, labeled with different colors. These areas share similar dynamics, in the sense of similar symbolic transition probabilities. The continents in the two hemispheres are in the same community and a large coherent area is detected in the ENSO basin, while the oceans are divided in tropical and extratropical. A detailed analysis of these communities is provided in the SI.

It is important to remark that such community structure can not be inferred from networks that are constructed from correlation analysis (by using Pearson coefficient or mutual information). As our goal is that regions with similar climate belong to the same community, the classic tools are not useful, because they would not provide direct connections among extratropical regions. In order to belong to the same community two nodes must be part of the same group of strongly interconnected nodes, and in the correlation approach, where the links are prominently local, direct teleconnections across hemispheres are scarce (see SI for more details).

It is interesting to compare how these communities are related to those found in Fig. 1 through the seasonal cycle. There are borders among different communities that are indeed shared by the two sets, such as the extra-tropical coastlines, or the separation of northern from southern Australia and of southern South America from the rest of the continent.

The Infomap algorithm automatically converges towards a certain number of communities that cannot be directly controlled, as they are defined by the network structure. The number of communities depends on the network density, which is in turn modified by the threshold $W$ used to construct the network. Increasing the network density makes the network to look like a giant coherent cluster, and the Infomap algorithm will detect a smaller number of communities. Decreasing the density will break the network in many small parts, and Infomap will detect them as many separate communities (see SI for details).

Figure 4 displays the communities extracted for GH anomalies at 1000 and 300 hPa. As it can be seen, increasing the height of the field implies a more zonal distribution of the communities: at 300 hPa the tropics form a belt that differentiates from the extratropical areas, which belong to the same community, the two are separated by strip-like communities, probably a signature of the subtropical jet. At 1000 hPa, instead, the effect of tropical convection is dominant, separating the low latitudes in two areas, the Maritime Continent together with the ENSO basin (perhaps a signature of the Walker circulation), and the rest of the tropics. The extratropics instead are grouped in the same community, regardless of the presence of landmasses.

## Discussion

We have presented two methods to identify communities in dynamical complex systems using the properties of observed time-series. We tested the methods with climate data (surface air temperature and the geopotential height at two pressure levels), and uncovered communities that are consistent with main large-scale climate phenomena. The first method, based on computing mutual lags among the time-series through correlation analysis, uncovered communities formed by geographical regions with synchronous seasonal cycles. The second method, based on symbolic analysis, identified communities formed by geographical regions where the climate variability displays similar symbolic patterns.

Practical applications of the proposed methods include the analysis of specific geographical regions, to uncover sub-areas with similar micro-climate. Moreover, the two methods can be used to analyse other real-world, dynamical complex systems. Because the methods were demonstrated with space-embedded, short and noisy datasets, they can be used to analyse brain signals, to uncover brain regions with in-phase dynamics or with similar symbolic dynamics.

## References


1. R. Albert and A. L. Barabasi, Rev. Mod. Phys. **74**, 47 (2002).
2. M. E. J. Newman, SIAM Rev. **45**, 167 (2003).
3. S. Boccaletti, V. Latora, Y. Moreno, M. Chavez and D.-U. Hwang, Phys. Rep. **424**, 175 (2006).
4. G. Palla, I. Derenyi, I. Farkas, and T. Vicsek, Nature **435**, 214 (2005).
5. S. Fortunato, Phys. Rep. **486**, 75 (2010).
6. J. Reichardt and S. Bornholdt, Phys. Rev. Lett. **93**, 218701 (2004).
7. R. Rosvall and C. T. Bergstrom, PNAS **104**, 7327 (2007).
8. E. A. Leicht and M. E. J. Newman, Phys. Rev. Lett. **100**, 118703 (2008).
9. M. A. Serrano, M. Boguña and A. Vespignani, PNAS **106**, 6483 (2009).
10. P. J. Mucha, T. Richardson, K. Macon, M. A. Porter and J.-P. Onnela, Science **328**, 876 (2010).
11. R. R. Nadakuditi and M. E. J. Newman, Phys. Rev. Lett. **108**, 188701 (2012).
12. P. Expert, T. S. Evans, V. D. Blondel, and R. Lambiotte, PNAS **108**, 7663 (2011).
13. F. Cerina, A. Chessa, F. Pammolli and M. Riccaboni, Sci. Rep. **4**, 4546 (2014).
14. D. A. Vilhena and A. Antonelli, Nat. Comm. **6**, 6848 (2015).
15. S. Bialonski, M.-T. Horstmann, and K. Lehnertz, Chaos **20**, 013134 (2010).
16. V. M. Eguiluz, D. R. Chialvo, G. A. Cecchi, M. Baliki, and A. V. Apkarian, Phys. Rev. Lett. **94**, 018102 (2005).
17. E. Bullmore and O. Sporns, Nat. Rev. Neuroscience **10**, 186 (2009).
18. A. Tsonis and P. Roebber, Physica A **333**, 497 (2004).
19. K. Yamasaki, A. Gozolchiani, and S. Havlin, Phys. Rev. Lett. **100**, 228501 (2008).
20. A. A. Tsonis and K. L. Swanson, Phys. Rev. Lett. **100**, 228502 (2008).
21. J. F. Donges, Y. Zou, N. Marwan, and J. Kurths, EPL **87**, 48007 (2009).
22. M. Barreiro, A. C. Marti, and C. Masoller, Chaos **21**, 013101 (2011).
23. L. C. Carpi, P. M. Saco, O. A. Rosso, and M. G. Ravetti, Eur. Phys. J. B. **85**, 389 (2012).
24. Y. Berezin, A. Gozolchiani, O. Guez and S. Havlin, Sci. Rep. **2**, 666 (2012).
25. J. I. Deza, M. Barreiro, and C. Masoller, Eur. Phys. J. Spec. Top. **222**, 511 (2013).
26. J. Hlinka, D. Hartman, M. Vejmelka, D. Novotna and M. Palus, Clim. Dyn. 42, 1873 (2014).
27. T. Zerenner, P. Friederichs, K. Lehnertz and A Hense, Chaos 24, 023103 (2014).
28. I. Fountalis, A. Bracco and C. Dovrolis, Clim. Dyn. 45, 511 (2015).
29. J. F. Donges, I. Petrova, A. Loew, N. Marwan, and J. Kurths, Clim. Dyn. 54, 2407 (2015).
30. A. A. Tsonis, G. Wang, K. L. Swanson, F. A. Rodrigues, and L. D. F. Costa, Clim. Dyn. **37**, 933 (2011).
31. I. Fountalis, A. Bracco, and C. Dovrolis, Clim. Dyn. **42**, 979 (2014).
32. K. Steinhaeuser and A. A. Tsonis, Clim. Dyn. **42**, 1665 (2014).
33. A. Tantet and H. A. Dijkstra, Earth Syst. Dynam. **5**, 1 (2014).
34. R. Kistler, W. Collins, S. Saha, G. White, J. Woollen, E. Kalnay, M. Chelliah, W. Ebisuzaki, M. Kanamitsu, V. Kousky, et al., reanalysis: Monthly means cd-rom and documentation, Bull. of the Am. Meteor. Soc. **82**, 247 (2001).
35. G. Tirabassi and C. Masoller, EPL **102**, 59003 (2013).
36. E. Martin and J. Davidsen, Nonlin. Processes Geophys. **21**, 929 (2014).

**37.** As in Ref.[33], we remove the fast anomaly fluctuations by using a one-year running mean.

**38.** $W$ is chosen such that each node is connected, on average, to 5% of the Earth surface; see the *Supplementary Information*, SI, for a discussion of the role of $W$. In the SI we also demonstrate the robustness of the results by presenting the communities detected by different algorithms.

**39.** C. Bandt and B. Pompe, Phys. Rev. Lett. **88**, 174102 (2002).

**40.** M. Zanin, L. Zunino, O. A. Rosso and D. Papo, Entropy **14** 1553 (2012).

**41.** J. M. Amigo, K. Keller and J. Kurths, Eur. Phys. J-ST **222**, 241 (2013).

## Acknowledgements

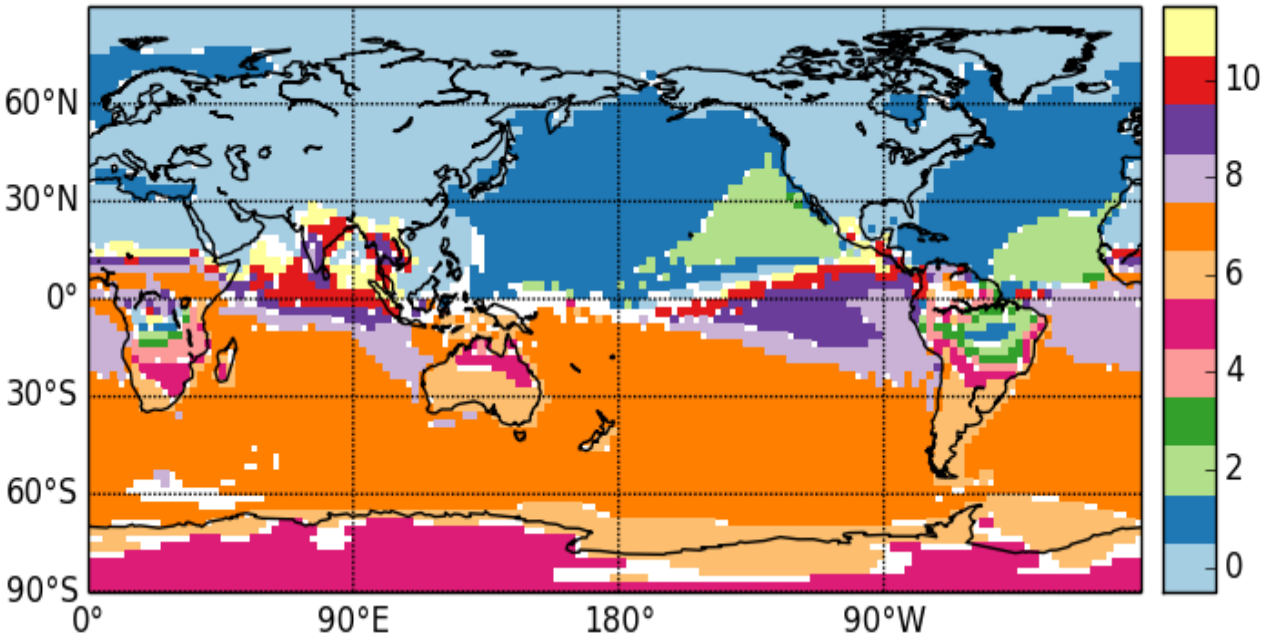
This work was supported by the LINC project (FP7-PEOPLE-2011-ITN, Grant No. 289447). C. M. also acknowledges partial support from Spanish MINECO (FIS2015-66503-C3-2-P) and ICREA ACADEMIA.

## Author contributions statement

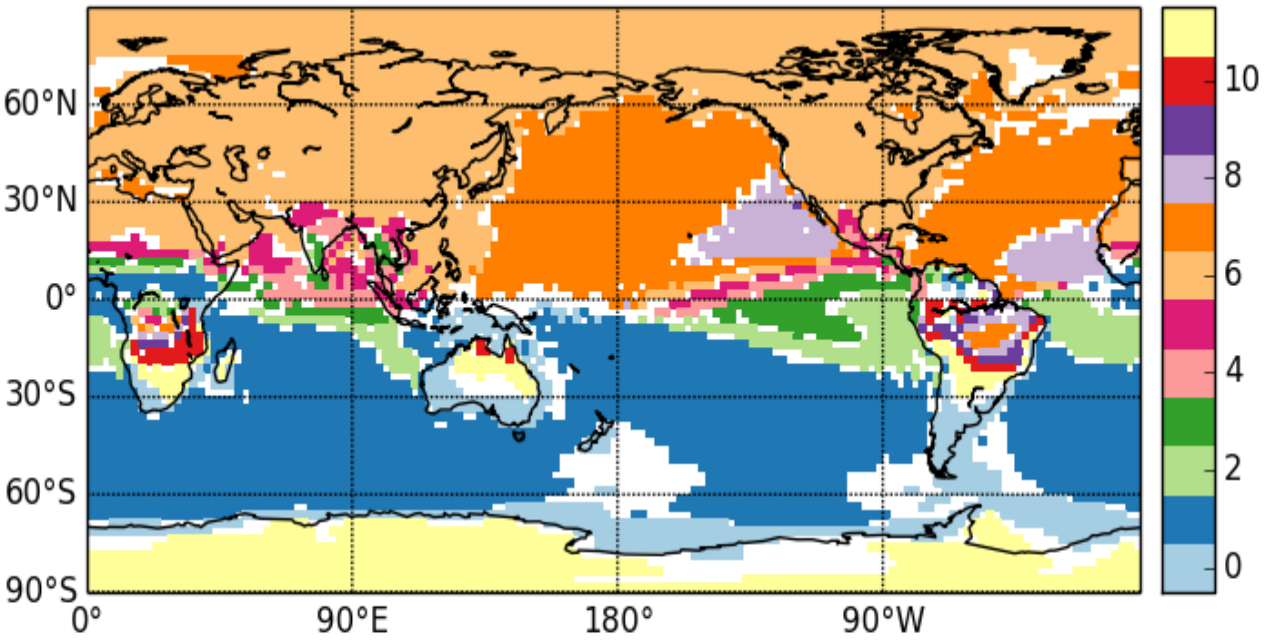
G.T. analysed the data. G.T. and C. M. wrote the manuscript.

## Additional information

**Supplementary information** accompanies this paper at http://www.nature.com/srep; **Competing financial interests** The authors declare no competing financial interests.




(a)




(b)

**Figure 1.** Communities obtained from computing mutual lags among SAT time-series. Regions depicted with the same color have a synchronous seasonal cycle, while the lag between two regions can be computed by subtracting the numbers associated with each color. Panel (a) was obtained by using a reference node located in continental Europe, while panel (b), a reference node in southern South America. Matlab software (version number 7.12.0.635) was used to create these maps, and all the other maps in this work.

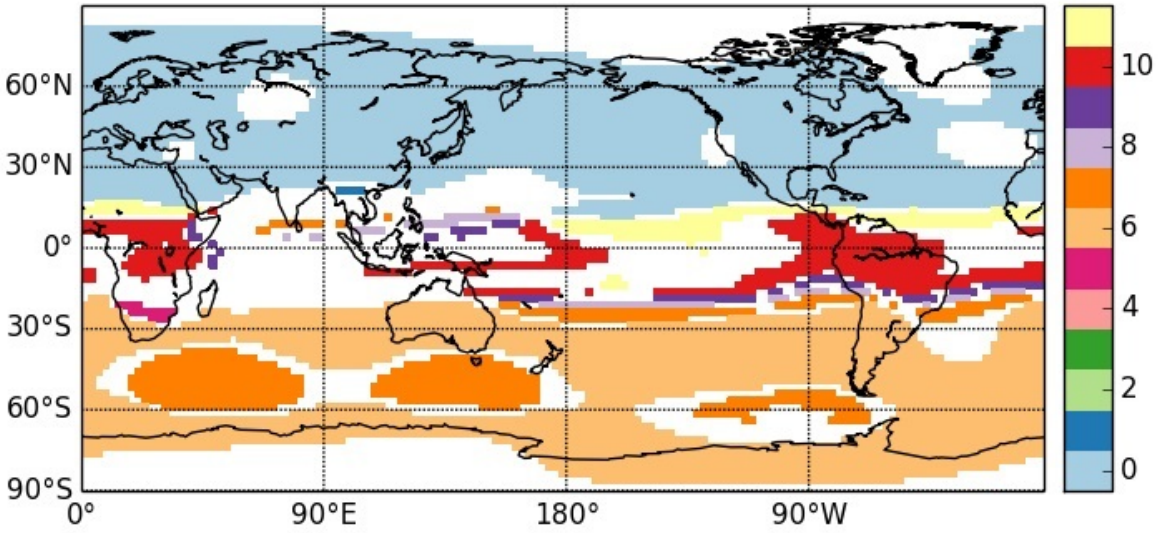



**Figure 2.** As Fig. 1a but computing the lag times from time-series of geopotential height at 500 hPa.

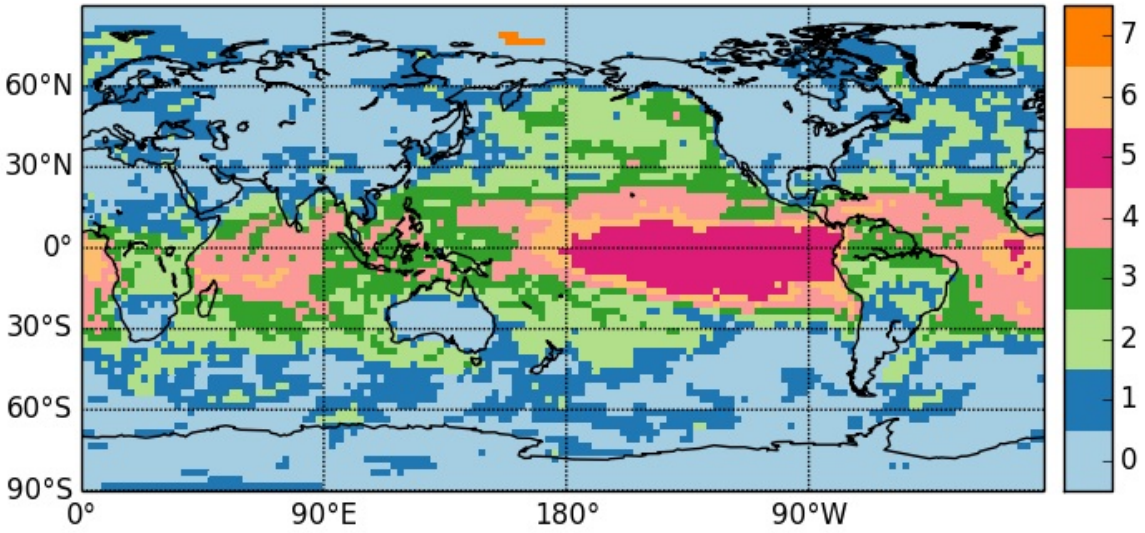
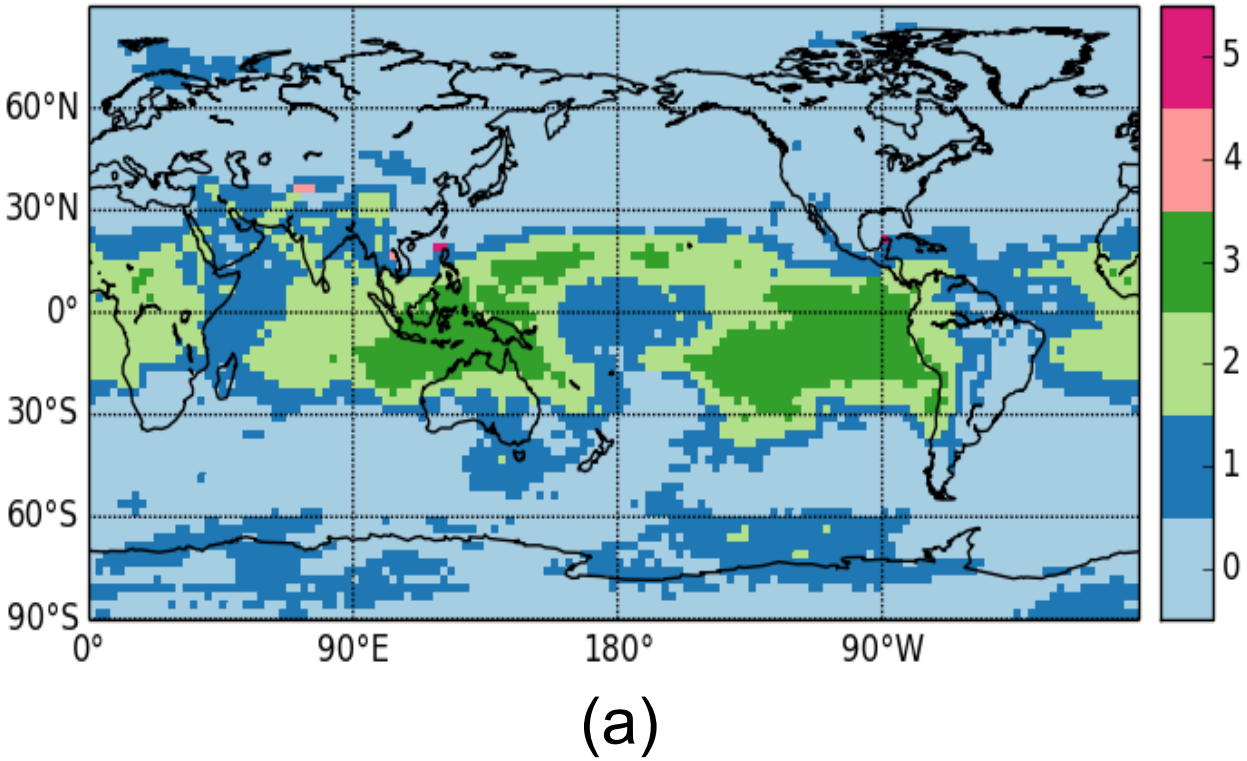
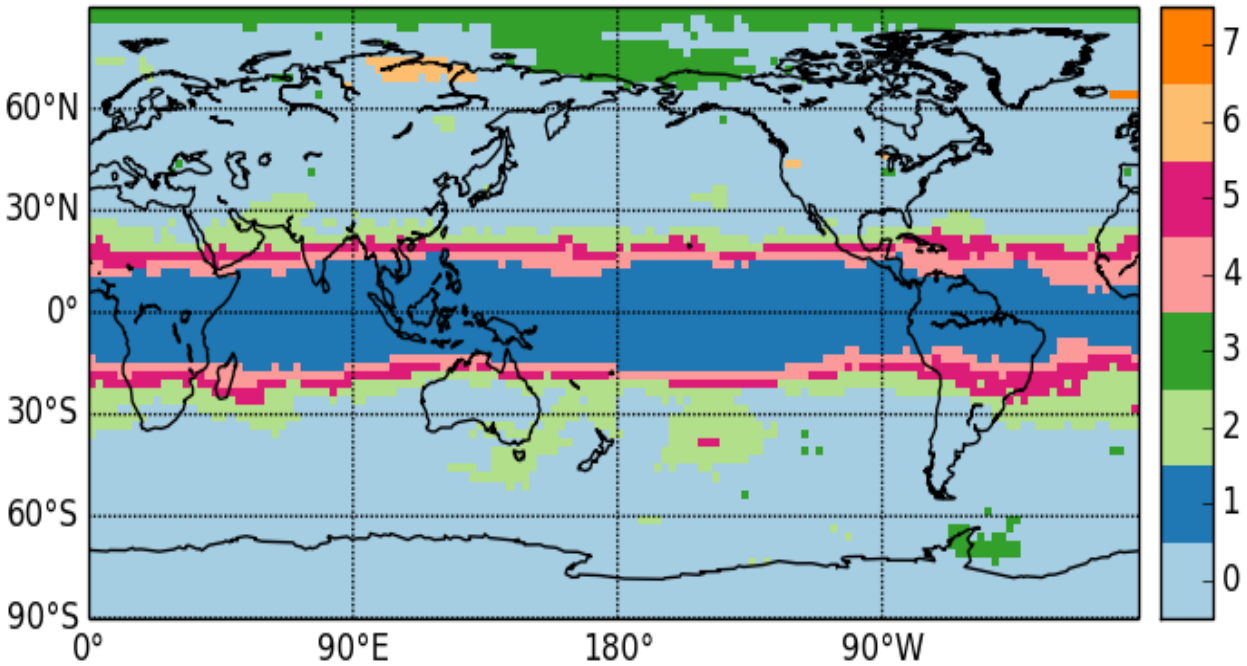



**Figure 3.** Communities obtained from the symbolic analysis of SAT anomalies. Regions depicted with the same color belong to the same community. Four macro-communities are identified: extratropical continents and oceans, tropical oceans and El Niño basin.




(a)




(b)

**Figure 4.** As Fig. 3 but for geopotential height anomalies at 1000 hPa (a) and 300 hPa (b).