# Analysis of Rayleigh-Bénard convection using latent Dirichlet allocation

B. Podvin and L. Soucasse
*EM2C, CentraleSupélec, CNRS, Université Paris-Saclay, 8-10 rue Joliot Curie, 91192 Gif-sur-Yvette, France*

F. Yvon
*LISN, CNRS, Université Paris-Saclay, Campus Universitaire, Rue Raimond Castaing bâtiment 650, 91190 Gif-sur-Yvette, France*

(Received 20 January 2024; accepted 29 May 2024; published 18 June 2024)

We apply a probabilistic clustering method, latent Dirichlet allocation (LDA), to characterize the large-scale dynamics of Rayleigh-Bénard convection. The method, introduced by Frihat *et al.* [J. Fluid Mech. **920**, A27 (2021)], is applied to a collection of snapshots in the vertical midplanes of a cubic cell for Rayleigh numbers in the range $[10^6, 10^8]$. For the convective heat flux, temperature, and kinetic energy, the decomposition identifies latent factors, called motifs, which consist of connex regions of fluid. Each snapshot is modeled with a sparse combination of motifs, the coefficients of which are called the weights. The spatial extent of the motifs varies across the cell and with the Rayleigh number. We show that the method is able to provide a compact representation of the heat flux and displays good generative properties. At all Rayleigh numbers the dominant heat flux motifs consist of elongated structures located mostly within the vertical boundary layers, at a quarter of the cavity height. Their weights depend on the orientation of the large-scale circulation. A simple model relating the conditionally averaged weight of the motifs to the relative strength of the corner rolls and of the large-scale circulation is found to predict well the average large-scale circulation reorientation rate. Application of LDA to the temperature fluctuations shows that temperature motifs are well correlated with heat flux motifs in space as well as in time, and to some lesser extent with kinetic energy motifs. The abrupt decrease of the reorientation rate observed at $10^8$ is associated with a strong concentration of plumes impinging onto the corners of the cell, which decrease the temperature difference within the corner structures. It is also associated with a reinforcement of the longitudinal wind through formation and entrainment of new plumes.

DOI: 10.1103/PhysRevFluids.9.063502

## I. INTRODUCTION

Rayleigh-Bénard convection, in which a fluid is heated from below and cooled from above, represents an idealized configuration to study thermal convection phenomena. These characterize a variety of applications ranging from industrial processes such as heat exchangers to geophysical flows in the atmosphere or the ocean. A central question is to determine how the heat transfer depends on nondimensional parameters such as the Prandtl number $\mathrm{Pr} = \nu/\kappa$, where $\nu$ is the kinematic viscosity and $\kappa$ the thermal diffusivity, and the Rayleigh number

$$\mathrm{Ra} = \frac{g\beta\Delta T H^3}{\nu\kappa}, \tag{1}$$

where $g$ is the gravity, $\beta$ is the thermal expansion coefficient, $\Delta T$ the temperature difference, and $H$ the cell dimension. The Grossmann-Lohse [1] theory constitutes a unified approach to address this question. It is based on a local description of the physics: the contributions from the bulk

 

averaged thermal and kinetic dissipation rate are split into two subsets, one corresponding to the boundary layers and one corresponding to the bulk. This theory was further refined by Grossmann and Lohse [2], where the thermal dissipation rate was split into a contribution from the plumes and a contribution from the turbulent background. Through the action of buoyancy, the thermal boundary layers generate plumes which create a large-scale circulation, as evidenced by Xi *et al.* [3], also called "wind" [4]. The distribution of temperature fluctuations depends on plume clustering effects [5], but it is also affected by interactions with turbulent fluctuations in the bulk, resulting in fragmentation [6].

Shang *et al.* [7] showed that plume-dominated regions were located near the sidewalls and the conducting surfaces and that thermal plumes carry most of the convective heat flux, which contributes to the production of both kinetic and thermal fluctuations. The morphology of plumes and its effect on the heat transfer have been given careful attention. The plumes have a sheetlike structure near the boundary layer and progressively become mushroomlike as they move into the bulk region [8]. Shishkina and Wagner [9] found that very high values of the local heat flux were observed in regions where the sheetlike plumes merged, constituting "stems" for the mushroomlike plumes developing in the bulk. The relative contributions of the plumes and turbulent background vary with the Rayleigh number: Emran and Schumacher [10] have shown that the fraction of plume-dominated regions decreases with the Rayleigh number, while that of background-dominated regions increases.

The identification of local coherent structures such as plumes is therefore an essential step for the understanding of thermal convection flows. Several definitions have been used: some of the first criteria were based on the skewness of the temperature derivative [11] or the temperature difference [12]. Ching *et al.* [13] have proposed to use simultaneous measurements of the temperature and the velocity to define the velocity of the plumes using conditional averaging. Following Huang *et al.* [14], van der Poel *et al.* [15] identified plumes from both a temperature anomaly and an excess of convective heat flux. Zhou *et al.* [16] relied on cliff-ramp-like structures in the temperature signals to determine the spatial characteristics of plumes. Emran and Schumacher [10] and Vishnu *et al.* [17] separated the plume from the background regions based on a threshold on the convective heat flux. Shevkar *et al.* [18] have recently proposed a dynamic criterion based on the two-dimensional (2D) velocity divergence to separate plumes from boundary layers.

As pointed out by Chillà and Schumacher [19], this multiplicity of criteria illustrates the difficulty of identifying coherent structures in a consistent and objective manner, which is a long-running question in various types of turbulent flows. To this end, proper orthogonal decomposition (POD) [20] has proven a useful tool to analyze large-scale fluctuations in Rayleigh-Bénard convection. It has been used in particular to study reorientations of the large-scale circulation [21–25]. Through spectral decomposition of the autocorrelation tensor, POD provides a basis of spatial modes, also called empirical modes, since they originate from the data. The modes are energetically optimal to reconstruct the fluctuations. The POD modes typically have a global support, which is well suited to capture the large-scale organization of the flow. However, this can make physical interpretation difficult as there is no straightforward connection between a mode and a local coherent structure as a local structure is represented with a superposition of many POD modes, a situation also observed in Fourier analysis. Soucasse *et al.* [26] have used POD to study the dynamics of the large-scale circulation for Rayleigh numbers in the range $[10^6, 10^8]$. They found that although the reorientation rate varied with the Rayleigh number, the dominant structures remained similar across that range, albeit with some variations in their energy. A new dissipation-based POD, proposed by Olesen *et al.* [27] and applied to Rayleigh-Bénard convection [28], highlighted the importance of boundary layers for the dynamics, which points to the need for local descriptions.

As an alternative, Frihat *et al.* [29] have recently adapted a probabilistic method that can extract localized latent factors in turbulent flow measurements. This method, latent Dirichlet allocation or LDA [30,31], was originally developed in the context of natural language processing, where it aims to extract topics from a collection of documents. In this framework, documents are represented by a nonordered set of words taken from a fixed vocabulary. A word count matrix can be built for the

collection, where each column corresponds to a document, each line corresponds to a vocabulary word, and the matrix entry represents the number of times the word appears in the document. LDA provides a probabilistic decomposition of the word count matrix, based on latent factors called *topics*. Topics are defined by two distributions: the distribution of topics within each document (each document is associated with a mixture of topics; the coefficients of the mixture sum up to one) and the distribution of vocabulary words with each topic (each topic is represented by a combination of words, the coefficients of which also sum up to one).

The method has been adapted for turbulent flows as follows: we consider a collection of snapshots of a scalar field discretized into cells. The equivalent of a document is therefore a snapshot, and the cells (or snapshot pixels) constitute the vocabulary. The digitized values of the scalar field over the cells in a snapshot are gathered into a vector which is formally analogous to a column of the word count matrix. The "topics" produced by the decomposition, called *motifs*, correspond to fixed (in the Eulerian sense), spatially coherent regions of the flow. The method was found to be well suited for the representation of intermittent data [29,32]. It was succesfully applied to the analysis of the turbulent Reynolds stress in wall turbulence [29]. Moreoever, the method provides a local description that is insensitive to the existence of global symmetries. It proved a useful tool to identify synoptic objects in weather data [32].

In this paper, we apply this method to the analysis of fluctuations in a cubic Rayleigh-Bénard cell in the range of Rayleigh number $[10^6, 10^8]$. The goal is to track the local signature of the large-scale dynamics of the flow, and to determine whether changes can be identified as the Rayleigh number increases. To this end, the technique is applied to 2D snapshots extracted from three-dimensional numerical simulations of Raleigh-Bénard convection in a cubic cell in the range of Rayleigh number $[10^6, 10^8]$. The numerical configuration and the data set are described in Sec. II. We first present the method for the convective heat flux, using a comparison with POD to highlight the similarities and differences of the approach. POD and LDA are respectively presented in Secs. III and IV. We examine in Sec. V how LDA compares with POD and the extent to which it is able to capture the general features of the heat flux. The characteristics of heat flux motifs and their connection with the reorientations of the large-scale circulation are discussed in Sec. VI. The analysis is then extended to temperature fluctuations and to the kinetic energy in Sec. VII in order to provide further insight into the physics. A conclusion is given in Sec. VIII.

## II. NUMERICAL SETTING

### A. Setup

The numerical setup and associated data sets are the same as used in Refs. [25,26]. The configuration studied is a cubic Rayleigh-Bénard cell filled with air, with isothermal horizontal walls and adiabatic sidewalls. The air is assumed to be transparent and thermal radiation effects are disregarded. Direct numerical simulations have been performed at various values of the Rayleigh number. The Prandtl number is set to 0.707. All physical quantities are made dimensionless using the cell size $H$, the reference time $H^2/(\kappa\sqrt{\mathrm{Ra}})$, and the reduced temperature $\theta = (T - T_0)/\Delta T$, $T_0$ being the mean temperature between hot and cold walls. Spatial coordinates are denoted $x$, $y$, $z$ ($z$ being the vertical direction) and the origin is placed at a bottom corner of the cube.

Navier-Stokes equations under Boussinesq approximation are solved using a Chebyshev collocation method [33,34]. Computations are made parallel using domain decomposition along the vertical direction. Time integration is performed through a second-order semi-implicit scheme. The velocity divergence-free condition is enforced using a projection method. Numerical parameters are given in Table I for the four considered Rayleigh numbers $\mathrm{Ra} = \{10^6; 3\times10^6; 10^7; 10^8\}$. We have checked that the number of collocation points is sufficient to accurately discretize the boundary layers according to the criterion proposed by Shishkina *et al.* [35]. Details on the validation of the numerical method and of the discretization can be found in Ref. [36]. A total of 1000 snapshots have been extracted from the simulations for each Rayleigh number at a sampling period of 10 (at

TABLE I. Characteristics of the data sets at various Rayleigh numbers: spatial resolution $N_x$, $N_y$, $N_z$ in each direction of space, number of snapshot, $N_S$, snapshot sampling period $\Delta t$, and thermal boundary layer thickness $\delta_{BL}$.

| Ra | $(N_x, N_y, N_z)$ | $N_S$ | $\Delta t$ | $\delta_{BL}$ |
|---|---|---|---|---|
| $1\times10^6$ | (81,81,81) | 1000 | 10 | 0.056 |
| $3\times10^6$ | (81,81,81) | 1000 | 10 | 0.042 |
| $1\times10^7$ | (81,81,81) | 1000 | 10 | 0.0297 |
| $1\times10^8$ | (161,161,161) | 1000 | 5 | 0.0167 |

$\mathrm{Ra} = \{10^6; 3\times10^6; 10^7\}$) or 5 (at $\mathrm{Ra} = 10^8$), in dimensionless time units. It is worth noting that the time separation between the snapshots is sufficient to describe the evolution of the large-scale circulation but is not suited for a fine description of the plume emission or of the reorientation process. For each Rayleigh number, a data set satisfying the statistical symmetries of the flow was then constructed from these 1000 snapshots, as will be described in the next section.

### B. Construction of the data set

At each Rayleigh number, the data set consisted of a collection of $N_S = 1000$ snapshots $q(\underline{x}, t_m)$, $m = 1, \ldots, N_S$. Results will be presented first for the convective heat flux $q = \Phi = w\theta$, then for the temperature fluctuations $q = \theta' = \theta - \langle\theta\rangle$ ($\langle\theta\rangle$ being the time-averaged temperature) and for the kinetic energy $q = k = \frac{1}{2}(u^2 + v^2 + w^2)$, $u$, $v$, and $w$ being the velocity components. We note that due to the velocity reference scale, the nondimensional heat flux varies like $\mathrm{NuRa}^{-1/2}$. As in Ref. [25], the data set was first enriched by making use of the statistical symmetries of the flow [37]. In the cubic Rayleigh-Bénard cell, four quasistable states are available for the flow for this Rayleigh number range: the large-scale circulation settles in one of the two diagonal planes of the cube with clockwise or counterclockwise motion. The evolution of the large-scale circulation can be tracked through that of the $x$ and $y$ components of the angular momentum of the cell, $\underline{L} = \int(\underline{x} - \underline{x_0})\times\underline{u}d\underline{x}$, with respect to the cell center $\underline{x_0}$. As Fig. 1 shows at $\mathrm{Ra} = 10^7$, the angular momentum along each horizontal direction oscillates near a quasisteady position for long periods of time—several hundreds of convective timescales—before experiencing a rapid switch [$O(10)$ convective timescales] to the opposite value, which corresponds to a reorientation. On each plane we can define an indicator function $I$, which takes the value $\mathrm{sgn}(L)1$, where $L$ is the angular momentum component normal to the plane.

Reorientations from one state to another occur during the time sequence but each state is not necessarily equally visited. In order to counteract this bias, we have built enlarged snapshot sets, obtained by the action of the symmetry group of the problem on the original snapshot sets. The

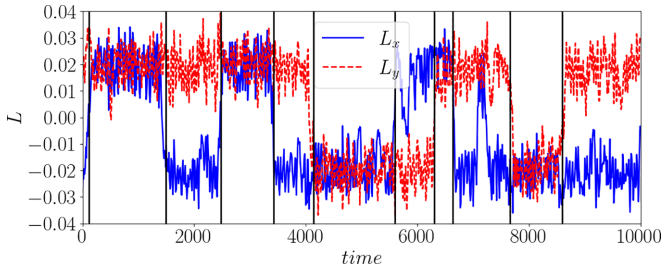


FIG. 1. Evolution of the horizontal components of the angular momentum at $\mathrm{Ra} = 10^7$. The vertical black lines correspond to reorientations of the large-scale circulation.

symmetries are based on four independent symmetries $S_x$, $S_y$, $S_z$, and $S_d$ with respect to the planes $x = 0.5$, $y = 0.5$, $z = 0.5$, and $x = y$. This generates a group of 16 symmetries for the cube, which should lead to a 16-fold number of snapshots. However, since we will exclusively consider the vertical midplanes $x = 0.5$ and $y = 0.5$, which are invariant planes for respectively $S_x$ and $S_y$, the increase is reduced. The data set aggregates 1000 snapshots on each of the planes $x = 0.5$ and $y = 0.5$, each of which undergoes a vertical flip, a horizontal flip, and a combination of the two, yielding a total of $N_S = 8000$ snapshots.

The LDA technique requires transforming the data into a non-negative integer field. The signal defined on a grid of $\tilde{N}_C$ cells was digitized using a rescaling factor $s$. If the field was not of constant sign (temperature, heat flux), positive and negative values were split onto two distinct grids, leading to a field defined on $N_C = 2\tilde{N}_C$ cells. For the heat flux, this gives

$$q(\underline{x}_j) = q(\underline{x}_j, t_m) = \mathrm{Max}[\mathrm{Int}[s\, w(\underline{x}_j, t_m)\theta(\underline{x}_j, t_m), 0], \tag{2}$$

$$q(\underline{x}_{j+\tilde{N}_C}) = q(\underline{x}_{j+\tilde{N}_C}, t_m) = -\mathrm{Max}[-\mathrm{Int}[s\, w(\underline{x}_j, t_m)\theta(\underline{x}_j, t_m), 0], \tag{3}$$

where $s > 0$, $m \in [1, N_S]$, and $j \in [1, \tilde{N}_C]$ and $\underline{x}_j$ represents the $j$th cell location on the midplanes $x = 0.5$ or $y = 0.5$. We note that throughout the paper, the total field will directly be represented on the physical grid of size $\tilde{N}_C$ from the renormalized difference $[q(\underline{x}_j, t_m) - q(\underline{x}_{j+\tilde{N}_C}, t_m)]/s$. It is worth noting that the temperature variance (always positive) could be used to lighten the LDA analysis on the temperature field. Yet, we chose to work on the signed temperature fluctuation in order to discriminate between leaving and impinging thermal patterns near the horizontal walls as it is often done in plume detection [12,15].

## III. POD ANALYSIS

### A. Method

Proper orthogonal decomposition (POD) [38] makes it possible to write a collection of $N_S$ spatial fields $q(\underline{x}_j, t_m)$ defined on $N_C$ grid points, as a superposition of spatial modes $\varphi_n(\underline{x})$, the amplitude of which varies in time:

$$q(\underline{x}_j, t_m) = \sum_{n=1}^{N_S} \sqrt{\lambda_n} a_n(t_m)\varphi_n(\underline{x}_j), \tag{4}$$

with $m \in [1, N_S]$ and $j \in [1, N_C]$. The spatial modes $\varphi_n(\underline{x})$ are orthonormal:

$$\sum_{j=1}^{N_C} \varphi_n(\underline{x}_j)\varphi_m(\underline{x}_j) = \delta_{nm}. \tag{5}$$

The amplitudes $a_n(t_m)$ are normalized eigenvectors of the eigenvalue problem

$$C_{mp} a_n(t_p) = \lambda_n a_n(t_m), \tag{6}$$

where $C$ is the temporal autocorrelation matrix

$$C_{mp} = \frac{1}{N_S} \sum_{j=1}^{N_C} q(\underline{x}_j, t_m) q(\underline{x}_j, t_p). \tag{7}$$

The eigenvalues $\lambda_n$, such that $\lambda_1 > \lambda_2 > \lambda_3 > \cdots$, represent the respective contribution of the modes to the total variance. If we consider the $p$ most energetic modes, the reconstruction based on $p$ modes minimizes the $L_2$-norm error between the set of snapshots and the projection of the set of snapshots onto a basis of size $p$.

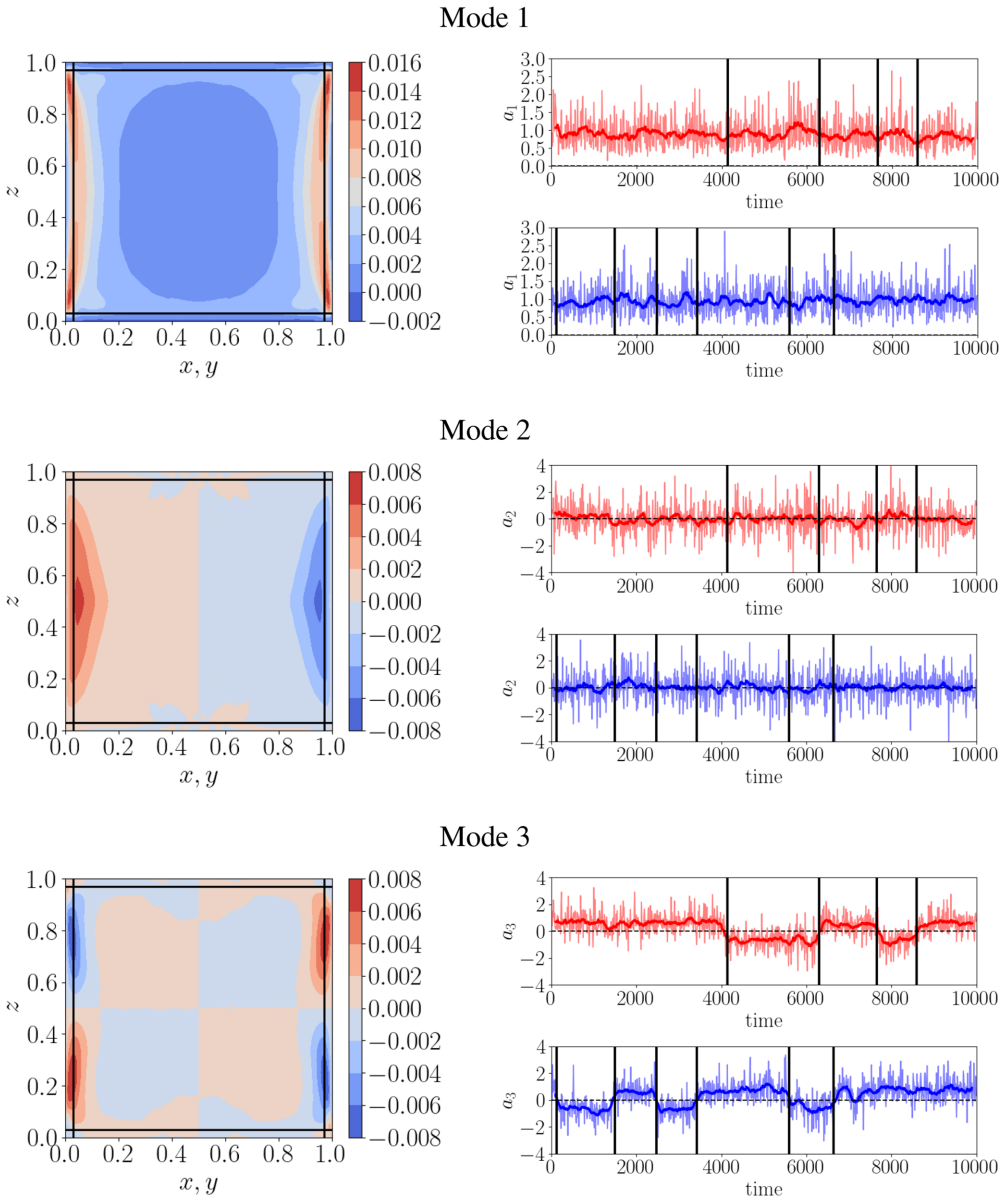


FIG. 2. POD dominant modes and amplitudes in the vertical midplane at Ra $= 10^7$. Left: POD modes $\varphi_n$. Right: POD amplitudes $a_n$ associated with plane $x = 0.5$ (in blue) and plane $y = 0.5$ (in red). The vertical black lines correspond to changes in the component of the angular momentum. The darker line corresponds to a moving average over 200 convective units.

### B. Application to the convective heat flux

POD is applied to the digitized heat flux signal $q = \Phi$ defined in Eqs. (2) and (3). The first three POD modes and POD coefficients are shown in Fig. 2 for Ra $= 10^7$, where black vertical and horizontal lines indicate the thickness of the boundary layers. We checked that the first mode corresponds to the mean flow. The mode is most important in a region close to the wall, with a maximum within the vertical boundary layer at a height of $z \approx 0.1$. The second mode corresponds

to a dissymmetry between the vertical sides and is most important at mid-height in the region outside the boundary layers. The third mode is antisymmetric in both the vertical and horizontal directions. It is maximum at the edge of the vertical boundary layers, at a vertical distance of about 0.25 from the horizontal surfaces. The pattern it is associated with corresponds to a more intense flux along a diagonal (bottom of one side and top of the opposite side) and a less intense flux along the opposite diagonal. As evidenced by application of a moving average performed over 200 convective time units (about four times the recirculation time $T_c$, as was determined in Ref. [25]), the evolution of the amplitude at large timescales matches that of the horizontal angular momentum components $L_x$ and $L_y$ (compare with Fig. 1), unlike the two dominant modes. This mode therefore appears to be the signature of the large-scale circulation, where the flux is more intense in the lower corner of the cell as hot plumes rise on one side and in the upper corner of the opposite side of the cell as cold plumes go down.

## IV. LATENT DIRICHLET ALLOCATION

### A. Principles

We briefly review the principles of latent Dirichlet allocation and refer the reader to Ref. [29] for more details. LDA is an inference approach to identify latent factors in a collection of observed data, which relies on Dirichlet distributions as priors.

We first recall the definition of a Dirichlet distribution $\vartheta$, which is a multivariate probability distribution over the space of multinomial distributions. It is parametrized by a vector of positive-valued parameters $\underline{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_N)$ as

$$p(\vartheta_1, \ldots, \vartheta_N; \alpha_1, \ldots, \alpha_N) = \frac{1}{B(\underline{\alpha})} \prod_{n=1}^{N} \vartheta_n^{\alpha_n - 1}, \tag{8}$$

where $B$ is a normalizing factor, which can be expressed in terms of the gamma function $\Gamma$:

$$B(\underline{\alpha}) = \frac{\prod_{n=1}^{N} \Gamma(\alpha_n)}{\Gamma\left(\sum_{n=1}^{N} \alpha_n\right)}. \tag{9}$$

The components $\{\alpha_n, n = 1, \ldots, N\}$ of $\underline{\alpha}$ control the sparsity of the distribution: values of $\alpha_n$ larger than unity correspond to evenly dense distributions, while values lower than unity correspond to sparse distributions.

As mentioned above, the data to which LDA is applied consist of a collection of non-negative integer fields that are defined in Eqs. (2) and (3). For each snapshot $m$, the integer value $q_m(\underline{x}_j)$ measured at cell $j$ is interpreted as an integer count of the cell $j$. The key is to interpret this integer count as the number of times cell $j$ appears in the composition of snapshot $m$. The idea is to construct a model for the probability $p(q_m, \underline{x}_j)$ of observing the cell $\underline{x}_j$ in the snapshot $q_m$, which is directly proportional to $q(\underline{x}_j)$. The model is based on the hypothesis that each snapshot of the collection $\{q_m, m = 1, \ldots, N_S\}$ consists of a mixture of $N_T$ latent factors $\{z_n, n = 1, \ldots, N_T\}$ called motifs, $N_T$ being a user-defined parameter analogous to a number of clusters. The probability $p(q_m, \underline{x}_j)$ therefore is written

$$p(q_m, \underline{x}_j) = p(q_m) \sum_n p(z_n|q_m) p(\underline{x}_j|z_n), \tag{10}$$

where $p(q_m)$ is the probability of observing the snapshot $q_m$ in the collection, $p(z_n|q_m)$ is the conditional probability of observing motif $z_n$ given the presence of snapshot $q_m$, and $p(\underline{x}_j|z_n)$ is the conditional probability of observing cell $\underline{x}_j$ given the latent factor $z_n$.

A formal analogy with POD and Eq. (4) can be seen by using the Bayes rule and rewriting $p(q_m, \underline{x}_j)$ as

$$p(q_m, \underline{x}_j) = \sum_n p(z_n) p(q_m|z_n) p(\underline{x}_j|z_n), \tag{11}$$

where $p(z_n)$ is the equivalent of the rms contribution $\sqrt{\lambda_n}$, $p(\underline{x}_j|z_n)$ is the equivalent of the POD spatial mode $\varphi_n(\underline{x}_j)$, and $p(q_m|z_n)$ is the equivalent of the temporal amplitude $a_n(t_m)$.

We emphasize that, unlike POD, all quantities in the LDA model are probabilities and therefore non-negative.

Latent Dirichlet allocation is therefore based on the following representation:

(1) Each motif $z_n$ is associated with a multinomial distribution $\psi_n$ over the grid cells so that the probability to observe the $j$th grid cell located at $\underline{x}_j$ given the motif $n$ is $p(\underline{x}_j|z_n) = \psi_n(x_j)$. The distribution $\psi_n$ is modeled with a Dirichlet prior parametrized with an $N_C$-dimensional vector $\underline{\eta}$. Low values of $\eta_l$ mean that the motif is distributed over a small number of cells.

(2) Each snapshot $q_m$ is associated with a distribution $b_n$ over the motifs such that the probability that motif $n$ is present in snapshot $m$ will be denoted $p(q_m|z_n) = b_n(t_m)$. This distribution is modeled with an $N_T$-dimensional Dirichlet distribution of parameter $\underline{\alpha}$. The magnitude of $\alpha$ characterizes the sparsity of the distribution. Low values of $\alpha_n$ mean that relatively few motifs are observed in each snapshot.

### B. Implementation

The snapshot-motif distribution $b_n$ and the motif-cell distribution $\psi_n$ are determined from the observed snapshots $q(\underline{x})$ and constitute $N_T$-and $N_C$-dimensional categorical distributions. Finding the distributions $b_n$ and $\psi_n$ that are most compatible with the observations constitutes an inference problem. The problem can be solved either with a Markov chain Monte Carlo (MCMC) algorithm such as Gibbs sampling [30], or by a variational approach [31], which aims to minimize the Kullback-Leibler divergence between the true posterior and its variational approximation. In both cases, the computational complexity of the problem is of the order of $N_C \times N_S \times N_T$.

The solution *a priori* depends on the number of motifs, $N_T$, as well as on the values of the Dirichlet parameters $\underline{\alpha}$ and $\underline{\eta}$. Special attention was therefore given to establish the robustness of the results reported here. Noninformative default values were used for the Dirichlet parameters; i.e., the prior distributions were taken with symmetric parameters equal to $\forall n, \alpha_n = 1/N_T$ and $\forall j, \eta_j = 1/N_T$. Practical implementation was performed in PYTHON using Gensim [39]. No significant change was observed in the results when the value of the quantization $s$ was high enough (however, it had to be kept reasonably low in order to limit the computational time). Although multiple tests were carried out for varying values of $s \in [40, 600]$, all results reported in this paper were obtained with $s = 600$ for the heat flux. Values of $s = 40$ and $s = 50$ were respectively used for the temperature fluctuations and for the kinetic energy. Analyses were also performed for varying numbers of motifs, $N_T$, ranging from 50 to 400.

### C. LDA as a generative process

The standard generative process performed by LDA with $N_T$ motifs is the following.

(i) For each motif $n$, an $N_C$-dimensional cell-motif distribution $\psi_n$ is drawn from the Dirichlet distribution of parameter $\underline{\eta}$.

(ii) To generate snapshot $m$:

(a) An $N_T$-dimensional snapshot-motif distribution $b_n$ is drawn according to a Dirichlet distribution parameterized by $\underline{\alpha}$.

(b) A total integer count $q_T(t_m)$ is drawn. This number corresponds to the total number of cell integer counts associated with snapshot $m$, i.e., $\sum_j q(\underline{x}_j, t_m)$. $q_T(t_m)$ is typically sampled from a Poisson distribution that matches the statistics of the original database.

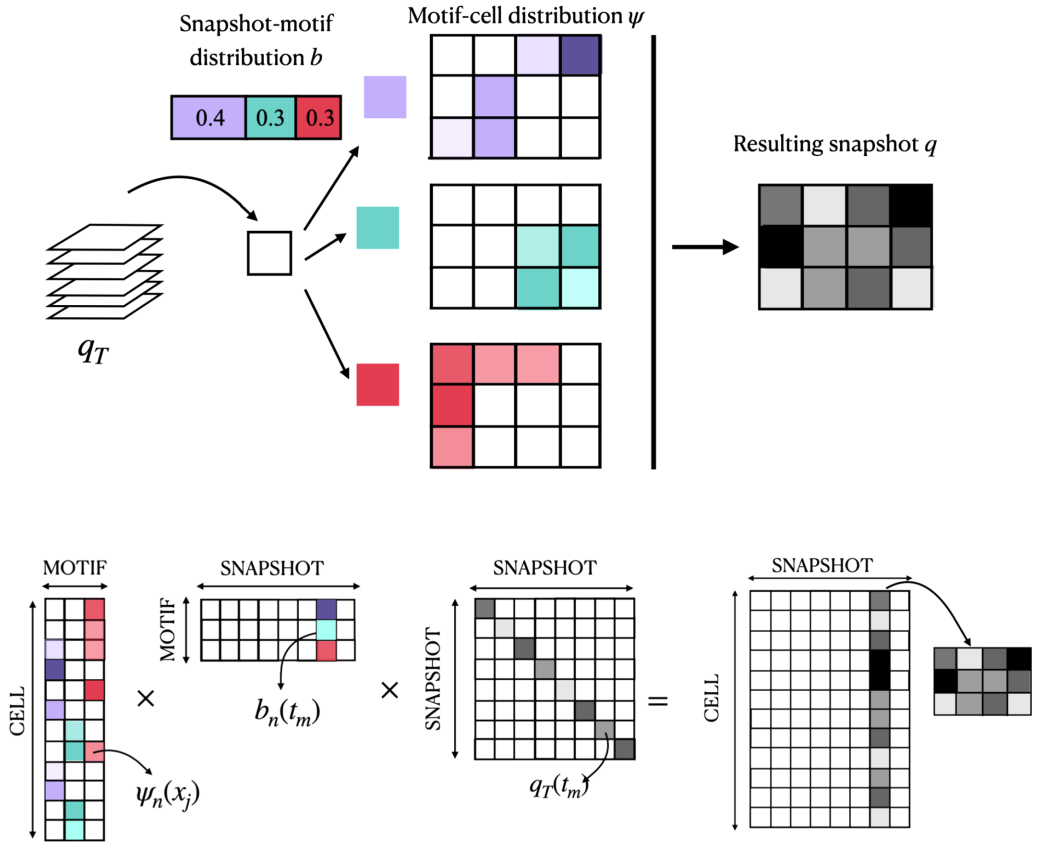


FIG. 3. Schematics of the LDA generative model illustrated here for a field defined on 12 cells and generated with 3 motifs (corresponding to purple, green, and red). A snapshot $m$ is represented as a set of integer values defined on an array of cells (see also text). Top: Probabilistic construction of a snapshot. Let us consider a stack of $q_T$ tokens of unit value. Each token is assigned to a cell as follows: a motif $n$ is selected by sampling the snapshot-motif distribution $b(t_m)$ corresponding to this snapshot. In the example shown, the probabilities for the purple, green, and red motifs are respectively 40%, 30%, and 30%. Once the motif $n$ is chosen, a cell $j$ is selected by sampling the motif-cell distribution $\psi_n$. At the end of the process, the number of tokens at cell $j$ yields the value of the field $q(\underline{x}_j, t_m)$. Bottom: Matrix-based reconstruction. Each snapshot $m$ is obtained by summing the contributions of all distributions $\psi_n(x_j)$ weighted by the corresponding probabilities $b_n(t_m)$, and rescaling the sum with a factor $q_T(t_m)$.

(c) For each $i = 1, \ldots, q_T(t_m)$:

(1) A motif $n$ is selected from $b_n(t_m)$ (since it represents the probability that motif $n$ is present in the snapshot $m$).

(2) Once this motif $n$ is chosen, a cell $j$ is selected from $\psi_n(\underline{x}_j)$ (since it represents the probability that cell $j$ is present in motif $n$).

The snapshot $m$ then represents the set of $q_T$ cells $j$ that have been drawn and can be reorganized as a list of $N_C$ cells with integer counts $q(\underline{x}_j, t_m)$. Figure 3 (top) illustrates the LDA generative process on a $4\times3$ grid for three topics.

In fluid mechanics applications [29,32], sampling from the motif-cell distribution [step (c)] can be replaced with a faster step, where the contribution of each motif $n$ to snapshot $m$ is directly obtained from the motif-cell distribution $\psi_n$ and the distribution $b_n(t_m)$ and expressed

as $q_T(t_m)b_n(t_m)\psi_n(\underline{x}_j)$. The reconstructed field is then the sum of the motif contributions. This matrixlike form of the reconstruction is summarized in the bottom part of Fig. 3.

### D. Interpretation and evaluation criteria

By construction, the decomposition identifies *fixed* regions of space over which the intensity of the scalar field is likely to be important at the same time. The connection between temperature motifs and plumes should be examined with caution since plumes are Lagrangian structures traveling and possibly changing in shape and orientation through the shell. LDA motifs only aim to detect the Eulerian signature of structures.

Each motif $n$ can be characterized in space through the motif-cell distribution $\psi_n$ (which integrates to 1 over the cells) and which will sometimes be referred to as the motif in the absence of ambiguity. Each distribution has a maximum value $\psi_n^{\max}$ and a maximum location $\underline{x}_n^{\max}$ such that $\psi_n(\underline{x}_n^{\max}) = \psi^{\max}$. One can also define a characteristic area $\Sigma_n$ using

$$\Sigma_n = \int_\Omega 1_{\{\psi_n \geq \psi_n^{\max}/e\}} d\Omega, \tag{12}$$

where $\Omega$ represents the plane of analysis and the factor $1/e \sim 0.606$ is an arbitrary factor chosen by analogy with a Gaussian distribution. If $\psi_n$ were a Gaussian of standard deviation $\sigma$, this value would delimit an area of size $2\pi\sigma^2$. Other choices could be made such as the full width at half maximum corresponding to a factor of 1/2. Moderate changes in the choice of the factor did not affect the trends reported below. Characteristic dimensions $l_i$ for the motif $n$ in the direction $i$ can also be defined using $l_i^n = [\int \psi_n(x_{n,i} - x_{n,i}^{\max})^2 dx_i]^{1/2}$. Each motif can also be characterized in time through the snapshot-motif distribution $b_n$, which will be called the motif *weight* throughout the paper. The motifs can be ordered by their time-averaged weight, also called *prevalence*, defined as $\langle b_n \rangle = \frac{1}{N_S} \sum_{m=1}^{N_S} b_n(t_m)$, where $\langle \cdot \rangle$ represents a time average.

LDA decompositions were carried out independently for the heat flux $\Phi = w\theta$, temperature fluctuations $\theta'$, and the total kinetic energy $k = \frac{1}{2}(u^2 + v^2 + w^2)$. To differentiate between these quantities, the motif topics and weights are denoted respectively as $\psi_n^\Phi$, $\psi_n^\theta$, and $\psi_n^k$ and $b_n^\Phi$, $b_n^\theta$, and $b_n^k$. A useful tool for comparing the motifs associated with two different quantities is to compute the correlation coefficient matrix between the corresponding motif weights (for instance, if we compare the heat flux and the temperature motifs, each $(n, n')$ entry of the matrix will correspond to the correlation coefficient between $b_n^\Phi$ and $b_{n'}^\theta$).

As noted above, a reconstruction of the field can be obtained by using the inferred motif-cell distribution and snapshot-motif distribution to provide what we will call the LDA-reconstructed field, defined as

$$q_R(\underline{x}_j, t_m) = \sum_{n=1}^{N_T} q_T(t_m) b_n^q(t_m) \psi_n^q(\underline{x}_j). \tag{13}$$

This equation can be compared to Eq. (11) for a probabilistic interpretation and to Eq. (4) for an analogy with POD. To evaluate the relevance of the decomposition, one can compute for each snapshot $m$ the instantaneous spatial correlation coefficient $C_m$ between a given field $q$ and its reconstruction $q_R$ defined as

$$C_m(q, q_R) = \frac{\int (\tilde{q}(\underline{x}, t_m)\tilde{q}_R(\underline{x}, t_m) d\underline{x}}{\left( \int \tilde{q}^2(\underline{x}, t_m) d\underline{x} \int \tilde{q}_R^2(\underline{x}, t_m) d\underline{x} \right)^{1/2}}, \tag{14}$$

where $\tilde{q}$ represents the fluctuation $\tilde{q}(\underline{x}, t_m) = q(\underline{x}, t_m) - 1/|\Omega| \int_\Omega q(\underline{x}, t_m) d\underline{x}$. A global measure of the reconstruction is then given by $\langle C \rangle = \frac{1}{N_S} \sum_{m=1}^{N_S} C_m$, the average value of $C$ over all snapshots.

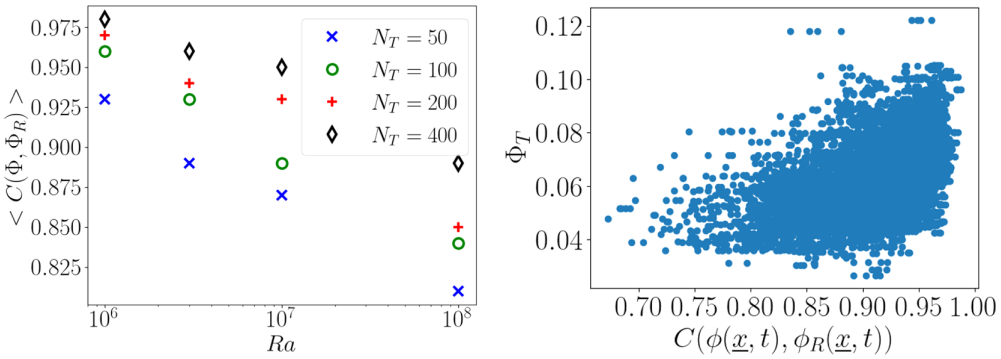


FIG. 4. Left: Instantaneous correlation coefficient between the projected and the true field as a function of the integral convective heat flux for $N_T = 100$ and $\mathrm{Ra} = 10^7$. Right: Average correlation coefficient $\langle C(\Phi, \Phi_R) \rangle$ as a function of the Rayleigh number and of the number of topics considered for both midplanes.

## V. EVALUATION OF LDA FOR RECONSTRUCTION AND GENERATION OF THE HEAT FLUX

### A. Reconstruction

We first evaluate to which extent the LDA decomposition provides an adequate reconstruction of the heat flux $\Phi$. Figure 4 (left) shows how the instantaneous value of the correlation coefficient $C_m(\Phi, \Phi_R)$ depends on the discrete integral of the field $q_T(t_m) = \sum_j \Phi(\underline{x}_j, t_m)$. The Rayleigh number considered is $\mathrm{Ra} = 10^7$ and the number of topics is $N_T = 100$, but the same trend was observed for all other Rayleigh numbers as well as all other values of $N_T$. Lower values of the correlation were associated with lower values of the total integrated heat flux, which illustrates that the LDA representation is suited to capture extreme events.

Figure 4 (right) presents the mean correlation coefficient $\langle C_m(\Phi, \Phi_R) \rangle$ for different numbers of motifs and different Rayleigh numbers on the vertical planes. Unsurprisingly, the correlation increases with the number of topics. It also decreases with the Rayleigh number, which is consistent with an increase in the complexity of the flow. However, the minimum value for the lower number of topics and the highest Rayleigh number was 0.8, which shows the relevance of the decomposition.

Figure 5 compares an original snapshot at $\mathrm{Ra} = 10^7$ (based on the digitized signal) with different reconstructions: (i) the LDA reconstruction based on $N_T = 100$ motifs, (ii) the reconstruction limited to the 20 most prevalent topics (for this particular snapshot), and (iii) the POD-based reconstruction based on the first 20 modes. By construction, POD provides the best approximation of the field for a given number of modes. Since the distribution of the heat flux is intermittent in space and time, only a limited number of motifs is necessary to reconstruct the flow. We note that little difference was observed between the full LDA reconstruction and the reconstruction limited

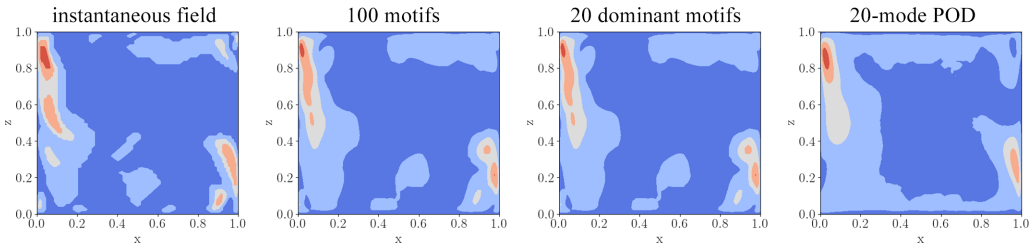


FIG. 5. Example of an instantaneous snapshot and its reconstructions at $\mathrm{Ra} = 10^7$. From left to right: original field, LDA-reconstructed field using $N_T = 100$ motifs, LDA-reconstructed field using the 20 (instantaneously) most prevalent motifs, and POD-reconstructed field using the 20 (on average) most energetic modes.

to the 20 most prevalent motifs, which highlights the intermittent nature of the field. The relative error between the original and the reconstructed field is 29% for the full LDA reconstruction, 34% when the 20 most prevalent modes are retained in the reconstruction. In contrast, limiting the POD to 20 global modes slightly lowers the quality of the reconstruction, with a global error of 38%. It should be noted that the 20 dominant POD modes correspond to an average over all snapshots, while the 20 most prevalent LDA modes are selected for that specific snapshot. On average, the reconstructed field based on keeping the 20 most prevalent motifs differed by less than 10% from the full 100-mode reconstruction and the average correlation coefficient $C = \langle C_m(\Phi, \Phi_R)\rangle$ decreased from 0.89 to 0.83. This shows that LDA can provide a compact representation of the local heat flux that compares reasonably well with POD.

### B. Generation

The ability to generate statistically relevant synthetic fields is of interest for a number of applications, such as accelerating computations or developing multiphysics models. As a generative model, LDA makes it possible to produce such a set of fields, the statistics of which can be compared with those of the original fields used to extract the motifs, as well as with those of the corresponding LDA-reconstructed fields. It would also be useful to compare the generated LDA data set with one generated using POD. To this end, we generated two sets of 4000 new fields using both LDA and POD, following the procedure described in Sec. IV C and illustrated in Fig. 3. The same number $N_T = 100$ of POD modes and LDA motifs was used to generate the data sets. The plane in which the data are generated is assumed to be the $y = 0.5$ plane. The different fields to be compared are therefore the following:

(1) The original (digitized) field $\Phi$ is defined in Sec. II B with Eqs. (2) and (3).

(2) The LDA-reconstructed (LDA-R) field is defined in Eq. (13).

(3) The LDA-generated (LDA-G) field, as described in Sec. IV C, is constructed by sampling weights $\tilde{b}_n(t_m)$ from snapshot-motif distributions and then reconstructing

$$\Phi^{\text{LDA-G}}(\underline{x}_j, t_m) = \Phi_T(t_m) \sum_{n=1}^{N_T} \tilde{b}_n^{\Phi}(t_m)\psi_n^{\Phi}(\underline{x}_j), \tag{15}$$

where $\Phi_T$ represents the $L_1$ spatial norm of the heat flux. For the snapshots of the original database, $\Phi_T(t_m) = \sum_j |\Phi(\underline{x}_j, t_m)|$. For the synthetic fields, $\Phi_T$ is modeled as a random variable obtained by sampling a Poisson distribution with the same mean and variance as the original database.

(4) The POD-generated (POD-G) field is constructed by independently sampling $N_T$ POD mode amplitudes $\tilde{a}_n$ from the POD amplitudes of the original database:

$$\Phi^{\text{POD-G}}(\underline{x}_j, t_m) = \sum_{n=1}^{N_T} \sqrt{\lambda_n}\tilde{a}_n(t_m)\varphi_n(\underline{x}_j). \tag{16}$$

The time-averaged fields corresponding to the different databases are compared in Fig. 6. A good agreement is observed for all data sets, with global errors of 4%, 8%, and 3% for respectively the LDA-reconstructed, the LDA-generated, and the POD-generated data sets. Although it provides the lowest error (as could be expected), the POD-generated data set overestimates negative values in the core of the cell.

For a given location $(y_0, z_0)$, we defined spatial autocorrelation functions in the horizontal and vertical directions as

$$R_y(y, y_0, z_0) = \frac{\langle \Phi(y, z_0, t)\Phi(y_0, z_0, t)\rangle}{\langle \Phi(y_0, z_0, t)^2\rangle}, \tag{17}$$

$$R_z(z, y_0, z_0) = \frac{\langle \Phi(y_0, z, t)\Phi(y_0, z_0, t)\rangle}{\langle \Phi(y_0, z_0, t)^2\rangle}. \tag{18}$$

The autocorrelation functions are displayed in Fig. 7 for the selected locations indicated in Fig. 6, which correspond to regions of high heat flux. We can see that, in all cases, the flux

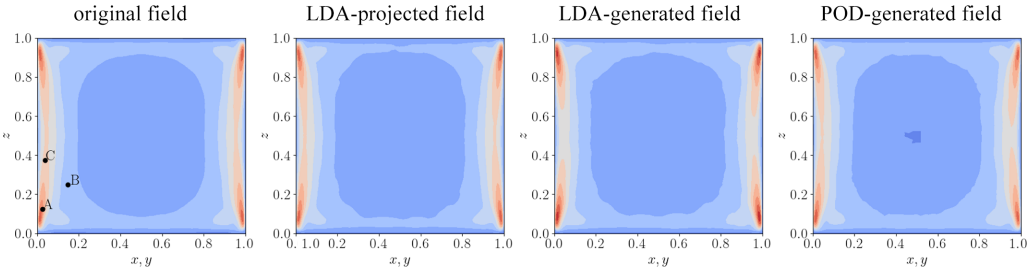


FIG. 6. Time-averaged value of the convective heat flux for different databases at $\mathrm{Ra} = 10^7$. From left to right: original fields, LDA-reconstructed (LDA-R) fields using $N_T = 100$ motifs, LDA-generated (LDA-G) fields using $N_T = 100$ motifs, and POD-generated (POD-G) fields using 100 modes.

remains correlated over much longer vertical extents than in the horizontal direction. Both the LDA-reconstructed and the POD-generated autocorrelations approximate the original data well; again, by construction, POD-based fields are optimal to reconstruct second-order statistics. The LDA-generated autocorrelation is not as close to the original one, but still manages to capture the characteristic spatial scale over which the fields are correlated.

One-point probability density functions (PDFs) of the flux $\Phi$ are represented in Fig. 8 for the same selected locations (again, indicated in Fig. 6). POD-generated fields tend to overpredict lower values and underpredict higher values, which means that they do not capture well the intermittent features of the heat flux. The LDA-generated fields display a better agreement with the original fields and are in particular able to reproduce the exponential tails of the distributions.

## VI. HEAT FLUX MOTIFS

### A. Spatial organization

We now describe the spatial organization of the motifs through the motif-cell distribution $\psi_n$. The general trends reported below held for all values of $N_T$ considered, which ranged from 50 to 400. For all Rayleigh numbers, most LDA motifs were found to be associated with a positive flux (i.e., they

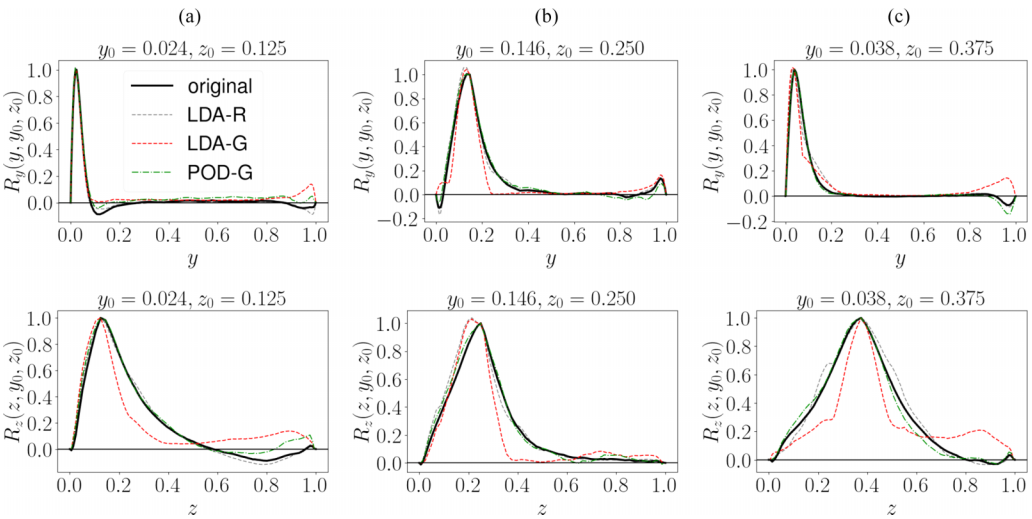


FIG. 7. Autocorrelation of the convective heat flux at selected locations (see Fig. 6).

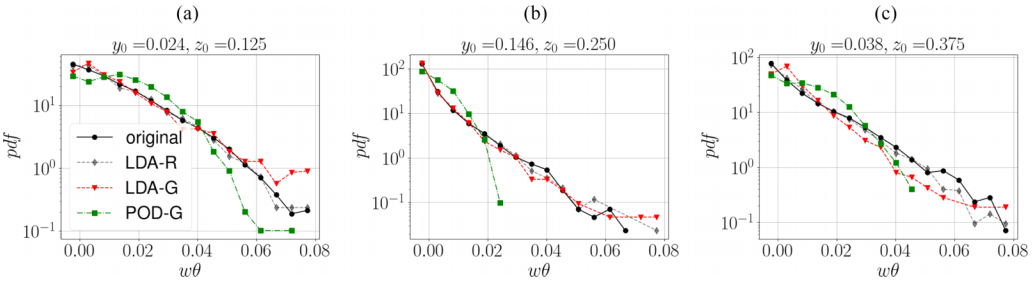


FIG. 8. Probability density function of the convective heat flux at the selected locations indicated in Fig. 6.

were associated with the first $\tilde{N}$ cells in the decomposition). A few negative (countergradient) motifs were also identified, but their average weight was generally very small (at most 10% of that of the dominant motif). We therefore chose to focus only on the motifs making a positive contribution to the heat flux. Figure 9 (left) displays these motifs for three different Rayleigh numbers for $N_T = 100$. The case $\mathrm{Ra} = 3\times10^6$ was omitted as it did not show significant differences with the case $\mathrm{Ra} = 10^6$. The motif-cell distribution is materialized by a black line corresponding to the isoprobability contour of $0.606\psi_n^{\max}$, which can be compared with the average value of the heat flux at this location. For all Rayleigh numbers, the motifs are clustered in the regions of high heat flux, close to the vertical walls. Within the vertical boundary layers, motifs are elongated in shape. Outside the vertical boundary layers, the motifs are more isotropic and tend to increase in shape as one moves away from the walls. Outside the horizontal boundary layers, the motif-cell distributions are elongated in the direction of the wind, with a horizontal orientation in the center of the cell, and a gradual vertical shift closer to the walls. Large motifs are found in the bulk at $\mathrm{Ra} = 10^6$ and $\mathrm{Ra} = 10^7$ (it was also the case at $\mathrm{Ra} = 3\times10^6$). In contrast, fewer, smaller motifs are found in the bulk at $\mathrm{Ra} = 10^8$ in the central region $x/y \in [0.2, 0.8]$, signaling a loss of spatial coherence in the bulk at this Rayleigh number.

In general, the motif size seems to decrease with the Rayleigh number. This is confirmed by Fig. 10, which represents the average motif area as a function of their distance from the vertical walls. In order to avoid the influence of the horizontal plates, we only considered the motifs located at a vertical distance larger than 0.07 from the horizontal walls (i.e., outside the horizontal boundary layer). The size of the symbols shown in the picture is proportional to the fraction of motifs over which the average was performed. Results were relatively robust with respect to the number of topics, $N_T$, although some dependence on $N_T$ is observed in the center of the cell. Within the

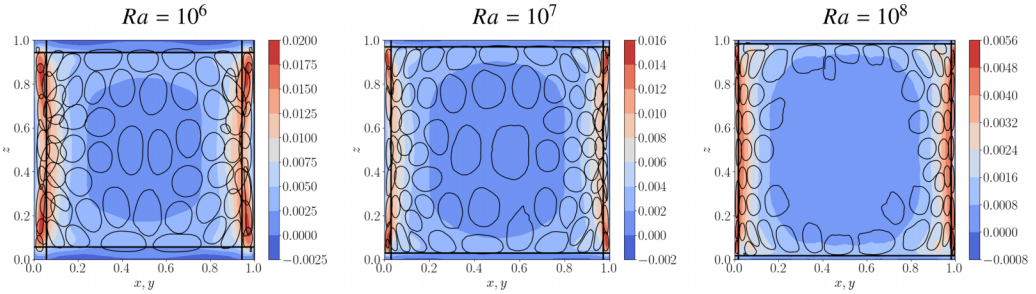


FIG. 9. Spatial distribution of the positive heat flux motifs $\psi_n^{\Phi}$ in the vertical midplane for $N_T = 100$. The motifs are materialized by a black line corresponding to a probability contour of $0.606\psi_n^{\max}$. The vertical lines correspond to the boundary layer thickness. The time-averaged convective heat flux is represented in the background.

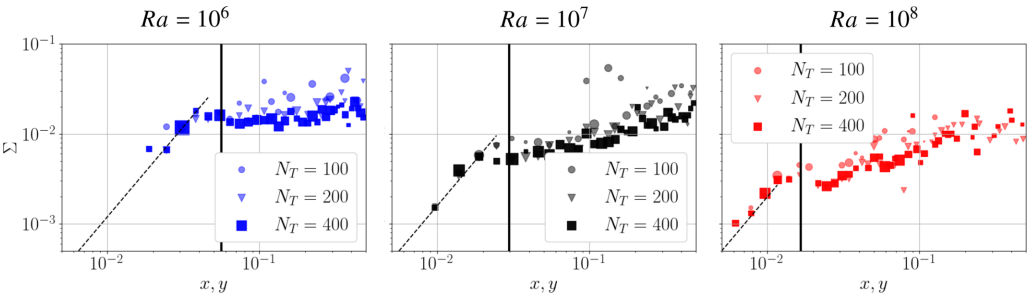


FIG. 10. Distribution of motif areas [see definition in Eq. (12)] in the vertical midplane with the distance from the lateral walls at varying Rayleigh numbers. The size of the symbols shown in the picture is proportional to the fraction of motifs over which the average was performed. The black solid lines indicate the boundary layer thickness. The dashed lines have slope 2.

boundary layer, the motif area grows roughly quadratically (a power-law fit yielded exponents in the range 1.6–2 at all Rayleigh numbers), which means that the characteristic motif size of the motif essentially grows like the wall distance. We note that a similar scaling was found for turbulent eddies in pressure-gradient-driven turbulence such as channel flow [29]. Farther away from the vertical wall, after a short plateau at the edge of the boundary layer, a slower increase in the motif size was observed with a rate that increased with the Rayleigh number, so that the motif area was about the same (on the order of 0.02) for all Rayleigh numbers in the center of the cell. This suggests the presence of a double scaling for the motifs: one based on the boundary layer thickness, and one based on the cell size. The decrease in size with the Rayleigh number appears consistent with a dependence on the boundary layer thickness but also with an increase of the fragmentation by the bulk turbulent fluctuations, in agreement with the literature [6,15]. The difference observed at the highest Rayleigh number also signals that the flow is still evolving and has not reached an asymptotic state.

### B. Dominant motifs

#### *1. Spatial description*

Owing to the symmetry of the database (see Sec. II B), the motifs in the vertical plane $(x, z)$ $[(y, z)]$ should approximate the symmetry $S_x : x \to 1 - x$ $(S_y : y \to 1 - y)$, and $S_z : z \to 1 - z$ (complete symmetry cannot be expected owing to the stochastic nature of the decomposition).

To help interpret the heat flux motifs, we compare them with LDA motifs corresponding to temperature fluctuations. The eight most prevalent heat motifs are represented in Fig. 11 (green lines). The prevalence of each motif is indicated at the top of each plot. Most motifs have similar sizes and are located close to the sidewalls at about a similar height, except for motifs 4 and 6, which have a smaller extent and are located closer to the horizontal wall. The same value of $N_T = 100$ was used for both heat flux and temperature.

For a heat flux motif $n$ with weight $b_n^\Phi$, we identified the temperature motif $j$ that maximized the correlation coefficient between the heat flux and the temperature motif weights $C(b_n^\Phi, b_{n'}^\theta)$. The maximal value of this coefficient, denoted $c$, is represented on each plot and is generally very high (about 0.7), especially in view of the intermittent nature of the weights. The best correlated heat flux and temperature motifs are close to each other in space, with a larger spread for temperature motifs. In all cases, flux motifs in the lower (higher) portion of the sidewalls correspond to positive (negative) fluctuations. Dominant heat flux motifs can be therefore interpreted as the wall imprint of hot plumes rising in the boundary layer (cold plumes descending in the boundary layer). The same observations were made at all other Rayleigh numbers.

Four of these dominant motifs at $\mathrm{Ra} = 10^7$ are represented in Fig. 12 (left) for $N_T = 100$. As noted above, they consist of elongated structures lying mostly in the boundary layer, and located at

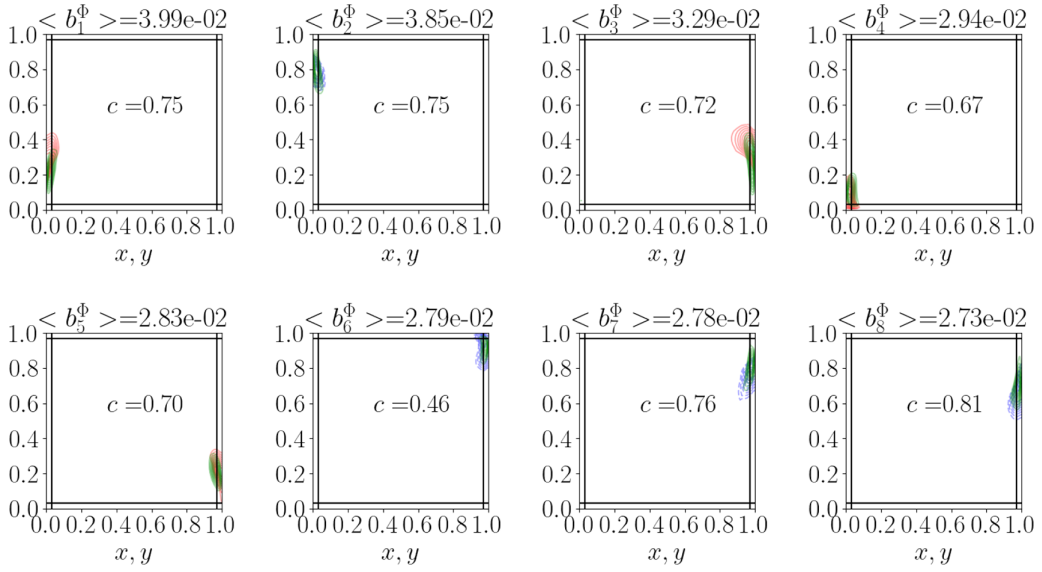


FIG. 11. Dominant heat flux motifs $\psi_n^\Phi$ (green lines) ordered by prevalence and associated temperature motifs $\psi_n^\theta$ (blue for negative and red for positive fluctuations) at $\mathrm{Ra} = 10^7$. Contour levels go from $0.2\psi_n^{\max}$ to $0.9\psi_n^{\max}$ with increments of $0.1\psi_n^{\max}$. $c$ is the maximum correlation coefficient between the heat flux and temperature motif weights.

a vertical distance of about 0.25 from the horizontal walls. Although the positions and sizes of the four identified motifs may slightly vary from one to the other, their features are generally similar and a characteristic motif can be obtained from taking the average over all four motifs. Figure 12 (right) represents this characteristic motif for the various Rayleigh numbers. We can see that the dominant motifs are always located mostly within the boundary layer, with a maximum at a height of about 0.25. Their characteristic width $l_y$ was found to decrease as $\mathrm{Ra}^{-0.23\pm0.04}$, which matches the scaling of the boundary layer thickness.

### *2. Temporal dynamics*

The evolution of the snapshot-motif distribution, or motif weight, is represented in Fig. 13 for $\mathrm{Ra} = 10^7$. We can see that the behavior of the motif weight depends on the sign of the global

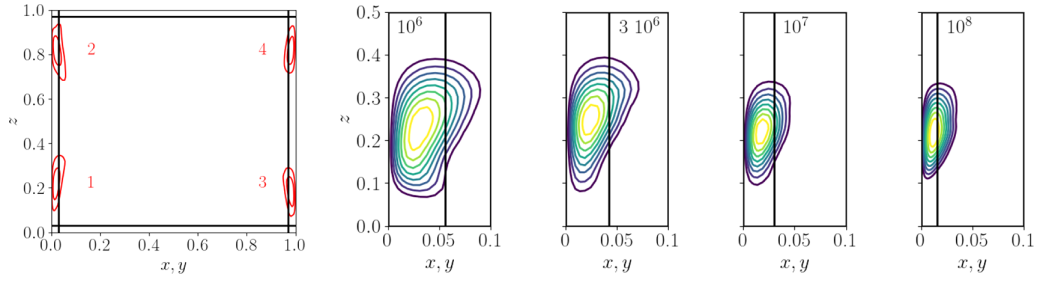


FIG. 12. Left: Dominant heat flux motifs $\psi_n^\Phi$ at Rayleigh number $\mathrm{Ra} = 10^7$ for $N_T = 100$. The contour lines correspond to $0.1\psi^{\max}$ and $0.3\psi^{\max}$. The motif labels correspond to those of Fig. 13. Right: Characteristic dominant motif at different Rayleigh numbers, $N_T = 100$. Isocontours of $\psi_1$ at $[0.2, 0.3, \ldots, 0.9]\psi_1^{\max}$. The black lines correspond to the boundary layer thickness.

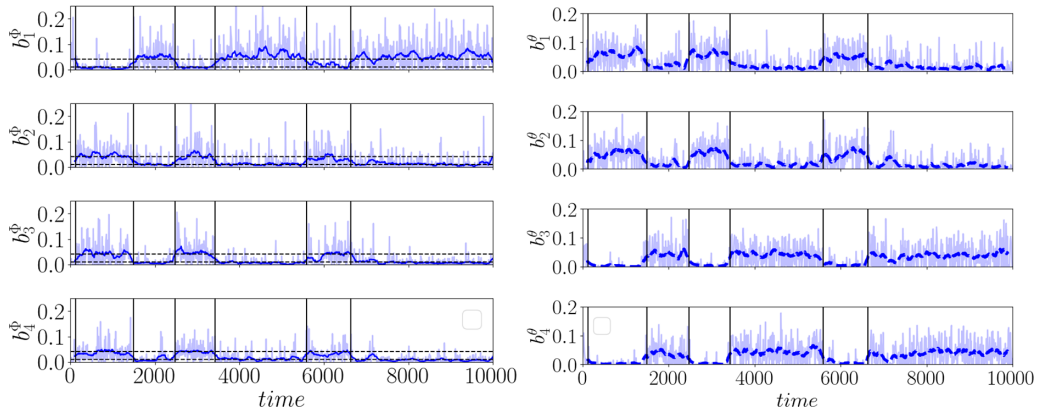


FIG. 13. Evolution of the snapshot-motif distributions $b_n^\Phi$ for the four dominant heat flux motifs (see Fig. 12 for labels) at Ra $= 10^7$ and for $N_T = 100$. Left: plane $x = 0.5$. Right: plane $y = 0.5$. The thick line corresponds to a moving average over 200 convective units (four recirculation times $T_c$). The horizontal dashed lines correspond to the values $b_- = 0.017$ and $b_+ = 0.035$. The vertical lines correspond to the changes in angular momentum.

momentum represented in Fig. 1. When a moving average of $T_f = 200$ time units, corresponding to four recirculation times $T_c$, was applied, two quasistationary states $b_+$ and $b_-$ could be identified in each plane (they are materialized by the dashed horizontal black lines indicated in Fig. 13). The two states appear to correspond to the sign of the angular momentum component, i.e., the orientation of the large-scale circulation $I$. Streamlines of the flow conditionally averaged on the higher weight value of $b_1^\Phi$ are represented in Fig. 14 (left). They indicate that for the higher characteristic value of the weight, $b_+$, the motif is associated with the large-scale circulation while it is associated with the corner vortex on the opposite side for the lower weight value, $b_-$, as summarized in Fig. 14 (right).

This indicates that information about the large-scale reorientation can be extracted from local measurements. Two states, $I_+$ and $I_-$, respectively corresponding to the large-scale circulation and

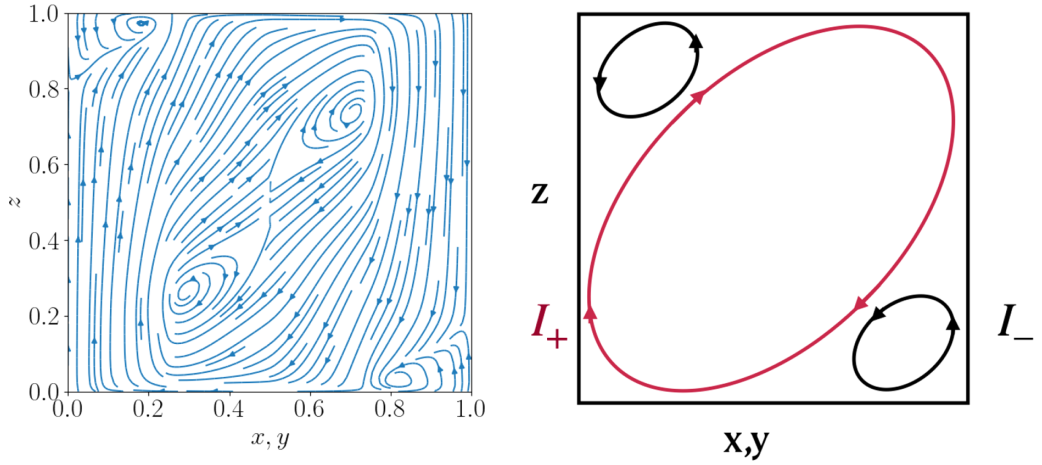


FIG. 14. Left: Streamlines of the flow conditionally averaged on the high weight value of $b_1^\Phi$. Right: Schematics of the cell organization in the vertical midplane: the large-scale circulation (in red) corresponds to the $I_+$ state while the corner structure (in black) corresponds to the $I_-$ state.

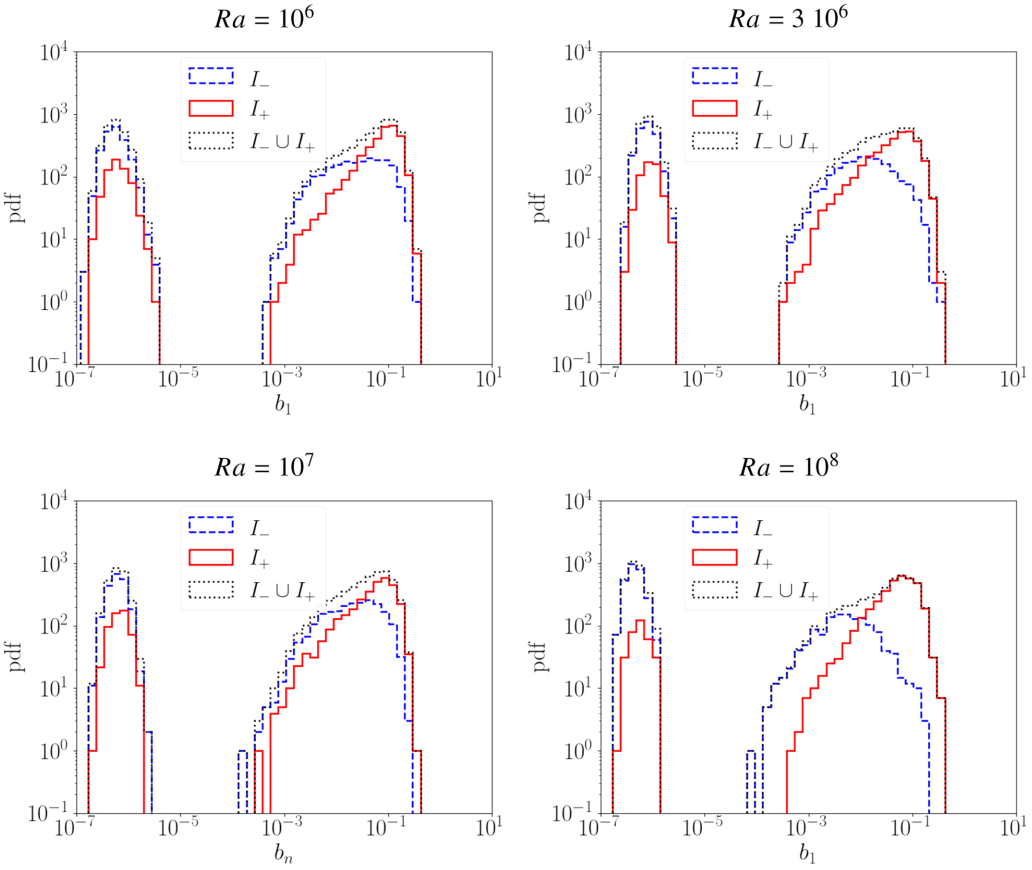


FIG. 15. Distribution of the dominant motif weight $b_1^\Phi$ for different Rayleigh numbers and $N_T = 100$.

corner vortex can be defined from the weight of the dominant motif $b_1^\Phi$ using

$$I_+ = \left\{m \middle| \left\langle b_1^\Phi(t_m)\right\rangle_{T_f} > \left\langle b_1^\Phi\right\rangle\right\} \quad \text{and} \quad I_- = \left\{m \middle| \left\langle b_1^\Phi(t_m)\right\rangle_{T_f} < \left\langle b_1^\Phi\right\rangle\right\}, \tag{19}$$

where $\langle\cdot\rangle_{T_f}$ represents the moving average over $T_f$. The average weights conditioned on $I_+$ and $I_-$ are respectively $b_+$ and $b_-$.

Figure 15 displays the histogram of the weight of the dominant motif $b_1^\Phi$ (motifs 2–4 displayed similar features). At all Rayleigh numbers, the total distribution is characterized by two distinct lobes, which correspond to the absence and the presence of the motif in the snapshot. The relative importance of the lobes therefore provides an indirect measure of the motif intermittency, which can be related to plume emission. The ratio of motif presence to motif absence was about 0.5–0.6 in the range of Rayleigh numbers considered and no significant variation was observed with the Rayleigh number.

However, further insights can be obtained by examining the respective contributions of the $I_+$ and $I_-$ states to the distribution of $b_1^\Phi$, which are also represented in Fig. 15. For all Rayleigh numbers, $I_+$ states contribute more to the higher-value lobe than $I_-$ states, while $I_-$ contributes more to the lower-value lobe. This shows that the rate of buoyancy production is less intense in the corner rolls than in the large-scale circulation, or equivalently that plumes are emitted at a lower frequency in the corner rolls than in the large-scale circulation. Moreover, the relative contributions of the $I_+$ and the $I_-$ states vary nonmonotonically with the Rayleigh number. In the higher-value lobe, the relative contribution of $I_-$ appears to increase relatively to $I_+$ with more high values of $I_-$ at Ra $= 3\times10^6$,

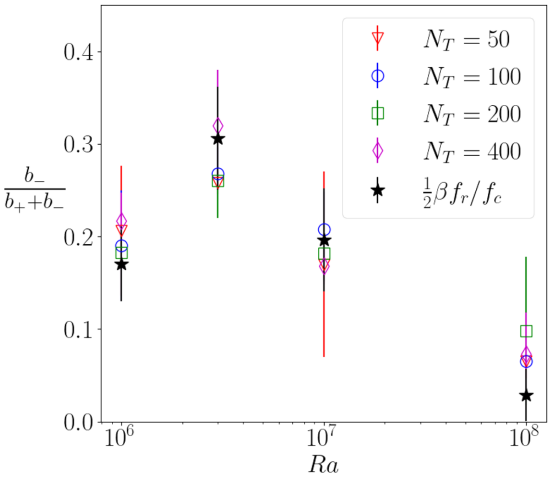


FIG. 16. Probability $p(T_+ > T_-)$ (see text) and comparison with ratio of reorientation to recirculation timescale at different Rayleigh numbers. The rescaling factor is $\beta = 5.6$.

while $I_-$ represents more low values at Ra $= 10^8$. In the lower-value lobe, the contribution of $I_+$ is least at Ra $= 3\times10^6$ and largest at Ra $= 10^8$. These observations suggest that both the intensity of the large-scale circulation and that of the corner roll appear to change with the Rayleigh number, in agreement with the findings of Vishnu *et al.* [40].

#### *3. A model for the reorientation timescale*

A simple model can be made to link these observations with the dynamics of reorientations. The conditionally averaged weight of the dominant motif in the region close to the wall, $b_\pm$, represents the rate of buoyancy production, which can be linked to the emission rate of plumes and can be modeled as a Poisson point process. This means that the time separating two plume ejections, $T_\pm$, follows an exponential distribution with mean $1/b_\pm$, where $+$ and $-$ respectively characterize the large-scale circulation ($I_+$) and the corner vortex ($I_-$) states. $b_\pm$ therefore represents the parameter of the exponential distribution. A reorientation can be associated with the event where the corner vortex becomes stronger than the large-scale circulation state; i.e., the time separating two emissions in the corner vortex state becomes smaller than that separating two emissions in the large-scale circulation state. This event can occur independently in either one of the two horizontal directions $x$ or $y$.

One can show that the probability $p$ that this event occurs at any given time is given by

$$p = p(T_- > T_+) = \frac{b_-}{b_+ + b_-}. \tag{20}$$

Owing to the memoryless nature of the exponential distribution, this holds for the time separating an arbitrary number of emissions, in particular over a characteristic time $T_s$ sufficiently long to reverse the circulation in that direction. $T_s$ should be on the order of the recirculation time $T_c$ so that we have $T_s = \beta T_c$ with $\beta = O(1)$. If $f_c$ is the recirculation frequency, one would then expect the frequency between reorientations, $f_r$, to depend on $p$ and $f_c$ following

$$f_r = 2p\beta^{-1} f_c, \tag{21}$$

where the factor 2 comes from the fact that a reorientation can occur in each direction. Figure 16 (right) compares for different Rayleigh numbers the probability $p$ with the ratio of the frequency between reorientations and the recirculation frequency estimated in Ref. [26]. We see that a very good agreement is obtained between the variations of the average reorientation rate and the measure of the relative intensity of the large-scale circulation and corner vortices. We note that the largest

TABLE II. Average correlation coefficient between the original and the reconstructed field for the temperature, kinetic energy, and heat flux.

| $\langle C(q, q_R)\rangle$ | $N_T$ | $\mathrm{Ra} = 10^6$ | $\mathrm{Ra} = 3\times10^6$ | $\mathrm{Ra} = 10^7$ | $\mathrm{Ra} = 10^8$ |
|---|---|---|---|---|---|
| $\langle C(\theta, \theta_R)\rangle$ | 100 | 0.90 | 0.86 | 0.84 | 0.66 |
| $\langle C(\theta, \theta_R)\rangle$ | 400 | 0.94 | 0.92 | 0.90 | 0.78 |
| $\langle C(k, k_R)\rangle$ | 100 | 0.91 | 0.88 | 0.85 | 0.78 |
| $\langle C(k, k_R)\rangle$ | 400 | 0.94 | 0.92 | 0.89 | 0.82 |
| $\langle C(\Phi, \Phi_R)\rangle$ | 100 | 0.96 | 0.93 | 0.89 | 0.84 |
| $\langle C(\Phi, \Phi_R)\rangle$ | 400 | 0.98 | 0.96 | 0.95 | 0.89 |

discrepancy is observed for the highest Rayleigh number, for which the reorientation rate is very low and therefore cannot be determined with good precision from the direct numerical simulations. The value of $\beta$ used in the figure was determined empirically and was found to be 5.6, which makes $T_s$ close to the filtering timescale $T_f = 4T_c$. This suggests that an estimate for the reorientation rate can be obtained by comparing directly the average weight of the motif associated with the large-scale circulation with that of its counterpart in the corner structure. This could be of particular interest in cases where the observation time is smaller than the expected reorientation time, a situation that is often encountered in (but not limited to) numerical simulations at higher Rayleigh numbers, as the simulation cost increases and the reorientation frequency decreases.

## VII. TEMPERATURE AND VELOCITY MOTIFS

In this section we try to understand the physics associated with the lower reorientation rate observed as the Rayleigh number increases. For this we turn to temperature and velocity fluctuations, to which we independently applied LDA. Although these are not intermittent quantities, and therefore might not be considered *a priori* appropriate for LDA application, Table II shows that the temperature and kinetic energy fields are relatively well reconstructed.

### A. Temperature fluctuations

Figure 17 shows the temperature motifs at three different Rayleigh numbers, along with the variance of the fluctuations, for $N_T = 100$. As mentioned above, some symmetry is expected but not perfectly enforced, due to the statistical character of the method. As for heat flux motifs there is a clear difference between the boundary layers and the bulk, as well as a strong decrease of motifs in the central part of the cell at $\mathrm{Ra} = 10^8$. We can see that temperature fluctuations are also important close to the horizontal walls. The bottom row of Fig. 17 shows a close-up of the lower part of the cell. The maximum of the motif spatial distribution is located at the edge of the boundary layer. The height of the motifs scales with the boundary layer height in the center of the cell, with negative motifs shorter and wider than positive ones in the bottom layer. Analogous observations can be made for the top wall, by swapping the role of cold and hot fluctuations.

Figure 18 represents the first four dominant motifs for the temperature at $\mathrm{Ra} = 10^6$ (similar observations can be made at $\mathrm{Ra} = 3\times10^6$). Although the most likely heat flux motifs corresponded to hot plumes near the bottom wall and cold plumes near the top wall, this is not the case for the temperature motifs. For the two lower Rayleigh numbers, temperature motifs are as likely to be found near the bottom wall as near the top wall. However, at $\mathrm{Ra} = 10^7$, Fig. 19 shows that the most likely temperature motifs correspond to hot fluctuations along the bottom sidewalls and cold near the top sidewall, corresponding to late-stage plumes arriving at the opposite wall.

Figure 20 shows the evolution of the temperature motif weights $b_n^\theta$ on both planes along with their filtered representation $\langle b_n^\theta\rangle_{T_f}$. As observed for the heat flux (Fig. 13), the importance of the weights

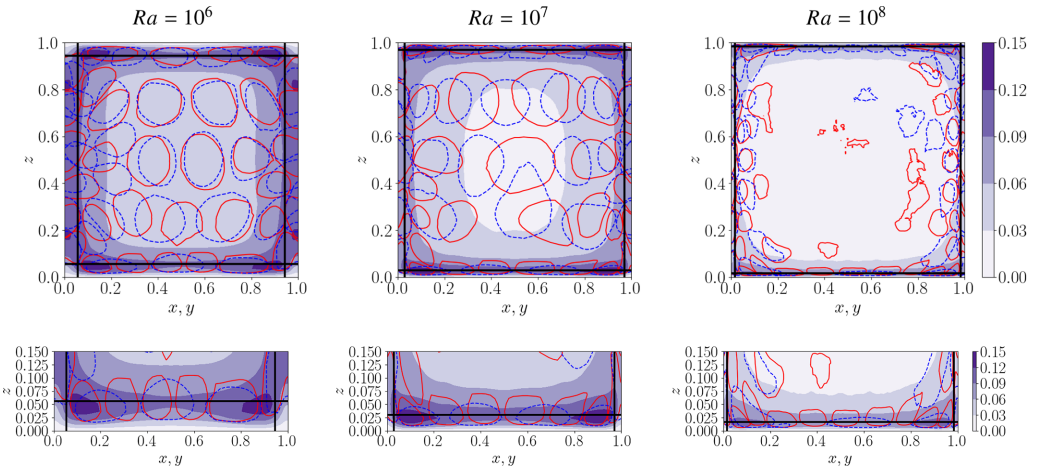


FIG. 17. Top: Distribution of temperature motifs $\psi_n^\theta$ in the cell midplane at different Rayleigh numbers; The motifs are materialized by a black line corresponding to a probability contour of $0.606\psi_n^{\max}$. Contours of the time-averaged variance are represented in the background. Bottom: Magnification of the bottom part of the cell.

depends on the orientation of the large-scale circulation $I$. Similar evolutions were observed at the lower Rayleigh numbers (not shown).

Strong differences can be observed when comparing Figs. 19 and 21. At Ra $= 10^8$, the most likely temperature motifs are no longer located within the vertical boundary layers, but extend from the corner of the cell along the horizontal walls. The first eight dominant structures consist of two types of corner motifs: large, predominantly horizontal ones, and small, vertical ones located within the boundary layers. Motifs near the top (bottom) wall are hot (cold) and therefore correspond to late-stage plumes. This is confirmed by the evolution of the motif weights shown in Fig. 22 for the plane $x = 0.5$. These motifs correspond to hot fluid being brought from the bottom layer by the large-scale circulation next to the top wall and into the corner structure, thus decreasing buoyancy effects there. These observations are consistent with the reduction in intensity of the corner roll and the significant decrease in the reorientation rate observed at this Rayleigh number. We note that although the small vertical temperature motifs are similar to heat flux motifs 4 and 6 identified in Fig. 11 at Ra $= 10^7$, they represent fluctuations of the opposite sign, and they are well correlated (or anticorrelated) with the orientation $I$ of the large-scale circulation. This confirms the dominance of the impinging plumes in the corners of the cell.

### B. Kinetic energy

More details about the structure of the large-scale circulation can be obtained by examining kinetic energy motifs. Figure 23 shows the spatial distribution of the velocity motifs for the different

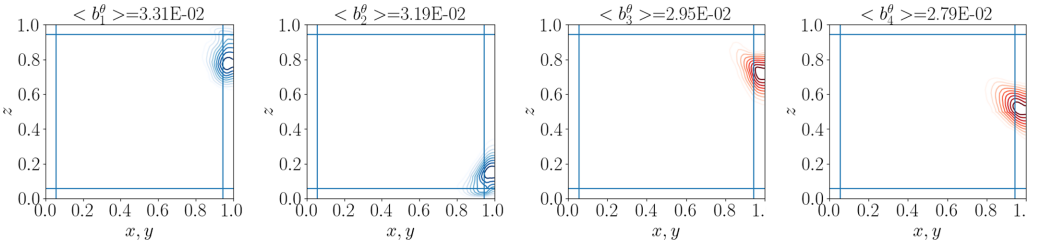


FIG. 18. First four dominant temperature motifs $\psi_n^\theta$ at Ra $= 10^6$.

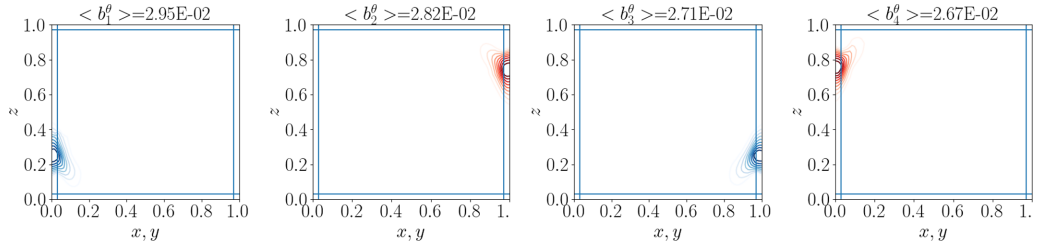


FIG. 19. First four dominant temperature motifs $\psi_n^\theta$ at Ra $= 10^7$.

Rayleigh numbers and $N_T = 100$. The spatial distribution of the time-averaged kinetic energy is also represented on the same plot. The size of the core (low-velocity region) appears to increase with the Rayleigh number. The size of the motifs did not appear to change significantly with the Rayleigh number, except for horizontal corner structures that seem to scale with the boundary layer thickness. The kinetic energy motifs have elongated shapes along the walls, with a significantly higher extent along the horizontal walls, which shows the importance of entrainment in the horizontal boundary layers, in particular in the middle of the cell. It is lowest at Ra $= 3\times10^6$ and highest at Ra $= 10^8$, which varies like the time between reorientations, $T_r$. The question is whether this reinforcement of the large-scale circulation can be associated with characteristic temperature fluctuations.

In Figs. 24–26 the 16 most prevalent kinetic energy motifs are represented at Rayleigh numbers $10^6$, $10^7$, and $10^8$ (the case Ra $= 3\times10^6$, not shown, was found generally similar to $10^6$ and $10^7$). The motifs were organized according to the location of their maximum: within the horizontal or vertical boundary layers, which we will refer to as respectively HBL or VBL motifs, at the corners of the horizontal and the vertical boundary layer (CBL motifs), and outside the boundary layers in the horizontal or vertical entrainment zones, which were termed HEZ or VEZ motifs. The different locations are shown in the top right illustration of Fig. 24. For each category the motifs are ordered according to their prevalence, indicated at the top of each plot. Generally speaking, the prevalence of the motifs increased with the Rayleigh number, which is consistent with a strengthening of the large-scale circulation.

For each kinetic energy motif $n$ (represented with green lines), we determined the temperature motif $j$ (represented with blue or red lines, depending on its sign) for which the correlation coefficient $C(b_n^k, b_{n'}^\theta)$ is maximal. The maximal value $c$ and the temperature motif are represented on each

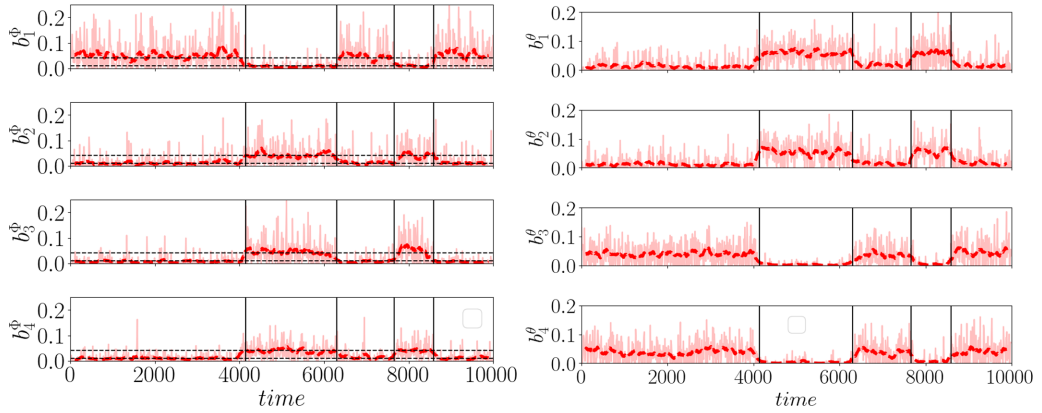


FIG. 20. First four dominant temperature motif weights $b_n^\theta$ at Ra $= 10^7$. Left: Plane $x = 0.5H$. Right: Plane $y = 0.5H$. The thick line corresponds to a moving average over 200 convective units (four recirculation times $T_c$). The vertical lines correspond to the changes in angular momentum.

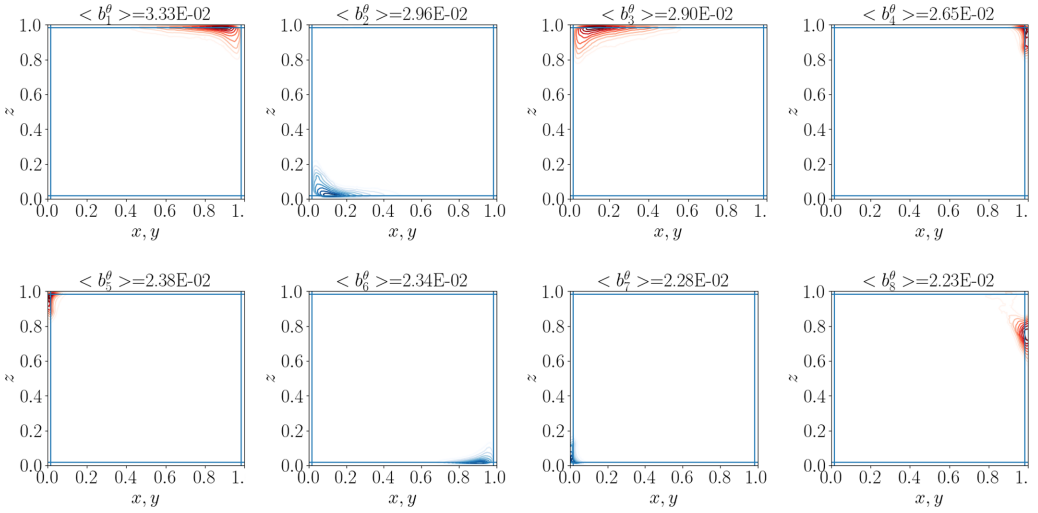


FIG. 21. First eight dominant temperature motifs $\psi_n^\theta$ at Ra $= 10^8$.

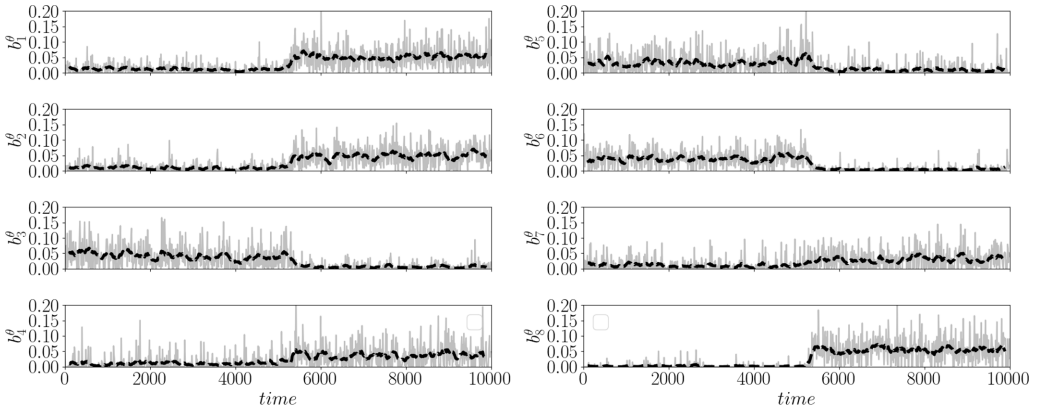


FIG. 22. First eight temperature motif weights $b_n^\theta$ at Ra $= 10^8$ in the plane $x = 0.5$.

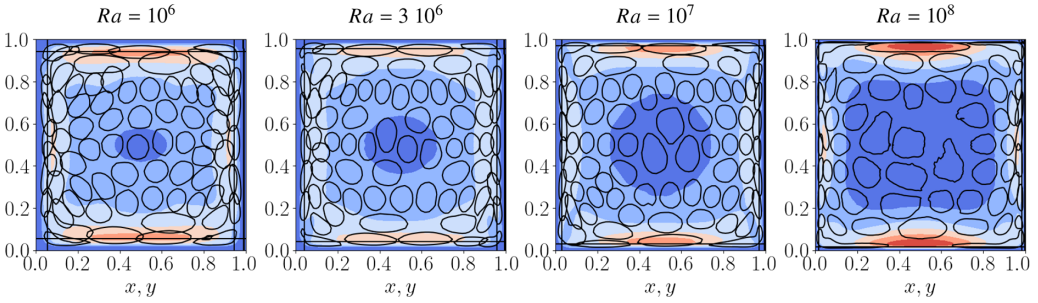


FIG. 23. Spatial distribution of kinetic energy motifs in the cell midplane at different Rayleigh numbers. The motifs are materialized by a black line corresponding to a probability contour of $0.606\psi_n^{\max}$. Contours of the time-averaged kinetic energy are represented in the background.

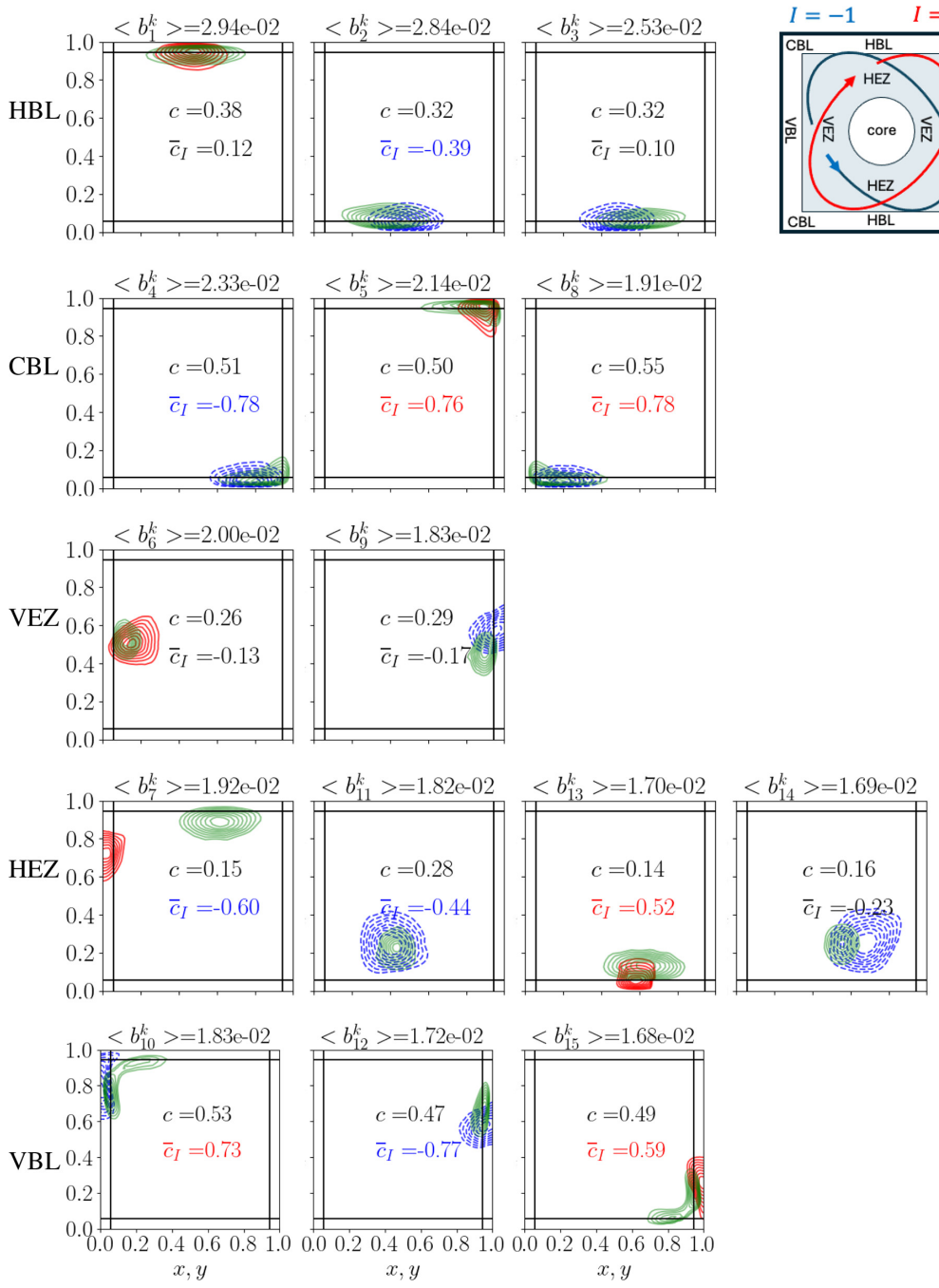


FIG. 24. Dominant kinetic energy motifs $\psi_n^k$ at Ra $= 10^6$ (green lines) ordered by prevalence and location as indicated at top right. Temperature motifs $\psi_n^\theta$ with the highest correlation coefficient $c$ are shown in blue (red) for negative (positive) fluctuations. Motif contour levels range from $0.2\psi_n^{\max}$ to $0.9\psi_n^{\max}$ with increments of $0.1\psi_n^{\max}$. $\overline{c}_I$ is the correlation coefficient between the heat flux motif weight and the LSC indicator $I$. Values of $\overline{c}_I$ larger than 0.3 (lower than $-0.3$) are represented in red (blue).

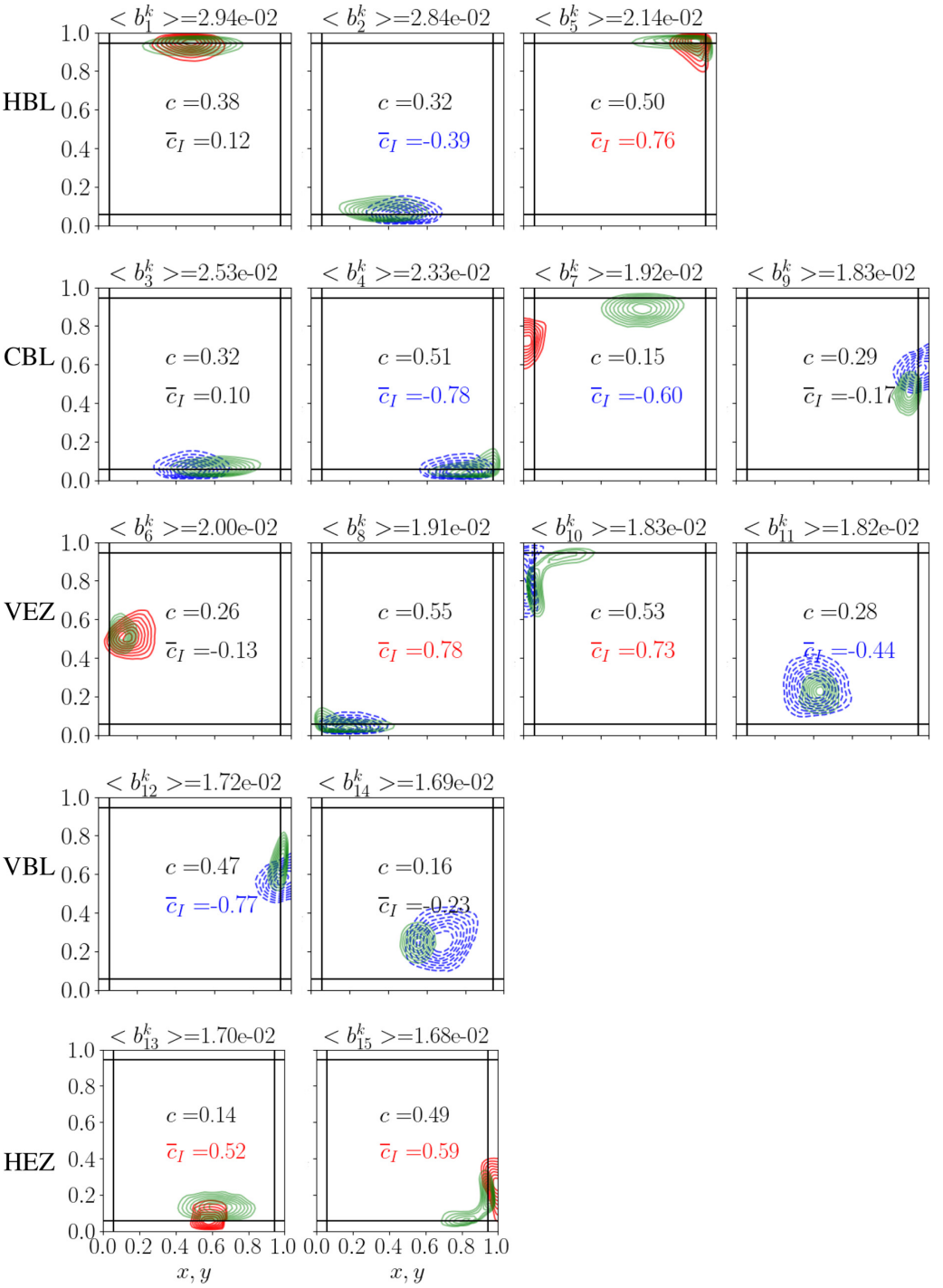


FIG. 25. Comparison between kinetic energy and temperature motifs at Ra $= 10^7$. See legend of Fig. 24.

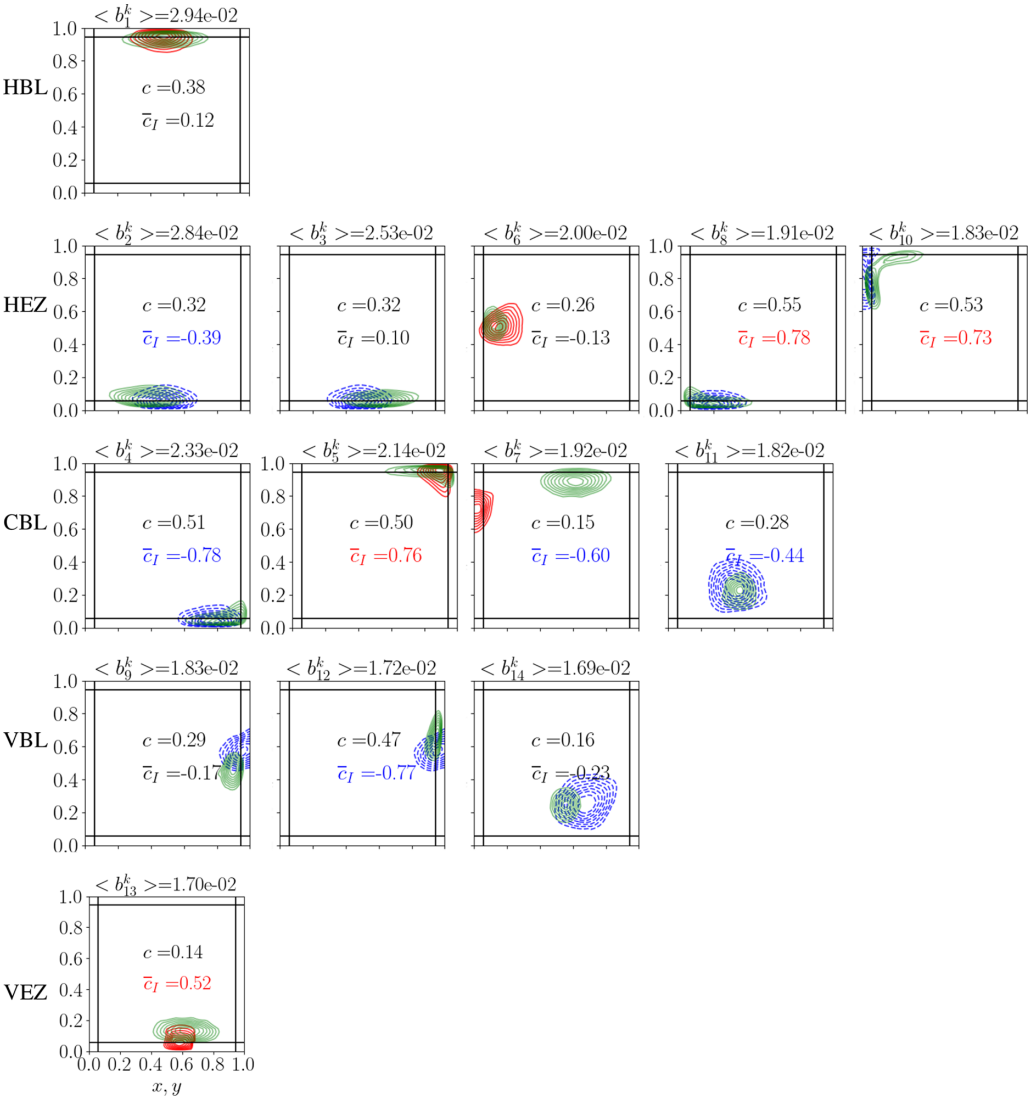


FIG. 26. Comparison between kinetic energy and temperature motifs at Ra $= 10^8$. See legend of Fig. 24.

plot, except in two cases corresponding to HBL motifs, for which the associated temperature motif had a very low prevalence and was considered to be irrelevant. In almost all cases, the kinetic energy and temperature motifs are located close to each other in space. Although the correlation coefficients are typically lower than those between the flux and temperature motifs represented in Fig. 11, several are high enough to associate kinetic energy patterns with specific temperature fluctuations. We also represented on each plot the correlation coefficient $\overline{c}_I$, defined as $\overline{c}_I = C(\langle b_n^k \rangle_{T_f}, I)$, where $\langle b_n^k \rangle_{T_f}$ is the low-pass-filtered kinetic energy motif weight (using $T_f$) and $I$ is the large-scale circulation indicator defined in Eq. (19) (see also Fig. 24 top right). High positive (negative) values of $\overline{c}_I$ are indicated in red (blue) for each motif, and show that the motif can be associated with a specific orientation of the large-scale circulation.

*a. Horizontal boundary layers and corners.* In all cases, the most frequent motifs consist of centered motifs close to the edge of the HBLs. Evidence of weak correlation (0.3) for some motifs

suggested possible association with impinging plumes; however, generally low values of $|\bar{c}_I|$ suggest that the weights of the motifs do not depend on the orientation of the large-scale circulation. In contrast, high values of $c$ and $|\bar{c}_I|$ were found for corner (CBL) motifs, that were best correlated with impinging plumes. Corner motifs have a relatively high prevalence, which shows that impinging plumes make a significant contribution to the horizontal wind at the edge of the boundary layer. The correlation coefficient $\bar{c}_I$ increased in absolute value with the Rayleigh number, and was larger than 0.9 at $\mathrm{Ra} = 10^8$. In contrast, the maximum correlation coefficient $c$ tended to decrease (but remained significant) at $\mathrm{Ra} = 10^8$.

*b. Vertical entrainment zone.* The next prevalent category of motifs at $\mathrm{Ra} = 10^6$ and $\mathrm{Ra} = 10^7$ consisted of motifs in the vertical entrainment zone (VEZ motifs). They were generally weakly correlated with temperature motifs of a slightly larger size ($c \approx 0.2$–$0.3$) and were still less correlated with the orientation of the large-scale circulation ($\bar{c}_I$ close to zero), which is consistent with their mid-height location. Two of the motifs at $\mathrm{Ra} = 10^7$ (third and fourth motifs) were located closer to a horizontal wall and showed a stronger correlation with $I$. They were found to be correlated with "upstream" temperature fluctuations originating from the opposite wall (arriving plumes). At $\mathrm{Ra} = 10^8$, only one VEZ motif, with a lower prevalence (compared with the other motifs), was identified. It also corresponded to an arriving plume and was strongly correlated with the orientation of the large-scale circulation.

*c. Vertical boundary layers.* High values of $|\bar{c}_I|$ were also observed for motifs within the VBLs, as well as significant values of $c$. The corresponding temperature motifs were also located within the vertical boundary layers and consisted of hot (cold) temperature fluctuations close to the bottom (top plate), suggesting that they correspond to plumes in the early formation stage (leaving plumes).

*d. Horizontal entrainment zone.* At $\mathrm{Ra} = 10^6$ and $\mathrm{Ra} = 10^7$, the last category of motifs consisted of motifs in the horizontal entrainment zone (HEZ). At the lowest Rayleigh number $\mathrm{Ra} = 10^6$, two of the HEZ motifs (second and fourth motifs in the last row in Fig. 24) have a predominantly vertical shape and are associated with large temperature motifs originating from the opposite (here, top) wall. They are therefore likely to represent coalescing plumes drifting towards the center of the cell as they reach the opposite wall. In contrast, all other HEZ motifs at all Rayleigh numbers have a horizontal shape and are associated with smaller temperature motifs originating from the closest wall. They are very well correlated with the orientation of the large-scale circulation. Significant changes were observed at $\mathrm{Ra} = 10^8$, with a much larger number of HEZ motifs and a noticeable increase in their prevalence—the prevalence of the dominant HEZ motif is twice as large at $\mathrm{Ra} = 10^8$ than at $\mathrm{Ra} = 10^7$.

To sum up, a significant difference is observed between $10^7$ and $10^8$. At the highest Rayleigh number, the large-scale circulation is largely reinforced in the horizontal direction due to the formation of new plumes, while stronger impinging plumes remain confined to the corner boundary layers.

## VIII. CONCLUSION

We have applied an analysis technique, latent Dirichlet allocation, to characterize the spatiotemporal organization of fluctuations in Rayleigh-Bénard convection. The method is based on the inference of probabilistic latent factors, spatially localized motifs, from a collection of instantaneous fields. It provides a local yet compact description of the flow in terms of quantitative indicators such as the (spatial) size and the (temporal) weight of the motifs. The technique was applied to the vertical midplane of a Rayleigh-Bénard cubic cell in a range of Rayleigh numbers in $[10^6, 10^8]$. The method was found to be robust with respect to the user-defined parameters. When applied to the heat flux, it was found to provide good reconstructions of the snapshots and was able to generate new data sets that reproduced key statistics of the original one.

For all Rayleigh numbers, dominant heat flux motifs consisted of elongated vertical structures located mostly within the vertical boundary layer, at a height of a quarter of the cell. The width of these motifs scaled with the boundary layer thickness. These motifs were found to be very

well correlated with temperature motifs corresponding to plumes in their early formation stage (leaving plumes). The motif weights were found to depend on the large-scale organization of the flow: two states could be identified, one corresponding to the large-scale circulation and one to a corner roll structure. The two states were characterized by different average weights which varied nonmonotonically with the Rayleigh number. A simple model was able to relate the weights of the dominant heat flux motif associated with the two states with the average reorientation rate of the large-scale circulation in the cell. This suggests that the model could be used as a predictor of this rate in cases where few or even no reorientations are observed.

Additional insight about the flow physics was obtained by examining dominant motifs for the temperature and the kinetic energy. While dominant heat flux motifs seemed to be associated with early-stage (leaving) plumes, dominant temperature motifs were associated with later-stage (arriving) plumes. In contrast with the lower Rayleigh numbers, dominant temperature motifs at $\mathrm{Ra} = 10^8$ were no longer within the vertical boundary layers, but consisted of plumes impinging onto the corners of the horizontal boundary layers, which led to a reduction of temperature gradients within the corner structure and a decrease in its potential energy. This is consistent with the significant drop in the large-scale reorientation rate observed at this Rayleigh number. LDA analysis of the kinetic energy showed that corner impinging plumes contributed to the kinetic energy of both the corner structure and the large-scale circulation. The reduction of the reorientation rate at $\mathrm{Ra} = 10^8$ was also associated with a reinforcement of the horizontal wind in the central part of the cell due to the formation and entrainment of new plumes. Changes in the dynamics of the large-scale circulation could thus be directly connected with local modifications of its structure. The LDA model therefore appears as a promising statistical tool that can help track subtle transitions in the spatiotemporal organization of turbulent flows. An interesting direction of investigation, suggested by one of the anonymous reviewers, would be to explore the connection between the LDA representation and structure function analysis, which could provide insight into local energy transfer mechanisms at different scales.

## ACKNOWLEDGMENTS

This work was granted access to the HPC resources of IDRIS under the allocation 2023-AD012A62062R1 made by GENCI. We thank Anouar Soufiani and Philippe Rivière for helpful discussions about the manuscript. We are grateful to Jean-Michel Dupays, Rémy Dubois, and Camille Parisel for technical support and useful discussions. We are also thankful to the anonymous reviewers for their insightful suggestions.

[1] S. Grossmann and D. Lohse, Scaling in thermal convection: A unifying theory, J. Fluid Mech. **407**, 27 (2000).

[2] S. Grossmann and D. Lohse, Fluctuations in turbulent Rayleigh-Bénard convection: The role of plumes, Phys. Fluids **16**, 4462 (2004).

[3] H.-D. Xi, S. Lam, and K. Q. Xia, From laminar plumes to organized flows: The onset of large-scale circulation in turbulent thermal convection, J. Fluid Mech. **503**, 47 (2004).

[4] B. Castaing, G. Gunaratne, F. Heslot, L. Kadanoff, A. Libchaber, S. Thomae, X.-Z. Wu, S. Zaleski, and G. Zanetti, Scaling of hard thermal turbulence in Rayleigh-Bénard convection, J. Fluid Mech. **204**, 1 (1989).

[5] Y. Wang, Y. Wei, P. Tong, and X. He, Collective effect of thermal plumes on temperature fluctuations in a closed Rayleigh-Bénard convection cell, J. Fluid Mech. **934**, A13 (2022).

[6] J. Bosbach, S. Weiss, and G. Ahlers, Plume fragmentation by bulk interactions in turbulent Rayleigh-Bénard convection, Phys. Rev. Lett. **108**, 054501 (2012).

[7] X.-D. Shang, X.-L. Qiu, P. Tong, and K.-Q. Xia, Measured local heat transport in turbulent Rayleigh-Bénard convection, Phys. Rev. Lett. **90**, 074501 (2003).

[8] Q. Zhou, C. Sun, and K.-Q. Xia, Morphological evolution of thermal plumes in turbulent Rayleigh-Bénard convection, Phys. Rev. Lett. **98**, 074501 (2007).
[9] O. Shishkina and C. Wagner, Analysis of sheet-like thermal plumes in turbulent Rayleigh-Bénard convection, J. Fluid Mech. **599**, 383 (2008).
[10] M. S. Emran and J. Schumacher, Conditional statistics of thermal dissipation rate in turbulent Rayleigh-Bénard convection, Eur. Phys. J. E **35**, 108 (2012).
[11] A. Belmonte and A. Libchaber, Thermal signature of plumes in turbulent convection: The skewness of the derivative, Phys. Rev. E **53**, 4893 (1996).
[12] Q. Zhou and K.-Q. Xia, Physical and geometrical properties of thermal plumes in turbulent Rayleigh-Bénard convection, New J. Phys. **12**, 075006 (2010).
[13] E. S. Ching, H. Guo, X. Shang, P. Tong, and K.-Q. Xia, Extraction of plumes in turbulent thermal convection, Phys. Rev. Lett. **93**, 124501 (2004).
[14] S.-D. Huang, M. Kaczorowski, R. Ni, and K.-Q. Xia, Confinement-induced heat-transport enhancement in turbulent thermal convection, Phys. Rev. Lett. **111**, 104501 (2013).
[15] E. P. van der Poel, R. Verzicco, S. Grossmann, and D. Lohse, Plume emission statistics in turbulent Rayleigh-Bénard convection, J. Fluid Mech. **772**, 5 (2015).
[16] S.-Q. Zhou, Y.-C. Xie, C. Sun, and K.-Q. Xia, Statistical characterization of thermal plumes in turbulent thermal convection, Phys. Rev. Fluids **1**, 054301 (2016).
[17] V. T. Vishnu, A. K. De, and P. K. Mishra, Statistics of thermal plumes and dissipation rates in turbulent Rayleigh-Bénard convection in a cubic cell, Int. J. Heat Mass Transfer **182**, 121995 (2022).
[18] P. P. Shevkar, R. Vishnu, S. K. Mohanan, V. Koothur, M. Mathur, and B. A. Puthenveettil, On separating plumes from boundary layers in turbulent convection, J. Fluid Mech. **941**, A5 (2022).
[19] F. Chillà and J. Schumacher, New perspectives in turbulent Rayleigh-Bénard convection, Eur. Phys. J. E **35**, 58 (2012).
[20] J. Lumley, The structure of inhomogeneous turbulent flows, in *Atmospheric Turbulence and Radio Wave Propagation*, edited by A. Iaglom and V. Tatarski (Nauka, Moscow, 1967), pp. 221–227.
[21] J. Bailon-Cuba, M. S. Emran, and J. Schumacher, Aspect ratio dependence of heat transfer and large-scale flow in turbulent convection, J. Fluid Mech. **655**, 152 (2010).
[22] N. Foroozani, J. J. Niemela, V. Armenio, and K. R. Sreenivasan, Reorientations of the large-scale flow in turbulent convection in a cube, Phys. Rev. E **95**, 033107 (2017).
[23] B. Podvin and A. Sergent, A large-scale investigation of wind reversal in a square Rayleigh-Bénard cell, J. Fluid Mech. **766**, 172 (2015).
[24] B. Podvin and A. Sergent, Precursor for wind reversal in a square Rayleigh-Bénard cell, Phys. Rev. E **95**, 013112 (2017).
[25] L. Soucasse, B. Podvin, P. Rivière, and A. Soufiani, Proper orthogonal decomposition analysis and modelling of large-scale flow reorientations in a cubic Rayleigh-Bénard cell, J. Fluid Mech. **881**, 23 (2019).
[26] L. Soucasse, B. Podvin, P. Rivière, and A. Soufiani, Low-order models for predicting radiative transfer effects on Rayleigh-Bénard convection in a cubic cell at different rayleigh numbers, J. Fluid Mech. **917**, A5 (2021).
[27] P. J. Olesen, A. Hodžić, S. J. Andersen, N. N. Sørensen, and C. M. Velte, Dissipation-optimized proper orthogonal decomposition, Phys. Fluids **35**, 015131 (2023).
[28] P. J. Olesen, L. Soucasse, B. Podvin, and C. M. Velte, Dissipation-based proper orthogonal decomposition of turbulent Rayleigh-Bénard convection flow, Phys. Fluids **36**, 035109 (2024).
[29] M. Frihat, B. Podvin, L. Mathelin, Y. Fraigneau, and F. Yvon, Coherent structure identification in turbulent channel flow using latent dirichlet allocation, J. Fluid Mech. **920**, A27 (2021).
[30] T. L. Griffiths and M. Steyvers, A probabilistic approach to semantic representation, in *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, edited by W. D. Gray and C. Schunn, Sponsors: Cognitive Science Society (George Mason University Fairfax, Virginia, 2002).
[31] D. M. Blei, A. Y. Ng, M. I. Jordan, and J. Lafferty, Latent Dirichlet allocation, J. Mach. Learn. Res. **3**, 993 (2003).

[32] L. Fery, B. Dubrulle, B. Podvin, F. Pons, and D. Faranda, Learning a weather dictionary of atmospheric patterns using latent Dirichlet allocation, Geophys. Res. Lett. **49**, e2021GL096184 (2022).
[33] S. Xin and P. Le Quéré, An extended Chebyshev pseudo-spectral benchmark for the 8:1 differentially heated cavity, Int. J. Numer. Methods Fluids **40**, 981 (2002).
[34] S. Xin, J. Chergui, and P. Le Quéré, 3D spectral parallel multi-domain computing for natural convection flows, in *Parallel Computational Fluid Dynamics*, Lecture Notes in Computational Science and Engineering Vol. 74 (Springer, Berlin, 2008), pp. 163–171.
[35] O. Shishkina, R. J. A. M. Stevens, S. Grossmann, and D. Lohse, Boundary layer structure in turbulent thermal convection and its consequences for the required numerical resolution, New J. Phys. **12**, 075022 (2010).
[36] M. Delort-Laval, L. Soucasse, P. Rivière, and A. Soufiani, Rayleigh-Bénard convection in a cubic cell under the effects of gas radiation up to $\mathrm{Ra} = 10^9$, Int. J. Heat Mass Transfer **187**, 122453 (2022).
[37] D. Puigjaner, J. Herrero, C. Simó, and F. Giralt, Bifurcation analysis of steady Rayleigh-Bénard convection in a cubical cavity with conducting sidewalls, J. Fluid Mech. **598**, 393 (2008).
[38] P. Holmes, J. Lumley, G. Berkooz, and C. Rowley, *Turbulence, Coherent Structures, Dynamical Systems and Symmetry* (Cambridge University Press, Cambridge, UK, 2002).
[39] R. Řehuřek and P. Sojka, Software framework for topic modelling with large corpora, in *Proceedings of the LREC 2010 Workshop New Challenges for NLP Frameworks* (University of Malta, 2010), pp. 46–50.
[40] V. T. Vishnu, A. K. De, and P. K. Mishra, Dynamics of large-scale circulation and energy transfer mechanism in turbulent Rayleigh–Bénard convection in a cubic cell, Phys. Fluids **32**, 095115 (2020).