Evaluation of the Atmospheric Circulation of CMIP6 Models for Extreme Temperature Events Using Latent Dirichlet Allocation

Nemo Malhomme, Davide Faranda, Lionel Mathelin, Bérengère Podvin

Evaluation of the Atmospheric Circulation of CMIP6 Models for Extreme Temperature Events Using Latent Dirichlet Allocation

NEMO MALHOMME,[a,b] BÉRENGÈRE PODVIN,[c] DAVIDE FARANDA,[a] AND LIONEL MATHELI[b]

[a] *ESTIMR, Université Paris-Saclay, CNRS, CEA, UVSQ, Laboratoire des sciences du climat et de l'environnement, Gif-sur-Yvette, France*
[b] *LISN, CNRS, Université Paris-Saclay, Orsay, France*
[c] *Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire EM2C, Gif-sur-Yvette, France*

(Manuscript received 1 December 2023, in final form 16 September 2024, accepted 2 January 2025)

ABSTRACT: For climate models to continue improving, we need to uncover as many discrepancies they have with reality as possible. In particular, evaluating the representation of extreme events is important but challenging owing to their rarity. Here, we study how general circulation models reproduce large-scale atmospheric circulation associated with extreme temperature events. To this end, we apply latent Dirichlet allocation (LDA), a dimensionality reduction method, to a set of sea level pressure ERA5 maps over the North Atlantic region. LDA provides a basis of sparse latent modes called "motifs" that consist of localized objects at the synoptic scale. Any pressure map can be approximated by a generally sparse combination of motifs, whose coefficients are called the weights, containing local information about large-scale circulation. Weight statistics can be used to locally characterize circulation patterns, in general and during extreme events, allowing for detailed comparison of datasets. For four CMIP6 models and reanalyses, we quantify local circulation errors and identify model-agnostic and model-specific biases. On average, large-scale circulation is well predicted by all models, but model errors are increased for heat waves and cold spells. Significant errors were found to be associated with Mediterranean motifs for all models in all cases. In addition, the combination of motif and temperature error can discriminate between models in the general and cold spell cases, while models perform similarly on heat waves. The sparse characterization provided by the LDA analysis is, therefore, well suited for the model preselection for the study of extreme events.

KEYWORDS: Extreme events; Heat wave; Statistical techniques; General circulation models; Model evaluation/performance

1. Introduction

Heat waves and cold spells both can significantly increase public health and safety risks (Weilnhammer et al. 2021), as well as cause infrastructure damage (Añel et al. 2017). They are generally defined as temperature events significantly higher or below average over a period of at least several days. Studies have shown that both the number and duration of heat waves in the European region have increased by up to 15% since preindustrial times (Frich et al. 2002; Alexander et al. 2006). Examples of severe heat waves include the European heat wave of 2003 (Fink et al. 2004) or that of 2018 (McCarthy et al. 2019). Both events have caused tens of thousands of deaths. While cold spell frequency and intensity have decreased since preindustrial times (Seneviratne et al. 2021), they still represent a hazard (López-Bueno et al. 2021). For instance, we can cite the cold spell of 2017 over the Balkans (Anagnostopoulou et al. 2017), which had consequent socioeconomic impacts. In addition, when occurring during spring, cold spells can have a devastating impact on the development of plants and cause major losses of agricultural yields (Papagiannaki et al. 2014). One such example is the cold spell of April 2021 described in Vautard et al. (2023b).

Heat waves and cold spells produce anomalies reaching up to ±15°C for several consecutive days. This implies that these events cannot be due to local thermodynamic drivers alone. They are explained in large part by changes in atmospheric circulation patterns (Rousi et al. 2022), namely, the ensemble of cyclones and anticyclones affecting a certain region at a given time. Cyclones and anticyclones advect warm or cold air from polar to tropical latitude and vice versa through the mechanism of baroclinic instability (Wallace and Hobbs 2006). With the temperature difference between pole and equator reaching up to 60°, cyclones and anticyclones can advect warm and cold air and trigger heat waves or cold spells. These cyclones and anticyclones evolve most of the time from west to east because they are embedded in the jet stream. Sporadically, the jet stream creates large meanders that trap cyclones and anticyclones in the same position for several days (Krishnamurti 1961). This phenomenon, called blocking, can cause persistence of warm or cold conditions in the same areas and trigger heat waves and cold spells (Faranda et al. 2016; Lupo 2021). Conditions of atmospheric circulation patterns that can cause extreme temperature events are often referred to as their dynamic drivers (Chan et al. 2022). Simulating the large excursions from the mean temperature responsible for hot and cold prolonged periods in Europe is crucial to understand, anticipate, and mitigate the impacts of heat waves and cold spells. Global and regional climate models are extensively used for this purpose both in present, past, and future climate conditions (Eyring et al. 2016).

However, models still face severe limitations in performing this task. According to the Coupled Model Intercomparison

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/JCLI-D-23-0719.s1.

Corresponding author: Nemo Malhomme, nemo.malhomme@lsce.ipsl.fr

DOI: 10.1175/JCLI-D-23-0719.1

Project (CMIP; Meehl et al. 2000), the statistical properties of extreme events are reasonably well captured by the models, but challenges persist in reproducing their frequencies and intensities, as well as in capturing local specificities (Kharin et al. 2013; Li et al. 2021). For example, Vautard et al. (2023a) show that models underestimate the trend of evolution of heat waves, and Jeong et al. (2021) show that models still underestimate the frequency of cold spells. Models are still unable to accurately reproduce the behavior of the atmosphere and ocean. In particular, they tend to underestimate the warming induced by climate change (van Oldenborgh et al. 2009) and still contain inaccuracies that affect local circulation patterns (Scaife et al. 2010), including those linked with extreme heat (D'Andrea et al. 2024) and extreme cold (Davini and D'Andrea 2020). Despite these biases compared to reality, models have made significant progress over the years. There have been increases in grid resolution, reaching a resolution as high as 1 km in regional models (Lucas-Picher et al. 2021). Tuning techniques have been developed to reduce biases, such as regarding Arctic sea ice cover extent, or the amplitude of Atlantic meridional overturning circulation (Mignot et al. 2021). Such improvements of the models have resulted in an increased ability to represent observed circulation patterns (Rodrigues et al. 2018). To assess the advances and the remaining challenges, it is necessary to develop evaluation methodologies that give a comprehensive and accurate measure of a model's ability to capture extremes and their drivers.

Regarding dynamic drivers, it is difficult to study directly atmospheric patterns, owning to their high dimensionality. Several methods attempt to produce a reduced-order representation of atmospheric circulation. One option is to categorize circulation fields into a set of weather regimes, large-scale quasi-stable states of atmospheric circulation [as first introduced in Rex (1950)]. Regimes are effective to describe persistent weather patterns [such as in Vautard (1990)]. This is useful to the study of extreme events, since some weather patterns, such as the abovementioned blockings, can induce extreme events such as cold spells or heat waves. However, by construction, weather regimes are not localized in space (Michelangeli et al. 1995). They combine various atmospheric structures that are local, such as, for example, cyclones or anticyclones, into large-scale atmospheric states. This loses the locality and the ability to differentiate between the components. They also typically exist at a time scale too large to define individual extreme events. Another technique is to use climate indices. Climate indices are one-dimensional variables that characterize the state of large-scale patterns, typically oscillations in oceanic circulation patterns that have a large influence over the global and regional climate (Stenseth et al. 2003). They condense information into a parameter that can be directly studied, and its correlation with all kinds of observables and events measured (de Freitas and Grigorieva 2017). This is especially useful to study oceanic oscillations, for example, Hanley et al. (2003). However, since climate indices aggregate a lot of data in a single variable, relevant information about the underlying circulation can be missing.

A solution to these difficulties might be offered by machine learning, which has already shown promise in a variety of applications in climate sciences (Reichstein et al. 2019; Chen et al. 2023). Significant contributions to modeling and prediction problems have been obtained through the application of machine learning, such as in the case of hydrology (Ardabili et al. 2020), or predictions on subseasonal to seasonal time scales (Cohen et al. 2019). Deep learning models have been used to model climate change and predict its consequences (Ren et al. 2020). Machine learning has also been applied to the study of extreme weather events [see Salcedo-Sanz et al. (2024) for a review]. Extreme events applications that make use of these methods range from detection and classification (Liu et al. 2016; Gardoll and Boucher 2022) to prediction (Fang et al. 2021). However, machine learning and especially deep learning have the drawback of being black boxes. Their inner workings can be unknown, which causes the output to be difficult to interpret, and available physical knowledge on the matter to be unused. There are several ways to tackle this issue. One such way is to use "explainable artificial intelligence (AI)," i.e., machine learning methods that offer ways to understand the processes behind the model predictions. Such methods have seen applications to the subject of extreme weather events (Bommer et al. 2024). However, as noted by O'Loughlin et al. (2024), to increase our trust in the models, it is important not only to focus on the mapping between model inputs and outputs but also to provide component-level understanding that makes it possible to attribute the performance of a model to a specific part of its architecture. As a step toward reaching this goal, we present an alternative solution, a dimensionality reduction technique facilitating physical insight through the use of easily interpretable latent factors inferred from the data. An advantage of the learned representation is that it is both sparse in the physical and the parameter space.

In this study, we show that a technique introduced in Fery et al. (2022) can provide new insight on the atmospheric circulation of extreme weather events and give both local and global quantitative measures of the performance of climate models. The technique relies on a statistical learning tool known as latent Dirichlet allocation (LDA) (Blei et al. 2003). Originally developed for text analysis, it has shown promise in capturing latent structures within complex datasets outside of natural language processing, such as in fluid mechanics (Frihat et al. 2021) or environmental sciences (Valle et al. 2018). In Fery et al. (2022), application of the LDA method to NCEP/NCAR sea level pressure (SLP) maps led to the identification of latent variables, or "motifs." Those motifs consist of synoptic objects, spatially localized pressure anomalies of the scale of 1000 km. Each map can be represented by a weighted combination of motifs. By monitoring the temporal evolution of the weights, they identified trends in impacts-defined extreme events.

As seen in Fery et al. (2022), the LDA methodology offers advantages in the context of the atmospheric circulation analysis. This methods reduces the dimension of the data to a limited number of modes while preserving most of the underlying information. Because LDA modes are localized and physically interpretable, information on the localization of relevant synoptic components follows directly from LDA application. Additionally, compared to other statistical learning methods of

dimensionality reduction, LDA is doubly sparse. The learned motifs are sparse, with significant values attributed only to a limited number of grid points, and the data representation is also sparse, with only a few significant motifs weights in the LDA composition of any given map.

In this article, we show that because of these properties, LDA can be used to locally characterize large-scale atmospheric circulation. In turn, this characterization can be used to comparatively evaluate climate models on their representation of circulation during extreme events. The paper is organized as follows. In section 2, we present the datasets to be analyzed and our methods of analysis. In section 3, motifs extracted from the ERA5 SLP dataset are used to study the synoptic configuration of hot and cold temperature extremes occurring in France. A comparison between the reanalysis and climate models using this synoptic representation is reported in section 4. An evaluation of the climate models is carried out in section 5, based on the joint analysis of the synoptic representation error and the average temperature discrepancy. Conclusions are given in section 6.

2. Methods

a. Climate data

We choose the reanalysis dataset ERA5 (Hersbach et al. 2020) as the ground truth to train LDA on and compare the models too. Our variable of study is the SLP, which contains the synoptic information relevant to a meteorological study, specifically the positions and extents of cyclones and anticyclones. An alternative for these properties would be 500-hPa geopotential height ($z500$). However, in ERA5 reanalysis data, $z500$ is computed from SLP rather than simulated directly. An average map is computed for each day of the year. To eliminate the seasonal cycle from the data, the corresponding average is subtracted from each map for each day of the year. The result is called anomalies.

We chose to evaluate general circulation models because they represent the physical detail of the atmospheric circulation. At the time of writing, the CMIP6 project contains the state of the art in general circulation models. We select four CMIP6 models for which a high number of runs are available: IPSL-CM6A-LR (33 runs) (Boucher et al. 2020), MIROC6 (50 runs) (Tatebe et al. 2019), ACCESS-ESM1.5 (29 runs) (Ziehn et al. 2020), and CanESM5 (25 runs) (Swart et al. 2019).

b. Extreme event definition

Among extreme weather events, we study specifically cold spells and heat waves. It is generally agreed upon that these terms refer to periods of temperatures significantly higher or below average for at least several days. However, any definition more precise is somewhat arbitrary, and there is no general consensus on a specific definition. A definition can be based on socioeconomic impacts, on physical indicators, or on the events that can be automatically categorized through machine learning methods trained on data categorized by hand [such as in Liu et al. (2016)].

Since we are interested in evaluating model dynamics, while Fery et al. (2022) uses a definition based on impacts, we prefer to use a physics-based definition. In particular, we define a cold spell (respectively, heat wave) as at least 3 consecutive days with average daily temperature below the 0.03 quantile (respectively, beyond the 0.97 quantile) of average temperatures over the studied period, from 1950 to 2021. Extreme events are defined for specific regions by considering the average temperatures over that region. To illustrate the method, we will consider cold spells and heat waves occurring in France. The results for five other countries, Italy, Spain, Poland, Germany, and the United Kingdom, are available in the online supplemental material.

c. Latent Dirichlet allocation

LDA is an unsupervised statistical learning method originally devised in the field of natural language processing (Blei et al. 2003). Its purpose is to extract, from a corpus of D written documents, a set of latent variables called "topics" that describe their content. Under the "bag of words" assumption, in which the ordering of words in the document is assumed to be irrelevant, each document is summarized by calculating the frequency of each word. The corpus of documents can then be represented as a word count matrix, where each column corresponds to a document, each line corresponds to a vocabulary word, and each matrix entry contains the number of times the word appears in the document.

The number of topics K is a hyperparameter of the method, equivalent to the number of clusters. Topics are characterized by a distribution over the vocabulary. For each $k \in [1, K]$, the topic of index k is associated with the multinomial distribution parameterized by the vector $\boldsymbol{\beta}_k$. We define $\mathbf{B}$ as the matrix containing all the $\boldsymbol{\beta}_k$ parameters. Assuming the vocabulary of possible words to be of size V, $\boldsymbol{\beta}_k$ are of length V and $\mathbf{B}$ of shape $K \times V$. An additional assumption of this approach is that $\boldsymbol{\beta}_k$ are drawn from a Dirichlet prior of parameter $\boldsymbol{\eta}$.

LDA is a soft clustering technique: It can associate to any document, included or not in the corpus, a distribution over topics indices. The distribution associated to a document indexed by d is a multinomial distribution over topic indices parameterized by the vector $\mathbf{c}(d)$, of length K, assumed to be drawn from a Dirichlet distribution of parameter $\boldsymbol{\alpha}$.

Given the parameters $\boldsymbol{\alpha}$ and $\mathbf{B}$, LDA assumes that each document d of the corpus has been generated as follows:

- A total number of word positions N in the document d is drawn from a Poisson distribution.
- A topic composition $\mathbf{c}(d)$ is drawn for the document d (see Fig. 1) from a Dirichlet distribution of parameter $\boldsymbol{\alpha}$.
- For each word position n in the document,
  - a topic index $z_{d,n}$ is drawn from the document-topic distribution parameterized by $\mathbf{c}(d)$ and
  - a word $w_{d,n}$ is drawn from the topic-word distribution parameterized by $\boldsymbol{\beta}_{z_{d,n}}$

The generative process is summarized in Fig. 2.

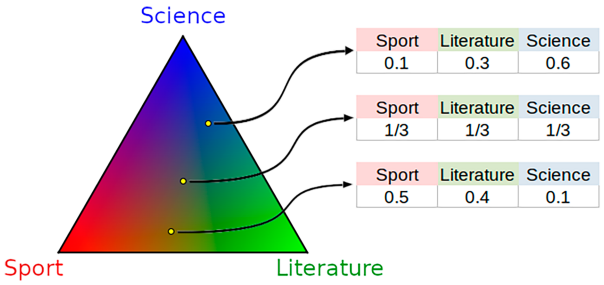

FIG. 1. Example representation of the space of possible topic compositions, on which we define the Dirichlet probability distribution parameterized by $\boldsymbol{\alpha}$.

The joint distribution of all observable and hidden variables, knowing the parameters $\boldsymbol{\alpha}$ and $\mathbf{B}$, is

$$p[(w_{d,n})_{d\in[1,D],n\in[1,N]}|\boldsymbol{\alpha},\mathbf{B}] = \prod_{d=1}^{D}\int_{c} p[\mathbf{c}(d)|\boldsymbol{\alpha}]\prod_{n=1}^{N}\sum_{z_{d,n}=1}^{K} p[z_{d,n}|\mathbf{c}(d)] \times p(w_{d,n}|z_{d,n},\mathbf{B})d\mathbf{c}(d), \quad (1)$$

where

- $\mathbf{c}(d)$ is drawn from the Dirichlet distribution of parameter $\boldsymbol{\alpha}$:

$$p[\mathbf{c}(d)|\boldsymbol{\alpha}] = \frac{1}{F(\boldsymbol{\alpha})}\prod_{k=1}^{K} c_k(d)^{\alpha_k-1}, \qquad F(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^{K}\Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}, \quad (2)$$

- $z_{d,n}$ is drawn from the multinomial distribution parameterized by $\mathbf{c}(d)$:

$$p[z_{d,n} = k|\mathbf{c}(d)] = c_k(d), \quad (3)$$

- $w_{d,n}$ is drawn from the multinomial distribution parameterized by $\boldsymbol{\beta}_{z_{d,n}}$:

$$p(w_{d,n} = i|z_{d,n},\mathbf{B}) = \boldsymbol{\beta}_{z_{d,n},i}. \quad (4)$$

This method is applied to datasets of bidimensional climate variables maps where each spatial map is reinterpreted as a document. Grid points, or cells, are reinterpreted as the words, with the list of cells taking the role of the vocabulary. Field values at each cell are reinterpreted as word counts. In this case, the topic-cell distributions, parameterized by $\boldsymbol{\beta}_k$, are defined over space and are called motifs. Since the climate variable values are interpreted by LDA as word counts, they have to be digitized and made nonnegative. To ensure nonnegativity, the real variable maps are separated into two channels, one for positive and one for negative values. This is equivalent to doubling the grid size over which the maps are defined (see also Fery et al. 2022). Moreover, a rescaling factor A is applied to the data before digitization in order to manage computation times, which depend on the total sum of field values in the maps. To sum up, if p_i designates the real

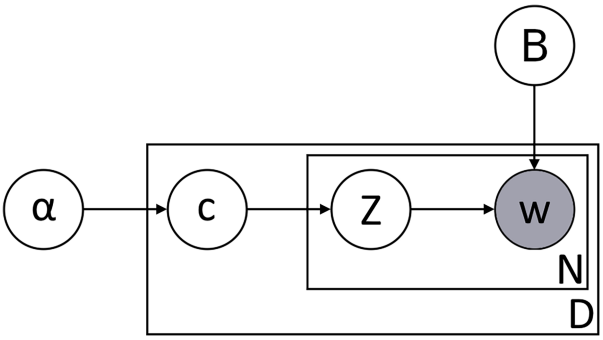

FIG. 2. Graphical model representation of the LDA generative process. The circles represent variables, grayed-out circles being observed variables. The arrows represent the process of drawing a random variable from a distribution. The rectangles represent reiteration of the process, with D being the number of documents and N being the number of words in a document.

value of an original map $\mathbf{p}$ at cell i, of a total number of I cells, we define a new map, noted $\mathbf{P}$, with nonnegative, integer values defined on $V = 2I$ cells. It is defined as follows:

$$\forall i \in \{1, \ldots, I\}, \begin{cases} P_i = \max[\text{int}(A \times p_i), 0] \\ P_{i+I} = \max[-\text{int}(A \times p_i), 0] \end{cases}, \quad (5)$$

where int() designates the truncate to integer function. The motifs $\boldsymbol{\beta}_k$ will then be represented on each cell i of the original grid as the difference of the two channels: $\beta_{k,i} - \beta_{k,i+I}$. We found that no motif attributes nonnegligible probability to both channels at any grid point, which reflects the anticorrelation of the two channels.

The analysis of a corpus of documents with LDA consists in examining the posterior distribution of the motifs $\mathbf{B}$, motif proportions $\mathbf{c}$, and motif assignments z. These are determined via a variational Bayes approach aiming to maximize the evidence lower bound, which is related to the likelihood of the observed data. Small values for the Dirichlet parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$, respectively, ensure the sparsity of the document-topic and the topic-word distributions: There are generally few topics in each document, and each topic is characterized by high occurrences of a few vocabulary words. This sparsity property makes LDA particularly suited to provide models and decompositions that can be interpreted easily. For more information, see Hoffman et al. (2010).

For a given set of D maps, LDA returns motif distributions over grid cells $(\boldsymbol{\beta}_k)_{k\in[1,K]}$, as well as the map compositions $[c_k(d)]_{d\in[1,D],k\in[1,K]}$. Throughout the paper, $c_k(d)$ will be called the weights of motif k in map d. We note that $\forall d \in [1, D]$, $\sum_{k=1}^{K} c_k(d) = 1$. The motif weights $\mathbf{c}(d)$ are always positive, unlike other decompositions such as the principal component analysis. The set of distributions $(\boldsymbol{\beta}_k)_{k\in[1,K]}$ can be considered as a basis of motifs. Any map $\mathbf{P}$ defined on the grid (but not necessarily part of the learning set) can be approximated in this basis by its K-dimensional motif composition $\mathbf{c}(\mathbf{P})$. Different sets of maps can thus be compared efficiently through

examination of their motif compositions. In practice, numerical implementation of LDA is carried out with the Python module Gensim (Řehůřek and Sojka 2010).

d. Application of LDA

We apply LDA to ERA5 SLP data from the North Atlantic region between 22.5° and 70° latitude and 80° and 50° longitude. Although higher resolutions are available, we used a spatial resolution of 1° as it was found to be sufficient to contain all relevant information about circulation patterns on the synoptic scale while maintaining manageable computation times. Our resolution is 48 points in latitude and 130 points in longitude, and we have two channels for positive and negative values. Therefore, the total number of values per map, noted N, is 12 480. The temporal correlation time of synoptic circulation patterns is approximately 5 days. The full dataset (which will be referred to as general data) consists of daily averaged SLP anomaly fields from 1950 to 2021. The number of motifs was set to $K = 28$, as previous work (Fery et al. 2022) showed, using a methodology from the field of dynamic systems (Faranda et al. 2017), that this was the average local dimension of SLP anomaly data. The rescaling factor is set to $A = 0.5$ to alleviate computation time. Some arbitrariness exists in the choice of the factor. However, increasing A to 1 did not significantly change the basis.

These 28 motifs are shown in Fig. 3 and sorted by their average weights in decreasing order. To make discussion easier, names based on their signs and geographical locations were assigned to the motifs. Several motifs in the basis are approximate opposites of one another, such as Labrador high (1) and Labrador low (17), Genoa low (25), and Mediterranean anticyclone (18). The resulting basis is similar to the one obtained in Fery et al. (2022), which was obtained for different reanalysis datasets at a lower resolution (NCEP/NCAR). Most of the motifs have recognizable equivalents from one basis to the other, although some geographical locations may occasionally differ by a few hundred kilometers. Motifs can be seen to be analogous to localized synoptic objects of a given sign, such as cyclones and anticyclones. Therefore, motif weights in a SLP anomaly map directly measure the contribution of the relevant synoptic objects.

LDA offers the possibility of reconstructing maps from a motif composition. The reconstruction of map $\mathbf{P}$, noted $\mathbf{P}^*$, is obtained based on Eq. (6):

$$\mathbf{P}^* = \|\mathbf{P}\|_1 \sum_{k=1}^{K} c_k(\mathbf{P})\boldsymbol{\beta}_k, \tag{6}$$

where

- $\boldsymbol{\beta}_k$ is the spatial distribution associated with motif k,
- $c_k(\mathbf{P})$ is the weight of the kth motif in the weight vector associated with the pressure map $\mathbf{P}$, and
- $\|\mathbf{P}\|_1 = \sum_{i=1}^{V}|\mathbf{P}_i|$, i iterating over the V grid points, is the ℓ_1 norm of map $\mathbf{P}$. This term is a renormalization factor, allowing for direct comparison with physical fields.

In this article, we reconstruct the average compositions of maps of cold spells and heat waves in a given model. In this case, $c_k(\mathbf{P})$ is replaced with $\langle\langle c_k(\mathbf{P})\rangle\rangle$, where $\langle\langle\cdot\rangle\rangle$ designates the average over all time steps associated with a given type of extreme event, and $\|\mathbf{P}\|_1$ is replaced with $\|\langle\langle\mathbf{P}\rangle\rangle\|_1$.

3. Synoptic configuration of extreme events

We first use the decomposition into synoptic objects given by LDA to identify the atmospheric circulation patterns associated with cold spells and heat waves. The patterns associated with extreme temperature events in one country are expected to differ from those that would cause such events in another. As mentioned above, we focus our study on extreme temperature events occurring in France. The average synoptic configuration of reanalysis fields corresponding to cold spells (respectively, heat waves) is represented and compared to the average configuration of all reanalysis data in Fig. 4. Uncertainties are estimated by a resampling method: Many alternative sets of cold spell (respectively, heat wave) days are generated by randomly sampling with replacements from the original cold spell (respectively, heat wave) data. The average motif weights in these datasets are computed, and the 0.05 and 0.95 quantiles weights for each motif are used as lower and upper errors. We found that statistical convergence was reached with 500 datasets, with quantiles chosen to have a 90% confidence interval.

The synoptic configuration of extreme events is different from the average configuration of the general data. Cold spell circulation is dominated by northern anticyclones such as Greenland high, Scandinavian anticyclone, and U.K. high, with more than 6% weights each. Correspondingly, the low pressure objects over those regions have less than half the weights they have in the general data. Genoa low is also a key motif in French cold spells, being the fourth most represented motif. Its opposite, the Mediterranean anticyclone, also has during cold spells half the weight it has in general. Heat wave circulation is dominated by a smaller set of high-weight motifs, mainly consisting of Scandinavian anticyclone and central European high. The U.K. high is also more prevalent during heat waves than in general. Both types of extremes are associated with an above-average weight of Scandinavian anticyclone and of U.K. high.

4. Evaluation of model representation

a. Robustness of the basis

We first establish that a unique basis can be used to compare models with reanalysis data. Figure 5 shows the correlation matrix between the reanalysis data basis and that obtained from a run from the IPSL-CM6A-LR model, which are, respectively, associated with cell–motif distributions $\mathbf{B}$ and $\mathbf{B}'$. Since motifs are defined by a spatial distribution, we use spatial correlation to measure their similarity.

The correlation matrix is obtained as follows: All fields are set to the same 1° resolution by linear interpolation. For each matrix entry, the Pearson correlation coefficient ρ_{kl} between motif k of basis $\mathbf{B}$, $\boldsymbol{\beta}_k$, and motif l of basis $\mathbf{B}'$, $\boldsymbol{\beta}'_l$ is computed as shown in Eq. (7):

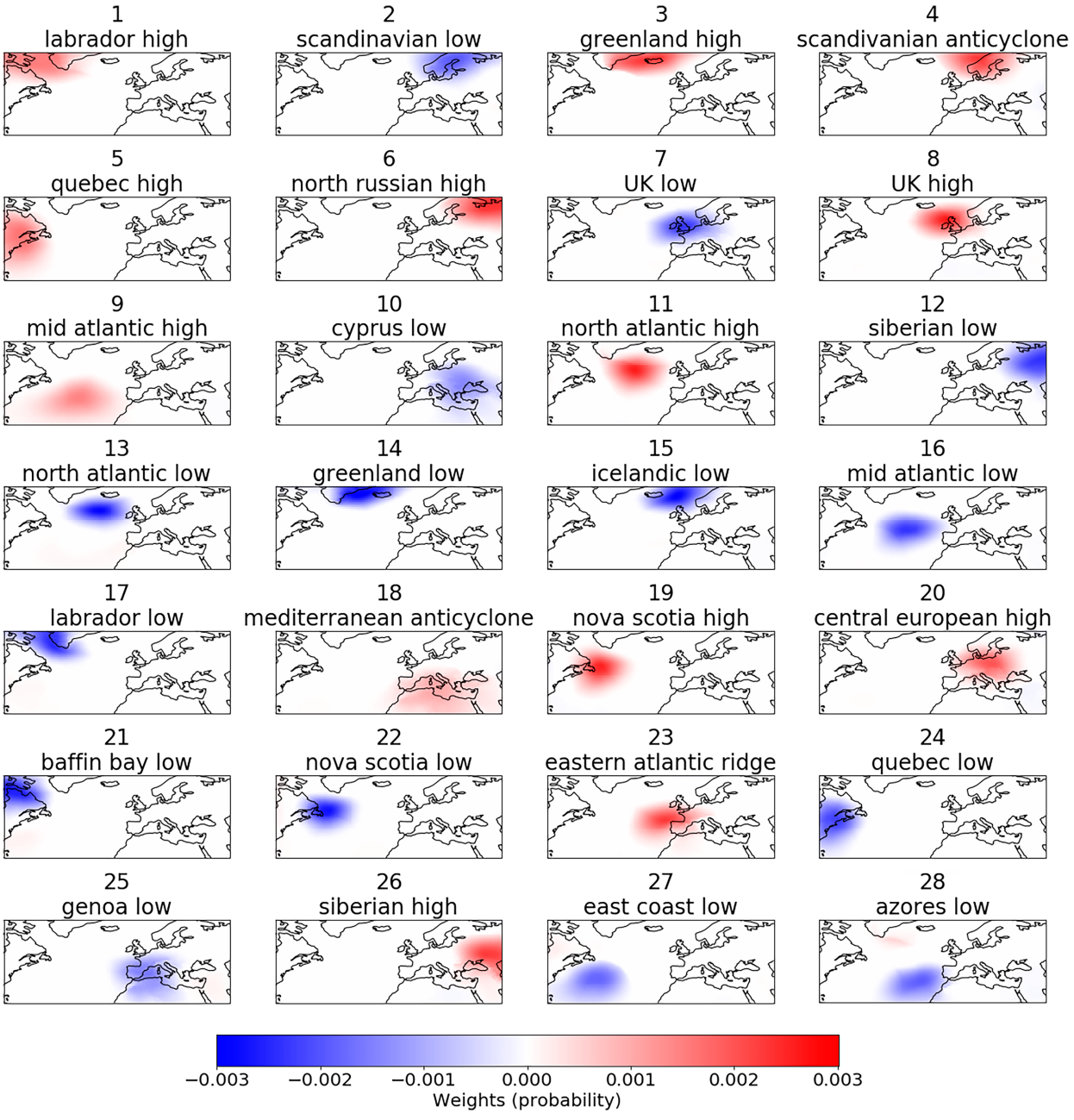

FIG. 3. The basis of 28 motifs learned by LDA from ERA5 SLP anomaly fields. Each motif is defined as a probability distribution over space, with positive and negative channels. The negative channel is subtracted from the positive channel. The names were given based on sign and geographical location.

$$\rho_{kl} = \frac{\sum_{i=1}^{V}[(\beta_{k,i} - \overline{\beta_k})(\beta'_{l,i} - \overline{\beta'_l})]}{\sqrt{\sum_{i=1}^{V}(\beta_{k,i} - \overline{\beta_k})^2}\sqrt{\sum_{i=1}^{V}(\beta'_{l,i} - \overline{\beta'_l})^2}}, \tag{7}$$

where $\overline{\ }$ designates the average over all grid points in the two channels: $\overline{\beta_k} = (1/V)\sum_{i=1}^{V}\beta_{k,i}$.

Motifs were reordered in order to give the same rank in the basis to the motifs with the highest correlation. For the case considered, 22 out of 28 motifs have a clear equivalent in the other basis with correlation of at least 0.7 (other choices of models gave similar results). Based on these results, we consider that the motif basis learned from ERA5 is relevant to represent all model data.

b. General data case

We project each run of the four models onto the motif basis learned from ERA5 and then average the resulting synoptic

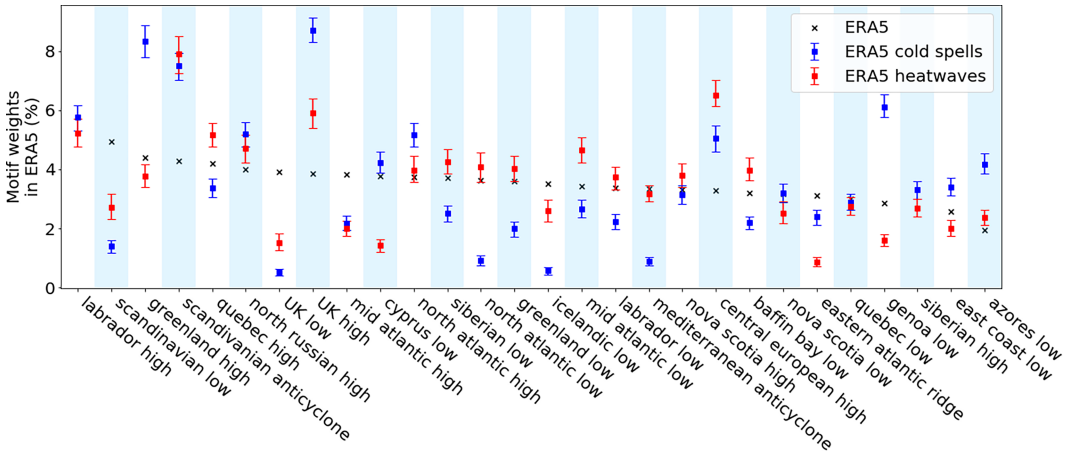

FIG. 4. Average motif weight in the configuration of ERA5 SLP anomaly fields, in the general case (black), in the case of cold spells in France (blue), and in the case of heat waves in France (red). The 90% confidence uncertainties are determined by resampling.

configuration of the fields over each run. We first consider all fields in the datasets. For each run, the relative difference between the K motif weights in the model and that in the reanalysis is computed following Eq. (8):

$$\forall k \in [1, K],\ E_k = \frac{\langle c_k(\mathbf{P}^{m,r})\rangle - \langle c_k(\mathbf{P})\rangle}{\langle c_k(\mathbf{P})\rangle}, \tag{8}$$

where $\mathbf{P}$ corresponds to the reanalysis maps, $\mathbf{P}^{m,r}$ corresponds to the maps from run r of model m, and $\langle\cdot\rangle$ designates the average over all maps in the dataset (model run or reanalysis). For each model, the statistics of the error computed for each model run are shown in Fig. 6, using box plots. The mean weight of the motifs in the reanalysis data is also indicated for comparison.

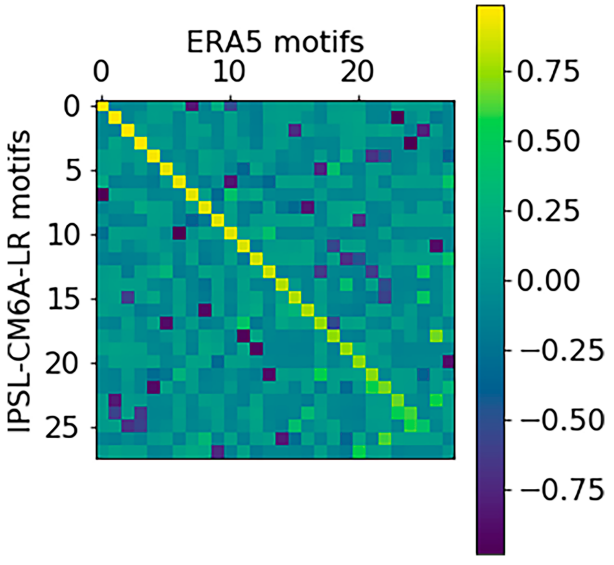

FIG. 5. The spatial correlation between the motifs of the bases obtained by applying LDA on ERA5 (vertical) and on IPSL-CM6A-LR run 1 (horizontal). The order of the motifs has been adjusted to put the highest correlations on the diagonal.

The median relative errors, materialized by the black lines within the boxes, are relatively small. In particular, the error is less than 15% for the eight most prevalent motifs in the reanalysis. Overall, models represent well the reanalysis synoptic configuration. Relative errors made by IPSL-CM6A-LR, MIROC6, and ACCESS-ESM1.5, which have resolutions of, respectively, 38 × 53, 34 × 92, and 39 × 69, are all below 20%. We note that the largest error (25%) is observed for CanESM5, which has a resolution of 17 × 46. It is possible that these larger errors could be due to its coarser resolution. Moreover, the inner variability of the models (corresponding to the width of the boxes) is typically much smaller than the error [in 96 cases out of the 112 (87.5%), the model's internal variability is lower than its bias]. This shows that all runs make similar predictions and also indicates the presence of a bias inherent to each model.

In addition, the motifs associated with the largest relative errors tend to be the same from one model to another. A multimodel ensemble mean would therefore not eliminate these biases. The largest errors are made on motifs located on the Mediterranean region. The Cyprus low and Mediterranean anticyclone motifs are overrepresented in all runs of all four models. Every model run also overrepresents Genoa low and underrepresents U.K. high and low. Finally, the Scandinavian anticyclone is the fourth most prevalent motif in the reanalysis, with an average weight of more than 4% of all models, but ACCESS-ESM1.5 systematically underrepresents it. These similarities in the model errors suggest that the origin of the errors could be common to all models.

c. *Model representation of cold spells*

We study how models capture the circulation patterns of extreme events. For this part, we focus on cold spells occurring in France. The datasets are filtered following the definition proposed in section 2.

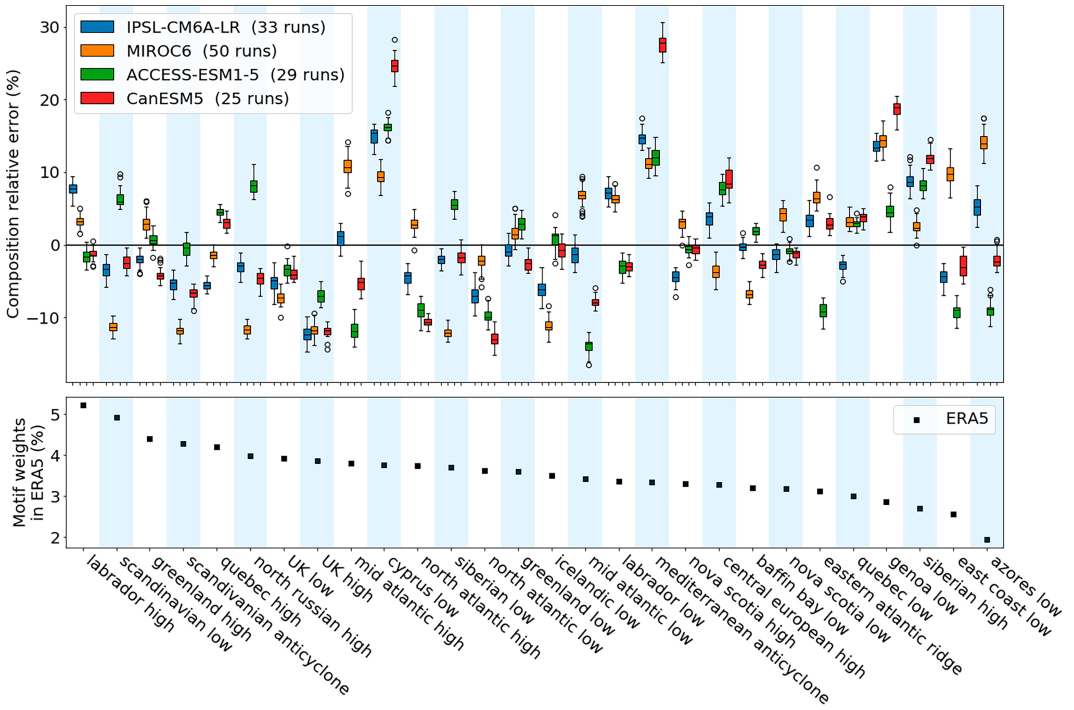

FIG. 6. (top) Relative error on average motif weight between models and ERA5 reanalysis. The box edges correspond to first and third quartiles. The black line is the median. The whiskers extend to the furthest data point, up to 1.5 times the difference between the first and third quartiles. Data points beyond the whiskers are represented as colorless circles. (bottom) Average motif weight in the synoptic configuration of ERA5 fields.

The fields corresponding to the real and reconstructed averages are represented in Fig. 7. The real average is obtained by taking a conditional average over all daily fields associated with a cold spell. The reconstructed average is obtained from the average motif compositions of the daily fields included in the conditional average, using Eq. (6). To identify the most significant motifs associated with each model, the two most prevalent cyclonic and the two most prevalent anticyclonic motifs in each case are annotated in the figure.

The overall synoptic structure associated with French cold spells consists of an anticyclonic structure in the north and a cyclonic structure in the south, with a corridor between the two slanted northeast–southwest, passing through the middle of France. For all models, the real average is generally similar to its reconstructed average, which shows that LDA captures the synoptic information contained in the real fields.

The model average fields are also in a good agreement with those of ERA5. They have the same two most prevalent cyclones as ERA5, Cyprus low and Genoa low, and reproduce motif 8, U.K. high, as a dominant motif. However, some discrepancies are present: All models underestimate the westward extent of the anticyclonic structure over the Atlantic.

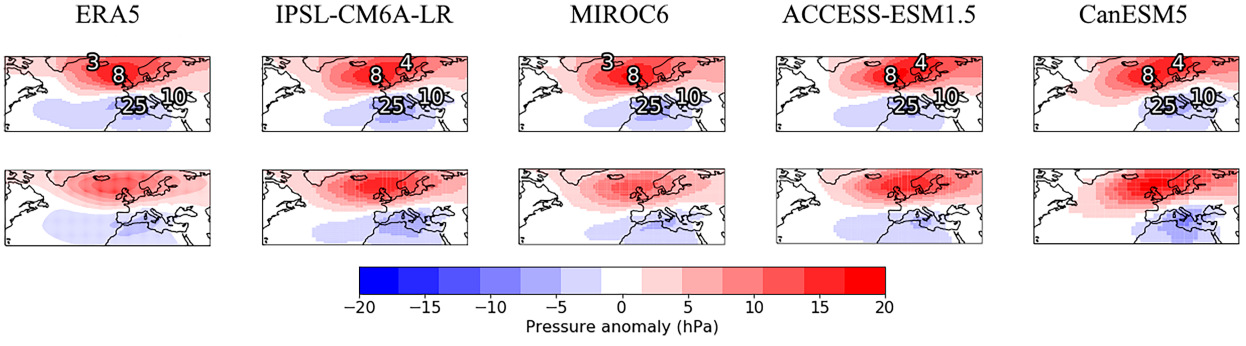

FIG. 7. (top) Reconstruction of the average motif composition of cold spells in France according to different models (columns). The two cyclones and the two anticyclones with highest average weights in each case are annotated. (bottom) Average SLP field for cold spells in France according to different models (columns).

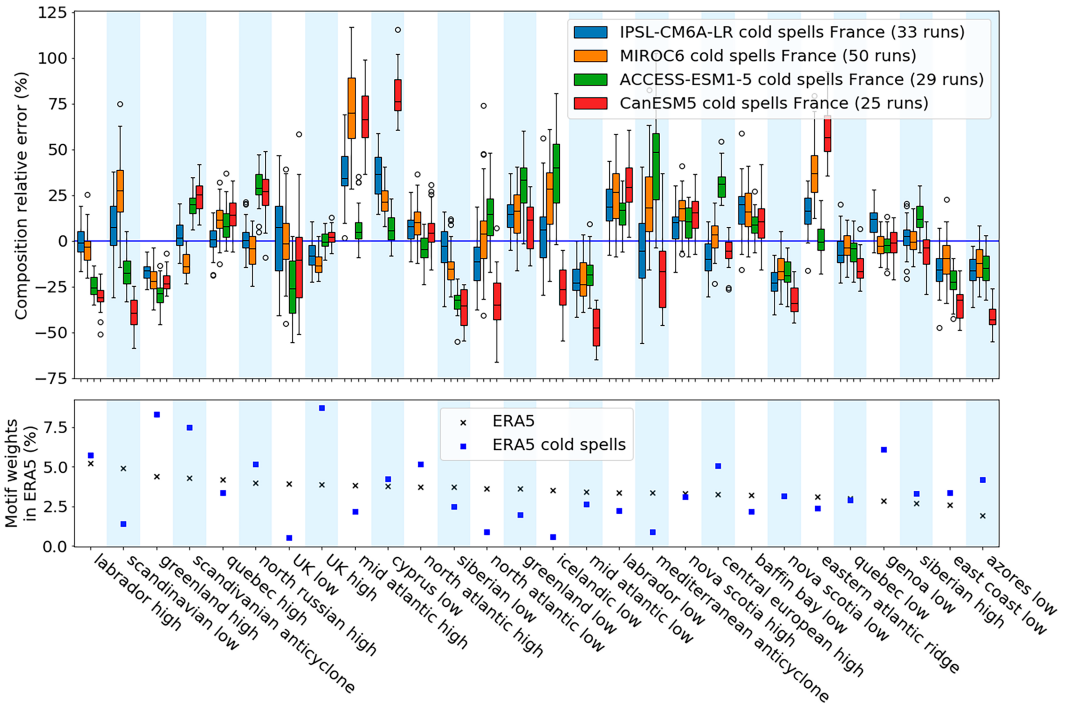

FIG. 8. (top) Relative error on average motif weight between models and ERA5 reanalysis in the case of cold spells occurring in France. (bottom) Average motif weight in the synoptic configuration of ERA5 fields, for cold spells and in the general case.

Only MIROC6 captures the fact that Greenland high (motif 3) is more prevalent than Scandinavian anticyclone (motif 4), though as seen in section 3, Greenland high and Scandinavian anticyclone are both relevant for French cold spells (near 8% weights). In addition, on CanESM5, Genoa low is too intense, and the cyclonic structure sees no extension to the west of the Mediterranean Sea.

For a more detailed analysis, we show for each motif the relative errors in weights between the reanalyses and the models in the case of cold spells occurring in France, in Fig. 8. The biases are significantly higher for the cold extremes than for the general case. The variability among the runs of each models is also higher than for the general case. The five most prevalent reanalysis motifs during French cold spells are U.K. high, Greenland high, Scandinavian anticyclone, Genoa low, and central European high. Most of these motifs are correctly represented by the models. The significantly higher weights of U.K. high and Genoa low during cold spells are well captured by all models with an error within the internal variability of all four models. The central European high is also well represented by all models except by ACCESS-ESM1.5 which overestimates it by 25%. The weight of Scandinavian anticyclone high is well captured by IPSL-CM6A-LR and MIROC6, while it is overestimated by 25% by the two other models. All models make about 25% error on Greenland high. Higher errors are made on less relevant motifs where the reanalysis values are lower. The most overrepresented motifs are Cyprus low and mid-Atlantic high for all models except ACCESS-ESM1.5. We note that larger errors are generally observed for the lower-resolution model CanESM5.

d. *Model representation of heat waves*

We now focus on heat waves occurring in France. We represent the real and reconstructed average heat wave fields in Fig. 9, using the same methodology as in the previous section.

The SLP anomaly values are weaker than in the case of cold spells. This is because heat waves are more varied in configuration, leading to average error values closer to zero. There are differences between the real and reconstructed fields. In ERA5 and all models, the anticyclonic structure over Europe has a more crescent-like shape around the Atlantic cyclone that changes into an arrow-like shape in LDA reconstruction. Still, the overall structure consisting of anticyclones over northern and central Europe with a depression over the Atlantic is preserved by LDA reconstruction.

Models reproduce the overall structure of ERA5 circulation, with anticyclonic conditions on northern and central Europe and cyclones over the Atlantic. Models disagree, with ERA5 and each other, on the shape of those cyclones and the extent of the anticyclonic structure over northern Atlantic. The most prevalent anticyclones in the reanalysis are the Scandinavian anticyclone (motif 4) and the central European high (motif 20). Only CanESM5 reproduces this property. For the other models, this leads to an anticyclonic structure that is weaker in the north for IPSL-CM6A-LR, in the south for MIROC6, and less intense overall for ACCESS-ESM1.5. The most prevalent cyclones are Siberian low (motif 12) and

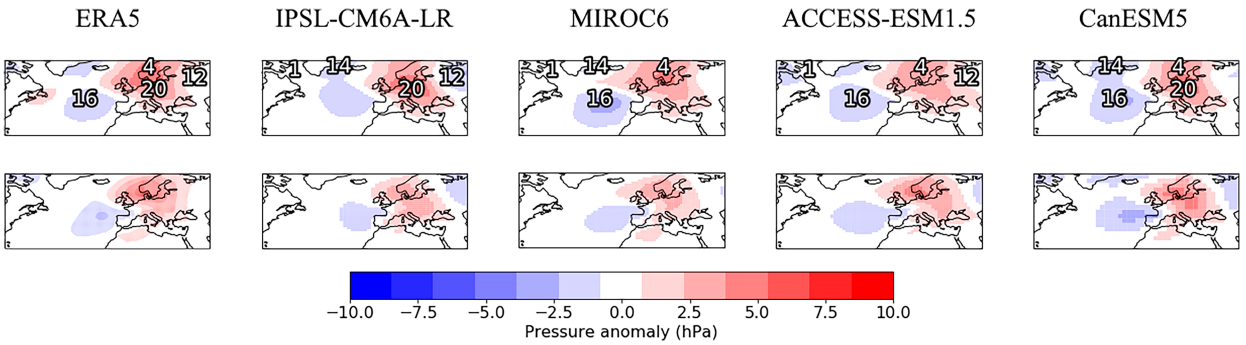

FIG. 9. (top) Reconstruction of the average motif composition of heat waves in France according to different models (columns). The two cyclones and the two anticyclones with highest average weights in each case are annotated. (bottom) Average SLP field for heat waves in France according to different models (columns).

mid-Atlantic low (motif 16). Only ACCESS-ESM1.5 reproduces this property.

For a more detailed analysis, we computed relative errors in motif weights between the reanalyses and the models for heat waves occurring in France. They are shown in Fig. 10.

In the case of heat waves too, model biases and internal variability are higher than in the general case. Which motifs are or are not relevant is generally well captured by the models. However, the most relevant motifs tend to be underpredicted by the models. All models except ACCESS-ESM1.5 underrepresent by 20% on average the contribution of the most prevalent motif, which is the Scandinavian anticyclone. The second most prevalent motif, the central European high, is well represented by IPSL-CM6A-LR and CanESM5 but underrepresented by about 20% by MIROC6 and ACCESS-ESM1.5. The U.K. high, the third most prevalent motif, is underrepresented by 20% or more by almost all runs of all models. In general, motifs that have higher weights than in the general case tend to be underrepresented (as, for instance, Quebec high and north Russian high), while motifs that have lower weights (U.K. low, Nova Scotia low, and Genoa low) are overrepresented. This shows that models underestimate the changes in atmospheric circulation associated with heat waves.

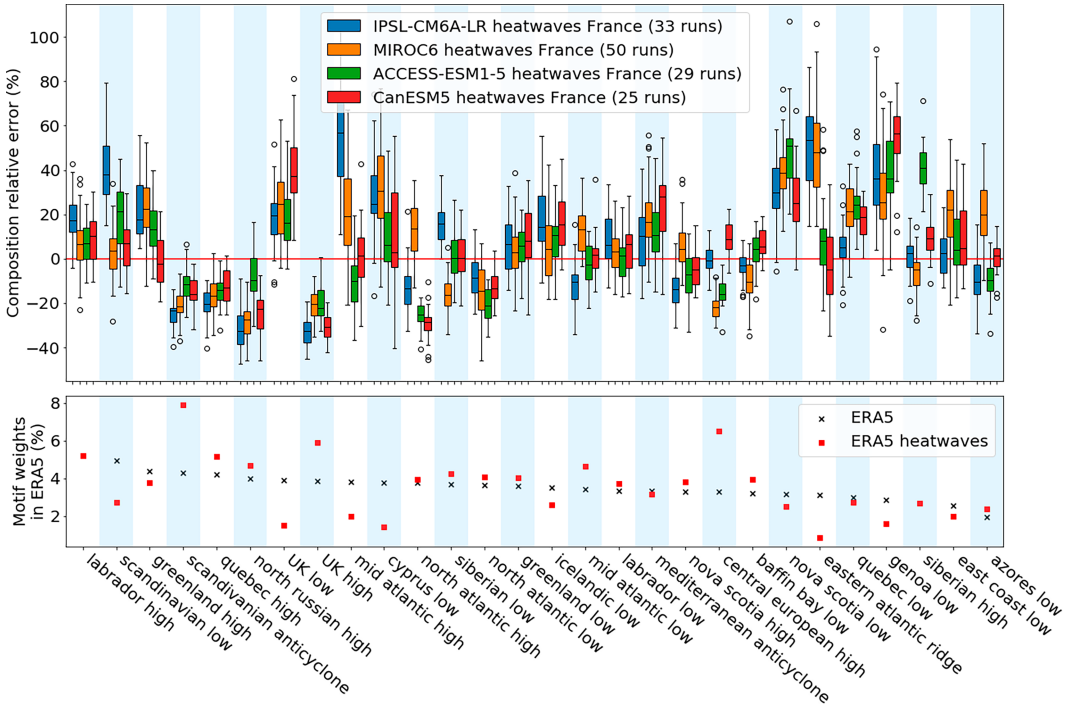

FIG. 10. (top) Relative error on average motif weight between models and ERA5 reanalysis in the case of heat waves occurring in France. (bottom) Average motif weight in the synoptic configuration of ERA5 fields, for heat waves and in the general case.

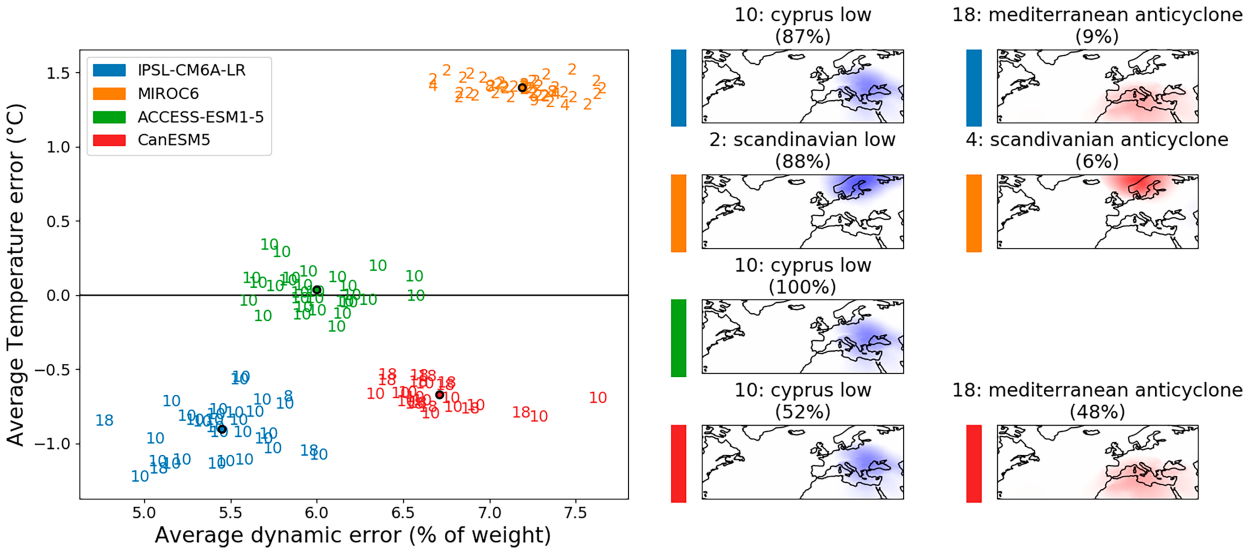

FIG. 11. The run-average temperature model error (average temperature difference with reanalysis) vs the run-average dynamic model error (average motif weights difference with reanalysis). The colored dots indicate the average of all runs of a model. Each number corresponds to the motif contributing the most to the dynamic error in a given run. The two most frequent such motifs for each model are displayed on the right.

5. Global dynamic and temperature error

a. General data case

LDA provides a decomposition of circulation patterns into motifs. As seen in section 4, differences in motif weights provide a quantitative measure of the model's ability to reproduce dynamics observed in reanalysis data. The dynamic error of run r of model m $E_{\mathbf{P}}^{m,r}$ is defined as the sum of individual motif errors, where the error associated with each motif is defined as the absolute difference between the average weight of this motif in the model and that in the reanalysis data. It is therefore expressed as

$$E_{\mathbf{P}}^{m,r} = \sum_{k=1}^{K} |\langle c_k(\mathbf{P}^{m,r})\rangle - \langle c_k(\mathbf{P})\rangle|. \tag{9}$$

The dynamic error can be used to evaluate models comparatively and produce rankings. Models can also be evaluated based on the temperature error, i.e., the temperature difference between models and reanalysis data. Two sources can contribute to the temperature error: errors in the underlying thermodynamic processes (regulating the exchanges of heat) and errors in the dynamic processes (regulating the circulation of air masses in the atmosphere Wehrli et al. 2018). The dynamic error, as previously defined, may propagate to the temperature error, but it is unclear to what extent the latter is determined by the former. Therefore, an important question is to determine whether the measure of temperature error brings additional information to the comparison of the model's performance. For run r of model m, the temperature error is computed as shown in Eq. (10), with $\mathbf{T}$ denoting reanalysis temperature fields and $\mathbf{T}^{m,r}$ denoting those from run r of model m.

$$E_{\mathbf{T}}^{m,r} = \langle \overline{T^{m,r}} \rangle - \langle \overline{T} \rangle. \tag{10}$$

Each model run is represented as a point in the error plane $(E_{\mathbf{P}}^{m,r}, E_{\mathbf{T}}^{m,r})$ shown in Fig. 11. In addition, we annotate for each run the index of the motif with the highest contribution to the dynamic error: $\max_k |\langle c_k(\mathbf{P}^{m,r})\rangle - \langle c_k(\mathbf{P})\rangle|$. For each model, we show on the right side of the figure the two motifs that appear most frequently as the largest contributor to the dynamic error of a run (the proportion of runs each motif corresponds to is indicated between parentheses)—except in the case of ACCESS-ESM1.5, where the largest contributor is always Cyprus low.

Although some overlap between the models would be observed if only one kind of error was considered, each model can be associated with a well-identified cluster in the 2D error plane. MIROC6 is the model with the highest dynamic and temperature error but with the lowest temperature variability. Unlike other models, it overpredicts the temperature. In contrast, the IPSL-CM6A-LR model has the highest temperature variability for a relatively low error (similar to that of CanESM5), and it also corresponds to the lowest dynamic error. ACCESS-ESM1.5 has the lowest temperature error for a relatively low dynamic error.

As mentioned earlier, each run is annotated with the index of the motif contributing the most to the dynamic error, which makes it possible to attribute the error to specific motifs and regions in space. Cyprus low (motif 10) is the least well-represented motif for all or almost all runs of ACCESS-ESM1.5 and IPSL-CM6A-LR, as well as most runs of CanESM5. Another motif that is occasionally the least well represented in runs of CanESM5 and IPSL-CM6A-LR is Mediterranean anticyclone (motif 18), the opposite of Cyprus low. Both are eastern Mediterranean motifs.

We note that these motifs, which contribute the most to the error, are, however, not the most prevalent motifs. The

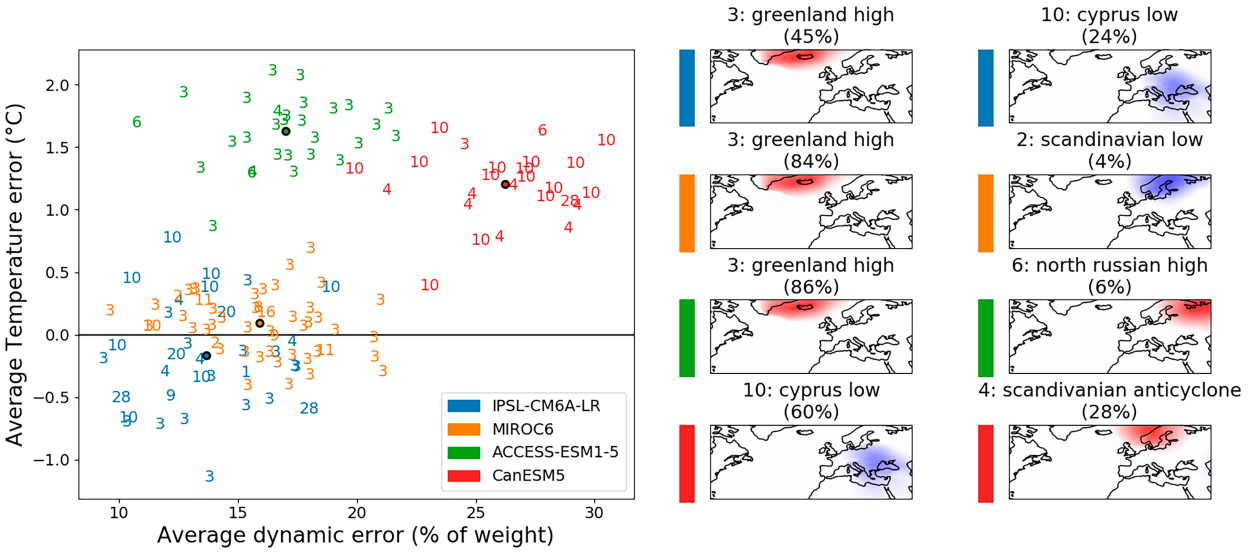

FIG. 12. The run-average temperature model error (average temperature difference with reanalysis) on cold spells in France vs run-average dynamic model error (average motif weights difference with reanalysis) on same extremes. We eliminate the errors computed in the general case, so as to look only at errors specific to extreme events. The colored dots indicate the average of all runs of a model. Each number corresponds to the motif contributing the most to the dynamic error in a given run. The two most frequent such motifs for each model are displayed on the right.

associated relative error is, therefore, necessarily large. This confirms that the representation of atmospheric circulation over the eastern Mediterranean region is a significant issue for all models, particularly for models IPSL-CM6A-LR, ACCESS-ESM1.5, and CanESM5. MIROC6 appears to differ from other models, as its error on the mean temperature is significantly higher, and its dynamic error is attributed to different motifs than other models, the Scandinavian low and Scandinavian anticyclone (motifs 2 and 4). This points to there being different sources of error between MIROC6 and the other models.

b. *Model representation of extreme events*

We now consider extreme temperature events and compute the dynamic and temperature errors associated with heat waves as well as cold spells. In that case, we eliminate the average bias, so as to only look at the component specific to extreme events. We, therefore, define the anomalous dynamic error $E^{m,r}_{\mathbf{P},\mathrm{ex}}$ similarly for heat waves and cold spells following Eq. (11):

$$E^{m,r}_{\mathbf{P},\mathrm{ex}} = \sum_{k=1}^{K} |\langle\langle c_k(\mathbf{P}^{m,r})\rangle\rangle - \langle\langle c_k(\mathbf{P})\rangle\rangle| - E^{m,r}_{\mathbf{P}}. \tag{11}$$

The anomalous temperature error $E^{m,r}_{\mathbf{T},\mathrm{ex}}$ is defined for heat waves and cold spells, for run r of model m following Eq. (12):

$$E^{m,r}_{\mathbf{T},\mathrm{ex}} = \langle\langle \overline{T^{m,r}}\rangle\rangle - \langle\langle \overline{T}\rangle\rangle - E^{m,r}_{\mathbf{T}}. \tag{12}$$

In subsequent figures, the dynamic and temperature errors represented are only the anomalous errors defined above. The average errors studied in Fig. 11 are eliminated. However, we note that the general conclusions reported below did not change when these errors were taken into account.

Figure 12 shows the model anomalous temperature error against the model anomalous dynamic error in the case of cold spells occurring in France.

The inner model variability is higher in the cold extreme case than in the full dataset case, for both dynamic and temperature errors. There are three distinct clusters in error space, the differences between them being bigger than internal model variabilities. The first cluster corresponds to model CanESM5. It is the model with the highest dynamic error and has a high temperature error. The second cluster corresponds to model ACCESS-ESM1.5. ACCESS-ESM1.5 has the highest temperature error, underpredicting the lowering of temperature due to cold spells by more than 1.5°C on average. Its dynamic error is comparable to that of MIROC6, and both are made in majority on the same motif (Greenland high). The third cluster consists of two models, IPSL-CM6A-LR and MIROC6. With the general bias removed, the temperature value from the reanalysis is within the internal variability of both these models. They are also associated with the lowest dynamic error. This cluster appears to be closest to the reanalysis. On average, IPSL-CM6A-LR has a slightly lower dynamic error than MIROC6, but the difference is lower than internal variability.

Greenland high (motif 3) is the least well-represented motif on more than 80% of MIROC6 and ACCESS-ESM1.5 runs, as well as 45% of ISPL-CM6A-LR runs. However, this does not signify a major model error in the local atmospheric circulation, as the relative error is small, and the significant contribution simply reflects the predominance of the motif in the

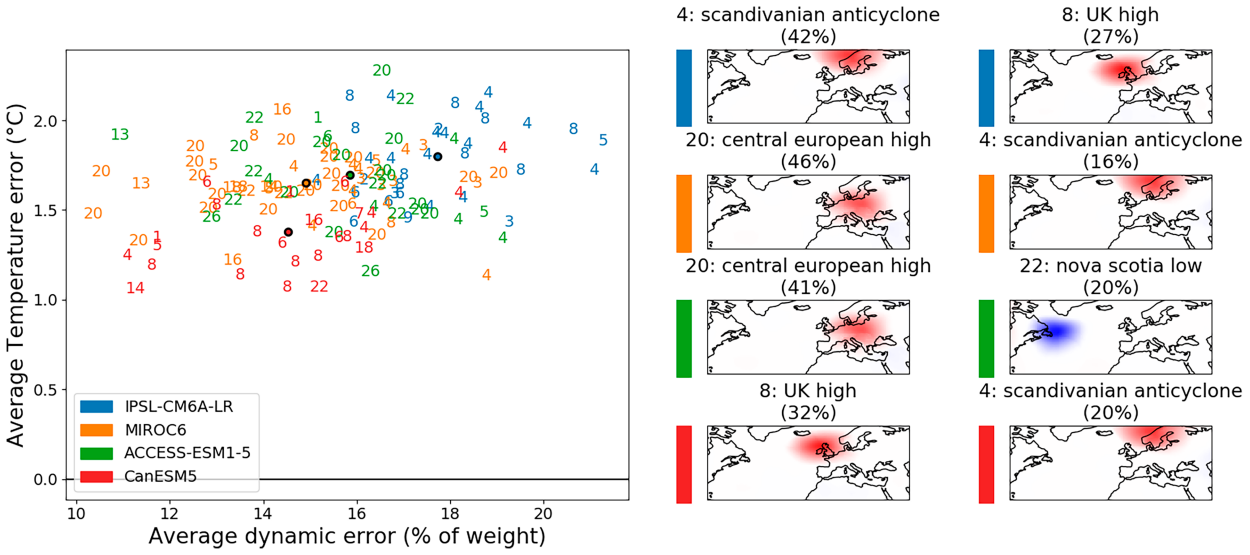

FIG. 13. The run-average temperature model error (average temperature difference with reanalysis) on heat waves in France vs the run-average dynamic model error (average motif weights difference with reanalysis) on same extremes. We eliminate the errors computed in the general case, so as to look only at errors specific to extreme events. The colored dots indicate the run-average value. Each number corresponds to the motif contributing the most to dynamic error in a given run. The two most frequent such motifs for each model are displayed on the right.

composition of cold spells. In contrast, for a majority of CanESM5 runs, as well as 24% of IPSL-CM6A-LR runs, the largest contribution to the dynamic error is due to Cyprus low (motif 10). It is not a particularly dominant motif, but one on which the model makes a significant relative error (75% on median, see Fig. 8). Again, this suggests a major flaw in the model representation of local circulation over the Mediterranean.

In Fig. 13, we plot the model anomalous temperature error against the model anomalous dynamic error in the case of heat waves occurring in France. The inner variability of the models for heat waves is similar to the cold spell case. However, both temperature and dynamic biases associated with the models are closer, so that in the 2D error space, regions occupied by each model are overlapping. All four models are associated with similar temperature errors, between +1.0° and +2.5°C—as these biases are all positive, they cannot be removed by the use of a multimodel mean.

Still, there were important differences between these models. CanESM5 has the lowest of both types of error on average and IPSL-CM6A-LR the highest, but the differences are lower than model internal variabilities. In addition, motifs that contribute the most to the error vary significantly more from run to run than for both general data and cold spells. In particular, no motif dominates the error in a majority of runs of any model, although some appear more often than others. The central European high (motif 20) appears most frequently as the most significant contributor to the error in runs of both MIROC6 and ACCESS-ESM1.5, while Scandinavian anticyclone (motif 4) makes the largest error contributions in multiple runs of IPSL-CM6A-LR, MIROC6, and CanESM5. However, we note that both central European high and Scandinavian anticyclone are dominant motifs in heat waves, so their presence does not reflect a significant relative motif error in the models. To sum up, all models appear to perform comparably for the representation of heat waves, and it seems difficult to identify specific error characteristics in the models.

6. Conclusions

In this paper, we use a statistical learning method called latent Dirichlet allocation (LDA) to study the circulation dynamics of ERA5 reanalysis data and CMIP6 general circulation models. Applied to sea level pressure fields of the North Atlantic region from ERA5 data, LDA yields a set of sparse latent variables called "motifs" that are recognizable localized synoptic-scale meteorological objects, such as cyclones and anticyclones. By projecting daily sea level pressure data onto this basis, we obtain the motif weights, which provide a sparse, low-dimensional representation of atmospheric circulation that can be physically interpreted as the associated synoptic configuration. We showed that synoptic configurations averaged over cold spells and heat waves were both different from each other and from the average taken over the full data.

Using this reanalysis motif basis, we computed the synoptic configuration of runs from four different CMIP6 models. Evaluation of the models was based on comparing the statistics of model synoptic configurations with that of reanalysis ones. Differences between models and reanalyses could then be directly attributed to changes in the average weights of individual motifs. This local characterization of the circulation could help discriminate between model predictions and also help identify the origin of model limitations. Generally speaking, a good agreement was found for general data, while

discrepancies were larger for extreme events. In all cases, the largest source of model error was due to the circulation over the eastern Mediterranean region. Moreover, all models tended to underestimate the changes in atmospheric circulation associated with heat waves.

A global dynamic error, based on synoptic configuration differences with reanalysis, was compared with a temperature error, defined as differences in average temperature. These two indicators were found to be sufficient to help discriminate between models when considering general data. Discriminating between models was still possible in the cold spell case, while models performed comparably on heat waves. This method could, therefore, be used to determine whether specific models are best suited to the study of a given type of event. Characterization of the error is also relevant to knowing how to aggregate model data and identifying the biases that can be eliminated this way.

Acknowledgments. This work is supported by CNRS-MITI (80 PRIME project ACLIM). DF was funded by COST Action FutureMed CA22162, supported by COST (European Cooperation in Science and Technology); an INSU-CNRS-LEFE-MANU grant (project CROIRE); the European Union Horizon 2020 research and innovation program under Grant Agreement 101003469 (XAIDA); and Marie Skłodowska-Curie Grant Agreement 956396 (EDIPI). We thank Robin Noyelle, Camille Cadiou, and Mireia Ginesta-Fernandez for their help in processing the data. We also thank Lucas Fery for his work on the application of LDA to climate data.

Data availability statement. This work makes use of the Gensim Python module, which is publicly available for download through the pip interface. ERA5 reanalysis data are made publicly available by the Copernicus program (https://doi.org/10.24381/cds.143582cf). Model datasets used in this article are simulations from the CMIP6 project for the "historical" experiment, by the models IPSL-CM6A-LR, MIROC6, ACCESS-ESM1.5, and CanESM5. The data are publicly available, thanks to the World Climate Research Programme, and can be found at https://esgf-node.llnl.gov/search/cmip6.

REFERENCES

Alexander, L. V., and Coauthors, 2006: Global observed changes in daily climate extremes of temperature and precipitation. *J. Geophys. Res.*, **111**, D05109, https://doi.org/10.1029/2005JD006290.

Anagnostopoulou, C., K. Tolika, G. Lazoglou, and P. Maheras, 2017: The exceptionally cold January of 2017 over the Balkan Peninsula: A climatological and synoptic analysis. *Atmosphere*, **8**, 252, https://doi.org/10.3390/atmos8120252.

Añel, J. A., M. Fernández-González, X. Labandeira, X. López-Otero, and L. De la Torre, 2017: Impact of cold waves and heat waves on the energy production sector. *Atmosphere*, **8**, 209, https://doi.org/10.3390/atmos8110209.

Ardabili, S., A. Mosavi, M. Dehghani, and A. R. Várkonyi-Kóczy, 2020: Deep learning and machine learning in hydrological processes climate change and Earth systems a systematic review. *Engineering for Sustainable Future*, A. R. Várkonyi-Kóczy, Ed., Springer International Publishing, 52–62, https://doi.org/10.1007/978-3-030-36841-8_5.

Blei, D. M., A. Y. Ng, and M. I. Jordan, 2003: Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, **3**, 993–1022.

Bommer, P. L., M. Kretschmer, A. Hedström, D. Bareeva, and M. M.-C. Höhne, 2024: Finding the right XAI method—A guide for the evaluation and ranking of explainable AI methods in climate science. *Artif. Intell. Earth Syst.*, **3**, e230074, https://doi.org/10.1175/AIES-D-23-0074.1.

Boucher, O., and Coauthors, 2020: Presentation and evaluation of the IPSL-CM6A-LR climate model. *J. Adv. Model. Earth Syst.*, **12**, e2019MS002010, https://doi.org/10.1029/2019MS002010.

Chan, P. W., J. L. Catto, and M. Collins, 2022: Heatwave–blocking relation change likely dominates over decrease in blocking frequency under global warming. *npj Climate Atmos. Sci.*, **5**, 68, https://doi.org/10.1038/s41612-022-00290-2.

Chen, L., and Coauthors, 2023: Artificial intelligence-based solutions for climate change: A review. *Environ. Chem. Lett.*, **21**, 2525–2557, https://doi.org/10.1007/s10311-023-01617-y.

Cohen, J., D. Coumou, J. Hwang, L. Mackey, P. Orenstein, S. Totz, and E. Tziperman, 2019: S2S reboot: An argument for greater inclusion of machine learning in subseasonal to seasonal forecasts. *Wiley Interdiscip. Rev.: Climate Change*, **10**, e00567, https://doi.org/10.1002/wcc.567.

D'Andrea, F., and Coauthors, 2024: Summer deep depressions increase over the eastern North Atlantic. *Geophys. Res. Lett.*, **51**, e2023GL104435, https://doi.org/10.1029/2023GL104435.

Davini, P., and F. D'Andrea, 2020: From CMIP3 to CMIP6: Northern Hemisphere atmospheric blocking simulation in present and future climate. *J. Climate*, **33**, 10 021–10 038, https://doi.org/10.1175/JCLI-D-19-0862.1.

de Freitas, C. R., and E. A. Grigorieva, 2017: A comparison and appraisal of a comprehensive range of human thermal climate indices. *Int. J. Biometeor.*, **61**, 487–512, https://doi.org/10.1007/s00484-016-1228-6.

Eyring, V., S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor, 2016: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.*, **9**, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016.

Fang, W., Q. Xue, L. Shen, and V. S. Sheng, 2021: Survey on the application of deep learning in extreme weather prediction. *Atmosphere*, **12**, 661, https://doi.org/10.3390/atmos12060661.

Faranda, D., G. Masato, N. Moloney, Y. Sato, F. Daviaud, B. Dubrulle, and P. Yiou, 2016: The switching between zonal and blocked mid-latitude atmospheric circulation: A dynamical system perspective. *Climate Dyn.*, **47**, 1587–1599, https://doi.org/10.1007/s00382-015-2921-6.

——, G. Messori, and P. Yiou, 2017: Dynamical proxies of North Atlantic predictability and extremes. *Sci. Rep.*, **7**, 41278, https://doi.org/10.1038/srep41278.

Fery, L., B. Dubrulle, B. Podvin, F. Pons, and D. Faranda, 2022: Learning a weather dictionary of atmospheric patterns using Latent Dirichlet Allocation. *Geophys. Res. Lett.*, **49**, e2021GL096184, https://doi.org/10.1029/2021GL096184.

Fink, A. H., T. Brücher, A. Krüger, G. C. Leckebusch, J. G. Pinto, and U. Ulbrich, 2004: The 2003 European summer heatwaves and drought—Synoptic diagnosis and impacts. *Weather*, **59**, 209–216, https://doi.org/10.1256/wea.73.04.

Frich, P., L. V. Alexander, P. Della-Marta, B. Gleason, M. Haylock, A. M. G. K. Tank, and T. Peterson, 2002: Observed coherent changes in climatic extremes during the second half

of the twentieth century. *Climate Res.*, **19**, 193–212, https://doi.org/10.3354/cr019193.

Frihat, M., B. Podvin, L. Mathelin, Y. Fraigneau, and F. Yvon, 2021: Coherent structure identification in turbulent channel flow using Latent Dirichlet allocation. *J. Fluid Mech.*, **920**, A27, https://doi.org/10.1017/jfm.2021.444.

Gardoll, S., and O. Boucher, 2022: Classification of tropical cyclone containing images using a convolutional neural network: Performance and sensitivity to the learning dataset. *Geosci. Model Dev.*, **15**, 7051–7073, https://doi.org/10.5194/gmd-15-7051-2022.

Hanley, D. E., M. A. Bourassa, J. J. O'Brien, S. R. Smith, and E. R. Spade, 2003: A quantitative evaluation of ENSO indices. *J. Climate*, **16**, 1249–1258, https://doi.org/10.1175/1520-0442(2003)16<1249:AQEOEI>2.0.CO;2.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quart. J. Roy. Meteor. Soc.*, **146**, 1999–2049, https://doi.org/10.1002/qj.3803.

Hoffman, M., F. Bach, and D. Blei, 2010: Online learning for Latent Dirichlet Allocation. *Advances in Neural Information Processing Systems 23*, Curran Associates, Inc., 856–864, https://proceedings.neurips.cc/paper_files/paper/2010/hash/71f6278d140af599e06ad9bf1ba03cb0-Abstract.html.

Jeong, D. I., B. Yu, and A. J. Cannon, 2021: Links between atmospheric blocking and North American winter cold spells in two generations of Canadian Earth System Model large ensembles. *Climate Dyn.*, **57**, 2217–2231, https://doi.org/10.1007/s00382-021-05801-0.

Kharin, V. V., F. W. Zwiers, X. Zhang, and M. Wehner, 2013: Changes in temperature and precipitation extremes in the CMIP5 ensemble. *Climatic Change*, **119**, 345–357, https://doi.org/10.1007/s10584-013-0705-8.

Krishnamurti, T. N., 1961: The subtropical jet stream of winter. *J. Atmos. Sci.*, **18**, 172–191, https://doi.org/10.1175/1520-0469(1961)018<0172:TSJSOW>2.0.CO;2.

Li, C., F. Zwiers, X. Zhang, G. Li, Y. Sun, and M. Wehner, 2021: Changes in annual extremes of daily temperature and precipitation in CMIP6 models. *J. Climate*, **34**, 3441–3460, https://doi.org/10.1175/JCLI-D-19-1013.1.

Liu, Y., and Coauthors, 2016: Application of deep convolutional neural networks for detecting extreme weather in climate datasets. arXiv, 1605.01156v1, https://doi.org/10.48550/arXiv.1605.01156.

López-Bueno, J. A., M. Á. Navas-Martín, J. Díaz, I. J. Mirón, M. Y. Luna, G. Sánchez-Martínez, D. Culqui, and C. Linares, 2021: The effect of cold waves on mortality in urban and rural areas of Madrid. *Environ. Sci. Europe*, **33**, 72, https://doi.org/10.1186/s12302-021-00512-z.

Lucas-Picher, P., D. Argüeso, E. Brisson, Y. Tramblay, P. Berg, A. Lemonsu, S. Kotlarski, and C. Caillaud, 2021: Convection-permitting modeling with regional climate models: Latest developments and next steps. *Wiley Interdiscip. Rev.: Climate Change*, **12**, e731, https://doi.org/10.1002/wcc.731.

Lupo, A. R., 2021: Atmospheric blocking events: A review. *Ann. N. Y. Acad. Sci.*, **1504**, 5–24, https://doi.org/10.1111/nyas.14557.

McCarthy, M., and Coauthors, 2019: Drivers of the UK summer heatwave of 2018. *Weather*, **74**, 390–396, https://doi.org/10.1002/wea.3628.

Meehl, G. A., G. J. Boer, C. Covey, M. Latif, and R. J. Stouffer, 2000: The Coupled Model Intercomparison Project (CMIP). *Bull. Amer. Meteor. Soc.*, **81**, 313–318.

Michelangeli, P.-A., R. Vautard, and B. Legras, 1995: Weather regimes: Recurrence and quasi stationarity. *J. Atmos. Sci.*, **52**, 1237–1256, https://doi.org/10.1175/1520-0469(1995)052<1237:WRRAQS>2.0.CO;2.

Mignot, J., and Coauthors, 2021: The tuning strategy of IPSL-CM6A-LR. *J. Adv. Model. Earth Syst.*, **13**, e2020MS002340, https://doi.org/10.1029/2020MS002340.

O'Loughlin, R., D. Li, and T. O'Brien, 2024: Moving beyond post-hoc XAI: Lessons learned from dynamical climate modeling. *EGUsphere*, **2024**, 1–24, https://doi.org/10.5194/egusphere-2023-2969.

Papagiannaki, K., K. Lagouvardos, V. Kotroni, and G. Papagiannakis, 2014: Agricultural losses related to frost events: Use of the 850 hPa level temperature as an explanatory variable of the damage cost. *Nat. Hazards Earth Syst. Sci.*, **14**, 2375–2386, https://doi.org/10.5194/nhess-14-2375-2014.

Řehůřek, R., and P. Sojka, 2010: Software framework for topic modelling with large corpora. *Proc. LREC 2010 Workshop on New Challenges for NLP Frameworks*, Valletta, Malta, ELRA, 45–50, https://is.muni.cz/publication/884893/en.

Reichstein, M., G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and Prabhat, 2019: Deep learning and process understanding for data-driven Earth system science. *Nature*, **566**, 195–204, https://doi.org/10.1038/s41586-019-0912-1.

Ren, X., L. Li, Y. Yu, Z. Xiong, S. Yang, W. Du, and M. Ren, 2020: A simplified climate change model and extreme weather model based on a machine learning method. *Symmetry*, **12**, 139, https://doi.org/10.3390/sym12010139.

Rex, D. F., 1950: Blocking action in the middle troposphere and its effect upon regional climate: II. The climatology of blocking action. *Tellus*, **2A** (4), 275–301, https://doi.org/10.3402/tellusa.v2i4.8603.

Rodrigues, D., M. C. Alvarez-Castro, G. Messori, P. Yiou, Y. Robin, and D. Faranda, 2018: Dynamical properties of the North Atlantic atmospheric circulation in the past 150 years in CMIP5 models and the 20CRv2c reanalysis. *J. Climate*, **31**, 6097–6111, https://doi.org/10.1175/JCLI-D-17-0176.1.

Rousi, E., K. Kornhuber, G. Beobide-Arsuaga, F. Luo, and D. Coumou, 2022: Accelerated western European heatwave trends linked to more-persistent double jets over Eurasia. *Nat. Commun.*, **13**, 3851, https://doi.org/10.1038/s41467-022-31432-y.

Salcedo-Sanz, S., and Coauthors, 2024: Analysis, characterization, prediction, and attribution of extreme atmospheric events with machine learning and deep learning techniques: A review. *Theor. Appl. Climatol.*, **155** (1), 1–44, https://doi.org/10.1007/s00704-023-04571-5.

Scaife, A. A., T. Woollings, J. Knight, G. Martin, and T. Hinton, 2010: Atmospheric blocking and mean biases in climate models. *J. Climate*, **23**, 6143–6152, https://doi.org/10.1175/2010JCLI3728.1.

Seneviratne, S. I., and Coauthors, 2021: Weather and climate extreme events in a changing climate. *Climate Change 2021: The Physical Science Basis*, V. Masson-Delmotte et al., Eds., Cambridge University Press, 1513–1766, https://doi.org/10.1017/9781009157896.013.

Stenseth, N. C., G. Ottersen, J. W. Hurrell, A. Mysterud, M. Lima, K.-S. Chan, N. G. Yoccoz, and B. Ådlandsvik, 2003: Studying climate effects on ecology through the use of climate indices: The North Atlantic Oscillation, El Niño Southern Oscillation and beyond. *Proc. Roy. Soc.*, **270B**, 2087–2096, https://doi.org/10.1098/rspb.2003.2415.

Swart, N. C., and Coauthors, 2019: The Canadian Earth System Model version 5 (CanESM5.0.3). *Geosci. Model Dev.*, **12**, 4823–4873, https://doi.org/10.5194/gmd-12-4823-2019.

Tatebe, H., and Coauthors, 2019: Description and basic evaluation of simulated mean state, internal variability, and climate sensitivity in MIROC6. *Geosci. Model Dev.*, **12**, 2727–2765, https://doi.org/10.5194/gmd-12-2727-2019.

Valle, D., P. Albuquerque, Q. Zhao, A. Barberan, and R. J. Fletcher Jr., 2018: Extending the Latent Dirichlet Allocation model to presence/absence data: A case study on North American breeding birds and biogeographical shifts expected from climate change. *Global Change Biol.*, **24**, 5560–5572, https://doi.org/10.1111/gcb.14412.

van Oldenborgh, G. J., S. Drijfhout, A. van Ulden, R. Haarsma, A. Sterl, C. Severijns, W. Hazeleger, and H. Dijkstra, 2009: Western Europe is warming much faster than expected. *Climate Past*, **5** (1), 1–12, https://doi.org/10.5194/cp-5-1-2009.

Vautard, R., 1990: Multiple weather regimes over the North Atlantic: Analysis of precursors and successors. *Mon. Wea. Rev.*, **118**, 2056–2081, https://doi.org/10.1175/1520-0493(1990)118<2056:MWROTN>2.0.CO;2.

——, and Coauthors, 2023a: Heat extremes in Western Europe increasing faster than simulated due to atmospheric circulation trends. *Nat. Commun.*, **14**, 6803, https://doi.org/10.1038/s41467-023-42143-3.

——, and Coauthors, 2023b: Human influence on growing-period frosts like in early April 2021 in central France. *Nat. Hazards Earth Syst. Sci.*, **23**, 1045–1058, https://doi.org/10.5194/nhess-23-1045-2023.

Wallace, J. M., and P. V. Hobbs, 2006: *Atmospheric Science: An Introductory Survey*. Vol. 92. Elsevier, 504 pp.

Wehrli, K., B. P. Guillod, M. Hauser, M. Leclair, and S. I. Seneviratne, 2018: Assessing the dynamic versus thermodynamic origin of climate model biases. *Geophys. Res. Lett.*, **45**, 8471–8479, https://doi.org/10.1029/2018GL079220.

Weilnhammer, V., J. Schmid, I. Mittermeier, F. Schreiber, L. Jiang, V. Pastuhovic, C. Herr, and S. Heinze, 2021: Extreme weather events in Europe and their health consequences—A systematic review. *Int. J. Hyg. Environ. Health*, **233**, 113688, https://doi.org/10.1016/j.ijheh.2021.113688.

Ziehn, T., and Coauthors, 2020: The Australian Earth system model: ACCESS-ESM1.5. *J. South. Hemisphere Earth Syst. Sci.*, **70**, 193–214, https://doi.org/10.1071/ES19035.