Learning a weather dictionary of atmospheric patterns using Latent Dirichlet Allocation

Lucas Fery[1,2], Berengere Dubrulle[3], Berengere Podvin[4], Flavio Pons[1], and Davide Faranda[1,5,6,*]

[1]Laboratoire des Sciences du Climat et de l'Environnement, CEA Saclay l'Orme des Merisiers, UMR 8212 CEA-CNRS-UVSQ, Université Paris-Saclay & IPSL, 91191, Gif-sur-Yvette, France
[2]Department of Physics, Ecole Normale Supérieure de Lyon, 69364, Lyon, France
[3]SPEC, CEA, CNRS, Université Paris-Saclay, F-91191 CEA Saclay, Gif-sur-Yvette, France
[4]LISN, CNRS, Université Paris-Saclay, 91405, Orsay, France
[5]London Mathematical Laboratory, 8 Margravine Gardens, London, W6 8RH, UK
[6]LMD/IPSL, Ecole Normale Superieure, PSL research University, 75005, Paris, France
[*]davide.faranda@lsce.ipsl.fr

ABSTRACT

Mid-latitude circulation dynamics is often described in terms of weather regimes, represented by atmospheric field configurations extracted using pattern recognition techniques. Each pattern is given by a given combination of distinct elements, corresponding to synoptic objects (cyclones and anticyclones). Such intrication makes it arduous to detect or quantify shifts in atmospheric circulation - possibly due to anthropogenic forcings - impacting recurrence and intensity of climate extremes. Here we apply Latent Dirichlet Allocation (LDA), typically used for topic modeling in linguistic studies, to build a weather dictionary: in analogy with linguistics, we define daily maps of a gridded target observable as documents, and the grid-points composing the map as words. LDA provides a representation of documents in terms of a combination of spatial patterns named *motifs*, which are latent patterns inferred from the set of snapshots. For atmospheric data, we find that *motifs* correspond to pure synoptic objects (cyclones and anticyclones), that can be seen as building blocks of weather regimes. We show that LDA weights provide a natural way to characterize the impact of climate change on the recurrence of regimes associated with extreme events.

Supplementary information

Beside the measures based on the area of the motifs, we also define a measure of the spatial standard deviation of all motifs (1) – which characterises how distinct are the motifs from a uniform one – and the average distance between reconstructed and reference maps on all snapshots (2) – which characterises the reprensentativeness of the basis of motifs. We observe that these two metrics seems to converge with an increasing number of motifs (see Fig.S3 and S4), but at larger values of N than the measures based on the area of the motifs.

$$\sigma(\mu) \equiv \frac{1}{l}\sum_{l}\sigma^{l} \quad \text{where} \quad \sigma^{l} \equiv \overline{\left((\mu^{l})^{2} - \overline{\mu}^{l^{2}}\right)}^{1/2} \quad \text{and} \quad \overline{\mu}^{l} \equiv \frac{1}{mn}\sum_{i,j}\mu^{l}_{i,j} \tag{1}$$

$$D(a,\hat{a}) \equiv \frac{1}{S}\sum_{s=1}^{S} d(a^{s},\hat{a}^{s}) \quad \text{with} \quad d(a^{s},\hat{a}^{s}) \equiv \overline{|a^{s}-\hat{a}^{s}|} = \frac{1}{mn}\sum_{i,j}\left|a^{s}_{i,j}-\hat{a}^{s}_{i,j}\right| \tag{2}$$

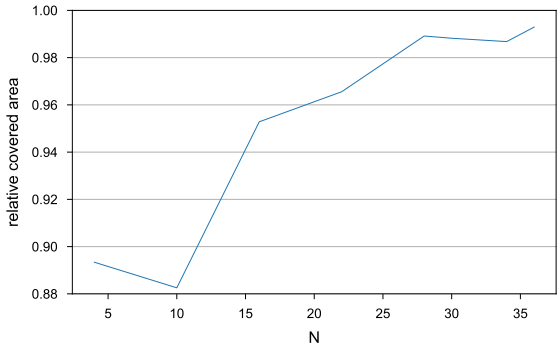

Figure 1. Measure of the area covered by the topics, compared to the total area of the pair of positive and negative anomaly maps. The relative covered area increases as we increase the number of motifs and reaches 99% with N=28. Thus taking more motifs would not improve significantly the coverage of the map by the motifs.

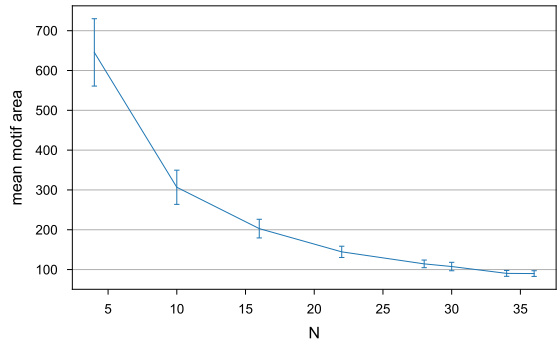

Figure 2. Average area of motifs for different total number N of motifs. The error bars correspond to 2 times the standard error. The average area decreases as we increase the number of motifs and converges around 100 which corresponds to the typical diameter (2000-3000 km) of cyclones and anticyclones.

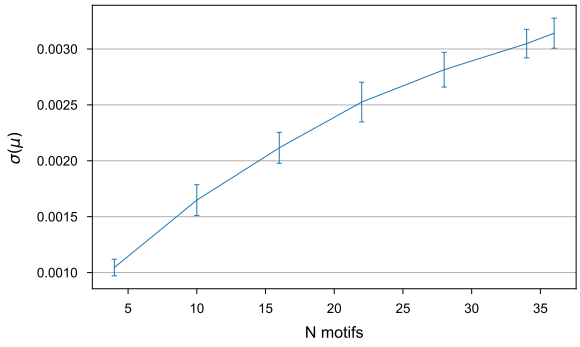

Figure 3. Measure of the spatial standard deviation of motifs for different total number N of motifs. The error bars correspond to 2 times the standard error. It indicates how distinct are the motifs from a uniform one. We observe that the mean increases as N increases, which stemmed from the fact that motifs tends to be more and more localized patterns with a more and more dominant sign (either mostly positive or negative). However, the average value increases more and more slowly. Thus, we expect not to improve significantly the relevance of motifs by increasing N further.

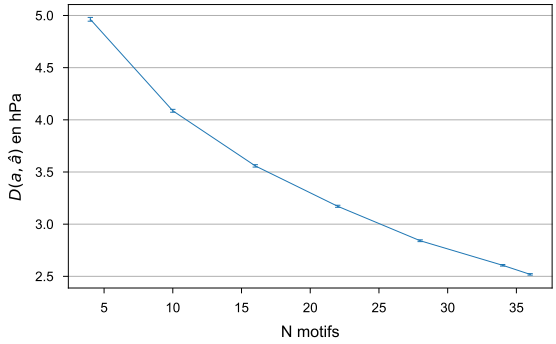

Figure 4. Distance between reconstructed and reference maps for different total number N of motifs. The error bars correspond to 2 times the standard error. It characterizes the accuracy of reconstruction of the reference map, or in other words the reprensentativeness of the basis of motif. We observe that the mean decreases as N increases, which is expected because we can represent more precisely a map with a higher number of degrees of freedom. However, the average value decreases more and more slowly. Thus, we expect not to improve significantly the reconstruction of a map by increasing N further.

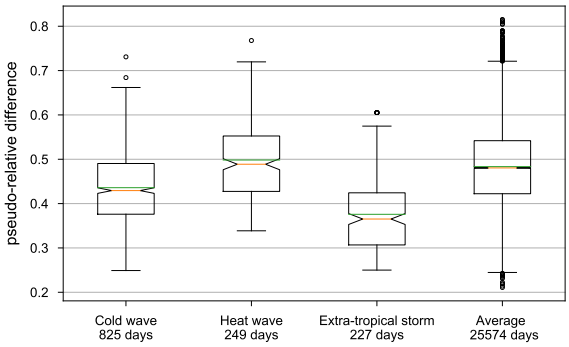

Figure 5. Pseudo-relative difference distribution between reconstructed map and reference map for the considered events in comparison to all the data. By "pseudo" we mean that we divide the distance $d\left(a^s, \hat{a}^s\right)$ by the average value of the reference map $\overline{|a|}^s$, and that we do not compute $d(\mathbb{I}, \hat{a}^s / a^s)$ because a^s can have zero elements. On the one hand, we notice that extra-tropical storms are represented by themotifs significantly better than the average map. On the other hand, the representation of heat waves is slightly worse than on average. Finally, the representation of cold waves are sligthly better than on average. These results can be explained by the fact that extra-tropical storms can be described by few degrees of freedom, and heat waves by more degrees of freedom. Fig. S4-6 highlight these observations if we look at the number of dominant motifs.

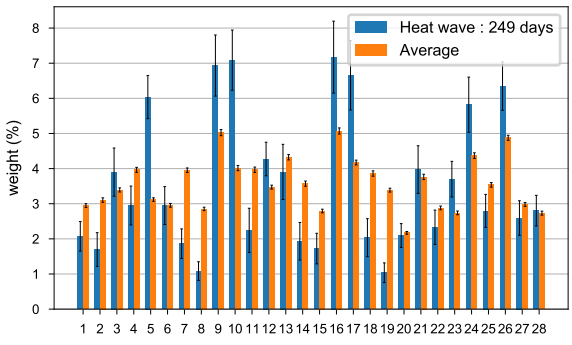

Figure 6. Comparison of average heat waves distribution and average distribution on all the data. The errors bars correspond to 2 times the standard error.

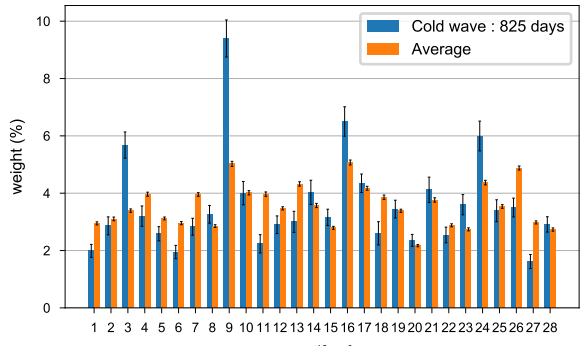

Figure 7. Comparison of average cold waves distribution and average distribution on all the data. The errors bars correspond to 2 times the standard error..

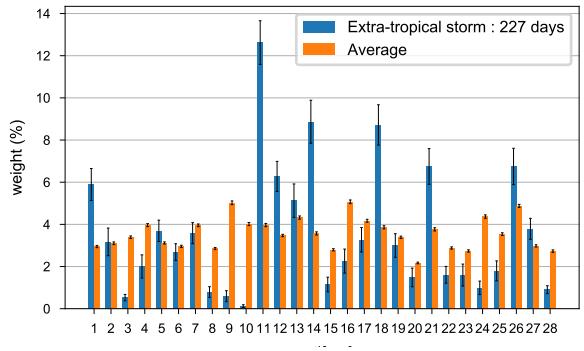

Figure 8. Comparison of average extratropical storms distribution and average distribution on all the data. The errors bars correspond to 2 times the standard error.

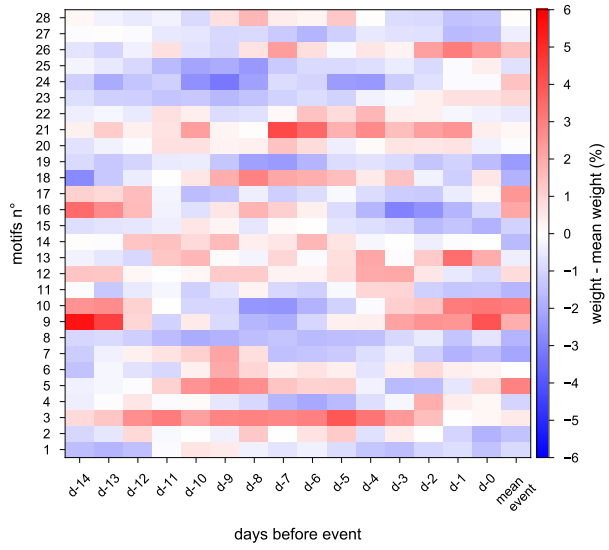

Figure 9. Evolution of the average distribution on motifs for heat waves from 14 days before the event to the first day. The average is made over 28 events. Each column corresponds to a day and each row corresponds to a motif. The color represents the difference of the average weight of the time slot and the average weight over all the data. It is red if the weight is higher than the overall average and blue instead. The mean event (last column) corresponds to the average over all days (249 in total) of the events.

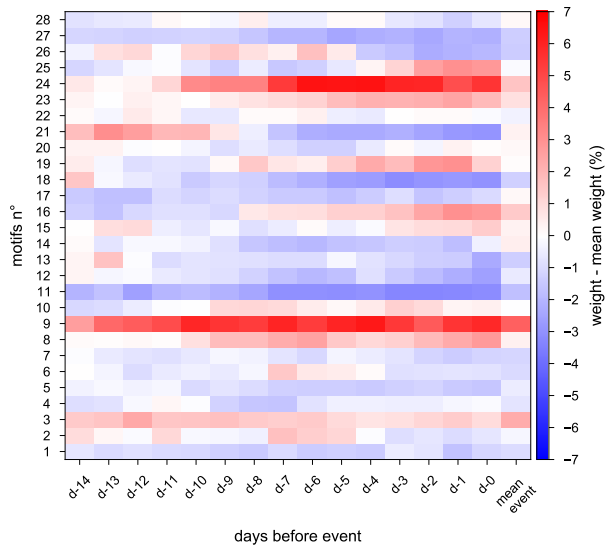

Figure 10. Evolution of the average distribution on motifs for cold waves from 14 days before the event to the first day. The average is made over 71 events. Each column corresponds to a day and each row corresponds to a motif. The color represents the difference of the average weight of the time slot and the average weight over all the data. It is red if the weight is higher than the overall average and blue instead. The mean event (last column) corresponds to the average over all days (825 in total) of the events. We observe that motifs 9 and 24 are still strongly higher than average many days before the beginning of the cold wave, which is in agreement with the typical anticyclonic blocking conditions associated with cold spells.

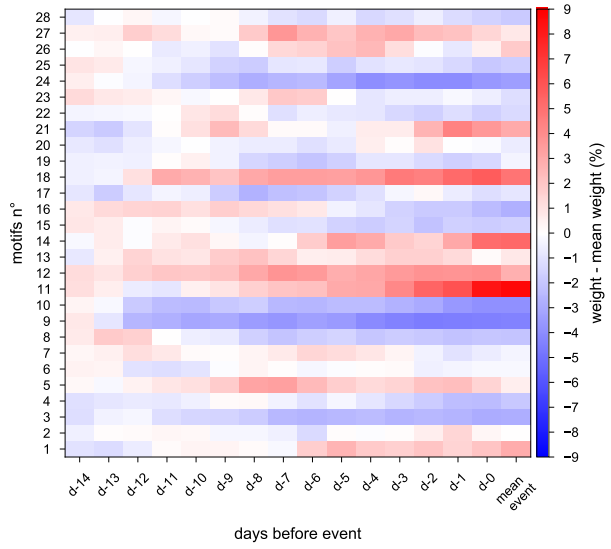

Figure 11. Evolution of the average distribution on motifs for extratropical storms from 14 days before the event to the first day. The average is made over 102 events. Each column corresponds to a day and each row corresponds to a motif. The color represents the difference of the average weight of the time slot and the average weight over all the data. It is red if the weight is higher than the overall average and blue instead. The mean event (last column) corresponds to the average over all days (227 in total) of the events.

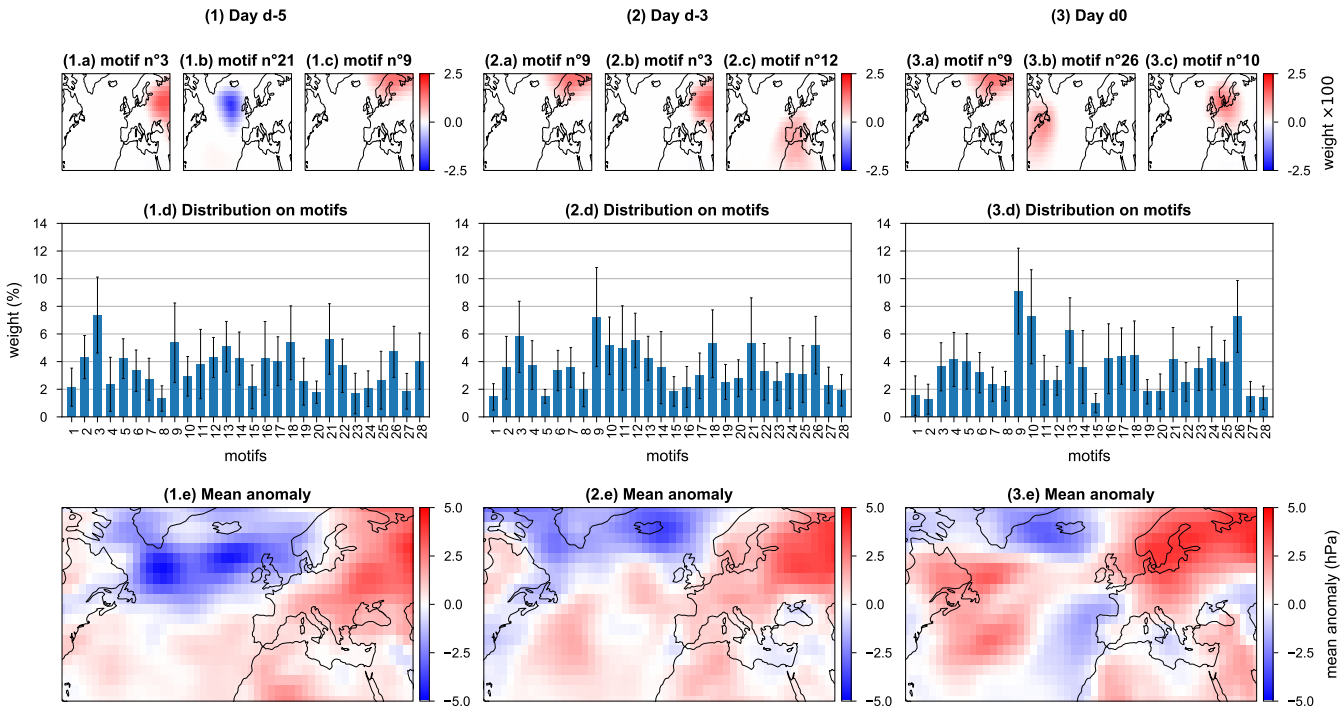

Figure 12. Mean anomaly for heat waves over Europe for five days before the event **(1)**, 3 days before **(2)** and the first day **(3)**. **(a-b-c)** shows the three leading motifs in the LDA representation of the anomaly for the corresponding day. **(d)** shows the whole representation of the anomaly based on a weight assigned to each motif. **(e)** shows the mean anomaly field for the corresponding events in the EM-DAT database. The error bars represent 2 times the standard error.

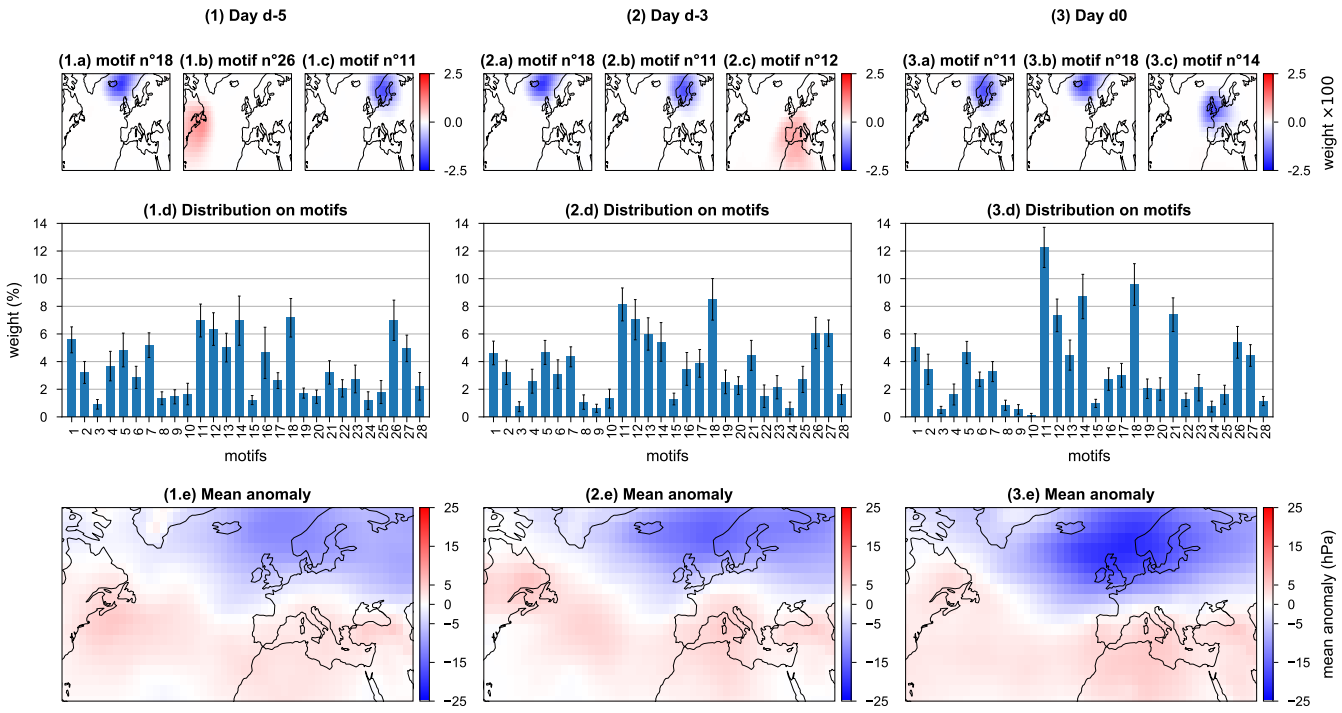

Figure 13. Mean anomaly for extratropical storms over Europe for five days before the event **(1)**, 3 days before **(2)** and the first day **(3)**. **(a-b-c)** shows the three leading motifs in the LDA representation of the anomaly for the corresponding day. **(d)** shows the whole representation of the anomaly based on a weight assigned to each motif. **(e)** shows the mean anomaly field for the corresponding events in the EM-DAT database. The error bars represent 2 times the standard error.