Boundary-Layer Meteorology (2020) 176:1–12
https://doi.org/10.1007/s10546-020-00517-w

RESEARCH LETTER

The Structure of the Convective Boundary Layer as Deduced from Topological Invariants

José Licón-Saláiz[2] · Cedrick Ansorge[1] · Yaping Shao[1] · Angela Kunoth[2]

Received: 30 October 2019 / Accepted: 24 March 2020 / Published online: 23 April 2020

Abstract
We study the convective boundary layer (CBL) through low-order topological properties of updrafts and downdrafts, that is, based solely on the sign of the vertical velocity. The geometric representation of the CBL as a pair of two-dimensional cubical complexes, one each for updrafts and downdrafts, is exemplarily obtained from two simulations of the CBL, a realistic daily cycle and an idealized quasi-steady CBL growing into linear stratification. Each cubical complex is defined as a set of grid cells that have the same sign of vertical velocity, either positive or negative. Low-order topological invariants, namely the Betti numbers of the cubical complexes, are found to capture key aspects of the boundary-layer organization and evolution over the diurnal cycle. An unsupervised-learning algorithm is trained using the topological invariants in order to classify the spatio–temporal evolution of convection over a whole day. The successful classification of the CBL by using this approach illustrates the potential of such simplified representation of turbulent flow for data reduction and boundary-layer parametrization approaches.

1 Introduction

The convective boundary layer (CBL) is a canonical case of geophysical turbulence that is described in terms of bulk quantities (Deardorff 1970a, b), mixed-layer and surface-layer profiles based on similarity theory (Kaimal et al. 1976; Sorbjan 1986; Mellado 2012; Mellado et al. 2016; Garcia and Mellado 2014), turbulent fluxes (Fernandes and Adrian 2002), and spectra (Finnigan and Kaimal 1994; Mellado et al. 2016). Resulting low-dimensional models describe the temporal evolution of CBL height and temperature (Fedorovich and Mironov 1995; Pino et al. 2006). Also the vertical profiles of mean quantities are determined relatively well as manifest in the great success of mixed-layer models of the CBL (Lilly 1968; van Heerwaarden et al. 2009; Vilà-Guerau de Arellano et al. 2012). This works well because the CBL is primarily driven by the energy that becomes available due to the insolation of the

✉ Cedrick Ansorge
cedrick@posteo.de

1 Institut für Geophysik und Meteorologie, Universität zu Köln, Pohligstr. 3, 50969 Cologne, Germany

2 Mathematisches Institut, Universität zu Köln, 50923 Cologne, Germany

surface; turbulence distributes energy in the vertical and across the range of scales that is available given the geometry of the flow.

Whether global operations, such as averaging or spectral transforms, appropriately represent the actual physical mechanisms in buoyant flow is questionable: the CBL is driven by local density differences creating an instability with respect to the environment. Limitations of conventional averaging and spectral approaches have become evident when the interaction of complex surface patterns with the atmosphere aloft is considered (Liu et al. 2017), but also in the context of deviations from similarity theory (Fodor et al. 2019). A common approach to simplifying the complex morphology of the CBL is to consider the CBL as composed of individual plumes and interspersed downdrafts, where a plume is a connected region with positive buoyancy. Indeed, such a simplification of the complex flow improves our capability to understand and model the CBL (Adrian 2007; Shah and Bou-Zeid 2014).

Coherent structures are often studied based on spectral transformation or conditional averaging—not necessarily optimal given the well-defined confinement of plumes and surrounding regions with sinking air. The emerging field of computational topology makes it possible to describe coherent structures based on their connectivity, and we pursue this approach here to yield a novel representation of CBL turbulence. The idea of building a geometric object from the scalar fields produced by the numerical model has been used in numerous applications of topology to data analysis (Wasserman 2018). Krishan et al. (2007) investigate non-Boussinesq effects in Rayleigh–Bénard convection, a first application that illustrates the utility of topological characterization for turbulent fluid flow. In fact, topological invariants associated with the complex shapes that emerge naturally from the physical systems under consideration evince strong regularity. Sometimes, this allows the establishment of relationships between numerical values of topological invariants and dynamical properties of the system, in spite of its seemingly chaotic nature.

Here, the horizontally homogeneous CBL is topologically characterized in terms of its vertical and temporal evolution—building upon the geometric representation of the three- and four-dimensional flow structures. We exemplarily base our analysis on the sign of the vertical velocity and decompose the flow into updrafts and downdrafts, a well-defined and physically meaningful binary partition of the flow domain. This geometric representation of the updraft and downdraft domains is characterized by elementary topological descriptors measuring the connectivity (order-zero Betti number β_0) and interspersion (order-one Betti number β_1) of the updrafts and downdrafts. We find that the Betti numbers separate the different subpartitions of the CBL—as it evolves over the course of a day—according to the different turbulence regimes encountered. Based on this separation, we eventually demonstrate the skill of a machine-learning approach for detection of the height-local turbulence regime in the boundary layer, based on the low-order topological characterization of the vertical velocity field only.

2 Homological Representation of the Convective Boundary Layer

Topological analysis requires a geometric representation of the problem; primarily, the flow has to be partitioned. We choose here the vertical velocity w to partition the flow domain into updrafts and downdrafts. While $w = 0$ would provide a sharp binary partition, we use $|w| > \varepsilon$, with $\varepsilon > 0$, a quasi-binary partition where the region around $w = 0$ is excluded for not being physically uniquely attributable. Moreover, since the goal of our method is to characterize the spatial patterns produced by turbulent convection, the areas with $|w| < \varepsilon$ can

be disregarded as corresponding to small fluctuations. Our criterion for choosing the value of ε has thus been to produce a symmetric interval about zero, which contains a small fraction of the total data points, in order to minimize any potential impact on the final results. In this study, we use $\varepsilon = 0.01\,\mathrm{m\,s^{-1}}$ for the LES data and $\varepsilon = 0.01 L_0 N$ for the DNS data, where L_0 is the Ozmidov scale as defined in Garcia and Mellado (2014, their Eq. 3) and N is the Buoyancy frequency imposed through the external stratification. After thresholding the data we are left with two distinct sets of grid points, which correspond to updrafts and downdrafts. We use these points to define the cubical complexes

$$\mathcal{C}^{\pm}(z,t) \equiv \left\{ \left[i - \tfrac{1}{2}, i + \tfrac{1}{2} \right] \times \left[j - \tfrac{1}{2}, j + \tfrac{1}{2} \right] \ \forall\ (i,j) \in \mathcal{P}^{\pm} \right\} \tag{1a}$$

with

$$\mathcal{P}^{\pm}(z,t) \equiv \left\{ (i,j) \mid (x_i, y_i) \in \mathcal{M}^{\pm}(z,t) \right\}, \tag{1b}$$

$$\mathcal{M}^{+}(z,t) \equiv \left\{ (x_i, y_j, z, t) \in \Omega \mid w(x_i, y_j, z, t) > \varepsilon \right\}, \tag{1c}$$

$$\mathcal{M}^{-}(z,t) \equiv \left\{ (x_i, y_j, z, t) \in \Omega \mid w(x_i, y_j, z, t) < -\varepsilon \right\}, \tag{1d}$$

where Ω is the flow domain and $(i, j) \in \mathbb{N}^2$ are indices on the regular, Cartesian grid, i.e., $i \in \{1 \ldots N_x\}$ and $j \in \{1 \ldots N_z\}$ where $N_{(\cdot)}$ stands for the number of collocation points along the axis $(\cdot)$. That is, $\mathcal{M}^+$ is the set of data points contained in updraft motions and $\mathcal{P}^+$ refers to the corresponding set of indices on the two-dimensional horizontal slice; eventually, $\mathcal{C}^+$ is the cubical complex formed by the unit cubes around all indices contained in $\mathcal{P}^+$. Each of these unit cubes is a plane grid cell, since we are working on two-dimensional subdomains. The word "unit" here refers to the fact that we treat the data as if they were defined on an integer lattice; the grid information gets encoded in cell indices.

Given the geometric representation of the CBL by cubical complexes $\mathcal{C}^{\pm}(t, z)$, we seek a quantitative description of its topological properties of physical relevance. One such measure is given by the Betti numbers β_i that describe the connectivity of a space at different levels: β_0 counts the independent connected components, β_1 counts the loops or "holes" (e.g., $\beta_1 = 1$ for a circle), β_2 counts the voids or cavities ($\beta_2 = 1$ for a hollow sphere). Higher-dimensional analogues of these structures exist, but need not be considered here.[1] We compute here the Betti numbers $\beta_0^{\pm}$ and $\beta_1^{\pm}$ for the two-dimensional horizontal cross-sections $\mathcal{C}^{\pm}(z, t)$ only, but point out that it is possible to understand an object of dimension n by its components of dimension $n - 1$ together with the relationship between them—a well-established concept exploited in tomographic methods. In the two-dimensional case of $\mathcal{C}^{\pm}(t, z)$, the zeroth Betti number β_0 is the number of connected components in $\mathcal{C}^{\pm}$, i.e., the number of updrafts and downdrafts, and the first Betti number β_1 corresponds to the number of "holes" or loops in these components, i.e., the interspersion of updrafts by downdrafts (β_1^+) and vice versa (β_1^-). In our analysis, we will resort to the logarithmic ratio of these numbers, $\ell^{\pm} = \ln(\beta_1^{\pm}/\beta_0^{\pm})$, that is a non-dimensional parameter characterizing the relative interspersion of updrafts by downdrafts and vice versa. The case $\beta_1^{\pm} \gg \beta_0^{\pm}$ would indicate significant interspersion, as one of the domains is made up by few connected components, with many holes in them that correspond to components of the complementary domain. On the other hand, $\beta_1^{\pm} \ll \beta_0^{\pm}$ signals the relative absence of interspersion for the corresponding domain.

[1] The nth Betti number of a cubical complex is formally defined as the rank of the nth homology group of that cubical complex. [computed here using CHOMP, a C++ code library (Mischaikow et al. 2019)].

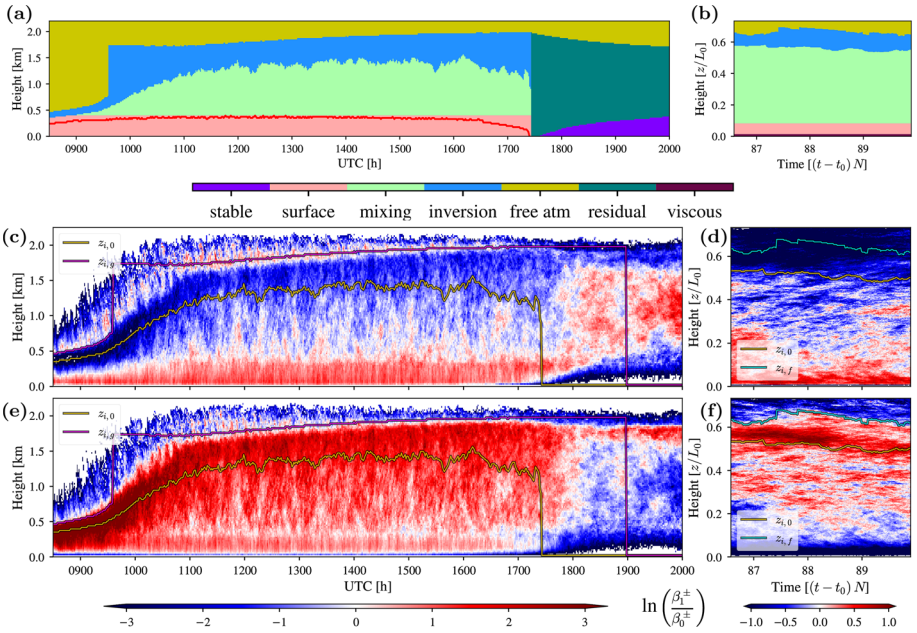

Fig. 1 **a** and **b** show classification of the different CBL subregions by bulk analysis of the flow (not invoking topological descriptors). The thick red line in panel (**a**) corresponds to the surface-layer height estimated by the maximum buoyancy gradient. Contour plot of the height-time section of $\ln(\beta_1^+/\beta_0^+)$ (panel **c**), $\ln(\beta_1^-/\beta_0^-)$ (panel **e** for LES, and corresponding profiles from the DNS (panels **d** and **f**)

3 The Convective Boundary Layer as Described by Betti Numbers of Horizontal Slabs

We use data from a realistic large-eddy simulation (LES) for a sheared CBL observed on 5 August, 2009 (based on a radio sounding at 0800 UTC, Case SP4 of Liu et al. 2017) and direct numerical simulation (DNS) data from a shear-free CBL growing into a linearly stratified atmosphere (Garcia and Mellado 2014, Case Re100). The characterization of the CBL by ℓ^+ and ℓ^- is given in Fig. 1c–f; for comparison, Fig. 1a, b shows the partitioning obtained by computing bulk statistics of the flow where the following criteria were used to obtain the partitioning for the LES dataset (Fig. 1a):

- The surface-layer height is determined by finding the height at which the mean buoyancy gradient vanishes. Since this value is essentially constant throughout most of the CBL regime, we use the maximum value it acquires over the course of the day as surface-layer height where the evolution over time is shown in Fig. 1a and illustrates the appropriateness of our approximation (differences occur mainly during morning and evening transition where a unique bulk classification becomes challenging anyhow).
- The base of the inversion layer is given by the zero-crossing height $z_{i,0}$ (where the total buoyancy flux becomes negative), and its top is given by the gradient-based height $z_{i,g}$ (where the mean buoyancy gradient is maximized).
- The mixing layer is located between the surface layer and the inversion-layer base, and the free atmosphere is the region above the inversion-layer top.

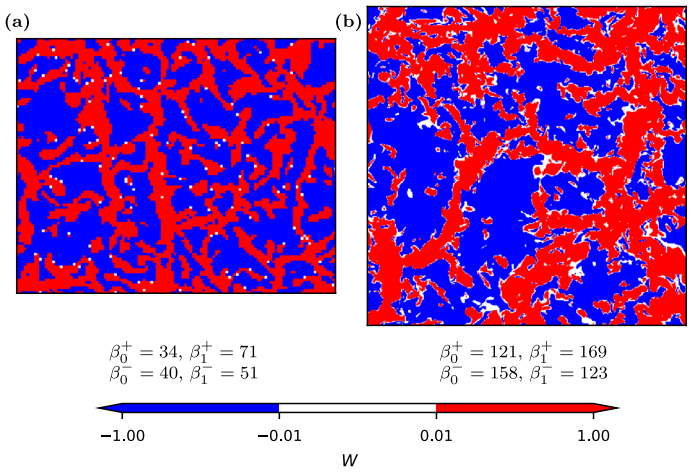

Fig. 2 Two-dimensional horizontal cross-sections of the vertical velocity from the LES (**a**), and DNS (**b**) simulations described in the text. In both cases, the data are taken from the transition region between the surface and mixing layers, and preserve some of the network-like pattern visible in the former. The ternary partitioning described in the text has also been applied, with the updrafts shown in red, the downdrafts in blue, and white representing the region where $|w| < \varepsilon$. The four Betti numbers for each cross-section are also shown

- The evening transition is signaled by the total buoyancy flux becoming positive over the entire vertical domain.
- After this point, the mixing and inversion layers are defined as a layer of residual turbulence, with the height of the free atmosphere modelled by an exponentially decaying function.
- The stable surface layer is defined as the region where mean temperature is lower than the temperature at the height at which the mean buoyancy gradient vanishes.

For the DNS data, an equivalent set of rules is employed as for the convective period of the LES, except that the inversion-layer top is given by the flux-based height $z_{i,f}$ (where total buoyancy flux is minimum). Additionally, we note the existence of a viscous layer underneath the surface layer, directly adjacent to the surface. It is characterized by $\ell^+,\ \ell^- < 0$, with ℓ^- being very small (smaller than in the free atmosphere, i.e., more connected components, less loops). The value of ℓ^+ grows rapidly as the updraft network pattern begins to form and transition occurs to the surface layer. For the LES, we focus now on the quasi-stationary period from 1300 UTC (local time is UTC plus 1 h) to 1700 UTC to facilitate comparison with DNS: most notably, the surface layer in LES has $\ell^+,\ \ell^- > 0$, whereas in DNS it is $\ell^+ > 0$ but $\ell^- < 0$. A positive sign of both ℓ^+ and ℓ^- implies $\beta_1^\pm > \beta_0^\pm$, that is, both the updraft and downdraft domain contain more than one "hole" per connected component, indicating a complex, intertwined network-like pattern. In the surface layer of the DNS, the updrafts feature this pattern too, while downdrafts are mostly limited to acyclic (hole-free) components ($\ell^- < 0$), surrounded by updrafts on all sides—and thus not linked. This difference is due to the presence of shear in the LES disturbing the hexagonal cell pattern of convection. In both the LES and DNS, (ℓ^+, ℓ^-) scatter around the origin in the mixed layer.

Figure 2 shows horizontal slabs of the vertical wind velocity field from the LES (a) and DNS (b). Both slabs are taken from the transition region between the surface and mixing layers, so the characteristic network pattern of updrafts can still be seen, even as these updrafts

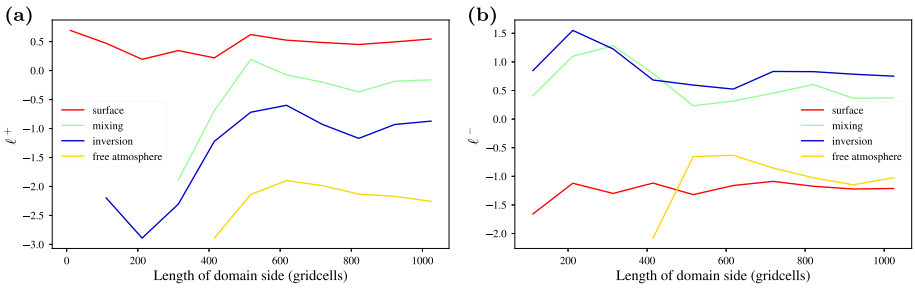

Fig. 3 Values of the log-quotients ℓ^+ (**a**) and ℓ^- (**b**), computed for two-dimensional horizontal sections of the DNS computational domain, visualized as functions of domain size (in grid cells). In each case, a representative section from each of the four main CBL subregions present in the simulation is shown

begin to coalesce into larger plumes. Also visible is the greater interspersion of updrafts and downdrafts present in the LES, which is reflected in the fact that $\ell^- > 0$ here. We also note that the areas with $|w| < \varepsilon$, shown in white in Fig. 2, are not distributed spatially at random but rather occur in the interface between updrafts and downdrafts. Therefore these areas do not contribute significantly to the values of $\beta_1^{\pm}$.

Moving higher up in the mixed layer, ℓ^+ decreases and ℓ^- increases in both LES and DNS, but the tendency is stronger for the DNS. That means, updrafts become fewer in number and more of them are acyclic—a geometric manifestation of smaller structures coalescing into larger plumes as described in Mellado et al. (2016, their Fig. 9). This trend continues into the inversion layer, where another difference between LES and DNS is found: the value of ℓ^- stays positive for the inversion layer of the LES case; for the DNS, on the contrary, ℓ^- decreases beyond a certain height and eventually becomes negative. The LES inversion layer is thus characterized by a large number of acyclic updrafts ($\ell^+ < 0$) within a large downdraft complex ($\ell^- > 0$). This is also the case for the lower part of the DNS inversion. Higher up in the DNS, however, ℓ^- becomes negative, i.e., updraft and downdraft domains are no longer intertwined but isolated acyclic components appear, separated by the mostly non-turbulent regions where $|w| < \varepsilon$. This two-layer structure of the inversion zone is indeed a manifestation of the explicit representation of small-scale entrainment (Garcia and Mellado 2014), a process that cannot be represented adequately by LES. The transition between inversion and free atmosphere is much smoother in DNS than in LES, a consequence of the strong capping inversion that is imposed on the LES by the initial condition. In the LES, the free atmosphere has points where both ℓ^+, $\ell^- > 0$; on the contrary, it is ℓ^+, $\ell^- < 0$ almost everywhere in the DNS for the free atmosphere.

While DNS data are available for a short quasi-stationary period, the LES data feature the daily cycle. From a thin boundary layer until 0930 UTC, the CBL grows rapidly until 1100 UTC. The above bulk description of the CBL in terms of its topological invariants also holds during this growth period: ℓ^- reaches much larger values in the inversion layer, implying that the updrafts are less interspersed while the downdrafts see more frequent interruptions by updrafts. This is consistent with the absence of a plume-merging layer in the morning hours and its subsequent formation. Starting at around 1730 UTC, the LES features an evening transition, that is, the CBL separates into a stably-stratified layer close to the surface and residual turbulence aloft. The stable layer is characterized by a predominance of very small fluctuations, which renders the updraft domain a set of small, acyclic components. Residual turbulence again features a network pattern of updrafts ($\ell^+ > 0$) and downdrafts ($\ell^- > 0$ occasionally).

Given the potential presence of very-large-scale structures, one might expect that $\ell^{\pm}$ is affected by domain size, despite the fact that we are normalizing here the Betti number to a non-dimensional parameter that does not fundamentally carry the dimension of space. This might happen if, for instance, the number of connected updraft components in the surface layer stays constant for a growing domain, since there is essentially one large connected component. The downdraft components would nonetheless increase in number with domain size. An analysis of this phenomenon for the DNS dataset (see Fig. 3) reveals that, while there is an effect on the values of $\ell^{\pm}$ for increasing domain size, this effect becomes negligible once a sufficiently large domain is considered. The complex nature of the flow produces many independent structures scattered across the domain, which in turn precludes significant changes in the value of $\ell^{\pm}$ for its non-dimensionalization. For the analysis presented here, we consider a horizontal subdomain of $(1.33 \times 1.33) L_O^2$[2] (in the DNS), which is large enough that scaling is not a problem.

4 Classification of Convective-Boundar-Layer Regimes Through Machine Learning

Given the intraregional similarity of the logarithmic ratios ℓ^+ and ℓ^- within CBL subregions, we expect that modes in their joint probability density correspond to different CBL subregions. We exploit this property to classify the CBL by (ℓ^+, ℓ^-) and fit the data to an unsupervised classification algorithm, namely the Gaussian mixture model (GMM)

$$\sum_{i=1}^{k} w_i \, \mathcal{N}_i(\mu_i, \Sigma_i) \qquad \text{with } w_i \in \mathbb{R}; \quad \sum_{i=1}^{k} w_i = 1, \tag{2}$$

where k is the number of clusters, w_i is the mixing weight of cluster $i \in \{1, \ldots, k\} \subset \mathbb{N}$, and $\mathcal{N}_i$ is a bivariate normal density with mean vector μ_i and covariance matrix Σ_i. The fit is obtained using expectation–maximization (EM). This procedure is a "soft" version of K-means clustering: the EM algorithm iteratively maximizes the likelihood of the data under a probabilistic cluster assignment (Hastie et al. 2001). The cluster boundaries thus obtained correspond to non-linear confidence ellipsoids of component Gaussians, as opposed to the Voronoi cells that would be obtained by K-means; we find these ellipsoids to give a better overall fit. To choose k, we calculate the silhouette score (Rousseeuw 1987), a combined measure of similarity within a cluster and dissimilarity between clusters, for $2 < k < 10$. The best silhouette score was obtained as $k_{\mathrm{LES}} = 7$ and $k_{\mathrm{DNS}} = 5$ for LES and DNS, where the $\ell^{\pm}$ values for the LES data are smoothed by a 5-min running average. Such a close match of $k_{\mathrm{LES,DNS}}$ with the number of regimes in the CBL (cf. Fig. 1) indicates that using ℓ^+ and ℓ^- as classification features captures the physically different state of the CBL subregions.

Beyond the mere match of cluster numbers, both datasets exhibit well-defined clusters representing the spatio–temporal organization of the CBL (Fig. 4c, d), a remarkably accurate representation of the physical CBL that shows the merit of such topological representation. While the clusters in themselves do not explicate the physical regime to which they belong, we can use their vertical positioning to attribute them unequivocally to a respective turbulence regime. The method uniquely attributes the mixed layer to a well-defined and isolated regime

[2] We use here a subdomain of 1024×1024 grid cells in the horizontal plane Oxy (out of a total of 5120×5120 grid cells). This subdomain is subsampled at every second grid point in both horizontal directions Ox and Oy to 512×512 grid cells. For the analysis in Fig. 3, a single instant for a subdomain of 2048×2048 grid points—subsampled equally—is used.

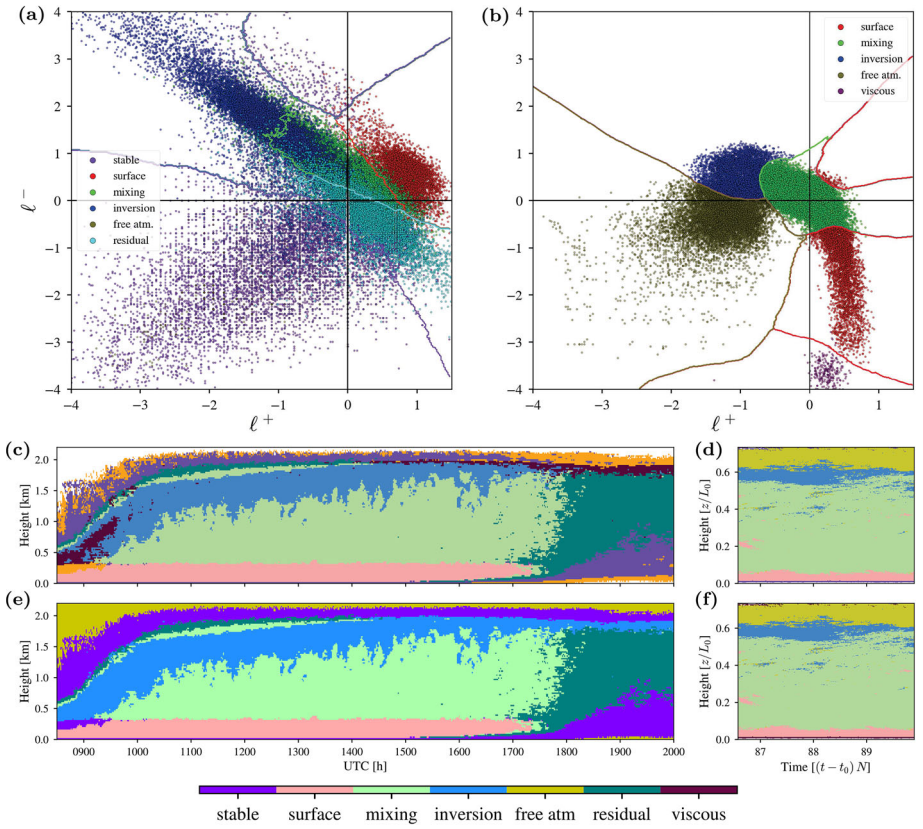

Fig. 4 Cluster assignment for the unsupervised GMM learning approach for the LES data (**a**) and DNS data (**b**). Colour lines show decision boundaries of k-nearest neighbour classifiers, thus delimiting the areas in the (ℓ^+, ℓ^-)-plane corresponding to each CBL subregion. **c** GMM-based partitioning of the CBL based on the ratios $\beta_1^{\pm}/\beta_0^{\pm}$ for LES, the corresponding partitioning for DNS is shown in (**d**). **e** shows the LES partitioning after applying physically motivated reassignment rules. **f** is identical to **d** as no such rules are applied for DNS. The classifications in (**a, b**) correspond to those in (**e, f**)

in both DNS and LES, but physical attribution of other subregions for LES requires final post-processing: (1) the inversion layer is formed by merging two clusters, shown in Fig. 4c in crimson and blue; (2) the stable layer is formed by merging the orange and violet clusters in Fig. 4c; (3) those points (t, z) in the domain for which $\max\left(\beta_0^+, \beta_0^-\right) = 0$ (and hence ℓ^+, ℓ^- are undefined), shown as the white area in Fig. 4c, are labeled as free atmosphere. This, altogether with imposing the physically intuitive vertical order of turbulence regimes in the atmospheric boundary layer, produces the classification shown in Fig. 4e and yields surprising classification accuracy (Table 1, left) where the bulk classification presented in Fig. 1a is used as 'truth'. For DNS, no such post-processing was carried out (the classification schemes shown in Fig. 4f, d are identical). The transition area between the LES inversion layer and free atmosphere is assigned to different clusters, which correspond to the mixing layer, residual turbulence, and stable layer in ascending order (height). This reflects the shift from turbulent dynamics to a largely turbulence-free zone, ending with a sharp break into the free atmosphere.

The decision regions attributed to individual regimes of PBL turbulence apparently differ when considered through LES versus DNS (Fig. 4a vs. b). This is, first, a consequence of a different representation of the turbulence (full cascade at reduced Reynolds number in DNS vs. reduced cascade and turbulence model in LES), i.e. a different effective Reynolds number. Second, the LES actually captures a different physical situation—namely the realistic evolution of the PBL over the course of a day—while the DNS is for a highly idealized scenario: a quasi-steady (slowly growing) convective PBL growing into a linear stratification. The focus here is hence not on the universality of the (ℓ^+,ℓ^-) regions for a given CBL regime but rather on the clarity of separation in between these different regimes when a single method is considered. This difference in absolute topological characterization of the physical system illustrates the utility of a topological approach to the problem: It also differentiates between different physical approaches to the physical system—not surprising in this particular case given the fundamental difference of small-scale representation among DNS and LES.

Comparing the assignment (Fig. 4e, f) with the sub-region partitioning in LES and DNS obtained from bulk analysis of the flow (Fig. 1a, b) reveals substantial qualitative agreement between the classical bulk classification and our approach. This is quantified in a classification accuracy of more than 65% (left and right matrix of Table 1) with the exception of the free atmosphere in the LES case and the inversion layer in the DNS case. The DNS inversion layer is problematic, as a significant portion of it is incorrectly labeled as part of the mixing layer, and we find that assigning the upper inversion layer to the free atmosphere results in a too-low boundary between these two regions.

In LES, the evening transition, as characterized by the clustering scheme, is not the sharp change indicated by the vanishing of $z_{i,0}$, but it happens gradually and at different times across the vertical direction. We interpret this as evidence that the Betti-number ratios correctly represent the structural properties of the turbulent mixing layer, which change gradually as the land surface ceases to act as an energy source in the early evening. The same is true for the growth of the stable layer. The point scattering for LES shown in Fig. 4a also shows the stable layer to be the one CBL subregion with the greatest variability of (ℓ^+, ℓ^-). While the stable layer is mostly confined to the lower-left quadrant in (ℓ^+, ℓ^-)-space, it undergoes transition into the residual layer (around the origin) and into the inversion layers (upper-left quadrant). Points in the free atmosphere start appearing towards the lower-left corner.

Comparing this to the point scattering for DNS highlights two aspects in which the simulation approaches differ: first, the characteristic convective cell pattern in the surface layer—illustrated by negative ℓ^- ($-3 < \ell^- < 1$) and approximately constant $\ell^+ \approx 0.5$ in the DNS data—is broken by the presence of shear in the LES where ℓ^- in the surface layer becomes positive. Second, we find that the inversion layer of the DNS data features a more physical transition to the free atmosphere, mainly due to gradual decrease of ℓ^- when moving out of the turbulent region of the boundary layer. In the LES, such a gradual transition that one might expect due to physical adjacency of these two compartments, is not found. We conjecture that this is due to the strong inversion that is imposed in the LES and decouples the turbulent boundary layer from the free atmosphere. Figure 4a also shows the decision boundaries for a k-nearest neighbours (kNN) classifier, trained on the labelled data points represented in Fig. 4(e, f). Specifically, by using the pair of values (ℓ^+, ℓ^-) as explanatory variables and the label assigned to each pair by the unsupervised clustering algorithm as the response, we can assign to any arbitrary point in the (ℓ^+, ℓ^-) plane the label that is most common amongst its k nearest neighbours (determined here with the Euclidean metric in $\mathbb{R}^2$). This aids visualization, and allows us to generalize the relationships between the values of $\ell^\pm$ to unseen data.

Table 1 Confusion Matrix for the flow classification from an un-supervised learning algorithm based on a Gaussian-mixture model and manually assigned names. The 'assigned labels' are based on bulk-analysis on the flow while the predicted labels (shown in the columns) are those given to the respective clusters that originate from a Gaussian-mixture classification of the Betti-number ratios $\beta_1^{\pm}/\beta_0^{\pm}$. The cell at row i and column j thus shows the percentage of points in region i (as per the bulk classification) assigned to region j by our model. Corresponding decision surfaces are shown in Fig. 4. For LES, the left confusion matrix corresponds to the classification shown in Fig. 4e, whereas the right matrix is obtained by relabeling points above the inversion as free atmosphere, as explained in the main text

GMM-Predicted label

Growing and ceasing CBL (LES)

Bulk-Assigned Label	SF	MX	IN	FA	RS	ST	SF	MX	IN	FA	RS	ST	%
SF	**70**	13	3	0	2	12	**70**	13	3	0	2	12	13
MX	0	**89**	11	0	0	0	0	**89**	11	0	0	0	28
IN	0	22	**69**	0	2	6	0	18	**69**	12	0	0	19
FA	0	3	6	**39**	8	45	0	0	6	**92**	1	0	22
RS	0	16	4	0	**65**	15	0	16	4	0	**65**	15	16
ST	0	0	0	10	0	**90**	0	0	0	10	0	**90**	02

Quasi-stationary CBL (DNS)

Bulk-Assigned Label	SF	MX	IN	VI	FA	%
SF	**67**	33	0	0	0	10
MX	1	**92**	6	0	1	65
IN	0	10	**50**	0	40	15
FA	0	0	0	**95**	5	09
VI	0	0	0	0	**100**	01

Cluster labels: VI–viscous layer ST–stable layer SF–surface layer MX–mixed layer
RS–residual layer IN–inversion layer FA–free Atmosphere %–% of domain

So far, the classification scheme for both DNS and LES has been based only on knowledge of the two logarithmic quotients $\ell^{\pm}$ at different (t, z) points in the simulations. However, in the case of LES our a-priori knowledge of the vertical organization of the flow can be used to further refine the classification. We employ our classification as an interface-detection method and attribute all compartments of the flow above the inversion layer to the free atmosphere. This amounts to assigning the light green, dark green, and violet areas above the blue inversion layer in Fig. 4e to the free atmosphere above them (in yellow). This relabelling dramatically reduces misclassification of free atmosphere as stable boundary and residual layer, as can be seen in Table 1 (centre).

5 Discussion and Conclusion

We introduced a non-dimensional geometric representation of turbulent structures in the atmospheric boundary layer based on the sign of vertical velocity only. Topological invariants, specifically the logarithmic ratio of Betti numbers in updrafts ($\ell^+ = \ln(\beta_1^+/\beta_0^+)$) and downdrafts ($\ell^- = \ln(\beta_1^-/\beta_0^-)$) have a remarkable descriptive power. While based only on the number of connected components and loops in the geometric representation of updrafts and downdrafts, they reach a classification accuracy of about 80% when used as features in an unsupervised learning setting to classify the boundary-layer regimes. This underlines the crucial role of the vertical velocity component w as carrier of structural information about the CBL state. Indeed, the notion of connectivity is central both in what the topological invariants measure and in the concept of spatial as well as temporal coherence. Therefore the topological representation of coherent structures appears natural. In comparison with approaches based on spectral transforms, this approach is both simpler and more general. No assumption of periodicity or smoothness is necessary; no set of basis functions is required, but only the construction of the cubical complexes $\mathcal{C}^{\pm}(z, t)$ from binary arrays. This also means that the structural properties found by the method can be immediately linked to other features in physical space, such as boundary conditions at the surface.

While the investigation put forth above is mainly of illustrative character, we see merit in this methodological approach for a number of applications. First, our results suggest that a classification of boundary-layer turbulence by its source as introduced, for instance, by Harvey et al. (2013) and Manninen et al. (2018) can also be based on the present approach, which would dramatically simplify the number of thresholds and the amount of data involved in such classification. Second, the algorithmic approach put forth here can be used in the emerging field of Doppler scans from lidars where resolved fields of atmospheric flow, in particular of the vertical velocity, are available (Lothon et al. 2009; Barlow et al. 2011). In this context, the threshold (used here for physical reasons) might be used to increase the resilience of the algorithm to instrument noise. Finally, the fact that key topological characteristics of the flow are preserved under a quasi-binary representation of the vertical velocity field may be utilized for physically-inspired data compression of atmospheric data.

Acknowledgements Open Access funding provided by Projekt DEAL. We thank Michael Hintz for providing the LES data used in this analysis, Shaofeng Liu for sharing his surface-layer model, and Jade Rachele Garcia and Juan Pedro Mellado for sharing their DNS results as initial condition. This work was supported by the German Research Foundation through the Collaborative Research Centre 32, Project C7. Computing time was provided through the project HKU24 of the Gauss Centre for Supercomputing at Jülich Supercomputing Centre on the supercomputers juqueen and juwels. CA acknowledges support through a UoC PostDoc Grant of the University of Cologne.

References

Adrian RJ (2007) Hairpin vortex organization in wall turbulence. Phys Fluids 19:041301

Barlow JF, Dunbar TM, Nemitz EG, Wood CR, Gallagher MW, Davies F, O'Connor E, Harrison RM (2011) Boundary layer dynamics over London, UK, as observed using Doppler lidar during REPARTEE-II. Atmos Chem Phys 11:2111–2125

Deardorff JW (1970a) Convective velocity and temperature scales for the unstable planetary boundary layer and for Rayleigh convection. J Atmos Sci 27:1211–1213

Deardorff JW (1970b) Preliminary results from numerical integrations of the unstable planetary boundary layer. J Atmos Sci 27:1209–1211

Fedorovich E, Mironov DV (1995) A model for a shear-free convective boundary layer with parameterized capping inversion structure. J Atmos Sci 52(1):83–96

Fernandes RLJ, Adrian RJ (2002) Scaling of velocity and temperature fluctuations in turbulent thermal convection. Exp Therm Fluid Sci 26(2–4):355–360

Finnigan JJ, Kaimal JC (1994) Atmospheric boundary layer flows: their structure and measurement. Oxford University Press, Oxford

Fodor K, Mellado JP, Wilczek M (2019) On the role of large-scale updrafts and downdrafts in deviations from Monin–Obukhov similarity theory in free convection. Boundary-Layer Meteorol 172(3):371–396

Garcia JR, Mellado JP (2014) The two-layer structure of the entrainment zone in the convective boundary layer. J Atmos Sci 71(6):1935–1955

Harvey NJ, Hogan RJ, Dacre HF (2013) A method to diagnose boundary-layer type using doppler lidar. Q J R Meteorol Soc 139(676):1681–1693

Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning, 2nd edn. Springer series in statistics. Springer, New York

Kaimal JC, Wyngaard JC, Haugen DA, Coté OR, Izumi Y, Caughey SJ, Readings CJ (1976) Turbulence structure in the convective boundary layer. J Atmos Sci 33(11):2152–2169

Krishan K, Kurtuldu H, Schatz MF, Gameiro M, Mischaikow K, Madruga S (2007) Homology and symmetry breaking in Rayleigh–Bénard convection: experiments and simulations. Phys Fluids 19(11):1–6 arXiv:nlin/0701043

Lilly DK (1968) Models of cloud-topped mixed layers under a strong inversion. Q J R Meteorol Soc 94(401):292–309

Liu S, Shao Y, Kunoth A, Simmer C (2017) Impact of surface-heterogeneity on atmosphere and land-surface interactions. Environ Modell Softw 88:35–47

Lothon M, Lenschow DH, Mayor SD (2009) Doppler lidar measurements of vertical velocity spectra in the convective planetary boundary layer. Boundary-Layer Meteorol 132:205–226

Manninen AJ, Marke T, Tuononen M, O'Connor EJ (2018) Atmospheric boundary layer classification with Doppler lidar. J Geophys Res Atmos 123(15):8172–8189

Mellado JP (2012) Direct numerical simulation of free convection over a heated plate. J Fluid Mech 712:418–450

Mellado JP, van Heerwaarden CC, Garcia JR (2016) Near-surface effects of free atmosphere stratification in free convection. Boundary-Layer Meteorol 159(1):69–95

Mischaikow K, Kokubu H, Mrozek M, Pilarczyk P (2019) CHomP—computational homology project. http://chomp.rutgers.edu/Projects/Computational_Homology/OriginalCHomP/software/

Pino D, Vilà-Guerau de Arellano J, Kim SW, Pino D, Kim SW (2006) Representing sheared convective boundary layer by zeroth- and first-order-jump mixed-layer models: large-eddy simulation verification. J Appl Meteorol 45(9):1224–1243

Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65

Shah SK, Bou-Zeid E (2014) Very-large-scale motions in the atmospheric boundary layer educed by snapshot proper orthogonal decomposition. Boundary-Layer Meteorol 153:355–387

Sorbjan Z (1986) On similarity in the atmospheric boundary-layer. Boundary-Layer Meteorol 34(4):377–397

van Heerwaarden CC, Vilà-Guerau de Arellano J, Moene AF, Holtslag AAM (2009) Interactions between dry-air entrainment, surface evaporation and convective boundary-layer development. Q J R Meteorol Soc 135(642):1277–1291

Vilà-Guerau de Arellano J, van Heerwaarden CC, Lelieveld J (2012) Modelled suppression of boundary-layer clouds by plants in a CO2-rich atmosphere. Nat Geosci 5(10):701–704

Wasserman L (2018) Topological data analysis. Annu Rev Stat Appl 5(1):501–532