



HAL
open science

Analysis of Rayleigh-Bénard convection using latent Dirichlet allocation

Bérengère Podvin, L. Soucasse, F. Yvon

► **To cite this version:**

Bérengère Podvin, L. Soucasse, F. Yvon. Analysis of Rayleigh-Bénard convection using latent Dirichlet allocation. *Physical Review Fluids*, 2024, 9 (6), pp.063502. 10.1103/PhysRevFluids.9.063502. hal-04729077

HAL Id: hal-04729077

<https://hal.science/hal-04729077v1>

Submitted on 10 Oct 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 , ,

2 **On the analysis of Rayleigh-Bénard convection using Latent Dirichlet**

3 **Allocation**

4 B. Podvin and L. Soucasse

5 *EM2C, Centralesupélec, CNRS, Université Paris-Saclay*

6 F. Yvon

7 *LISN, CNRS, Université Paris-Saclay*

Abstract

We apply a probabilistic clustering method, Latent Dirichlet Allocation (LDA), to characterize the large-scale dynamics of Rayleigh-Bénard convection. The method, introduced in Frihat *et al.* [1], is applied to a collection of snapshots in the vertical mid-planes 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 modelled 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.

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 $Pr = \nu/\kappa$ where ν is the kinematic viscosity and κ the thermal diffusivity, and the Rayleigh number

$$Ra = \frac{g\beta\Delta TH^3}{\nu\kappa}, \quad (1)$$

where g is the gravity, β is the thermal expansion coefficient, ΔT the temperature difference and H the cell dimension. The Grossmann and Lohse [2] theory constitutes a unified approach to address this question. It is based on a local description of the physics: the contributions from the bulk

37 averaged thermal and kinetic dissipation rate are split into two subsets, one corresponding to the
38 boundary layers, and one corresponding to the bulk. This theory was further refined in Grossmann
39 and Lohse [3], where the thermal dissipation rate was split into a contribution from the plumes
40 and a contribution from the turbulent background. Through the action of buoyancy, the thermal
41 boundary layers generate plumes which create a large-scale circulation, as evidenced by Xi *et al.* [4],
42 also called "wind" [5]. The distribution of temperature fluctuations depends on plume clustering
43 effects [6], but it is also affected by interactions with turbulent fluctuations in the bulk, resulting in
44 fragmentation [7].

45 Shang *et al.* [8] showed that plume-dominated regions were located near the sidewalls and
46 the conducting surfaces and that thermal plumes carry most of the convective heat flux, which
47 contributes to the production of both kinetic and thermal fluctuations. The morphology of plumes
48 and its effect on the heat transfer have been given careful attention. The plumes have a sheet-
49 like structure near the boundary layer and progressively become mushroom-like as they move
50 into the bulk region [9]. Shishkina and Wagner [10] found that very high values of the local
51 heat flux were observed in regions where the sheet-like plumes merged, constituting "stems" for
52 the mushroom-like plumes developing in the bulk. The relative contributions of the plumes and
53 turbulent background vary with the Rayleigh number: Emran and Schumacher [11] have shown
54 that the fraction of plume-dominated regions decreases with the Rayleigh number, while that of
55 background-dominated regions increases.

56 The identification of local coherent structures such as plumes is therefore an essential step for
57 the understanding of thermal convection flows. Several definitions have been used: some of the
58 first criteria were based on the skewness of the temperature derivative [12] or the temperature dif-
59 ference [13]. Ching *et al.* [14] have proposed to use simultaneous measurements of the temperature
60 and the velocity to define the velocity of the plumes using conditional averaging. Following Huang
61 *et al.* [15], van der Poel *et al.* [16] identified plumes from both a temperature anomaly and an excess
62 of convective heat flux. Zhou *et al.* [17] relied on cliff-ramp-like structures in the temperature
63 signals to determine the spatial characteristics of plumes. Emran and Schumacher [11] and Vishnu
64 *et al.* [18] separated the plume from the background regions based on a threshold on the convective
65 heat flux. Shevkar *et al.* [19] have recently proposed a dynamic criterion based on the 2D velocity
66 divergence to separate plumes from boundary layers.

67 As pointed out by Chilla and Schumacher [20], this multiplicity of criteria illustrates the difficulty
68 of identifying coherent structures in a consistent and objective manner, which is a long-running

69 question in various types of turbulent flows. To this end, Proper Orthogonal Decomposition
70 (POD) [21] has proven a useful tool to analyze large-scale fluctuations in Rayleigh-Bénard convec-
71 tion. It has been used in particular to study reorientations of the large-scale circulation [22–26].
72 Through spectral decomposition of the autocorrelation tensor, POD provides a basis of spatial
73 modes, also called empirical modes, since they originate from the data. The modes are energet-
74 ically optimal to reconstruct the fluctuations. The POD modes typically have a global support,
75 which is well suited to capture the large-scale organization of the flow. However, this can make
76 physical interpretation difficult as there is no straightforward connection between a mode and a lo-
77 cal coherent structure as a local structure is represented with a superposition of many POD modes,
78 a situation also observed in Fourier analysis. Soucasse *et al.* [27] have used POD to study the dy-
79 namics of the large-scale circulation for Rayleigh numbers in the range $[10^6, 10^8]$. They found that
80 although the reorientation rate varied with the Rayleigh number, the dominant structures remained
81 similar across that range, albeit with some variations in their energy. A new dissipation-based POD
82 decomposition, proposed by Olesen *et al.* [28] and applied to Rayleigh-Bénard convection [29],
83 highlighted the importance of boundary layers for the dynamics, which points to the need for local
84 descriptions.

85 As an alternative, Frihat *et al.* [1] have recently adapted a probabilistic method that can extract
86 localized latent factors in turbulent flow measurements. This method, Latent Dirichlet Allocation
87 or LDA [30, 31], was originally developed in the context of natural language processing, where
88 it aims to extract topics from a collection of documents. In this framework, documents are
89 represented by a non-ordered set of words taken from a fixed vocabulary. A word count matrix can
90 be built for the collection, where each column corresponds to a document, each line corresponds
91 to a vocabulary word and the matrix entry represents the number of times the word appears in the
92 document. LDA provides a probabilistic decomposition of the word count matrix, based on latent
93 factors called *topics*. Topics are defined by two distributions: the distribution of topics within each
94 document (each document is associated with a mixture of topics, the coefficients of the mixture
95 sum up to one) and the distribution of vocabulary words with each topic (each topic is represented
96 by a combination of words, the coefficients of which also sum up to one).

97 The method has been adapted for turbulent flows as follows: we consider a collection of
98 snapshots of a scalar field [discretized into cells](#). The equivalent of a document is therefore a
99 snapshot, and the cells (or snapshot pixels) constitute the vocabulary. The digitized values of the
100 scalar field over the cells in a snapshot are gathered into a vector which is formally analogous to

101 a column of the word count matrix. The "topics" produced by the decomposition, called *motifs*,
102 correspond to fixed (in the Eulerian sense), spatially coherent regions of the flow. The method was
103 found to be well suited for the representation of intermittent data (Frihat *et al.* [1], Fery *et al.* [32]).
104 It was successfully applied to the analysis of the turbulent Reynolds stress in wall turbulence [1].
105 Moreover, the method provides a local description that is insensitive to the existence of global
106 symmetries. It proved a useful tool to identify synoptic objects in weather data [32].

107 In this paper, we apply this method to the analysis of fluctuations in a cubic Rayleigh-Bénard
108 cell in the range of Rayleigh number $[10^6, 10^8]$. The goal is to track the local signature of
109 the large-scale dynamics of the flow, and to determine whether changes can be identified as the
110 Rayleigh number increases. To this end, the technique is applied to [2D snapshots extracted from](#)
111 [3D numerical simulations of Rayleigh-Bénard convection in a cubic cell](#) in the range of Rayleigh
112 number $[10^6, 10^8]$. The numerical configuration and the data set are described in Sec. II. We first
113 present the method for the convective heat flux, using a comparison with POD to highlight the
114 similarities and differences of the approach. POD and LDA are respectively presented in Sec. III
115 and IV. We examine in Sec. V how LDA compares with POD and the extent to which it is able
116 to capture the general features of the heat flux. The characteristics of heat flux motifs and their
117 connection with the reorientations of the large-scale circulation are discussed in Sec. VI. The
118 analysis is then extended to temperature fluctuations and to the kinetic energy in Sec. VII. in order
119 to provide further insight into the physics. A conclusion is given in Sec. VIII.

120 II. NUMERICAL SETTING

121 A. Set-up

122 Numerical setup and associated datasets are the same as used in Soucasse *et al.* [26, 27]. The
123 configuration studied is a cubic Rayleigh-Bénard cell filled with air, with isothermal horizontal
124 walls and adiabatic side walls. The air is assumed to be transparent and thermal radiation effects are
125 disregarded. Direct numerical simulations have been performed at various values of the Rayleigh
126 number. The Prandtl number is set to 0.707. All physical quantities are made dimensionless using
127 the cell size H , the reference time $H^2/(\kappa\sqrt{Ra})$ and the reduced temperature $\theta = (T - T_0)/\Delta T$, T_0
128 being the mean temperature between hot and cold walls. Spatial coordinates are denoted x, y, z (z
129 being the vertical direction) and the origin is placed at a bottom corner of the cube.

Ra	(N_x, N_y, N_z)	N_S	Δt	δ_{BL}
10^6	(81,81,81)	1000	10	0.056
$3 \cdot 10^6$	(81,81,81)	1000	10	0.042
10^7	(81,81,81)	1000	10	0.0297
10^8	(161,161,161)	1000	5	0.0167

TABLE I. Characteristics of the datasets 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 Δt and thermal boundary layer thickness δ_{BL} .

130 Navier–Stokes equations under Boussinesq approximation are solved using a Chebyshev collo-
131 cation method [33, 34]. Computations are made parallel using domain decomposition along the
132 vertical direction. Time integration is performed through a second-order semi-implicit scheme.
133 The velocity divergence-free condition is enforced using a projection method. Numerical param-
134 eters are given in Table I for the four considered Rayleigh numbers $Ra = \{10^6; 3 \cdot 10^6; 10^7; 10^8\}$.
135 We have checked that the number of collocation points is sufficient to accurately discretize the
136 boundary layers according to the criterion proposed by Shishkina *et al.* [35]. [Details on the vali-](#)
137 [dation of the numerical method and of the discretisation can be found in Ref. \[36\].](#) A number of
138 1000 snapshots have been extracted from the simulations for each Rayleigh number at a sampling
139 period of 10 (at $Ra = \{10^6; 3 \cdot 10^6; 10^7\}$) or 5 (at $Ra = 10^8$), in dimensionless time units. It is
140 worth noting that the time separation between the snapshots is sufficient to describe the evolution
141 of the large-scale circulation but is not suited for a fine description of the plume emission or of the
142 reorientation process. For each Rayleigh number, a dataset satisfying the statistical symmetries of
143 the flow was then constructed from these 1000 snapshots, as will be described in the next section.

144 B. Construction of the data set

145 At each Rayleigh number, the data set consisted of a collection of $N_S = 1000$ snapshots $q(x, t_m)$,
146 $m = 1, \dots, N_S$. Results will be presented first for the convective heat flux $q = \Phi = w\theta$, then for
147 the temperature fluctuations $q = \theta' = \theta - \langle \theta \rangle$ ($\langle \theta \rangle$ being the time-averaged temperature) and for
148 the kinetic energy $q = k = \frac{1}{2}(u^2 + v^2 + w^2)$, u, v and w being the velocity components. We note
149 that due to the velocity reference scale, the non-dimensional heat flux varies like $NuRa^{-1/2}$. As
150 in Soucasse *et al.* [26], the data set was first enriched by making use of the statistical symmetries

151 of the flow [37]. In the cubic Rayleigh-Bénard cell, four quasi-stable states are available for the
152 flow for this Rayleigh number range: the large-scale circulation settles in one of the two diagonal
153 planes of the cube with clockwise or counterclockwise motion. The evolution of the large-scale
154 circulation can be tracked through that of the x and y components of the angular momentum of
155 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 $Ra = 10^7$, the
156 angular momentum along each horizontal direction oscillates near a quasi-steady position for long
157 periods of times - several hundreds of convective time scales, before experiencing a rapid switch
158 ($\mathcal{O}(10)$ convective time scales) to the opposite value, which corresponds to a reorientation. On
159 each plane we can define an indicator function I , which takes the value $\text{sgn}(L)$ where L is the
160 angular momentum component normal to the plane.

161 Reorientations from one state to another occur during the time sequence but each state is not
162 necessarily equally visited. In order to counteract this bias, we have built enlarged snapshot sets,
163 obtained by the action of the symmetry group of the problem on the original snapshot sets. The
164 symmetries are based on four independent symmetries S_x , S_y , S_z and S_d with respect to the planes
165 $x = 0.5$, $y = 0.5$, $z = 0.5$ and $x = y$. This generates a group of 16 symmetries for the cube, which
166 should lead to a 16-fold in the number of snapshots. However, since we will exclusively consider
167 the vertical mid-planes $x = 0.5$ and $y = 0.5$, which are invariant planes for respectively S_x and S_y ,
168 the increase is reduced. The data set aggregates 1000 snapshots on each of the planes $x = 0.5$ and
169 $y = 0.5$, each of which undergoes a vertical flip, a horizontal flip and a combination of the two,
170 yielding a total of $N_S = 8000$ snapshots.

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

$$175 \quad q(\underline{x}_j) = q(\underline{x}_j, t_m) = \text{Max} \left[\text{Int}[s w(\underline{x}_j, t_m)\theta(\underline{x}_j, t_m), 0], \right] \quad (2)$$

$$176 \quad q(\underline{x}_{j+\tilde{N}_C}) = q(\underline{x}_{j+\tilde{N}_C}, t_m) = -\text{Max} \left[-\text{Int}[s w(\underline{x}_j, t_m)\theta(\underline{x}_j, t_m), 0], \right] \quad (3)$$

177 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 mid-planes
178 $x = 0.5$ or $y = 0.5$. We note that throughout the paper, the total field will directly be represented
179 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
180 is worth noting that the temperature variance (always positive) could be used to lighten the LDA
181 analysis on the temperature field. Yet, we chose to work on the signed temperature fluctuation in

182 order to discriminate between leaving and impinging thermal patterns near the horizontal walls as
 183 it is often done in plume detection [13, 16].

184 III. POD ANALYSIS

185 A. Method

186 Proper Orthogonal Decomposition (POD) [38] makes it possible to write a collection of N_S
 187 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
 188 amplitude of which varies in time:

$$189 \quad q(\underline{x}_j, t_m) = \sum_{n=1}^{N_S} \sqrt{\lambda_n} a_n(t_m) \varphi_n(\underline{x}_j), \quad (4)$$

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

$$191 \quad \sum_{j=1}^{N_C} \varphi_n(\underline{x}_j) \varphi_m(\underline{x}_j) = \delta_{nm}. \quad (5)$$

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

$$193 \quad C_{mp} a_n(t_p) = \lambda_n a_n(t_m), \quad (6)$$

194 where C is the temporal autocorrelation matrix

$$195 \quad C_{mp} = \frac{1}{N_S} \sum_{j=1}^{N_C} q(\underline{x}_j, t_m) q(\underline{x}_j, t_p). \quad (7)$$

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

200 B. Application to the convective heat flux

201 POD is applied to the digitized heat flux signal $q = \Phi$ defined in equations (2) and (3). The first
 202 three POD modes and POD coefficients are shown in Fig. 2 for $Ra = 10^7$, where black vertical
 203 and horizontal lines indicate the thickness of the boundary layers. We checked that the first mode

204 corresponds to the mean flow. The mode is most important in a region close to the wall, with
 205 a maximum within the vertical boundary layer at a height of about $z \sim 0.1$. The second mode
 206 corresponds to a dissymmetry between the vertical sides and is most important at mid-height in the
 207 region outside the boundary layers. The third mode is both antisymmetric in the vertical and in the
 208 horizontal direction. It is maximum at the edge of the vertical boundary layers, at a vertical distance
 209 of about 0.25 from the horizontal surfaces. The pattern it is associated with corresponds to a more
 210 intense flux along a diagonal (bottom of one side and top of the opposite side) and a less intense
 211 flux along the opposite diagonal. As evidenced by application of a moving average performed over
 212 200 convective time units (about 4 times the recirculation time T_c , as was determined in Soucasse
 213 *et al.* [26]), the evolution of the amplitude at large time scales matches that of the horizontal angular
 214 momentum components L_x and L_y (compare with Fig. 1), unlike the two dominant modes. This
 215 mode therefore appears to be the signature of the large-scale circulation, where the flux is more
 216 intense in the lower corner of the cell as hot plumes rise on one side and in the upper corner of the
 217 opposite side of the cell as cold plumes go down.

218 IV. LATENT DIRICHLET ALLOCATION

219 A. Principles

220 We briefly review the principles of Latent Dirichlet Allocation and refer the reader to Frihat
 221 *et al.* [1] for more details. LDA is an inference approach to identify latent factors in a collection of
 222 observed data, which relies on Dirichlet distributions as priors.

223 We first recall the definition of a Dirichlet distribution ϑ , which is a multivariate probability
 224 distribution over the space of multinomial distributions. It is parameterized by a vector of positive-
 225 valued parameters $\underline{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_N)$ as follows

$$226 \quad p(\vartheta_1, \dots, \vartheta_N; \alpha_1, \dots, \alpha_N) = \frac{1}{B(\underline{\alpha})} \prod_{n=1}^N \vartheta_n^{\alpha_n - 1}, \quad (8)$$

227 where B is a normalizing factor, which can be expressed in terms of the Gamma function Γ :

$$228 \quad B(\underline{\alpha}) = \frac{\prod_{n=1}^N \Gamma(\alpha_n)}{\Gamma(\sum_{n=1}^N \alpha_n)}. \quad (9)$$

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

232 As mentioned above, the data to which LDA is applied consists of a collection of non-negative,
 233 integer fields that are defined in equations (2) and (3). For each snapshot m , the integer value
 234 $q_m(\underline{x}_j)$ measured at cell j is interpreted as an integer count of the cell j . The key is to interpret
 235 this integer count as the number of times cell j appears in the composition of snapshot m . The idea
 236 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 ,
 237 which is directly proportional to $q(\underline{x}_j)$. The model is based on the hypothesis that each snapshot of
 238 the collection $\{q_m, m = 1, \dots, N_S\}$ consists of a mixture of N_T latent factors $\{z_n, n = 1, \dots, N_T\}$
 239 called motifs, N_T being a user-defined parameter analogous to a number of clusters. The probability
 240 $p(q_m, \underline{x}_j)$ therefore writes

$$241 \quad p(q_m, \underline{x}_j) = p(q_m) \sum_n p(z_n|q_m)p(\underline{x}_j|z_n), \quad (10)$$

242 where:

- 243 • $p(q_m)$ is the probability of observing the snapshot q_m in the collection,
- 244 • $p(z_n|q_m)$ is the conditional probability of observing motif z_n given the presence of snapshot
 245 q_m ,
- 246 • $p(\underline{x}_j|z_n)$ is the conditional probability of observing cell \underline{x}_j given the latent factor z_n .

247 A formal analogy with POD and equation (4) can be seen by using the Bayes rule and rewriting
 248 $p(q_m, \underline{x}_j)$ as

$$249 \quad p(q_m, \underline{x}_j) = \sum_n p(z_n)p(q_m|z_n)p(\underline{x}_j|z_n), \quad (11)$$

250 where

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

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

256 Latent Dirichlet allocation is therefore based on the following representation:

- 257 1. Each motif z_n is associated with a multinomial distribution ψ_n over the grid cells so that the
 258 probability to observe the j^{th} grid cell located at \underline{x}_j given the motif n is $p(\underline{x}_j|z_n) = \psi_n(x_j)$.
 259 The distribution ψ_n is modelled with a Dirichlet prior parameterized with an N_C -dimensional
 260 vector $\underline{\eta}$. Low values of η_l mean that the motif is distributed over a small number of cells.
- 261 2. Each snapshot q_m is associated with a distribution b_n over the motifs such that the probability
 262 that motif n is present in snapshot m will be denoted $p(q_m|z_n) = b_n(t_m)$. This distribution
 263 is modelled with a N_T -dimensional Dirichlet distribution of parameter $\underline{\alpha}$. The magnitude of
 264 α characterizes the sparsity of the distribution. Low values of α_n mean that relatively few
 265 motifs are observed in each snapshot.

266 B. Implementation

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

274 The solution *a priori* depends on the number of motifs N_T as well as on the values of the Dirichlet
 275 parameters $\underline{\alpha}$ and $\underline{\eta}$. Special attention was therefore given to establish the robustness of the results
 276 reported here. Non-informative default values were used for the Dirichlet parameters i.e. the prior
 277 distributions were taken with symmetric parameters equal to $\forall n, \alpha_n = 1/N_T$ and $\forall j, \eta_j = 1/N_T$.
 278 Practical implementation was performed in Python using `gensim` [39]. No significant change was
 279 observed in the results when the value of the quantization s was high enough (however it had to
 280 be kept reasonably low in order to limit the computational time). Although multiple tests were
 281 carried out for varying values of $s \in [40, 600]$, all results reported in this paper were obtained with
 282 $s = 600$ for the heat flux. Values of $s = 40$ and $s = 50$ were respectively used for the temperature
 283 fluctuations and for the kinetic energy. Analyses were also performed for varying numbers of
 284 motifs N_T , ranging from 50 to 400.

285 C. LDA as a generative process

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

287 (i) For each motif n , a N_C -dimensional cell–motif distribution ψ_n is drawn from the Dirichlet
288 distribution of parameter $\underline{\eta}$.

289 (ii) To generate snapshot m :

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

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

295 (c) for each $i = 1, \dots, q_T(t_m)$:

296 * a motif n is selected from $b_n(t_m)$ (since it represents the probability that motif n is
297 present in the snapshot m)

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

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

303 In fluid mechanics applications ([1, 32]), sampling from the motif-cell distribution (step c))
304 can be replaced with a faster step, where the contribution of each motif n to snapshot m is
305 directly obtained from the motif-cell distribution ψ_n and the distribution $b_n(t_m)$ and expressed as
306 $q_T(t_m)b_n(t_m)\psi_n(\underline{x}_j)$. The reconstructed field is then the sum of the motif contributions. [This](#)
307 [matrix-like form of the reconstruction is summarized in the bottom part of Fig. 3.](#)

308 D. Interpretation and evaluation criteria

309 By construction, the decomposition identifies *fixed* regions of space over which the intensity of
310 the scalar field is likely to be important at the same time. The connection between temperature
311 motifs and plumes should be examined with caution since plumes are Lagrangian structures

312 travelling and possibly changing in shape and orientation through the shell. LDA motifs only aim
 313 to detect the Eulerian signature of structures.

314 Each motif n can be characterized in space through the motif-cell distribution ψ_n (which
 315 integrates to 1 over the cells) and which will sometimes referred to as the motif in the absence of
 316 ambiguity. Each distribution has a maximum value ψ_n^{max} and a maximum location \underline{x}_n^{max} such that
 317 $\psi_n(\underline{x}_n^{max}) = \psi_n^{max}$. One can also define a characteristic area Σ_n using

$$318 \quad \Sigma_n = \int_{\Omega} 1_{\{\psi_n \geq \psi_n^{max}/e\}} d\Omega, \quad (12)$$

319 where Ω represents the plane of analysis and the factor $1/e \sim 0.606$ is an arbitrary factor chosen
 320 by analogy with a Gaussian distribution. If ψ_n were a Gaussian of standard deviation σ , this
 321 value would delimit an area of size $2\pi\sigma^2$. Other choices could be made such as the full width at
 322 half-maximum corresponding to a factor of $1/2$. Moderate changes in the choice of the factor did
 323 not affect the trends reported below. Characteristic dimensions l_i for the motif n in the direction i
 324 can also be defined using $l_i^n = \left[\int \psi_n(x_{n,i} - x_{n,i}^{max})^2 dx_i \right]^{1/2}$. Each motif can also be characterized in
 325 time through the snapshot-motif distribution b_n , that will be called the motif *weight* throughout the
 326 paper. The motifs can be ordered by their time-averaged weight, also called *prevalence*, defined as
 327 $\langle b_n \rangle = \frac{1}{N_S} \sum_{m=1}^{N_S} b_n(t_m)$ where $\langle \cdot \rangle$ represents a time average.

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

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

$$338 \quad q_R(\underline{x}_j, t_m) = \sum_{n=1}^{N_T} q_T(t_m) b_n^q(t_m) \psi_n^q(\underline{x}_j). \quad (13)$$

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

342 its reconstruction q_R defined as

$$343 \quad C_m(q, q_R) = \frac{\int (\tilde{q}(\underline{x}, t_m) \tilde{q}_R(\underline{x}, t_m) d\underline{x})}{\left(\int \tilde{q}^2(\underline{x}, t_m) d\underline{x} \int \tilde{q}_R^2(\underline{x}, t_m) d\underline{x} \right)^{1/2}}, \quad (14)$$

344 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
 345 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.

346 V. EVALUATION OF LDA FOR RECONSTRUCTION AND GENERATION OF THE HEAT 347 FLUX

348 A. Reconstruction

349 We first evaluate to which extent the LDA decomposition provides an adequate reconstruction
 350 of the heat flux Φ . Figure 4 (left) shows how the instantaneous value of the correlation coefficient
 351 $C_m(\Phi, \Phi_R)$ depends on the discrete integral of the field $q_T(t_m) = \sum_j \Phi(\underline{x}_j, t_m)$. The Rayleigh
 352 considered is $Ra = 10^7$ and the number of topics is $N_T = 100$, but the same trend was observed
 353 for all other Rayleigh numbers as well as all other values of N_T . Lower values of the correlation
 354 were associated with lower values of the total integrated heat flux, which illustrates that the LDA
 355 representation is suited to capture extreme events.

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

361 Figure 5 compares an original snapshot at $Ra = 10^7$ (based on the digitized signal) with different
 362 reconstructions: i) the LDA-reconstruction based on $N_T = 100$ motifs; ii) the reconstruction limited
 363 to the 20 most prevalent topics (for this particular snapshot); iii) the POD-based reconstruction
 364 based on the first 20 modes. By construction, POD provides the best approximation of the field for
 365 a given number of modes. Since the distribution of the heat flux is intermittent in space and time,
 366 only a limited number of motifs is necessary to reconstruct the flow. We note that little difference
 367 was observed between the full LDA reconstruction and the reconstruction limited to the 20 most
 368 prevalent motifs, which highlights the intermittent nature of the field. The relative error between
 369 the original and the reconstructed field is 29% for the full LDA reconstruction, 34% when the 20

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

378 B. Generation

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

- 389 1. the original (digitized) field Φ defined in Sec. II B with equations (2) and (3)
- 390 2. the LDA-Reconstructed field (LDA-R) as defined in equation (13)
- 391 3. the LDA-Generated field (LDA-G): as described in Sec. IV C, the field is constructed by
 392 sampling weights $\tilde{b}_n(t_m)$ from snapshot-motif distributions and then reconstructing

$$393 \quad \Phi^{LDA-G}(\underline{x}_j, t_m) = \Phi_T(t_m) \sum_{n=1}^{N_T} \tilde{b}_n^\Phi(t_m) \psi_n^\Phi(\underline{x}_j), \quad (15)$$

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

398 4. the POD-Generated field (POD-G): the field is constructed by independently sampling N_T
 399 POD mode amplitudes \tilde{a}_n from the POD amplitudes of the original database

$$400 \quad \Phi^{POD-G}(\underline{x}_j, t_m) = \sum_{n=1}^{N_T} \sqrt{\lambda_n} \tilde{a}_n(t_m) \varphi_n(\underline{x}_j). \quad (16)$$

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

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

$$408 \quad R_y(y, y_0, z_0) = \frac{\langle \Phi(y, z_0, t) \Phi(y_0, z_0, t) \rangle}{\langle \Phi(y_0, z_0, t)^2 \rangle}, \quad (17)$$

$$409 \quad R_z(z, y_0, z_0) = \frac{\langle \Phi(y_0, z, t) \Phi(y_0, z_0, t) \rangle}{\langle \Phi(y_0, z_0, t)^2 \rangle}. \quad (18)$$

411 The autocorrelation functions are displayed in Fig. VIII for the selected locations indicated in
 412 Fig. 6, which correspond to regions of high heat flux. We can see that that in all cases, the flux
 413 remains correlated over much longer vertical extents than in the horizontal direction. Both the
 414 LDA-reconstructed and the POD-generated autocorrelations approximate the original data well -
 415 again, by construction, POD-based fields are optimal to reconstruct second-order statistics. The
 416 LDA-generated autocorrelation is not as close to the original one, but still manages to capture the
 417 characteristic spatial scale over which the fields are correlated.

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

423 VI. HEAT FLUX MOTIFS

424 A. Spatial organization

425 We now describe the spatial organization of the motifs through the motif-cell distribution ψ_n .
426 The general trends reported below held for all values of N_T considered, which ranged from 50 to
427 400. For all Rayleigh numbers, most LDA motifs were found to be associated with a positive flux
428 (i.e. were associated with the first \tilde{N} cells in the decomposition). A few negative (counter-gradient)
429 motifs were also identified, but their average weight was generally very small (at most 10% of
430 that of the dominant motif). We therefore chose to focus only on the motifs making a positive
431 contribution to the heat flux. Figure 9 (left) displays these motifs for three different Rayleigh
432 numbers for $N_T = 100$. The case $Ra = 3 \cdot 10^6$ was omitted as it did not show significant differences
433 with the case $Ra = 10^6$. The motif-cell distribution is materialized by a black line corresponding
434 to the iso-probability contour of $0.606 \psi_n^{max}$, which can be compared with the average value of the
435 heat flux at this location. For all Rayleigh numbers, the motifs are clustered in the regions of high
436 heat flux, close to the vertical walls. Within the vertical boundary layers, motifs are elongated in
437 shape. Outside the vertical boundary layers, the motifs are more isotropic and tend to increase in
438 shape as one moves away from the walls. Outside the horizontal boundary layers, the motif-cell
439 distributions are elongated in the direction of the wind, with a horizontal orientation in the center
440 of the cell, and a gradual vertical shift closer to the walls. Large motifs are found in the bulk at
441 $Ra = 10^6$ and $Ra = 10^7$ (it was also the case at $Ra = 3 \cdot 10^6$). In contrast, fewer, smaller motifs
442 are found in the bulk at $Ra = 10^8$ in the central region $x/y \in [0.2, 0.8]$, signalling a loss of spatial
443 coherence in the bulk at this Rayleigh number.

444 In general, the motif size seems to decrease with the Rayleigh number. This is confirmed by
445 Fig. 10, which represents the average motif area as a function of their distance from the vertical
446 walls. In order to avoid the influence of the horizontal plates, we only considered the motifs
447 located at a vertical distance larger than 0.07 from the horizontal walls (i.e. outside the horizontal
448 boundary layer). The size of the symbols shown in the picture is proportional to the fraction of
449 motifs over which the average was performed. Results were relatively robust with respect to the
450 number of topics N_T , although some dependence on N_T is observed in the center of the cell. Within
451 the boundary layer, the motif area grows roughly quadratically (a power law fit yielded exponents
452 in the range 1.6 – 2 at all Rayleigh numbers) which means that the characteristic motif size of

453 the motif *essentially grows like the wall distance*. We note that a similar scaling was found for
 454 turbulent eddies in pressure-gradient driven turbulence such as channel flow [1]. Further away
 455 from the vertical wall, after a short plateau at the edge of the boundary layer, a slower increase
 456 in the motif size was observed with a rate that increased with the Rayleigh number, so that the
 457 motif area was about the same (on the order of 0.02) for all Rayleigh numbers in the center of the
 458 cell. This suggests the presence of a double scaling for the motifs: one based on the boundary
 459 layer thickness, and one based on the cell size. The decrease in size with the Rayleigh number
 460 appears consistent with a dependence on the boundary layer thickness but also with an increase of
 461 the fragmentation by the bulk turbulent fluctuations, in agreement with the literature [7, 16]. The
 462 difference observed at the highest Rayleigh number also signals that the flow is still evolving and
 463 has not reached an asymptotic state.

464 **B. Dominant motifs**

465 *1. Spatial description*

466 Owing to the symmetry of the database (see Sec. **II B**), the motifs in the vertical plane (x, z)
 467 (resp. (y, z)) should approximate the symmetry $S_x : x \rightarrow 1 - x$ (resp. $S_y : y \rightarrow 1 - y$), and
 468 $S_z : z \rightarrow 1 - z$ (complete symmetry cannot be expected owing to the stochastic nature of the
 469 decomposition).

470 To help interpret the heat flux motifs, we compare them with LDA motifs corresponding to
 471 temperature fluctuations. The eight most prevalent heat motifs are represented in Fig. 11 (green
 472 lines). The prevalence of each motif is indicated at the top of each plot. Most motifs have similar
 473 sizes and are located close to the side walls at about a similar height, except for motifs 4 and
 474 6, which have a smaller extent and are located closer to the horizontal wall. The same value of
 475 $N_T = 100$ was used for both heat flux and temperature.

476 For a heat flux motif n with weight b_n^Φ , we identified the temperature motif j that maximized the
 477 correlation coefficient between the heat flux and the temperature motif weights $C(b_n^\Phi, b_{n'}^\theta)$. The
 478 maximal value of this coefficient, denoted c , is represented on each plot and is generally very high
 479 (about 0.7), especially in view of the intermittent nature of the weights. The best correlated heat
 480 flux and temperature motifs are close to each other in space, with a larger spread for temperature
 481 motifs. In all cases, flux motifs in the lower (resp. higher) portion of the side walls correspond to

482 positive (resp. negative) fluctuations. Dominant heat flux motifs can be therefore interpreted as
 483 the wall imprint of hot plumes rising in the boundary layer (resp. cold plumes descending in the
 484 boundary layer). The same observations were made at all other Rayleigh numbers.

485 Four of these dominant motifs at $Ra = 10^7$ are represented in Fig. 12 (left) for $N_T = 100$. As
 486 noted above, they consist of elongated structures lying mostly in the boundary layer, and located at
 487 a vertical distance of about 0.25 from the horizontal walls. Although the positions and sizes of the
 488 four identified motifs may slightly vary from one to the other, their features are generally similar
 489 and a characteristic motif can be obtained from taking the average over all four motifs. Figure 12
 490 (right) represents this characteristic motif for the various Rayleigh numbers. We can see that the
 491 dominant motifs are always located mostly within the boundary layer, with a maximum at a height
 492 of about 0.25. Their characteristic width l_y was found to decrease as $Ra^{-0.23 \pm 0.04}$, which matches
 493 the scaling of the boundary layer thickness.

494 2. Temporal dynamics

495 The evolution of the snapshot-motif distribution, or motif weight, is represented in Fig. 13 for
 496 $Ra = 10^7$. We can see that the behavior of the motif weight depends on the sign of the global
 497 momentum represented in Fig. 1. When a moving average of $T_f = 200$ time units, corresponding
 498 to 4 recirculation times T_c , was applied, two quasi-stationary states b_+ and b_- could be identified in
 499 each plane (they are materialized by the dashed horizontal black lines indicated in Fig. 13). The two
 500 states appear to correspond to the sign of the angular momentum component i.e. the orientation of
 501 the large-scale circulation I . Streamlines of the flow conditionally averaged on the higher weight
 502 value of b_1^Φ are represented in Fig. 14 (left). They indicate that for the higher characteristic value
 503 of the weight, b_+ , the motif is associated with the large-scale circulation while it is associated with
 504 the corner vortex on the opposite side for the lower weight value, b_- , as summarized in Fig. 14
 505 (right).

506 This indicates that information about the large-scale reorientation can be extracted from local
 507 measurements. Two states, I_+ and I_- , respectively corresponding to the large-scale circulation and
 508 corner vortex can be defined from the weight of the dominant motif b_1^Φ using

$$509 \quad I_+ = \{m | \langle b_1^\Phi(t_m) \rangle_{T_f} > \langle b_1^\Phi \rangle\} \text{ and } I_- = \{m | \langle b_1^\Phi(t_m) \rangle_{T_f} < \langle b_1^\Phi \rangle\}, \quad (19)$$

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

511 are respectively b_+ and b_- .

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

519 However, further insights can be obtained by examining the respective contributions of the I_+
520 and I_- states to the distribution of b_1^Φ , which are also represented in Fig. 15. For all Rayleigh
521 numbers, I_+ states contribute more to the higher-value lobe than I_- states, while I_- contributes
522 more to the lower-value lobe. This shows that the rate of buoyancy production is less intense in the
523 corner rolls than in the large-scale circulation, or equivalently that plumes are emitted at a lower
524 frequency in the corner rolls than in the large-scale circulation. Moreover, the relative contributions
525 of the I_+ and the I_- states vary non-monotonically with the Rayleigh number. In the higher-value
526 lobe, the relative contribution of I_- appears to increase relatively to I_+ with more high values of
527 I_- at $Ra = 3 \cdot 10^6$, while I_- represents more low values at $Ra = 10^8$. In the lower-value lobe, the
528 contribution of I_+ is least at $Ra = 3 \cdot 10^6$ and largest at $Ra = 10^8$. These observations suggest that
529 both the intensity of the large-scale circulation and that of the corner roll appear to change with the
530 Rayleigh number, in agreement with the findings of Vishnu *et al.* [40].

531 3. A model for the reorientation time scale

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

542 x or y .

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

544
$$p = p(T_- > T_+) = \frac{b_-}{b_+ + b_-}. \quad (20)$$

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

550
$$f_r = 2p\beta^{-1}f_c, \quad (21)$$

551 where the factor 2 comes from the fact that a reorientation can occur in each direction. Figure 16
552 (right) compares for different Rayleigh numbers the probability p with the ratio of the frequency
553 between reorientations and the recirculation frequency estimated in Ref. [27]. We see that a
554 very good agreement is obtained between the variations of the average reorientation rate and the
555 measure of the relative intensity of the large-scale circulation and corner vortices. We note that
556 the largest discrepancy is observed for the highest Rayleigh number, for which the reorientation
557 rate is very low and therefore cannot be determined with good precision from the DNS. The value
558 of β used in the figure was determined empirically and was found to be 5.6, which makes T_s close
559 to the filtering time scale $T_f = 4T_c$. This suggests that an estimate for the reorientation rate can
560 be obtained by comparing directly the average weight of the motif associated with the large-scale
561 circulation with that of its counterpart in the corner structure. This could be of particular interest
562 in cases where the observation time is smaller than the expected reorientation time, a situation that
563 is often encountered in (but not limited to) numerical simulations [at higher Rayleigh numbers, as](#)
564 [the simulation time increases and the reorientation frequency decreases.](#)

565 VII. TEMPERATURE AND VELOCITY MOTIFS

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

$\langle C(q, q_R) \rangle$	N_T	$Ra = 10^6$	$Ra = 3 \cdot 10^6$	$Ra = 10^7$	$Ra = 10^8$
$\langle C(\theta, \theta_R) \rangle$	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

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

571 A. Temperature fluctuations

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

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

589 Figure 20 shows the evolution of the temperature motif weights b_n^θ on both planes along with
590 their filtered representation $\langle b_n^\theta \rangle_{T_f}$. As observed for the heat flux (Fig. 13), the importance of
591 the weights depends on the orientation of the large-scale circulation I . Similar evolutions were

592 observed at the lower Rayleigh numbers (not shown).

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

607 B. Kinetic energy

608 More details about the structure of the large-scale circulation can be obtained by examining
609 kinetic energy motifs. Figure 23 shows the spatial distribution of the velocity motifs for the
610 different Rayleigh numbers and $N_T = 100$. The spatial distribution of the time-averaged kinetic
611 energy is also represented on the same plot. The size of the core (low-velocity region) appears to
612 increase with the Rayleigh number. The size of the motifs did not appear to change significantly
613 with the Rayleigh number, except for horizontal corner structures that seem to scale with the
614 boundary layer thickness. The kinetic energy motifs have elongated shapes along the walls, with a
615 significantly higher extent along the horizontal walls, which shows the importance of entrainment
616 in the horizontal boundary layers, in particular in the middle of the cell. It is lowest at $Ra = 3 \cdot 10^6$
617 and highest at $Ra = 10^8$, which varies like the time between reorientations T_r . The question
618 is whether this reinforcement of the large-scale circulation can be associated with characteristic
619 temperature fluctuations.

620 In Figs. 24 to 26 the 16 most prevalent kinetic energy motifs are represented at Rayleigh numbers
621 10^6 , 10^7 and 10^8 (the case $Ra = 3 \cdot 10^6$, not shown, was found generally similar to 10^6 and 10^7).

622 The motifs were organized according to the location of their maximum: within the horizontal or
623 vertical boundary layers, which we will refer to as respectively HBL or VBL motifs, at the corners
624 of the horizontal and the vertical boundary layer (CBL motifs), and outside the boundary layers in
625 the horizontal or vertical entrainment zones, which were termed HEZ or VEZ motifs. The different
626 locations are shown in the top right illustration of Fig. 24. For each category the motifs are ordered
627 according to their prevalence, indicated at the top of each plot. Generally speaking, the prevalence
628 of the motifs increased with the Rayleigh number, which is consistent with a strengthening of the
629 large-scale circulation.

630 For each kinetic energy motif n (represented with green lines), we determined the temperature
631 motif j (represented with blue or red lines, depending on its sign) for which the correlation
632 coefficient $C(b_n^k, b_n^\theta)$ is maximal. The maximal value c and the temperature motif are represented
633 on each plot, except in two cases corresponding to HBL motifs, for which the associated temperature
634 motif had a very low prevalence and was considered to be irrelevant. In almost all cases, the
635 kinetic energy and temperature motifs are located close to each other in space. Although the
636 correlation coefficients are typically lower than those between the flux and temperature motifs
637 represented in Fig. 11, several are high enough to associate kinetic energy patterns with specific
638 temperature fluctuations. We also represented on each plot the correlation coefficient \bar{c}_I , defined
639 as $\bar{c}_I = C(\langle b_n^k \rangle_{T_f}, I)$, where $\langle b_n^k \rangle_{T_f}$ is the low-pass-filtered kinetic energy motif weight (using T_f)
640 and I is the large-scale circulation indicator defined in equation (19) (see also Fig. 24 top right).
641 High positive (resp. negative) values of \bar{c}_I are indicated in red (resp. blue) for each motif, and
642 show that the motif can be associated with a specific orientation of the large-scale circulation.

643 *a. Horizontal boundary layers and corners* In all cases, the most frequent motifs consist
644 of centered motifs close to the edge of the horizontal boundary layers (HBL). Evidence of weak
645 correlation (0.3) for some motifs suggested possible association with impinging plumes, however
646 generally low values of $|\bar{c}_I|$ suggest that the weights of the motifs do not depend on the orientation
647 of the large-scale circulation. In contrast, high values of c and $|\bar{c}_I|$ were found for corner (CBL)
648 motifs, that were best correlated with impinging plumes. Corner motifs have a relatively high
649 prevalence, which shows that impinging plumes make a significant contribution to the horizontal
650 wind at the edge of the boundary layer. The correlation coefficient \bar{c}_I increased in absolute value
651 with the Rayleigh number, and was larger than 0.9 at $Ra = 10^8$. In contrast, the maximum
652 correlation coefficient c tended to decrease (but remained significant) at $Ra = 10^8$.

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

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

668 *d. Horizontal entrainment zone* At $Ra = 10^6$ and $Ra = 10^7$, the last category of motifs
669 consisted of motifs in the horizontal entrainment zone (HEZ). At the lowest Rayleigh number
670 $Ra = 10^6$, two of the HEZ motifs (second and fourth motif in the last row in Fig. 24) have a
671 predominantly vertical shape and are associated with large temperature motifs originating from
672 the opposite (here, top) wall. They are therefore likely to represent coalescing plumes drifting
673 towards the center of the cell as they reach the opposite wall. In contrast, all other HEZ motifs at
674 all Rayleigh numbers have a horizontal shape and are associated with smaller temperature motifs
675 originating from the closest wall. They are very well correlated with the orientation of the large-
676 scale circulation. Significant changes were observed at $Ra = 10^8$, with a much larger number of
677 HEZ motifs and a noticeable increase in their prevalence - the prevalence of the dominant HEZ
678 motif is twice as large at $Ra = 10^8$ than at $Ra = 10^7$.

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

683 VIII. CONCLUSION

684 We have applied a new analysis technique, Latent Dirichlet Allocation, to characterize the
685 spatio-temporal organization of fluctuations in Rayleigh-Bénard convection. The method is based
686 on the inference of probabilistic latent factors, spatially localized motifs, from a collection of
687 instantaneous fields. It provides a local yet compact description of the flow in terms of quantitative
688 indicators such as the (spatial) size and the (temporal) weight of the motifs. The technique was
689 applied to the vertical mid-plane of a Rayleigh-Bénard cubic cell in a range of Rayleigh numbers in
690 $[10^6, 10^8]$. The method was found to be robust with respect to the user-defined parameters. When
691 applied to the heat flux, it was found to provide good reconstructions of the snapshots and was able
692 to generate new datasets that reproduced key statistics of the original one.

693 For all Rayleigh numbers, dominant heat flux motifs consisted of elongated vertical structures
694 located mostly within the vertical boundary layer, at a height of a quarter of the cell. The width
695 of these motifs scaled with the boundary layer thickness. These motifs were found to be very
696 well correlated with temperature motifs corresponding to plumes in their early formation stage
697 (leaving plumes). The motif weights were found to depend on the large-scale organization of the
698 flow: two states could be identified, one corresponding to the large-scale circulation and one to a
699 corner roll structure. The two states were characterized by different average weights which varied
700 non-monotonically with the Rayleigh number. A simple model was able to relate the weights of
701 the dominant heat flux motif associated with the two states with the average reorientation rate of
702 the large-scale circulation in the cell. This suggests that the model could be used as a predictor of
703 this rate in cases where few or even no reorientations are observed.

704 Additional insight about the flow physics was obtained by examining dominant motifs for the
705 temperature and the kinetic energy. While dominant heat flux motifs seemed to be associated
706 with early-stage (leaving) plumes, dominant temperature motifs were associated with later-stage
707 (arriving) plumes. In contrast with the lower Rayleigh numbers, dominant temperature motifs at
708 $Ra = 10^8$ were no longer within the vertical boundary layers, but consisted of plumes impinging
709 onto the corners of the horizontal boundary layers, which led to a reduction of temperature
710 gradients within the corner structure and a decrease in its potential energy. This is consistent with
711 the significant drop in the large-scale reorientation rate observed at this Rayleigh number. LDA
712 analysis of the kinetic energy showed that corner impinging plumes contributed to the kinetic energy
713 of both the corner structure and the large-scale circulation. The reduction of the reorientation rate

714 at $Ra = 10^8$ was also associated with a reinforcement of the horizontal wind in the central part of
715 the cell due to the formation and entrainment of new plumes. [Changes in the dynamics of the large-](#)
716 [scale circulation could thus be directly connected with local modifications of its structure.](#) The
717 LDA model therefore appears as a promising statistical tool that can help track subtle transitions
718 in the [spatio-temporal organization of turbulent flows.](#) [An interesting direction of investigation,](#)
719 [suggested by one of the anonymous Reviewers, would be to explore the connection between the](#)
720 [LDA representation and structure function analysis, which could provide insight into local energy](#)
721 [transfer mechanisms at different scales.](#)

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

-
- 727 [1] M. Frihat, B. Podvin, L. Mathelin, Y. Fraigneau, and F. Yvon, Coherent structure identification in
728 turbulent channel flow using latent dirichlet allocation, *Journal of Fluid Mechanics* **920** (2021).
- 729 [2] S. Grossmann and D. Lohse, Scaling in thermal convection: a unifying theory, *Journal of Fluid*
730 *Mechanics* **407**, 27–56 (2000).
- 731 [3] S. Grossmann and D. Lohse, Fluctuations in turbulent rayleigh-benard convection: The role of plumes,
732 [PHYSICS OF FLUIDS](#) **16**, 4462 (2004).
- 733 [4] H.-D. Xi, S. Lam, and K. Xia, From laminar plumes to organized flows: the onset of large-scale
734 circulation in turbulent thermal convection, [Journal of Fluid Mechanics](#) **503**, 47–56 (2004).
- 735 [5] B. Castaing, G. Gunaratne, F. Heslot, L. Kadanoff, A. Libchaber, S. Thomae, X.-Z. Wu, S. Zaleski,
736 and G. Zanetti, Scaling of hard thermal turbulence in Rayleigh-Bénard convection, *Journal of Fluid*
737 *Mechanics* **204**, 1 (1989).
- 738 [6] Y. Wang, Y. Wei, P. Tong, and X. He, Collective effect of thermal plumes on temperature fluctuations
739 in a closed rayleigh-bénard convection cell, [Journal of Fluid Mechanics](#) **934**, A13 (2022).
- 740 [7] J. Bosbach, S. Weiss, and G. Ahlers, Plume fragmentation by bulk interactions in turbulent rayleigh-
741 Bénard convection, [Phys. Rev. Lett.](#) **108**, 054501 (2012).
- 742 [8] X.-D. Shang, X.-L. Qiu, P. Tong, and K.-Q. Xia, Measured local heat transport in turbulent rayleigh-

- 743 Bénard convection, [Phys. Rev. Lett. **90**, 074501 \(2003\)](#).
- 744 [9] Q. Zhou, C. Sun, and K.-Q. Xia, Morphological evolution of thermal plumes in turbulent rayleigh-
745 benard convection, [PHYSICAL REVIEW LETTERS **98**, 10.1103/PhysRevLett.98.074501 \(2007\)](#).
- 746 [10] O. Shishkina and C. Wagner, Analysis of sheet-like thermal plumes in turbulent rayleigh-benard
747 convection, [JOURNAL OF FLUID MECHANICS **599**, 383 \(2008\)](#).
- 748 [11] M. S. Emran and J. Schumacher, Conditional statistics of thermal dissipation rate in turbulent rayleigh-
749 benard convection, [EUROPEAN PHYSICAL JOURNAL E **35**, 10.1140/epje/i2012-12108-8 \(2012\)](#).
- 750 [12] A. Belmonte and A. Libchaber, Thermal signature of plumes in turbulent convection: The skewness
751 of the derivative, [Phys. Rev. E **53**, 4893 \(1996\)](#).
- 752 [13] Q. Zhou and K.-Q. Xia, Physical and geometrical properties of thermal plumes in turbulent
753 rayleigh–bénard convection, [New Journal of Physics **12**, 075006 \(2010\)](#).
- 754 [14] E. S. Ching, H. Guo, X. Shang, P. Tong, and K.-Q. Xia, Extraction of plumes in turbulent thermal
755 convection, [Physical Review Letters **93** \(2004\)](#).
- 756 [15] S.-D. Huang, M. Kaczorowski, R. Ni, and K.-Q. Xia, Confinement-induced heat-transport enhancement
757 in turbulent thermal convection, [Phys. Rev. Lett. **111**, 104501 \(2013\)](#).
- 758 [16] E. P. van der Poel, R. Verzicco, S. Grossmann, and D. Lohse, Plume emission statistics in turbulent
759 Rayleigh-Bénard convection, [Journal of Fluid Mechanics **772**, 5 \(2015\)](#).
- 760 [17] S.-Q. Zhou, Y.-C. Xie, C. Sun, and K.-Q. Xia, Statistical characterization of thermal plumes in turbulent
761 thermal convection, [PHYSICAL REVIEW FLUIDS **1**, 10.1103/PhysRevFluids.1.054301 \(2016\)](#).
- 762 [18] V. T. Vishnu, A. K. De, and P. K. Mishra, Statistics of thermal plumes and dissipation rates in turbulent
763 rayleigh-benard convection in a cubic cell, [INTERNATIONAL JOURNAL OF HEAT AND MASS
764 TRANSFER **182**, 10.1016/j.ijheatmasstransfer.2021.121995 \(2022\)](#).
- 765 [19] P. P. Shevkar, R. Vishnu, S. K. Mohanan, V. Koothur, M. Mathur, and B. A. Puthenveetil, On separating
766 plumes from boundary layers in turbulent convection, [Journal of Fluid Mechanics **941**, A5 \(2022\)](#).
- 767 [20] F. Chilla and J. Schumacher, New perspectives in turbulent rayleigh-benard convection, [EUROPEAN
768 PHYSICAL JOURNAL E **35**, 10.1140/epje/i2012-12058-1 \(2012\)](#).
- 769 [21] J. Lumley, The structure of inhomogeneous turbulent flows, in *Atmospheric Turbulence and Radio
770 Wave Propagation*, edited by A. Iaglom and V. Tatarski (Nauka, Moscow, 1967) pp. 221–227.
- 771 [22] J. Bailon-Cuba, M. S. Emran, and J. Schumacher, Aspect ratio dependence of heat transfer and large-
772 scale flow in turbulent convection, [Journal of Fluid Mechanics **655**, 152 \(2010\)](#).
- 773 [23] N. Foroozani, J. J. Niemela, V. Armenio, and K. R. Sreenivasan, Reorientations of the large-scale flow

- 774 in turbulent convection in a cube, *Physical Review E* **95**, 033107 (2017).
- 775 [24] B. Podvin and A. Sergent, A large-scale investigation of wind reversal in a square Rayleigh-Bénard
776 cell, *Journal of Fluid Mechanics* **766**, 172 (2015).
- 777 [25] B. Podvin and A. Sergent, Precursor for wind reversal in a square Rayleigh-Bénard cell, *Physical*
778 *Review E* **95** (2017).
- 779 [26] L. Soucasse, B. Podvin, Ph. Rivière, and A. Soufiani, Proper orthogonal decomposition analysis
780 and modelling of large-scale flow reorientations in a cubic Rayleigh-Bénard cell, *Journal of Fluid*
781 *Mechanics* **881**, 23 (2019).
- 782 [27] L. Soucasse, B. Podvin, Ph. Rivière, and A. Soufiani, Low-order models for predicting radiative
783 transfer effects on Rayleigh-Bénard convection in a cubic cell at different rayleigh numbers, submitted
784 to *Journal of Fluid Mechanics* **917**, A5 (2021).
- 785 [28] P. J. Olesen, A. Hodžić, S. J. Andersen, N. N. Sørensen, and C. M. Velte, Dissipation-optimized proper
786 orthogonal decomposition, *Physics of Fluids* **35**, 015131 (2023).
- 787 [29] P. J. Olesen, L. Soucasse, B. Podvin, and C. M. Velte, Dissipation-based proper orthogonal decompo-
788 sition of turbulent rayleigh-bénard convection flow (2024).
- 789 [30] T. L. Griffiths and M. Steyvers, A probabilistic approach to semantic representation, in *Proceedings of*
790 *the 24th Annual Conference of the Cognitive Science Society* (2002).
- 791 [31] D. Blei, A. Ng, and M. Jordan, Latent Dirichlet allocation, *Journal of Machine Learning Research* **3**,
792 993 (2003).
- 793 [32] L. Fery, B. Dubrulle, B. Podvin, F. Pons, and D. Faranda, Learning a weather
794 dictionary of atmospheric patterns using latent dirichlet allocation, *Geophys-*
795 *ical Research Letters* **49**, e2021GL096184 (2022), e2021GL096184 2021GL096184,
796 <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2021GL096184>.
- 797 [33] S. Xin and P. Le Quéré, An extended Chebyshev pseudo-spectral benchmark for the 8:1 differentially
798 heated cavity, *Numerical Methods in Fluids* **40**, 981 (2002).
- 799 [34] S. Xin, J. Chergui, and P. Le Quéré, 3D spectral parallel multi-domain computing for natural convec-
800 tion flows, in *Parallel Computational Fluid Dynamics, Lecture Notes in Computational Science and*
801 *Engineering book series*, Vol. 74, edited by Springer (2008) pp. 163–171.
- 802 [35] O. Shishkina, R. J. A. M. Stevens, S. Grossmann, and D. Lohse, Boundary layer structure in turbulent
803 thermal convection and its consequences for the required numerical resolution, *New Journal of Physics*
804 **12**, 075022 (2010).

- 805 [36] M. Delort-Laval, L. Soucasse, P. Rivière, and A. Soufiani, Rayleigh–Bénard convection in a cubic cell
806 under the effects of gas radiation up to $Ra=10^9$, *Int. J. Heat Mass Transfer* **187**, 122453 (2022).
- 807 [37] D. Puigjaner, J. Herrero, C. Simó, and F. Giralt, Bifurcation analysis of steady Rayleigh–Bénard
808 convection in a cubical cavity with conducting sidewalls, *Journal of Fluid Mechanics* **598**, 393 (2008).
- 809 [38] P. Holmes, J. Lumley, G. Berkooz, and C. Rowley, *Turbulence, Coherent Structures, Dynamical Systems*
810 *and Symmetry* (Cambridge University Press, 2002).
- 811 [39] R. Rehurek and P. Sojka, Gensim–python framework for vector space modelling, NLP Centre, Faculty
812 of Informatics, Masaryk University, Brno, Czech Republic **3** (2011).
- 813 [40] V. T. Vishnu, A. K. De, and P. K. Mishra, Dynamics of large-scale circulation and energy transfer
814 mechanism in turbulent rayleigh–bénard convection in a cubic cell, *Physics of Fluids* **32**, 095115
815 (2020).

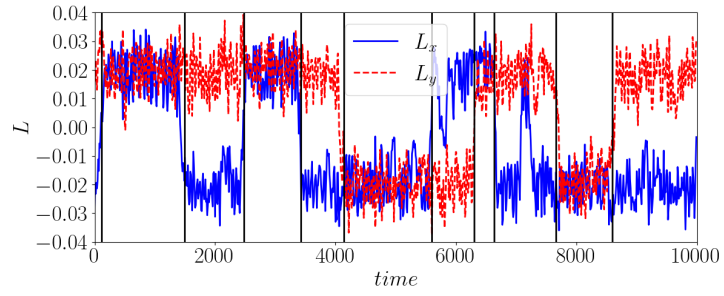
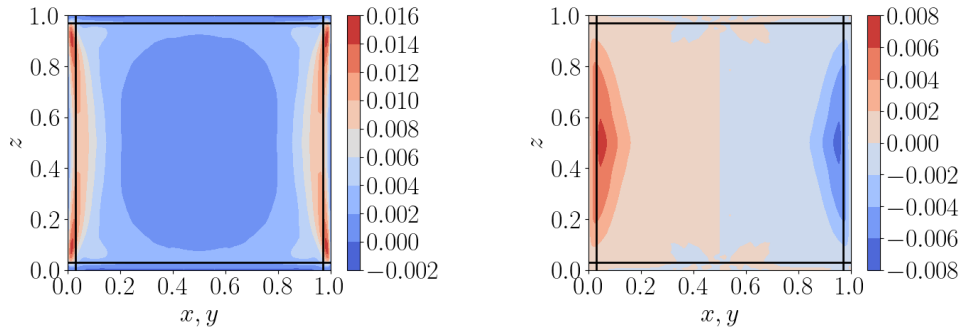
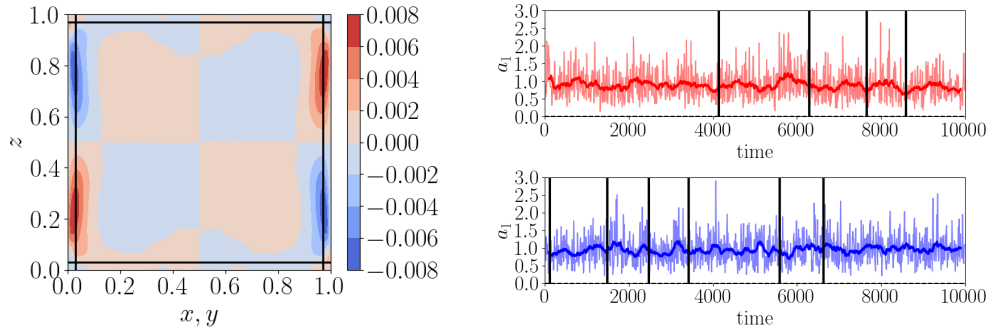


FIG. 1. Evolution of the horizontal components of the angular momentum at $Ra = 10^7$. The vertical black lines correspond to reorientations of the large-scale circulation.

Mode 1



Mode 2



Mode 3

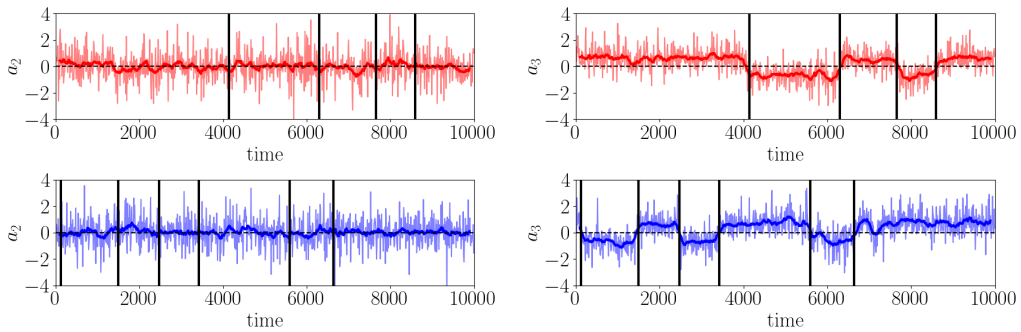


FIG. 2. POD dominant modes and amplitudes in the vertical mid-plane at $Ra = 10^7$. Left: POD modes φ_n , Right: POD amplitudes a_n associated with plane $x = 0.5$ (in blue) and plane $y = 0.5$ (in red). The vertical black lines correspond to changes in the component of the angular momentum. The darker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

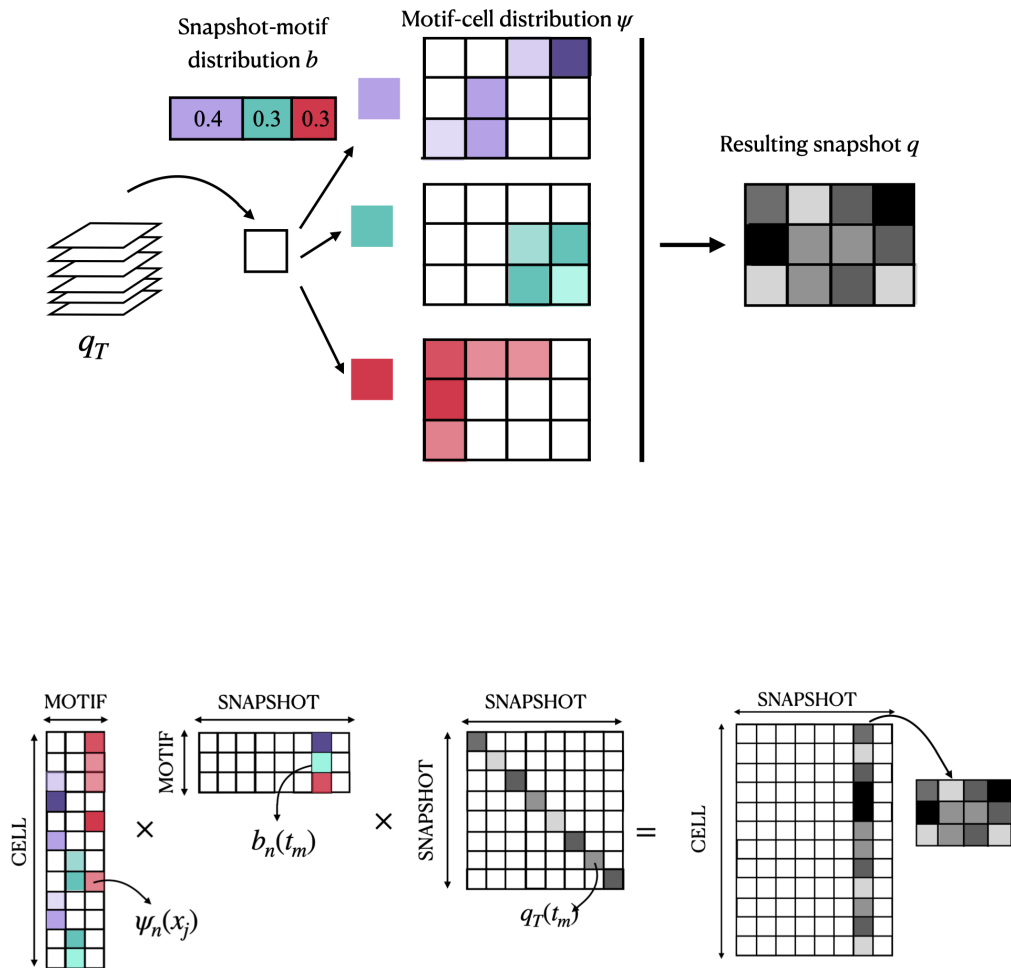


FIG. 3. Schematics of the LDA generative model illustrated here for a field defined on 12 cells and generated with 3 motifs (corresponding to the purple, green and red colours). A snapshot m is represented as a set of integer values defined on an array of cells (see also text). **Top: Probabilistic construction 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 ψ_n . At the end of the process, the number of tokens at cell j yields the value of the field $q(x_j, t_m)$. **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 $q_T(t_m)$.**

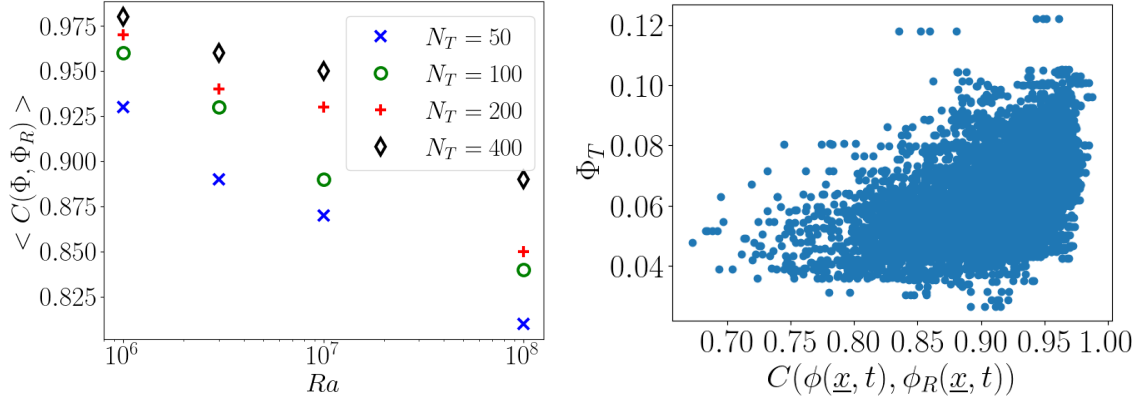


FIG. 4. Left: Instantaneous correlation coefficient between the projected and the true field as a function of the integral convective heat flux for $N_T = 100$ and $Ra = 10^7$. Right: Average correlation coefficient $\langle C(\Phi, \Phi_R) \rangle$ as a function of the Rayleigh number and of the number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

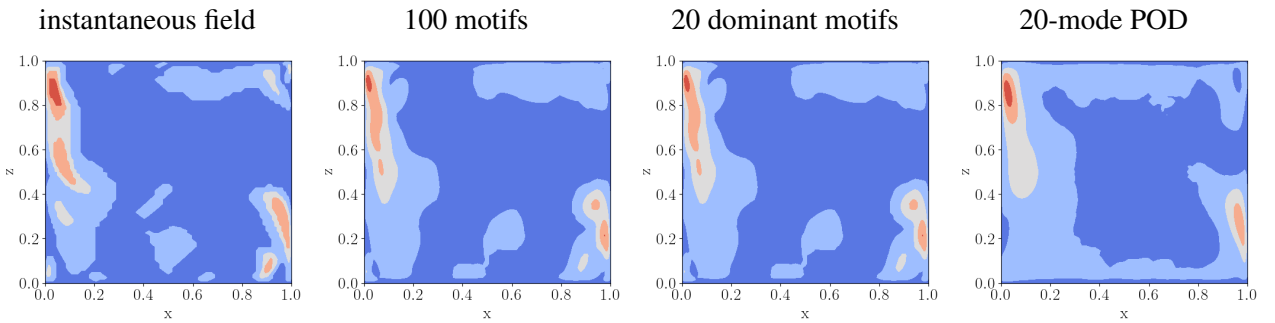


FIG. 5. Example of an instantaneous snapshot and its reconstructions at $Ra = 10^7$. From left to right: original field, LDA-reconstructed field using $N_T = 100$ motifs, LDA-reconstructed field using the 20 (instantaneously) most prevalent motifs, POD-reconstructed field using the 20 (on average) most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

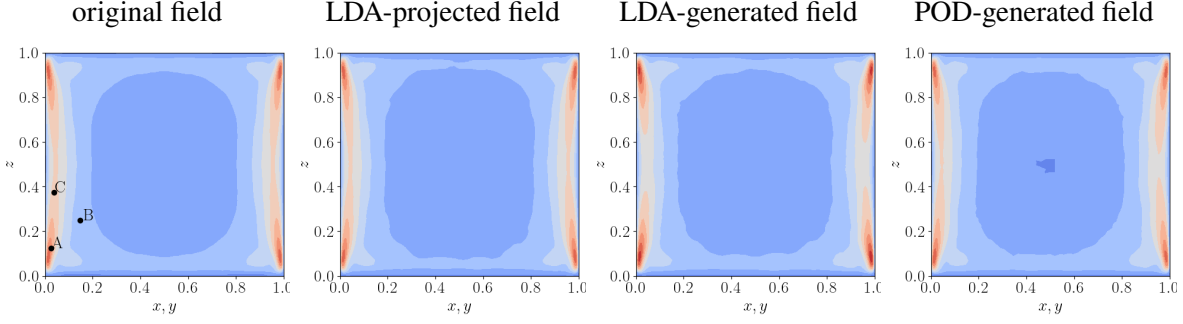


FIG. 6. Time-averaged value of the convective heat flux for different databases at $Ra = 10^7$. From left to right: original fields, LDA-reconstructed (LDA-R) fields using $N_T = 100$ motifs, LDA-generated (LDA-G) fields using $N_T = 100$ motifs, POD-generated (POD-G) fields using 100 modes.

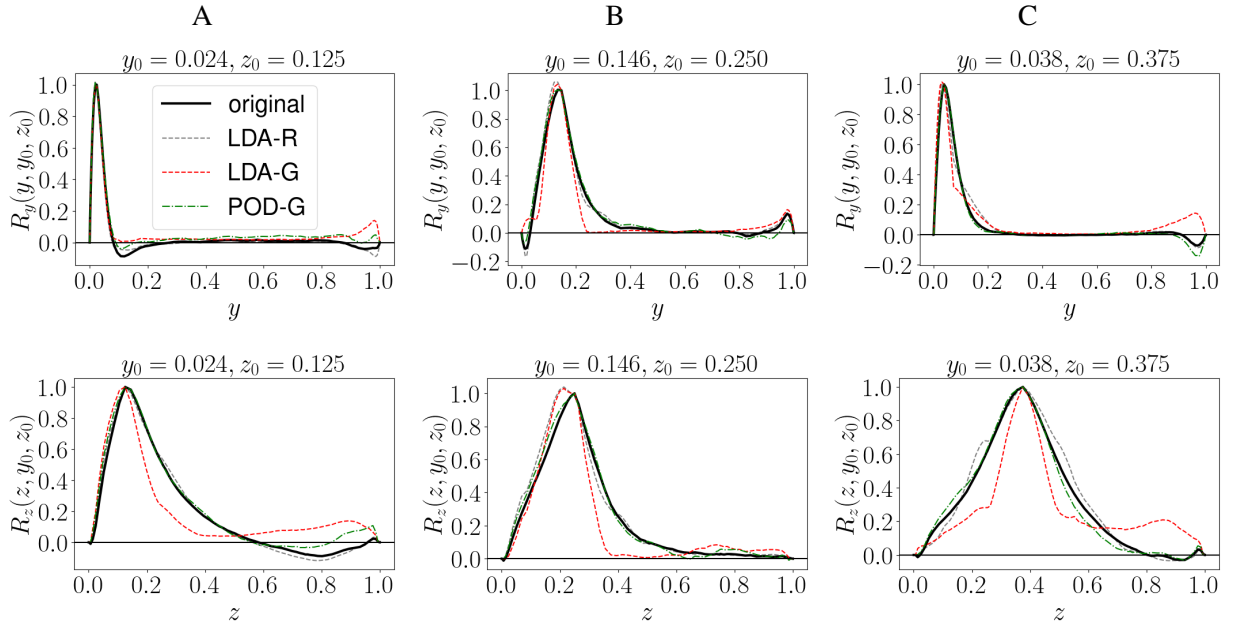


FIG. 7. Autocorrelation of the convective heat flux at selected locations (see Fig. 6).

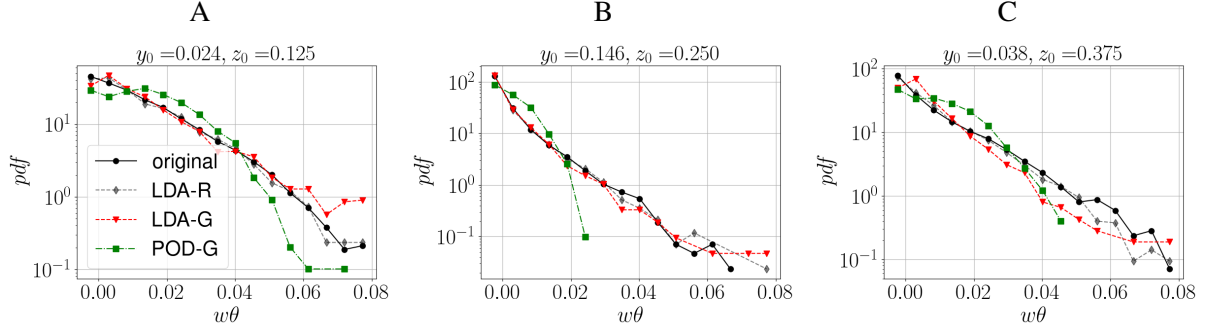


FIG. 8. Probability density function of the convective heat flux at the selected locations indicated in Fig. 6.

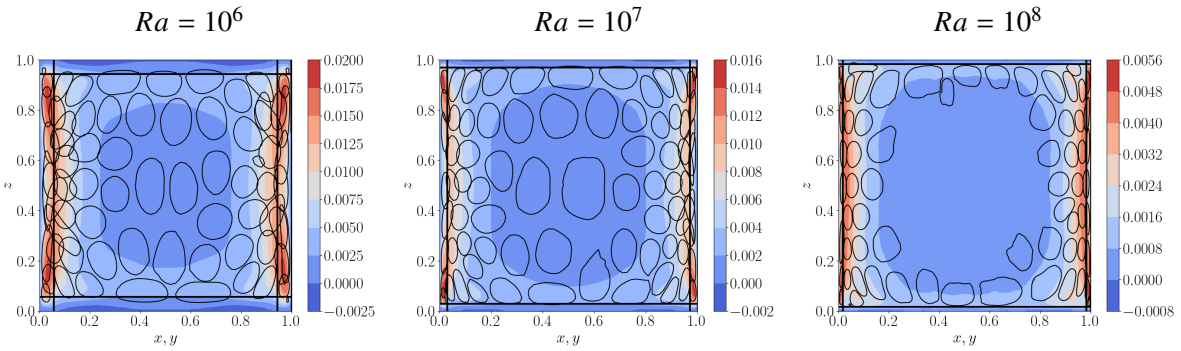


FIG. 9. Spatial distribution of the positive heat flux motifs ψ_n^Φ in the vertical mid-plane for $N_T = 100$. The motifs are materialized by a black line corresponding to a probability contour of $0.606 \psi_n^{max}$. The vertical lines correspond to the boundary layer thickness. The time-averaged convective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

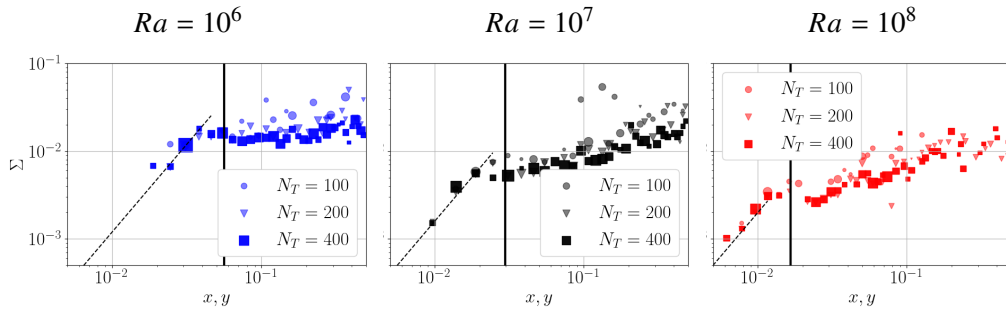


FIG. 10. Distribution of motif areas (see definition in equation (12)) in the vertical mid-plane with the distance from the lateral walls at varying Rayleigh numbers. The size of the symbols shown in the picture is proportional to the fraction of motifs over which the average was performed. The black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

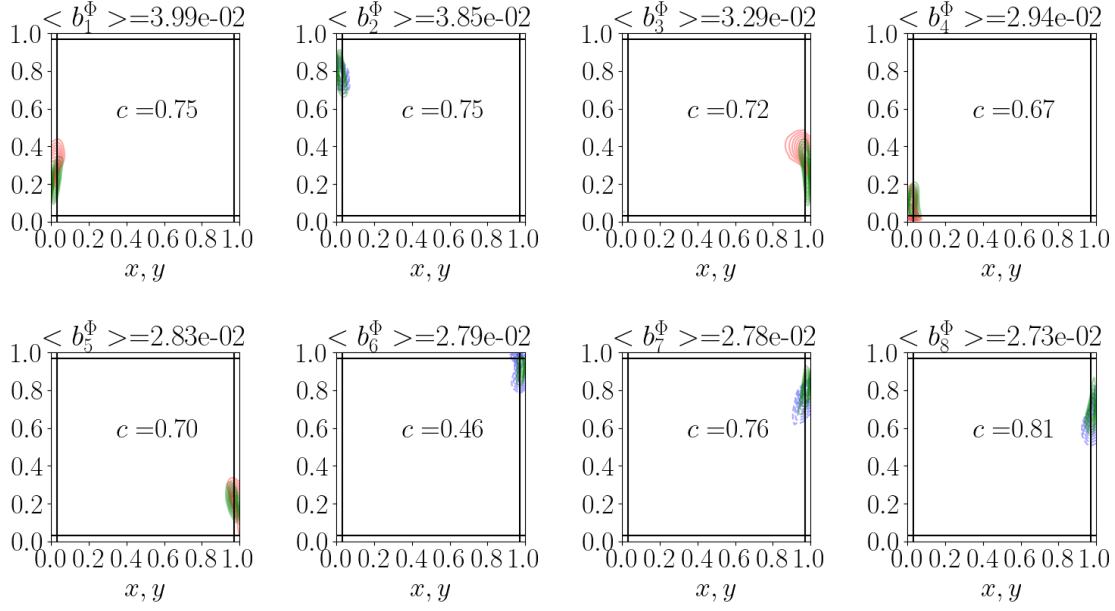


FIG. 11. Dominant heat flux motifs ψ_n^Φ (green lines) ordered by prevalence and associated temperature motifs ψ_n^θ (blue for negative and red for positive fluctuations) at $Ra = 10^7$. Contour levels go from 0.2 to $0.9 \psi_n^{max}$ with increments of $0.1 \psi_n^{max}$. c is the maximum correlation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

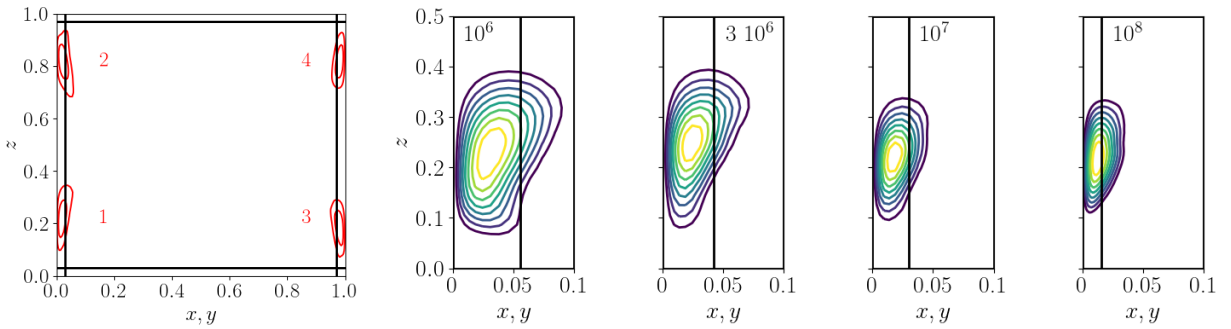


FIG. 12. Left: Dominant heat flux motifs ψ_n^Φ at Rayleigh numbers $Ra = 10^7$ for $N_T = 100$. The contour lines correspond to $0.1 \psi^{max}$ and $0.3 \psi^{max}$. The motif labels correspond to those of Fig. 13. Right: Characteristic dominant motif at different Rayleigh numbers $N_T = 100$. Isocontours of ψ_1 at $[0.2, 0.3, \dots, 0.9] \psi_1^{max}$. The black lines correspond to the boundary layer thickness.

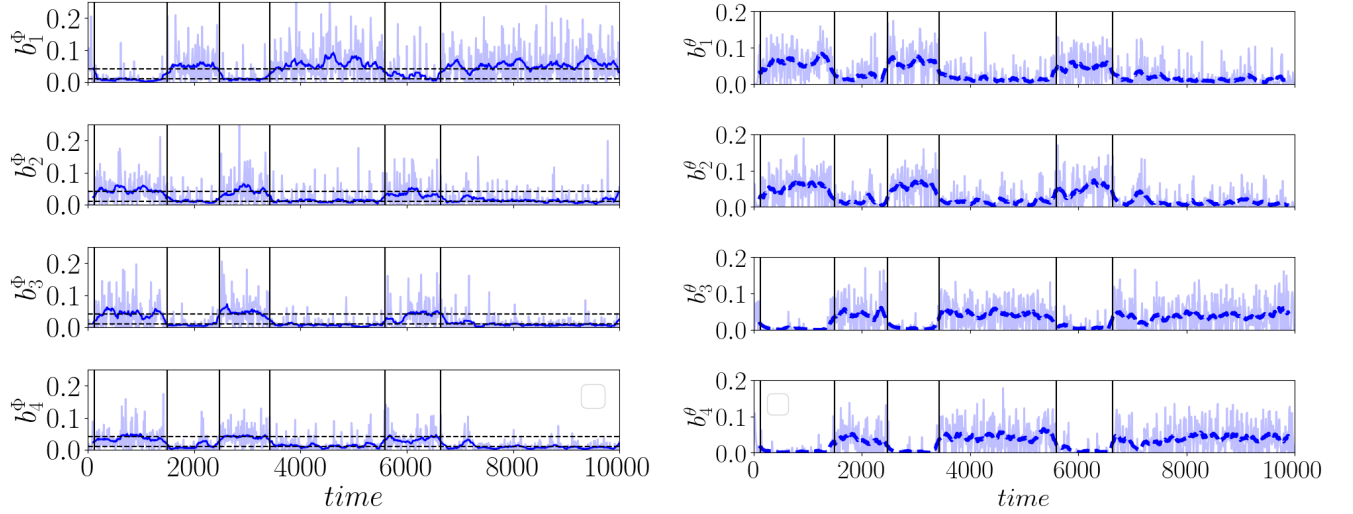


FIG. 13. Evolution of the snapshot-motif distributions b_n^Φ for the four dominant **heat flux** motifs (see Fig. 12 for labels) at $Ra = 10^7$ and for $N_T = 100$. Left: plane $x = 0.5$. Right: plane $y = 0.5$. The thick line corresponds to a moving average over 200 convective units (4 recirculation 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.

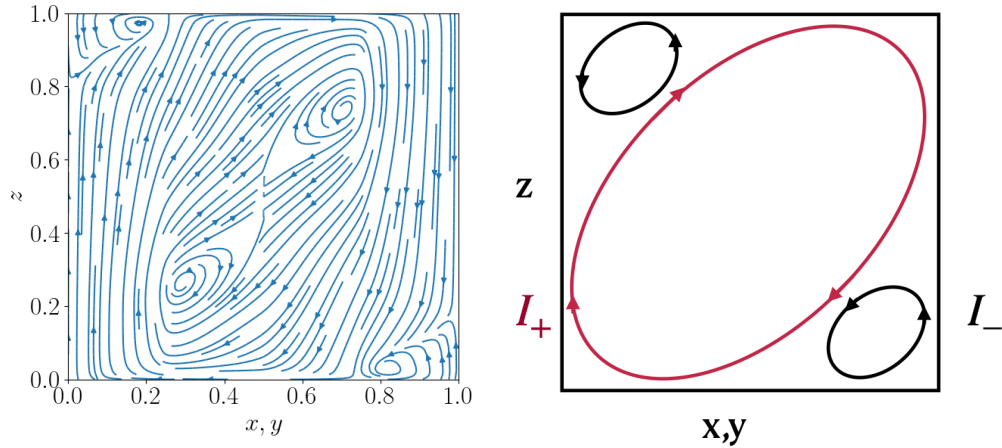


FIG. 14. Left: Streamlines of the flow conditionally averaged on the high weight value of b_1^Φ . Right: Schematics of the cell organization in the vertical mid-plane: the large-scale circulation (in red) corresponds to the I_+ state while the corner structure (in black) corresponds to the I_- state.

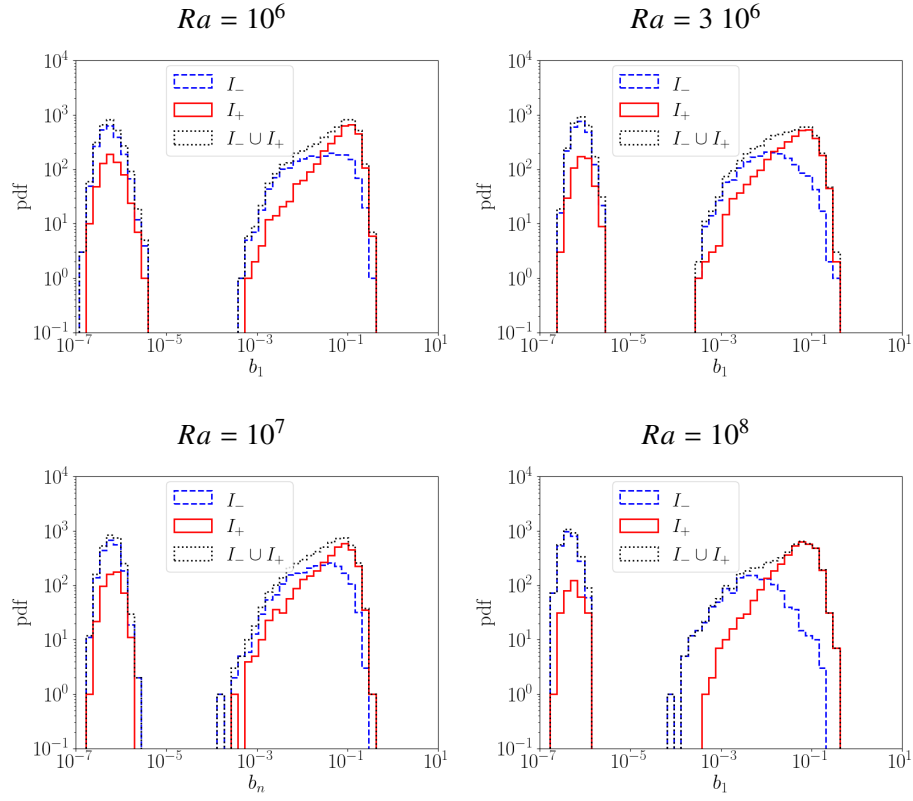


FIG. 15. Distribution of the dominant motif weight b_1^Φ for different Rayleigh numbers and $N_T = 100$.

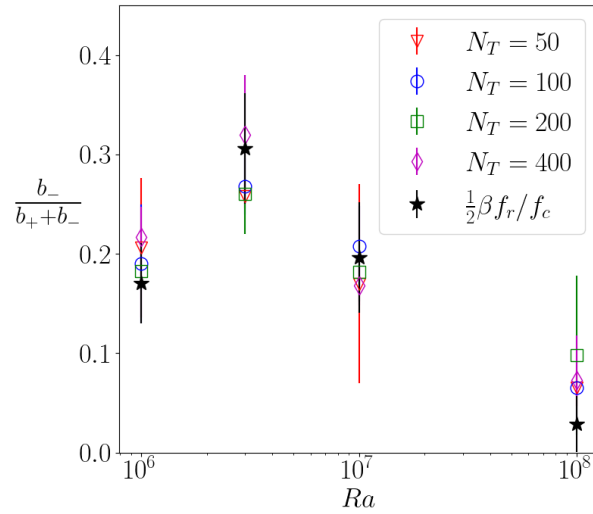


FIG. 16. Probability $p(T_+ > T_-)$ (see text) and comparison with ratio of reorientation to recirculation time scale at different Rayleigh numbers - the rescaling factor is $\beta = 5.6$.

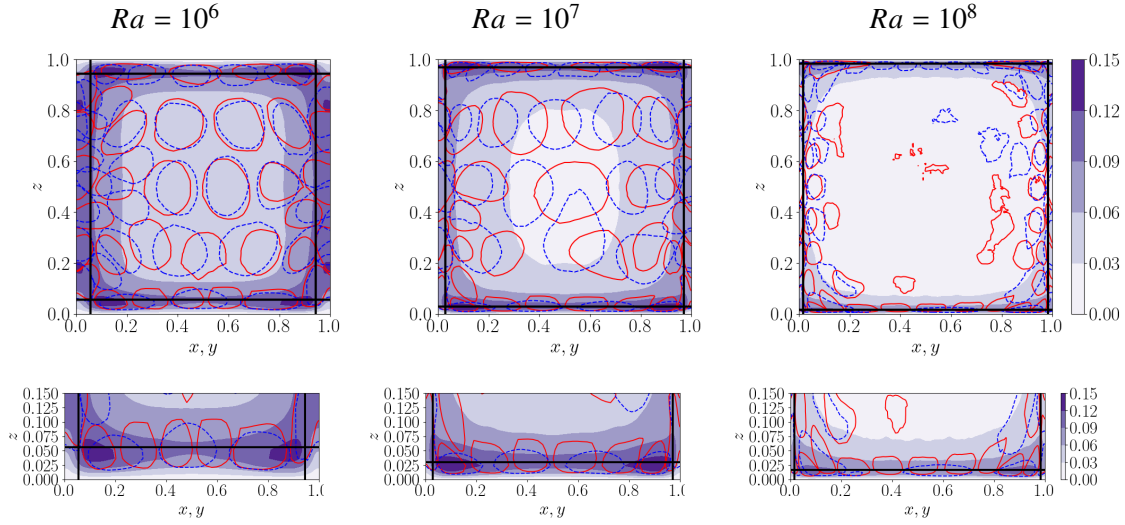


FIG. 17. Top row: Distribution of temperature motifs ψ_n^θ in the cell mid-plane at different Rayleigh numbers; The motifs are materialized by a black line corresponding to a probability contour of $0.606 \psi_n^{max}$. Contours of the time-averaged variance are represented in the background. Bottom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

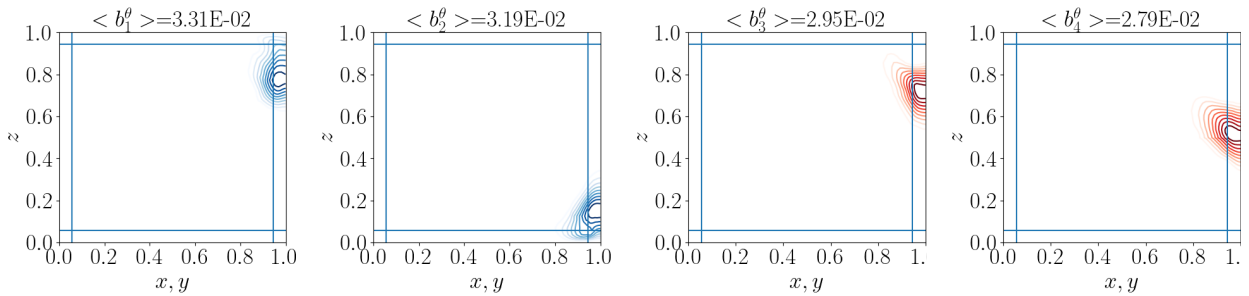


FIG. 18. First four dominant temperature motifs ψ_n^θ at $Ra = 10^6$.

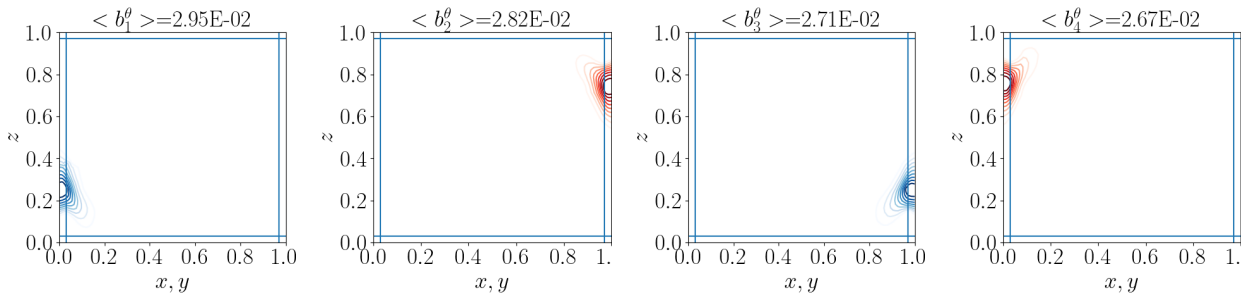


FIG. 19. First four dominant temperature motifs ψ_n^θ at $Ra = 10^7$.

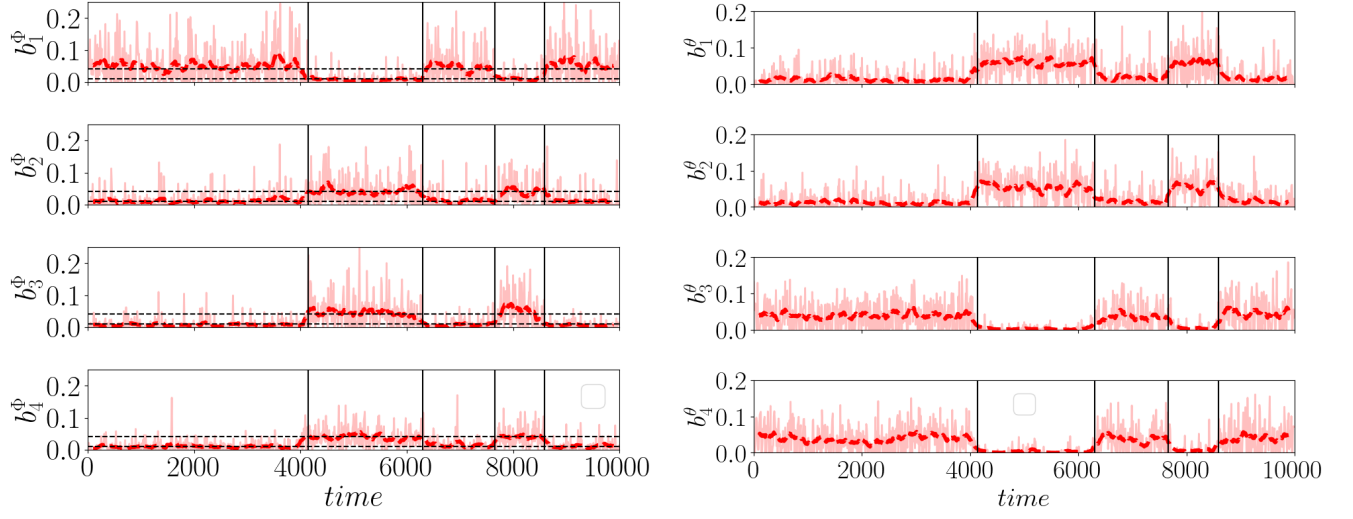


FIG. 20. First four dominant temperature motif weights b_n^θ at $Ra = 10^7$. Left: plane $x = 0.5H$. Right: plane $y = 0.5H$. The thick line corresponds to a moving average over 200 convective units (4 recirculation times T_C). The vertical lines correspond to the changes in angular momentum.

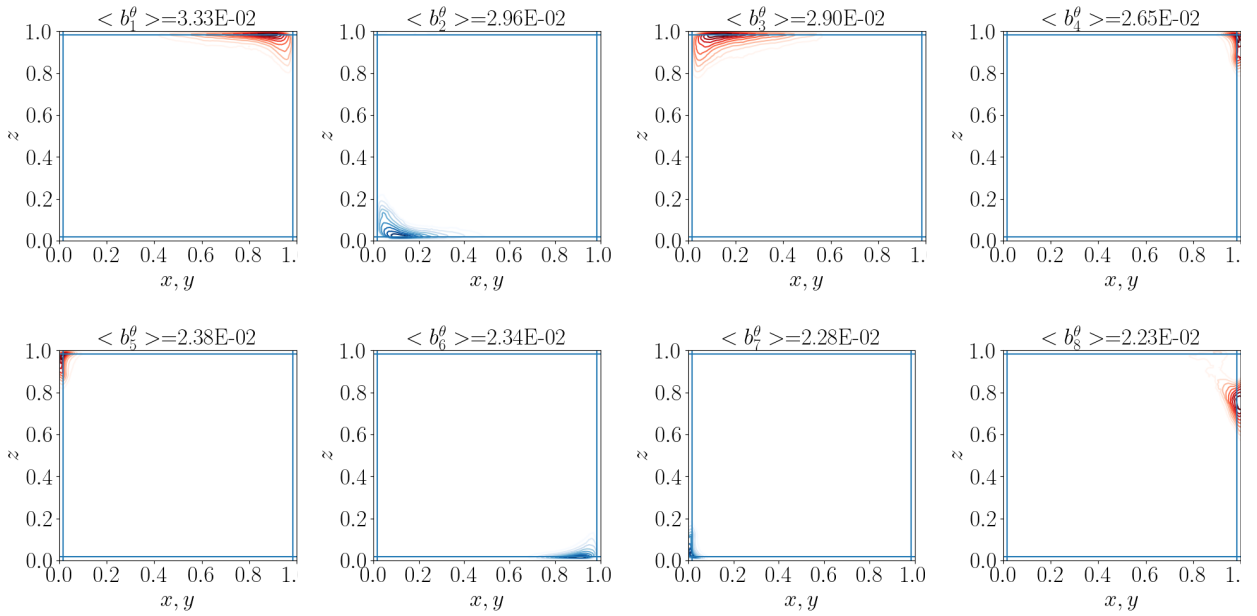


FIG. 21. First eight dominant temperature motifs ψ_n^θ at $Ra = 10^8$.

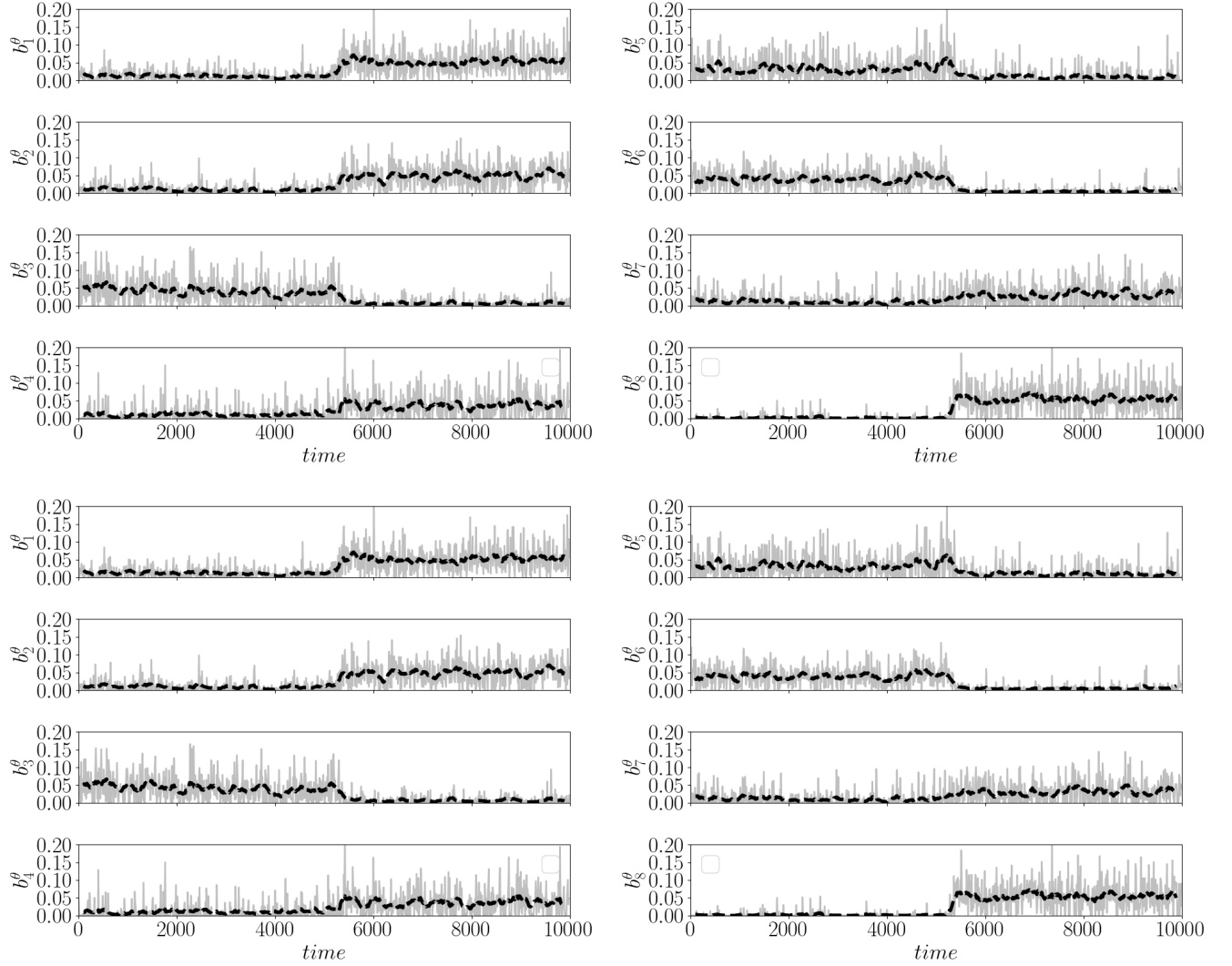


FIG. 22. First eight temperature motif weights b_n^θ at $Ra = 10^8$ in the plane $x = 0.5$.

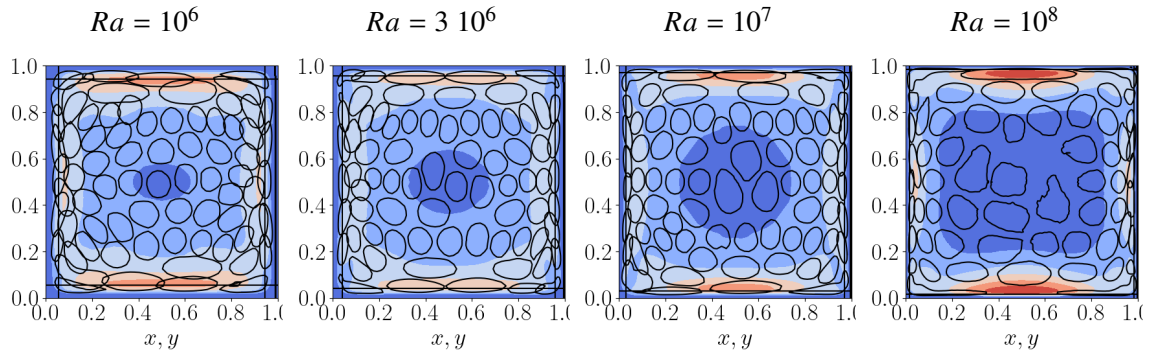


FIG. 23. Spatial distribution of kinetic energy motifs in the cell mid-plane at different Rayleigh numbers. The motifs are materialized by a black line corresponding to a probability contour of $0.606 \psi_n^{max}$. Contours of the time-averaged kinetic energy are represented in the background.

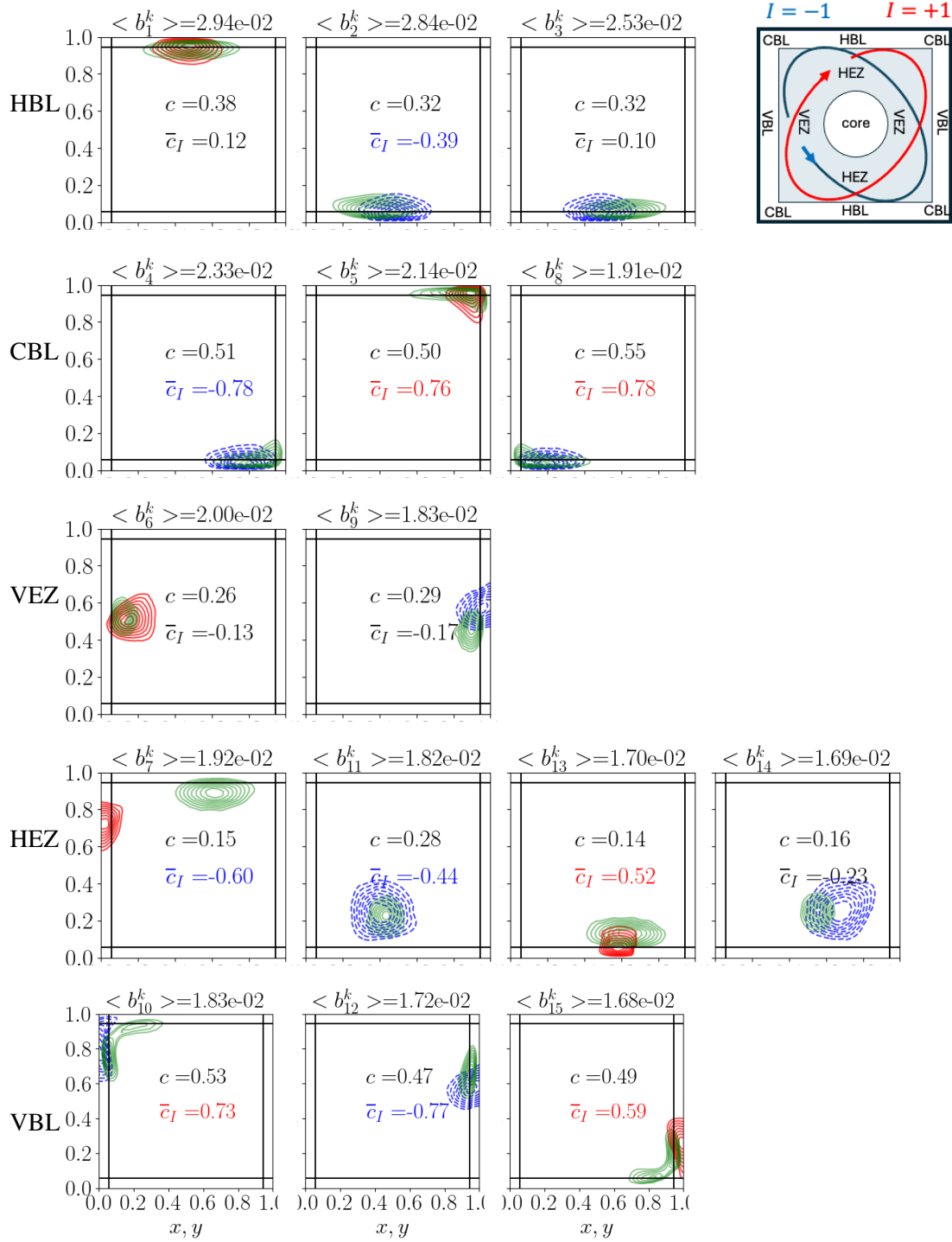


FIG. 24. Dominant kinetic energy motifs ψ_n^k at $Ra = 10^6$ (green lines) ordered by prevalence and location as indicated at top right. Temperature motifs ψ_n^θ with the highest correlation coefficient c are shown in blue (resp. red) for negative (resp. positive) fluctuations. Motif contour levels range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

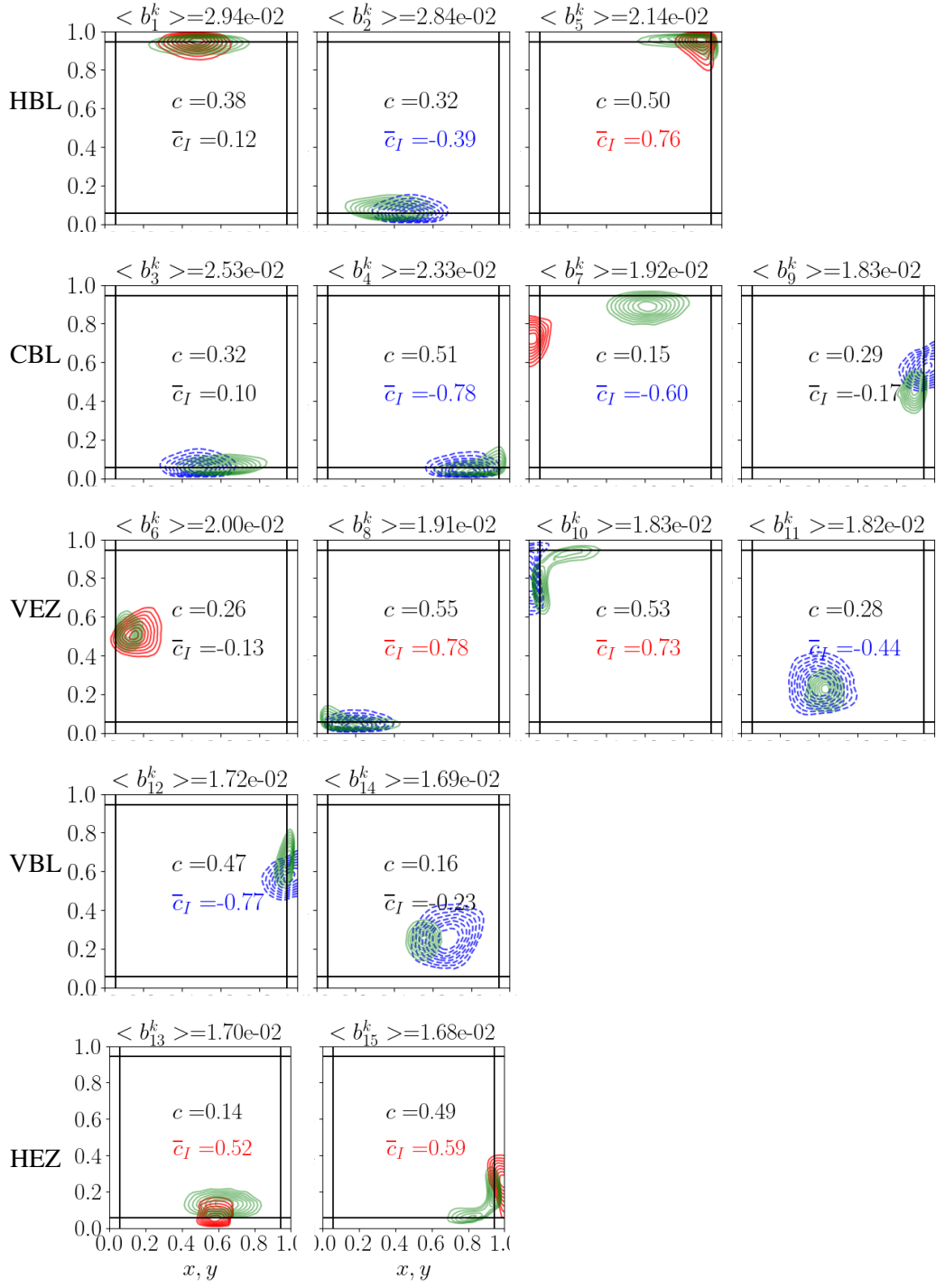


FIG. 25. Comparison between kinetic energy and temperature motifs at $Ra = 10^7$. See legend of Fig. 24.

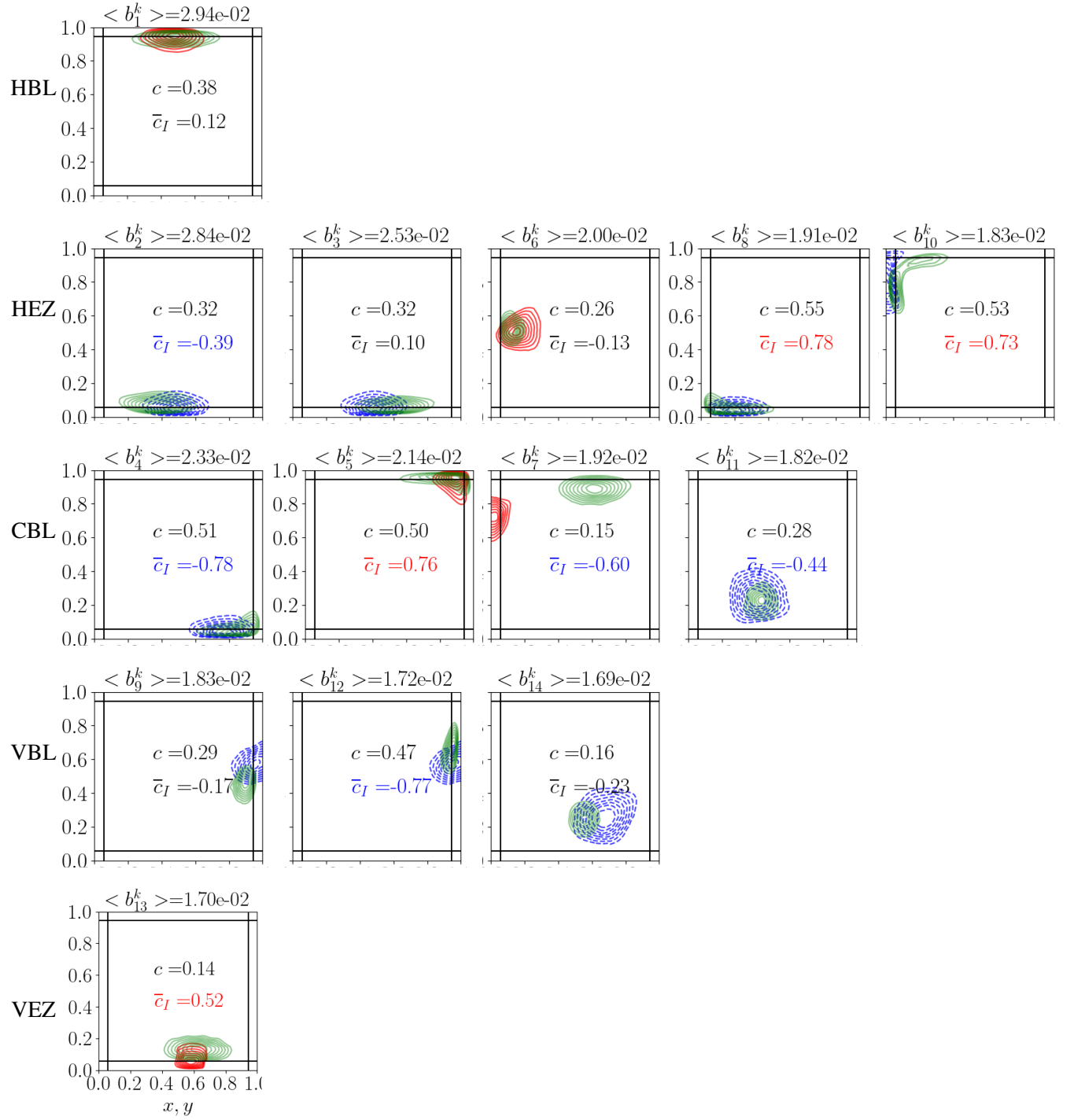


FIG. 26. Comparison between kinetic energy and temperature motifs at $Ra = 10^8$. See legend of Fig. 24.