scientific reports

OPEN

Metastability and teleconnection of atmospheric circulation via hidden Markov models and network modularity

Dmitry Mukhin[1]✉, Roman Samoilov[1] & Abdel Hannachi[2]✉

The low-frequency variability of the mid-latitude atmosphere involves complex nonlinear and chaotic dynamical processes posing predictability challenges. It is characterized by sporadically recurring, often long-lived patterns of atmospheric circulation of hemispheric scale known as weather regimes. The evolution of these circulation regimes in addition to their link to large-scale teleconnections can help to extend the limits of atmospheric predictability. They also play a key role in sub- and inter-seasonal weather forecasting. Their identification and modeling remains an issue, however, due to their intricacy, including a clear conceptual picture. In recent years, the concept of metastability has been developed to explain regimes formation. This suggests an interpretation of circulation regimes as communities of states in the neighborhood of which the atmospheric system remains abnormally longer than typical baroclinic timescales. Here we develop a new and effective method to identify such communities by constructing and analyzing an operator of the system's evolution via hidden Markov model (HMM). The method makes use of graph theory and is based on probabilistic approach to partition the HMM transition matrix into weakly interacting blocks – communities of hidden states – associated with regimes. The approach involves nonlinear kernel principal component mapping to consistently embed the system state space for HMM building. Application to northern winter hemisphere using geopotential heights from reanalysis yields four persistent and recurrent circulation regimes. Statistical and dynamical characteristics of these circulation regimes and surface impacts are discussed. In particular, unexpected high correlations are obtained with EL-Niño Southern Oscillation and Pacific decadal oscillation with lead times of up to one year.

Humans have, since the dawn of civilization from the Babylonian to present day, sought to predict the weather far in the future. The theory of modern mathematical models of nonlinear dynamical systems show that weather cannot be forecast accurately beyond $1 - 2$ weeks ahead. However, this is not the full story. In fact, the atmosphere exhibits complex variation on a wide spectrum of spatio-temporal scales ranging from synoptic to decadal and longer time scales originating from internal nonlinear dynamics and also from the boundary conditions such as sea surface temperature (SST) variation. These sources can contribute to extending the chaotic predictability limits to include subseasonal-to-seasonal (S2S) timescales.

Low-frequency variability (LFV) of the atmosphere, spanning spatio-temporal scales greater than typical synoptic and baroclinic timescales is poorly understood and is difficult to predict. Extratropical LFV, in particular, involves recurring and persistent large-, e.g. hemispheric-scale structures. These structures of atmospheric circulation are often referred to as teleconnection patterns or regimes and are known to have profound impact on weather predictability on intraseasonal and longer timescales[1,2]. With such scales, the ensuing circulation anomalies strongly modulate the tracks and intensity of synoptic disturbances in the atmosphere, operating like waveguides, and hence determine the long-term behavior of weather over large areas. The modeling and analysis of the mid-latitude climate is closely related to recurring circulation regimes. Both the identification and the understanding of their dynamics continue to be a debatable issue due to the complexity and nonlinearity of the problem beside the unsettled concept of regimes.

Quasi-geostrophic theory of large-scale atmospheric circulation shows that LFV is essentially the result of nonlinear dynamics[3]. Experimental analyses based on quasi-geostrophic barotropic and baroclinic models reveal the high-dimensional nature and chaotic dynamics of the LFV atmospheric flow[4]. There is suggestion in

[1]Institute of Applied Physics of RAS, Nizhny Novgorod 603950, Russia. [2]Department of Meteorology, Stockholm University, Stockholm SE-106 91, Sweden. ✉email: mukhin@ipfran.ru; abdel.hannachi@misu.su.se

the literature[5] that in an infinite set of saddle-invariant manifolds of a chaotic attractor, the most stable of those manifolds form the skeleton of the dynamics, and sporadic adherence of the system state in their neighborhood can cause regime persistence. This nonlinear perspective leads to the understanding of regimes as metastable regions in the system's phase space where the system slows down. To consistently adopt this paradigm of *persistent* regimes, an extra requirement on regime detection methodology has to be added, namely, it should account for temporal relationships of observed states rather than just their similarity in a selected projection of phase space. A number of existing methods[6] for identifying recurrent patterns, e.g. based on PDF decomposition or partitioning networks of similarity, do not satisfy this requirement.

An appropriate and elegant way to address this issue is to construct and analyze an evolution operator of the system using hidden Markov model (HMM)[7,8]. A conventional Markov chain is a stochastic model based on 'observed' states of the chain. An HMM[7,9] is an extension of the Markov chain by including other (*hidden*) states, which generate the observed sequence. Because of the finite and short-memory representation of the system, HMM is a consistent method for detecting metastability as well as identifying the basins of metastable states. Essentially, with HMM, nonlinear effects, like metastability and trace of multiple equilibria, can be captured by linear transfer operator pretty much like dynamic mode decomposition (DMD)[10]. HMM has been used to study atmospheric regimes via spectral analysis of the Markovian operator[8,11].

Inspired by ideas from[8], we elaborate a new methodology for constructing and analyzing HMM aimed at detecting persistent and recurrent circulation regimes determined as metastable communities of hidden states. We derive and use a method of partitioning the HMM transition probability matrix into blocks, providing significantly low probabilities of transitions between them. The method is based on maximizing a cost-function for a given partition and has a clear probabilistic meaning. In fact, the approach finds that partition with abnormally high mean probability that the state will remain within its community, i.e. persistent. We show that this optimization problem effectively reduces to the Newman iterative algorithm, widely used in graph theory for graph decomposition[12]. We pay special attention to the significance test, which rejects the null hypothesis that the detected communities could have been obtained by chance from a sample of a randomized process with similar spectral properties.

Another key part of the methodology is the choice of an embedding space, i.e. a set of dynamic variables for dimension reduction and model construction. This is of paramount importance because it determines the projection of the phase space in which the model operates. The dimension of this space is desired to be as low as possible to ensure stable results during the HMM learning process, but at the same time it should be informative enough for regime detection and identification. Kernel principal component analysis (KPCA)[6,13] is considered here, which belongs to the class of nonlinear data transformation, and has been shown to enable a reliable and robust separation of states based on similarity measure in a space of just few leading components. Recent works[14,15] used it to study recurrent patterns of the mid-latitude atmospheric circulation. In this manuscript, KPCA serves as a prepossessing step for the HMM to get both recurrent[1] and persistent circulation flows.

The paper is organized as follows. In Section "Methods" we describe the methodology of HMM building, regimes detection and analysis, with an illustration using a low-order chaotic dynamical system. Section "Results" presents the properties of circulation regimes obtained from historical climate data. Finally, in Section "Discussion", we summarize and discuss advantages and prospects of the presented methodology and results. Model examples together with additional details on the methodology as well as additional illustrations are detailed in the Supplementary Material.

Methods
Hidden Markov model
In the present setup the observed atmospheric variability is represented by time series of some dynamic variables $X = (x_1, x_2, \ldots, x_d)$. For example, in case of spatially distributed data, as is the case here, the embedding space for X can be constructed by means of projecting the data onto a lower dimensional space such as the linear empirical orthogonal functions (EOFs) or using a more sophisticated nonlinear functions, e.g., kernel principal components (see Sec. 2.6). We assume that the evolution of these variables is driven by transitions between a finite number of hidden states of the system, $J = 1, 2, \ldots, K$, where J denotes a state number. The transitions are random and are modeled by a discrete-time Markov chain defined by the transition probability matrix $Q_{ij} = p(J_i|J_j)$, which is a stochastic matrix, i.e. has a unit sum along each column ($\sum_i Q_{ij} = 1$). Each hidden state is connected with the dynamic variables via emission probabilities $P(X|J)$. Here we use Gaussian probability density function (PDF) for the emissions:

$$P(X|J) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_J|}} \exp\left(-\frac{1}{2}(X - a_J)^T \Sigma_J^{-1}(X - a_J)\right), \tag{1}$$

although more complex PDFs can be employed. For example, using Gaussian mixture models as emission probabilities would allow us to obtain hidden states that have a more complex distribution in the data space. One would expect a smaller optimal number of hidden states than for the simple Gaussian emissions due to the more efficient approximation of phase space. The price to pay is an increase in the number of parameters corresponding to each hidden state that may lead to an increase in the optimization problem complexity. Such a generalization of HMM in context of atmospheric circulation modeling is an interesting task for future investigations.

The described HMM is fully determined by its parameters $\mu = (Q, \Sigma, a)$, which are to be estimated from the observed time series $(X_1, X_2, \ldots, X_N)$. The likelihood function for HMM contains the sum over huge

number of possible sequences of the hidden states $S_1, S_2, \ldots$, where each S_k contains N_k hidden states $S_k = \{J_1, J_2, \ldots, J_{N_k}\}$, with $J_{N_k} < N$:

$$P(X_1, \ldots, X_N | \mu) = \sum_k P(X_1, \ldots, X_N, |S_k, \Sigma, a) P(S_k | Q),$$

(2)

and hence, cannot be calculated and maximized directly. Here we use Baum-Welch algorithm[16] that provides convergence to a local maximum of Eq. (2). This is an iterative EM-algorithm, based on repeating two steps: (i) estimating the state occupation probabilities (E-step) and (ii) estimating the HMM parameters μ using the occupation probabilities (M-step). Thus, at the end of this procedure, we obtain the estimates of transition probabilities, position and orientation of emission functions in the phase space of dynamic variables X, in addition to the occupation probability of the observation X_t in the hidden state $J = j$: $P(J = j | X_t)$.

Since we are to train the HMM from natural stochastic time series, we will almost surely obtain a Markov chain that does not have both fully closed communities (classes) of states and pure periodic states. This guarantees the existence of a unique stationary distribution π to which the Markov process converges starting from any initial distribution q:

$$\lim_{k \to \infty} Q^k q = \pi,$$

(3)

or, equivalently, π is the unique right eigenvector of the matrix Q corresponding to its largest eigenvalue which is exactly equal to 1: $Q\pi = \pi$. One more assumption we use in Eq. (3) is that the transition matrix does not depend on time, i.e. a stationary process. Although this assumption seems relevant in case of quite short atmospheric time series (after removing the regression on the CO2 trend; see the result section), it should be relaxed when studying a nonlinear response of the atmosphere to external forcing.

The number of hidden states K is a structural parameter of the HMM, that determines the model resolution. Generally speaking, it is recommended to set this parameter as large as possible, provided that it is adequate to available statistics, i.e. each hidden state is reasonably well visited in the observed sample. While a too-coarse model with unreasonably small K may miss important features of the metastable state space, an overfitted model with too large K will have a nearly binary transition matrix Q that describes the unique given sample rather than the underlying system, i.e. overfitting. Since the Baum-Welch algorithm used for learning HMM allows us to calculate the likelihood Eq. (2), the optimal K can be roughly estimated based on the conventional Akaike information criterion (AIC). The idea behind is that the large-scale structure of the metastable regimes we are trying to find should be insensitive to K when K greatly exceeds the number of regimes (e.g., saturation), and any value near the AIC optimum provides a good model for regime detection.

Modularity, graphs and persistent regimes

According to Franzke et al.[8], an atmosphere circulation regime is associated with a metastable state of the system, i.e. a region of slow evolution of the trajectory in the system's phase space. With the HMM representation of the atmospheric circulation, a regime can be sought as a block (a subset of the hidden states) of the transition matrix with small probabilities of transitions to the states out of the block. Decomposing the full transition matrix Q into such blocks, some sort of tiling, gives us a reduced matrix of transition probabilities between the basins of stability in the space of hidden states. Going this way, we treat the HMM hidden states as microstates of the system[11], meanwhile a regime is determined as a community of the microstates. Such an approach helps reduce the influence of a specific sample on the result. The coverage of the data space by the hidden states, their size and location, can depend on the data sample we have. For example, the longer the time series on which the model is trained, the more hidden states we can expect provided that each of them is well visited by the observations. This means that each individual microstate reflects particular observations rather than the dynamical properties of the system and therefore lacks physical interpretation. In our study, the microstates serve only as a support for representing fundamental, data-independent invariant structures in the system's phase space, that underlie the metastable regimes. In other words, the regimes must be less sensitive to the sample, since they are determined as unions of many microstates, and, therefore not so sensitive to their specific positions and sizes.

Now we can safely assume that a sign of metastability in a Markov chain appears when there exists a partition $\{A_k\}$ of the hidden states such that the average probability of remaining in a given subset A_k is larger, compared to that in a completely randomized chain having the same stationary distribution π. Thus, to find the best partition $\{A_k\}$, we should maximize the quantity:

$$\langle P_Q(J \in A_k | J \in A_k) - P_{Q^\infty}(J \in A_k | J \in A_k) \rangle_{A_k}$$
$$= \sum_k [P_Q(J \in A_k | J \in A_k) - \pi(A_k)] \pi(A_k),$$

(4)

where $P_Q(a|b)$ is the probability of an event a given an event b at the previous state in the Markov chain Q; $\pi(A_k)$ is the stationary probability of a subset A_k, i.e. $\pi(A_k) = \sum_{i \in A_k} \pi_i$; the chain Q^∞ is the white-noise process with the stationary distribution π, which is generated by the transition matrix Q^∞ having identical columns given by π. The obtained largest value of Eq. (4) should be positive to consider the partition viable, otherwise we conclude that there are no persistent regimes states.

Noting that

$$P_Q(J \in A_k | J \in A_k) = \frac{1}{\pi(A_k)} \sum_{i,j \in A_k} Q_{ij}\pi_j,$$

we can rewrite Eq. 4 in the following form:

$$M = \sum_{A_k} \sum_{i,j \in A_k} [Q_{ij}\pi_j - \pi_i\pi_j] = \sum_{i,j} [Q_{ij}\pi_j - \pi_i\pi_j] g_{ij} = \sum_{i,j} B_{ij}g_{ij}, \tag{5}$$

where $g_{ij} = 1$ if hidden states $J = i$ and $J = j$ belong to the same subset A_k and $g_{ij} = 0$ otherwise. Written this way, and this is the novelty here, the expression M in Eq. (5) is precisely similar to the modularity introduced by Newman[12], which stems from graph theory[17,18]. The modularity was previously used by[14] for detecting atmospheric regimes in connection with recurrence networks. One of the attractive feature of this formulation, Eq. (5), is the possibility of using the iterative splitting algorithm to maximize M simply via solving for an eigenvalue problem. The algorithm proceeds by iteratively splitting each subset A_k into two communities so that each split provides the maximal positive increment of the modularity ΔM, until indivisible subsets are obtained. Splitting of some subset A into two groups can be represented by a classifier – a vector s of dimension $|A|$ (size of A) whose elements $s_i = -1$ (for the first group) and $s_i = 1$ (for the second group), $i = 1, \ldots |A|$). Now denoting by $B^{(A)}$ the submatrix of $B = (B_{ij})$ obtained by selecting the elements of B with the indices $i, j \in A$, it can be shown that the increment ΔM of the whole modularity, after this split, takes the form:

$$\Delta M = \frac{1}{2} \sum_{i,j} (B_A)_{ij}(s_i s_j + 1) = \frac{1}{2} s^T B_A s,$$

$$(B_A)_{ij} = B_{ij}^{(A)} - \delta_{ij} \sum_k B_{ik}^{(A)}, \tag{6}$$

where δ_{ij} is the Kronecker delta. The algorithm is initiated by setting A equal to the whole set of hidden states, that is $B_A = B$.

It is worth noting that the matrix B_A is symmetric only when the Markov chain obeys the detailed balance condition, i.e. $Q_{ij}\pi_j = Q_{ji}\pi_i$. However, this is not true for a general dynamical system and is not required in the HMM training procedure. The interesting point here, another advantage of this formulation, is that in the general non-symmetric situation, finding the binary vector s that maximizes ΔM, makes use of the fact that the quadratic form in Eq. (6) depends only on the symmetric part of the matrix B_A, that is $H = \frac{1}{2}\left(B_A + B_A^T\right)$, and the change in the modularity can simply be expressed in terms of the eigen-elements of H:

$$\Delta M_A = \frac{1}{2} s^T H s = \frac{1}{2} \lambda_1 (w_1^T s)^2 + \ldots, \tag{7}$$

where λ_1 is the largest eigenvalue of H and w_1 the corresponding eigenvector. Eq. (7) tells us that the binary vector maximizing the modularity increment ΔM_A is precisely given by the sign of w_1, $s = sign(w_1)$. In other words, the optimum vector s is the one whose $i'th$ component s_i is 1 if the corresponding component of w_1 is positive and -1 otherwise. A key conclusion follows from this, namely, if the matrix H has positive eigenvalues, two new subsets defined by the vector s emerge, provided that $\Delta M_A > 0$. Otherwise, its largest eigenvalue is zero, since this matrix has the zero sum over each row or column, and no further splitting of the community A is possible.

As provided by maximizing the modularity Eq. 4, the probability of staying in each regime A_k is relatively high, but the residence time of the system in individual hidden states belonging to A_k can differ. Eq. (7) shows that the absolute values $|w_{1i}|$ of the $i'th$ component of w_1 determines the contribution of states with the corresponding classifiers s_i to the modularity increment. This means that the system more often falls into states corresponding to large $|w_{1i}|$, and hence, this value can be treated as a measure of state stability. The most stable states form the cores of the obtained regimes and constitute the skeleton of metastable structure of the system phase space. The metastable states of dynamical systems determine in fact the structure of the system's attractor.

The splitting process continues until the increment ΔM_A ceases to be positive for the subset A. In practical terms a threshold is used for ΔM_A in order to eliminate insignificant splits. The significance test is based on surrogates obtained by random block shuffling of a time series generated by the HMM preserving main spectral properties of the original process (see the Supplementary Material). Given the fixed emission PDFs, Eq. (1), estimated from data, an ensemble of transition matrices is obtained by training the HMM on the ensemble of surrogates. These surrogates preserve the stationary distribution π, but may not preserve autocorrelations at large lags. The ensemble of the modularity increments ΔM_A based on the surrogate matrices is then used to construct a confidence interval of ΔM_A from the original sample. Specifically, when a split is carried out of some community with the original Q, the same block in every surrogate matrix is also split to obtain the null distribution of increments. Thus, the value of increment on the right tail of the surrogate ensemble, that corresponds to the significance level, is used as a threshold for accepting ΔM_A.

Time step selection

An important aspect of HMM is the choice of time step. For example, using daily data successive states are very likely to be highly correlated. As a result, the diagonal elements of the transition matrix will be dominant,

reflecting high persistence in the hidden states, and making partition difficult and the results prone to noise. Traditionally, this weakness is addressed by adjusting the time step via smoothing and/or resampling the data. But the HMM provides another solution to analyze the Markov operator across L days, that is, to decompose the transition matrix Q^L (instead of the one-step Q) where L represents the spatiotemporal resolution of regime determination. This approach is akin to diffusion maps, see e.g.[18–20], which selects the step size of the Markov process to determine the scale of the underlying structure. Moderately large L reduces regime sensitivity to noise. However, quite large L can lead to matrix degeneracy, eventually ending up with rank-1 transition matrix with columns composed of the stationary distribution π. We can expect that persistent states are reliably detected when L is comparable to the characteristic timescale of the regimes, hence the need to look for a range of L for which the regimes do not change significantly. It is worth noting that by adjusting the time step via L, we do not distort the original data as with smoothing and/or resampling, but rather involve all daily observations in model learning. A schematic representation of the method is sketched in Fig. 1a.

Illustration with a low-order chaotic model

Here we demonstrate the identification method using a low-order nonlinear chaotic dynamical system. The system is the classical Lorenz[21] model with stochastic forcing:

$$
\begin{aligned}
\dot{x} &= 10(y - x) + \sigma\xi_x \\
\dot{y} &= x(28 - z) - y + \sigma\xi_y \\
\dot{z} &= xy - \frac{8}{3}z + \sigma\xi_z.
\end{aligned}
\tag{8}
$$

Here $\sigma\xi$ are Gaussian uncorrelated processes with variance $\sigma^2 = 20$. From a 10,000-length trajectory generated by this system with time step 0.1, we consider the time series of x and z as our "observations" shown in Fig. 1b. The (unforced) Lorenz system has 3 equilibria; an unstable center at the origin and two metastable equilibria located at the centers of the butterfly wings. The high density around the origin (Fig. 1b) is not genuine, and is partly due to the fact that this region corresponds to transition path between the two metastable equilibria, and partly to the stochastic forcing. Next, an HMM has been constructed using these time series. As it can be seen from Fig. 1b, the obtained two communities of HMM hidden states represent well the metastable regions, and the most stable hidden states (see Sec. 2.2) lie close to, and around their centers. It is interesting that parts of the outer region of each of the "wings" are assigned to the regime dominating the other wing. This means that starting from these regions, the system goes to another wing with substantially higher probability than to the current wing.

Probabilistic analysis of regimes

The obtained block structure of the matrix Q enables the calculation of reduced matrices $\overline{Q^L}$ with elements defined by:

$$
\overline{Q_{kl}^L} = \left[\sum_{j \in A_l} \pi_j\right]^{-1} \sum_{i \in A_k, j \in A_l} Q_{ij}^L \pi_j.
\tag{9}
$$

The columns of this matrix represent the probability distributions of transitions between different regime states and allow us to predict the preferred transition path from any current regime. In particular, we can estimate the expected lifetime τ_k of the $k'th$ regime:

$$
\tau_k = \overline{Q_{kk}^1}/(1 - \overline{Q_{kk}^1}) + 1.
\tag{10}
$$

Note that the term $(\tau_k - 1)$ coincides with the mean of the geometric probability distribution

$$
P(\tau) = q^\tau(1 - q),
\tag{11}
$$

which represents the probability of a continuous series of length τ for an event with probability q.

When dealing with real high-dimensional data, as the case here, one practically maps the observations Y onto a lower dimensional space X, for HMM application, hence possibly compromising the inverse mapping onto the physical space. This is the case of many nonlinear projections like kernel principal components (see the next section), diffusion maps, etc., for which we cannot derive an exact image of the hidden states and their groups in the data space.

Here, to get anomalies associated with states A_k, we compute a composite pattern $\langle Y|A_k\rangle$, that is the mean of observed value Y given A_k via the obtained estimates of $P(J = j|Y_t)$, i.e. probability that the observation Y at time t belongs to the hidden state $J = j$ (see sec. 1.1). For each Y_t, the regime occupation probability is first calculated:

$$
P(A_k|Y_t) = \sum_{j \in A_k} P(J = j|Y_t),
\tag{12}
$$

which is then used to estimate the composite pattern:
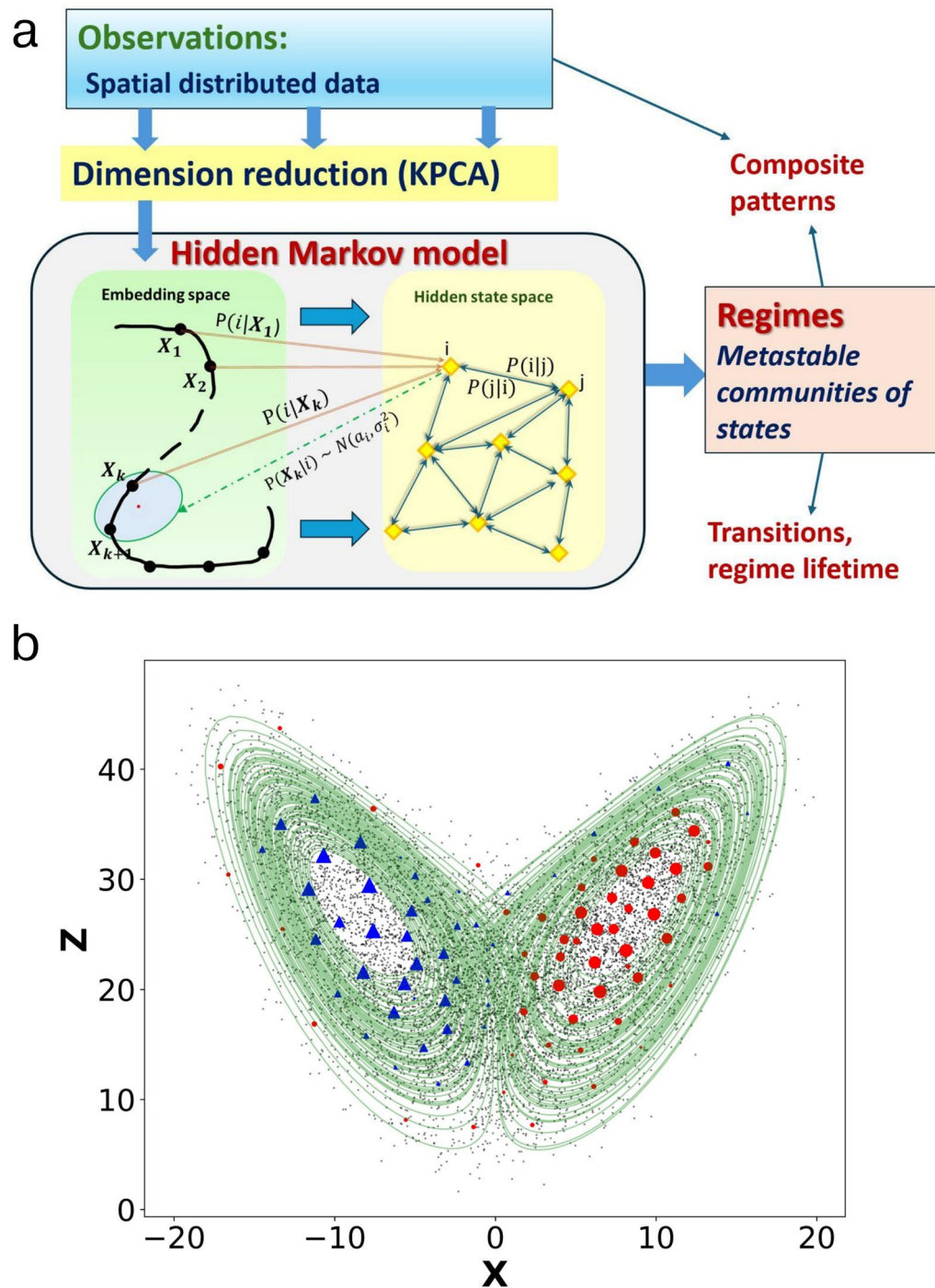
Fig. 1. (**a**) Schematic representation of the method for regime identification and analysis. (**b**) Illustration of the method with the Lorenz model, Eq. (8), showing the metastable states within the (x, z) phase space. The cloud of points (black dots) represent the "observations", whereas the filled circles and triangles represent the centers of the HMM hidden states (vectors a_J in Eq. 1) for regimes 1 and 2 respectively. The size of these symbols is proportional to the stability measure $|w_{1i}|$, see Sec. 2.2. The green line represents a fragment of the trajectory on the attractor of the deterministic Lorenz system corresponding to $\sigma = 0$.

$$\langle Y|A_k\rangle = \frac{\sum_t Y_t P(A_k|Y_t)}{\sum_t P(A_k|Y_t)}. \tag{13}$$

Note that the community A_k, Eq. (12), can be used wholly or replaced by a subset corresponding, for example, to the most stable states (with the largest $|w_{1i}|$). The occupation probability, Eq. (12), can also be used to derive other dynamical characteristics linked, for example, to long-term behavior such as mean number of days D_k within A_k, in a given winter T:

$$D_k(T) = \sum_{t \in T} P(A_k|Y_t). \tag{14}$$

Kernel PCA

To take into consideration the high-dimensional complex nonlinear structure of the atmospheric dynamical system, with possible nonlinear manifold of the dynamics an enable an efficient application of HMM, we apply nonlinear dimension reduction method based on kernel principal components (KPCs). Linear methods, such as conventional EOFs, are unable to identify a nonlinear low-dimensional embedding of the system (e.g.[15]).

Kernel principal component analysis (KPCA) has been shown[14] to efficiently separate families of recurrent patterns of the mid-latitude atmospheric circulation and identify low-dimensional nonlinear manifold[15]. KPCA belongs to the class of spectral methods and is based on embedding the data into a higher-, e.g. infinite-dimensional space where the system most likely becomes linear, and nonlinear manifolds and nonlinear features become linearly identifiable[15]. It allows implicit nonlinear mapping from the initial space by means of redefining the dot product, that tends to bring similar states closer together. Such a transformation makes it possible to embed the classes of similar states into low-dimensional subspace of a few leading scalar components. The measure of similarity of two states can be defined in the form of a Gaussian kernel

$$K(Y_i, Y_j) = \exp\left[-\frac{d^2(Y_i, Y_j)}{2\sigma^2}\right]. \tag{15}$$

Here $d()$ is a distance between the states, which should be properly selected depending on the problem. Focusing on similarity of the pattern's structures/shapes rather than their amplitudes, we use Euclidean distance between the normalized vectors:

$$d(x_i, x_j) = \left\|\frac{x_i}{\|x_i\|} - \frac{x_j}{\|x_j\|}\right\|, \tag{16}$$

where the norm of x, $\|x\| = (x^T \Lambda x)^{\frac{1}{2}}$, is based on a weight (diagonal) matrix Λ to reflect the uneven density of grid nodes especially near the poles. Typically, the weight matrix depends on the latitude θ_i of the $i'th$ grid cell. For example, if the focus is on the global grid then $\Lambda_{ii} = \cos \theta_i$, but if the focus is on midlatitude large scale dynamics then one uses (see the work[14] for details):

$$\Lambda_{ii} = \cos \theta_i \sin^2 \theta_i. \tag{17}$$

An eigenanalysis of the kernel matrix $K_{ij} = K(Y_i, Y_j)$ is then conducted after some kind of centering[15]

$$\mathbf{K_c} = \mathbf{C} \cdot \mathbf{K} \cdot \mathbf{C}, \tag{18}$$

with $C = I - \frac{1}{N}1$, I and 1 are respectively the identity matrix and the matrix of ones. The result of this centering operation can be interpreted as a dot product matrix calculated from a sample with zero mean. The eigenelements of K_c are then derived, yielding our kernel PCs used in the HMM.

We point out here that the hyperparameter σ in Eq. (15) entails the degree of nonlinearity of the mapping. For instance, the smaller σ is the better the local proximity preservation gets, that is nearby states in the input space map onto nearby states in the feature space. This is an attractive feature which allows easy cluster separation. Furthermore, when σ is large it can be seen via Taylor expansion that for centered time series Y we have $\mathbf{K_c}_{ij} = \frac{1}{\|Y_i\|\|Y_j\|}Y_i^T \Lambda Y_j + o(\sigma^{-2})$, yielding similar results to conventional linear principal component analysis.

The entire method described in this section is schematically represented in Fig. 1a. On the dimension reduction step the observed high-dimensional data vectors are mapped onto several KPCs, making thus a low-dimensional state space, which is used to separate classes of similar states. The HMM is then constructed and trained in this space, and communities of hidden states associated with metastable circulation regimes identified. Finally, dynamical and statistical characteristics of the circulation regimes are investigated along with association to large-scale teleconnections.

Results

Circulation regime identification

We have, first, applied the method to simulation from a simplified system of atmospheric dynamics, namely, a 3-layer quasi-geostrophic (QG3) model[3] and check stability of the method. The results of this experiment are presented in the Supplementary Material (see Figs. S1-S2 and text). In this section we discuss the application to daily geopotential heights at 500-hPa pressure level (hereinafter Z500) in winter seasons over the mid-latitude of the Northern Hemisphere. The Z500 fields are taken from the National Center of Environmental Prediction/ National Center of Atmospheric Research (NCEP/NCAR) reanalysis dataset[22]. This is a gridded time series with $2.5^o \times 2.5^o$ latitude-longitude resolution, that covers the period 1950–2023. We focus here on the Northern Hemisphere mid-latitudes by considering latitudes north of 30N. The annual cycle is removed by subtracting the daily seasonal means over the whole period smoothed in time with a Gaussian window with the standard deviation of 15 days. This smoothing suppresses the day-to-day noise in the resulting annual cycle while retaining the intra-annual seasonal structure. And only winter (December–January–February) segments are taken, making a sample size of 6588 days, between winter 1950-51 and winter 2022-23.

Relying on previous findings[14], for constructing of an HMM, we use a space of three leading KPCs, that provides good separating groups (high modularity) of similar states. Note that a more formal optimization algorithm for the number of KPCs has yet to be developed. The HMM that has been derived from the described dataset has optimal number of hidden states equal to 50 (see Fig. S1 in the Supplementary Material). We decomposed the obtained 50×50 transition matrix into metastable blocks for different powers L determining the time steps of the analyzed evolution operator (see Sec. 2.3). Only significant splits of communities at the 1% level were accepted during the splitting procedure. An example of how the states in the used KPC projection are arranged to the detected communities is presented in Fig. 2b, for the whole observed sample as well as for points attributed to the most stable hidden states within each circulation regime. It is seen that the communities are clearly separated in the KPC space, especially the most stable ones that form the cores of the regimes.

Structure of regimes and weather impact

Figure 2 illustrates the splitting process of the communities for $L = 7$. Other values of L have been investigated and are presented in figures S3-S9 of the Supplementary Material. These figures show the composite patterns of the obtained communities calculated using Eq. 13. For spatio-temporal resolution $L \leq 9$ days, the method yields consistently 4 highly significant communities. We notice, however, that the structure of the obtained circulation regimes with $L \leq 5$ and $5 < L \leq 9$ are somehow not quite similar. This dissimilarity can be explained by the existence of many short-lived recurrent fluctuations that corrupt the results at low L which cannot be correctly resolved due to lack of statistics. This is confirmed by examining an experiment by dropping the significance test and the threshold of the modularity increment ΔM. We found (see, for example, supplementary figure S4) that, in case of low values of L we get many more communities which vary with L, but for $L > 5$ such spurious communities are filtered out and we obtain a few communities.

Increasing the power L will eventually lead to degeneracy of the transition matrix. For instance, with $L > 9$ the degeneracy first starts by merging regimes 3 and 4 yielding a broader community. With $L = 11$ only two communities are obtained, which remain for larger values, corresponding to an indivisible matrix. As a result, this consistently suggests that reliable and statistically significant persistent communities are detected for $5 < L \leq 9$, and, since the set of patterns for this range of values are almost identical, we hereinafter discuss the results corresponding to $L = 7$.

The first split of the initial community of all hidden states provides about 89% of the relative community increment, and leading eventually to four communities (see Fig. 2). The communities resulting from the first split represent opposing Z500 anomalies over the Arctic, Greenland, with anomalies in Eastern Europe and Far Eastern Siberia. Roughly speaking, these two communities reflect a Rossby wave structure with zonal wavenumber 4, and project respectively on states of intensified and weakened westerly flow. The following (and final) split yields smaller, but still significant, contribution to the modularity, providing identification of qualitatively different families of circulation patterns within the basic communities, hereinafter referred to as regimes 1–4.

These communities, regarded as metastable states in the low-frequency range, have longer time scales than typical synoptic systems and can therefore affect surface conditions over periods of time longer than synoptic time scales. To investigate the effect on surface weather we consider NCEP/NCAR daily surface air temperature (SAT) anomalies at 0.995 sigma level. The anomalies are computed in the same way as with the Z500 anomalies (see sec. 3.1). The regime occupation probabilities obtained for Z500 anomalies, Eq. 12, are used as weights to get the SAT composites patterns associated with the four circulation regimes (e.g., Eq. 13). The obtained SAT composites are shown in Fig. 3.

Pattern 1 (Fig. 2), projects strongly on the positive phase of the Arctic Oscillation (AO) characterized by lower than normal pressure over the Arctic ocean. The negative anomalies over the Arctic extends occasionally to reach the North Pacific and eastern Europe resulting in a stronger, than normal, jet stream there bringing weather systems from the west, e.g. loaded with rainfall. The second circulation regime Regime 2 combines features of the positive phase of the North Atlantic Oscillation (NAO), manifested as a deepened northern-subtropical Atlantic dipole in atmospheric pressure, and the negative Pacific North America Pattern (PNA) with an area of high pressure over the northeastern Pacific.

A negative meridional pressure gradient in the North Atlantic, for both regimes, strengthens the zonal westerly flow in the Euro-Atlantic region, which may bring warm and wet conditions from Europe to Siberia (see SAT composites in Fig. 3). Over North America, the two patterns have opposite effects. Regime 1 causes a strengthening of the zonal flow bringing warm and wet oceanic air to this region, especially over North east
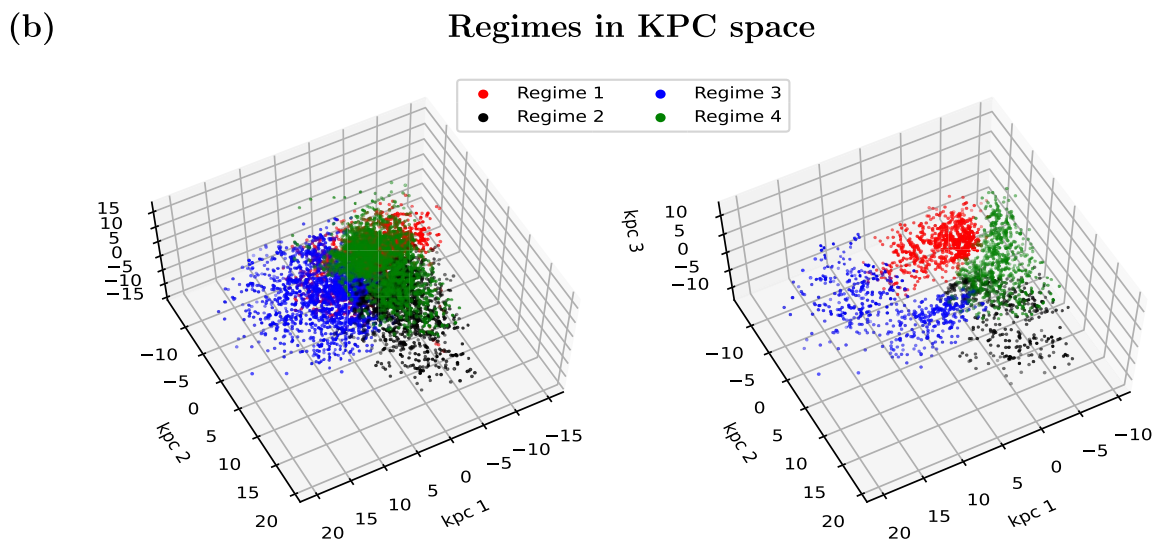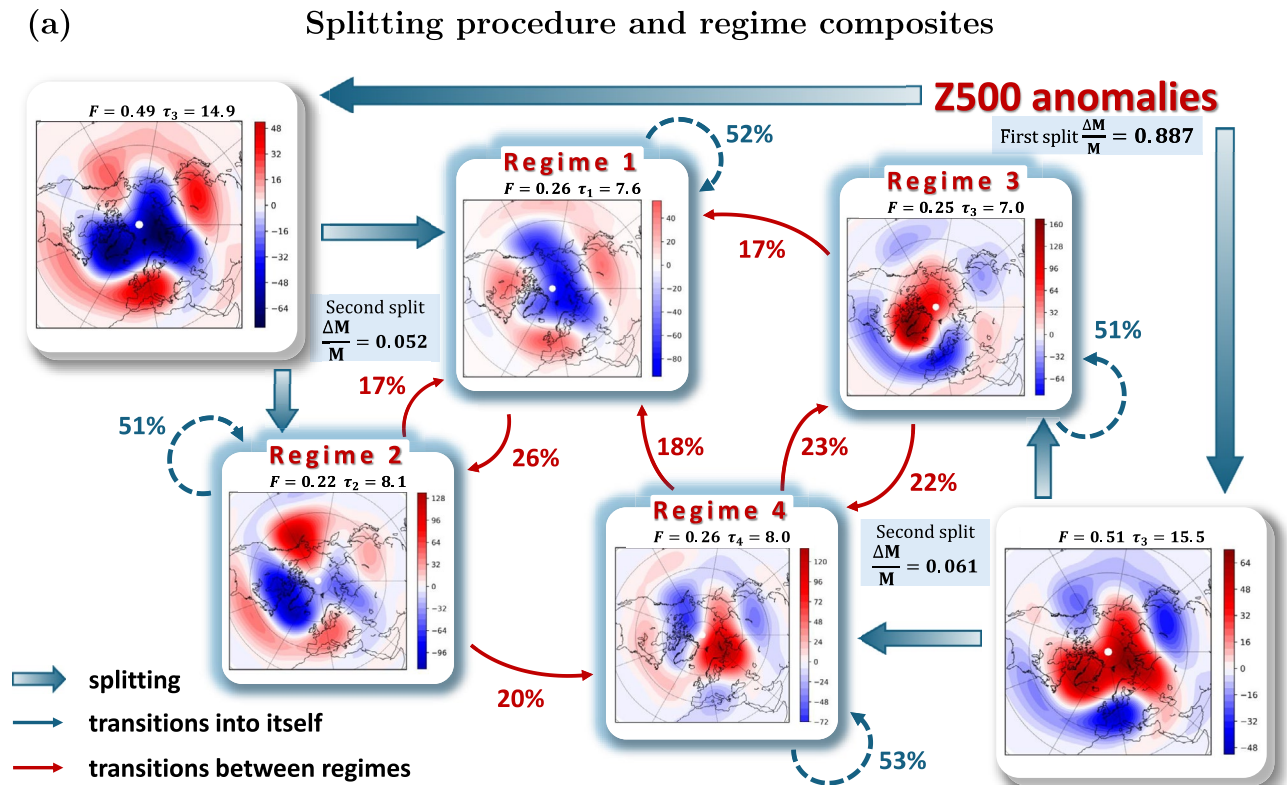
Fig. 2. Circulation regime representation in data and KPC spaces. **(a)** Composite patterns of communities obtained with $L = 7$ days. Each pattern is calculated as a soft average of observed Z500 anomalies over a community of hidden states according to Eq. 13. Parent and child patterns of emerging communities during the iterative splitting process (see text) are shown by thick arrows. For each split, the relative increment $\frac{\Delta M}{M}$ of modularity is indicated, where M is the final modularity after process completion. The four resulting communities are highlighted in the center of the diagram. Thin arrows indicate probabilities of transitions between regimes (red) as well as self transitions (blue). Additionally, for each community, the fraction of residence F (days) as well as its mean lifetime τ (see Eq. 10) are indicated. **(b)** States of the system within the three leading KPCs color-coded according to their regime assignment showing points with assignment probability (see Eq. 12) greater than 0.9 (left panel) and points belonging to 20% of the most stable hidden states, see Sec. 2.2, (right panel).

US/east Canada and the Canadian Archipelago. However, the blocking high pressure anomalies in the northern Pacific Ocean in regime 2 contributes to pumping arctic cold air to the region.

The remaining two circulation patterns 3 and 4 (Fig. 2) emerge from the splitting of the community characterized by high pressure over the Arctic, i.e. negative phase of AO. Pattern 3 clearly projects on the
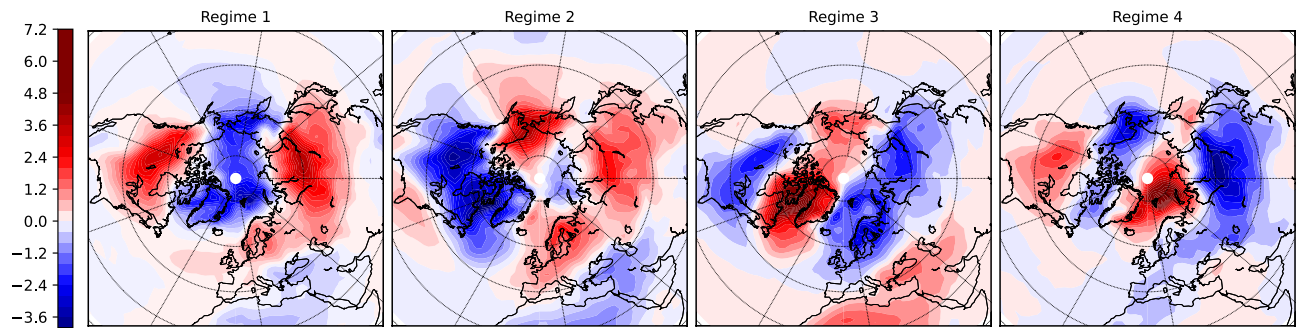
Fig. 3. Composite of surface air temperature anomalies based on the resulting Z500 circulation regimes.

negative NAO phase, which prevents moist and warm Atlantic air masses from reaching Europe and contributes to cooling northern Eurasia. At the same time, the northwest Atlantic area, including Greenland and Labrador sea, are getting warmer than normal, in addition to warmer and wetter, than normal, Mediterranean. In the circulation pattern 4, anticyclonic anomalies are shifted eastward, centered over the Barents sea. This pattern contributes to cooling northern Siberia, but also warms the Eurasian basin of the Arctic Ocean.

Transition probabilities, lifetimes and pattern diversity

Another, yet important characteristic feature of the identified patterns is their lifetime and transitions. The transition probabilities between the regimes, i.e. the reduced matrix calculated via Eq. 12 with $L = 7$ is

$$\overline{Q^L} \approx \begin{pmatrix} 0.52 & 0.17 & 0.17 & 0.18 \\ 0.26 & 0.51 & 0.1 & 0.06 \\ 0.13 & 0.12 & 0.51 & 0.23 \\ 0.09 & 0.2 & 0.22 & 0.53 \end{pmatrix} \tag{19}$$

Several points can be inferred from the above transition matrix. First and foremost we can derive the probability distribution of the likelihood of the transition to the next circulation regime. These distributions are clearly different from the uniform PDF, as it is seen from Fig. 2 (thin arrows). Circulation regimes 3 and 4, for example, are more likely to interact with each other than with the remaining regimes. Both these circulation patterns concern similar blocking structures in the Euro-Atlantic region, and during the coldest Eurasian winters they often alternate in time (see the Supplementary Animation S1 Video). The next likely transition, from these two patterns, is toward circulation pattern 1, which, in turn, prefers to resolve into regime 2. However, the preferred path from the latter is regime 4, although the probability of switching to regime 1 is also high. These transitions lead to a preferred loop of circulation regime dynamics, namely, $r1 \rightarrow (r2 \leftrightarrow r1) \rightarrow (r4 \leftrightarrow r3) \rightarrow r1$.

The probabilities of self transitions, given by the diagonal elements of the transition matrix, Q^L, Eq. 19, can be used to estimate a posteriori the distribution of regimes lifetime via Eq. 11. At the same time we can use the HMM to generate an ensemble of surrogate time series mimicking the observations. These surrogates can be used to derive various statistics, such as regime lifetime distribution, and can be used in hypothesis testing. Figure 4 shows the obtained lifetime distribution compared to that from the observations for each of the circulation patterns. Overall the figure shows comparable values of the expected lifetimes of all circulation regimes, ranging from 7 to 8 days, though an anomalous maximum, of 25-30 days, can be noted in the observational distribution of circulation pattern 3. In the meantime, Fig. 4 also shows that such rare events fit well into the model statistics and cannot be rejected based on the quasi-geometric model distribution. We caution, however, that the available statistics are not sufficient to cover the tail of the observational distribution well because of the sample size, and there is no decisive evidence to suggest that regime 3 tends to live systematically longer than other regimes.

In addition, the time series of the regime probabilities $P(A_k|Y_t)$ (see Eq. 12) permits to estimate non-Markovian predictability of a regime A_i by calculating the measure of convergence to stationary distribution starting from this regime at $t = 0$. This is the Kullback-Leibler (KL) divergence between probability distribution of regimes at time t given the regime A_i at $t = 0$, and stationary distribution π (see[23]). The result obtained for our four regimes shown in logarithmic scale reveals a sign of slower than Markov relaxation to the stationary distribution for Regimes 2 and 3 (see Fig. S10 in the Supplementary Material), indicating that these regimes, dominated by NAO, could be predicted on longer times than with one-step HMM.

To investigate the similarity of the atmospheric circulation within the patterns we show in Fig. 4 the relative variance of observations given a specific regime k:

$$R_k = \frac{\Omega_k - \overline{\Omega}}{\overline{\Omega}}, \tag{20}$$

where $\Omega_k = \left\langle Y^2|A_k \right\rangle - \left\langle Y|A_k \right\rangle^2$ is the variance of the $k'th$ regime (the operator $\langle\rangle$ is defined as in Eq. 13), and $\overline{\Omega}$ is the total sample variance of the observations. The quantity R_k provides a measure of intra-regime variance. Overall, we see low intra-regime variance in areas of high values of anomaly composites for regimes 1, 2, and 4 (shown in Fig. 2). But this is not the case for regime 3, which shows small variance in the Euro-Atlantic region,
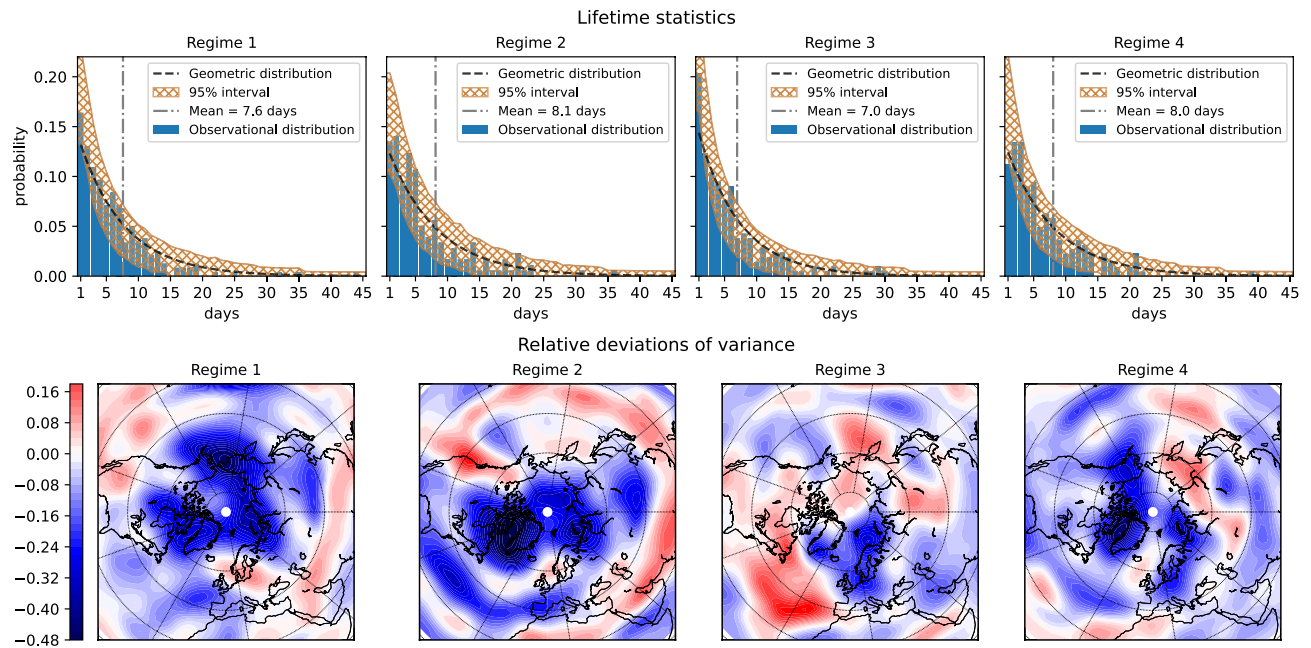
Fig. 4. Properties of the regimes. **Top panel**: distributions of regime lifetimes. Blue bars represent observational distribution, i.e. interval statistics of sequences of observations assigned to the regime with probability more than 0.95. Orange shade shows 95% confidence interval of an ensemble of 10,000 realizations generated by the HMM. Geometric distribution Eq. 11 with a parameter q equal to the probability of a regime to transit into itself for a one-day interval (i.e. the diagonal element of the one-day reduced transition matrix), as well as its mean, are indicated by dashed and dash-dotted line, respectively. **Bottom panel**: relative Z500 variance anomalies within a regime with respect to the sample variances (see Eq. 20).

while in the subtropics the variance significantly exceeds the total sample variance. This circulation regime projects onto −NAO for which the blocking high around Greenland is known by its high persistence[24]. But the subtropical low pressure system is not as persistent explaining the larger variance there. It is also possible that the weakening of the pressure gradient between the polar region and the midlatitudes yields unstable disturbances that can travel equatorward reaching the subtropics and modify the flow there (see Supplementary Animation S1 Video for the winter 2009-2010 with dominating regime 3). A similar situation occurs for regime 2 over the North Pacific, where the blocking high leads to increased variances further south.

Long-term dynamics of circulation regimes and teleconnection

One of the question of importance here is related to the characteristics and variability across scales of these patterns. Figure 5 shows the mean number of days per winter, i.e. regime frequencies, as in Eq. 14. The figure exhibits pronounced interannual (2–4 year) and decadal variability. Note that we attempted to assemble the time series into pairs based on the blocking action of the zonal flow in the Euro-Atlantic region resulting in either warm or cold winters over northern Eurasia (see Fig, 3). Although cross-regime correlations appear quite low on interannual timescales, we observe clear decadal correlations between the "warm" regimes 1 and 2 as well as the "cold" regimes 3 and 4. In particular, there is an approximate 10-year period of warmer than normal winters extending from the mid-80s to the mid-90s, a consequence of the abnormally high difference between the frequencies of "warm" and "cold" regimes. Opposite conditions arose in the 1960 s and early 2010s. Examples of atmospheric behavior in two winters with opposite anomalies in the Euro-Atlantic region are presented in the y Animations S1 Video and S2 Video, where it is shown that 2009-2010 and 1991-1992 winters are dominated by regimes 3 and 1, respectively.

Detailed analysis of whether this dynamical behavior can be attributed to decadal climate modes such as, e.g., Pacific decadal oscillation (PDO), or whether it is a manifestation of internal atmospheric variability known to be a characteristic feature of mid-latitudes, goes beyond the present paper, and is subject of further research. Here we only report significant correlations of regime frequency with El-Niño Southern Oscillation (ENSO) represented by the Niño 3.4 index[25] and PDO index[26]. We also leave beyond the scope of this article a detailed numerical analysis of directional (causal) relationships between processes, similar to the one that was performed, for example, in the recent work[27] for climate indices.

ENSO is one of the dominant modes of climate variability bringing about a complex network of teleconnections worldwide[28,29]. The Niño 3.4 index defined as mean sea surface temperature in the central tropical pacific region (5S-5N and 170-120W) is widely used for ENSO monitoring and forecast[30]. Correlation analyses between the Niño 3.4 index and regime frequency produce small but significant correlations with regimes 1,2 and 4 (see Fig. 5 and 6). Positive correlation with regime 1 is likely due to the influence of ENSO on the positive PNA-like anomaly[31] in the North Pacific (see Fig. 2). Interestingly, there is a lagged influence of regime 2 on ENSO, but
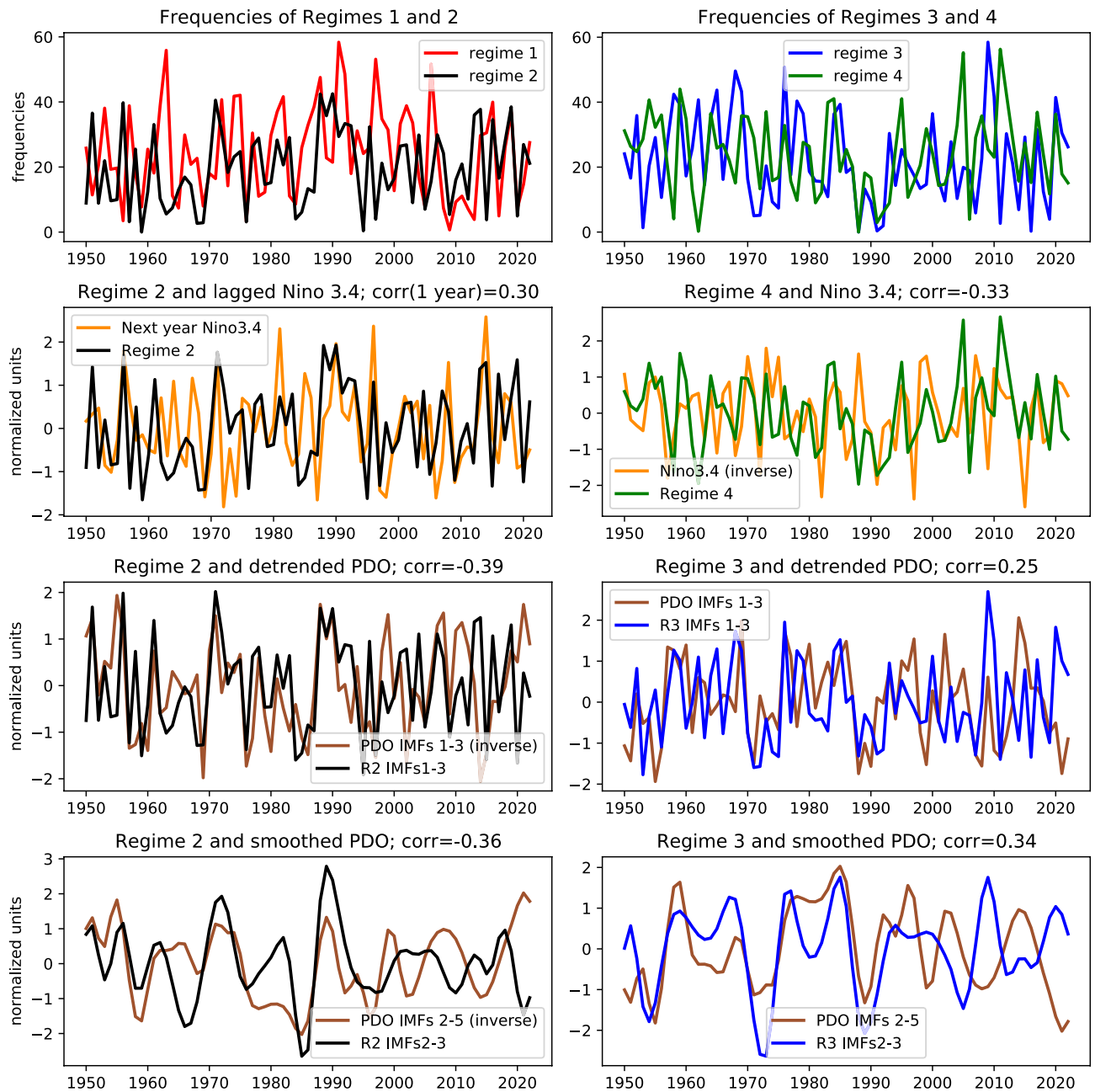
Fig. 5. Interannual fluctuation of regime frequency and link to large-scale teleconnection. **Top row**: Frequency of the number of days in each regime during winter, see Eq. 14. **Second row**: Regimes 2 and 4 frequency and ENSO showing the Niño 3.4 index with 1-yr lag (left) and (opposite) Niño 3.4 (right). **Third row**: Regimes 2 and 3 frequency and detrended PDO index filtered using the first three IMFs. **Fourth row**: Regimes 2 and 3 frequency and smoothed PDO index by removing the first (short time-scale) IMF. When plotted together with regime 2 frequency, the PDO index is multiplied by -1. Correlation coefficients of the regime frequency with the (non-inverted) indices are also indicated. The year in the x-axis refers to that of January.

with no significant contemporaneous correlation. This indicates that negative PNA, which projects onto regime 2, impacts North Pacific Oscillation mode in the following months making it favorable for developing El-Niño in the following season, see, e.g.[32,33]. Note that directed nonlinear influence of PNA and AO on ENSO was detected in[27]. Finally, there is a negative correlation between winter Niño 3.4 and circulation regime 4. Because this regime is dominated by high pressure anomaly over north/eastern Europe, its association with the culminating ENSO phase suggests an influence of the latter on weather in the Euro-Atlantic region.

The PDO, a dipole of SST pattern in the North Pacific, is driven by internal feedbacks from the complex interaction of coupled ocean-atmosphere system and by tropical/extratropical forcing. This results in a wide spectrum of PDO including interannual, decadal, and longer timescales. In an attempt to scrutinize the relation
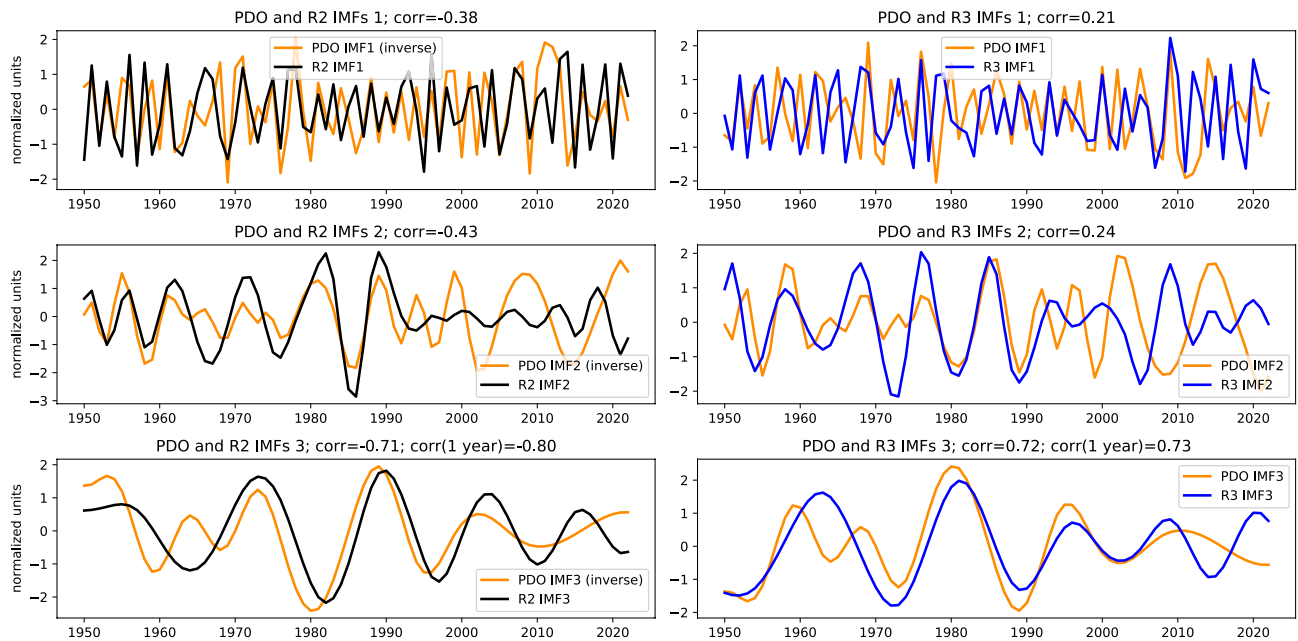
Fig. 6. Leading three IMFs of PDO winter mean index (orange), regimes 2 (black) and regime 3 (blue) frequencies. The PDO IMFs time series in the middle and bottom left panels are multiplied by −1 (inverted) for convenience. Correlation coefficients of regime IMFs with corresponding PDO IMFs (non-inverted) are also indicated. In the bottom panels correlation of PDO IMF3 with 1-year lagged IMFs 3 of regimes 2 and 3 are also shown. The year in the x-axis refers to that of January.

to teleconnections we filter out high-frequency variability and compare the obtained time series. We use Empirical Mode Decomposition (EMD)[34] – a powerful data-driven approach for extracting dominant intrinsic modes/components ranked by their characteristic timescales, referred to as Intrinsic Mode Functions (IMFs). These IMFs provide a consistent smoothing tool by discarding for example the fastest modes or by removing long-term trends. We identified highly significant correlations of the PDO index with regimes 2 and 3 frequency both in the short time-scale band, dominated by 1-10 year periods, and in the intermediate, 10-20 year, band (see Fig. 5). The highest correlation, around −0.4, is reached with the short time-scale component of the PDO index and regime 2 frequency time series, both represented by IMFs 1–3. This reflects a significant impact of PDO-induced SST anomalies in the eastern North Pacific on the mass field and geopotential height anomalies where regime 2 is pronounced. A significant positive, but lower, correlation is obtained with regime 3 frequency. For the intermediate time scale, we consider a smoothed PDO index by discarding the high-frequency IMF-1 and smooth the regime frequency time series using IMFs 2–3. Figure 5 shows highly significant negative and positive correlation with regimes 2 and 3, respectively. Since regime 3 is largely dominated by the NAO pattern, the obtained correlation suggests interactions between North Pacific and North Atlantic climate variability on decadal time-scales. To highlight the effect of these timescales, Figure 6 shows the correlations between PDO index and regimes 2 and 3 frequency using IMFs 1–3. Unexpectedly high correlations are obtained with IMF-3 of PDO and 1-year lag regimes frequency of the order −0.8 for regime 2 and 0.73 for regime 3. The substantial PDO influence on the low frequency variability of the NAO was also discussed in[27]. Correlations based on IMFs 1 and 2 are also highly significant for regime 2 with values −0.38 and −0.43 respectively. It is worth noting that the time period from the 70's to the 90's, identified as a positive PDO epoch[29,35], has the largest contribution to the obtained correlations.

Discussion

We present a novel methodology of data-driven detection of recurring atmospheric circulation patterns by considering these patterns as metastable communities of the nonlinear system dynamics. The nonlinear formulation of the problem effectively narrows the class of recurrent states of the atmospheric flow into circulation regimes. This is based on building a dynamic stochastic model of the system evolution operator, allowing one to determine areas in the state space of trajectory slowing down. Data-driven nonlinear modeling is known to be a challenging inverse problem, and nonlinear optimization is needed, see e.g.[36,37]. Here we adopt a simplified representation of the system via a finite-state Markov model whose complexity, or resolution, depends on the extent of observational statistics. The approach recasts the nonlinear dynamics of the original system into a linear operator on a set of probability distributions, and where metastability can be captured via stochastic matrix analytic methods.

The method uses HMM combined with nonlinear dimension reduction and community splitting based on graph theory. One of the main advantage of the method is that no prior assumptions about number of regimes are needed, as this number is recognized by the iterative division algorithm based on clear stopping condition.

The number of microstates K is also chosen to be optimal relative to the data and is not predefined. Given the available sample size, a substantial part of our analysis is devoted to investigating the reliability of the results. First, each community splitting must pass a fairly stringent significance test before acceptance. Next, the degree L of the transition matrix, which determines the model resolution, is adjusted, and the range corresponding to stable results is selected. This contributed to results reliability and helped yield roughly coherent metastable structures of the system dynamics. The method has been successfully applied to a low-order chaotic model and a 3-level quasi-geostrophic model, which is a simplified low-complexity atmospheric model of midlatitude dynamics. The method is then applied to the NCEP/NCAR Z500 over the NH during winters 1950-2023. Four (nonlinear) circulations patterns are identified. These patterns partially project onto NAO and PNA. Lifetime and likely transitions between these metastable patterns are then identified.

One important question that arises often in discussions is related to the connection between these patterns and "standard" modes of atmospheric teleconnections emerging in atmospheric LFV. The typical patterns, such as NAO, PNA, and AO, etc., have historically been isolated empirically by studying regional dynamical processes and by their influence on weather systems, and then confirmed via linear correlation analysis[38]. However, from a global perspective, all these modes are manifestation of a more general evolution of the planetary-scale atmospheric circulation involving many nonlinear and complex processes, including lower boundary conditions, nonlinear wave-wave interaction, and complex forcing patterns. The dynamical feature of atmospheric circulation manifests itself partly in those linear modes, but the full picture needs a deeper investigation into the nonlinear structure of the system trajectory within its state space. The metastable regimes constitute the backbone elements of the dynamics of this global system and are of paramount importance for understanding the dynamics of all regionally significant large-scale processes, and helps toward extending the chaotic limits of predictability. In particular, in Sec. 3 we show that each of the obtained regimes contributes to those well-known teleconnections, and can provide suggestions for predictability of various atmospheric flows. The presented method allows to extract areas of metastability, but also gives a consistent probabilistic assessment of their evolution via the transition matrix.

The method presented here can provide potential for subseasonal-to-seasonal (S2S) prediction. One pivotal direction of the method extension consists in adopting it for long-term regime prediction, including inter-seasonal forecasting. This requires a model that evolves continuously across the year involving the interaction between the circulation regimes and the annual cycle. Moreover, the model should be able to take into account the response of the system to forcings of different nature like CO_2 trend or long-period climate modes like ENSO and PDO. In its current version the method deals with the autonomous HMM regarding the hidden states and Markov operator. As a result, the metastable communities of states associated with regimes are defined for the whole dataset despite of its long-scale components (which certainly remain even after removing the additive trend/annual component). However, the method allows us to study such components by considering the time series of occupation probabilities of the observations in different hidden states. Such probability vectors are the new (transformed) dynamical variables, which store the information about the non-stationarity of the process. For example, using these variables, we can construct and study the time series of regime frequencies, which contain long-period components associated with decadal processes, as is shown in Sec. 3.4 and Figs. 5-6. Further enhancement of the approach could incorporate an external long-period signals directly in the HMM (see e.g.[39]). To expand the current stationary HMM, our future model extension will include forcing by nonstationary teleconnection and includes the annual variability.

Data availability

The analyzed NCEP/NCAR data of 500-hPa geopotential heights and surface atmosphere temperatures are open to the public. Niño 3.4 and PDO time series are provided by the NOAA Physical Sciences Laboratory, Boulder, Colorado, USA, from their website at https://psl.noaa.gov/. The package of Python programs developed by the authors that implements the described technique for identifying regimes is available here.

References

1. Hannachi, A., Straus, D. M., Franzke, C. L. E., Corti, S. & Woollings, T. Low-frequency nonlinearity and regime behavior in the northern hemisphere extratropical atmosphere. *Rev. Geophys.* **55**, 199–234. https://doi.org/10.1002/2015RG000509 (2017).
2. Ghil, M., Groth, A., Kondrashov, D. & Robertson, A. W. Chapter 6 - extratropical sub-seasonal to seasonal oscillations and multiple regimes: The dynamical systems view. In Robertson, A. W. & Vitart, F. (eds.) *Sub-Seasonal to Seasonal Prediction*, 119–142, https://doi.org/10.1016/B978-0-12-811714-9.00006-1 (Elsevier, 2019).
3. Marshall, J. & Molteni, F. Toward a dynamical understanding of planetary-scale flow regimes. *J. Atmospheric Sci.* **50**, 1792–1818. https://journals.ametsoc.org/view/journals/atsc/50/12/1520-0469_1993_050_1792_taduop_2_0_co_2.xml (1993).
4. Vannitsem, S. & Nicolis, C. Lyapunov vectors and error growth patterns in a t21l3 quasigeostrophic model. *J. Atmospheric Sci.* **54**, 347–361. https://journals.ametsoc.org/view/journals/atsc/54/2/1520-0469_1997_054_0347_lvaegp_2.0.co_2.xml (1997).
5. Gritsun, A. & Lucarini, V. Fluctuations, response, and resonances in a simple atmospheric model. *Phys. D: Nonlinear Phenom.* **349**, 62–76. https://doi.org/10.1016/j.physd.2017.02.015 (2017).
6. Hannachi, A. *Patterns Identification and Data Mining in Weather and Climate* (Springer, 2021).
7. Murphy, K. P. *Machine Learning: A Probabilistic Perspective* (MIT Press, 2012).
8. Franzke, C., Crommelin, D., Fischer, A. & Majda, A. J. A hidden markov model perspective on regimes and metastability in atmospheric flows. *J. Clim.* **21**, 1740–1757. https://doi.org/10.1175/2007JCLI1751.1 (2008).
9. Rabiner, L. R. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE* **17**, 257–286 (1989).
10. Brunton, S. L. & Kutz, J. N. *Data-Driven Science and Engeneering* (Machine LearningDynamical Systems and Control (CUP, 2022).

11. Springer, S., Laio, A., Galfi, V. M. & Lucarini, V. Unsupervised detection of large-scale weather patterns in the northern hemisphere via markov state modelling: from blockings to teleconnections. *npj Clim. Atmospheric Sci.* **7**, https://doi.org/10.1038/s41612-024-00659-5 (2024).

12. Newman, M. E. J. Modularity and community structure in networks. *Proc. Natl. Acad. Sci.* **103**, 8577–8582, https://doi.org/10.1073/pnas.0601602103 (2006). https://www.pnas.org/content/103/23/8577.full.pdf.

13. Schölkopf, B., Smola, A. & Müller, K.-R. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Comput.* **10**, 1299–1319, https://doi.org/10.1162/089976698300017467 (1998). https://direct.mit.edu/neco/article-pdf/10/5/1299/813905/089976698300017467.pdf.

14. Mukhin, D., Hannachi, A., Braun, T. & Marwan, N. Revealing recurrent regimes of mid-latitude atmospheric variability using novel machine learning method. *Chaos: An Interdiscip. J. Nonlinear Sci.* **32**, 113105, https://doi.org/10.1063/5.0109889 (2022). https://pubs.aip.org/aip/cha/article-pdf/doi/10.1063/5.0109889/18108937/113105_1_5.0109889.pdf.

15. Hannachi, A. & Iqbal, W. Bimodality of hemispheric winter atmospheric variability via average flow tendencies and kernel eofs. *Tellus A: Dyn. Meteorol. Oceanogr.* **71**, 1633847. https://doi.org/10.1080/16000870.2019.1633847 (2019).

16. Baum, L. E., Petrie, T., Soules, G. & Weiss, N. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals Math. Stat.* **41**, 164–171. https://doi.org/10.1214/aoms/1177697196 (1970).

17. Chung, F. R. K. *Spectral Graph Theory* (CBMS Regional Conference Series in Mathematics (Am. Math. Soc.), Providence, RI, 1997).

18. Lafon, S. & Lee, A. B. Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization. *EEE Transactions on Pattern Analysis and Mach. Intell.* **28**, 1393–1403 (2006).

19. Nadler, B., Lafon, S., Coifman, R. R. & Kevrekidis, I. G. Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. *Appl. Comput. Harmon. Analysis* **21**, 113–127 (2006).

20. Belkin, M. & Niyogi, P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.* **15**, 1373–1396 (2003).

21. Lorenz, E. N. Deterministic nonperiodic flow. *J. Atmospheric Sci.* **20**, 130–141. https://journals.ametsoc.org/view/journals/atsc/20/2/1520-0469_1963_020_0130_dnf_2_0_co_2.xml (1963).

22. Kalnay, E. et al. The NCEP/NCAR 40-year reanalysis project. *Bull. Am. Meteorol. Soc.* **77**, 437–471. https://journals.ametsoc.org/view/journals/bams/77/3/1520-0477_1996_077_0437_tnyrp_2_0_co_2.xml (1996).

23. Vannitsem, S. Toward a phase-space cartography of the short- and medium-range predictability of weather regimes. *Tellus A: Dyn. Meteorol. Oceanogr.* https://doi.org/10.3402/tellusa.v53i1.12180 (2001).

24. Woollings, T., Hannachi, A. & Hoskins, B. Variability of the north atlantic eddy-driven jet. *Q. J. Roy. Meteorol. Soc.* **136**, 856–868 (2010).

25. Rayner, N. A. *et al.* Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res. Atmospheres* **108**, https://doi.org/10.1029/2002JD002670 (2003). https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2002JD002670.

26. Newman, M. et al. The pacific decadal oscillation, revisited. *J. Clim.* **29**, 4399–4427. https://doi.org/10.1175/JCLI-D-15-0508.1 (2016).

27. Vannitsem, S., Liang, X. S. & Pires, C. A. Nonlinear causal dependencies as a signature of the complexity of the climate dynamics. *Earth Syst. Dyn.* **16**, 703–719. https://doi.org/10.5194/esd-16-703-2025 (2025).

28. Alexander, M. A. et al. The atmospheric bridge: The influence of ENSO teleconnections on air-sea interaction over the global oceans. *J. Clim.* **15**, 2205–2231. https://journals.ametsoc.org/view/journals/clim/15/16/1520-0442_2002_015_2205_tabtio_2.0.co_2.xml (2002).

29. Trenberth, K. E. El Niño southern oscillation (ENSO). *Encycl. Ocean. Sci.* 420–43 https://doi.org/10.1016/B978-0-12-409548-9.04082-3 (2019).

30. Barnston, A. G., Tippett, M. K., L'Heureux, M. L., Li, S. & Dewitt, D. G. Skill of real-time seasonal ENSO model predictions during 2002–11: Is our capability increasing?. *Bull. Am. Meteorol. Soc.* **93**, 631–651. https://doi.org/10.1175/BAMS-D-11-00111.1 (2012).

31. Mo, K. C. & Livezey, R. E. Tropical-extratropical geopotential height teleconnections during the northern hemisphere winter. *Mon. Weather. Rev.* **114**, 2488–2515. https://journals.ametsoc.org/view/journals/mwre/114/12/1520-0493_1986_114_2488_teghtd_2_0_co_2.xml (1986).

32. Vimont, D. J., Wallace, J. M. & Battisti, D. S. The Seasonal Footprinting Mechanism in the Pacific: Implications for ENSO*. *J. Clim.* **16**, 2668–2675, https://journals.ametsoc.org/view/journals/clim/16/16/1520-0442_2003_016_2668_tsfmit_2.0.co_2.xml (2003). https://journals.ametsoc.org/jcli/article-pdf/16/16/2668/3780359/1520-0442(2003)016_2668_tsfmit_2_0_co_2.pdf.

33. Mukhin, D., Gavrilov, A., Seleznev, A. & Buyanova, M. An Atmospheric Signal Lowering the Spring Predictability Barrier in Statistical ENSO Forecasts. *Geophys. Res. Lett.* **48**, e2020GL091287, https://doi.org/10.1029/2020GL091287 (2021).

34. Huang, N. E. & Wu, Z. A review on hilbert-huang transform: Method and its applications to geophysical studies. *Rev. Geophys.* **46**, https://doi.org/10.1029/2007RG000228 (2008). https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2007RG000228.

35. Mukhin, D., Gavrilov, A., Loskutov, E., Feigin, A. & Kurths, J. Nonlinear reconstruction of global climate leading modes on decadal scales. *Clim. Dyn.* **51**, 2301–2310. https://doi.org/10.1007/s00382-017-4013-2 (2018).

36. Mukhin, D., Gavrilov, A., Loskutov, E., Kurths, J. & Feigin, A. Bayesian Data Analysis for Revealing Causes of the Middle Pleistocene Transition. *Sci. Reports* **9**, https://doi.org/10.1038/s41598-019-43867-3 (2019).

37. Seleznev, A. & Mukhin, D. Improving statistical prediction and revealing nonlinearity of enso using observations of ocean heat content in the tropical pacific. *Clim. Dyn.* **60**, 1–15. https://doi.org/10.1007/s00382-022-06298-x (2022).

38. Wallace, J. M. & Gutzler, D. S. Teleconnections in the geopotential height field during the northern hemisphere winter. *Mon. Weather. Rev.* **109**, 784–812. https://journals.ametsoc.org/view/journals/mwre/109/4/1520-0493_1981_109_0784_titghf_2_0_co_2.xml (1981).

39. Hughes, J. P., Guttorp, P. & Charles, S. P. A non-homogeneous hidden markov model for precipitation occurrence. *J. Royal Stat. Soc. Ser. C: Appl. Stat.* **48**, 15–30, https://doi.org/10.1111/1467-9876.00136 (2002). https://academic.oup.com/jrsssc/article-pdf/48/1/15/48750484/jrsssc_48_1_15.pdf.

Acknowledgements

Author contributions

D.M. and A.H. conceived the project and wrote the manuscript. D.M. and R.S. implemented the methods and performed calculations. All authors analyzed the results and reviewed the manuscript.

Funding

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-14696-4.

Correspondence and requests for materials should be addressed to D.M. or A.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.