Learning a weather dictionary of atmospheric patterns using Latent Dirichlet Allocation

Lucas Fery, Berengere Dubrulle, Berengere Podvin, Flavio Pons, Davide Faranda

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.

Introduction

Describing mid-latitude atmospheric circulation is challenging because of the turbulent[1] and chaotic[2] nature of the underlying flow, driven by the jet stream[3]. The phase space of general turbulent geophysical flows is usually thought to be very large, with its dimension scaling as the Reynolds number[4]. However, previous work based on dimensionality reduction methods[5] have shown that the mid-latitude dynamics can, in fact, be understood through a limited number of degrees of freedom, such as superpositions of cyclonic and anticyclonic structures in observable fields, like sea-level pressure or geopotential height[6–8]. Indeed, while cyclones and anticyclones normally progress West to East transported by the jet, they can sometimes organize in persistent or intense structures[9]. These structures yield to extreme events such as cold spells, heat waves or extratropical storms, whose frequency can be modified by anthropogenic forcing[10]. At first, the characterization of these patterns was based on atmospheric indices built upon physical arguments. For example, the North Atlantic Oscillation[11] or the Arctic Oscillation have been devised with the goal of separating blockings from the zonal flow. Later, unsupervised pattern recognition based on classification or dimensionality reduction methods, such as k-means[12], empirical orthogonal functions[13] or proper orthogonal decomposition[14], have allowed the detection of patterns (weather regimes[15]) in arbitrary situations. Such patterns can be used to separate modes of natural weather variability from extreme events and, sometimes with open debates, detect the effects of climate change[16,17]. However, those methods provide a field description of patterns, made of a given combination of distinct elements, corresponding to synoptic objects (cyclones and anticyclones) : e.g. the NAO+ pattern includes both the Azores anticyclone and the Icelandic low pressure system. In a changing climate, shifts in these patterns are difficult to detect or analyse, as it is difficult to disentangle changes due to shifts in the synoptic objects properties, or shifts in their relative weights. Similarly, a description based on entire fields fails to identify the genesis and dynamical properties of extreme weather events, as the latter are often associated with specific cyclonic or anticyclonic structures. In addition, we point out that each field can only belong to a single cluster with the k-means method.

In this manuscript, we offer a change of perspective on atmospheric pattern recognition by adapting a machine learning technique used in linguistics to geophysical flows. Recently, several efforts have been made to apply machine learning to the

prediction of geophysical data[18], to learn parameterizations of subgrid processes in climate models[19–27], to the forecasting[28–31] and nowcasting (i.e. extremely short-term forecasting) of weather variables[32–34], and to quantify the uncertainty of deterministic weather prediction[35]. We specifically use a soft clustering technique called Latent Dirichlet Allocation (LDA), a generative probabilistic model usually applied to collections of discrete data[36]. In particular, LDA is widely used to describe *corpora* of text documents: each *document* is assumed to be a mixture of a small number of *topics* which are characterized by a distribution over *words* of a finite vocabulary. Here, we apply LDA to a set of daily sea-level pressure anomaly maps over North-Atlantic from 1948 to 2018, considering grid-points as the equivalent of words and looking for a classification of the pressure anomaly maps in terms of distinct patterns equivalent to the topics, as presented in Fig. 1. We show that LDA is capable to write any sea-level pressure map (*documents*) from the dataset (*corpus*) in terms of well-known cyclonic or anticyclonic structures (*motifs*, which are here the equivalent of *topics*). We choose here to think in terms of *motifs* instead of *topics* as these latent variables correspond to spatial patterns in our case. First, we show that building a weather dictionary requires a number of motifs compatible with the estimate of the number of active degrees of freedom of the gridded fields; this quantity should be ideally known prior to the application of LDA, and thus estimated with different techniques. Then, we show the usefulness of this mid-latitude weather dictionary to detect trends in occurrence of certain motifs caused by anthropogenic forcing and to find common precursors of different class of climate extremes, an important element to assess their predictability.

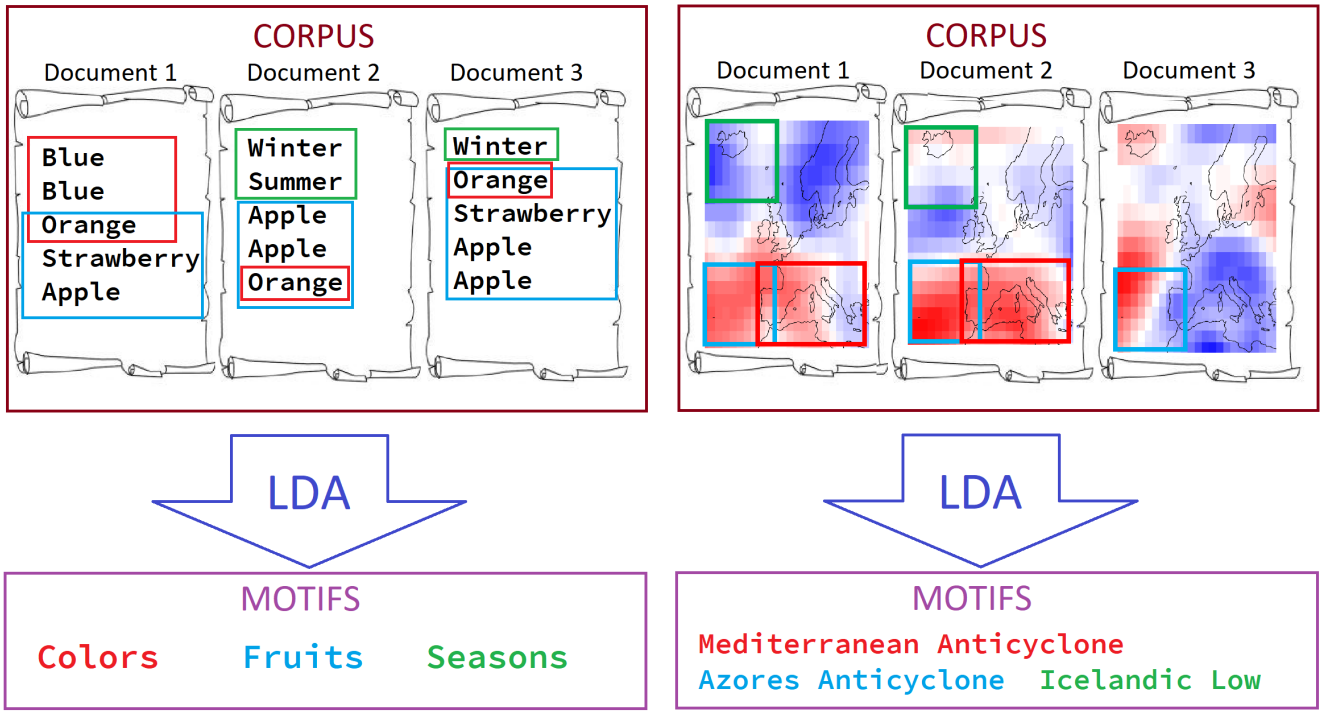

Figure 1. Diagram showing the analogy between the application of LDA to text documents and to a gridded observable map – here sea level pressure anomaly data. The equivalent of words in the latter case are the grid-points of the map and the value of the observable on the grid-points corresponds to the number of occurrences of the words. LDA identifies latent motifs – or topics in the case of text documents – in a corpus of documents that are defined by a distribution over the words of a finite vocabulary (see Methods for a detailed explanation), and then interpreted as meaningful patterns (e.g. "Colors", "Fruits" and "Seasons" for the text documents or "Mediterranean Anticyclone", "Azores Anticyclone" and "Icelandic low" for sea level pressure maps). Each word of the vocabulary, or equivalently each grid-point, can be associated with different motifs as represented by the word "orange" that can be seen as a fruit or a color.

Results

When applying LDA, one first needs to select the number of motifs that will be necessary to describe the whole dataset. Other pattern recognition methods allow to chose the optimal number of components, if this is not know a priori: for example, the scree plot defined by deviance within groups for k-means, or the proportion of explained variance for PCA. For LDA, we propose two measures : the relative area covered by the motifs and the motifs average area (see eqs. (6) and (7) in Methods).

We observe that these two metrics converge with an increasing number of motifs (see Fig.S1 and S2). Moreover, the length scale computed from the average area corresponds to the typical diameter of cyclones and anticyclones. This preliminary inspection suggests that taking 28 motifs is sufficient to get relevant patterns which are representative of the dynamics of mid-latitude circulation. This value is in very good agreement with the upper bound of the number of degrees of freedom obtained computing the local attractor dimensions for the same dataset[37]. This suggests that each motif corresponds to an active degree of freedom in the maps, reinforcing the claim that cyclones and anticyclones are the building blocks of the dynamics of mid-latitude flows. This may also indicate that LDA could be a method more suitable than others to define the modes of variability associated with the number of active degrees of freedom estimated by the local dimension. The latter can indeed provide a suggestion for optimal dimensionality reduction, but does not provide a method to project the dataset on the reduced dimension. The 28 motifs obtained by training the LDA model on the daily sea-level pressure anomaly maps over North-Atlantic from 1948 to 2018 are presented in Fig.2 (see Methods for more details). First, we observe that they consist of localized anomaly patterns which are mostly exclusively positive or negative, with a radius between 1000 and 2000 km i.e. they truly correspond to cyclonic or anticyclonic anomalies. Furthermore, they reproduce the typical locations of pressure structures previously named by meteorologists studying North Atlantic circulation, e.g. the Genoa low (motif 8), the Icelandic low (motif 18), the Siberian high (motif 3) or the Azores anticyclone (motif 7). This strengthens our confidence that the motifs are not only a practical way to represent a complex map, but also that they correspond to actual coherent sea level pressure patterns.

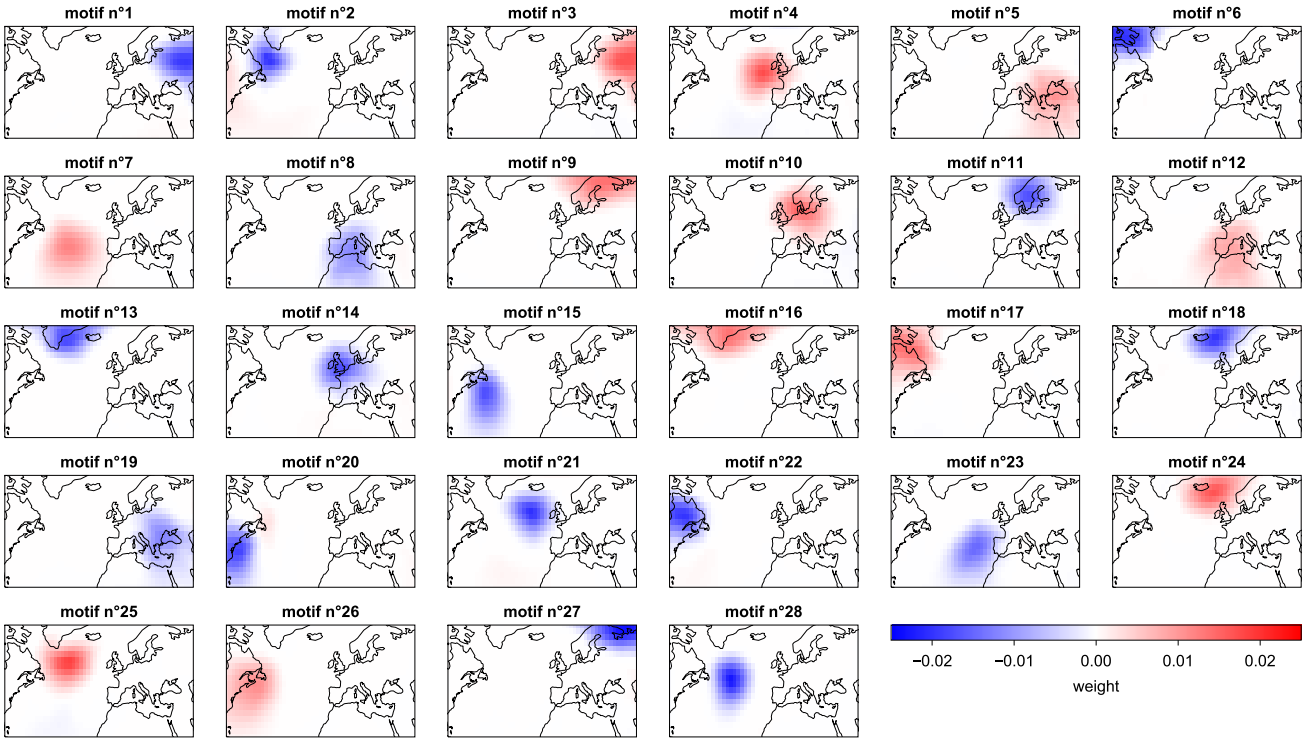

Figure 2. The 28 motifs identified by LDA by using as documents the daily sea-level pressure anomaly maps over North-Atlantic from 1948 to 2018. Positive anomalies and negative anomalies are respectively represented in red and blue. LDA outputs normalized topics, so the color intensity only represents a relative weight.

Internal variability and climate change

As each daily anomaly is represented by a mixture of motifs with given weights, we can study the evolution over years of the weight for each motif. The average values of the motifs weights for yearly intervals are presented in Fig.3. Most of the motifs show oscillations of weights of the order of few percents with signs alternating with interdecadal frequency, coherently with the claim that chaotic behaviour of mid-latitude circulation at climate scales dominate. On the contrary, motifs 5, 8, 12 and 19 stand out with a visible coherent evolution over time. They correspond to pressure structures over Mediterranean sea. Anticyclonic anomalies associated with motifs 5 and 12 have been already linked to heatwaves on the Mediterranean basin[38,39]. During the analysed period, there is an increase of high pressure patterns average weights and a corresponding decrease of low pressure ones. The augmented weight of motifs 5 and 10 could be explained via the strengthening of subtropical anticyclones

to higher latitudes and a corresponding weakening low pressure systems (motifs 8 and 19). This phenomenon, investigated in[40] at the global scale, has received little attention for the Mediterranean basin, although few signals are coherent with our findings, namely the fact that this area will be warmer and dryer[41], and that heatwaves are projected to increase[42,43].

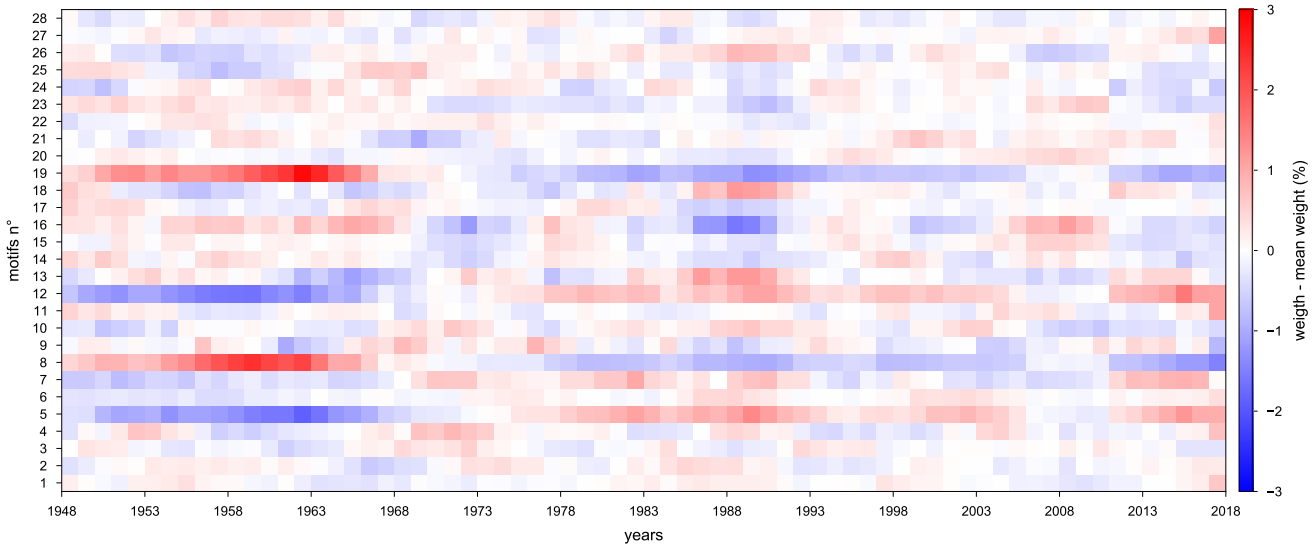

Figure 3. Evolution of mean motifs weights over the years. Each column corresponds to a time slot of one year from 1948 to 2018 and each row corresponds to a motif. The color represents the difference of the average weight of the time slot and the average weight over the whole interval. It is red if the weight is higher than the overall average and blue instead. We notice that at least four motifs differentiate themselves by a coherent evolution over time (motifs 5, 8, 12 and 19). They correspond to structures on Mediterranean sea, with an increase of high pressure patterns average weights and a decrease of low pressure ones.

Extreme events

We also investigate the distribution on motifs for particular events such as heat waves, cold waves and extra-tropical storms which are associated with large scale pressure anomaly patterns. We use the EM-DAT[44] database to identify these events and analyse their average properties in the motifs representation. This analysis has two main goals: first identify the relevant motifs for each class of extreme events and then verify whether some of the motifs are precursors of those extreme events. Note that the extreme events identified by EM-DAT are defined on the basis of their impacts and not by applying extreme value theory or statistical analysis of atmospheric variables. It is therefore interesting to see whether such extremes are associated to specific motifs[45]. Fig.4, S12 and S13 each show the average anomaly map **(e)** and motifs weights **(d)** along with the corresponding three leading motifs **(a,b,c)** for 5 days before the beginning of the event **(1)**, 3 days before **(2)**, and on the first day of the event as reported in the EM-DAT databases **(3)** for the 3 types of events mentioned before. The distributions on the motifs for the first day **(3)** highlight the fact that we can identify dominant motifs that are consistent with each type of event. Indeed, we find positive anomalies i.e. high pressures for heat waves and cold waves, and negative anomalies i.e. low pressure for extra-tropical storms, as expected. What's more, this representation might shed light on teleconnections between different patterns[46,47]. As an example Fig.S12 shows that heat waves on Europe tend to be associated with high pressure over Canada (motif 26). This teleconnection is a robust feature already identified with other techniques such as importance sampling[48]. We also notice that LDA is capable of identifying motifs that act as precursors for these events. Let us analyse in details the case of cold-spells (Fig.4 and S10), for which motifs 9 and 24 are precursors several days in advance. These motifs show that cold spells over Europe are generally associated with anticyclonic blocking conditions located to the North of continental Europe, consistently with previous results[49]. Looking at 5 days precursors of heat waves (Fig. S9 and S12), a cyclonic mid-Atlantic pattern appears. This is consistent with previous analysis[50] showing that a wavy pattern of wave number 7 with alternating cyclonic-anticyclonic structures favor heatwaves over Europe[51].

Discussion

We have discussed the application of an unsupervised pattern recognition method based on Latent Dirichlet Allocation (LDA), a generative probabilistic model for collections of discrete data. The method has been applied to a set of snapshots featuring daily

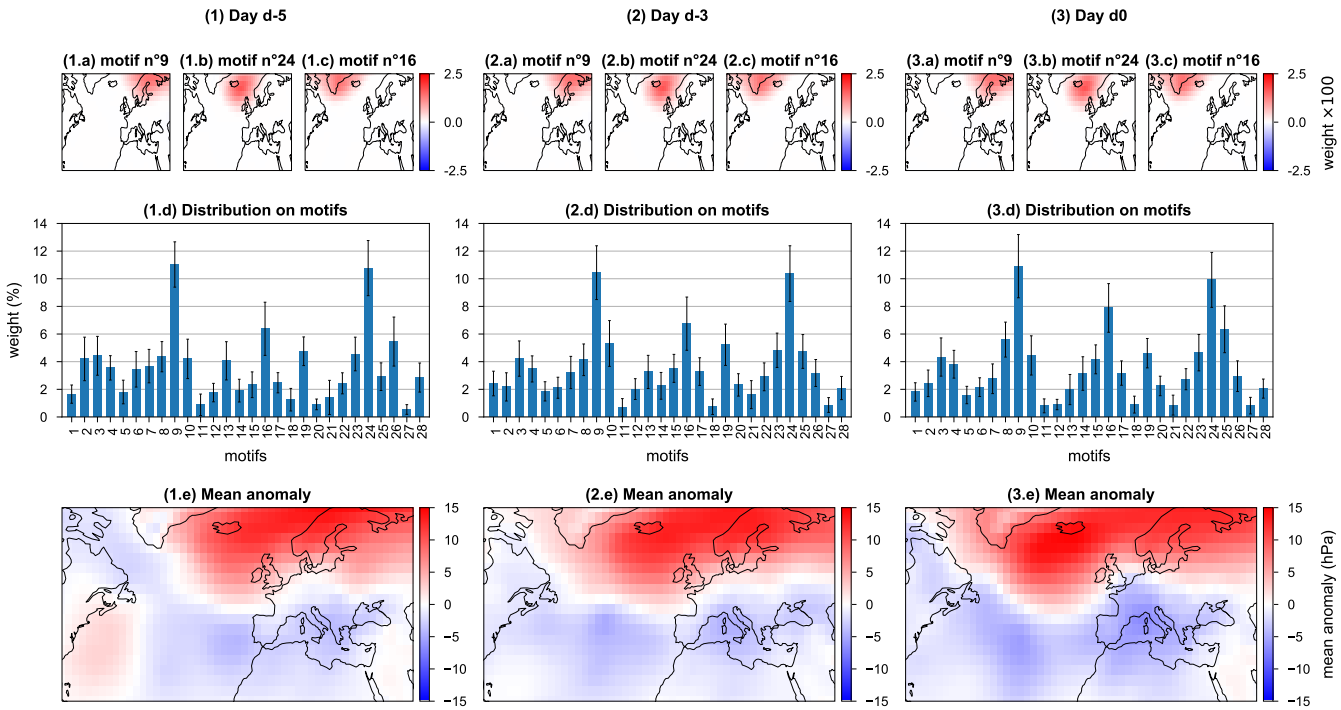

Figure 4. Mean anomaly for cold waves over Europe. The anomaly is presented for five days before the event **(1)**, 3 days before **(2)** and on the first day of the event as identified in EM-DAT **(3)**. **(a-b-c)** shows the three leading motifs in the LDA representation of the anomaly for the corresponding day. **(d)** shows the whole representation of the anomaly based on a weight assigned to each motif. **(e)** shows the mean anomaly map the corresponding events in the EM-DAT database. The error bars represent 2 times the standard error.

sea level pressure anomaly over North-Atlantic from 1948 to 2018. The main advantage of LDA is the representation of a map of a gridded observable in terms of a combination of motifs, which are latent spatial elements inferred from the set of snapshots. We have shown that the motifs correspond to intelligible localized patterns in contrast with other techniques[12–14] like those leading to the definition of weather regimes[15]. Indeed motifs correspond to single cyclonic or anticyclonic patterns and are, therefore, of immediate interpretation for meteorologists and climatologists. Using motifs as a weather dictionary, we have been able to naturally address questions about climate change and precursors of weather extreme events which have required advanced analyses techniques in other studies[8,9,39]. Our illustrative analysis of the capabilities of LDA for climate science has shown that subtle changes in the frequency or intensity of cyclonic structures over the Mediterranean can be distinguished from the general interdecadal variability of the chaotic atmospheric circulation. Furthermore, we have determined that only a few number of motifs are relevant to describe weather extreme events such as cold spells, heat waves or extratropical storms even if these have been defined only on the bases of their impacts in the EM-DAT database. This suggests a partially positive answer to the question raised by van der Wiel[45] on whether extreme events defined by their impacts can be connected to the underlying climate dynamics: the connection exists but with specific motifs rather than with maxima or minima of meteorological variables. We have also been able to identify precursors of these events through motifs and explain their origin, consistently with previous results[49–51].

Our study paves the way for several exciting applications of LDA in the context of present and future climates, namely i) a comprehensive study of precursors of extreme events, whose existence hints to a possibly unexploited predictive power of LDA ii) the intercomparison of climate models using as skill score their ability of representing specific motifs, e.g. those relevant for weather extreme events, iii) the analysis of the emergence or disappearance of motifs in different climates conditions.

Methods

We use daily sea level pressure data from NCEP/NCAR reanalysis[52] over the period 01/01/1948 – 06/01/2018 (S = 25,574 days). The data is mapped on a $(m = 20, n = 53)$ grid of resolution $2.5^\circ \times 2.5^\circ$ representing a region from 80W to 50E and 22.5N to

70N which corresponds to North Atlantic and Europe. We compute the daily seasonal standards of sea level pressure based on these data for every day of a year and then the daily anomaly by subtracting it to the original map. We then split each anomaly map into a positive anomaly map and a negative one. which is equivalent to doubling the size of the grid (or grid-points). The anomaly map constitute a pair of 2D arrays where each element of the (m,n) array (or grid-point) is characterized by a positive value. In the LDA framework, each grid-point is considered to be a word of a vocabulary. As already done by Frihat[53] for the identification of coherent structures in turbulent channel flow, the positive value of each element is assimilated as the number of occurrences of the word found in each document, i.e. the anomaly map, to within a potential rescaling factor and a digitization. Here we took a rescaling factor of 1 for the raw sea level pressure anomaly in Pa and we also put a threshold value of 100 Pa to consider only significant anomalies and to speed-up the training of the model by reducing the word count of documents. We train the LDA model with its implementation in the Gensim Python library[54]. The LDA model takes several parameters as arguments, such as the number of *passes* and *iterations* (we refer to Gensim LDA documentation for a definition of these parameters). These have been chosen high enough so that the *perplexity*[36] – a quantity defined to evaluate the performance of the model – converges. Here we set the number of *iterations* as 2000, the number of *passes* as 20, the *chunksize* as 2048, *alpha* and *eta* as 'auto', *random state* as 100 to get reproducible results and other parameters to defaults values. The number of *topics* N is also a parameter which has to be set before training.

The program produces two arrays corresponding in the text mining applications to the *document-topic distributions* w (1) and the *topic-word distributions* t (2) which are normalized probability distributions. Thus, each document – here corresponding to a daily anomaly map $(a^s_{i,j})_{i,j}$ with $s \in [1,S]$ denoting the snapshot in the dataset and $i \in [1,m]$ and $j \in [1,n]$ indicating the rows and columns of the map – is associated with a combination of *topics* with given weights w^s_l where $l \in [1,N]$ indicates the corresponding *topic*. Each *topic* t^l is defined the same way as a mixture of words of the vocabulary – i.e. grid-points of the positive and negative anomaly maps in our case – with given weights $t^l_{i,j,k}$ where $k \in \{+,-\}$ indicates the positive or negative map. We can thus think of topics as a normalized distribution on a set of two maps. We then combine the positive and negative maps of each *topic* to a single map which is interpreted as a spatial anomaly pattern that we call motif, which we represent by an array μ (3). From the *document-topic distribution* and the motifs, we can reconstruct the anomaly map to within a rescaling factor. To be able to compare the value of the reference anomaly a^s and the reconstructed one $\hat{a}^s$, we also multiply the combination of motifs by the sum of the reference absolute anomaly values on the whole map (4).

$$w = (w^s_l)^{s\in[1,S]}_{l\in[1,N]} \quad \text{with} \quad \forall (s,l), w^s_l \in [0,1] \quad \text{and} \quad \forall s, \sum_l w^s_l = 1 \tag{1}$$

$$t = \left(t^l_{i,j,k}\right)^{l\in[1,N]}_{i\in[1,m],\ j\in[1,n],\ k\in\{+,-\}} \quad \text{with} \quad \forall (l,i,j,k), t^l_{i,j,k} \in [0,1] \quad \text{and} \quad \forall l, \sum_{i,j,k} t^l_{i,j,k} = 1 \tag{2}$$

$$\mu = \left(\mu^l_{i,j}\right)^{l\in[1,N]}_{i\in[1,m],\ j\in[1,n]} \equiv \left(t^l_{i,j,+} - t^l_{i,j,-}\right)^{l\in[1,N]}_{i\in[1,m],\ j\in[1,n]} \tag{3}$$

$$\hat{a}^s \equiv \left(\sum_{i,j} |a^s_{i,j}|\right)\left(\sum_l w^s_l \mu^l\right) \tag{4}$$

As we want to find an optimal number of motifs, we consider the following indicators : the relative area covered by the motifs (7) and the motifs average area (6). To do so, we define the area A^l of a motif l by the number of elements of t^l (or grid-points of the pair of maps) whose values are above a threshold, parametrized by a coefficient $\alpha \in [0,1]$ (5). Here we take $\alpha = 2/3$ to get values of area close to the value one would get by counting the colored pixels on the maps of motifs in Fig. 2. On the one hand, Figure S1 shows that $r_c(\mu)$ increases and reaches about 99% at $N = 28$. Thus, we choose to consider $N = 28$ motifs for further analysis as taking more motifs would not improve significantly the coverage of the map by the motifs. Moreover, Figure S2 shows that $A(\mu)$ decreases and converges near 100 grid-points which corresponds to the typical diameter (2000-3000 km) of cyclones and anticyclones[55]. This choice is further motivated by an independent analysis on the same dataset with a dynamical system angle[37]. Indeed, this value of N is in very good agreement with the upper bound of the number of active degrees of freedom obtained computing the local attractor dimensions.

$$A^l = \text{card}\left\{ t^l_{i,j,k} \,\middle|\, t^l_{i,j,k} > \max_{i,j,k}\{t^l_{i,j,k}\} - \alpha\left(\max_{i,j,k}\{t^l_{i,j,k}\} - \min_{i,j,k}\{t^l_{i,j,k}\}\right)\right\} \quad \text{and} \quad \alpha \in [0,1] \tag{5}$$

$$A(\mu) = \frac{1}{l}\sum_{l} A^{l} \tag{6}$$

$$r_c(\mu) = \frac{A_c(\mu)}{2mn} = \frac{1}{2mn}\mathrm{card}\left\{(i,j,k)\,\middle|\,\exists l \in [1,N],\, t^l_{i,j,k} > \max_{i,j,k}\{t^l_{i,j,k}\} - \alpha\left(\max_{i,j,k}\{t^l_{i,j,k}\} - \min_{i,j,k}\{t^l_{i,j,k}\}\right)\right\} \tag{7}$$

References

1. Marino, R., Pouquet, A. & Rosenberg, D. Resolving the paradox of oceanic large-scale balance and small-scale mixing. *Phys. review letters* **114**, 114504 (2015).
2. Lorenz, E. N. Deterministic nonperiodic flow. *J. atmospheric sciences* **20**, 130–141 (1963).
3. Hilborn, R. C. Sea gulls, butterflies, and grasshoppers: A brief history of the butterfly effect in nonlinear dynamics. *Am. J. Phys.* **72**, 425–427 (2004).
4. Dubrulle, B., Daviaud, F., Faranda, D., Marié, L. & Saint-Michel, B. Lewis fry richardson medal lecture – how many modes are needed to predict climate bifurcations ? : Lessons from an experiment. *Nonlin. Process. Geophys. Discuss.* (2021). Preprint at https://doi.org/10.5194/npg-2021-19.
5. Gámez, A. J., Zhou, C., Timmermann, A. & Kurths, J. Nonlinear dimensionality reduction in climate data. *Nonlinear Process. Geophys.* **11**, 393–398 (2004).
6. Rivière, G., Arbogast, P., Lapeyre, G. & Maynard, K. A potential vorticity perspective on the motion of a mid-latitude winter storm. *Geophys. Res. Lett.* **39** (2012).
7. Shepherd, T. G. Atmospheric circulation as a source of uncertainty in climate change projections. *Nat. Geosci.* **7**, 703 (2014).
8. Cassou, C. & Cattiaux, J. Disruption of the european climate seasonal clock in a warming world. *Nat. Clim. Chang.* **6**, 589 (2016).
9. Ferranti, L., Corti, S. & Janousek, M. Flow-dependent verification of the ecmwf ensemble over the euro-atlantic sector. *Q. J. Royal Meteorol. Soc.* **141**, 916–924 (2015).
10. Rummukainen, M. Changes in climate and weather extremes in the 21st century. *Wiley Interdiscip. Rev. Clim. Chang.* **3**, 115–129 (2012).
11. Folland, C. K. *et al.* The summer north atlantic oscillation: past, present, and future. *J. Clim.* **22**, 1082–1103 (2009).
12. Vrac, M. & Yiou, P. Weather regimes designed for local precipitation modeling: Application to the mediterranean basin. *J. Geophys. Res. Atmospheres* **115** (2010).
13. Zampieri, M., Toreti, A., Schindler, A., Scoccimarro, E. & Gualdi, S. Atlantic multi-decadal oscillation influence on weather regimes over europe and the mediterranean in spring and summer. *Glob. Planet. Chang.* **151**, 92–100 (2017).
14. Wang, Z., Akhtar, I., Borggaard, J. & Iliescu, T. Proper orthogonal decomposition closure models for turbulent flows: a numerical comparison. *Comput. Methods Appl. Mech. Eng.* **237**, 10–26 (2012).
15. Vautard, R. Multiple weather regimes over the north atlantic: Analysis of precursors and successors. *Mon. weather review* **118**, 2056–2081 (1990).
16. Cohen, J. *et al.* Recent arctic amplification and extreme mid-latitude weather. *Nat. geoscience* **7**, 627 (2014).
17. Screen, J. A. The missing northern european winter cooling response to arctic sea ice loss. *Nat. communications* **8**, 14603 (2017).
18. Wu, J.-L., Xiao, H. & Paterson, E. Physics-informed machine learning approach for augmenting turbulence models: A comprehensive framework. *Phys. Rev. Fluids* **3**, 074602 (2018).
19. Krasnopolsky, V. M., Fox-Rabinovitz, M. S. & Chalikov, D. V. New approach to calculation of atmospheric model physics: Accurate and fast neural network emulation of longwave radiation in a climate model. *Mon. Weather. Rev.* **133**, 1370–1383 (2005).
20. Krasnopolsky, V. M. & Fox-Rabinovitz, M. S. Complex hybrid models combining deterministic and machine learning components for numerical climate modeling and weather prediction. *Neural Networks* **19**, 122–134 (2006).

21. Rasp, S., Pritchard, M. S. & Gentine, P. Deep learning to represent subgrid processes in climate models. *Proc. Natl. Acad. Sci.* **115**, 9684–9689 (2018).

22. Gentine, P., Pritchard, M., Rasp, S., Reinaudi, G. & Yacalis, G. Could machine learning break the convection parameterization deadlock? *Geophys. Res. Lett.* **45**, 5742–5751 (2018).

23. Brenowitz, N. D. & Bretherton, C. S. Prognostic validation of a neural network unified physics parameterization. *Geophys. Res. Lett.* **45**, 6289–6298 (2018).

24. Brenowitz, N. D. & Bretherton, C. S. Spatially extended tests of a neural network parametrization trained by coarse-graining. *J. Adv. Model. Earth Syst.* **11**, 2728–2744 (2019).

25. Yuval, J. & O'Gorman, P. A. Stable machine-learning parameterization of subgrid processes for climate modeling at a range of resolutions. *Nat. communications* **11**, 1–10 (2020).

26. Gettelman, A. *et al.* Machine learning the warm rain process. *J. Adv. Model. Earth Syst.* **13**, e2020MS002268 (2021).

27. Krasnopolsky, V. M., Fox-Rabinovitz, M. S. & Belochitski, A. A. Using ensemble of neural networks to learn stochastic convection parameterizations for climate and numerical weather prediction models from data simulated by a cloud resolving model. *Adv. Artif. Neural Syst.* **2013** (2013).

28. Liu, J. N., Hu, Y., He, Y., Chan, P. W. & Lai, L. Deep neural network modeling for big data weather forecasting. In *Information Granularity, Big Data, and Computational Intelligence*, 389–408 (Springer, 2015).

29. Grover, A., Kapoor, A. & Horvitz, E. A deep hybrid model for weather forecasting. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 379–386 (ACM, 2015).

30. Haupt, S. E. *et al.* Machine learning for applied weather prediction. In *2018 IEEE 14th International Conference on e-Science (e-Science)*, 276–277 (IEEE, 2018).

31. Weyn, J. A., Durran, D. R. & Caruana, R. Can machines learn to predict weather? using deep learning to predict gridded 500-hpa geopotential height from historical weather data. *J. Adv. Model. Earth Syst.* **11**, 2680–2693 (2019).

32. Xingjian, S. *et al.* Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, 802–810 (2015).

33. Shi, X. *et al.* Deep learning for precipitation nowcasting: A benchmark and a new model. In *Advances in neural information processing systems*, 5617–5627 (2017).

34. Sprenger, M., Schemm, S., Oechslin, R. & Jenkner, J. Nowcasting foehn wind events using the adaboost machine learning algorithm. *Weather. Forecast.* **32**, 1079–1099 (2017).

35. Scher, S. & Messori, G. Predicting weather forecast uncertainty with machine learning. *Q. J. Royal Meteorol. Soc.* (2018).

36. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent dirichlet allocation. *J. machine Learn. research* **3**, 993–1022 (2003).

37. Faranda, D., Messori, G. & Yiou, P. Dynamical proxies of north atlantic predictability and extremes. *Sci. reports* **7**, 1–10 (2017).

38. Stefanon, M., D'Andrea, F. & Drobinski, P. Heatwave classification over europe and the mediterranean region. *Environ. Res. Lett.* **7**, 014023 (2012).

39. Alvarez-Castro, M. C., Faranda, D. & Yiou, P. Atmospheric dynamics leading to west european summer hot temperatures since 1851. *Complexity* **2018** (2018).

40. Li, W., Li, L., Ting, M. & Liu, Y. Intensification of northern hemisphere subtropical highs in a warming climate. *Nat. Geosci.* **5**, 830–834 (2012).

41. Drobinski, P. *et al.* How warmer and drier will the mediterranean region be at the end of the twenty-first century? *Reg. Environ. Chang.* **20**, 1–12 (2020).

42. Bador, M. *et al.* Future summer mega-heatwave and record-breaking temperatures in a warmer france climate. *Environ. Res. Lett.* **12**, 074025 (2017).

43. Perkins-Kirkpatrick, S. & Gibson, P. Changes in regional heatwave characteristics as a function of increasing global temperature. *Sci. Reports* **7**, 1–12 (2017).

44. CRED, D. G.-S. Em-dat: The emergency events database. www.emdat.be. Université catholique de Louvain (UCL) , Brussels, Belgium.

45. van der Wiel, K., Selten, F. M., Bintanja, R., Blackport, R. & Screen, J. A. Ensemble climate-impact modelling: extreme impacts from moderate meteorological conditions. *Environ. Res. Lett.* **15**, 034050 (2020).

46. Agarwal, A. *et al.* Network-based identification and characterization of teleconnections on different scales. *Sci. reports* **9**, 1–12 (2019).

47. Wazneh, H., Gachon, P., Laprise, R., de Vernal, A. & Tremblay, B. Atmospheric blocking events in the north atlantic: trends and links to climate anomalies and teleconnections. *Clim. Dyn.* 1–23 (2021).

48. Ragone, F., Wouters, J. & Bouchet, F. Computation of extreme heat waves in climate models using a large deviation algorithm. *Proc. Natl. Acad. Sci.* **115**, 24–29 (2018).

49. Raymond, F., Ullmann, A., Camberlin, P., Oueslati, B. & Drobinski, P. Atmospheric conditions and weather regimes associated with extreme winter dry spells over the mediterranean basin. *Clim. Dyn.* **50**, 4437–4453 (2018).

50. Kornhuber, K. *et al.* Extreme weather events in early summer 2018 connected by a recurrent hemispheric wave-7 pattern. *Environ. Res. Lett.* **14**, 054002 (2019).

51. Petoukhov, V., Rahmstorf, S., Petri, S. & Schellnhuber, H. J. Quasiresonant amplification of planetary waves and recent northern hemisphere weather extremes. *Proc. Natl. Acad. Sci.* **110**, 5336–5341 (2013).

52. Kalnay, E. *et al.* The ncep/ncar 40-year reanalysis project. *Bull. Am. meteorological Soc.* **77**, 437–472 (1996).

53. Frihat, M., Podvin, B., Mathelin, L., Fraigneau, Y. & Yvon, F. Coherent structure identification in turbulent channel flow using latent dirichlet allocation. *arXiv preprint arXiv:2005.10010, accepted by J. Fluid Mech.* (2021).

54. Řehůřek, R. & Sojka, P. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50 (ELRA, Valletta, Malta, 2010). http://is.muni.cz/publication/884893/en.

55. Ulbrich, U., Leckebusch, G. & Pinto, J. G. Extra-tropical cyclones in the present and future climate: a review. *Theor. Appl. Climatol.* **96**, 117–131 (2009).

Acknowledgements

This work is supported by the CNRS INSU-LEFE-MANU grant "DINCLIC" and by the French ANR-T-ERC grant "BOREAS". We thank Aglaé Jézéquel, Théo Mandonnet and the ESTIMR team for useful discussions.

Author contributions

D.F and B.D. conceived the study; L.F. performed all dynamical and statistical analyses by adapting the numerical code written by B.P.; F.P. suggested statistical tests to check the robustness of the findings. All authors participated in writing and reviewing the manuscript.

Competing Interests

The authors report no conflict of interest.

Supplementary Information

See detached document