Topological Aspects of Turbulence in the Planetary Boundary Layer

José Luis Licón-Saláiz

JOSÉ LUIS LICÓN-SALÁIZ

TOPOLOGICAL ASPECTS OF TURBULENCE IN THE PLANETARY BOUNDARY LAYER

Topological aspects of turbulence in the planetary boundary layer

INAUGURAL-DISSERTATION

zur
Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität zu Köln

vorgelegt von

JOSÉ LUIS LICÓN-SALÁIZ
aus
Ciudad de México, Distrito Federal
México

Köln 2020

BERICHTERSTATTER:
Prof. Dr. Angela Kunoth
Prof. Dr. Roel Neggers

TAG DER LETZTEN MÜNDLICHEN PRÜFUNG:
04.02.2020

ABSTRACT

The second decade of the 21st century has seen significant advances in the nascent field of Topological Data Analysis (TDA), both from the theoretical and applied standpoints. The subject of this dissertation is the investigation, by topological means, of the spatial structure of turbulent convection in the planetary boundary layer (PBL) of the Earth. It thus serves a dual purpose: first, to introduce atmospheric science as a source of rich and challenging problems to the applied topology community; second, to introduce topological ideas and methods to the atmospheric science community, as a complement to commonly applied methods such as global bulk measurements and spectral transforms.

The first problem considered here is the interaction of the atmosphere with complex land surface patterns, where the limitations of classical approaches are known. Low-order topological invariants, the Betti numbers, are computed for the vertical wind velocity fields of Large-Eddy Simulation (LES) models, and used to quantify the structural changes in convective flow induced by different surface patterns. These invariants are also shown to capture structural properties of the boundary layer and its temporal evolution, as they induce a physically meaningful partition of the model domain into the corresponding boundary layer subregions. Here, the results obtained from LES model data are compared with those from a direct numerical simulation (DNS).

Next, the connectivity of updrafts in an LES is determined algorithmically, and used in the quantitative analysis of the hierarchical organization of convective flow. An empirical law of updraft scaling is derived, which agrees with the expected Kolmogorov scaling. A tree-like representation is used to quantify the rate of plume coalescence, and its dependence on land surface heterogeneity. This representation is then used to assess the relative efficiency of different land surface types in sustaining convection.

Finally, the spatial distribution of shallow cumulus clouds is analyzed. To this end, the stable rank invariant from persistent homology is interpreted as a homological density function, which justifies its use as a spatially descriptive statistical measure. This allows direct comparison of the spatial pattern of a cloud field with different spatial point processes. An index for spatial organization is introduced based on persistent homology, and comparison with the classical I_{org} index shows it is more resilient to spatial noise. Finally, a morphological classification of cloud fields based on the persistence contour formalism is described.

ZUSAMMENFASSUNG

Topological Data Analysis (oder Topologische Datenanalyse), ein neuer Forschungsbereich der Angewandten Mathematik, hat im zweiten Jahrzent des XXI. Jahrhunderts rapide Fortschritte aufgewiesen, sowohl in der Theorie als auch in der Anwendung. Das Ziel dieser Dissertation ist, die räumliche Struktur turbulenter Konvektion in der planetaren Grenzschicht der Erde mit topologischen Werkzeugen zu untersuchen. Sie dient also einem doppelten Zweck: erstens, der Einführung der Meteorologie als reiche Quelle von praktischen Problemen für die Angewandte Topologie; und zweitens, der Einführung von Ideen und Methoden topologischer Natur in die Meteorologie, als Ergänzung anderer, gewöhnlicher Datenauswertungsmethoden.

Als erstes Problem wird ist die Interaktion zwischen Atmosphäre und die Erdoberfläche, bei der die Einschränkungen der gängigen Analysemethoden schon bekannt sind, betrachtet. Topologische Invarianten niedriger Ordnung, die Bettizahlen, werden für Datensätze aus Large-Eddy Simulationen (Grobstruktursimulationen) und Direct Numerical Simulations (Direkte Numerische Simulationen) berechnet. Anhand dieser Zahlen können strukturellen Veränderungen in der turbulenten Strömung gemessen werden, die durch die Einflüsse der Landoberfläche entstehen. Darüber hinaus kann auch die inhärente Struktur der planetaren Grenzschicht ausschließlich durch diese Bettizahlen bestimmt werden.

Als nächstes wird eine räumliche Aufspaltung der Aufwinde, die im Modell entstehen, in Zusammenhangskomponenten algorithmisch bestimmt. Diese Partitionierung beschreibt quantitativ die hierarchische Organisation der Strömung. Ein empirisches Skalengesetz wird abgeleitet, welches mit der Kolmogorovtheorie für den Inertialbereich des turbulenten Spektrums übereinstimmt. Eine baumartige Darstellung der Aufwinddaten ermöglich die Quantifizierung der Koaleszenzrate konvektiver Strukturen und

deren Zusammenhang mit der Heterogeneität der Landoberfläche. Damit kann auch die relative Effektivität einzelner Landoberflächentypen in der Erhaltung der Kovektion quantifiziert werden.

Danach wird die räumliche Verteilung von Schönwetterwolken (Cumulus humilis/ Cumulus mediocris; Cu hum/med) analyisiert. Hierfür wird die stabile Ranginvariante aus der persistenten Homologie als eine Homologiedichtefunktion interpretiert, wodurch die Invariante als räumliche Statistik benutzt werden kann. Damit kann die Raumverteilung von Wolkenfeldern mit der von Punktprozessen verglichen werden. Ein Index für die räumliche Organisation von räumlichen Objekten wird, basierend auf der Ranginvariante, definiert, und mit dem gängigen I_{org} Index verglichen. Die topologische Version weist eine höhere Resilienz gegnüber Rauschen auf. Schließlich wird eine morphologische Klassifikation diverser Wolkenfelder mittels des persistenten Konturformalismus definiert.

Algebra is the offer made by the devil to the mathematician. The devil says: I will give you this powerful machine, it will answer any question you like. All you need to do is give me your soul: give up geometry and you will have this marvelous machine.

— Sir Michael Atiyah

PROEMIUM

There exist within mathematics different sources of tension between disparate modes of thought: between the discrete and the continuous, between the pure and the applied, or between the algebraic language of structures and the geometric language of shapes, to name but a few. This dissertation, in some sense, straddles the diffuse boundaries between all of these, which is a rather strange place to be in, but at the same time a very exciting one.

This project was born out of my fascination with the subject of applied topology, which on first impression can sound somewhat like an oxymoron. Topology is, after all, regarded as one of the more esoteric areas of mathematics. A quick review of the literature, however, reveals that topology is the source of applicable ideas and methods, in particular algebraic topology with its emphasis on finite and computable (sometimes even in reasonable time) invariants. One finds that different people, working in different scientific fields, can and do use these concepts in problems of data analysis, and derive new insights from them. This arcane field thus springs forth bearing a new language to talk about data. Thus, when I was faced with a new, challenging problem in the form of pattern analysis in turbulent convection, the temptation of attacking it with these techniques proved irresistible for me.

A work such as this, spanning several years, has inevitably been influenced by many people, both directly and indirectly. Here I will attempt to recount those people.

Firstly, I thank my family in Mexico for their continuous support and encouragement. This goes especially for my parents, Araceli

and José Luis, who have been a guiding example and inspiration throughout my life.

On this side of the ocean, I would like to express my gratitude to my supervisor, Prof. Angela Kunoth, for placing the trust in me to undertake this project, even when the subject matter was a departure from her usual field of work. In the end I came to appreciate the immense intellectual freedom she gave me to determine my own path, as it has been an invaluable learning experience. I would also like to thank Dr. Cedrick Ansorge, with whom I spent many hours in detailed discussions which were always a great aid in formalizing my ideas and interpreting the results obtained. In the process I learned a lot, not only about boundary layer meteorology and fluid mechanics, but also about good scientific practice. For this I am also deeply grateful. I also thank Prof. Yaping Shao for his input and comments on my work, and for the inspiration behind one of the methods discussed in this dissertation. My gratitude goes also to Prof. Roel Neggers, who has agreed to review this work on relatively short notice (and provided some of the simulation data studied here), and to Prof. Kathrin Bringmann, president of my examination committee.

I would like to acknowledge the German Research Foundation (DFG) for the funding granted to me during most of my work on this project, through the SFB/TR32 research initiative "Patterns in Soil–Vegetation–Atmosphere Systems: Monitoring, Modelling and Data Assimilation". The Gauss Centre for Supercomputing e.V. (www.gauss-centre.eu) is gratefully acknowledged for funding this project by providing computing time through the John von Neumann Institute for Computing (NIC), project ID HKU24 on the GCS Supercomputer JUWELS at Jülich Supercomputing Centre (JSC).

My gratitude goes to the TDA community at large, and especially to the organizers of all the conferences and workshops I had the privilege of attending in the past years, namely

- Applied Topology in Będlewo, 2017.
- Linking Topology to Algebraic Geometry and Statistics. Leipzig, 2018.
- Summer School on TDA and PH. Levico Terme, 2018.

- Dynamics, Topology, and Computations. Będlewo, 2017.
- Applied Topology: Methods, Computations, and Science. Vienna, 2018.
- 8th International Climate Informatics Workshop. Boulder, 2018.
- European Machine Learning Conference. Würzburg, 2019.

I also thank the many wonderful people I met during these events, with whom I could share my enthusiasm for applied and computational topology: Dejan, Siargey, Oliver, Barbara, Grzegorz, Janis. And especially my dear friend Henri Riihimäki for all the valuable time spent together, be it working out mathematical issues or at a pool table (sometimes even at the same time). Besides, cruising along the Finnish countryside while computing pullbacks is no doubt among the most memorable moments of my whole PhD experience, thank you for that!

Finally, the one person who has stood next to me, during the best and the worst moments of the past years, and whose seemingly unquenchable desire for adventure and discovery never failed to arouse my own: Thirza, thank you for your love and encouragement, which have done more for me than I will ever be able to express.

Darmstadt, December 7th, 2019

José Luis Licón-Saláiz

CONTENTS

1

INTRODUCTION

1.1 ATMOSPHERIC PATTERNS AND TOPOLOGY

The idea of *patterns* and pattern formation has long played an important role in natural science, and has experienced a renewed interest during the first decades of the XXIst century, as quantitative investigations into the nature of patterns in natural systems become more prevalent. A pattern in the sense discussed here is understood as a connected region of the spacetime domain with similar characteristics, and which exhibits a certain degree of spatial and temporal coherence. There are two important aspects involved in this definition: The first is the geometric aspect, according to which we expect a pattern to exhibit regularity in space which makes its presence visually and intuitively clear, even if it escapes a precise geometric definition. From mathematics we know of spatial patterns which exhibit regularity, for example sphere packings and tessellations of the plane. We can hardly hope to find such perfect patterns in nature, however—after all, clouds are not spheres and rocks are not cubes. Yet, as we will see, there is much that can be known about patterns in nature without need for this degree of geometric precision. The second aspect in the definition is a statistical one, where we expect the values of a given quantity, such as temperature or wind velocity, to be regular in a statistical sense within the region of spacetime that is spanned by a pattern. This is a manifestation of the underlying dynamics which result in pattern formation.

The importance of patterns stems from several factors: 1) their simplicity, which contrasts with the complexity of the systems

in which they emerge; 2) their universality, insofar as dissimilar systems will exhibit the same patterns close to their respective bifurcation regimes; and finally, 3) the resilience of patterns to environmental noise [Goehring, 2013]. Patterns are understood as a manifestation of self-organization in a non-linear dynamical system driven away from equilibrium, or as an emergent phenomenon [Vereecken *et al.*, 2016]. In either case they aid in understanding the dynamics and interactions of the underlying systems. A classic example of patterns as a characteristic of self-organization is given by the ideal situation of Rayleigh-Bénard or Bénard-Marangoni convection, where the interplay of buoyancy and fluid viscosity leads to stable configurations which display characteristic spatial patterns (Figure 1.1 shows an example of this). This idealized setting has been extensively studied from a theoretical [Mizushima, 1994; Golovin *et al.*, 1995], empirical [Meyer *et al.*, 1987; Cerisier *et al.*, 1996], and numerical [Lee *et al.*, 1989; Mizushima, 1995; Gelfgat, 1999] perspective, and is indeed the expression of a more general phenomenon: convective systems will naturally tend to form such geometrical patterns, whether the convective medium is oil on a hot pan, the Earth's atmosphere, or the surface of the Sun, to name a few examples.

This dissertation will focus on the special case of convection in the Earth's atmosphere. A key component in this study is the presence of *turbulence*. Atmospheric turbulence is associated by many of us with a bumpy airplane ride, but it is a much richer and deeper subject than this anecdotal association would suggest. It is not only one of the great open problems in classical physics, but has also attracted the attention of numerous people in the mathematical community. Some notable examples of this include the pioneering work by Hopf [1948], who advanced the hypothesis that, as the Reynolds number[1] of a fluid increases, its flow undergoes a large, possibly infinite, number of bifurcations in phase space, each adding different periodic modes. The result of this process is the seemingly random, chaotic motions of fully-developed turbulence. An equivalent theory was independently proposed by Soviet physicist Lev Landau [1944].

1 The Reynolds number of a fluid is defined as the ratio of its velocity to its viscosity. See Section 2.4.

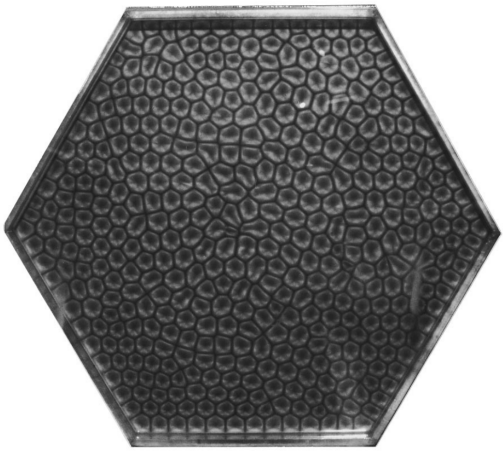


Figure 1.1: Cellular structure observed experimentally in Bénard-Marangoni convection when uniformly heating a thin layer of oil from below (Figure from Cerisier *et al.* [1996]).

Another notable example is the work of Ruelle and Takens [1971], who instead of an infinite sequence of bifurcations, proposed only three: from steady flow to periodic flow, followed by quasi-periodic flow, and thence to turbulent flow. This approach would prove to be more successful and longer-lived than the Hopf-Landau theory, not least because it is easier to test experimentally. Of special significance in the mathematical study of turbulence is the name of A.N. Kolmogorov, who developed a statistical theory of turbulence which we will briefly comment on in Chapter 2.

Returning now to the setting of atmospheric convection, we find that atmospheric turbulence exists almost exclusively in the vicinity of the Earth's surface. This part of the atmosphere is the *planetary boundary layer* (PBL), sometimes also called the atmospheric boundary layer (ABL). Assuming a day with fair-weather conditions, two distinct regimes will exist in the PBL: first, a stable surface layer during nighttime and in the early morning hours. Then, after sunrise, the surface is continuously fed with energy via sunlight, and this energy is then reflected or radiated into

the atmosphere. This energetic forcing initiates convection, and the system rapidly transitions from the stable regime to an unstable one characterized by fully-developed turbulence, which destroys the simple, elegant patterns seen in idealized convection illustrated in Figure 1.1, but gives rise to new ones: vortex streets in wake turbulence [Henderson, 1997], hairpin vortices [Adrian, 2007], convective plumes [Bizon *et al.*, 1997], or large-scale organized motion Jiménez [2012]. These are examples of *coherent structures*: connected regions of the flow domain with similar characteristics, which exhibit spatiotemporal coherence, i.e. they occupy relatively large volumes of space and exist for a significant time span [Antonia, 1980; Shah and Bou-Zeid, 2014]. Such large-scale structures also play an important role in the atmosphere system, and are of central importance in meteorology [Agee, E. M. and Chen, T. S. and Dowell, 1973; Rosmond, 1973; Ray, 1986]. Being able to characterize and describe these structures and their dynamics in an objective, quantitative fashion is essential in order to improve current understanding of the Earth's changing climate and its future.

Given the fact that the land surface acts as an energy source for atmospheric convection, it is to be expected that the precise configuration and features of the surface will have a direct impact on the structure of the resulting turbulent flow. It is known that the structure and shape of convective plumes in isolated settings is modulated by the shape of the heat source driving convection [Kondrashov *et al.*, 2016]. More generally, the PBL is not merely forced by surface conditions, but is in a state of constant feedback: atmospheric turbulence affects the conditions of the land surface, which responds and in turn affects the state of the turbulent flow, and so on. Since the land surface is not a collection of simple geometric shapes but rather a mosaic of diverse land types arranged in intricate patterns, we can expect that its effect on atmospheric convection will reflect this complexity. Indeed, the effect of land surface heterogeneity has been studied for realistic Shao *et al.* [2001, 2013] and idealized set-ups Rieck *et al.* [2014]; van Heerwaarden *et al.* [2014]. In all these scenarios, another crucial feature is the hierarchical organization of the turbulent flow, in the sense that smaller convective plumes merge into larger plumes close

to the surface, and this merging process continues as positively buoyant air moves up. This process eventually gives rise to large, dominant structures Mellado *et al.* [2016], and is also influenced by surface conditions.

A very different source of complexity is another phenomenon which is familiar to us all: the formation of clouds. These, particularly shallow cumulus clouds, are an important component in the climate system, due to their feedback with the rest of the land-atmosphere system. They form when convective updrafts have enough energy to transport moist air beyond the lifting condensation level. The presence of cloud will then shade a part of the land surface and reflect some of the incoming solar radiation back into space, thus reducing the total influx of energy into the atmosphere and decreasing the potential for future cloud formation. The representation of these shallow cumuli in climate models remains a major source of error [Bony and Dufresne, 2005]. This is due to the fact that these clouds exist at spatial and temporal scales smaller than the resolution of current models, therefore they need to be parametrized. Such parametrization depends on a correct description of the spatial structure and dynamical behavior of cloud fields. This covers different scale ranges, from small cloud fields over areas no larger than a few square kilometers to large domains hundreds of kilometers across, where mesoscale circulations can produce spatial patterns of great complexity (see for example Figure 1.2). In this respect the study of spatial cloud patterns is an area of active research [Pankiewicz, 1995; Heinle *et al.*, 2010; Seifert and Heus, 2013], as well as the influence exerted on these patterns by the land-surface [Garcia-Carreras and Parker, 2011; Gentine *et al.*, 2013; Rieck *et al.*, 2014].

We can now appreciate that the PBL, characterized by the complexity of turbulent flow and the presence of exquisitely intricate patterns, is a fertile field for the applied mathematician. The starting point for this dissertation can be formulated thus: **the search for a mathematical representation of patterns in the PBL, and of its interaction with the underlying land surface**. The focus is especially on spatial patterns, which are often neglected in modelling these systems, but are nevertheless important sources of variability in the resulting models [Koch *et al.*, 2017]. It quickly

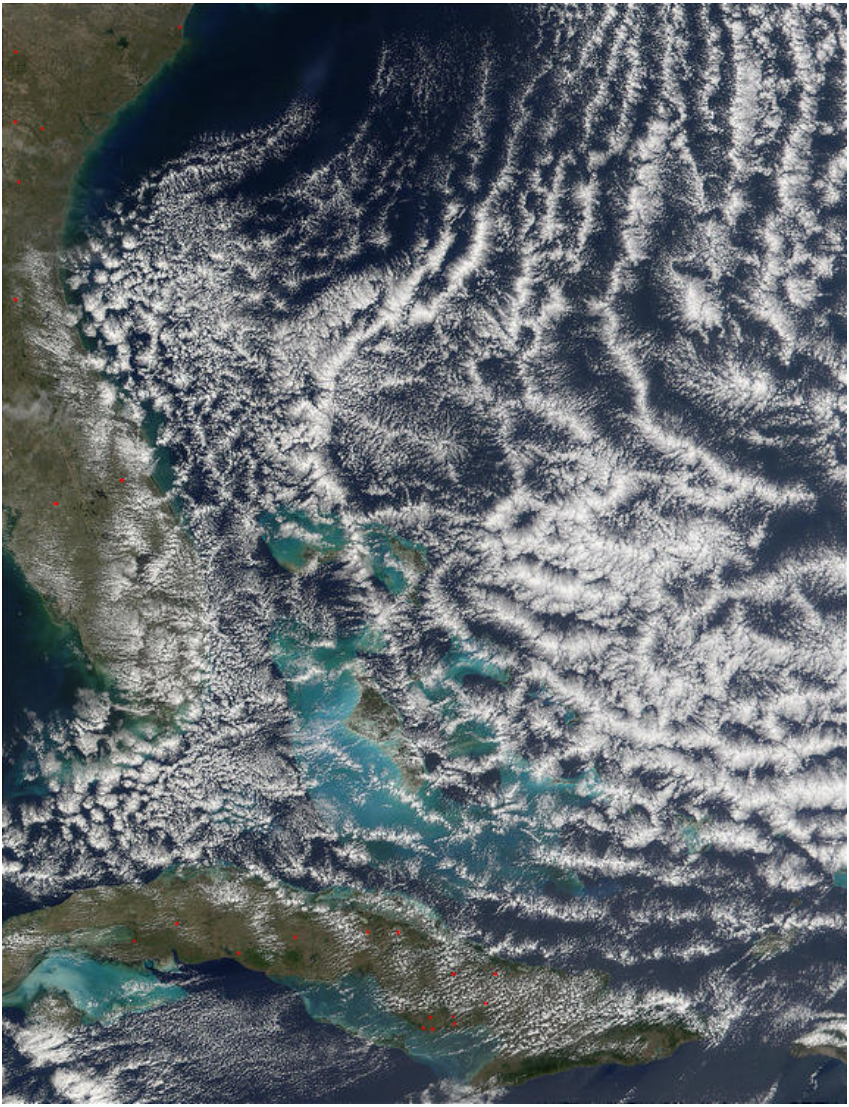


Figure 1.2: A cloud pattern resulting from open-cell mesoscale convection over the tropical Atlantic. Open-cell refers here to the fact that rising motion happens at the border of convective cells, while sinking motions happens in their interior. This results in the cloudy areas mirroring the network structure of cell boundaries, as seen here. Mesoscale is, as the name implies, the range that lies between the microscale, where small and localized systems such as storms exist, and the synoptic scale, where weather systems that span thousands of kilometers exist. For reference, the island of Cuba can be seen in the lower part of the picture. Image credit: Jacques Descloitres, MODIS Land Rapid Response Team, NASA/GSFC.

becomes clear that, given the complex geometries displayed by these patterns, we will need a very general and flexible language to describe them. Consider, by way of example, the cloud field shown in Figure 1.2: a certain regularity is apparent in the spatial arrangement of the individual clouds, owing to the open-cell convection mechanism at work here. This regularity is a consequence of clouds forming in the updraft regions, which in this regime correspond to the boundaries of convective cells. These cells are, as in the ideal case, arranged in a hexagonal cell pattern, which is perturbed here by atmospheric dynamics. Owing to this perturbation, any attempt to draw meaningful conclusions from the precise metric relationships between the clouds appears futile—the cells are far from being perfect hexagons, and their sizes and side lengths vary seemingly at random. Yet spatial coherence exists, even if it escapes a simple geometric definition. This brings us to the one word in the title of this dissertation we must yet elucidate: *topology*.

Topology is a classical field in mathematics, concerned with the study of abstract properties of space which are invariant under smooth transformations, such as twisting or bending. One of the most common examples is that of a donut and a coffee mug being topologically equivalent, since (assuming a sufficiently soft material) each can be obtained from the other via a continuous deformation, without needing to tear or puncture the material. It would not be possible, however, to obtain a donut from a solid ball: this would require carving out a hole from its center. The property of having a hole is one of many possible topological invariants.

Topology traces its origins to the late XVIIth century, and has remained for most of its existence strongly anchored in the realm of abstract mathematics. It is only in the early XXIst century that it has experienced a renaissance as a viable part of applied mathematics. This evolution has been made possible largely by the computable nature of some topological invariants, but also because numerous researchers, working independently in different scientific fields, have found striking and unexpected relationships between topological invariants which can be computed from data, and the underlying physical systems. Some notable examples are

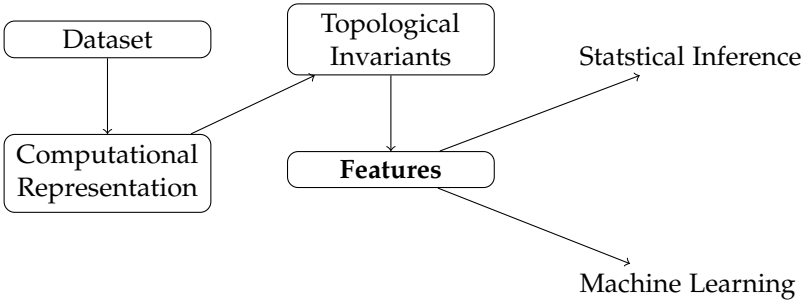


Figure 1.3: The different components of the Topological Data Analysis pipeline.

its applications to chaotic dynamics [Gameiro *et al.*, 2004], plasma physics [Garcia *et al.*, 2009], biology [Chan *et al.*, 2013; Nanda and Sazdanović, 2014; Máté *et al.*, 2014], materials science [Dłotko and Wanner, 2016; Lee *et al.*, 2017], cosmology [Pranav *et al.*, 2017], neurology [Bendich *et al.*, 2016], and medicine [Ellis and Klein, 2014; Wu *et al.*, 2017]. These are all comprised within the emerging discipline of *topological data analysis* (TDA), and this dissertation is to be seen as one step in its development. More precisely, the research question at the core of this work is:

can topology be leveraged to represent PBL patterns, and if so, does it offer any new information not provided by classical methods?

Crucially, TDA does not offer a unified approach to attack a given research problem, given the many different topological invariants that exist and the multiple representations which can be chosen for a given dataset. The general program in TDA is as follows: given a pair of topological spaces X, Y, and a finite set of points $S \subset Y$ sampled with noise from X, to recover the topology of X using the information contained in S [Zomorodian, 2012]. The analysis pipeline is illustrated in Figure 1.3. At each step of this process several questions face the practitioner, namely:

1. Which computational representation to use? This amounts to translating data defined in continuous space and time to a discrete data structure that can be stored and processed by a computer.

2. Which topological invariants can be computed efficiently from that representation?

3. How can these be mapped to meaningful features for use as inputs in further analysis?

The answers to these questions are largely domain-dependent. Throughout this dissertation we will provide the answers for the case of pattern analysis in land–atmosphere systems, thus formalizing an analysis pipeline which allows for an objective and quantitative representation of spatial patterns in the data, and the use of these in studying the underlying system's dynamics.

1.2 OUTLINE AND CONTRIBUTIONS

The main focus of this work is bridging the divide between applied topology and meteorology, by introducing ideas from the former into the latter and vice versa. As this is in essence an interdisciplinary enterprise, the first thing we need is to establish a common language. After this introduction, Chapters 2 and 3 begin by presenting the necessary background information from both fields. Chapter 2 covers boundary layer meteorology, in particular the case of the radiatively-driven convective boundary layer (CBL). Chapter 3 covers some elementary aspects of algebraic topology, namely simplicial and cubical homology, as well as persistent homology. Both chapters are intended as reference material for non-specialists in either area.

After the necessary definitions have been established, 3 research chapters with original work form the core of the dissertation. Chapter 4 focuses on the analysis of spatial patterns in coupled land–atmosphere systems by using specific topological invariants, the *Betti numbers*. We show that these invariants constitute more informative features than descriptors obtained from bulk analysis of the flow, when used in the task of discriminating between

different land surface patterns. They also allow us to classify the different subregions of the PBL with high accuracy, which can be shown by training an unsupervised classification model on the topological data and comparing the results with physical definitions.

In Chapter 5 we specialize the analysis to flow connectivity. Connectivity is, perhaps, the most elementary topological property, and is often neglected in the TDA community in favor of other, more sophisticated invariants. Nevertheless, its computation is simple, and can be very informative. Here we use this property to offer a quantitative description of the effect of land surface heterogeneity on the hierarchical organization of the flow. Most notably, this allows us to recover a key element of the Kolmogorov theory of turbulence alluded to earlier, without needing to compute spectral transforms of the data. This is a significant advantage of the methods introduced here, and as a further illustration we show how land surface heterogeneity also impacts the relative effectiveness of different land types in sustaining convection.

Chapter 6 addresses the description of spatial patterns, both regular and random, by means of persistent homology, more specifically, of the stable rank function, which will be defined later. We show how this function differs from classical first- and second-order statistics, in the sense that it provides different information about the underlying spatial distributions. Using this we can then describe the spatial distribution of shallow cumulus cloud fields, and derive from it conclusions regarding the cloud size distribution which agree with recent studies on the subject. Further, we introduce a persistent-homology based index for spatial organization and compare it with the I_{org} index. Finally, we show how the persistence contour formalism can be used to produce a morphological classification of cloud fields.

Chapter 7 contains an overview and discussion of the main results, closing remarks, and an outlook on future avenues of research.

The main contributions of this dissertation are:

1. The use of the normalized quotient of Betti numbers as a descriptive parameter for the dynamics of a system, in this

case a convective boundary layer. In a purely mathematical sense, this is a coarser descriptor than the individual Betti numbers, as some information is lost when considering their quotients. In the physical sense, however, this quantity has the advantage of being adimensional. This simplifies analysis by reducing the number of parameters needed to describe the state of the system.

2. The use of the stable rank invariant as the analogue of a cumulative distribution function for persistent homology. The result of computing persistent homology is a collection of persistence intervals, each described by a pair of real numbers. The values in these pairs are not distributed uniformly at random, but rather respond to the spatial distribution of the point cloud from which they were computed. This fact allows us to use the stable rank invariant to express higher-order relationships between points in space. This functional summary is shown to be strictly more informative than first-order (e.g. nearest-neighbor distances) and second-order (e.g. Ripley's K function) statistics.

2

METEOROLOGICAL BACKGROUND

Outline

This chapter gives an overview of the general information concerning the planetary boundary layer (PBL) needed throughout this dissertation (Section 2.1). The structure and temporal evolution of the PBL is described in Section 2.2, and the appearance of boundary layer clouds is discussed in Section 2.3. Section 2.4 recounts some key aspects of turbulent motion in the PBL, and Section 2.5 discusses different numerical simulations used to generate the data used for analysis in later chapters. The presentation of PBL structure and dynamics is based on Stull [1988], while the overview of turbulent dynamics draws from Pope [2000].

2.1 PLANETARY BOUNDARY LAYER

The planetary boundary layer (PBL), also called atmospheric boundary layer (ABL), is the part of the Earth's atmosphere closest to its land or ocean surface. It is therefore here that the surface exerts greatest influence upon atmospheric dynamics. The PBL is a subset of the *troposphere*, which is the lowest layer of the atmosphere. The troposphere has an average height of 11 km, out of which the PBL occupies anywhere between a few hundred meters and a few kilometers. The rest of the troposphere is the *free atmosphere*. The difference between the PBL and the free atmosphere above it is precisely that the former responds directly to forcings from the Earth surface, such as those induced by friction

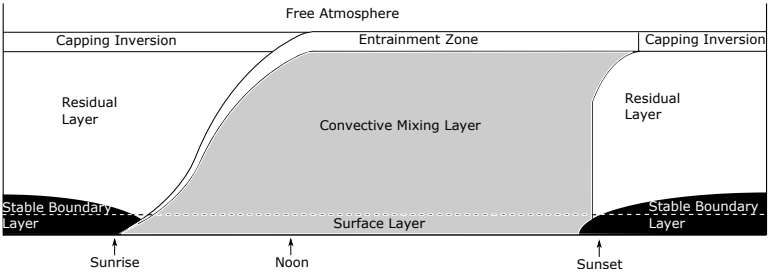


Figure 2.1: Diagram showing the structure and temporal evolution of the planetary boundary layer (PBL) over a day with fair-weather conditions over land. During daytime, a convective boundary layer (CBL) develops, with fully-developed turbulence occurring in the mixing layer, and separated from the free atmosphere by an entrainment zone located at a height of 1–2 km above the surface. At night, a stable surface layer forms adjacent to the land surface, with a layer of residual turbulence aloft (based on Stull [1988]).

with terrain, thermal radiation, and evaporation. This response happens within short time scales (in the order of a few hours).

The most important of these processes in the case studied here is energetic forcing as a result of solar heating. This results in the existence of strong gradients in air velocity within the PBL, which in turn produce *turbulence*. This is a crucial component of PBL dynamics, and is the main driver of molecular transport and mixing in the atmosphere. This turbulence-driven transport happens primarily in the vertical direction, that is, orthogonally to the land surface.

2.2 STRUCTURE OF THE PBL

As mentioned above, the structure of the PBL depends largely on surface forcings. The clearest example of this is the boundary layer at mid-latitudes over land, which will be the case studied in this dissertation. In this situation, PBL structure varies periodically with time in what is termed the *diurnal cycle*, which is a consequence of the presence or absence of sunlight. The general

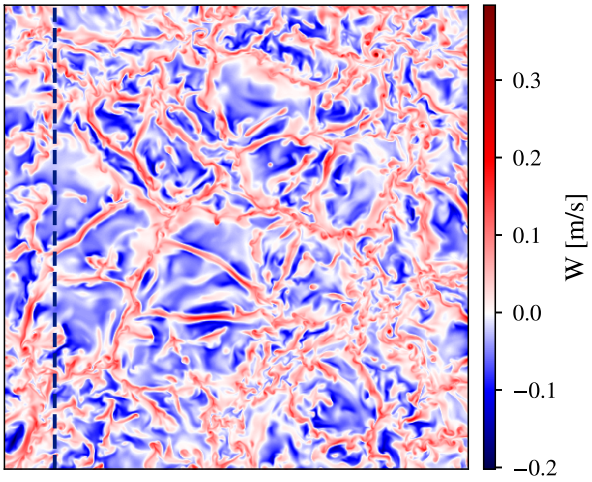


Figure 2.2: Horizontal cross section of the vertical wind velocity field (W) within the surface layer of a CBL. Data is taken from a direct numerical simulation (DNS). The characteristic surface layer structures, namely updraft curtains, can be clearly seen here. The dashed line indicates the location of the vertical cross section in Figure 2.3.

form of this cycle is shown in Figure 2.1. In the early morning hours, solar radiation begins to heat the ground, and this heat is gradually radiated into the atmosphere. As time progresses, the amount of heat fed into the system increases as well, which in turn increases the temperature gradient between land surface and atmosphere. This gradient results in heat being transferred to the atmosphere, which gives positive buoyancy to the air adjacent to the surface. This buoyancy is the main driving force behind turbulence, and is responsible for the formation of an *unstable surface layer*, where convective structures begin to form. These structures assume the form of *updraft curtains* arranged in a honeycomb-like pattern which is reminiscent of Rayleigh-Bénard convection. At the intersection of these curtains we find greater buoyancy and thus a stronger rising motion of air. An example of these structures is shown in Figure 2.2, which depicts a horizontal section of

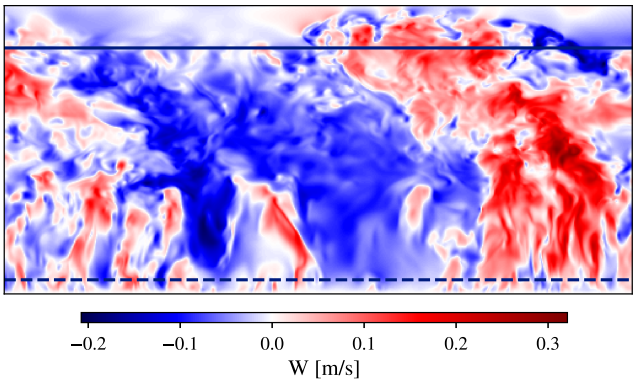


Figure 2.3: Vertical cross section of the vertical wind velocity field (W) from direct numerical simulation (DNS) of a convective boundary layer (CBL). Horizontal lines indicate the location of the cross sections for the surface layer (dashed line, Figure 2.2) and the entrainment zone (solid line, Figure 2.5). The space between those lines corresponds to the turbulent mixing layer.

the vertical wind velocity field from a direct numerical simulation (DNS, see Section 2.5).

Assuming no mixing with its surrounding environment happens, this positively buoyant, rising air gradually cools down as it travels upward. If it is buoyant enough, it continues its upward motion until encountering an area with a strong gradient in temperature and moisture: the *inversion layer*. The positive buoyancy of rising air is lost here, but given enough momentum it can break through the inversion layer and penetrate the free atmosphere aloft, where it mixes with drier, cooler air before descending back through the inversion layer. This is the *entrainment* process, and is a key factor behind boundary layer growth. This part of the PBL is sometimes also called the *entrainment zone*.

The space between surface layer and entrainment zone is where rising, positively buoyant air meets the descending, negatively buoyant air being entrained. This region is aptly called the *mixing layer*. We can therefore distinguish three main boundary layer subdomains during the day: the surface layer, the mixing region,

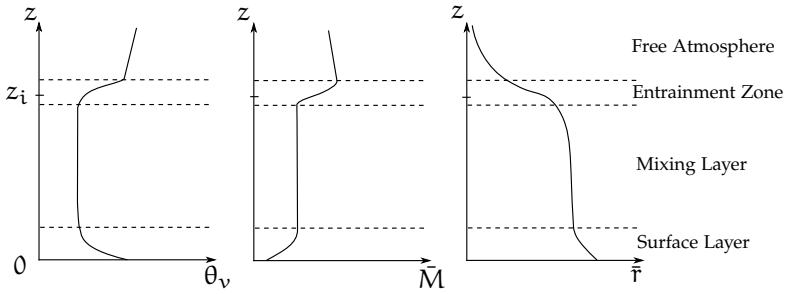


Figure 2.4: Mean profiles of temperature (left), mean horizontal wind (center, $\bar{M}^2 = \bar{u}^2 + \bar{v}^2$), and water vapor mixing ratio (right) in the daytime PBL.

and the entrainment zone. Together these constitute the convective boundary layer (CBL). Figure 2.3 shows a vertical cross section of a CBL from the same DNS as before.

The diurnal cycle consists of boundary layer growth throughout the day, as a continuous influx of solar radiation increases the energy available for air rising from the surface layer to rise through the inversion and feed the entrainment process. Boundary layer depth (i.e., its height) increases as additional air from the free atmosphere is entrained, and in the first hours of the morning this happens very rapidly due to a weaker stable layer capping at the top. Boundary layer growth gradually slows down and eventually stops after reaching a maximum height of 1–1.5 km. After sunset the influx of energy into the system stops, and the land surface stops being an energy source to become an energy sink. The temperature gradient between land and atmosphere thus begins to dissipate, and a *stable surface layer* or *nocturnal stable layer* forms, where turbulent convection is replaced by weaker motion of air. Higher up, however, the mixing layer becomes a region where residual turbulence, feeding off the energy still stored in the system, can continue for several hours after sundown, now decoupled from the surface layer. The energy in the residual layer gradually decays and the inversion height with it. The cycle recommences on the following morning at sunrise, with convection starting again beneath the nocturnal stable layer.

As a consequence of turbulent mixing, physical quantities such as wind velocity and temperature are relatively homogeneous with height in the mixing layer. This can be seen in Figure 2.4, which also shows how the remaining PBL subregions can be characterized by the gradients of these quantities. The turbulent conditions in the mixing layer destroy the updraft patterns that arise within the surface layer, as some individual structures are absorbed by their neighbors, which then grow in size. This gives rise to a new type of coherent structure, the *convective plume*. The size of these structures is significantly larger, as they can span the entirety of the mixing layer, and are thus responsible for most of the momentum exchange in the CBL, as well as for boundary layer growth through entrainment.

Beyond the mixing layer lies a stable layer, which is where the entrainment zone is located. The precise location of this layer changes both in space and time. Temporal changes occur in accordance to the diurnal cycle of the PBL. Spatial variations are due to the nature of the entrainment process, whereby rising thermals from the mixing layer go through the capping inversion in the entrainment zone, only to later sink back into the mixing layer carrying air from the free atmosphere with them. We then see a core of positively buoyant air overshooting the capping inversion, surrounded by a region of negatively buoyant air going into the mixing layer from the free atmosphere. The boundary layer height (or depth), denoted in the meteorological literature[1] by z_i, is the average height of the inversion layer base, with the average computed over the horizontal extent of the domain. A cross section of the inversion layer base is shown in Figure 2.5, where we can see the tops of at least three distinct overshooting thermals, surrounded by downdrafts. The vertical wind velocity in a large part of the domain exhibits much smaller fluctuations than it does in the up- and downdraft areas, which would suggest that it belongs to the free atmosphere.

1 Contrary to its use in mathematics, the subscript i refers here to the word *inversion*, and is not to be understood as an index ranging over a discrete set of values.

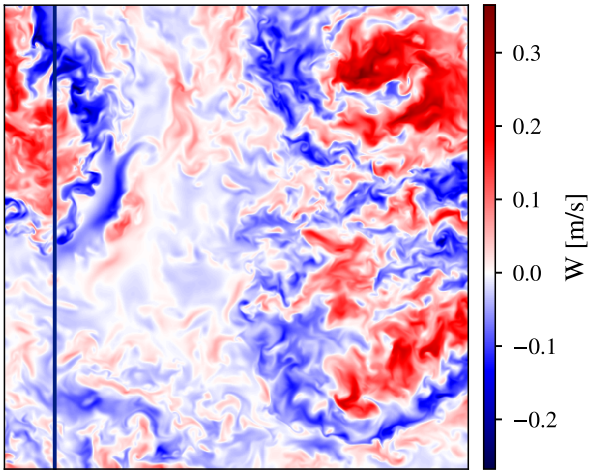


Figure 2.5: Horizontal cross section of the vertical wind velocity field (W) in the entrainment zone of a CBL, showing the qualitative difference between regions of overshooting thermals and the stably stratified free atmosphere. The solid line indicates the location of the vertical cross section in Figure 2.3.

2.3 BOUNDARY LAYER CLOUDS

Our discussion so far has assumed a dry boundary layer, which has as a consequence the absence of cloud. If we now assume the presence of a significant amount of water vapour in the atmosphere, the same basic dynamics of the CBL still apply: solar radiation heats the land surface, and radiation of this heat into the atmosphere powers buoyancy-driven convection. Given enough buoyancy, the air will rise adiabatically (and thus cooling) until the water vapour it contains condenses and produces a boundary layer cloud. The height at which this happens is the *cloud base height*, or *lifting condensation level*, and for boundary-layer cloud, it is in general close to the top of the mixing layer. The type of clouds formed by this process are *shallow cumulus clouds* (Cumulus humilis/ Cumulus mediocris; Cu hum/med), sometimes also called *fair-weather clouds* due to their occurrence at daytime in otherwise

clear skies. There is another important type of boundary layer cloud, namely stratocumulus. In this dissertation, however, we will only encounter shallow cumulus clouds. These can be treated as discrete spatial objects, on which the methods we will develop later (see Chapter 6) can be then applied. Stratocumulus, on the other hand, tend to form large unbroken sheets in diverse spatial arrangements. The topological methods developed here can be extended to deal with such spatial patterns as well, but such an extension lies beyond the scope of this dissertation.

Shallow cumuli start their existence atop a convective plume, where the air is taken to a sufficient height that condensation of its water vapor content can take place. In some cases this can happen after the plume has overshot the inversion layer and is now therefore negatively buoyant. In either case, a cloud formed by this process remains coupled to the land surface by this convective plume. Their continued existence of a shallow cumulus cloud thus depends initially on the plume's ability to supply moist air from below. In this state it is still a *forced* cloud. Condensation is an exothermic process, and in some cases the increase in temperature caused by condensing water vapor is sufficient to make the cloud positively buoyant again. This causes the cloud to draw in more air through its base, and can thus now exist and grow independently of the plume that created it. The cloud thus reaches the *active* state. This type of clouds directly impact CBL dynamics by transferring mass and energy from the mixing layer into the free atmosphere. When the cloud is no longer able to draw in air through its base, it can retain positive buoyancy for some time, but it begins to dissolve from the base upwards. The cloud is now decoupled from the surface, and this *passive* state is the last step in its existence [Stull, 1985].

The significance of shallow cumuli lies in the impact they can have on boundary layer dynamics via feedbacks. For example, such a cloud will both absorb and reflect solar radiation and thus reduce the energy available to heat the land surface, and therefore to drive convection. The net result is that the existing convective plumes will be weakened, and new ones will be less likely to form, which has a negative impact on further cloud formation. Clouds which have reached an active state can also inhibit further cloud

formation, as the process by which they transport air from the turbulent mixing region into the free atmosphere weakens boundary layer growth, thus reducing the likelihood that convective plumes will reach the point where they can produce new clouds. On the other hand, the air taken from the mixing layer will tend to increase the environmental humidity in the free atmosphere, which will result in a slower dissipation rate for the clouds already there.

Shallow cumuli are also an important subgrid-scale process for regional or global climate models. These processes occur over spatial and temporal scales which are too small to be explicitly resolved by these models, and therefore must be approximated by using a parametrization scheme. An important component in designing an accurate parametrization scheme for shallow cumulus clouds is an accurate representation of the organization of their spatial distribution, as well as of the cloud size distribution.

2.4 TURBULENCE IN THE PBL

The dynamics in the PBL are governed by the Navier-Stokes equations of fluid mechanics. When studying convection it is common to employ a simplified version of these equations, the so-called Boussinesq approximation, which amounts to neglecting differences in density throughout the fluid (for an overview of this approximation and its derivation from the equations of motion, see Stull [1988, §3.3]). The equations then become

$$\begin{aligned} \frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla)\mathbf{u} &= -\nabla p + \nu\nabla^2\mathbf{u} + b\mathbf{k}, \\ \nabla\cdot\mathbf{u} &= 0, \\ \frac{\partial b}{\partial t} + \nabla\cdot(\mathbf{u}b) &= \kappa\nabla^2 b, \end{aligned} \tag{2.1}$$

where $\mathbf{u}(\mathbf{X}, t)$ is the velocity field; $b(\mathbf{X}, t)$ is the buoyancy; ν is the kinematic viscosity of the fluid, κ is its molecular diffusivity. Here we denote vectors by boldface letters, and in particular $\mathbf{X} = (x, y, z)$ denotes the three coordinates of Euclidean space. In conformity with the common practice in physics, we will use x, y for the two horizontal directions, and z for the vertical direction

(perpendicular to the surface). The two horizontal (i.e., parallel to the boundary) components of the velocity field are conventionally denoted by $u(\mathbf{X}, t)$ and $v(\mathbf{X}, t)$, whereas the vertical component of the velocity field is denoted by . We will similarly follow this convention throughout the present text. The velocity field can be written as

$$\mathbf{u}(\mathbf{X}, t) = \begin{pmatrix} u(\mathbf{X}, t) \\ v(\mathbf{X}, t) \\ w(\mathbf{X}, t) \end{pmatrix}. \tag{2.2}$$

A fundamental quantity which characterizes flows governed by these equations is the ratio of inertial forces to viscous forces, known as the *Reynolds number* and defined as

$$\mathrm{Re} = \frac{\mathcal{U}\,\mathcal{L}}{\nu},$$

where $\mathcal{U}$ and $\mathcal{L}$ are characteristic velocity and length scales of the fluid, and ν its kinematic viscosity [Pope, 2000]. For very high values of Re, that is, high fluid velocity coupled with relatively low viscosity, turbulent motions arise. The appearance of turbulence has a dramatic effect on the complexity of the flow when compared to non-turbulent (i. e., laminar) flows, as turbulent motions occupy a wide range of spatiotemporal scales. This fact is expressed by the principle of *energy cascading* in turbulent motion, whereby the energy put into the fluid by external forcings is first transferred to the largest eddies[2], which in turn transfer their energy to smaller eddies and so on, until the effects of viscosity stop this process and force the dissipation of energy into heat. The scale of the largest eddies, the so-called *integral scale* L, is determined by the domain size. On the other end of the spectrum

2 An *eddy* in fluid mechanics "eludes precise definition, but it is conceived to be a turbulent motion, localized within a region of size ℓ, that is at least moderately coherent over this region. The region occupied by a large eddy can also contain smaller eddies." [Pope, 2000]

is the *Kolmogorov scale* η, which was derived by A. N. Kolmogorov as

$$\eta = \left(\frac{\nu^3}{\varepsilon}\right)^{1/4},$$

where ε is the rate of dissipation of kinetic energy and ν is again the kinematic viscosity of the fluid [Kolmogorov, 1941b,a]. It can be shown that these scales are related:

$$\frac{L}{\eta} \sim Re^{3/4},$$

where Re is the fluids Reynold's number. The separation between the two extreme scales can thus be significant: several orders of magnitude for large Reynolds numbers. In the space between these two extremes, viscosity does not play a significant role, and the dynamics are thus dominated by inertial effects. This space is known as the *inertial subrange*.

Another fundamental result of the Kolmogorov theory of turbulence concerns the distribution of kinetic energy across these various spatial scales. The energy, it turns out, is not dispersed uniformly at random throughout the inertial subrange, but follows the distribution dictated by an *energy spectrum function* $E(k)$, which can be expressed as

$$\frac{1}{2}\langle u^2 + v^2 + w^2\rangle = \int_0^\infty E(k)\,dk.$$

The left hand side represents the turbulence kinetic energy[3], computed as the average over the three directional components of the velocity field. The wavenumber k corresponds to spatial frequency, with small wavenumbers associated with the largest eddies in the flow. Kolmogorov then showed, via dimensional analysis, that the spectrum function $E(k)$ must have the form

$$E(k) = C\varepsilon^{2/3}k^{-5/3}, \tag{2.3}$$

for some constant C. The key component in this expression is the exponent $-5/3$, which has been subject to numerous experimental verifications [Pope, 2000].

3 The expression $\langle Q\rangle$ is used here to represent the expectation or the average of a variable Q, for consistency with the literature on fluid mechanics and meteorology.

Analogously for the temporal scales, we can go from the *large-eddy turnover time* t_L, defined as $t_L = L/U$ for the mean velocity U, to the *Kolmogorov time scale* t_η, which is

$$t_\eta = \left(\frac{\nu}{\varepsilon}\right)^{1/2}.$$

These two extremes are again connected by the Reynold's number:

$$\frac{t_L}{t_\eta} \sim Re^{1/2}.$$

As a consequence of these fundamental principles, the computational cost for performing direct numerical simulations (DNS) of a turbulent PBL, where all the necessary spatiotemporal scales are explicitly resolved, is prohibitive. It is thus common practice to employ different approximations and simplifications of the underlying equations. One of the most widely-used approximations is the large-eddy simulation (LES), which is obtained by taking the convolution of the prognostic variables of the model (e.g. velocity, temperature) with a filtering kernel, which is usually taken to be a compactly-supported or rapidly-decaying smooth function. The resulting model equations are then discretized and solved. This has the effect of removing the smallest scales from the flow, which greatly reduces the computational cost. In order to maintain a realistic behaviour, however, the information contained in those scales must still be included. This is done by subgrid-scale modelling.

2.5 MODELING OF THE PBL

In this dissertation we will limit the data used for analysis to datasets obtained from numerical simulations of the PBL. These simulations solve a discretized version of the Boussinesq approximation to the Navier-Stokes equations (Equations 2.1), either directly or after applying a filtering kernel for an LES.

In all, data from three different simulations will be used: one DNS of a CBL growing into a linearly stratified atmosphere [Garcia and Mellado, 2014], and two LES models: the first is the Large-Eddy Simulation Atmosphere–Land-Surface Model [LES-ALM,

Table 2.1: Description of LES land-surface patterns.

SURFACE PATTERN	LAND USE	DOMINANT SCALE
SP1	Original	360 m
SP2	SP1 without small-scale features	1500 m
SP3	SP1 randomized	Δx (=60m)
SP4	Uniform pasture	∞

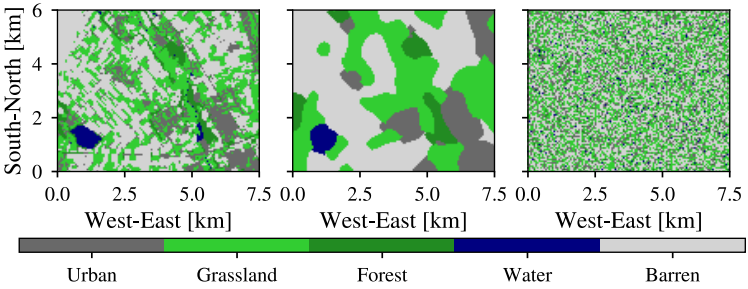


Figure 2.6: Surface patterns for the three cases SP1-SP3 with heterogeneous land use.

Shao *et al.*, 2013], which models a dry boundary layer coupled with an underlying land surface of varying degrees of heterogeneity. The second one is the Dutch Atmospheric Large-Eddy Simulation [DALES Heus *et al.*, 2010], which models a moist boundary layer and in this case is used to study the diurnal evolution of cloud size distribution for shallow convection over land. In all cases, the space discretization used is a sixth-order finite difference scheme. Both LES models use a third-order Runge-Kutta scheme for time discretization, while the DNS uses a fourth-order Runge-Kutta scheme.

We now briefly summarize the main model parameters of each simulation.

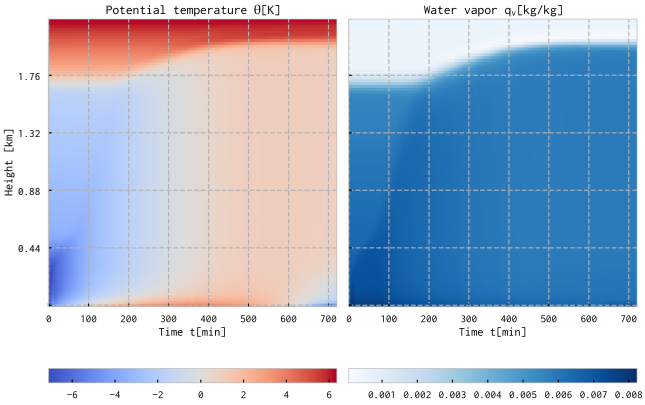


Figure 2.7: Evolution of potential temperature θ and water vapor mixing ratio q_v for simulation SP1.

DNS These simulations are carried out on a Cartesian grid of $5120\times5120\times840$ points, that is, 5120 points in each of the horizontal directions and 840 in the vertical. This grid has been subsampled and coarse-grained for our analysis. We will consider a subset of the grid formed by $512\times512\times234$ points. Out of the original 5000 timesteps computed, only 256 are used. These are taken at intervals of 20 close to the end of the simulation, which features a fully-developed CBL. The system is non-dimensionalized, so all time and space coordinates are given in relation to a characteristic length $L_0 = \sqrt{B_0\, N^{-3}}$ and characteristic time $T_0 = N^{-1}$, where N^2 is the background stratification and B_0 is the surface buoyancy flux. The initial conditions are thus a small random perturbation on the velocity, used to trigger the instability that will lead to turbulence, and the background profile $b(z) = N^2 z$ for the buoyancy (here the symbol z again denotes the vertical dimension in the grid, i.e. orthogonal to the direction of gravity). The simulations were ran by Cedrick Ansorge, with initial and boundary conditions as specified by Garcia and Mellado [2014].

LES-ALM The simulation domain features a cube of (7.5 x 6.0 x 2.2) km^3, discretized on a cubical grid of 125x100x100 cells, with horizontal resolution of 60 m, and the lowest grid point located at $z \simeq 2$ m. Periodic boundary conditions are prescribed for the lateral boundaries. The simulation period is 0800-2000 UTC, Aug. 5th 2009, with a timestep of 0.2 s, of which a regular subset of 720 timesteps is used (one for every 15 minutes). The LES simulations were ran by Michael Hintz using the surface model developed by Liu *et al.* [2017], using an implementation of the surface closure taking into account the fast response of the surface to atmospheric eddies. The initial state of the atmospheric model (wind speed, potential temperature, water mixing ratio, etc.) is estimated from a sounding at 08:00 UTC, 5 Aug 2009 and is characterized by a capping inversion around $z \simeq$ 1600 m with dried air aloft. At the surface, a weak inversion is found below $z \simeq 300$ m corresponding to weakly stable conditions preceding the onset of convection. Four different model runs are used, each with a different land surface pattern. These patterns, which we will denote by SP1-4, feature different degrees of heterogeneity as described in Table 2.1. SP1 is the original land use pattern from the Selhausen site in Western Germany, and was one of the designated study sites for the TR32 project; SP2 is as SP1 but with the small-scale features filtered out; SP3 is a fully randomized pattern. These three land surface patterns are illustrated in Figure 2.6, and all three have the same relative proportion of each land surface type. The fourth pattern, SP4, is a uniform grassland pattern (not shown here). Figure 2.7 shows the evolution of temperature and water vapor mixing ratio for SP1, averaged over horizontal slabs, throughout the entire simulation. The changes in these quantities, expressed as a function of height and time, illustrate the process of boundary layer growth over time (cf. Figure 2.1).

DALES The domain size is (12.8 x 12.8 x 5) km^3, with horizontal resolution of 50 m, vertical resolution of 40 m, and periodic boundary conditions in the lateral directions. The

land surface conditions are specified as a homogeneous land surface (grassland). Each simulation covers one day, and the data are saved at 15 min time intervals for analysis. The simulations were performed by Roel Neggers with a model setup as described by van Laar *et al.* [2019]. Several days are simulated, and the location corresponds to a site at mid-latitude, over land (continental Europe). Additional forcings were added to each model run, corresponding to large-scale weather processes affecting the model site at various points throughout each simulation day. The DALES model also features a coupled land–atmosphere scheme, but unlike LES-ALM this model is not designed to account for the effect of land surface heterogeneity on energy transfers between land surface and atmosphere. It does, however, include different cloud microphysics schemes, which allow it to model cloud formation, in contrast to the dry (i.e., no liquid water content) boundary layer simulated by LES-ALM.

3

TOPOLOGICAL BACKGROUND

Outline

This chapter gives an overview of the topological concepts used throughout the thesis, with an emphasis on breadth of scope rather than depth. After some historical remarks (Section 3.1) the basic definitions from point-set topology are reviewed in Section 3.2. Some important geometrical constructions are introduced in Section 3.3, which will form the basis for the discussion of homology theory in Section 3.4. Section 3.5 gives a brief discussion on the computable nature of homology, and Section 3.6 closes with an overview of persistent homology.

Topology is the branch of mathematics that deals with the abstract properties of space which remain unchanged through continuous transformations. Its main subfields are:

POINT-SET TOPOLOGY The study of the main properties of connectivity, compactness, and continuity from the perspective of set theory.

ALGEBRAIC TOPOLOGY As its name implies, it is a hybrid of algebra and topology: the study of topological properties of abstract spaces by using algebraic structures associated with these.

DIFFERENTIAL TOPOLOGY The study of topological properties of smooth manifolds, and differentiable maps defined on them.

COMPUTATIONAL TOPOLOGY This field of study is concerned with the study and development of algorithms for topological problems, as well as their application to other areas of mathematics and science.

3.1 HISTORICAL REMARKS

Topology has a long history: its genesis is usually traced back to the negative solution given by Leonhard Euler to the Seven Bridges of Königsberg problem. This solution was first presented in 1735 but published six years layer [Euler, 1741]. The problem setting is as follows: consider the layout of the Prussian city of Königsberg (now Kaliningrad, Russia) in the early 17th century, as shown in Figure 3.1. The city was built on both sides of the Pregel river (now Pregolya river), and included two islands. A total of seven bridges joined the islands with the mainland and each other. Is it then possible to walk through the city in such a way that each bridge is crossed once, and only once? The only admissible way to cross the river is by using one of the bridges, that is, swimming or sailing across the river are not allowed. The solution given by Euler to this problem rests on a key idea: all the information relevant to the problem is contained in the connections between the land masses and the seven bridges. In other words, the exact shape of the land masses and bridges, their sizes and relative positions do not matter. The paths traced out over land when moving from bridge to bridge are also irrelevant. In modern terms, we can model the city as a *graph*, shown in Figure 3.1 (right), and thus the question on the existence of a path through the city can be reformulated in terms of a question regarding the structure of this graph. Specifically, Euler showed that such a path exists on a graph G if and only if all vertices of G have even degree, which is not the case for the Königsberg graph.

The solution to the Seven Bridges problem was thus an important step towards the formalization of one of the key ideas in topology: connectivity. Further work in this direction centered on extending this idea to the study of geometric surfaces. The birth of *algebraic topology*, the subfield which we will concentrate on in this dissertation, only happened in the time between the 19th and 20th

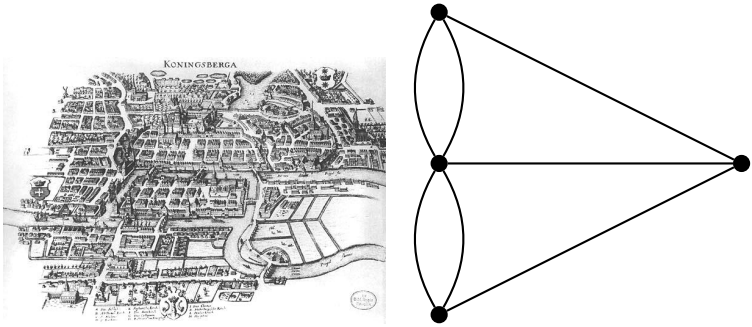


Figure 3.1: Left: map of Königsberg in 1652 showing the seven bridges that join the mainland and the two islands (image credit: Merian Erben, public domain). Right: graph representation of the seven bridges problem.

centuries, in the work of Henri Poincaré. The subject first developed along two directions: first, it was noted by Poincaré that all possible loops on a manifold can be given a group structure, with loop addition defined by traversing both loops in sequence. This group, called the *fundamental group*, turned out to be a topological invariant of the manifold [Poincaré, 1895], and this direction later developed into *homotopy theory*. The second direction, also explored by Poincaré, centered on the study of polyhedra, and under which conditions these can be used to approximate arbitrary closed manifolds. The polyhedra can thus provide a combinatorial representation of manifolds, which can be used to define a different type of topological invariant [Poincaré, 1895]. This second direction eventually led to *homology theory*. Both homotopy and homology are closely related, although the former tends to convey more topological information, while the latter has less intuitive definitions but is more amenable to computation. It is this relative ease of computation which makes homology one of the cornerstones of topological data analysis as it exists today. Consequently, we will only focus on homology in the sequel.

3.2 GENERAL TOPOLOGY

For the basic concepts of topology we refer to the book by Munkres [2000], on which the following presentation is based.

Definition 3.2.1. A topology on a set X is a collection τ of subsets of X satisfying the following properties:

1. $\varnothing, X \in \tau$.
2. The arbitrary union of elements of τ is again an element of τ.
3. The *finite* intersection of elements of τ is again an element of τ.

The pair (X, τ) is a *topological space*; the elements of τ are the *open sets* of X.

Definition 3.2.2. Let X be a set, and $\mathcal{B}$ an arbitrary collection of subsets of X which satisfies the following conditions:

1. For all $x \in X$, there exists at least one set $B \in \mathcal{B}$ such that $x \in B$.
2. If $x \in B_1 \cap B_2$ for some $x \in X$, $B_1, B_2 \in \mathcal{B}$, then there exists $B_3 \in \mathcal{B}$ such that $x \in B_3 \subset B_1 \cap B_2$.

Then $\mathcal{B}$ generates a topology $\tau_{\mathcal{B}}$ on X, defined as the collection of all unions of elements of $\mathcal{B}$. The collection $\mathcal{B}$ is also referred to as a basis for the topology $\tau_{\mathcal{B}}$.

This concept generalizes familiar properties of open sets in $\mathbb{R}$, the real line, to arbitrary spaces. Recall that an open interval (a, b) on the real line is the set of real numbers x such that $a < x < b$. These are the fundamental open sets in $\mathbb{R}$, in the sense that any generic open set can be expressed as a union of open intervals. Indeed, the set of all open intervals on $\mathbb{R}$,

$$\mathcal{B} = \{(a, b) \mid a, b \in \mathbb{R}, a < b\},$$

can be used to generate a topology on $\mathbb{R}$, the *standard topology*.

This simple characterization of a topology in terms of open sets allows us to state one of the most fundamental topological properties a space can have, and which will play a very important role in the remainder of this dissertation.

Definition 3.2.3. A set X that can be written as the union of two disjoint open sets, $X = U \bigcup W$ and $U \bigcap W = \varnothing$, is called *disconnected*. A set is *connected* if it is not disconnected.

Example. Let $X = (-1, 0) \cup (0, 1)$. Since X is clearly the union of two open sets with no points in common, it is a disconnected space.

Definition 3.2.4. Let (X, τ) be a topological space, and $Y \subset X$ a subset. The collection of sets defined by

$$\tau_Y = \{Y \cap U \mid U \in \tau\}$$

is the *subspace topology on* Y inherited from X.

Example. Let $(\mathbb{R}, \tau)$ be the real line with the standard topology, and $Y = [0, 1] \subset \mathbb{R}$. The subspace topology on Y is generated by the collection

$$\{(a, b) \cup [0, 1] \mid (a, b) \in \tau\}.$$

Topologies defined in these terms appear very abstract, but there is another way of defining a topology which is more intuitively clear.

Definition 3.2.5. Let X be a set. A *metric on* X is a function $d : X \times X \to \mathbb{R}$ such that

1. $d(x, y) \geqslant 0$ for all x, $y \in X$; $d(x, y) = 0$ if and only if $x = y$.
2. $d(x, y) = d(y, x)$ for all x, $y \in X$.
3. $d(x, y) \leqslant d(x, z) + d(z, y)$ for all x, y, $z \in X$.

A common example of such a function is the *Euclidean distance* on $\mathbb{R}^n$, defined by

$$d(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^2 \right)^{1/2}$$

for $x, y \in \mathbb{R}^n$. An important set defined in terms of a metric is the *open ε-ball* centered on a point $x \in X$. This set is defined as

$$B_\varepsilon(x) = \{z \in X \mid d(x,z) < \varepsilon\}.$$

The importance of open balls lies in the fact that they generate a topology on their ambient space X, called the *metric topology* induced by d. In the special case of $\mathbb{R}$, the open balls are all the sets of the form

$$B_\varepsilon(x) = (x - \varepsilon, x + \varepsilon).$$

The metric topology generated by these sets, τ_d, can be shown to be equivalent to the standard topology, τ_s, in the following sense: if (a, b) is a basis element in the standard topology, we can always find a basis element in the metric topology, $B_\varepsilon(x)$, such that $B_\varepsilon(x) \subset (a, b)$. Specifically, we only need to set $x = \frac{b-a}{2}$ and $\varepsilon < \frac{b-a}{4}$. Conversely, given any basis element $B_\varepsilon(x)$ in the metric topology, we can always find an open interval (a, b) such that $(a, b) \subset B_\varepsilon(x)$, for example $(x - \varepsilon/2, x + \varepsilon/2)$. Thus, by showing that both collections of basis elements are each one contained by the other, we show that each topology is *finer* than the other one, therefore both are equivalent.

We come to another way of constructing topological spaces which will be of importance in the sequel. These are the so-called *identification spaces*, which we can imagine as the result of taking a topological space and "gluing" several of its points together to obtain a new space. For example, if we take a line segment and glue its endpoints together, we would obtain a circle. The mathematically precise statement of this fact requires several concepts to be defined.

Definition 3.2.6. Let X be a set. A *binary relation* on X is a subset of its Cartesian product, $R \subset X \times X$.

Example. If X is the set of integers, $\mathbb{Z}$, define a relation R on it as follows: $a \, R \, b$, meaning a is related to b by R, if and only if $a = 2\,b$. This is the same as saying that the relation is the set of ordered pairs

$$\{(a, b) \mid a = 2\,b\}.$$

Definition 3.2.7. Let X be a set, and $\sim$ a binary relation on X. $\sim$ is an *equivalence relation* if the following conditions hold:

1. $a \sim a$ for all $a \in X$ (reflexivity).
2. $a \sim b \Rightarrow b \sim a$ (symmetry).
3. If $a \sim b$ and $b \sim c \Rightarrow a \sim c$ (transitivity).

Such a relation forms a partition of the set X into *equivalence classes*, with each element $a \in X$ belonging to one and only one equivalence class.

Example. Consider again the set of integers $\mathbb{Z}$, and the equivalence relation defined by congruence modulo 3, namely $a \sim b$ if and only if $a \equiv b \pmod 3$. This relation forms three equivalence classes:

$$\begin{aligned}
\bar{0} &= \{\ldots, -3, 0, 3, 6, 9, \ldots\} &&= \{3k \mid k \in \mathbb{Z}\} \\
\bar{1} &= \{\ldots, -2, 1, 4, 7, 10, \ldots\} &&= \{3k+1 \mid k \in \mathbb{Z}\} \\
\bar{2} &= \{\ldots, -1, 2, 5, 8, 11, \ldots\} &&= \{3k+2 \mid k \in \mathbb{Z}\}.
\end{aligned}$$

The symbol $\bar{a}$ in this case signifies that a is the *representative* of its equivalence class.

Definition 3.2.8. Given a topological space (X, τ), define a relation $\sim$ on X by $x \sim y$ if there exists a connected subspace of X which contains both x and y. This is an equivalence relation, and its equivalence classes are the *connected components* of X.

Definition 3.2.9. Let (X, τ) be a topological space, and $\sim$ an equivalence relation on X. The *quotient space* $X/_\sim$ is defined as the set of equivalence classes of $\sim$ on X equipped with the *quotient topology* $\tau_\sim$. This is the topology formed by all sets with an open preimage under the map

$$q : X \to X/_\sim$$

which sends a point x to its equivalence class in $X/_\sim$:

$$\tau_\sim = \{U \subset X/_\sim \mid q^{-1}(U) \in \tau\}.$$

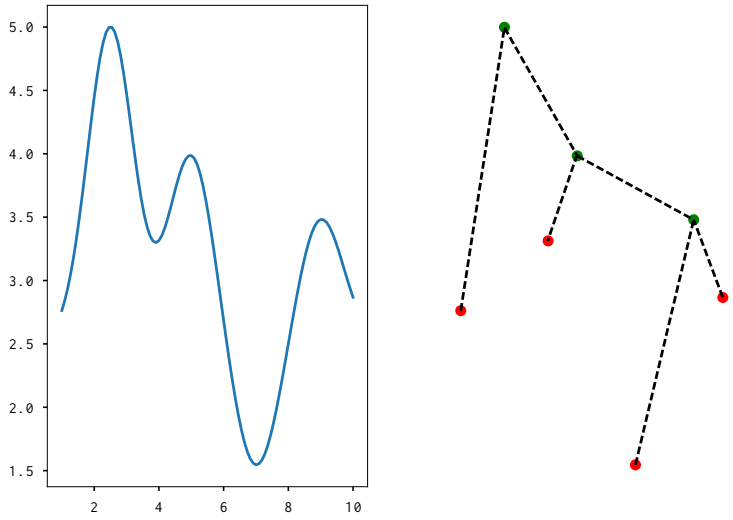
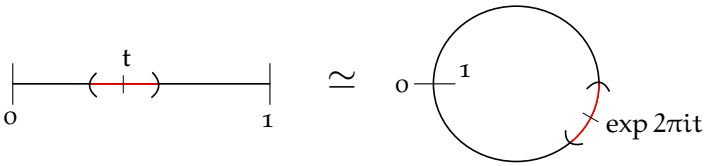


Figure 3.2: A merge tree (right) obtained from a scalar function on $\mathbb{R}$ (left).

Example. Consider the space $X = [0, 1]$ with the subspace topology inherited from the real line $\mathbb{R}$ with its standard topology. Define the following equivalence relation on X: $x \sim x$ for all $0 < x < 1$, and $0 \sim 1$. Each point $x \in (0, 1)$ is its own equivalence class, and there is a two-point equivalence class formed by $\{0, 1\}$. By effectively reducing these two points to a single entity, this has the same effect as the "gluing" of the endpoints referred to above, and indeed the quotient space $[0, 1]/_\sim$ is homeomorphic to the unit circle, S^1, via the map $\exp(2\pi i) : [0, 1] \to S^1$.

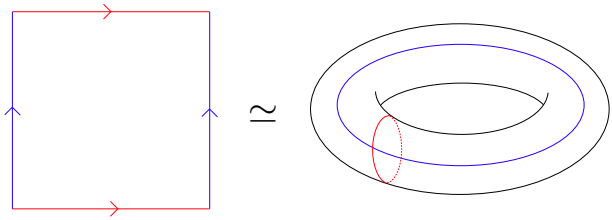


Example. Consider now the space $X = [0, 1] \times [0, 1]$, the unit square, and define an equivalence relation on it by $(x, y) \sim (x, y)$ for all (x, y) such that $0 < x,\ y < 1$, $(x, 0) \sim (x, 1)$ for all $x \in [0, 1]$,

and $(0, y) \sim (1, y)$ for all $y \in [0, 1]$. We again have the interior points of X as their own equivalence classes, and now identify each pair of opposite sides of the square with one another. We obtain the quotient space $[0, 1]^2/_\sim \cong (S^1)^2$, via the map

$$\phi(\alpha, \beta) = ((\sin(2\pi\alpha), \cos(2\pi\alpha)), (\sin(2\pi\beta), \cos(2\pi\beta)),$$

and as illustrated here, the two pairs of opposite sides become two independent loops on the surface of the torus. Independent here means that they cannot be continuously deformed into one another, in a sense to be made more precise in Section 3.4.



The last example of a quotient space presented here is of a different nature than the previous ones, as the starting object is different, but the idea behind it is the same as before.

Example. Consider a smooth scalar function, $f : \mathbb{R}^n \longrightarrow \mathbb{R}$, and the graph of f, defined by

$$\mathrm{gr}(f) = \{(x, f(x)) \mid x \in \mathbb{R}^n\}.$$

Define an equivalence relation on $\mathrm{gr}(f)$ by $x \sim y$ if and only if both points belong to the same level set of f, namely $f(x) = f(y)$, and to the same connected component of the sublevel set $f^{-1}(-\infty, f(x)]$. The quotient space $\mathrm{gr}(f)/_\sim$ is the *merge tree* of the function f. This is illustrated in Figure 3.2.

3.3 SIMPLICES, CUBES, AND COMPLEXES

We now describe two ways to construct combinatorial representations[1] of geometric objects, which depend on the joining of simpler

1 The use of the word *combinatorial* in this context refers to two key properties of the representation we seek: first, the geometric object must be represented by a

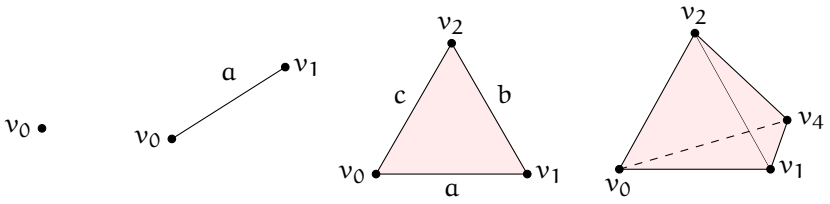
pieces under adequate conditions. These simpler pieces can be either simplices, which are the n-dimensional generalizations of triangles, or elementary cubes, the n-dimensional generalizations of cubes. The resulting representations are the simplicial or cubical complexes, respectively. For a more detailed treatment of these concepts, we refer to Munkres [1984, Ch. 1] for the case of simplicial complexes, and Kaczynski *et al.* [2004, Ch. 2] for the cubical case.

We first give the definitions of the atomic components.

Definition 3.3.1 (Simplex)**.** Let $u_0, u_1, \ldots, u_k \in \mathbb{R}^d$. An affine combination of the u_i is a point of the form $x = \sum_i \lambda_i u_i$, with $\lambda_i \in \mathbb{R}$ such that $\sum_i \lambda_i = 1$. The u_i are *affinely independent* if two affine combinations, $x = \sum_i \lambda_i u_i$, $y = \sum_i \mu_i u_i$, are equal if and only if $\lambda_i = \mu_i$ for $i = 0, 1, \ldots, k$. An affine combination is a *convex combination* if $\lambda_i \geqslant 0$ for all i. The *convex hull* of a set of points is the set of its convex combinations. A k-simplex is the convex hull of $k+1$ affinely independent points, $u_0, u_1, \ldots, u_k$. We say that the k-simplex is *spanned* by the u_i, and write it as

$$\sigma = [u_0, u_1, \ldots, u_k].$$

Example. We illustrate k-simplices, for $k = 0, \ldots, 3$. The 0-simplex $[v_0]$ is the one-point set $\{v_0\}$. The 1-simplex $[v_0, v_1]$ is a line, the 2-simplex $[v_0, v_1, v_2]$ is a triangle, and the 3-simplex $[v_0, v_1, v_2, v_3]$ is a tetrahedron.

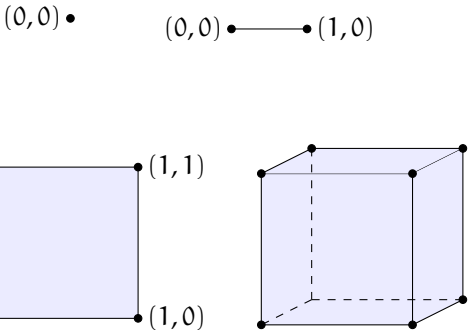


countable set of element, indeed a finite countable set if we intend to perform computations with it. Second, in this set we will have not only individual, point-like elements (i.e. the vertices of the object), but also combinations or permutations of these, which will represent structures such as edges, faces, and so on. This will become clearer once the definition of an abstract simplicial complex is introduced (Definition 3.3.7).

Definition 3.3.2 (Elementary cube)**.** An *elementary interval* is a closed interval $I \subset \mathbb{R}$, with $I = [l, l+1]$ or $I = [l, l] = [l]$, for some $l \in \mathbb{Z}$. The latter is referred to as a *degenerate* elementary interval. An *elementary cube* is a product of elementary intervals,

$$Q = I_1 \times I_2 \times \ldots \times I_d \subset \mathbb{R}^d.$$

Example. We now illustrate the cubical analogues of the low-dimensional simplices shown above. A 0-cube is again a one-point set, $[0] = \{(0,0)\}$. A 1-cube is an interval, in this case $[0,1]$. A 2-cube is a square, here $[0,1]^2$, and a 3-cube is a geometrical cube, in this case $[0,1]^3$.



Definition 3.3.3. Let $\sigma = [u_0, u_1, \ldots, u_k]$ be a k-simplex. The points u_i that span σ are the *vertices* of σ. The number k is the *dimension* of σ. A *face* of σ is a simplex spanned by any non-empty subset of the u_i. If this is a proper subset, the spanned simplex is a *proper face*. The union of all proper faces of σ is the *boundary of* σ, and is denoted by $\partial\sigma$ or $\operatorname{Bd}\sigma$.

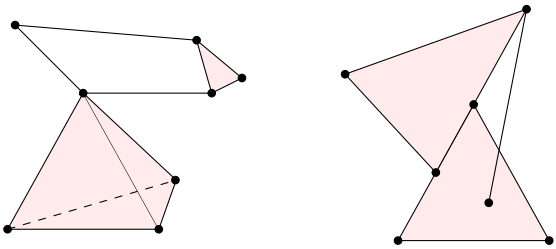
Definition 3.3.4. Let $Q \subset \mathbb{R}^d$ be an elementary cube. The *dimension* of Q is defined as the number of non-degenerate intervals that constitute Q. If P is any elementary cube such that $P \subset Q$, then P is a *face* of Q. If P is a proper subset of Q, then it is a *proper face*.

Definition 3.3.5 (Simplicial complex)**.** Let K be a (finite) collection of simplices in $\mathbb{R}^n$. K is a *simplicial complex in* $\mathbb{R}^n$ if the following two conditions hold:

1. If τ is the face of any simplex $\sigma \in K$, then $\tau \in K$.
2. If $\sigma_1, \sigma_2 \in K$, then $\sigma_1 \cap \sigma_2$ is either empty or a face of both σ_1 and σ_2.

If a subcollection $L \subset K$ is also a simplicial complex, it is a *subcomplex* of K. The collection of all simplices of K of dimension at most j is the j-*skeleton* of K. The points of the 0-skeleton of K are the *vertices* of K.

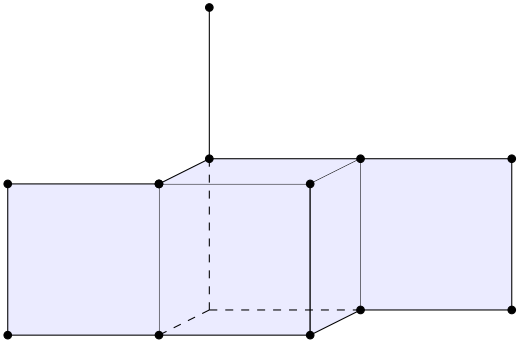
The picture shows a simplicial complex (left), formed by different simplices with vertices as their intersection points. The right figure shows a set formed by simplices which is not a simplicial complex, since the individual simplices intersect along proper subsets of their faces.



Definition 3.3.6 (Cubical complex)**.** A *cubical complex* X is a finite union of elementary cubes Q_i: $X = \bigcup_{i=1}^{k} Q_i$. A *subcomplex* A of a cubical complex X is a subset $A \subset X$ that is a cubical complex on its own, that is, it can again be expressed as a finite union of elementary cubes.

Observe that, by construction, the intersection of any pair of elementary cubes within a cubical complex is either empty or is a

face of both of them, since they all must lie on an integer-valued grid. This is illustrated in the following figure.



Finally, we define a concept which will become important soon.

Definition 3.3.7. An *abstract simplicial complex* is a finite collection of sets A such that $\sigma \in A$ and $\tau \subset \sigma$ implies $\tau \in A$. If K is a simplicial complex, and V its vertex set, let $\mathcal{K}$ be the collection of all subsets $\{u_0, u_1, \ldots, u_k\} \subset V$ that span any simplex of K. $\mathcal{K}$ is an abstract simplicial complex, the *vertex scheme* of K. Correspondingly, K is the *geometric realization* of $\mathcal{K}$.

3.4 SIMPLICIAL HOMOLOGY

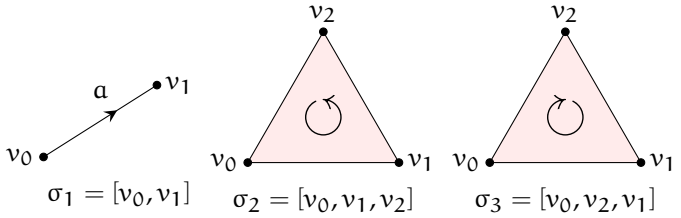
We will now present the main ideas behind homology theory for simplicial complexes. It is possible to define different homology theories emanating from different types of fundamental objects; for example, cubical homology for cubical complexes, or cellular homology for CW complexes. We will only present the simplicial version here for ease of exposition and to avoid repetition. The other case of interest to us, cubical homology, is developed along very similar lines, and can actually be shown to be equivalent to simplicial homology [for more details on this, see Kaczynski *et al.*, 2004, §11.2].

As was mentioned in the previous section, homology is easier to compute than homotopy, but this comes at the cost of a less intuitive mathematical machinery, which we will now go into.

Since the ultimate goal is to produce an algebraic structure associated with a given topological space, we start by defining how to operate on a simplicial complex. This requires us to, first of all, orient our simplices.

Definition 3.4.1. Let $\sigma = [u_0, u_1, \ldots, u_k]$ be a k-simplex, $k > 0$. An *orientation* of σ is a total order on its vertex set. Two orientations of σ are defined to be equivalent if they differ by an even permutation. This results in two different orientations being possible for any k-simplex if $k > 0$. If $k = 0$, only one orientation is possible.

Example. Here we show an oriented one-simplex as the directed line from v_0 to v_1, and the two simplex $[v_0, v_1, v_2]$ with its two possible orientations. Each one of these orientations corresponds to one of the two possible directions in which we can travel through the vertices of the triangle.



Definition 3.4.2. Let K be a simplicial complex, and $p \in \mathbb{N}$. A *p-chain* on K is a formal sum of p-simplices of K. Technically speaking it is a function c going from the set of oriented p-simplices of K to the integer numbers, such that the following two conditions hold:

1. If σ, σ' are opposite orientations of the same simplex, then $c(\sigma) = -c(\sigma')$.
2. $c(\tau) = 0$ for all but a finite number of oriented p-simplices τ.

A general p-chain is denoted by $c = \sum_i a_i \sigma_i$, whence it follows that c assigns the value $a_i \in \mathbb{Z}$ to the oriented p-simplex σ_i.

Two p-chains, $c = \sum_i n_i \sigma_i$ and $d = \sum_i m_i \sigma_i$, can be added componentwise:

$$c + d = \sum_i n_i \sigma_i + \sum_i m_i \sigma_i = \sum_i (n_i + m_i)\sigma_i.$$

With this operation, the set of all p-chains on K is a group, denoted by $C_p(K)$. If $p < 0$, or p is larger than the dimension of K, the group $C_p(K)$ is trivial.

An important observation must be made at this point. We have just defined p-chains on a simplicial complex K as functions going from the set of oriented p-simplices of K to $\mathbb{Z}$, the integer numbers. This is indeed the usual definition of p-chains, and using it will lead us to homology with coefficients in $\mathbb{Z}$. This is not the only possibility, as we can use coefficients in any ring or field. Here, as we are ultimately looking for a computational use of homology, we will use $\mathbb{Z}_2$ coefficients. $\mathbb{Z}_2$, sometimes also denoted $\mathbb{Z}/2\mathbb{Z}$, is the set of integers modulo 2, and greatly simplifies computations[2].

Recall (Definition 3.3.3) that the boundary of a simplex is the union of its proper faces. We now formalize this using p-chains.

Definition 3.4.3 (Boundary map)**.** Let $\sigma = [u_0, u_1, \ldots, u_k]$ be a k-simplex. The union of its proper faces can be expressed as the sum of the (k-1)-chains defined on these faces, namely

$$\partial_k \sigma = \sum_{i=0}^{k} [u_0, \ldots, \hat{u}_i, \ldots, u_k],$$

where $[u_0, \ldots, \hat{u}_i, \ldots, u_k]$ represents the (k-1)-simplex spanned by the vertices of σ with u_i deleted or "left out". Being a sum of (k-1)-simplices, the boundary of σ is a (k-1)-chain. Moreover, for a general k-chain $\sum a_i \sigma_i$, its boundary is the sum of the boundaries of its components,

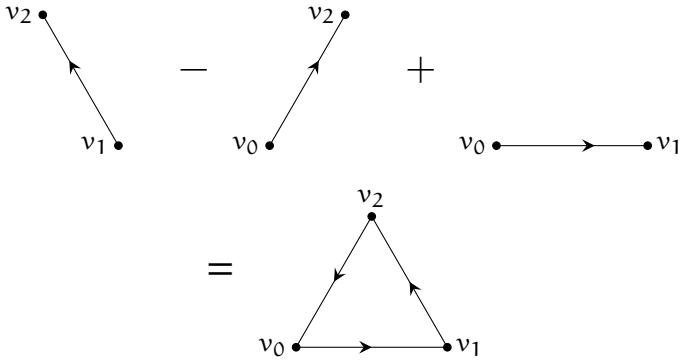
$$\partial_k c = \sum a_i \partial_k \sigma_i,$$

2 That is, $\mathbb{Z}_2 = \{0, 1\}$, with addition defined as

$$0 + 0 = 0$$
$$0 + 1 = 1 + 0 = 1$$
$$1 + 1 = 0$$

and it holds that $\partial_k(c + d) = \partial_k c + \partial_k d$ for any k-chains c, d. The map $\partial_k : C_k \to C_{k-1}$ thus defines a group homomorphism, the *boundary map*.

Example. We can illustrate the computation of the boundary of a 2-simplex, namely $\partial\,[v_0, v_1, v_2] = [v_1, v_2] - [v_0, v_2] + [v_0, v_1]$. The minus sign amounts to a change of orientation (see Definition 3.4.2).



We were able to disregard the orientation of all simplices involved in this by virtue of the $\mathbb{Z}_2$ coefficients being used.

Proposition 3.4.1. $\partial_{k-1} \circ \partial_k = 0$.

Proof. Consider an arbitrary k-simplex, $\sigma = [v_0, v_1, \ldots, v_k]$. By definition, its boundary is

$$\partial_k \sigma = \sum_{i=0}^{k} (-1)^i [v_0, \ldots, \hat{v}_i, \ldots, v_k].$$

Thus

$$\partial_{k-1}\partial_k \sigma = \sum_{i=0}^{k} (-1)^i \partial_{k-1} [v_0, \ldots, \hat{v}_i, \ldots, v_k],$$

and in general we have that

$$\partial_{k-1} [v_0, \ldots, \hat{v}_i, \ldots, v_k] = \sum_{j<i} (-1)^j [v_0, \ldots, \hat{v}_j, \ldots, \hat{v}_i, \ldots, v_k] + \sum_{j>i} (-1)^{j-1} [v_0, \ldots, \hat{v}_i, \ldots, \hat{v}_j, \ldots, v_k].$$

Therefore

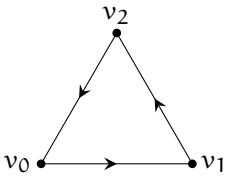
$$\partial_{k-1}\partial_k\sigma = \sum_{j<i}(-1)^i(-1)^j[v_0,\ldots,\hat{v}_j,\ldots,\hat{v}_i,\ldots,v_k] + \sum_{j>i}(-1)^i(-1)^{j-1}[v_0,\ldots,\hat{v}_i,\ldots,\hat{v}_j,\ldots,v_k],$$

where summing over both indices results in each term appearing twice, each time with opposite sign. □

The geometric notion of boundary allows us to distinguish two important types of chains, which also play a central role in the algebraic construction of the homology group.

Definition 3.4.4. Let K be a simplicial complex, and $c \in C_p(K)$. If c has an empty boundary, that is if $\partial_p c = 0$, it is a *p-cycle*. The set of all p-cycles on K is the kernel of the boundary map $\partial_p : C_p(K) \to C_{p-1}(K)$, and as such is a subgroup of $C_p(K)$, denoted by $Z_p(K) = \ker \partial_p$.

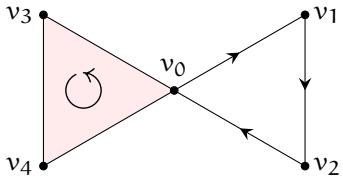
Example. The boundary of a 2-simplex, as computed above, is an example of a cycle.



Definition 3.4.5. Let K be a simplicial complex, and $c \in C_p(K)$. If there exists $d \in C_{p+1}(K)$ such that $c = \partial_{k+1} d$, then c is a *p-boundary*. The set of all p-boundaries on K is the image of the boundary map $\partial_{p+1} : C_{p+1}(K) \to C_p(K)$, and as such is a subgroup of $C_p(K)$, denoted by $B_p(K) = \operatorname{im} \partial_{p+1}$.

Example. The following picture illustrates a simplicial complex K formed by a 2-simplex and a 1-cycle intersecting at a point. The

cycle is not a boundary of anything, whereas the boundary of the 2-simplex would be an element of the group $B_1(K)$.



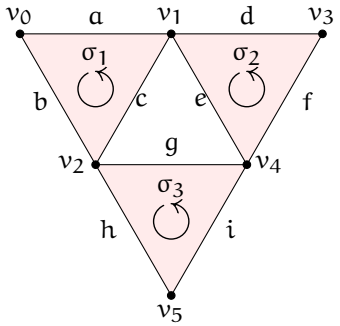
Proposition 3.4.2. *The group of p-boundaries,* $B_p(K)$*, is a subgroup of the group of p-cycles,* Z_p*.*

Proof. Follows immediately from Proposition 3.4.1. □

We now define an important relationship between p-chains as follows.

Definition 3.4.6. Let $c, c' \in C_p$. If there exists $d \in C_{p+1}$ such that $c - c' = \partial_{p+1} d$, the two chains c and c' are said to be *homologous*. If $c = \partial_{p+1} d$, then c is *homologous to zero*. Alternatively, we say that c is a *bounding cycle*.

Example. Consider the following simplicial complex K:



which is formed by three 2-simplices joined pairwise at different vertices. The group of 1-boundaries, B_1, is generated here by the three boundaries of these 2-simplices. These are also elements of the group of 1-cycles, Z_1. We also have a 1-cycle, $c + g + e$, which is not the boundary of anything. Moreover, if we form the new

cycle $(c, g, e) + (d, e, f)$, which can be seen as traversing first the non-bounding cycle and then the boundary of a 2-simplex, we can see that this is homologous to $c + g + e$, since their difference is an element of B_1. This is the algebraic expression of the geometrical fact that we can continuously deform the cycle $(c + g + e) + (d + e + f)$ to $c + g + e$ within the simplicial complex K.

Proposition 3.4.3. *The relation $\sim$, with $c \sim c'$ if they are homologous, is an equivalence relation.*

Proof. Reflexivity follows from the fact that, for any k-cycle $z \in Z_k$, $z = z + 0$, where 0 represents the trivial boundary. If two k-cycles are homologous, $z_1 - z_2 = \partial d$ for some $d \in B_{k+1}$. But then

$$z_2 - z_1 = -\partial d = \partial(-d),$$

which shows the symmetry. Finally, if $z_1 - z_2 = \partial d_1$ and $z_2 - z_3 = \partial d_2$, then

$$(z_1 - z_2) - (z_2 - z_3) = z_1 - z_3 = \partial d_1 - \partial d_2 = \partial(d_1 - d_2).$$

□

Since the p-boundaries of a simplicial complex K are a subgroup of its p-cycles, we can form the quotient group Z_p/B_p. We have thus arrived at the central concept in this section:

Definition 3.4.7. The p-th *homology group* of a simplicial complex K is

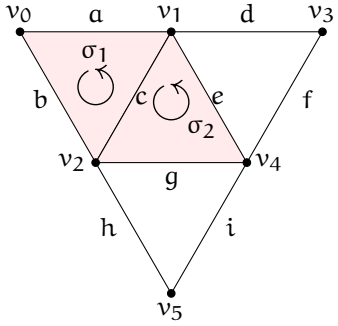
$$H_p(K) = Z_p(K)/B_p(K).$$

The elements of the p-th homology group $H_p(K)$ are, by definition, the cosets of B_p in Z_p. One such coset would have the form

$$z + B_p = \{z + d \mid d \in B_p\},$$

for an arbitrary $z \in Z_p$. If we take a second cycle belonging to the same coset, say $z' = z + d'$ with $d' \in B_p$, then we can see that $z' + B_p = z + d' + B_p = z + B_p$, thus we get the same coset. But $z - z' = d' \in B_p$, so z and z' are homologous, as per Definition 3.4.6. We can thus see that the elements of the p-th homology group are equivalence classes of p-cycles which differ by a p-boundary.

Example. Let the simplicial complex K now be as illustrated here:



There are now two 2-simplices joined at a face, and a series of 1-simplices which form two different 1-cycles: $z_1 = e + f + d$ and $z_2 = h + i + j$. If we consider the boundaries of the 2-simplices, namely $b_i = \partial\sigma_i$ $(i = 1, 2)$, we can see that $z_1 \sim z_1 + b_1 \sim z_1 + b_2$, and similarly that $z_2 \sim z_2 + b_1 \sim z_2 + b_2$. It is not the case, however, that z_1 is homologous to z_2. We have thus shown that each of the non-bounding cycles has its own homology class, and thus they are both generators of the first homology group H_1.

For a finite simplicial complex K, the p-chain group has finite rank, where the rank of a group G is defined as the number of generators of G. This then implies that the p-cycle group Z_p, being a subgroup of C_p, also has finite rank. The p-th homology group is thus a finitely generated abelian group. These groups are fully characterized in the following sense:

Theorem 3.4.4. *Let* G *be a finitely generated abelian group. Then* G *is isomorphic to*

$$(\mathbb{Z} \oplus \mathbb{Z} \oplus \ldots \oplus \mathbb{Z}) \oplus \mathbb{Z}_{t_1} \oplus \mathbb{Z}_{t_2} \oplus \ldots \oplus \mathbb{Z}_{t_k}.$$

Proof. See Theorem 4.3 in Munkres [1984, §1.4]. □

Based on this result, we can define a very important concept:

Definition 3.4.8. Let G be a finitely generated abelian group, with decomposition given by

$$G \simeq (\mathbb{Z} \oplus \mathbb{Z} \oplus \ldots \oplus \mathbb{Z}) \oplus \mathbb{Z}_{t_1} \oplus \mathbb{Z}_{t_2} \oplus \ldots \oplus \mathbb{Z}_{t_k}.$$

Then the number of copies of $\mathbb{Z}$ in the decomposition, denoted by β, is the *Betti number* of G. The $t_1, \ldots, t_k$ are the *torsion coefficients* of G, and are such that $t_1 \mid t_2 \mid \ldots \mid t_k$.

The Betti number of a homology group $H_p(K)$, denoted by β_p, indicates the number of generators of p-dimensional homology in the simplicial complex K. Referring to our discussion above, a generator of p-dimensional homology is to be understood as the equivalence class corresponding to a non-bounding p-cycle in the complex. In this sense, 1-dimensional homology is generated by the holes or loops present in K; 2-dimensional homology is generated by the voids or cavities enclosed by K; higher-dimensional homology is harder to interpret geometrically, but the underlying principle is the same.

The zero-dimensional homology group is special, in the sense specified by the following result:

Theorem 3.4.5. *Let K be a simplicial complex. Its zero-dimensional homology group $H_0(K)$ is a free abelian group. If $\{v_i\}$ is a set consisting of one vertex from each connected component of K, then the homology classes of the chains v_i are a basis for $H_0(K)$.*

The zeroth Betti number, β_0, is thus a count of the connected components that make up K. Together, the set of Betti numbers of a simplicial complex K are topological invariants of K: if K' is another simplicial complex homeomorphic to K, then their Betti numbers coincide. These numerical invariants are usually not the most important property of a space in algebraic topology, but for us they will be a central object of study. It is also important to note here the fact that, for any simplicial complex K, its homology is completely determined by its vertex scheme, which is a combinatorial object, and therefore independent of its geometric realization.

3.5 ON THE COMPUTABILITY OF HOMOLOGY

In the previous section we were intentionally vague regarding the actual computation of homology groups. As mentioned before, we are not so much interested in the algebraic structure of homology

groups, but rather on their numerical invariants represented by the Betti numbers. Computing these numbers can be simpler than computing the group itself, as we will now show.

Let K be a finite simplicial complex. Recall that we are using $\mathbb{Z}_2$ coefficients, so that a p-chain on K is a sum $\sum a_i \sigma_i$, where $a_i \in \{0, 1\}$. Suppose, moreover, there are a total of n p-simplices in K. The group of p-chains on K, $C_p(K)$, will then have cardinality 2^n. This is easy to see: imagine the set of p-simplices as a collection of binary flags or switches, which then makes a p-chain into a specific setting of the switches, with individual simplices being either "on" or "off". Each of the p-simplices of K is thus a generator of $C_p(K)$. The cardinality of the group is its *order*, and we denote it by $\operatorname{ord} C_p(K)$.

The group $C_p(K)$ is isomorphic to $\mathbb{Z}_2^n$, which is an n-dimensional vector space. Its dimension, also called its *rank*, is then the number of its generators, namely

$$\operatorname{rank} C_p(K) = \log_2 \operatorname{ord} \mathbb{Z}_2^n = n.$$

Recall that a homology class in $H_p(K)$ is of the form $z + B_p(K)$ for some $z \in Z_p(K)$. The number of cycles in this class is equal to the number of p-boundaries, that is, $\operatorname{ord} B_p$. Since each element of $Z_p(K)$ appears in exactly one of the cosets of B_p (by definition of the quotient group), we then have

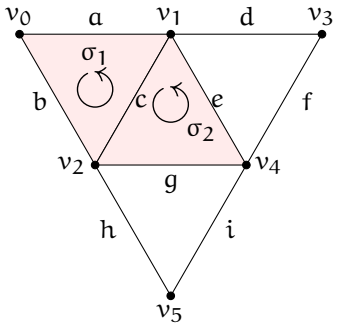
$$\operatorname{ord} H_p(K) = \operatorname{ord} Z_p(K) / \operatorname{ord} B_p(K),$$

or equivalently,

$$\begin{aligned} \operatorname{rank} H_p(K) &= \log_2 \operatorname{ord} H_p(K) \\ &= \log_2 (\operatorname{ord} Z_p(K) / \operatorname{ord} B_p(K)) \\ &= \log_2 \operatorname{ord} Z_p(K) - \log_2 \operatorname{ord} B_p(K) \\ &= \operatorname{rank} Z_p(K) - \operatorname{rank} B_p(K) \\ &= \beta_p. \end{aligned}$$

We thus see that the Betti numbers of a simplicial complex can be computed as the difference in the rank of its p-cycle group and its p-boundary group.

Example. To illustrate this, we consider again the simplicial complex K we encountered before:



It contains nine 1-simplices, thus

$$\begin{aligned} \operatorname{ord} C_1 &= 2^9 \\ \operatorname{rank} C_1 &= 9. \end{aligned}$$

These simplices form four 1-cycles, thus

$$\begin{aligned} \operatorname{ord} Z_1 &= 2^4 \\ \operatorname{rank} Z_1 &= 4, \end{aligned}$$

and each 1-simplex has a boundary, thus

$$\begin{aligned} \operatorname{ord} B_1 &= 2^2 \\ \operatorname{rank} B_1 &= 2. \end{aligned}$$

But then

$$\begin{aligned} \operatorname{ord} H_1 &= \operatorname{ord} Z_1 / \operatorname{ord} B_1 = 2^4/2^2 = 4 \\ \operatorname{rank} H_1 &= \operatorname{rank} Z_1 - \operatorname{rank} B_1 = 4 - 2 = 2. \end{aligned}$$

This tells us that there must be four homology classes in the first homology group of K. Two of these, as we saw before, are those corresponding to each of the two non-bounding cycles. The third one is the trivial class, and the fourth one is that formed by the addition of the two cycles, $z_1 + z_2$. There are thus only two generators of H_1, namely z_1 and z_2 themselves. Thus the Betti number, $\beta_1 = 2$, corresponds to the number of "holes" in K, each one represented by one of these two cycles.

Suppose now that there are n_p p-simplices and n_{p-1} (p-1)-simplices in a simplicial complex K, thus making $C_p(K)$ and $C_{p-1}(K)$ vector spaces of dimension n_p and n_{p-1} respectively. The boundary map between them, $\partial_p : C_p(K) \to C_{p-1}(K)$ is a linear map between two finite-dimensional vector spaces, and as such can be represented by a matrix. Assume an arbitrary ordering on both the set of p-simplices, $\{\sigma_1, \ldots, \sigma_p\}$ and the set of (p-1)-simplices, $\{\tau_1, \ldots, \tau_{p-1}\}$. The *boundary matrix* for ∂_p is defined by $D_p = (d_{ij})$, where

$$d_{ij} = \begin{cases} 1 & \text{if } \tau_i \text{ is a face of } \sigma_j, \\ 0 & \text{otherwise.} \end{cases} \tag{3.1}$$

The boundary of an arbitrary p-chain $c = \sum a_i \sigma_i$, represented by its vector of coefficients $\bar{a} = (a_1, \ldots, a_{n_p})$ can then be computed by multiplication: $\partial_p c = D_p \bar{a}$. The number of columns of D_p is equal to the rank of $C_p(K)$), and its number of rows is equal to the rank of $C_{p-1}(K)$.

We recall the following fundamental result from linear algebra:

Theorem 3.5.1 (Rank-nullity theorem)**.** *Let V, W be finite-dimensional vector spaces, and $T : V \to W$ a linear transformation. Then*

$$\dim V = \dim(\operatorname{im} T) + \dim(\ker T).$$

In particular, this result implies that, if $z_p = \operatorname{rank} Z_p(K)$ and $b_{p-1} = \operatorname{rank} B_{p-1}(K)$, then

$$n_p = z_p + b_{p-1}. \tag{3.2}$$

It is possible to reduce the boundary matrix D_p to *Smith normal form*, which in the case of $\mathbb{Z}_2$ coefficients is

$$D'_p = \left[\begin{array}{ccc|c} 1 & & O & \\ & \ddots & & O \\ O & & 1 & \\ \hline & O & & O \end{array}\right]. \tag{3.3}$$

This matrix has b_{p-1} non-zero columns to the left, one for each basis element of the vector space $B_{p-1}(K)$, and z_p zero columns to

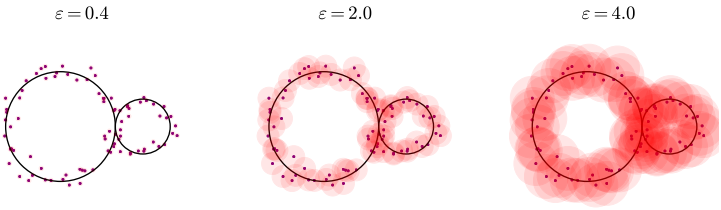


Figure 3.3: 85 points drawn with noise from a manifold $\mathcal{M}$, two circles joined at a point. When taking the union of balls of radius ε centered at these points, for various values of ε, different approximations to $\mathcal{M}$ can be obtained.

the right, one for each basis element of $Z_p(K)$. Obtaining the Betti numbers of a simplicial complex K thus amounts to constructing the necessary boundary matrices, and reducing them to their normal forms.

3.6 PERSISTENT HOMOLOGY

One of the guiding ideas behind topological data analysis is that we can think of a point cloud in $\mathbb{R}^n$ not only as a set of discrete points with trivial homology, but we can also see it as representing a higher-dimensional object, and use its associated topological information to make inference about the processes and structure underlying noisy data.

This is the original idea behind persistent homology (PH), which first appeared in the seminal works of Robins [2002] and Edelsbrunner *et al.* [2002], and it is based on finding not a single representation for this higher-dimensional object, but rather a large collection of representations, such that the necessary information can be inferred from the relationships between them. Before giving the formal definitions, we will give a motivating example.

Consider the manifold $\mathcal{M} \subset \mathbb{R}^2$ shown in Figure 3.3, the wedge sum of two circles. Also shown is a set of points $\{x_i\}$ sampled with noise from $\mathcal{M}$. As before, the first step is to convert the set

of points to a geometric object from which topological invariants are actually computable. The difference is that now there is no recourse to the concept of anchor points in order to build a cubical complex, because the points no longer lie on a regular grid. Instead, we consider the union of closed balls of radius ε centered on each of the points x_i, for a given value $\varepsilon > 0$:

$$\mathcal{B}_\varepsilon = \bigcup_{i=1}^{n} B_\varepsilon(x_i).$$

In this case the distance parameter ε will play the same role played by the threshold value in the construction of a cubical complex from point data. For $\varepsilon = 0.4$, the set $\mathcal{B}_\varepsilon$ consists of mostly non-intersecting balls, save for a few cases in which the points in $\{x_i\}$ are pairwise closer than $r = 0.4$. $\mathcal{B}_\varepsilon$ therefore still has the homotopy type of a finite set of points, and would therefore give us no useful information about the underlying manifold $\mathcal{M}$. With $\varepsilon = 2$ we can see that $\mathcal{B}_\varepsilon$ is now a single connected component which overlaps the smaller circle, and indeed it is no longer a contractible space. There is a small gap in the lower right part of the larger circle, so we still don't recover the true homotopy type of $\mathcal{M}$. Increasing the parameter to $\varepsilon = 4$ now yields a topologically correct representation of the larger circle, but the smaller one is destroyed. One might ask what the optimal value of ε is in this case: a value $\varepsilon^* \in \mathbb{R}$ such that the corresponding set $\mathcal{B}_{\varepsilon^*}$ has exactly the same homotopy type as $\mathcal{M}$. This is in general unknown for arbitrary datasets, and depending on the structure of the noise in the sampling process such an optimal value can be very hard or impossible to find. The answer given by PH is that it is not necessary to find this optimum, and that the relevant information can be inferred by computing the topological features at different scales, as well as the relationships between them.

It can be computationally prohibitive to work with unions of closed balls, but it is possible to use other, simpler geometric representations in the form of abstract simplicial complexes.

Definition 3.6.1 (Čech complex)**.** Let $\mathcal{P} = \{x_1, x_2, \ldots, x_n\}$ be a set of points in a metric space (X, d), and $\varepsilon > 0$. Construct an abstract simplicial complex by taking the points $x_i \in \mathcal{P}$ as the

0-simplices, and adding all k-simplices $[x_{i_0}, \ldots, x_{i_k}]$ such that $\cap_{j=0}^{k} B_\varepsilon(x_{i_j}) \neq \varnothing$, for $k = 1, \ldots, n$. This is the *Čech complex* for $\mathcal{P}$ and ε, denoted by $C(\mathcal{P}, \varepsilon)$.

The following classical result shows the utility of the Čech complex in manifold reconstruction. For a more thorough discussion, see [Edelsbrunner and Harer, 2010, §III.2].

Theorem 3.6.1 (Nerve theorem)**.** *Let $E = \{E_1, \ldots, E_m\}$ be a finite collection of closed, convex subsets of $\mathbb{R}^n$. Then the abstract simplicial complex defined by all non-empty subcollections of E, $\{E_{i_0}, \ldots, E_{i_k}\}$, such that $\cap_{j=0}^{k} \neq \varnothing$, has the same homotopy type as $\cup_{i=1}^{m} E_i$.*

Proof. See McCord [1967]. □

In particular, the sets used in the definition of the Čech complex, closed balls in $\mathbb{R}^n$, are convex sets. Thus the Čech complex has the same homotopy type as the union of these balls. This implies that, if the points in $\mathcal{P}$ are sampled from an underlying manifold X with sufficient density, and if the value of ε is chosen judiciously, then we can indeed recover the homotopy type of X from the associated Čech complex.

This result is in the same spirit as the discussion surrounding the relationship between abstract simplicial complexes and their geometric realizations given in Section 3.4, in the sense that the relevant topological information about a given space X is contained in a combinatorial structure, in this case a Čech complex. In practice, however, computing this complex can be a challenging task, as we would require to keep track of a very large number of point subsets for high-dimensional simplices. Another combinatorial structure which is computationally simpler is the following:

Definition 3.6.2 (Vietoris-Rips complex)**.** Let $\mathcal{P} = \{x_1, x_2, \ldots, x_n\}$ be a set of points in a metric space (X, d), and $\varepsilon > 0$. Construct an abstract simplicial complex by adding the points of $\mathcal{P}$ as its 0-simplices, and all k-simplices $[x_{i_0}, \ldots, x_{i_k}]$ whenever $d(x_{i_l}, x_{i_m}) < \varepsilon$, for all $0 \leqslant l < m \leqslant k$, and $k = 1, \ldots, n$. This is the *Vietoris-Rips complex*, or simply *Rips complex*, for $\mathcal{P}$ and ε, denoted by $R(\mathcal{P}, \varepsilon)$.

The main advantage of the Rips complex over the Čech complex is the fact that the former is fully specified by the pairwise

distances of points in $\mathcal{P}$, even if the number of simplices in the total complex can still be very large. In general, given a set of points $\mathcal{P}$ and a value $\varepsilon > 0$, the complexes $C(\mathcal{P}, \varepsilon)$ and $R(\mathcal{P}, \varepsilon)$ will be different, but can coincide in some cases, such as when the metric space is $\mathbb{R}^n$ equipped with the L^∞ metric instead of the usual Euclidean metric [Adler *et al.*, 2010]. In light of the Nerve Theorem, we can see that the Rips complex will tend to introduce gratuitous homological information. There is a positive result that connects the two kinds of complexes:

Theorem 3.6.2. *Let $\mathcal{P}$ be a finite set of points in Euclidean space $\mathbb{R}^n$, and $C(\mathcal{P}, \varepsilon)$, $R(\mathcal{P}, \varepsilon)$ its associated Čech and Rips complexes for some $\varepsilon > 0$. Then, if $\varepsilon' > 0$ is such that*

$$\frac{\varepsilon'}{\varepsilon} \leqslant \sqrt{\frac{2\,d}{d+1}},$$

the following chain of inclusions holds:

$$R(\mathcal{P}, \varepsilon) \hookrightarrow C(\mathcal{P}, \varepsilon') \hookrightarrow R(\mathcal{P}, \varepsilon').$$

Proof. See de Silva and Ghrist [2007]. □

The idea behind this result is that, despite sacrificing topological accuracy for computational simplicity in going from the Čech complex to the Rips complex, it is still possible to recover the "true" homology represented by the Čech complex by taking an appropriate sequence of Rips complexes. Specifically, we know that any homology class that appears in both $R(\mathcal{P}, \varepsilon)$ and $R(\mathcal{P}, \varepsilon')$ must also be present in $C(\mathcal{P}, \varepsilon)$ under the inclusions shown above. This brings us to the central idea behind persistence: knowing the changes in homology induced by an inclusion, in this case $R(\mathcal{P}, \varepsilon) \hookrightarrow R(\mathcal{P}, \varepsilon')$, can be more informative than merely knowing the homology of either complex independently.

Recall that for a given simplicial complex K, its chain complex is

$$\cdots \xrightarrow{\partial_{p+2}} C_{p+1} \xrightarrow{\partial_{p+1}} C_p \xrightarrow{\partial_d} C_{p-1} \xrightarrow{\partial_{p-1}} \cdots,$$

and its p-th homology group is defined as $H_p(K) = Z_p(K)/B_p(K)$. The elements of the $H_p(K)$ are the equivalence classes of non-

bounding p-cycles in K. Consider now a nested sequence of simplicial complexes indexed by a parameter i,

$$K_0 \subset K_1 \subset K_2 \subset \ldots \subset K_n = K.$$

This is a *filtration* or a *filtered complex*, and at each step we will have an inclusion $\iota : K_{i-1} \hookrightarrow K_i$. More generally, for $p \in \mathbb{N}$ and $i \leqslant j$, the inclusion $\iota : K_i \hookrightarrow K_j$ induces a homomorphism of the homology groups:

$$f_p^{i,j} : H_p(K_i) \to H_p(K_j).$$

Definition 3.6.3 (Persistent homology group)**.** The p*-th persistent homology groups* for a filtered complex K, denoted by $H_p^{i,j}$, are the images of the induced homomorphisms,

$$H_p^{i,j} = \operatorname{im} f_p^{i,j}.$$

The ranks of these groups are the *persistent Betti numbers* $\beta_p^{i,j}$,

$$\beta_p^{i,j} = \operatorname{rank} H_p^{i,j}.$$

This definition makes sense because a continuous map between topological spaces maps cycles to cycles and boundaries to boundaries (cycles can become boundaries, all boundaries are cycles). The inclusion is a continuous map. For a simplicial map $f : |K| \to |L|$, $p \in \mathbb{N}$, there is an induced homomorphism between chain groups

$$f_\# : C_p(K) \to C_p(L),$$

which in turn induces a homomorphism on homology groups

$$f_* : H_p(K) \to H_p(L).$$

Observe that $H_p^{i,i} = H_p(K_i)$, so the persistent homology groups generalize the ordinary homology groups.

Assume a homology class $\alpha \in H_p(K_i)$. If it happens that it is the first time that it appears in the filtration, that is to say $\alpha \notin H_p^{i-1,i}$, then we say that α is *born at* i. Similarly, if α is born at i and

merges with an older class when passing from $j-1$ to j, then both of the following hold:

$$f_p^{i,j-1}(\alpha) \notin H_p^{i-1,j-1} \text{ (α is born at i)}$$
$$f_p^{i,j}(\alpha) \in H_p^{i-1,j} \text{ (α merges into the image of an older class).}$$

The first condition ensures that $j-1$ is indeed the last moment in which α exists as an independent class.

Definition 3.6.4. The *persistence* of a given homology class α that is born at i and dies entering j is defined as $j-i$. The *persistence interval* of α, denoted by $[b_\alpha, d_\alpha)$, is equal to $[i, j)$. The endpoints are the birth and death points of α respectively.

The number of homology classes that are born at K_i and die upon entering K_j is given by

$$\mu_p^{i,j} = (\beta_p^{i,j-1} - \beta_p^{i,j}) - (\beta_p^{i-1,j-1} - \beta_p^{i-1,j}).$$

The topological information provided by the persistent homology groups is usually represented in either a *persistence diagram* or a *barcode*. A persistence diagram is a dot plot of pairs (b_i, d_i) for each homology dimension $p \in \mathbb{N}$, with their multiplicities $\mu_p^{i,j}$. A persistence barcode is a diagram in which each homology class occurring throughout the filtration is represented by a line, spanning the filtration interval during which it exists independently. To put all these definitions together, we revisit the example shown in Figure 3.3 at the beginning of this section. Figure 3.4 shows the same point sample from $\mathcal{M}$, and the Vietoris-Rips complexes VR_ε for the three values $\varepsilon = 1.0, 2.0, , 4.0$. The full barcode representing the persistent homology groups $H_1^{i,j}$ is also shown. This representation of the persistence of homology classes makes the existence of two cycles in the underlying manifold clear, appearing here as long bars, as opposed to three much shorter-lived cycles which can be disregarded as topological noise introduced by the sampling.

The significance of the barcode (or persistence diagram) representation is given by the following result:

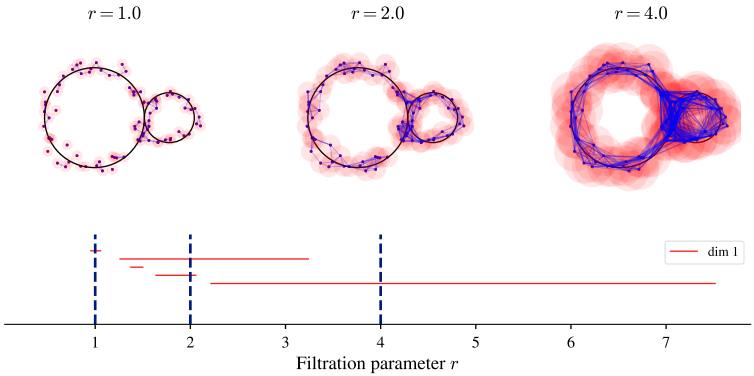


Figure 3.4: Barcode representing the persistent homology of the Vietoris-Rips filtration built from the set of points in Figure 3.3. The blue lines represent the 1-skeleta of the Vietoris-Rips complexes at each filtration step.

Theorem 3.6.3. *The rank of the persistence homology group* $H_p^{i,j}$ *is equal to the number of intervals in the barcode of* H_p *spanning the parameter interval* $[i, j]$. *The value of* $\beta_p(K_i)$ *is equal to the number of intervals containing* i.

Proof. See Zomorodian and Carlsson [2005, §3.3] on the correspondence between the algebraic structure of a persistence module and decomposition theorems in commutative algebra. □

This implies that the barcode representation, for all its simplicity, does contain all the homological information about the filtered complex. It thus offers a quick and reliable way to separate true topological features from noise. Both the barcode and the persistence diagram, however, have a significant disadvantage: since they both represent a multiset of pairs of numbers, doing any kind of statistics on them becomes unwieldy.

A further and very important property of persistence is its *stability*: small perturbations of the input data will lead to small changes in the output, either as persistence diagram or barcode. This crucial fact was shown by Cohen-Steiner *et al.* [2007].

4

HOMOLOGICAL SIGNATURE OF LAND SURFACE–ATMOSPHERE INTERACTION

Outline

This chapter applies techniques from cubical homology to the analysis of three- and four-dimensional data from numerical simulation of atmospheric flow. Section 4.1 provides an overview of similar applications of these techniques to the analysis of scientific data found in the literature. Section 4.2 details the analysis methodology, which is then applied to extract topological invariants, the Betti numbers, from model datasets in Section 4.3. These invariants are used as features in the problem of classifying land surface patterns based on known properties of atmospheric flow. Section 4.4 uses the Betti number data to furnish a structural classification of the planetary boundary layer, based on topological information only. Part of this chapter was published in the journal Boundary-Layer Meteorology as Licón-Saláiz *et al.* [2020].

As discussed in Chapter 1, a central issue in this dissertation is the investigation of the effect of land–atmosphere interaction on the flow structure of the daytime planetary boundary layer (PBL). This system is sometimes also referred to as a radiatively-driven convective boundary layer (CBL). Throughout this chapter we will use both terms, so it is important to keep the distinction in mind. The CBL is energetically forced by solar heating of the land surface. This energy is then radiated or reflected back into the atmosphere, and it drives the convective process. A land surface is in general not homogeneous, but it is made up by different

components, such as soils, vegetation, water, ice, and so on, and these also adopt various spatial configurations. Each surface type has its own heat transfer capacity, and thus impacts differently on the convective system as a whole. The total effect exerted by the surface on the atmosphere is determined not only by the different surface types present, or their relative sizes, but as it turns out, also by the specific spatial arrangement of these surface types. Here we will analyze the effects produced on the structure of turbulent flow in the PBL by the different land surface patterns introduced in Section 2.5 (see Figure 2.6).

The dynamics in a convective boundary layer (CBL) are characterized by distinct vertical profiles of the momentum exchange in the vertical direction which is caused by convective plumes. A plume is a connected region of the space domain with positive buoyancy, and thus upwardly accelerating air, with positive vertical velocity. In relation to their substantial horizontal extent, such plumes persist for a significant interval of time and can be understood as a coherent structure. These structures have proven important in reducing the complexity inherent to turbulent flows, and thereby improving our ability to understand and model them [Adrian, 2007; Shah and Bou-Zeid, 2014]. Methods to detect and quantify such coherent structures in model datasets most often rely on computing a spectral transform of the data, and performing the necessary analysis in spectral space. We present here a novel approach that constructs a geometric representation of flow structures, based on three- and four-dimensional data. This geometric representation is then characterized by topological descriptors, which can then be used as features in a statistical learning problem. To this end we apply methods from computational topology, specifically computational homology, to study the effects of land surface heterogeneity on the PBL. With these we obtain a a representation of the structure of the turbulent convective flow found in the PBL in terms of morphological flow features, and use this to parametrize the effects of land surface heterogeneity. As we will see, some of the questions that arise are topological in nature, giving good cause for using these methods. We identify the following components in the analysis pipeline:

DATASET This is the output of numerical simulations of the PBL.

REPRESENTATION The numerical methods used to generate the data use cubical grids, therefore the most natural combinatorial representation to use is the *cubical complex*.

INVARIANT Two different topological invariants that can be efficiently computed from the datasets as described here: Betti numbers and merge trees.

4.1 RELATED WORK

The idea of building a geometrical object from the scalar fields produced by the numerical model of a nonlinear physical system is at the heart of many of the earliest applications of topology to data analysis in the natural sciences. Some examples of this are:

Gameiro *et al.* [2004] The existence of chaotic spatiotemporal dynamics in non-linear systems, such as those described by the Gray-Scott and FitzHuhg-Nagumo models, can be ascertained by computing the Lyapunov exponents (see for example Strogatz [2014]). These chaotic dynamics are usually associated with the existence of intricate geometric patterns. Gameiro *et al.* thus propose a technique based on homology: cubical complexes are created by thresholding the computational domain, and topological invariants are computed from these complexes, the Betti numbers. The Lyapunov exponent can then be measured by using the time series of Betti numbers, specifically that of β_1. The other Betti numbers are not relevant, since β_0 shows a very stable behavior and $\beta_2 = 0$ in these cases.

Gameiro *et al.* [2005] Chemical phase separation in alloys is described by the Cahn-Hilliard equation:

$$\frac{u}{t} = -\Delta\left(\gamma\Delta u - \frac{\partial\Psi}{\partial u}(u)\right),$$

where u measures the relative concentration difference of the materials, and Ψ is the bulk free energy, given by a

double-well potential. This is another example of a nonlinear system where intricate geometric patterns emerge, and the authors again propose a technique based on the computation of Betti numbers to characterize the system's behavior via pattern morphology. As in Gameiro *et al.* [2004], the first step is to build a cubical complex by thresholding the computational domain. The authors find that these topological invariants provide a quantitative method to describe the effect of ambient noise on pattern morphology. They also allow for a comparison of the different patterns observed when the mass parameter changes.

Krishan *et al.* [2007] This study was the first to use topology to characterize the geometric patterns emerging in a system both in laboratory experiments and in numerical simulations. Specifically, the system under study is thermal Rayleigh-Bénard convection, in the state known as spiral defect chaos [Morris *et al.*, 1993].

The presence of non-Boussinesq effects in the system is reflected in the behavior of the first two Betti numbers for plane domains in both experiments and simulations. As the reduced Rayleigh number ε is increased, the time series for the Betti numbers of hot and cold regions become asymmetric, which is consistent with the known effects of the breakdown of the Boussinesq approximation.

Carreras *et al.* [2008] The authors study the behavior of turbulent plasmas in a toroidal domain, specifically the effect on transport induced by turbulence. The system is described by a set of reduced magnetohydrodynamic equations, so that the flow is determined by a velocity stream function $\Phi(\rho, \theta, \zeta)$ in toroidal coordinates. Then the velocity field is $\mathbf{V} = \nabla\Phi \times \mathbf{b}$, with $\mathbf{b} = \mathbf{B}/|\mathbf{B}|$ being the direction of the magnetic field.

A three-dimensional plasma flow is simulated, and the numerical data are analyzed using computational homology on the three-dimensional cubical complexes obtained by thresholding. These Betti numbers for the regions of space thus obtained give a good geometric characterization of the

number and type of structures present in the flow. In this case, since the model is three-dimensional and of high resolution, the authors need to devise a scheme to approximate the value of the Betti numbers, as computing on the full domain is impractical due to memory limitations. This depends on the possibility of deriving an analytical expression for the Betti numbers from the Φ function, something which is unique to this problem and does not generalize.

Garcia *et al.* [2009] This study builds on the techniques presented by Carreras *et al.* [2008], addressing the limitations in resolution which were brought up there. The approach presented here permits to distinguish between topological features of the flow which influence the transport process, and those that do not. Specifically, the authors observe the existence of dominant large-scale connected components and cycles, which can be taken to indicate a lack of homogeneity in the turbulent flow. The frequency distribution of these large-scale components is found to be in close agreement to the probability distribution given by a stochastic particle transport model, although no analytical relationship is given.

Muszynski *et al.* [2019] The focus of this study is again the characterization of geometric features in numerical simulations of a non-linear system. The scope is significantly larger than in the previous references in at least two ways: first, the volume of data involved is much larger, as the system under study is the Earth's atmosphere; second, the methodology developed here does not have the extraction of topological features as an end in itself, but they are used as input to a machine learning algorithm. The goal is the automated identification of *atmospheric rivers*, long, narrow connected regions in the atmosphere with a large water vapor content. Another implicit limitation of the method is addressed here: the dependence on the choice of threshold. The way around this is to extract the connected components for all possible threshold values, and track their size as the threshold changes. The result is then vectorized and used as features for an SVM classifier.

In all these studies, the notion of "shape" plays a central role. More precisely, topological invariants associated with the complex shapes that emerge naturally from the physical systems under consideration are shown, firstly, to display a strong regularity. Such regularity is especially surprising given the fact that some of these systems are archetypal examples of chaotic dynamics. Second, a relationship can be established between the numerical values of these invariants and dynamical properties of the underlying system. Thus, if we look at complex dynamics through the lens of topology, it is possible to cut through the seemingly chaotic motions and recover important information about the physical processes involved. In this chapter we show how topological characteristics of convective PBL flow change in response to different land surface patterns. To this end, we use data from simulations produced by a large-eddy simulation atmosphere–land model (LES-ALM) developed by Shao *et al.* [2013], forced by four different land surface patterns [Liu *et al.*, 2017].

4.2 GEOMETRIC REPRESENTATION

The numerical models used in the study of CBL structure contain several physical variables, so a first step in the methodology described here is to select a variable or group of variables that appropriately relate to the phenomenon under study, namely free convection.

4.2.1 *Variable selection*

As the system is dominated by the exchange of information in the vertical direction, we will consider model variables for which this physical direction is especially important:

VERTICAL WIND VELOCITY In an idealized convective system, energy and momentum are primarily exchanged in the vertical direction. The velocity of wind in this direction is thus an immediate representative of convective motion, and it will be useful in studying the structural properties of the flow, as the existence and spatial distribution of flow elements are

expressed by this variable. In the case of inhomogeneous convection the vertical wind velocity will carry information from the boundary condition at the surface into the atmosphere [Kondrashov *et al.*, 2016].

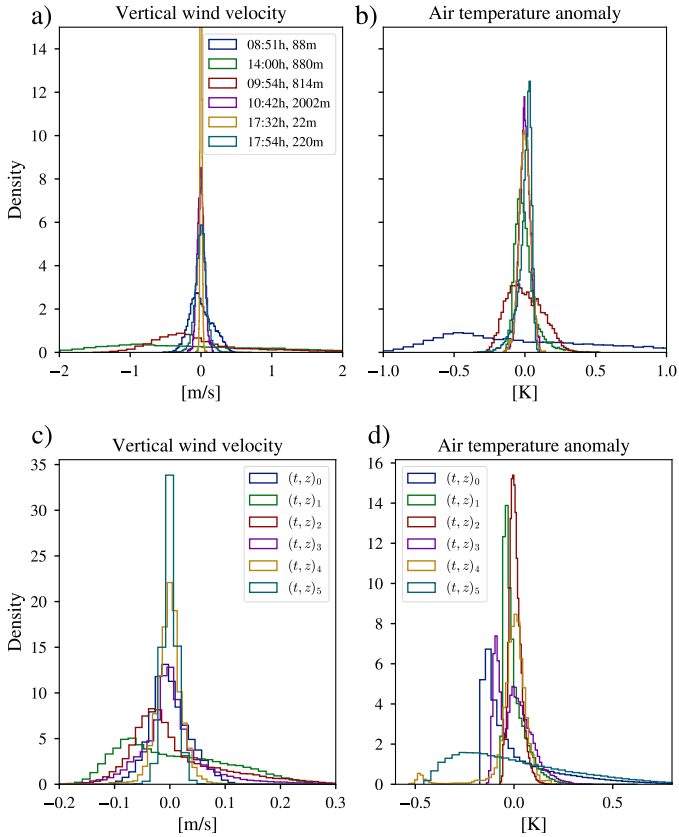
ENSTROPHY The magnitude of the rotational field of velocity is a good representation of the local intensity of turbulence. It is, however, not computed by most atmospheric models, and its direct measurement in the atmosphere is very hard. Additionally, it can also confound the effects of the three velocity components.

BUOYANCY This is only non-zero in the case of background stratification, thus might not be very informative in general. Like the vertical velocity, it also provides a direct view of energy transfer in the convective system.

Figure 4.1 shows the empirical probability density functions (PDF) of two of these variables, for both LES and DNS data. As can be seen in the accompanying table, the first and second order statistics of these variables are very similar, with the probable exception of those corresponding to the wind velocity at 14:00 h, 880 m for the LES data, for which the standard deviation is an order of magnitude higher than the rest. The corresponding statistics for the air temperature anomaly (not shown) exhibit the same pattern. All this highlights the existence of statistical regularities in the midst of complex turbulent motion, the phenomenon represented by both simulation datasets. The reasons behind the seemingly arbitrary choice of these data points will become clearer towards the end of this chapter, once we have developed different tools to better characterize the data shown here.

4.2.2 *Thresholding and anchor points*

To obtain the kind of geometrical representation we need, the first necessary step is to form a binary array from the original data which represents those regions of the spacetime domain that fulfill a given condition. In other words, we want to select areas of the domain based on a predefined threshold value. In



LES-ALM	μ	σ	DNS	μ	σ
08:51h, 88m	6.64e−4	1.59e−1	$(t,z)_0$	3.39e−4	3.61e−2
14:00h, 880m	1.97e−4	1.32	$(t,z)_1$	6.98e−3	9.53e−2
09:54h, 814m	2.19e−4	7.54e−1	$(t,z)_2$	7.83e−3	8.29e−2
10:42h, 2002m	9.21e−5	5.19e−2	$(t,z)_3$	3.83e−3	5.86e−2
17:32h, 22m	5.92e−6	1.17e−2	$(t,z)_4$	−6.78e−4	2.52e−2
17:54h, 220m	−4.10e−6	8.13e−2	$(t,z)_5$	2.16e−5	1.30e−2

Figure 4.1: Empirical probability density functions (PDFs) of vertical wind velocity and temperature anomaly for the LES-ALM SP4 simulation (top) and DNS (bottom) datasets, measured at different time points and heights. The table below shows mean and standard deviation for all the vertical velocity densities shown.

the case of vertical wind velocity and buoyancy, the choice of value for this threshold can be motivated physically: a value of 0 separates the spatial domain into regions of air moving or accelerating upwards from those where it moves or accelerates downwards. This will result in an array of points representing the physical locations of the convective plumes discussed above, and we will investigate the characteristics of this structure. Moreover, since the data from these numerical simulations have a natural representation in a cubical grid, the choice of cubical complex as the geometric representation is also natural.

We denote the computational domain of a given numerical simulation by $\Omega \subset \mathbb{Z}^4$. In general, we have

$$\Omega = \{1, 2, \ldots, N_t\} \times \{1, 2, \ldots, N_x\} \times \{1, 2, \ldots, N_y\} \times \{1, 2, \ldots, N_z\}, \tag{4.1}$$

where N_t represents the number of grid points in the time dimension, N_x, N_y, and N_z the number of grid points in each of the three spatial dimensions. We will consider the subsets of this domain formed by fixing the value of one or more of the four coordinates, for example $\mathcal{M}_{z^*,t^*} = \{(t, x, y, z) \in \Omega \mid z = z^*,\ t = t^*\}$ will represent the two-dimensional grid formed by points in the x, y directions for fixed z, t. We can then consider the values of a model variable, for example vertical wind velocity w, on this subset: $\{w(i, j) \in \mathbb{R} \mid (i, j) \in \mathcal{M}_{z^*,t^*}\}$. We can now make a conditional selection by using a threshold value θ and extracting the points of $\mathcal{M}_{z^*,t^*}$ where the value of w is greater than this threshold, $\mathcal{P} = \{(i, j) \mid w(i, j) \geqslant \theta\}$. This set of points can be used as the *anchor points* of a cubical complex: for each anchor point (i, j) in this set, include the elementary cube defined by $[i, i + 1] \times [j, j + 1]$. We can then denote the cubical complex associated to this set of points by $\mathcal{M}^+ = \{[i, i + 1] \times [j, j + 1] \mid (i, j) \in \mathcal{P}\}$.

As mentioned above, a threshold value of $\theta = 0$ is a natural choice when using vertical wind velocity, since this has a very direct physical interpretation in terms of splitting the domain $\mathcal{M} = \mathcal{M}_{z^*,t^*}$ into updrafts, $\mathcal{M}^+$, and downdrafts, $\mathcal{M}^-$. For the latter case, the steps are similar but we will take a negative threshold, and obtain the anchor points from the condition $\{(i, j) \mid w(i, j) \leqslant \theta\}$. Having done this we will end up with two distinct geometrical

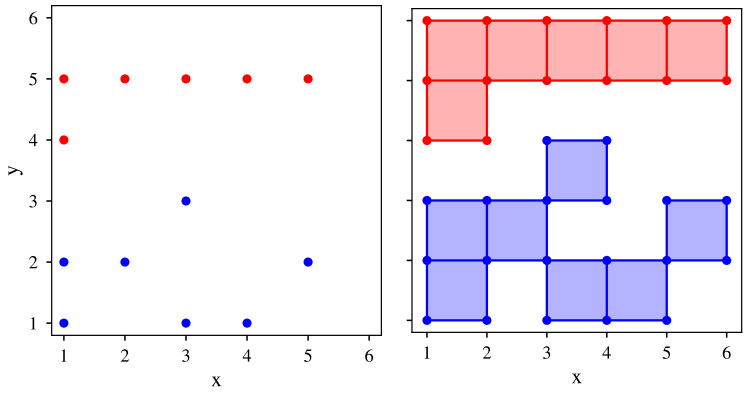


Figure 4.2: An example of cubical complexes built from data points on a grid. Left panel shows two distinct sets of anchor points, right panel shows the two-dimensional cubical complexes generated by each of these sets.

objects corresponding to each two-dimensional slice of the original scalar field. Figure 4.2 shows an example of this. It is by using these objects that we will describe the structural properties of updraft and downdraft regions. Specifically, we will compute the first two Betti numbers for these objects, β_0 and β_1.

4.3 BETTI NUMBERS FOR THE VERTICAL WIND VELOCITY FIELD

Once we have built geometric objects that represent domains of interest, in this case the regions of space associated to updrafts and downdrafts, we seek a quantitative description of their properties relevant to the physical problem at hand. Homology provides us with topological descriptors, the Betti numbers (see Section 3.4), which are inexpensive to compute in a setting such as the one dealt with here, and are conceptually simple to understand, as they quantify the connectivity of a space in terms of its components, loops, holes, and similar structures in higher dimensions. We will then show how these low-level topological descriptors can be used

to analyze the structure of the CBL, as well as its interaction with the underlying land surface. The Betti number computations were performed using CHOMP, a C++ code library for computational homology [Mischaikow *et al.*, 2019].

We will compute the Betti numbers for two-dimensional horizontal cross sections. The motivation behind this is not only computational simplicity, but also the fact that it is possible to understand a given object of dimension n by understanding its components of dimension $n-1$, as well as the relationships between them. A well-known example of this is a computed tomography scan of the brain, where a three-dimensional volume can be reconstructed from two-dimensional slices. Concretely, assume we have the three-dimensional cubical complexes $\mathcal{M}_t^{\pm}$ defined above, which correspond to the volume of space associated with updrafts and downdrafts at simulation timestep t. These objects will play the role of the brain in our analysis, although we will not deal with the issue of its three-dimensional structure until Chapter 5. In this section we focus on the slices obtained from this object by fixing a height z, and building from them the two-dimensional cubical complexes described in Section 4.2.2 to represent the areas covered by up- and downdrafts at time t and height z, as illustrated in Figure 4.4. The Betti numbers we compute are the ranks of the first two homology groups, H_0 and H_1, for these subcomplexes.

As mentioned in Section 4.2.2, an important advantage of using a variable like vertical wind velocity is that the value of 0 gives us a physically meaningful threshold. However, for a number of reasons it might be advantageous to consider a value $\varepsilon > 0$, and from it construct a ternary partition by thresholding symmetrically around 0.

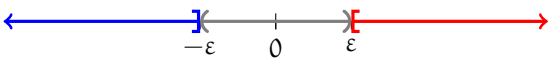


Figure 4.3: Symmetric thresholding of values around 0.

As shown in Figure 4.3, the positive domain (in red) is formed by all values $w > \varepsilon$, the negative domain (in blue) by $w < \varepsilon$, and the gray area around zero is left out. The first reason why this might be a good idea has to do with the fact that the data analyzed originates from numerical simulation: values small in magnitude could well be due to numerical error. Secondly, such values are also not significant from a physical perspective and can be ignored. Finally, leaving out those values also has the effect of simplifying the boundaries of the regions we are considering. Throughout this section we will be using threshold values of $(-\varepsilon, \varepsilon) = (-0.01\,\mathrm{m\,s^{-1}}, 0.01\,\mathrm{m\,s^{-1}})$.

We will then associate four numbers to each two-dimensional cross section: β_0^+ and β_1^+ for the subcomplex obtained from the positive region, β_0^- and β_1^- for the corresponding subcomplex from the negative region. Figure 4.4 shows an example of the partition into disjoint sets induced by the thresholding, and Table 4.1 shows the respective values of the Betti numbers.

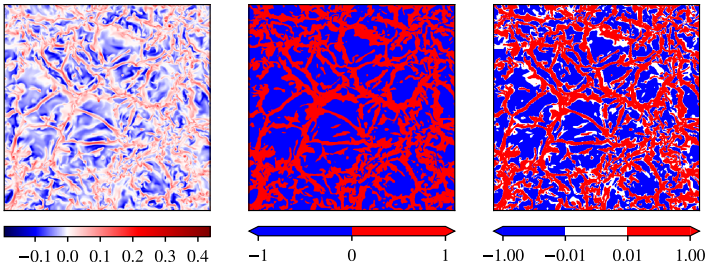


Figure 4.4: Left: vertical wind velocity w, in $\mathrm{m\,s^{-1}}$, within a horizontal cross section of the DNS domain. Center: binary partition based on values of w. Right: ternary partition of the same field, corresponding to the domains $\mathcal{M}^-$ (blue), $\mathcal{M}^+$ (red), and $\mathcal{M}^{\sim 0}$ (grey).

The physical interpretation of these numbers is clear: $\beta_0^\pm$ counts the number of connected updrafts (resp. downdrafts), while $\beta_1^\pm$ is the number of "holes" enclosed by these updrafts (resp. downdrafts). In general, one would expect that $\beta_0^+ = \beta_1^-$, since every connected component in the positive domain will necessarily cor-

Table 4.1: Betti numbers for the spatial regions shown in Figure 4.4 (right panel).

	$\mathcal{M}^+$	$\mathcal{M}^-$	$\mathcal{M}^{\sim 0}$
β_0	149	309	2245
β_1	233	93	183

respond to a "hole" in the negative domain, and $\beta_0^- = \beta_1^+$ for the same reason. In this case we observe that neither of these relationships holds. A possible explanation for this is the fact that we are using a *ternary* partition, and indeed we can see cycles within the positive domain in Figure 4.4, right panel, which do not enclose any connected components of the negative domain $\mathcal{M}^-$, but only of the white domain $\mathcal{M}^{\sim 0}$. We could think that equality would hold in case we had a binary partition of the domain, but this is again not true, as shown in Figure 4.5. The reason for this is a subtle but important point, consequence of the definitions in the framework of cubical homology. By way of example, let $X = [0, 7] \times [0, 7]$ be represented as a cubical complex generated by the elementary 2-cubes in its interior, and consider the two cubical subcomplexes $\mathcal{M}^+$ and $\mathcal{M}^-$ generated by the sets of anchor points P^+ and P^- shown in Figure 4.5 as red and blue dots respectively. We can see that on the one hand, $P = P^+ \cup P^-$ as a set of anchor points would also generate X, and $P^+ \cap P^- = \varnothing$. On the other hand, $\mathcal{M}^+ \cup \mathcal{M}^- = X$, but $\mathcal{M}^+ \cap \mathcal{M}^- \neq \varnothing$, since both sets have elementary 0- and 1-cubes in common. This means that $\mathcal{M}^- \supset X \setminus \mathcal{M}^+$, and $X \setminus \mathcal{M}^+$ would not even be a proper cubical complex, as it would contain 2-cubes which would lack some of their boundary 0- and 1-cubes. Each of the constituent squares of $\mathcal{M}^+$ share a 0-cell, hence $\mathcal{M}^+$ has a unique connected component, and $\beta_0^+ = 1$. The same 0-cells are also part of $\mathcal{M}^-$. But this then implies that that the four cycles shown in Figure 4.5 as $\gamma_1, \ldots, \gamma_4$ are four generators for the first homology group $H_1(\mathcal{M}^-)$, hence β_1^-. The "hole" that we would expect to get when carving out $\mathcal{M}^+$ is then a linear combination of these four independent cycles, $\Gamma = \sum_i \gamma_i$.

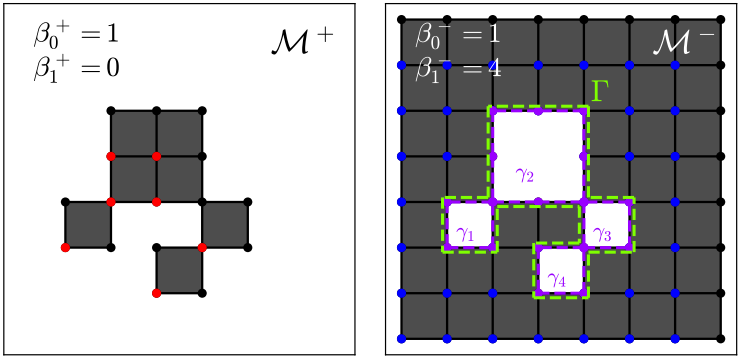


Figure 4.5: Binary partition of a rectangular domain. The black region on the left panel represents $\mathcal{M}^+$, the black region on the right panel is $\mathcal{M}^-$. Also shown on the right panel are the four generators of the first homology group H_1 for $\mathcal{M}^-$, $\gamma_1, \ldots, \gamma_4$, and the cycle Γ formed as a linear combination of them, which would correspond to "cutting $\mathcal{M}^+$ out" of $\mathcal{M}^-$.

The two-dimensional slices shown in Figure 4.6 are more illustrative of the general situation. Here we can see that the different surface patterns, with different characteristic length scales, induce vertical velocity fields which also exhibit clear qualitative differences. The intuition behind this is as follows: a large, contiguous area of common land type, for example forest, will tend to have a roughly constant heat flux into the atmosphere. If the scale of this patch of forest happens to be commensurate with that of a characteristic CBL plume, we would expect to find a correspondingly large plume above the forest patch, with its return circulation extending to the surface and merging smaller plume structures originating from the patch into the dominant plume [van Heerwaarden *et al.*, 2014]. If this forest patch is now broken up into many smaller, disjoint patches, each significantly smaller than the characteristic CBL length scale, this merging process does not necessarily occur, and we might expect to see a larger number of smaller, more localized plumes. Thus, we expect the value of β_0^+ to be negatively correlated to the characteristic length scale of the

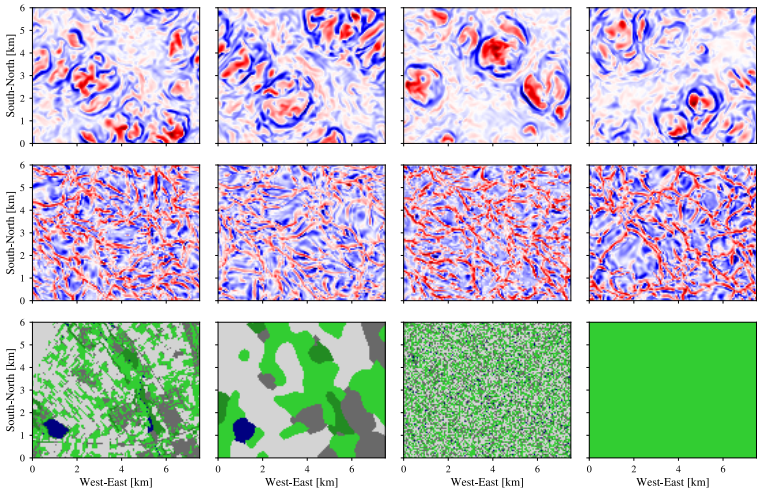


Figure 4.6: Comparison of land surface patterns from the four LES-ALM datasets (bottom), and two horizontal cross sections of vertical wind velocity w for each dataset (red: $w > 0$, white: $w \sim 0$, blue: $w < 0$). Middle row: vertical wind velocity at 220 m height. Top row: 1760 m. All slices show the state at 13:00 h.

surface pattern, at least close to the surface. This is the case in the example shown here. In the next section we will show how this statement can be made more precise.

4.3.1 *Betti profiles*

Definition 4.3.1. Let (t, z) be a pair of time and height values in the computational domain Ω, such that $\mathcal{M}^{\pm}_{(t,z)}$ are the two-dimensional cubical complexes for up- and downdrafts at time t and height z, both in grid units. Then $\beta^{+}_{0} : \{0, \ldots, N_t\} \times \{0, \ldots, N_z\} \longrightarrow \mathbb{Z}$ is a function of time and height which maps the pair (t, z) to the Betti number β_0 for the cubical complex $\mathcal{M}^{+}_{(t,z)}$. Its corresponding *Betti profile* at time τ is the function of z obtained when conditioning on $t = \tau$, that is $\beta^{+}_{0}(t, z)|_{t=\tau}$. The functions β^{+}_{1}, β^{-}_{0}, and β^{-}_{1} are defined analogously, as well as their corresponding profiles.

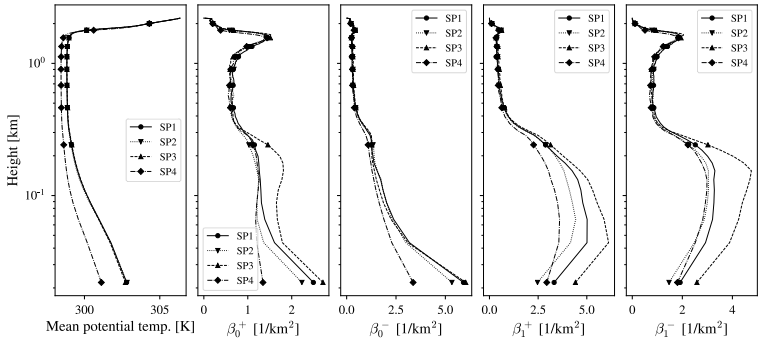


Figure 4.7: Vertical profiles for the LES-ALM datasets. First panel: temperature. Each of the following panels shows the four Betti numbers, $\beta_0^{\pm}$ and $\beta_1^{\pm}$, for one of the four simulations. Each profile is a one-hour average centered at 12:00h.

The use of the name *profile* is akin to that in meteorology, where it denotes the function of height that describes variables such as temperature, water vapor, or velocity. This is motivated by the *boundary layer hypothesis*, whereby the effects of viscosity on a fluid are significant only in close proximity to a solid barrier and are negligible away from it. This hypothesis, first proposed by Prandtl [1905], allows for the following approximation in case that the boundary layer thickness, in the direction normal to the solid barrier (z), is significantly smaller than the spatial extent of the domain parallel to the barrier (in the directions x and y):

$$\frac{\partial u}{\partial x} \ll \frac{\partial u}{\partial z} \quad \text{and} \quad \frac{\partial^2 u}{\partial x^2} \ll \frac{\partial^2 u}{\partial z^2},$$

with the same relationship between the derivatives with respect to y and z. Therefore we can assume that first- and second-order statistics change significantly only in the z and t coordinates. In this sense, Figure 4.7 shows the temperature profile for the LES-ALM datasets, as well as the corresponding Betti number profiles of them.

BETTI FEATURE MATRICES Given a simulation dataset, we calculate four different arrays of Betti numbers, each containing

the numerical values of the functions given in Definition 4.3.1. We can represent each of these as a matrix of dimension $N_t \times N_z$, where N_t is the number of simulation timesteps, and N_z is the number of vertical grid levels. For example, corresponding to the Betti number β_0^+ we have the matrix $(B_0^+)_{ij}$, which has as its (i,j)-th entry the number β_0 for the positive vertical wind velocity domain at the i-th timestep and j-th height level. We create a *feature matrix* via the following steps:

1. Select a range of timesteps to use, and a sampling frequency. Each timestep will be one observation. For example: in the LES-ALM simulations it would be reasonable to consider the time window between 10:00h and 18:00h as separate from the window between 18:00h and 21:00h, as these represent two different physical regimes. Additionally, since the simulation timesteps are of one minute, taking only every k-th timestep would reduce the temporal autocorrelation between the observations.

2. Select the range of height levels to use. Each height level z will be a feature, representing all the values of a given Betti number at height z throughout the simulation time.

3. This results in a matrix of dimension $N_t' \times N_z'$, with $N_t' \leqslant N_t$ and $N_z' \leqslant N_z$ being the dimensions of the computational domain (see Equation 6.13). In case the model to be considered uses the information from both Betti numbers, β_0 and β_1 (from either the positive or the negative domain), we concatenate these two matrices column-wise, effectively doubling the number of features, and resulting in a matrix of size $N_t' \times 2\,N_z'$.

4. Concatenate the matrices from all available simulations row-wise. In the case of LES-ALM, this would mean the four simulations SP1 to SP4. This results in a $4\,N_t' \times N_z'$ matrix (or $4\,N_t' \times 2\,N_z'$ if using both Betti numbers).

5. Normalize the features: for each column in the matrix, center and rescale it so that its mean is 0 and its standard deviation is 1.

4.3.2 *Classification of land surface patterns*

We proceed by classifying the four different land surface patterns in the LES-ALM simulations by using only the information contained in the vertical Betti number profiles. In other words, if all we know is the geometric information given the Betti numbers for a certain region of the space-time domain, what can we conclude about the land surface pattern that generated it? Moreover, how does this information compare to the information provided by the physical variables? Using the same set of four simulations discussed here, Liu *et al.* [2017] show that the bulk profiles of temperature can distinguish between homogeneous and inhomogeneous surface pattern, but they cannot distinguish between the three inhomogeneous surface patterns. This subsection will show to which extent the Betti numbers of the vertical velocity fields are better descriptors for this task.

We will try to answer this question by using the feature matrices X described in the previous paragraph. In total, we will use eight of these matrices: four corresponding to the four sets of Betti numbers $\beta_0^{\pm}$ and $\beta_1^{\pm}$, one for the concatenation of the two positive Betti numbers, $\beta_{0,1}^{+}$, one for the concatenation of the two negative Betti numbers, $\beta_{0,1}^{-}$. Finally, we also use the two matrices corresponding to the profiles for vertical wind speed, W, and temperature, T.

The response variable will be the label indicating the simulation each observation corresponds to, $y = \text{SP1, SP2, SP3, SP4}$. For constructing the feature matrices, we focus here on the quasi-stationary period, 10:00h - 17:00h, after the rate of boundary layer growth has stabilized, and before the evening transition begins. Timesteps are selected in 5 min intervals, to diminish the effect of temporal autocorrelation. This results in $N_t' = 1684$ observations. Two different classification schemes will be considered: a k-nearest neighbors classifier, and a multinomial logistic regression. In both cases we use the implementations available in the Python library *scikit-learn* [Pedregosa *et al.*, 2011].

K-NEAREST NEIGHBOR CLASSIFICATION The k-nearest neighbor (k-NN) classifier is a well-known method in machine learning

[Hastie *et al.*, 2001, §13.3]. Let $Z = \{(X_i, y_i)\}_{i=1}^N$ denote the training data set, with each X_i being an observation of the M explanatory variables, $X_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,M})$, and the response Y_i being a categorical variable with values in a finite set, $y_i \in \{1, 2, \ldots, K\}$. Given a new observation X_0, the method finds the k training points x_i closest to x_0, and assigns to it the value $\hat{y}$ to x by majority vote, i.e. the most common label within the set $\{x_i\}_{i=1}^k$, with ties being broken at random in case two or more labels have the same frequency within the group of k nearest neighbors. This method requires the features to be embedded in some metric space, which is the case here since we have translated the geometric properties of the flow into a vector of natural numbers. The value of k is commonly determined by cross-validation. The metric we use is the Euclidean metric, and the k observations are given uniform weight.

MULTINOMIAL LOGISTIC REGRESSION We give a brief summary of the method here. We follow the presentation given in Hastie *et al.* [2001, §4.4]. Let $Z = \{(X_i, y_i)\}_{i=1}^N$ be the training data set as before. If we let $p_{i,j} = P(Y_i = j)$ be the probability that the i-th observation falls into category j, and $Y_{i,j}$ be the indicator variable for observation i being in category j, the joint probability density of the response is given by

$$P(Y_{ij} = y_{ij}, \ldots, Y_{iK} = y_{iK}) = \prod_{j=1}^{K} p_{ij}^{y_{ij}}. \tag{4.2}$$

The values of the p_{ij} are unknown, and we want to obtain an estimate for them conditioned on the data $X = (X_1, \ldots, X_M)$.

Logistic regression estimates this via a linear model of the log-odds for class membership of observation x_i, as given by

$$\begin{aligned}
\log \frac{P(y_i = 1 \mid X_i = x_i)}{P(y_i = K \mid X = x_i)} &= \beta_{10} + \beta_1^T X_i \\
\log \frac{P(y_i = 2 \mid X_i = x_i)}{P(y_i = K \mid X = x_i)} &= \beta_{20} + \beta_2^T X_i \\
&\vdots \\
\log \frac{P(y_i = K-1 \mid X_i = x_i)}{P(y_i = K \mid X = x_i)} &= \beta_{(K-1)0} + \beta_{K-1}^T X_i.
\end{aligned} \tag{4.3}$$

The model is completely determined by this set of $K-1$ equations due to the additional restriction that $\sum_j p_{ij} = 1$. It can be shown that these equations are equivalent to

$$\begin{aligned}
P(y_i = k \mid X_i = x_i) &= \frac{\exp(\beta_{k0} + \beta_k^T X_i)}{1 + \sum_{\ell=1}^{K-1} \exp(\beta_{\ell 0} + \beta_\ell^T X_i)} \\
P(y_i = K \mid X_i = x_i) &= \frac{1}{1 + \sum_{\ell=1}^{K-1} \exp(\beta_{\ell 0} + \beta_\ell^T X_i)}
\end{aligned} \tag{4.4}$$

We denote the set of parameters by $\Theta = \{\beta_{10}, \beta_1^T, \ldots, \beta_{(K-1)0}, \beta_{K-1}^T\}$. These should not be confused with the Betti numbers, also denoted by β_i, $i \in \mathbb{N}$. We use the β symbol for the linear model parameters here for consistency with the statistical literature. Using Equation 4.2 and Equation 4.4 we can obtain the log-likelihood function for a given parameter vector β:

$$\ell(\Theta) = \sum_{i=1}^{N} \log P(y_i = k \mid X_i = x_i; \Theta) - \lambda \sum_{j=1}^{K} |\beta_j|, \tag{4.5}$$

where we have added an additional term for L_2-regularization. This is especially important in high-dimensional problems, as using larger values of the penalization parameter λ will tend to shrink the magnitude of the estimated parameters, thereby reducing model complexity. The estimates for the β parameters are obtained by maximizing the log-likelihood function, something typically done in practice by using an iterative optimization algorithm. For this work we used a L-BFGS solver, with the number

of iterations capped at 1000, and a tolerance value of 1e−10. Once we have obtained the parameter estimates, we can compute the posterior probabilities of class membership from Equation 4.4.

F_1 SCORE The F_1 score is a classical evaluation metric for classifiers [Powers, 2011]. It is based on the notions of precision and recall, which we now state:

Definition 4.3.2. Let $\{Y_i\}$ be a set of binary observations, and $\{\widehat{Y}_i\}$ the corresponding set of labels assigned to it by a classifier f. The *precision* of f is

$$\mathrm{pr} = \frac{|\{Y_i = 1\} \cap \{\widehat{Y}_i = 1\}|}{|\{Y_i = 1\} \cap \{\widehat{Y}_i = 1\}| + |\{Y_i = 0\} \cap \{\widehat{Y}_i = 1\}|}, \tag{4.6}$$

that is, the proportion of true positive results out of all the results produced by the classifier. The *recall* of f is

$$\mathrm{rec} = \frac{|\{Y_i = 1\} \cap \{\widehat{Y}_i = 1\}|}{|\{Y_i = 1\} \cap \{\widehat{Y}_i = 1\}| + |\{Y_i = 1\} \cap \{\widehat{Y}_i = 0\}|}, \tag{4.7}$$

the proportion of true positive results returned by f out of all the positive values in the population. The F_1 score (also called F measure) for f is the harmonic mean of precision and recall:

$$F_1 = \frac{2 \cdot \mathrm{pr} \cdot \mathrm{rec}}{\mathrm{pr} + \mathrm{rec}}. \tag{4.8}$$

For multiclass classification, we will use the *weighted average* F_1 *score*: the F_1 score is computed for each class individually, and they are averaged with a weight equal to the number of true instances of the class.

RESULTS We compared 8 Betti feature matrices, obtained from the following variables as described above: the four Betti numbers from the vertical wind velocity fields,β_0^+, β_1^+, β_0^-, β_1^-. Also $\beta_{0,1}^+$ and $\beta_{0,1}^-$, the concatenation of both Betti numbers for the positive and negative wind velocity domain respectively. Finally, also the physical variables W (vertical wind velocity profile), and T (temperature profile). For each feature matrix X:

1. Split the rows of X into X_{train} and X_{test}, where we use 70% of all rows as training set. A corresponding partitioning of the response Y into Y_{train} and Y_{test} is also performed.

2. Obtain a bootstrap estimate of the F_1 score for both the k-NN and the multinomial logistic classifier: take n samples with replacement from the rows of X_{train}, and compute the weighted average F_1 score on the holdout set, $(X_{\text{test}}, Y_{\text{test}})$. In our experiments, $n = 1000$.

Each classifier has a complexity parameter: for the k-NN classifier it is k, the number of neighboring points, whereas for the multinomial logistic classifier it is $C = 1/\lambda$, the inverse of the regularization parameter λ specified in Equation 4.5. The bootstrap estimation was performed over a range of parameter values, the results are shown in Figure 4.8. As can be seen, the classifiers trained on the combined feature sets of β_0 and β_1 tend to perform better than all other feature sets, especially the one corresponding to $\beta^+_{0,1}$, which achieves the maximum F_1 score in both cases: 0.65 for k-NN and 0.85 for logistic regression. Table 4.2 shows the maximum scores for all variables. This performance advantage shows that the updrafts, and in particular their geometric characteristics represented by their Betti numbers, are better at characterizing the state of the PBL than the vertical profiles of either wind velocity or temperature. Recalling the structure of the vertical profiles themselves, as seen in Figure 4.7, we note that their variation is greatest close to the surface, whereas their behavior for the higher regions of the boundary layer appears more similar. This motivates the use of subsets of the profile data, to see if this variation in the surface region indeed translates to a better classification of surface patterns by the corresponding features.

Following the same logic described before, a bootstrap estimate of the F_1 score was obtained for classifiers trained not on the full feature matrix using the 100 height levels of simulation data, but using only subsets of size 10, starting from ground level. That is, the values of the vertical profiles for levels 1 to 10 were used to train a set of classifiers, then the values for levels 11 to 20, and so on. The results of this are shown in Figure 4.9, where we show only the models for the combined Betti numbers, $\beta^+_{0,1}$ and $\beta^-_{0,1}$.

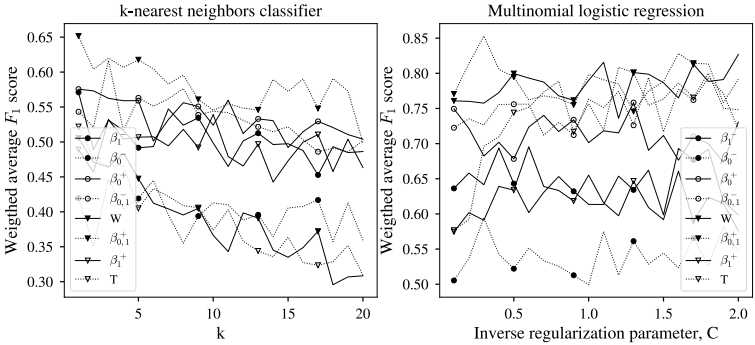


Figure 4.8: Bootstrap estimates for the F_1 score of k-NN (left) and multinomial logistic (right) classifiers, given the 8 feature matrices computed from vertical profile data.

Table 4.2: Maximum weighted average F_1 scores in 1000 bootstrap samples computed from the feature matrices for all 8 variables.

Variable	**k-nearest neighbors**		**Mult. logistic regression**	
	k	Max. F_1	C	Max. F_1
β_0^+	1	0.58	1.30	0.76
β_1^+	10	0.54	0.60	0.70
β_0^-	3	0.45	0.30	0.59
β_1^-	1	0.57	0.40	0.69
$\beta_{0,1}^+$	1	0.65	0.30	0.85
$\beta_{0,1}^-$	3	0.62	1.40	0.79
W	1	0.51	2.00	0.83
T	1	0.52	1.20	0.81

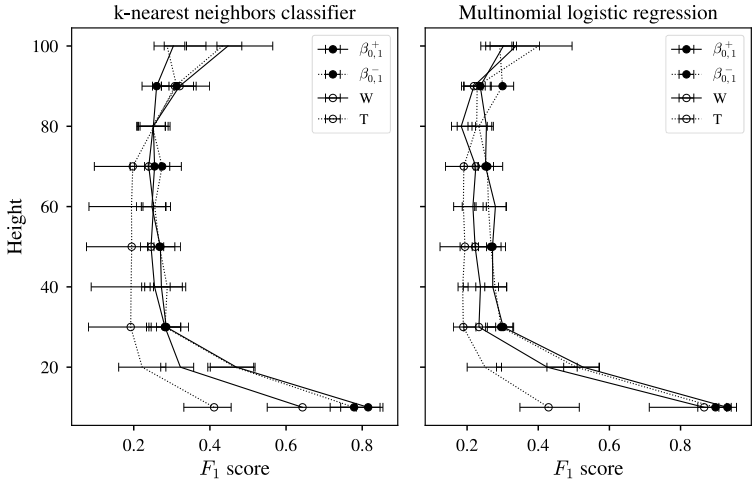


Figure 4.9: F_1 scores for k-NN and multinomial classifiers trained on subsets of size 10 of the 100 features available for LES-ALM datasets. The models at height 10 were trained using features corresponding to height levels 1 to 10, and so on.

Unsurprisingly, performance is significantly better close to the surface, and it degrades rapidly for higher levels. The models based on Betti number data display better performance up to level 70, where the models based on vertical wind velocity and temperature see an improvement in their F_1 scores. In the last region (levels 91 to 100) it is actually the classifiers based on temperature data which have the best scores overall (0.565 and 0.495), whereas for the most part they tend to perform worse than those based on the other three variables.

We can simplify the classification problem further, to obtain a better view of the difference in information carried by the Betti numbers on one hand, and the physical variables such as vertical velocity or temperature on the other. Figure 4.10 shows an example of this. Here we compare three variables: temperature, β_0^+, and $\beta_{0,1}^+$. For each variable, we have reduced the number of features to only two, which in the case of temperature and β_0^+ are the values of their profiles in the first two horizontal layers, and for

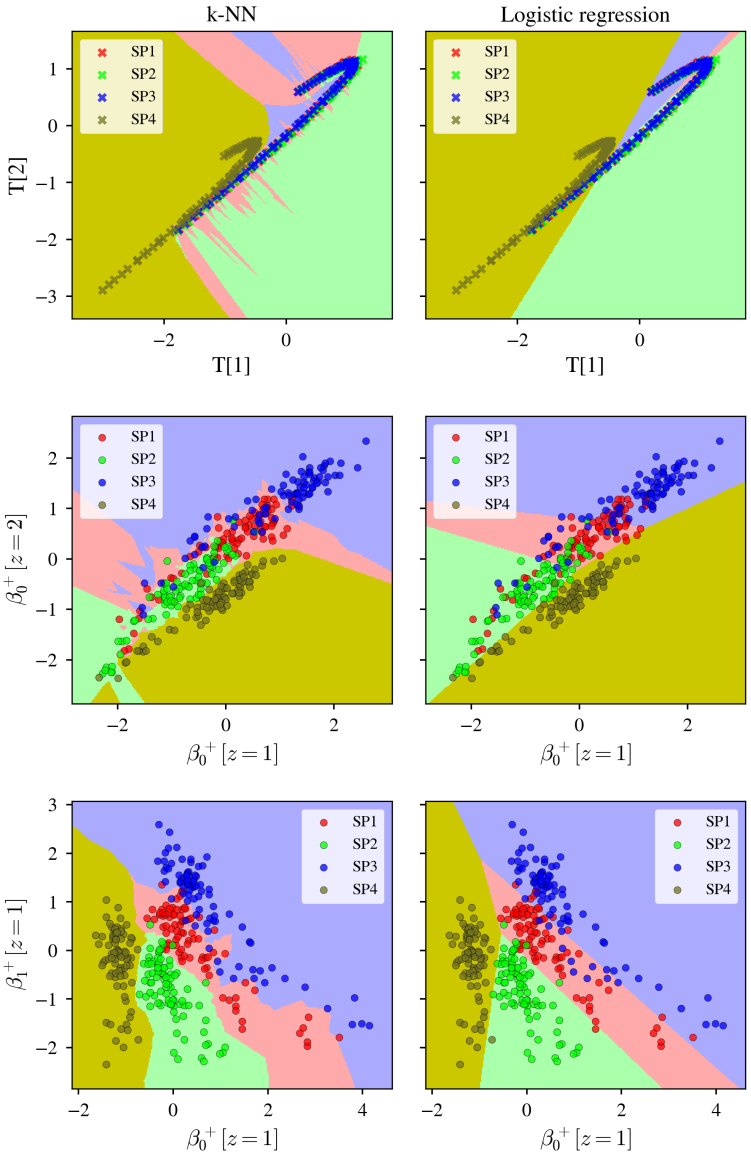


Figure 4.10: Classification models trained with data from the first two layers of the temperature profile (top row), from the first two layers of the β_0^+ profile (middle row), and from the first layer of β_0^+ and β_1^+ (bottom row). Left column shows a k-NN classifier, right column a multinomial logistic regression classifier, each with its respective decision boundaries. Each point in the plot represents one simulation timestep, with the axes showing the standardized feature values in each of the profile layers.

$\beta^+_{0,1}$ they are the values of both β^+_0 and β^+_1 in the first horizontal layer. Each point in the figures represents one of the simulation timesteps used as training set, where we again have sampled the time window between 10:00h and 17:00h at a 5 minute interval. The decision boundaries for each classifier are also shown. By looking at what happens close to the surface in this way, we can clearly see the significant differences in how the Betti numbers of the vertical velocity fields respond to different land surface patterns in a CBL regime. This confirms our earlier observation that the geometric-topological notions of connected components and "loops" or cycles quantified by these numbers are adept at characterizing the structural properties of turbulent flow, as well as the dependence of these properties on the forcing induced by land surface patterns of different heterogeneity. This is in line with the intuitive observation, made in Section 4.3, that if we break up a connected region of uniform land type and length scale $L \sim \mu_0$, with μ_0 representing the characteristic scale of a convective plume, then the spatial coherence of the turbulent flow above this surface region will also be broken. This breakup will result in a larger number of connected updraft regions than would be the case for a uniform land surface type. Indeed, as shown in the middle and bottom rows of Figure 4.10, the ordering of the β^+_0 values tends to agree with ordering the four simulations by increasing level of land surface heterogeneity, namely SP4, SP2, SP1, and SP3.

This is also in line with the findings reported by Liu *et al.* [2017], where data from the vertical temperature profiles are seen to represent the difference between the homogeneous case (SP4) and the three heterogeneous cases (SP1, SP2, and SP3). These temperature profiles, however, do not reveal any difference between the three heterogeneous cases (cf. [Liu *et al.*, 2017, Fig. 4] and our own Figure 4.7).

4.3.3 *Time series of Betti numbers*

It is also possible to use the information expressed in the Betti numbers to study the temporal evolution of the system they describe. We can, for example, fix a value $z = \zeta$ and look at the time series of values $\beta_i(t, \zeta)$. Given the existence of three distinct

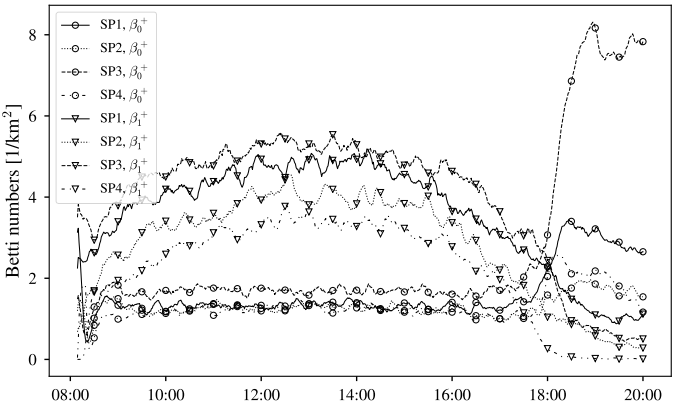


Figure 4.11: Time series of the 10-minute moving averages of β_0^+ and β_1^+ measured at 100 m above the surface, for all four LES-ALM simulations.

regions within a CBL, namely the surface, mixing, and inversion layers, it will be instructive to study time series which show the temporal evolution of the Betti numbers in each of these regions. For the case of the LES-ALM simulations we choose a height of $z_1 = 100$ m for the surface layer, $z_2 = 800$ m for the mixing layer, and $z_3 = 1600$ m for the inversion layer. As before, we compute four Betti numbers in each two-dimensional layer, one pair for both the positive and negative vertical wind subdomains. The time series for the Betti numbers measured at 100 m show a strong influence of the surface pattern, especially for the positive wind velocity domain (see Figure 4.11). The number of updrafts, β_0^+, becomes stationary throughout the CBL regime, which is in agreement with the observation made by Krishan *et al.* [2007] for the case of Rayleigh-Bénard convection. The mean value is similar for all simulations except SP3, which has a higher mean. At around 18 h the evening transition takes place as solar irradiation ceases, and the land surface is no longer an energy source for the convective system. The regularity with which convective plumes originated in the surface layer throughout the day is thereby destroyed, and the behavior of this part of the PBL now displays

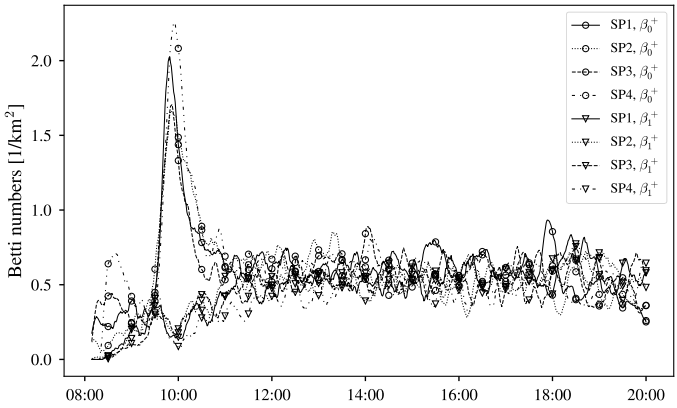


Figure 4.12: Time series of the 10-minute moving averages of β_0^+ and β_1^+ measured at 800 m above the surface, for all four LES-ALM simulations.

a strong sensitivity to the land surface pattern. This is reflected in a sharp increase in the value of β_0^+, and the magnitude of this increment is strongly dependent on surface conditions: the surface patterns with a smaller characteristic length scale will cause the convective structures to break down into a larger number of smaller components at surface level than those surface patterns with larger scales. The value of β_1^+, on the other hand, is not stationary but appears to follow a broad parabolic curve, peaking after noon and decaying towards the evening. Here again the surface conditions play an important role: more heterogeneous surface patterns will be associated with larger values of β_1^+, on average.

By comparison, the time series for the horizontal layer at an altitude of 800 m show a negligible effect of the surface pattern on the behavior of the Betti numbers (see Figure 4.12), which is not surprising given the scale of land surface heterogeneity in the model. They do show one salient feature, which was absent closer to the surface: a sharp peak in the value of β_0^+, happening at around 10 h. After this, both β_0^+ and β_1^+ become stationary for the remainder of the day. The reason for this peak, as we will

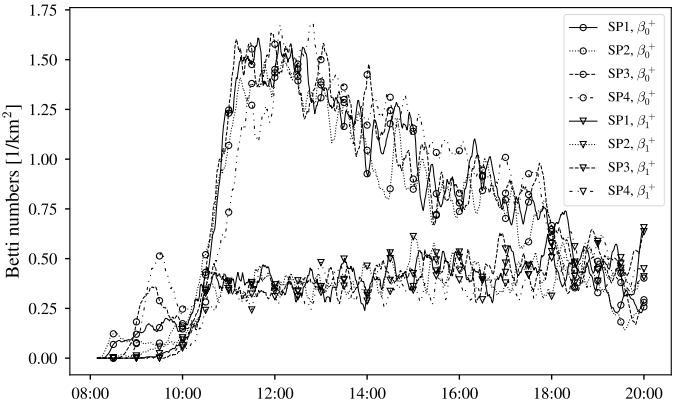


Figure 4.13: Time series of the 10-minute moving averages of β_0^+ and β_1^+ measured at 1600 m above the surface, for all four LES-ALM simulations.

see at the end of this section, is clear: it is a signal for the time at which the inversion layer grows past the altitude represented in this time series (800 m).

Finally, there is a negligible effect of surface pattern on the Betti number time series for the layer at 1600 m altitude (see Figure 4.13). For β_0^+, a transient peak at around 09:30 h is followed by a steep increase in value towards 10:30h, with all four time series reaching their maximum values between 11:00 h and 12:00 h. The rate of increment for SP4 is smaller than that observed in the other three simulations, hence the maximum value for this time series is observed later than for the other three. After reaching their maximum values, all four time series decrease during the afternoon. Thus, in contrast to what happens at 100 m and 800 m, the time series for β_0^+ is *not* stationary. Neither is β_1^+, with its variance increasing throughout the day, and a slight upwards trend noticeable in its 10-minute moving average.

In summary, the PBL system undergoes significant regime changes throughout its diurnal evolution. These changes in turn become manifest in qualitative differences in the convective structure present in the system at different times. The set of time series

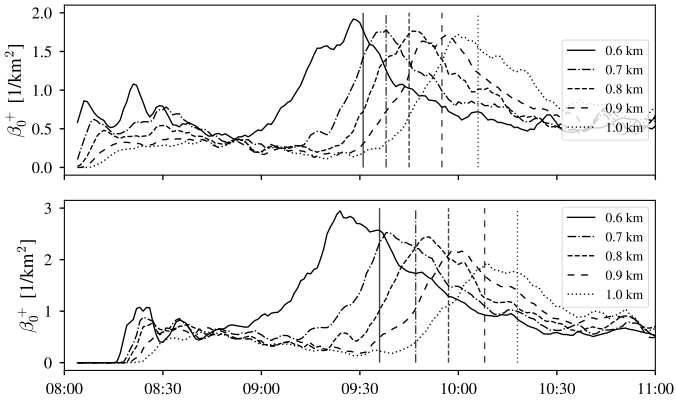


Figure 4.14: 5-minute moving averages of the time series for β_0^+, evaluated at different altitude levels. Two LES-ALM simulations are shown: SP3 (top) and SP4 (bottom). The vertical lines show the time at which the thermal inversion reached the altitude at which each time series was measured.

obtained here from the values of the Betti numbers give us a first glimpse at how these structural changes, product of a shift in the underlying physics, are reflected in the geometrical properties of the scalar field associated with vertical wind velocity. This is even clearer if we restrict ourselves to time series within the mixing layer, as seen in Figure 4.14. This shows time series of β_0^+ for five different altitude levels, in the case of SP3 and SP4: 600 m, 700 m, 800 m, 900 m, and 1000 m, between 08:00 h and 11:00 h. Here it is easy to see how the increasing altitude is reflected in the time at which the time series reaches its maximum: this happens at a later time for higher altitudes. The corresponding value of the maximum is also lower at higher altitudes, although this difference is more pronounced for the uniform case SP4, where the maximum goes from $2.93\,\mathrm{km}^{-2}$ at 600 m down to $1.93\,\mathrm{km}^{-2}$ at 1000 m, compared to the SP3 case where the change is between $1.91\,\mathrm{km}^{-2}$ at 600 m to $1.71\,\mathrm{km}^{-2}$ at 1000 m. As observed before, the time at which this value is achieved bears a close relation to the time at which the inversion layer grows to that altitude. This time is indicated by the vertical lines in Figure 4.14.

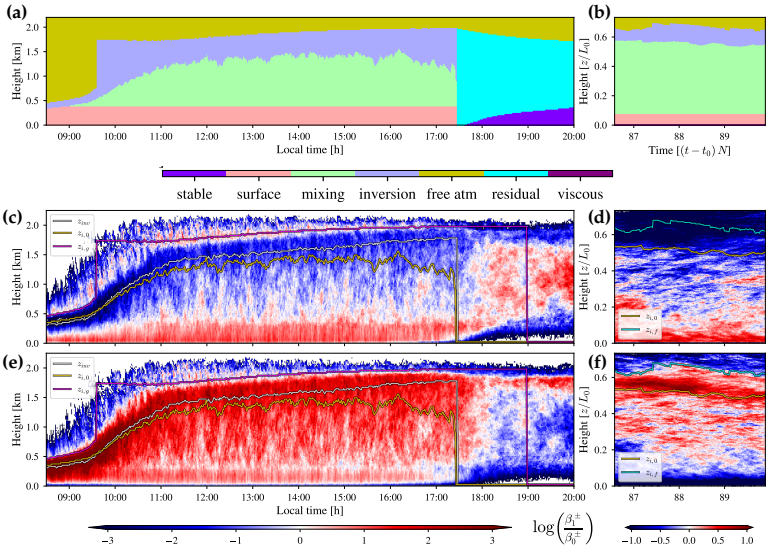


Figure 4.15: **(a)** and **(b)** The partitioning of the CBL into its components subregions by bulk analysis of the flow is shown in **(a)** for the LES dataset SP4, and in **(b)** for the DNS dataset. The contour plots show the height-time sections of $\ell^+ = \log(\beta_1^+/\beta_0^+)$ and $\ell^- = \log(\beta_1^-/\beta_0^-)$ for LES in panels **(c)** and **(e)**. The corresponding sections for DNS are shown in **(d)** and **(f)** (Figure from Licón-Saláiz *et al.* [2020]).

The temporal evolution of Betti numbers shows a limited sensitivity to land surface pattern, and this sensitivity is evident only at some points throughout the entire simulation. It also shows something which is hard to see if we consider vertical profiles: this temporal evolution differs, depending on which part of the boundary layer we look at. Some features are evident from the time series, such as the evening transition away from a CBL into stable stratification close to the surface, and the time of inversion crossing at higher altitudes. With this in mind, we can now think about what the topological data at our disposal reveal about the global structure of the PBL.

4.4 TOPOLOGICAL CHARACTERIZATION OF THE PBL

The general situation is more clearly visible in Figure 4.15 (c-f). Here we do not show the Betti numbers individually, but the value of their log-quotients, $\ell^{\pm} = \log(\beta_1^{\pm}/\beta_0^{\pm})$. The use of this quantity follows the observations made in Section 4.3.2, where the combination of β_0 and β_1 was shown to have more explanatory power than either number on its own. The quotient β_1/β_0, which can be understood as a normalization of the number of generators in the homology group by the number of data points, has been used successfully as a descriptor in the classification of plane shapes [Chacholski and Riihimäki, 2020]. In the present case, a further advantage of β_1/β_0 is that it is a dimensionless quantity. Furthermore, we also visualize the values of $\ell^{\pm}$ for both the LES (panels c and e) and DNS (panels d and f). Based on this visualization, the following observations can be made:

- The global structure for the four LES simulations (only SP4 is shown here) is similar. The radiatively-driven CBL structure is also similar in both the LES and DNS models. In the former, boundary layer growth starts in each at around 09:30h, slowing down at 11:00h, and persists until the late afternoon. Then, at around 17:30h, the inversion layer disappears. The time period between the start of boundary layer growth and the collapse of the thermal inversion corresponds to the quasi-stationary CBL regime. For the DNS dataset, only the CBL regime is shown here.

- Within the CBL, four distinct regions are apparent from the values of ℓ^{+}: the surface layer, the mixing layer, the inversion layer, and the free atmosphere.

- After the collapse of the inversion layer, and especially after 18:00h, the transition to a stably stratified boundary layer occurs. This is reflected in the increasing values of β_0^{+}, and as is clear from Figure 4.11, the magnitude of this increase is closely related to the length scale of the underlying surface pattern heterogeneity. Above this, the mixing layer becomes

a region of residual turbulence, decoupled from the surface layer as the energy input from the land surface disappears.

- In the LES, the white region at the top indicates the free atmosphere, characterized by an absence of large, turbulent fluctuations. In terms of the ternary partitioning of W introduced before, this means that no connected regions where vertical wind speed has a magnitude larger than the threshold $\varepsilon = 0.01\,\mathrm{m\,s^{-1}}$ are observed. In DNS this abrupt change does not happen.

Also shown are four different measurements of boundary layer height:

1. *Inversion height*, z_i: where the mean buoyancy gradient, $\frac{\partial \langle b \rangle}{\partial z}$, becomes positive.
2. *Zero-crossing height*, $z_{i,0}$: where the total buoyancy flux, B, becomes negative. B is defined as
$$B = \mathrm{Cov}(b, w) - \kappa \frac{\partial \langle b \rangle}{\partial z}.$$
3. *Variance-based height*, $z_{i,v}$: where the buoyancy variance, $\mathrm{Var}(b)$ is maximum away from the surface region.
4. *Flux-based height*, $z_{i,f}$: where the total buoyancy flux B is minimized [Garcia and Mellado, 2014].

The inversion layer is then considered to be the space between the minimum and the maximum of these different values.

At this point we also note the qualitative agreement between the values of the log-quotients of the Betti numbers on the one hand, and the structure and diurnal evolution of the PBL (see Figure 2.1). This motivates a qualitative classification of the regions of the (t, z) plane based on the values of $\ell^{\pm}$. More precisely, we want to know if it is possible to recover a partitioning of the PBL into its component subregions, such as the one shown in Figure 4.15 (a,b), from knowledge of the Betti number log-quotients only. This partitioning, obtained from bulk measurements of the flow, again shows the characteristic diurnal evolution of the PBL, at least

for the LES: starting at 9:30 h, and until the thermal inversion disappears in the afternoon, the first 300 meters above the surface constitute the surface layer (shown in red). Above this is the well-mixed layer (green), followed by the inversion layer (dark blue) and the free atmosphere above (yellow). The lower (resp. upper) bound of the inversion layer is given by $z_{i,0}$ (resp. $z_{i,\nu}$). Finally, after the inversion disappears and the atmosphere transitions to a stably stratified regime, we are left with residual turbulence (light blue). After this point in time, the height of the residual layer decreases gradually, we model this using an exponential decay function of the form

$$f(t) = z_0 + \delta e^{-\kappa t}.$$

In view of this clear structural partitioning, we can now leave aside the issue of determining the discriminative power of the Betti numbers with respect to the different land surface patterns, and focus instead on their capacity to distinguish between the different regions of the PBL shown above. In other words, we want to find out whether knowing only the two numerical values $\ell^{\pm}$ at a point (t, z) is enough to decide whether this point belongs to the surface layer, or the well-mixed layer, or any of the other regions.

4.4.1 *Supervised learning*

The first approach is again supervised learning, where we build a feature matrix X with one row for each pair (t, z) of time and height, excluding the first 90 minutes of simulation time. For simplicity, we will only focus on the LES case in this subsection. X will have two feature columns with the value of ℓ^+ or ℓ^- respectively. The response variable y will be the label of the boundary layer region to which the pair (t, z) belongs according to the partitioning shown in Figure 4.15 (a, b). In this case we consider a single feature matrix containing all data from the four LES-ALM simulations, which results in a matrix with 231869 rows. 70% of these were selected at random as training set, the remaining 30% as test set. The k-NN classifier (with k=15 in this case) has a F_1 score of .71 on the test set, while the multinomial classifier (with

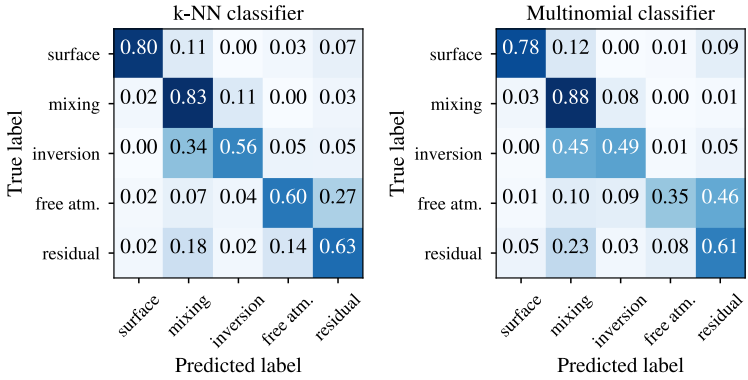


Figure 4.16: Normalized confusion matrices for k-NN and multinomial logistic classifiers, for the task of predicting the region of the PBL corresponding to a pair of values (ℓ^+, ℓ^-).

regularization parameter C=1.5) has a F_1 score of 0.67 on the test set. The classification results are summarized in Figure 4.16, which shows the confusion matrices for both classifiers.

We can see that both classifiers perform best at classifying both the surface and mixing layers, which makes sense since the surface layer appears well-defined by the values of ℓ^+ in Figure 4.15 (c), with its transition to the mixing layer being clearly defined and regular across time. Classification accuracy drops significantly in distinguishing the inversion layer from the mixing layer, and the free atmosphere from the residual layer. This is again reflected in the fact that the boundaries between these regions are more diffuse, as seen in Figure 4.15 (c, e).

4.4.2 *Unsupervised learning*

Supervised learning gives a good result in terms of classification accuracy. However, it is not especially surprising since we are imposing our previous knowledge on the classification model, in some sense, by defining what the values of the target variable y must be from looking at the values of the Betti numbers, and how these change over time. A more interesting question to ask

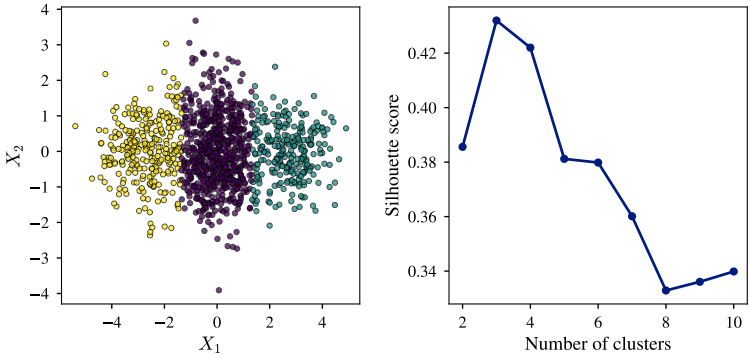


Figure 4.17: Left: set of 2400 points drawn from a mixture of three bivariate normal densities. The colors represent the cluster membership as determined by the K-means algorithm. Right: silhouette scores obtained from performing K-means clustering with $k = 2, \ldots, 10$.

would be: is it possible to recover this structural partition of the space-time domain by using only the values $\ell^{\pm}$, and not using any predefined labels? This is, of course, a problem of unsupervised learning. In contrast to the situation we had before, where the response variable y is defined by the values (t, z) that determine a two-dimensional slice from which Betti numbers were computed, we will now attempt to characterize the probability density of observations in our feature matrix X without an explicit reference to the underlying values of (t, z). Only after obtaining this abstract model for the probability density of X will we then compare it to our ground truth, which in this case is the classification of (t, z) values into the different regions of the PBL.

What kind of model this might be is already hinted at by the terminology used: what is needed is a partitioning of the available values of $\ell^{\pm}$ into groups, according to some similarity measure, and then we can compare this partition with the known partitioning of (t, z) space (see Figure 4.15 (a, b)). This is an example of *cluster analysis*, one of the most common techniques in unsupervised learning [Hastie *et al.*, 2001, §14.3]. A simple and popular technique for cluster analysis is the K-means algorithm, which

we briefly review here. Given a set $\{x_i\}$ of N observations, and k points $\bar{x}_j$ chosen as initial centroids at random:

1. Assign each point x_i to cluster j, where $\bar{x}_j$ is the centroid closest (in Euclidean metric) to x_i.
2. After assigning all observations to a cluster, update each $\bar{x}_j$ to be the centroid of the new clusters,

and this is repeated until convergence is achieved. Incidentally, we note that cluster analysis has been considered to be a technique in topological data analysis [Carlsson, 2009], as it amounts to determining the connected components that underlie a given data sample. While this is certainly a legitimate interpretation of cluster analysis, the technique is actually more general. Consider, for example, the situation illustrated in Figure 4.17 (left), which shows 2400 points drawn from

$$\sum_{i=1}^{3} w_i \, \mathcal{N}_i(\mu_i, \sigma_i^2 I_2),$$

a mixture of three bivariate normal densities, with $\mu_1 = (0, 0)$, $\mu_2 = (3, 0)$, $\mu_3 = (-3, 0)$, and $\sigma_1^2 = 1$, $\sigma_2^2 = \sigma_3^2 = 1/2$. The weights w_i are (2/3, 1/6, 1/6). The colors indicate the result of applying the K-means algorithm to this pointcloud, with $k = 3$. This also illustrates an important feature of K-means clustering: the decision boundaries found by the algorithm are linear (and indeed coincide with the Voronoi tessellation induced by the centroids $\bar{x}_j$). The right panel shows the *silhouette score* [Rousseeuw, 1987] for a clustering with k centroids, $k = 2, \ldots, 10$. The *silhouette coefficient* for observation x_i is defined as

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}},$$

where $a(i)$ is the mean distance from x_i to all other points in its own cluster, and $b(i)$ the minimum mean distance from x_i to all points in one cluster, with the minimum taken over all clusters C except for the one to which x_i belongs. The silhouette score is then the mean silhouette coefficient over all observations x_i,

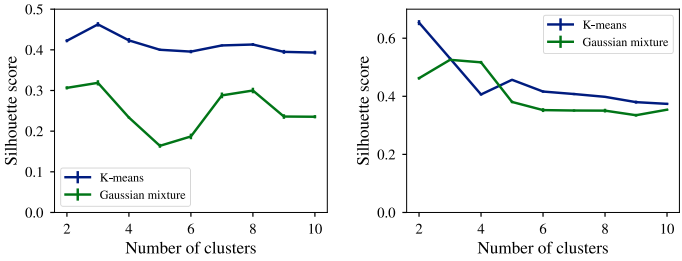


Figure 4.18: Silhouette scores for clustering of LES (left) and DNS (right) data.

and serves as a metric to evaluate a cluster assignment, since the coefficient $s(i)$ encodes both the proximity of x_i to other elements in its own cluster, which we would expect to be small, and the dissimilarity between x_i and all other clusters, which we would expect to be large. If indeed $b(i) \gg a(i)$, then we consider observation x_i "well-clustered", and $s(i) \approx 1$. In the example shown in Figure 4.17, the maximum silhouette score is achieved with $k = 3$ clusters, which agrees with the fact that 3 different centroids (corresponding to the means of the mixture component) underlie the pointcloud, even though it is hard to argue for the existence of three disconnected components. After this short parenthetical remark, we return to the problem of classifying the points in (t, z)-space from the values of their corresponding Betti numbers alone. This is similar to the situation in Figure 4.17, where instead of the mixture density we now have $X_1 = \ell^+$, and $X_2 = \ell^-$. As observations we use the 5-minute rolling means of the variables, and perform a K means clustering with $k = 6$. The choice of $k = 6$ is motivated by two considerations. First, the silhouette score for different values of k only shows a clear difference for $k = 3$ (see Figure 4.18), in the case of K-means clustering, and for $k = 3, 7, 8$ for gaussian mixture clustering. Second, we know that we would need more than 3 clusters if we hope to capture the structural differences in the data, if only because the transition from a convective boundary layer regime, with three distinct components, to a stably stratified regime already necessitates at least 4 clusters. We then focus on $k = 6$, with the result as shown in Fig-

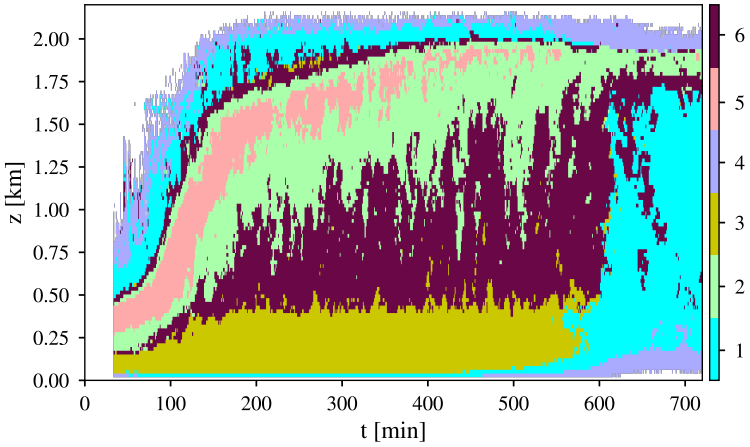


Figure 4.19: K-means clustering of the (t, z) domain based on values of ℓ^+ and ℓ^-.

ure 4.19. We can go one step further, and obtain a slightly more refined clustering by replacing K-means with another, closely related algorithm: Gaussian mixture estimation. In this method, the starting point is again an arbitrary assignment of k centroids, which now correspond to the means of the k components of a mixture of Gaussian densities. The observations x_i are then assigned to each cluster probabilistically. After assignment, the centroids are recomputed, and the process is iterated until convergence is achieved [Hastie *et al.*, 2001, §14.3.7] The result for this is shown in Figure 4.20, this time for $k = 7$ clusters.

We can draw two conclusions from the results obtained by both algorithms. In the first place, the fact that both classifications of points in (t, z) space are qualitatively similar, insofar as both exhibit regions which can be construed as a well-defined surface layer and mixing layer which transitions into an inversion layer, beyond which a more irregular behavior is observed. Both classifications also show a structural change for values of t after about 550 min, characterized by the vanishing of the three well-defined regions. The agreement of both classification schemes points to the fact that there exist regions of the ℓ^+, ℓ^--plane in which the data

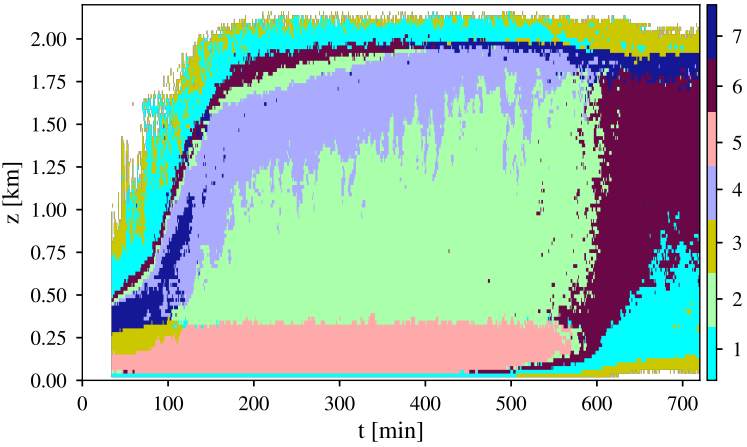


Figure 4.20: Gaussian mixture clustering of the (t, z) domain based on values of ℓ^+ and ℓ^-.

points tend to cluster together, and the data points in each of these regions correspond to points in each of the known PBL regions in (t, z) space. The second conclusion is that, since the different clusters obtained from GMM have clearer, less noisy boundaries in (t, z) space than the clusters obtained from K-means, we we can infer that the boundaries between the putative regions in the ℓ^+, ℓ^--plane are actually non-linear. To get a clearer understanding of this, the distribution of values for the two variables ℓ^+ and ℓ^- is shown in Figure 4.21. Here we visualize the individual values classified according to our ground truth (Figure 4.15 (a, b)), and the corresponding decision surfaces obtained from a k-NN classifier as described above. We can see that, as in the example with three normal densities we discussed before, there is no clear separation of the data into disjoint sets, but there do appear to be regions of higher data density. Moreover, at least three of these regions coincide with physical regions in the PBL: surface, mixing, and inversion layers. The clearest feature of the distribution is doubtlessly the scattering of data points about the line $X_2 = -X_1$, especially for points in the mixing layer, inversion layer, and residual turbulence. Most of the points from the surface layer sit within

a convex region in the semiplane $\{\ell^- > -\ell^+\}$, whereas points from the free atmosphere and a small part of the residual layer lie on the opposite semiplane, $\{\ell^- < -\ell^+\}$. We also indicate six points on this figure, labeled A-F, which correspond to each one of the six two-dimensional slices shown in Figure 4.23.

In addition, we can also consider the division into quadrants according to the signs of X_1 and X_2, and what this means in terms of the geometry encoded by the corresponding Betti numbers. By definition, $\ell^+ > 0$ implies that $\beta_1^+/\beta_0^+ > 1$, or equivalently $\beta_1^+ > \beta_0^+$. Accordingly, $\ell^+ < 0$ implies that $\beta_1^+ < \beta_0^+$. The same relations hold for ℓ^-. Recalling the geometric meaning of the Betti numbers, $\beta_1/\beta_0 > 1$ can then be interpreted as the existence of many "holes" or "loops" in the corresponding domain, whereas $\beta_1/\beta_0 < 1$ would indicate that most of the connected components in the domain are acyclic (they are contractible, or loop-free). The correspondence between these geometrical characteristics and flow morphology in the different PBL regions is illustrated in Figure 4.23. As can be seen in Figure 4.21, the points from each of the PBL regions are not scattered randomly, but tend to cluster in different parts of the ℓ^+, ℓ^--plane. This tendency to form clusters points to the different flow morphology present in each PBL region being represented by the two values ℓ^+ and ℓ^-. This is best illustrated by comparing the six points A-F from Figure 4.21 with the two-dimensional slices they represent, shown in Figure 4.23. Point A, representing the first quadrant, has both $\ell^+ > 0$ and $\ell^- > 0$. This means that for both the positive and negative wind velocity domains, it is the case that $\beta_1 > \beta_0$, i.e. that the number of cycles is larger than the number of components. Geometrically this means that both domains are intertwined in a complex network-like pattern as seen in Figure 4.23, panel A. For points B and C in the second quadrant, we have $\beta_1^- > \beta_0^-$, but $\beta_1^+ < \beta_0^+$. Most of the components in the positive domain can thus be expected to be acyclic, which means that the network pattern has been replaced by a small number of large components, which do not tend to encircle areas of negative wind velocity. Physically this can be interpreted as the coalescence of updrafts into large convective plumes, which is especially clear in the mixing layer (Figure 4.23, panel B). For points in the fourth quadrant

the opposite will be true: most of the components in the negative domain are acyclic, which is especially clear in the residual layer (Figure 4.23, F). Here we can see that the negative domain is made up by many small acyclic components ($\beta_1^- / \beta_0^- < 1$), all of which are surrounded by the positive domain, itself consisting of one large dominant component and many smaller, acyclic ones ($\beta_1^+ / \beta_0^+ > 1$). Physically this would mean that the convective plumes, deprived of energy from the surface, are now being slowly degraded by cooling air from above. Finally, the point E is located in the region with $\ell^{\pm} < 0$, which implies that most components of both domains are acyclic. Geometrically this is only possible if neither domain tends to encircle the other, but rather both are surrounded by the region consisting of near-zero values. We see that this is indeed the case in Figure 4.23, panel E. This also makes physical sense, as point E corresponds to the stable stratification close to the surface, where the velocity fluctuations have become very small. Figures 4.22 and 4.24 show the log-quotient scattering and representative two-dimensional slices for the DNS data, where the differences between the main CBL subregions are also expressed by changes in the mean log-quotient values found in each subregion.

4.4.3 *Semi-supervised learning*

The qualitative agreement exhibited by the unsupervised learning techniques with the expected structure of the atmospheric boundary layer is significant, yet two important questions remain:

1. How to give a quantitative evaluation of the clustering, in terms of classification accuracy?

2. Is it possible to improve on the automatic cluster assignment by leveraging our knowledge of the physical constraints on the data points? For example, since it is clear which clusters correspond to the inversion layer in Figure 4.20, we can merge all other clusters which are above it in physical space into a new cluster, and declare this to be the free atmosphere.

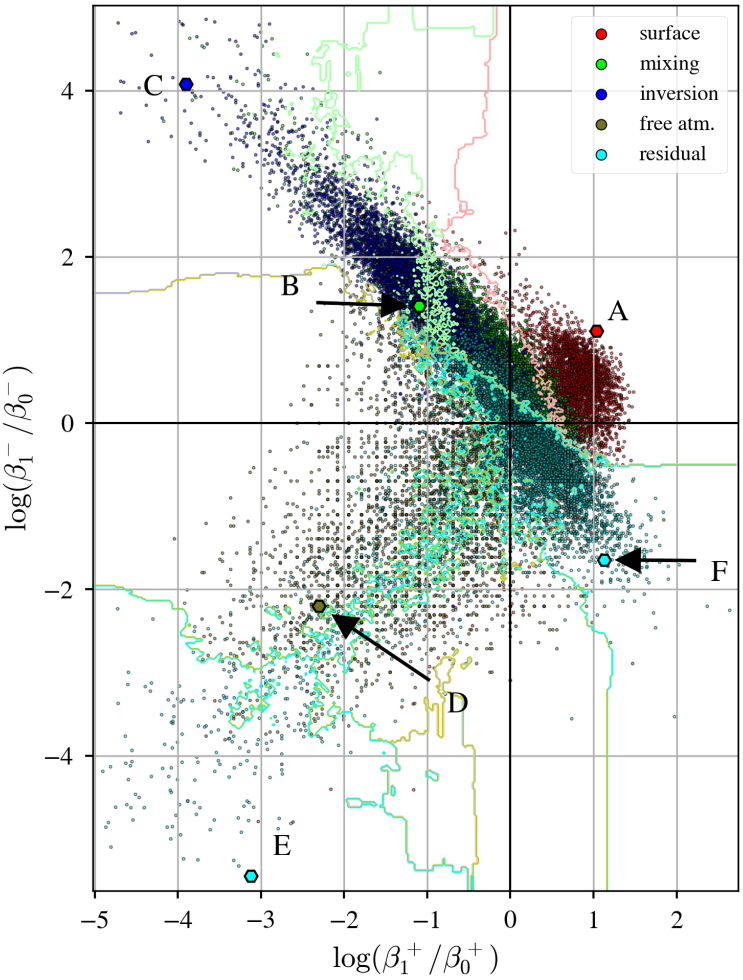


Figure 4.21: Scatterplot of the two Betti number log-quotients for the LES-ALM simulation SP4. The decision surfaces shown are computed using a k-NN classifier with $k = 15$ trained on the feature matrix X containing all observations of the two variables, with the response variable being the PBL subregion assigned by bulk analysis of the flow. The 6 points A-F correspond to the two-dimensional slices shown in Figure 4.23.

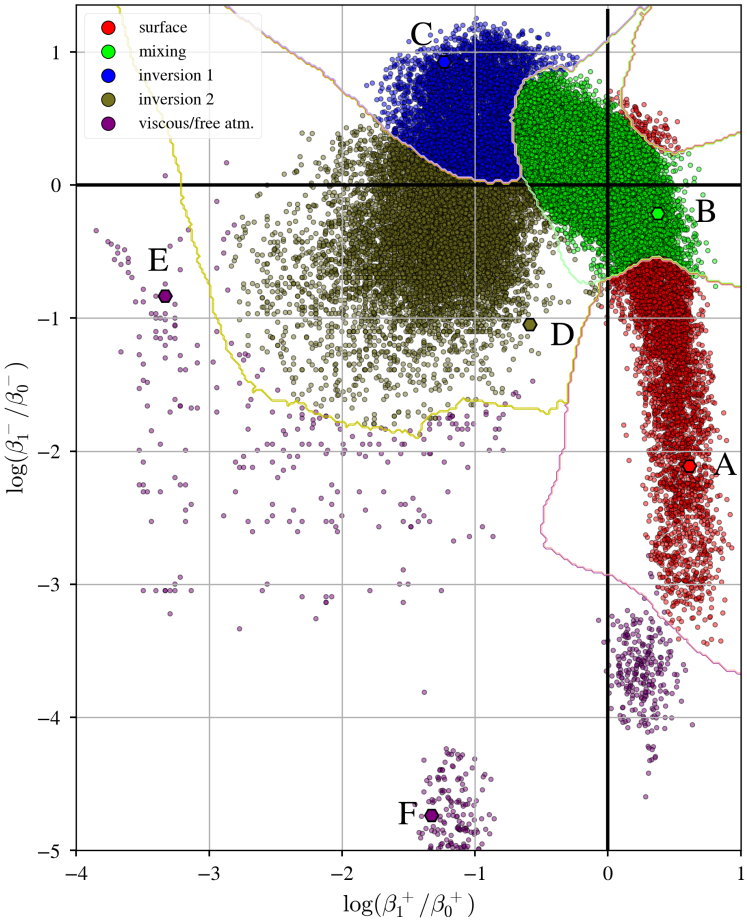


Figure 4.22: Scatterplot of the two Betti number log-quotients for the DNS dataset. The clustering was obtained using a gaussian mixture model with $k = 5$ components. The decision surfaces are then computed using a k-NN classifier with $k = 5$, trained on the feature matrix X containing all observations, and the response variable is the unsupervised classification induced by GMM clustering, for $k = 5$ clusters.

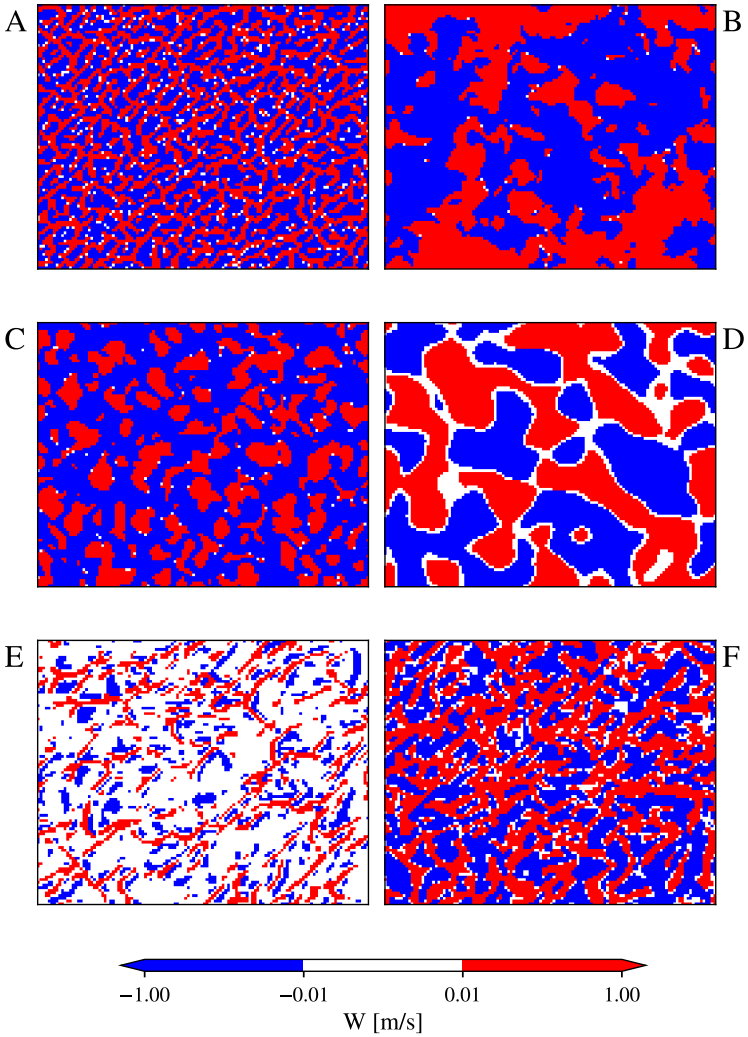


Figure 4.23: Two-dimensional slices of the vertical wind velocity field taken from LES-ALM simulation SP4. The values have been discretized according to the thresholding scheme discussed in Section 4.2.2. A: 08:51 h at 88 m, surface layer. B: 14:00 h at 880 m, mixing layer. C: 09:54 h at 814 m, inversion layer. D: 10:42 h at 2002 m, free atmosphere. E: 17:32 h at 22 m, residual layer, stably stratified. F: 17:26 h at 330 m, residual turbulence layer. These slices correspond to the same data shown in Figure 4.1.

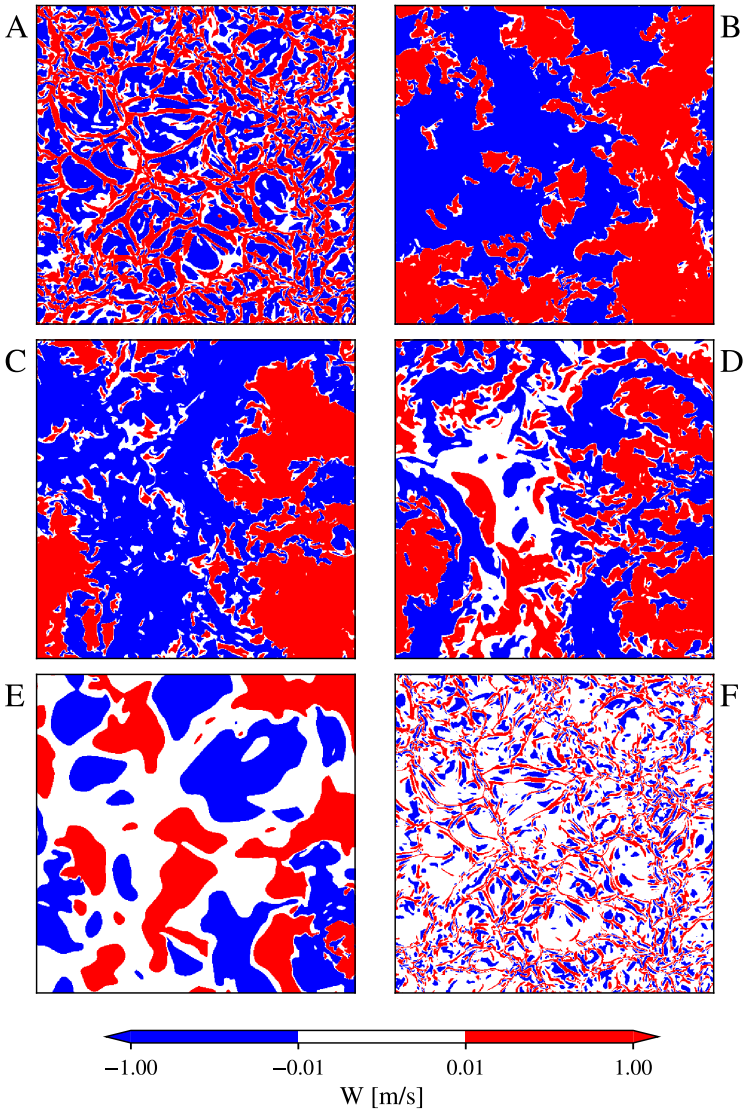


Figure 4.24: Two-dimensional slices of the vertical wind velocity field taken from the DNS dataset, with the same thresholding scheme as in Figure 4.23. A: surface layer. B: Mixing layer. C: Inversion 1. D: Inversion 2. E: Free atmosphere. F: Viscous layer.

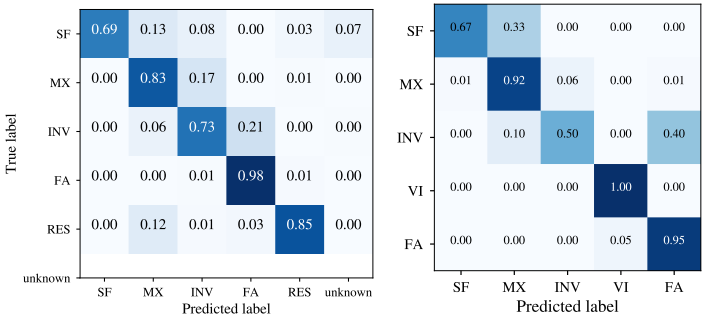


Figure 4.25: Confusion matrix for semi-supervised classification of CBL regions in LES (left) and DNS (right).

This process of mixing an unsupervised learning method with a small set of labeled data is aptly named *semi-supervised learning* in the machine learning literature. We can take as starting point any of the clusterings produced by either k-Means or GMM. We choose the latter (as shown in Figure 4.20), since it produces a less noisy cluster assignment, and shows a clearer separation of the surface, mixing, and inversion layers. The first task is to identify those clusters which can be reliably assigned to one of the known regions of the PBL, as shown in Figure 4.15. In the example shown here, this is the case for the surface, mixing, and inversion layers, as well for the residual turbulence during the evening. For each region, the correspondence of the different clusters is given by the percentage of (t, z) pairs for that region assigned to each cluster. If this percentage is high enough, we assign that cluster to the corresponding region. In this example we see that cluster 5 corresponds to the surface layer, cluster 2 to the mixing layer, cluster 4 to the inversion layer, and cluster 6 to the residual layer. It remains to be seen what to do with clusters 1, 3, and 7, which is where the physical considerations come into play. Clusters 3 and 7 include part of the inversion layer in the first part of the simulation. The first rule is therefore to merge these components into cluster 4. With this we can identify the inversion uniquely as cluster 4. The next step is to classify every data point above the top of the inversion layer as free atmosphere

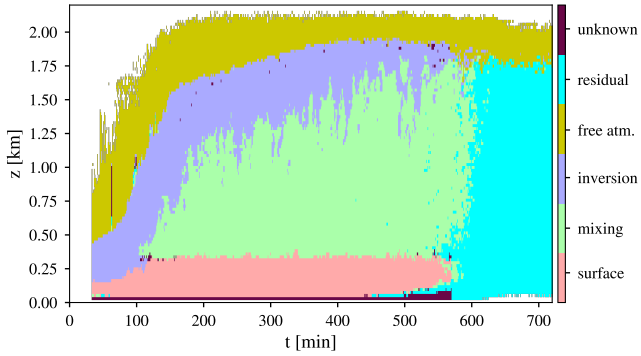


Figure 4.26: Result of applying semi-supervised classification to the log-quotient variables for the LES dataset SP4.

for that part of the simulation which corresponds to the CBL regime, represented here by the existence of well-defined surface, mixing, and inversion layers. For the second part of the day, after buoyancy-driven convection ceases, we classify everything above the already identified residual layer as free atmosphere as well. Finally, we merge the components of clusters 1 and 3 that sit beneath the residual layer into the residual layer itself.

We can see the result of these modifications to the original clustering in Figure 4.26. Figure 4.25 (left) shows the confusion matrix for this classification, which has a weighted average F_1 score of 0.82, which is even better than the supervised approach discussed earlier. This good performance can be attributed to the strong relationship of the two variables used, $\ell^{\pm}$, with morphological properties of the vertical wind velocity field in different regions of the PBL. In other words, these two features can be understood as strong descriptors of the physical state of the system. Further, this approach does not suffer from the ambiguities introduced by the ad-hoc partitioning sketched in Figure 4.15 (a, b).

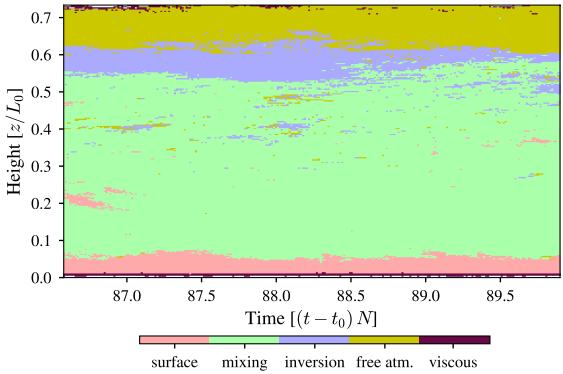


Figure 4.27: Unsupervised classification for DNS dataset, induced by GMM clustering with $k = 5$ clusters.

4.4.4 *Model comparison*

As a last step in the analysis of the Betti numbers as descriptors for a CBL, we will compare the results obtained in Section 4.4.2 using data from a large-eddy simulation, with those obtained from applying the same techniques to data from a direct numerical simulation (DNS). In this model all spatial and temporal scales of the turbulent flow are resolved, hence we should obtain a much more detailed picture of its structural properties than from a large-eddy representation. The dataset used here comprises 256 timesteps, with a spatial domain of 512×512 gridcells in the horizontal direction, and 235 gridcells in the vertical direction. This translates to 60160 two-dimensional slices of the vertical velocity field from which we then compute the same four Betti numbers as before, after applying the symmetric thresholding at -0.01 and 0.01. Figure 4.28 shows the vertical profiles for these four Betti numbers, averaged over 60 timesteps. As was also the case for the LES data, we again see that the largest variation amongst these numbers happens close to the surface. Throughout the mixing layer they are, on average, very similar, and start to diverge again as we approach the inversion layer. Taking this as

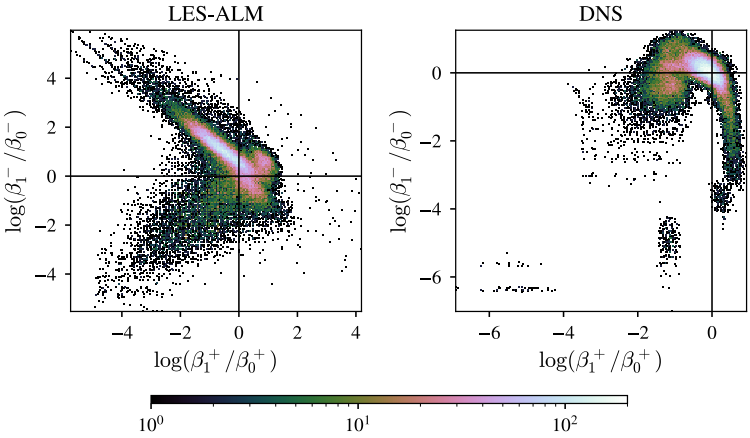


Figure 4.29: Two-dimensional histogram showing the two Betti number log-quotients for LES (left) and DNS (right) data. As before, the LES dataset is simulation SP4.

starting point, we might conjecture that clustering on the values of these Betti numbers would also yield sensible results.

We proceed as before and compute the two log-quotients, $\ell^{\pm}$, which will be used as features to be fed into the clustering algorithm. In this case, however, we do not compute the 5-minute moving average. The distribution of these two variables is illustrated in Figure 4.29. The most striking feature in this scatterplot is again the strong negative relationship between ℓ^+ and ℓ^-. This is similar to what happened in the LES case (see Figure 4.21), but the relationship here is more clearly non-linear. The LES data also had as a salient feature a clearly defined "lobe" which accumulated most of the values of ℓ^+, ℓ^- for the surface layer, but no such feature is apparent for the DNS data.

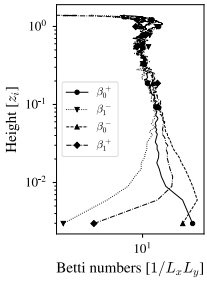


Figure 4.28: Betti number profiles for DNS data.

The figure shows a large accumulation of points around the origin. Two "arms" stretch out of this putative centroid into each of the two semiplanes defined by $\ell^+ > 0$ and $\ell^- < 0$. The latter "arm" exhibits a more irregular shape, with the data points becoming more loosely scattered towards its end. Finally, we can also see three distinct accumulations of points that go from the lower left part of the diagram towards the bottom part of the right "arm". These features are visually clear, and the shape apparent in the diagram is indeed very suggestive. We will now make these empirical observations more precise. We consider to this end the silhouette scores for different clusterings of the data, shown in Figure 4.18. According to this, a sensible clustering of the data would feature anywhere between 2 and 5 clusters. As a first approach, we compute a K-means clustering for $k = 3$ clusters (not shown here). The result is as we would expect: a division of the (t, z) space into a surface layer, a turbulent region, and the inversion layer. However, the qualitative features observed in the scattering of the variables suggest using a larger number of clusters. The result of running a gaussian mixture clustering with $k = 5$ components is shown in Figure 4.22 as decision surfaces on the ℓ^+, ℓ^- plane. We can see that the distinct regions of the bivariate scattering discussed above do agree with a physically meaningful partitioning of (t, z) space, as shown in Figure 4.27. Here we have declared the blue region as the inversion layer, and the yellow region above it as the free atmosphere. This has a significant impact in classification performance for the inversion layer (cf. Figure 4.25, right). Figure 4.27 also shows 6 points, labeled A-F, which correspond to the six two-dimensional slices of the vertical wind velocity field in Figure 4.24. As can be seen from the scatterplot, starting from the lower left there is a broad counterclockwise ordering of the distinct clusters, and this ordering corresponds to a bottom-up motion in the vertical direction in physical space. This again reveals how the different values of ℓ^+ and ℓ^- encode the structural properties of the various PBL subregions, and how the partitioning of the domain by these values alone broadly agrees with a partitioning obtained by bulk analysis of the flow (cf. Figure 4.15). We also find significant qualitative

agreement between the partitioning obtained for LES and DNS. The most salient differences encountered are:

1. The appearance, for DNS, of a *viscous layer* directly adjacent to the surface.

2. Ambiguity of the definition of the entrainment zone in DNS, as the unsupervised algorithm produces two distinct clusters for this region, and the transition between entrainment zone and free atmosphere is not very clear. This is actually in line with the two-layer structure of the entrainment zone as described by Garcia and Mellado [2014], and is a consequence of the explicit representation of small-scale entrainment in DNS, a process which cannot be explicitly represented by an LES model.

3. The values of ℓ^- in the surface layer are negative for DNS, but positive for LES. This reflect the presence of wind shear in the LES simulation, which changes the interspersion pattern between up- and downdrafts close to the surface. A quick comparison of Figure 4.23 (A) and Figure 4.24 (A) shows that, indeed, in the LES case the number of updrafts completely enclosed by downdrafts is greater than in DNS.

4. A more abrupt transition from the entrainment zone to the free atmosphere in LES. This fact reflects the strong capping inversion which is imposed as an initial condition in this model, and is absent from DNS.

5

CONNECTIVITY AND THE SPATIAL ORGANIZATION OF CONVECTIVE FLOW

Outline

We now focus on the special case of zero-dimensional homology, which requires simpler computational tools than the matrix operations necessary for its higher-dimensional counterparts. These tools are first introduced in Section 5.1, where it is shown how they can be applied to the case of data analysis for atmospheric models, by identifying the connected components that make up a given domain in space (updrafts in this case). Once we have identified these components, their sizes are found to exhibit self-similar scaling, and this scaling interacts with the underlying land-surface patterns (Section 5.2). A comparison of the predictive power of these scaling laws with that of the Betti numbers found in Chapter 4 is also made. Furthermore, another kind of topological invariant is introduced here, the merge tree, which makes use of connectivity information to furnish a representation of the three-dimensional coherent structures that form in the CBL (Section 5.3). The conference paper Licón-Saláiz and Ansorge [2019] is based on this chapter.

In Chapter 4, we study the interaction between land surface and atmosphere as well as the structure of the atmospheric boundary layer based on the numerical values of topological invariants: the Betti numbers. These invariants are coarse, in the sense that they only quantify the number of structures present, but give no further information regarding their properties, such as their size or

location. From the Betti numbers we can know that there are 2 cycles present in our domain, but we cannot know where in the domain or of what size they are. In technical terms, this would require identifying representatives for them. As it turns out, this is still a major open problem in computational topology, and subject of ongoing research [Escolar and Hiraoka, 2014, 2016; Obayashi, 2018]. At the heart of this problem lies the very definition of homology groups as equivalence classes of cycles. This means that the currently available algorithms for the computation of homology compute arbitrary representations of the cycles involved. This is sufficient to provide the ranks of the involved groups, i.e. the Betti numbers, but sheds no light on what structures those cycles could correspond to.

An example of this is shown in Figure 5.1, where we show a cubical complex X with $\beta_0(X) = 1$, $\beta_1(X) = 2$. Also shown are two cycles which both belong to the same equivalence class in the first homology group of X, $H_1(X)$. One of them corresponds exactly to the boundary of the hole that is one of the generators of H_1. Yet when using a matrix reduction algorithm to compute

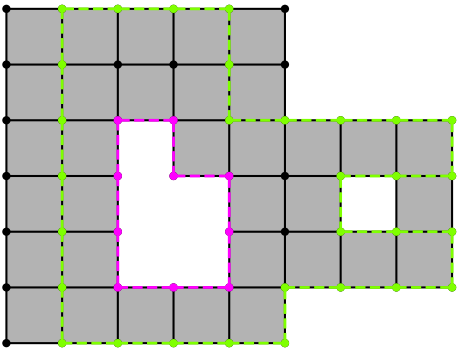


Figure 5.1: Cubical complex (grey) with two 1-cycles homologous to each other (green and magenta). Both are representatives of the same feature in the complex, yet only one (magenta) can be thought of as being "optimal" in terms of area or volume.

Betti numbers, there is no guarantee that such a representative will be obtained. It could also be the other cycle, shown in green, or indeed any other cycle homologous to it.

With this in mind, we can consider the special case of 0-homology, which turns out to be more amenable to computations. To see this, we recall the result stated in Section 3.4 according to which the zero-dimensional homology group of a complex X is a free abelian group with basis

$$\{\hat{x}_i \in C_0(X) \mid i = 1, \ldots, n\},$$

where $\{x_i \mid i = 1, \ldots, n\}$ is a collection of vertices such that there is one, and only one, vertex for each connected component of X (see Theorem 3.4.5). What this means is that computation of $H_0(X)$ is then equivalent to separating the connected components of X. This is a classical problem in computer science, which has a very efficient algorithmic solution as described by Hopcroft and Ullman [1973], making use of the Union-Find (UF) data structure.

5.1 THE UNION-FIND DATA STRUCTURE

For this dissertation, we implement the UF data structure in Python (see Listing 5.1), making use of the NumPy library. This allows us to work directly with the cubical complex representation of numerical simulation data. An additional component that needs to be implemented is the special case of periodic boundary conditions, since the two-dimensional slices obtained from LES models with doubly periodic boundary conditions are essentially cubical complexes embedded in the flat two-dimensional torus. Our implementation of the UF algorithm (see Algorithm 1) and data structure is based on the description given in Edelsbrunner and Harer [2010, §I]. It consists of a linear array of parent pointers of the same size as the cardinality of our vertex set V, together with a `Find` and a `Union` operation. The `Find` operation finds the parent node for a given vertex x; the `Union` operation, given two nodes puts them in the same connected component by pointing both to the same parent node. The algorithm starts from an array where each node is its own parent, i.e. it points to itself. All nodes are then visited sequentially, updating the parent pointers with

Listing 5.1: Python implementation of the Union-Find data structure.

```
class UnionFind:
    import numpy as np
    def __init__(self, n):
        """
        n : int
        """
        self.parent = np.array(range(n), dtype=np.uint32)
        self.size   = np.ones(n, dtype=np.uint32)

    def find(self, i):
        """
        i : int
        return p : int
        """
        a = i
        b = self.parent[i]
        if a == b:
            # Node is its own parent, do nothing.
            return i
        # Go up the hierarchy until root is found.
        while a != b:
            a = b
            b = self.parent[a]
            # Root is now stored in b.
            p = b
            # Return to starting level, go back up pointing everything to the root along the way.
            a = i
            b = self.parent[i]
        while a != b:
            self.parent[a]  = p
            a = b
            b = self.parent[a]
        # Return root.
        return p

    def union(self, i, j):
        """
        i : int
        j : int
        return : void
        """
        a = self.find(i)
        b = self.find(j)
        if a == b:
            return
        if self.size[a] > self.size[b]:
            self.parent[b] = a
            self.size[a] += self.size[b]
            self.size[b] = 0
        else:
            self.parent[a] = b
            self.size[b] += self.size[a]
            self.size[a] = 0
```

the `Union` operation according to node connectivity. An example of this is shown in Figure 5.2, where the top part shows a cubical complex, with its 2-cells labeled by numbers. In this case the 2-cells play the role of nodes, and we define them to be connected if they are adjacent to one another, that is, if they share an edge. The bottom part shows the state of the parent pointer array after visiting all nodes.

This algorithm is optimized in two ways:

1. Paths are compressed when they are traversed: when running `Find`(x), the corresponding root node found is also declared to be the parent node of x if it is not already.

2. Smaller sets are always absorbed by large sets. When calling `Union`(x, y), the sizes of the components they belong to are compared and the root of the largest component becomes the root of the new merged component. In a set with k distinct nodes, the length of paths will then be bounded above by $\log_2 k$.

Thanks to point 2 above, when running the UF algorithm to separate the connected components of a set, we get their sizes essentially "for free". Next we will look at the information carried by these sizes for a population of connected components, with respect to the problem of land–atmosphere interaction.

5.2 COMPONENT–SIZE DISTRIBUTION

Consider a two-dimensional cubical complex $\mathcal{C}$ built from thresholding the values of w (vertical wind velocity) in a two-dimensional slice corresponding to a pair (t, z) of time and height coordinates, as described in Section 4.2. The starting point is the empirical observation that the sizes of the connected components that make up the complex $\mathcal{C}$ are not distributed uniformly at random. Instead, they follow a specific pattern characterized by the existence of one large component, which accumulates most of the domain area, and a large number of much smaller components, as is shown in Figure 5.3.

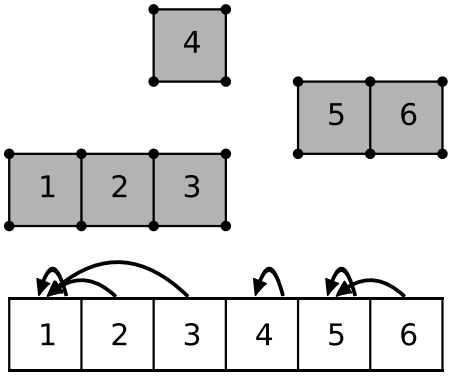


Figure 5.2: An example of the Union-Find algorithm.

This type of size distribution turns out to be a more general phenomenon. More formally, let $\mathcal{C}$ be a two-dimensional cubical complex as described above, and denote its connected components by c_i. Denote the size of a component by j, and the fraction of components with size equal to j by p_j. We can measure component size in grid cells, in which case $j \in \mathbb{Z}^+$, or rescale it to a physical dimension such as m^2 or km^2. Irrespective of the choice made in this regard, p_j as defined here is a discrete variable, as is the underlying grid on which the cubical complexes are built. The distribution of the quantity p_j, expressed here in km^2 is shown in Figure 5.4 for two different slices. In each case all component sizes observed in the 30 timesteps between 13:00 h and 13:30 h are shown, for the 4 LES-ALM simulations. The left panel shows data for the horizontal slice at a height of 44 m (surface layer), the right panel at a height of 1100 m (mixing layer). The general pattern becomes clear in these figures: the smallest sizes accumulate most of the density for p_j. This density then decreases regularly as size increases, until it is several orders of magnitude smaller for the very largest components. This regular behavior is clearer in the data from the surface layer (left panel), and especially in the sizes for simulation SP3 (randomized land surface

Algorithm 1: Separate connected components in an array using `Union-Find`

```
Function ConnectedComponents(A)        /* separate the
 connected components in binary array A */
    n = number of True cells in A
    UF = UnionFind(n)
    foreach cell x ∈ A do
        N = neighbors of cell x
        foreach x_n ∈ N do
            if x_n then
                UF.union(x, x_n)
```

pattern). The distribution appears to decay linearly in a log-log scale. There is a marked difference between the density curve for SP4 (homogeneous land surface pattern) and SP1-3, but the three heterogeneous surface patterns exhibit a mostly similar curve. This serves to illustrate an important difference between the homogeneous and heterogeneous cases, namely a change in the dominant physical mechanism. For SP1-3, heterogeneity in the PBL is maintained as a consequence of forcing over a range of spatial scales, whereas in SP4 structures are produced only as a consequence of turbulence internal to the boundary layer. In other words, the absence of forcing at the largest spatial scales in the case of SP4 becomes manifest here as a scale break at $\sim 0.7\,\mathrm{km}^2$.

For the data from the mixing layer (right panel) the first part of the distribution still appears to decay linearly on the log-log scale, but there are now much larger fluctuations at the tail. We recall from Section 4.3 that the Betti numbers of the positive vertical wind velocity domains, especially β_0^+, have a strong relationship with the underlying land surface pattern, and this is more clearly seen closer to the surface. We now inquire whether this is also the case for the size distribution of connected components. Specifically, before showing how to represent the three-dimensional structure of convection, we will address the following questions:

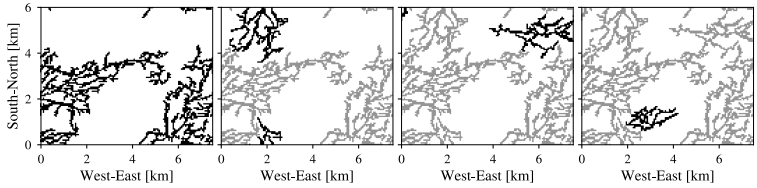


Figure 5.3: The four largest connected components of the updraft domain $\mathcal{M}^+$ in a two-dimensional cross section of LES-ALM simulation SP1, at 13 h, 44 m height. The first panel shows the largest connected component with an area of 7.88 km^2. The next panels show the second, third, and fourth largest components (in black), each having sizes of 1.96 km^2, 1.5 km^2, and 0.82 km^2 respectively. Figure from Licón-Saláiz and Ansorge [2019].

1. Does the empirical probability density of connected component size also carry information about the underlying land surface?
2. If so, how does this depend on time and height?
3. Does it provide a greater discriminating power than the zeroth Betti number for the updraft domains, β_0^+?

To this end, we note that the empirical distributions shown in Figure 5.4 suggest using a heavy-tailed parametric distribution, such as a log-normal or a power-law distribution, to analyze these data.

5.2.1 *Power-law distributions*

A power-law distribution is characterized by a probability density of the form

$$p(x) = Cx^{-\alpha}, \tag{5.1}$$

where α is the scaling parameter and C a normalization constant. A second parameter is the support of $p(x)$, represented by a value x_{min} which gives the lower bound for the power-law scaling. The

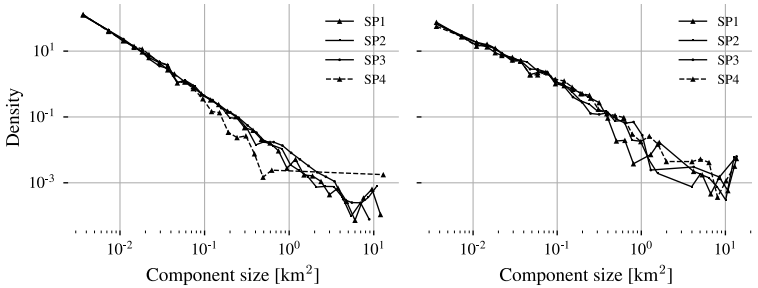


Figure 5.4: Empirical PDF of the variable p_j, the fraction of connected components of size j, in the positive vertical wind velocity domains for a two-dimensional cross section. Both panels show data for the 30 timesteps between 13:00 h and 13:30 h. Left: 44 m, right: 1100 m. Figure from Licón-Saláiz and Ansorge [2019].

exact value of C depends on whether the random variable X in question is discrete or continuous:

$$C = \begin{cases} 1/\zeta(\alpha, x_{\min}) & x \text{ discrete,} \\ (\alpha - 1)x_{\min}^{\alpha-1} & x \text{ continuous.} \end{cases} \tag{5.2}$$

In both cases $x_{\min}$ represents the lower bound for the power-law scaling, and in the discrete case $\zeta(\alpha, x_{\min})$ is the generalized zeta function,

$$\zeta(\alpha, x_{\min}) = \sum_{n=0}^{\infty} (n + x_{\min})^{-\alpha}.$$

Distributions of this form are important in different areas of science. Clauset *et al.* [2009] have proposed a statistical framework based on maximum-likelihood parameter estimation and the Kolmogorov-Smirnov goodness-of-fit test to study empirical data and ascertain whether it conforms to a power-law distribution. For this chapter, we use the Python implementation of these methods contained in the package *powerlaw* [Alstott *et al.*, 2014].

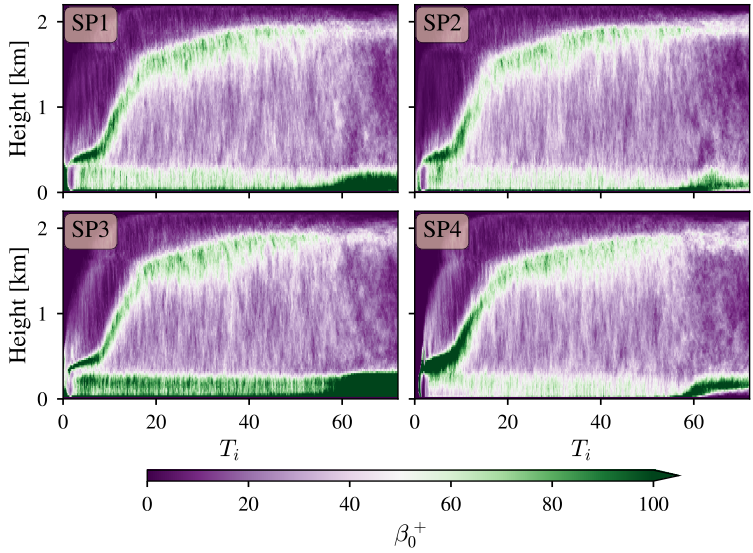


Figure 5.5: Values of β_0^+, four LES-ALM simulations.

5.2.2 *Parameter fitting*

The power-law distribution, Equation 5.1, has the scaling exponent α as its only parameter when its support is known a priori. Fitting a power-law distribution to given data then means to find the value of this parameter which gives the best fit to the data. Doing this with a classical least-squares approach involves fitting a line of the form

$$\log y = b + \alpha \log x,$$

and has been shown to be prone to different kinds of error [see Clauset *et al.*, 2009, Appendix A]. Hence we use maximum likelihood estimation, which is more computationally intensive but free of such systematic sources of error.

Given the data for w in a two-dimensional slice at (t, z) we will then obtain the optimal scaling parameter α that best describes the density of p_j, the fraction of components of a given size. An important consideration to keep in mind before doing this is the

amount of data points available. Clauset *et al.* [2009] give $n = 50$ as the minimum size to obtain reasonable fit results. If we look at the values of β_0^+ computed in Section 4.3, shown here in Figure 5.5, we see that the number of connected components in the cubical complexes for two-dimensional slices is less than 50 in most of the domain, exceeding this number only in some parts of the surface and inversion layers. This means that, by taking data from each slice individually, we would not have a sample size large enough to guarantee reasonable results for most of the domain. We will need to aggregate the data to enlarge the sample size.

Denote the range of simulation timesteps by $t = 0, 1, \ldots, N_t$. We begin by defining a time window t_w, and splitting the time range into the time intervals T_i defined as

$$\begin{aligned} T_1 &= \{0, 1, \ldots, t_w - 1\} \\ T_2 &= \{t_w, t_w + 1, \ldots, 2\,t_w - 1\} \\ &\vdots \\ T_M &= \{(M-1)\,t_w, \ldots, N_t\}. \end{aligned}$$

Given a height value z and a time interval T_i, we will have t_w two-dimensional slices, except perhaps for the last time interval, T_M. For each of these slices we filter the points for which $w \geqslant \theta$, resulting in a set of anchor points for a cubical complex, which we call $\mathcal{C}_\ell$, for $\ell \in T_i$. This cubical complex will be conformed by n_ℓ connected components which we denote by $c_{k,\ell}$, such that

$$\mathcal{C}_\ell = \bigcup_{k=0}^{n_\ell - 1} c_{k,\ell}.$$

Let $s(c_k)$ denote the size of component c_k. We then define $\mathcal{S}_\ell = \{s(c_k) \mid c_k \text{ conn. component of } \mathcal{C}_\ell\}$ as the set of all sizes of connected components that make up $\mathcal{C}_\ell$. Define furthermore the aggregate set

$$\mathcal{S}_{T_i} = \bigcup_{\ell \in T_i} \mathcal{S}_\ell, \tag{5.3}$$

which accumulates the sizes of all connected components observed during the time interval T_i, at height level z. We then fit a

power-law distribution to the values p_j for this set, which represent the fraction of the $n = \sum_{\ell \in T_i} n_\ell$ components with size equal to j. Doing this increases the sample size used to fit the power-law distribution, and further allows the time dependence of the power-law scaling behavior to be observed. In what follows, we use a time window of $t_w = 10\,\mathrm{min}$, and a threshold value of $\theta = 0.01$ for w. The LES-ALM simulations have $N_t = 721$ timesteps, which results in 73 different time periods for the specified time window. This means we will fit 7300 power law distributions for each simulation dataset, each characterized by its scaling parameter α.

The role of $x_{\min}$ in Equation 5.2 must also be made clear. There are two possibilities: either a fixed value is declared *ab initio* and used for fitting all the distributions throughout, or a principled method is used to determine a value of $x_{\min}$ for each distribution, essentially turning it into a second free parameter. This is reasonable because, when dealing with heavy-tailed distributions, random fluctuations at the smallest scales are not as important as the behavior at the tail, and may even display non-power-law behavior. Hence choosing a value of $x_{\min}$ that is strictly larger than the minimal value of x would allow the fitting process to be driven by tail behavior while disregarding the noise at the beginning of the value range. This is the approach we take here. The method for estimating the value of $x_{\min}$, described by Clauset *et al.* [2007], consists of selecting that value of $x_{\min}$ that minimizes the distance between the empirical cumulative distribution function (CDF) of the data and that of the power law model fit from the data. The distance between the two CDFs is measured by computing the Kolmogorov-Smirnov (KS) statistic, defined as

$$D(F, G; x_{\min}) = \max_{x \geqslant x_{\min}} |F(x) - G(x)|, \tag{5.4}$$

where F is the empirical CDF of the data, and G is the power law CDF. The estimated value for $x_{\min}$ is then

$$\hat{x}_{\min} = \arg\min_{x_{\min}} D(F, G; x_{\min}).$$

We now describe the results obtained from applying this methodology to the LES-ALM datasets.

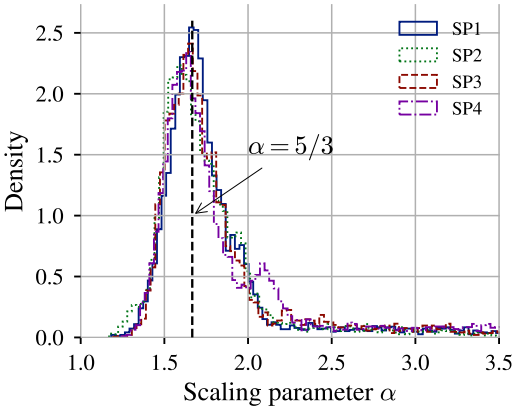


Figure 5.6: Normalized histograms of all power law exponent values. Figure from Licón-Saláiz and Ansorge [2019].

5.2.3 *Scaling parameter*

Table 5.1: Mean value of α for each simulation.

Simulation	Mean α	Stddev. α
SP1	1.79	0.37
SP2	1.76	0.35
SP3	1.82	0.41
SP4	1.85	0.46

The values of the scaling parameter α for all 7300 power law densities fitted for the LES-ALM datasets are shown in Figure 5.6. The distribution of these values is similar across the four simulations, with the mean values of α and the standard deviations shown in Table 5.1. The only difference apparent between these sets of values is that the distribution for SP4 has a more clearly defined bimodal structure. The values of the scaling parameter α appear reasonable, insofar as power law densities encountered

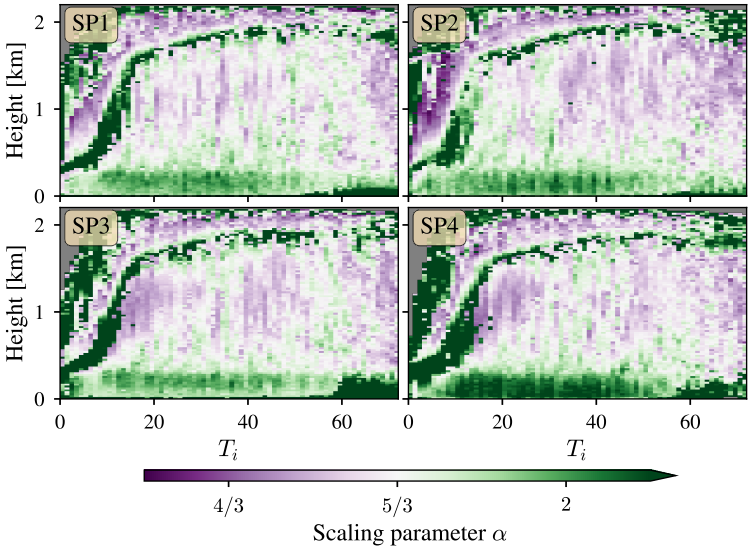


Figure 5.7: Values of the power law exponent for all z, T_i. The color scale, centered at 5/3, highlights the preponderance of this scaling factor within the mixing layer.

in practice tend to have a scaling parameter within the range $1 \leqslant \alpha \leqslant 2$ [Mitzenmacher, 2004]. We further note the fact that the maximum of the PDF in all four cases occurs in a narrow band around $\alpha \approx 5/3$, which is the value expected from the self-similarity of eddies cascading throughout the spectrum of motion (cf. Equation 2.3) [Kolmogorov, 1941b,a; Obukhov, 1941; Kaimal and Finnigan, 1994].

The time-height sections of the scaling parameter α are illustrated in Figure 5.7, as a function height and time. The distribution of α in the (t, z)-plane is clearly non-uniform, and displays a structure reminiscent of that encountered before in the Betti numbers (see Figure 4.15). The qualitative features seen here are:

1. The existence of a well-defined surface layer, encompassing the first 15 height levels closest to the land surface.

2. The evening transition is apparent as a sharp increase in the mean value of α close to the surface, starting around $T_i = 57$. This effect is most pronounced for SP3 and SP4.

3. The inversion layer is characterized by larger α values than those found either in the mixing layer below it, or in the free atmosphere above.

The mean values of α for each of the separate ABL regions is shown in Table 5.2. Here we have used the classification into distinct regions produced by the semi-supervised classification algorithm defined in Section 4.4.3. As can be seen, the variation across the 5 regions is greater than it is across the 4 simulations. It is therefore not possible, by using the information at hand, to state any conclusion regarding the effect of different land surface patterns on the scaling of connected component size.

Table 5.2: Mean value of α for each PBL region.

Region	SP1	SP2	SP3	SP4
Surface layer	1.91	1.93	1.89	2.01
Mixing layer	1.68	1.65	1.65	1.66
Inversion layer	1.89	1.81	1.89	1.97
Free atm.	1.66	1.67	1.72	1.77
Residual layer	1.67	1.64	1.76	1.71

5.2.4 *Goodness-of-fit*

As was noted by Clauset *et al.* [2009], it is difficult to accurately discriminate a power-law distribution from empirical data, due to the possibility of error introduced by the appearance of very large values at the tail of the distribution, which is one of the defining characteristics of a power law. Therefore, an exercise in fitting such distributions to data should be accompanied by a statistical measure of the goodness-of-fit to the data, as well as a comparison with alternative distributions.

The goodness-of-fit of a power law to a given data set can be obtained by the following procedure: let $X = \{x_1, \ldots, x_n\}$ be the data set, and $f(x) = f(x; \hat{\alpha}, \hat{x}_{min})$ the power-law density fitted to it, with corresponding CDF denoted by $F(x)$.

1. Compute the Kolmogorov-Smirnov (KS) statistic for $F(x)$ and $G(x)$, the empirical CDF for data set X.

2. Generate n sets of pseudo-random numbers, S_i, each of size m, distributed according to $f(x; \hat{\alpha}, \hat{x}_{min})$.

3. Fit a separate power-law density, $f'(x; \hat{\alpha}', \hat{x}'_{min})$ to each of the n data sets, and compute the KS statistic for the power-law CDF $F'(x)$ and $G_i(x)$, the empirical CDF for data set S_i.

4. Define the p-value of this test to be the fraction of the n datasets for which its KS value is larger than the KS value obtained for the original data set X and its power-law density.

This p-value represents the proportion of artificial datasets which have a poorer agreement with their respective power-law fits than does the original data set. If this proportion is "small enough", meaning that nearly all the random datasets display better agreement with their respective fitted distributions than does the original data set, we can rule out the power-law hypothesis. Following Clauset *et al.* [2009] we take "small enough" to be $p < 0.1$. According to Clauset *et al.*, the number n of data sets needed is a function of the desired precision. In particular, generating at least $n = 2500$ datasets gives a p-value that is accurate up to 2 decimal places. This is the criterion we employ here.

Such a test allows us to conclude whether there is enough evidence to reject the original hypothesis of power-law scaling in the data. However, it does not say anything about whether the hypothesis need be true. In order to get a clearer picture, it is useful to compare the power-law fit with an alternative heavy-tailed distribution, of which the most common are the exponential and the lognormal distributions. Even if the goodness-of-fit test gives no indication that we need reject the power-law hypothesis,

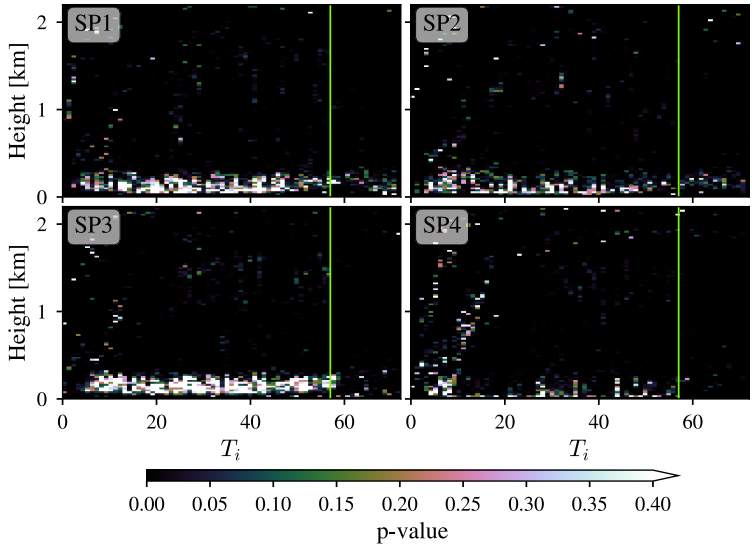


Figure 5.8: P-values for the Kolmogorov-Smirnov statistic. The green line at $T_i = 57$ marks the start of the evening transition.

it could well be the case that the alternative distribution gives a better fit to the data. One way to carry out this comparison is to evaluate the likelihood $\mathcal{L}_0 = \mathcal{L}_0(X)$ of the data under the power-law hypothesis, as well as the likelihood under the alternative hypothesis, $\mathcal{L}_1 = \mathcal{L}_1(X)$, and compute the log-ratio:

$$R = \log\left(\frac{\mathcal{L}_0}{\mathcal{L}_1}\right). \tag{5.5}$$

If $R > 0$, the power-law hypothesis is then the more likely one. To account for possible random fluctuations affecting the value of R, its standard deviation can be estimated from the available data to yield a p-value for the likelihood ratio test [Vuong, 1989]. In this case, "small" p-values would indicate that the observed value is unlikely to be the result of random fluctuations alone. If the p-value is larger than a certain threshold, we consider the data to be insufficient for this test.

We first take a look at the p-values for the KS goodness-of-fit test as a function of (z, T_i), as shown in Figure 5.8. These values indicate that the original power-law hypothesis does not hold for most of the domain, exception made of the surface layer. Outside this region the p-values are well below 0.1 for all but very few data points. Power-law scaling behavior, if it exists, is therefore concentrated in the surface layer. The four LES-ALM simulations also show a different behavior in this regard. The p-value for the test on data within the surface layer surpasses 0.1 more often for SP3, the randomized land surface pattern, than for the other three simulations. Conversely, it is for the SP4 surface layer that this happens least often. Overall this is in agreement with the data shown in Figure 5.4, where the scaling for the data from surface layer in case SP3 is the most regular of all four simulations. The KS p-values show this to be the case in general, if we restrict ourselves to the surface layer. This scaling behavior becomes less frequent, or vanishes altogether in the case of SP4, after the evening transition takes place ($T_i = 57$).

Evaluating the log-likelihood ratio R (Equation 5.5) for a power law and exponential distribution shows that the exponential distribution is not favored by the data, in the vast majority of the analysis domain (cf. Figure 5.9). Indeed, the p-values for the corresponding log-likelihood test (not shown here) are smaller than 0.01 for 79% of the (z, T_i) pairs from SP1, 82% for SP2, 78% for SP3, and 76% for SP4. Moreover, the points where this happens agree with those where the log-likelihood ratio is positive and greatest in magnitude (i. e., the solid red area in Figure 5.9, top).

The log-likelihood test for a power law against a log-normal distribution paints a more nuanced picture (Figure 5.9, bottom). Here we find only a small subset of (z, T_i) points for which the power-law hypothesis has greater likelihood, most of which are concentrated in the surface layer. The magnitude of the ratio itself is also lower than it was for the power law vs. exponential test. Thus, it becomes necessary to separate those instances where the test is actually conclusive based on the resulting p-value, which we show in Figure 5.10. Here we see that, at a significance level of 0.05, the test is actually not able to distinguish between either of the two hypothesis for the vast majority of the domain. More

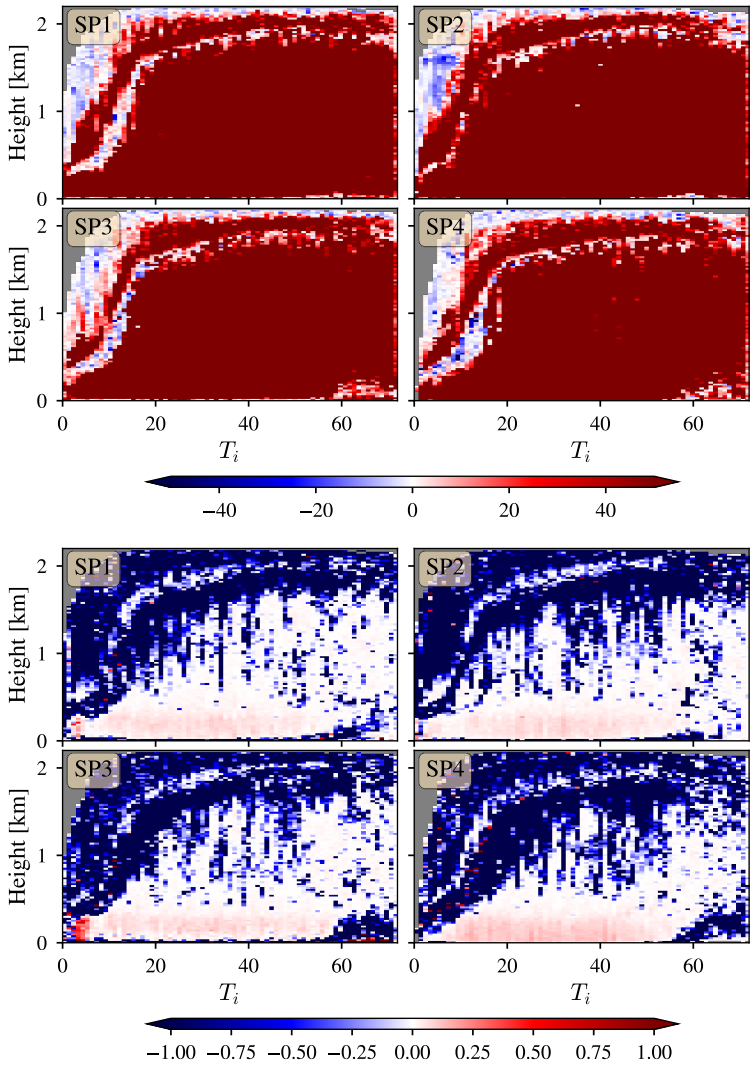


Figure 5.9: Distribution of log-likelihood ratios for the comparison of a power-law distribution with an exponential (top) and lognormal (bottom) distributions.

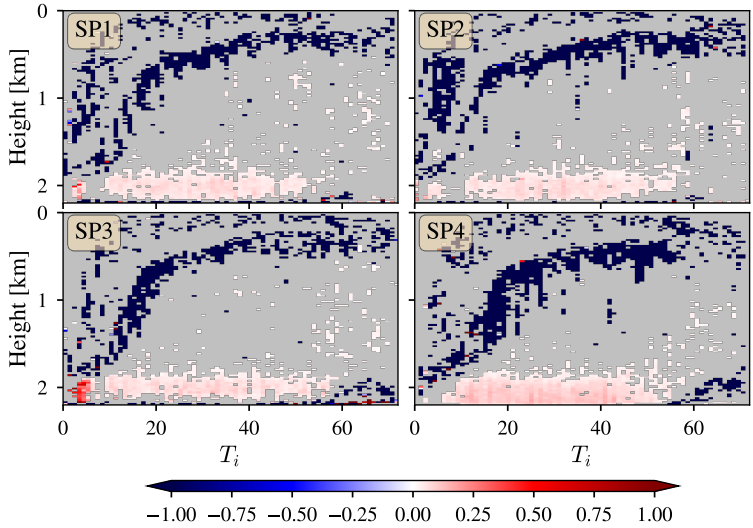


Figure 5.10: Value of the log-likelihood ratio for the power law vs. lognormal test, where only the datasets for which the test p-value is smaller than 0.05 are shown (i. e., gray represents a non-significant result).

important for our current analysis is that most of the surface layer points, for which the power-law hypothesis has the greater likelihood, fall within the statistically significant tests (at the 0.05 level). We take this as further evidence supporting the original hypothesis of power-law scaling in the distribution of sizes for connected updrafts. We summarize the results of this subsection so far:

1. The original observation made at the beginning, that the logarithm of frequency at which connected updraft regions appear in two-dimensional slices of the LES-ALM simulations decreases linearly with the logarithm of their size, finds limited evidential support across the totality of the simulation data.

2. There is a strong indication of power-law scaling for the size of connected updraft components within the surface

layer, and the degree to which a power-law distribution is a good fit for the data appears to be in direct relation to the heterogeneity of the underlying land surface. In all four simulations, the scaling parameter α for the surface layer data is, on average, in the range $1.9 \leqslant \alpha \leqslant 2.0$.

3. Comparison of the power-law distribution with the alternative of an exponential or a log-normal distribution via a log-likelihood ratio test shows that, for the surface layer, the power law is in general the most adequate to describe the distribution of connected component size.

4. The evening transition brings an end to the power-law scaling behavior in the surface layer. After this point in time, the power-law distribution appears only in one vertical level adjacent to the land surface for SP3.

It is not yet apparent, beyond the observation made in point 2 above, to which extent the information provided by the power-law scaling parameters can be used to separate the different land surface patterns. We address this question in the next subsection.

5.2.5 *Comparison with the updraft Betti number, β_0^+*

In order to determine whether or not knowing the values of the power-law scaling parameter α is sufficient to distinguish the four land-surface patterns, we will follow an approach analogous to that used in Section 4.3.2. That is, we will build a feature matrix with the values of α obtained for each of the vertical levels in the computational domain, with each level being a different feature vector. Each observation is the set of all features at time interval T_i. In light of the discussion at the end of Section 5.2.4, we will restrict the data to the surface layer, the 15 vertical levels closest to the land surface, and to the CBL regime, time intervals $T_i = 1, \ldots, 60$. This results in the two feature matrices X_α for the power-law exponent data and X_β for the zeroth Betti number, each of dimensions 240×15.

We will use these data to train a classification model, with the response variable y being the label SP1, ..., SP4. As a reference,

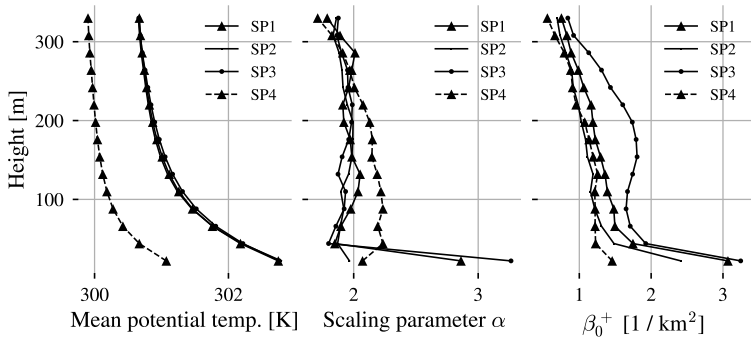


Figure 5.11: Vertical profiles of mean potential temperature T (left), power-law scaling parameter α (center), and zeroth Betti number β_0^+ (right) for the four LES-ALM simulations. The values of α correspond to time interval $T_i = 40$. For β_0^+ and T, the 10 corresponding one-minute timesteps are averaged. Figure from Licón-Saláiz and Ansorge [2019].

we will also compare the performance of this classifier to one trained on the values of β_0^+, which have already been shown to be a powerful descriptor for this classification task, and to one trained on mean potential temperature, averaged on horizontal slabs. In order to have the same number of observations in all models, we will use for $\beta_0^+(z, T_i)$ the average of all $\beta_0^+(z, t)$ values for $t \in T_i$, and analogously for the temperature data. Figure 5.11 shows a comparison of these three sets of features. The middle panel shows the values of α, and we note that, as was observed in the first example of the size distributions shown in this section (Figure 5.4), the scaling (i. e., the slope of the line in log-log scale) is different for SP4 than for the three heterogeneous surface patterns, but the difference in scaling between the three heterogeneous patterns is not evident. This appears to be the case throughout the surface layer. For the mean potential temperature, the similarity between the three heterogeneous patterns is even greater. However, as shown in the previous chapter, the values of β_0^+ are in a direct relationship to the heterogeneity scale of each land surface pattern,

Table 5.3: Performance of k-NN classifiers, $k = 3$, trained with feature matrices using the power law exponent α (left) and the zeroth Betti number β_0^+ (center). The temperature data (right) was used to train a k-NN classifier with $k = 7$.

	α			β_0^+			θ		
	Precision	Recall	F_1	Precision	Recall	F_1	Precision	Recall	F_1
SP1	0.42	0.45	0.43	0.72	0.74	0.72	0.18	0.22	0.19
SP2	0.53	0.52	0.52	0.63	0.68	0.65	0.18	0.19	0.18
SP3	0.60	0.61	0.60	0.97	0.96	0.96	0.16	0.14	0.14
SP4	0.83	0.77	0.80	0.89	0.81	0.85	0.88	0.80	0.83
avg.	0.60	0.58	0.59	0.81	0.79	0.80	0.35	0.33	0.34

especially close to the surface, which is reflected in a clearer separation of the four simulations.

The classification model used for both datasets is a k-NN classifier. The weighted average F_1 score for each is estimated by bootstrapping, for k neighbors in each model, $k = 2, \ldots, 15$. In both cases the best performance was achieved by the classifier with $k = 3$, as shown in Table 5.3 in terms of precision, recall, and F_1 score. The difference between the two models is significant, with an average F_1 score for X_α of 0.59, compared to 0.80 for X_β. This shows that the scaling law found to describe the size of connected updraft components is sensitive to the differences in land surface patterns. It is not, however, more sensitive to these differences than the *number* of connected components, β_0^+. In other words, when trying to determine which surface pattern produced a given set of surface-layer values, knowing how many connected components are present in the two dimensional slices is more informative than knowing how their sizes scale.

The classification metrics shown in the table also show that both sets of features have greatest discriminatory power when it comes to classifying the SP3 and SP4 data, with the power law model having greatest F_1 score for SP4, and the Betti number model for SP3. Both facts are consistent with the differences in power-law scaling across the four simulations, to wit: that the power law densities fit to size data can distinguish between the homogeneous land surface pattern SP4 and the three heterogeneous patterns,

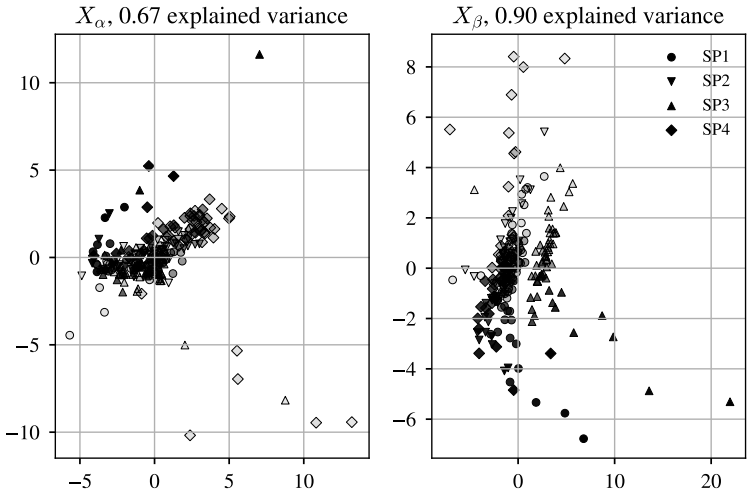


Figure 5.12: First two principal components of the surface layer values of the power law scaling parameter α (left) and the zeroth Betti number β_0^+ (right). Each point represents a row of the X_α, X_β matrices, projected onto the two dimensional subspace generated by the first two principal components. The time period T_i of each point is indicated by its level of transparency, with darker points corresponding to later time periods.

but do not distinguish between the latter three. They are also consistent with the observation made in the previous chapter that, throughout most of the day, the value of β_0^+ for the surface layer of simulation SP3 is on average higher than it is for any of the other three simulations (see Figure 4.11). An interpretation of these facts would be that land surface heterogeneity introduces an element of scale invariance to the size distribution of connected updraft regions, with this being clearest for the maximally heterogeneous SP3. Conversely, a purely uniform land surface, SP4, breaks away from this behavior, thus making its classification easier when only the scaling parameter α is known.

Another way to illustrate this is shown in Figure 5.12. Here the 240 observations of the 15 features contained in X_α and X_β

are projected onto the space generated by the first two principal components of each dataset. In the case of X_α (left panel) there is no clear separation between the four simulations. For the X_β, on the other hand, we see a clear separation between the SP_3 points and those of the remaining three simulations. The temporal evolution of the projected Betti numbers, shown by the level of transparency of each point in the figure, also appears regular throughout the four simulations. This is not the case for the projected power-law exponent data.

5.3 MERGE TREE REPRESENTATION

The core idea of a dominant connected component accumulating most of the updraft volume carries over when we pass from the two-dimensional slices which have been analyzed so far to the scalar field spanning the three spatial dimensions. In this section we will introduce another topological invariant, the *merge tree*, which will give us a representation of said dominant structure in terms of its connectivity along the vertical direction.

This invariant serves a double purpose: first, it gives us a simplified representation of the geometric structure of three-dimensional convective flow, which can be linked directly to the land surface. Second, it allows us to record and quantify the spatial coalescence of convective plumes. This is an important feature of free convection, which takes place in what has been called the *plume-merging layer* (PML) [Mellado *et al.*, 2016].

5.3.1 *Height function*

Recall from Section 3.2 the definition of the merge tree of a function: if $f : X \to \mathbb{R}$ is a smooth function defined on a manifold X, and we define the equivalence relation $\sim$ on X by $x \sim y$ if $f(x) = f(y)$ and both x and y belong to the same connected component of the sublevel set $f^{-1}(-\infty, f(x)]$. The merge tree of f is the quotient space $X/_\sim$.

It is not immediately clear from the definition how this concept can be applied to numerical data, such as that produced

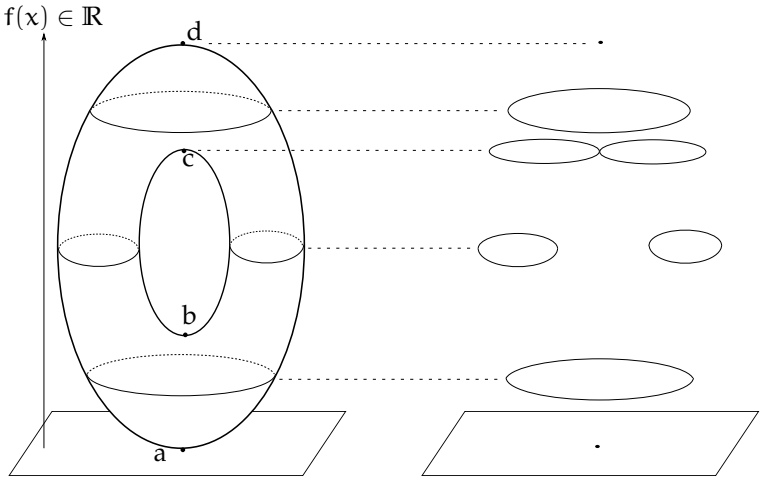


Figure 5.13: Height function for an upright torus (left), and its topologically distinct level sets (right). Figure inspired by Edelsbrunner and Harer [2010].

by an LES simulation. For this we need to introduce the idea of a *height function*. Consider for example the manifold $\mathcal{M}$ shown in Figure 5.13, an upright torus, depicted as resting on a plane. Define the function $f : \mathcal{M} \to \mathbb{R}$ by $f(x) = z$, where z is the height of point $x \in \mathcal{M}$ above the plane on which the torus rests. This is an example of a height function. The right part of the figure shows the level sets of f,

$$f^{-1}(r) = \{x \in \mathcal{M} \mid f(x) = r\}.$$

A key fact about this function is that, if we consider all its possible values, then the topology of its level sets only changes at a finite set of points, labeled a, b, c, d on the torus. The possible cases are:

1. For $r > f(d)$ or $r < f(a)$, the level set is empty.
2. If $r = f(a)$, the level set is a single point: $f^{-1}(a) = \{a\}$, and analogously for $r = f(d)$.

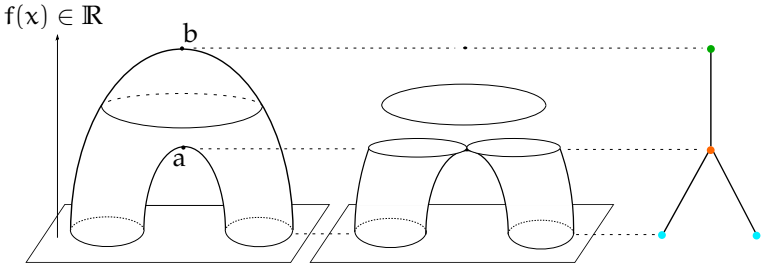


Figure 5.14: Capped torus resting on a plane (left), its topologically distinct level sets (middle), and the merge tree for its height function (right).

3. For $f(a) < r < f(b)$, and $f(b) < r < f(c)$, the level set becomes a circle.
4. Finally, if $f(b) \leqslant r \leqslant f(c)$, the level set is formed by two circles.

If we look at the evolution of the sublevel sets of f instead,

$$f^{-1}(-\infty, r] = \{x \in \mathcal{M} \mid f(x) \leqslant r\},$$

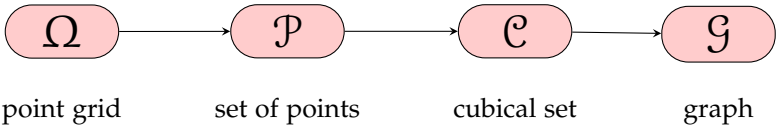
it is also true that their topology changes only when passing through the four points $a, \ldots, d$, which are called *critical points*. Since there is only one connected component in all sublevel sets throughout, however, the resulting merge tree is not very interesting to look at. Consider therefore the capped torus standing on a plane shown in Figure 5.14. In this case the sublevel sets have two connected components until we reach point a, where both components merge together, giving one connected component for all $r > f(a)$. The right part of the figure shows the merge tree for the height function of this capped torus. It has three kinds of nodes, according to the changes that occur in them: *birth* nodes, shown in blue, one *merge* node, in orange, and one *final* node, in green. Each of the edges of the graph is an element in the quotient space $X/\sim$ defined above, that is, a connected component in the corresponding sublevel set.

5.3.2 *Methodology*

Having seen how the merge tree for a height function can represent the evolution of sublevel sets of a given manifold, we now show how we can use it for the task of representing the geometric structure of convection. We start as before with the computational domain Ω, and consider now the three-dimensional subsets with a constant time coordinate, $\Omega_t = \{1, 2, \ldots, N_x\} \times \{1, 2, \ldots, N_y\} \times \{1, 2, \ldots, N_z\} \times \{t\}$. We will proceed as before and take the subset of points $\mathcal{P} \subset \Omega_t$ such that w, the vertical wind velocity, is greater than a threshold value at those points: $\mathcal{P} = \{(x, y, z) \in \Omega_t \mid w(x, y, z) \geqslant \theta\}$. These will be the anchor points for a three-dimensional cubical complex $\mathcal{C} \subset \mathbb{R}^3$ which represents the volume of space covered by updrafts at time t. If we think of this as representing a geometric object resting on the plane $z = 0$, we can define a height function on this set by

$$f(x, y, z) = z.$$

That is, for any point $(x, y, z) \in \mathcal{C}$, f returns its third component, which is its height above the ground. The level sets of this function are then the binary two-dimensional cross sections of the vertical velocity field which were studied in depth in the previous chapter. The merge tree of this function will be a graph which encodes the evolution of the sublevel sets of this geometric object $\mathcal{C}$ along the vertical direction. In particular, it will record the merge events which bring together two or more components at a given height. The general methodology then has the following steps:



The main distinction between this case and the one treated in Chapter 4 is the fact that the topological invariant used here is not the rank of a group or a vector space, but rather a graph. Although

developing a full statistical methodology for analyzing such topological invariants lies beyond the scope of this dissertation, we will still show how they can be used to yield information about the physical problem at hand, namely the interaction between the land surface and atmospheric convection.

The construction of the merge tree is carried out by Algorithm 2, which operates on a binary three-dimensional array such as the set of points $\mathcal{P}$ described above. This algorithm works as a nested UF data structure on the anchor points of $\mathcal{C}$: for each of the level sets of the height function f, denoted here by $\mathcal{C}_{t,z}$, a "local" instance of UF, UF_z, is created for that level, to store the connected components of that level set. Then, for $z \geqslant 2$, the connectivity of the components of $\mathcal{C}_{t,z}$ with those of $\mathcal{C}_{t,z-1}$ is stored. For each component $c_z^i \in \mathcal{C}_{T,z-1}$ one of the following will be true:

1. It still exists as an independent component in level z.
2. It no longer appears in level z as an independent component because it has been merged into another component at level z. This is the situation represented by the orange node in Figure 5.14 (right).
3. It no longer appears at all in level z, since it has stopped growing beyond level $z-1$. This is represented by the green node in Figure 5.14.

These events will be stored as nodes in the resulting graph $\mathcal{G}$. Event 2 will be a *merge* node, while event 3 will be a *terminal* node. We will also store *birth* nodes to represent the appearance of a new connected component when passing from level $z-1$ to z. A key point in our implementation is the selection of the new parent node after a merge event takes place. The usual criterion used in the UF algorithm is the so-called *elder rule*, where the oldest node will always be chosen as parent from all potential candidates. Since the ordering of the nodes depends on the direction the array is traversed, this could result in different tree structures for different directions. We hence choose a more physically meaningful rule, and declare as parent the node associated to the component with largest value of the variable being modeled (w in this case, so we choose the component with largest total flux).

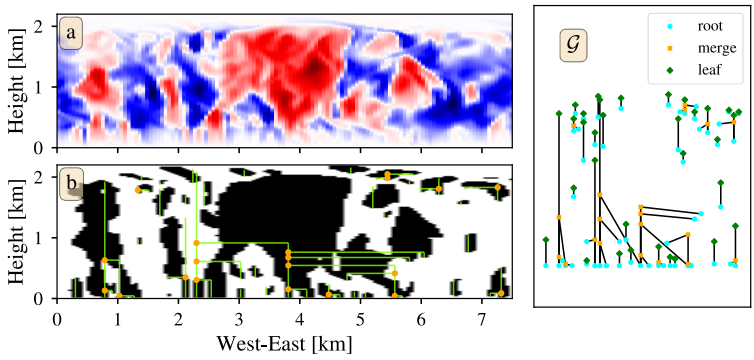


Figure 5.15: Example of applying the merge tree algorithm on a vertical two-dimensional slice. **(a)** shows the original values of vertical wind velocity w; **(b)** the cubical complex obtained by thresholding the data, together with its merge tree; $\mathcal{G}$ shows the graph representation of the tree. Figure from Licón-Saláiz and Ansorge [2019].

An example of this process is shown in Figure 5.15, where the algorithm has been run on a two-dimensional (vertical) cross section for clarity. The scalar field containing the values of w shown in panel (a) is first converted to a cubical set, as shown in (b), where the black region represents the area with $w > 0$. The vertical green lines show the lifespan of connected components of the sublevel sets $f^{-1}(-\infty, z]$ as z goes through the range $1, 2, \ldots, N_z$. The horizontal lines link these components to merge nodes, shown in orange, at the vertical levels where such mergers occur. The complete graph representation of the merge tree for this cubical set is shown as $\mathcal{G}$, with birth, merge, and terminal nodes.

5.3.3 *Results*

The first step is to inspect the number of merge nodes in the graphs produced for each timestep in the four LES-ALM simulations. These values are shown as a time series in Figure 5.16. It can be seen that the four time series exhibit a similar qualitative pattern: a brief initial transient, followed by a period of

Algorithm 2: Build a merge tree from a binary three-dimensional array

Function BUILDMERGETREE(A) `/* function to build a merge tree from a binary array */`

- Create a Union-Find data structure, UF
- **foreach** level z **do**
  - Create a hash map B of below joins, $(b : [p])$
  - **for (** $x \in z$ **)** {
    - **if** x **then**
      - UF.Find(x)
      - **if** x connected behind **then** UF.Union(node behind)
      - **if** x connected side **then** UF.Union(node side)
      - **if** x connected below **then**
        - Find the parent of x: x_p = UF.Find(x)
        - Find the parent of cell below: b_p = UF.Find(below)
        - Add entry b_p to $B[x_p]$
  - Level is finished
  - Perform all the Union operations for elements in B
  - If len$B[x_p] > 1$, select the parent with maximum value
  - Any new component at level z not joined to a parent below: add a *birth node*
  - Any component from level $z-1$ not present at z: add either a *merge node* or *terminal node*

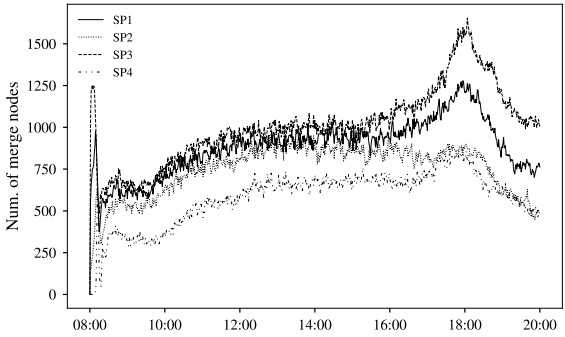


Figure 5.16: Number of merge nodes at each simulation timestep. The four LES-ALM simulations are shown. Figure from Licón-Saláiz and Ansorge [2019].

growth until noon, after which the value stabilizes. There is a second growth phase in the late afternoon, which intensifies at the evening transition, with a marked decline afterwards. Also noteworthy is the fact that the average values of the four series are different, and stand in an inverse relationship with land surface pattern heterogeneity. That is, SP3 has on average the largest number of merge nodes, with SP4 having the smallest.

A closer view of the surface layer, where most of the merge nodes are concentrated, is given by Figure 5.17. The clearest difference is again that between SP3 and SP4, with the former having most of its surface layer merge nodes distributed throughout the first 5 vertical levels. In the case of SP4, the merge nodes are concentrated in the first level adjacent to the land surface. Recall the size distribution shown in Figure 5.4, and the discussion surrounding the KS goodness-of-fit test for these size distributions. It was shown that the surface layer size distributions in SP4 have the worst power-law fit out of all four simulations. This fact is expressed in the empirical PDF (Figure 5.4, left) for SP4 having a scale break followed by a gap of one order of magnitude, reflecting the presence of the dominant component. What the distribution of merge nodes suggests is that, in SP4, this dominant component appears already at the first level, whereas for simulations

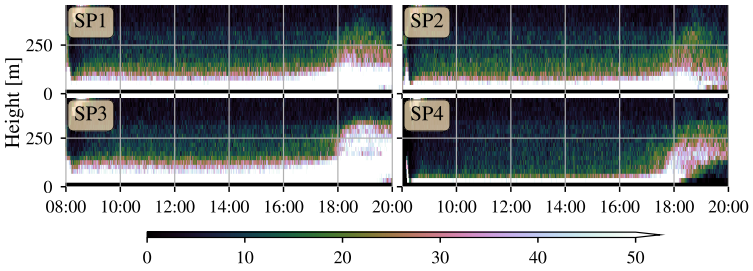


Figure 5.17: Number of merge nodes for each vertical level within the surface layer, for the four LES-ALM simulations. Figure from Licón-Saláiz and Ansorge [2019].

with increasing levels of land surface heterogeneity the process of component merging occupies a larger portion of the vertical direction.

This merging process results in the formation of a large connected structure which accumulates most of the volume in the updraft domain, close to 99%. This also happens very rapidly, within the first 30 simulation minutes in all four cases, as shown in Figure 5.18, and once the structure forms it persists for the remainder of the simulation. As a further point of interest, we inspect the merge tree root nodes at surface level, for these four simulations. Recall that each root node in a merge tree signals the appearance of a new equivalence class in the quotient space $X/_{\sim}$, i. e.the appearance of a new connected component in a sublevel set $f^{-1}(-\infty, z]$ for $z \in \mathbb{R}$. A root node at surface level, therefore, simply represents a connected component in the first sublevel set, $f^{-1}(-\infty, 1] = f^{-1}(1)$.

An advantage of this explicit geometrical representation is that it enables us to directly connect flow structures, in this case convective plumes, to other features in physical space, as there is no need to compute a spectral transform of the data. Thus, we can identify those points in the land surface to which the convective plume in connected. This is significant, as each cell in the land surface is associated with an energy flux into the atmosphere aloft. We can then assume that a land cell connected to the plume

is directly contributing energy to it. We can define the relative contribution of a given land surface type i as

$$p_i = \frac{\#\{\text{cells of type } i \text{ connected to plume}\}}{\#\{\text{cells of type } i\}}.$$

Since the total number of cells per land surface type are the same in the three heterogeneous LES simulations, we can directly compare the proportions p_i for these three simulations, and take them as a measurement of the effect of land surface heterogeneity on the effectiveness of a given land type in supporting convection. Figure 5.19 shows the 5-minute moving averages of the p_i for all land surface types throughout each of the three heterogeneous simulations. For reference, the 5-minute average of p_i for SP4 is also shown, which in this case is simply the percentage of the total land surface (grassland) that is connected to the dominant plume. There are appreciable differences in the values of p_i for each simulation, with the most notable being the average increase of p_{urban} as land surface heterogeneity increases. In SP2, urban land cells actually have the smallest relative contribution to the convective plume. In SP1 they are already the land type with the second-highest contribution, and here a diurnal evolution of p_{urban} also becomes clear: it gradually increases towards its maximum around noon, after which a gradual decrease ensues. In SP3, both urban and forest cells have the greatest relative contribution to the plume.

A second significant difference is the presence, in SP3, of a sharp upturn in p_{forest} in the early evening. To a much lesser extent, a similar increase in p_{forest} is also present in SP1, and it is entirely absent from SP2. This contrasts sharply with the case of SP4, for which the land surface contribution to the dominant plume actually vanishes before the simulation terminates, which represents a much more rapid decoupling of the nascent stable surface layer from the layer of residual turbulence (cf. Figure 5.16). This goes to show how differences in land surface geometry can modulate the effectiveness of diverse land types in the initiation and sustenance of atmospheric convection.

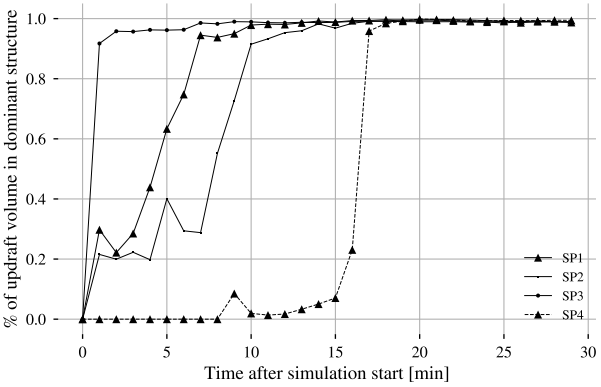
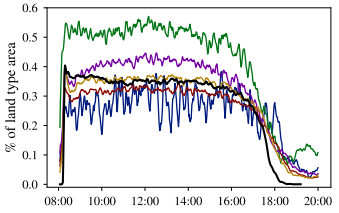
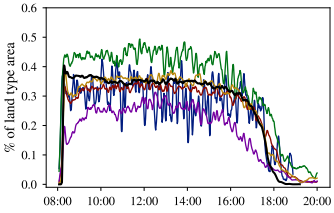


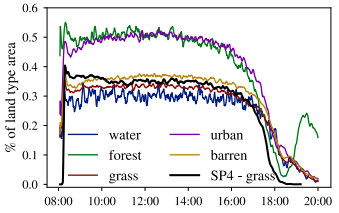
Figure 5.18: Formation time of dominant connected component.



(a) SP1



(b) SP2



(c) SP3

Figure 5.19: Proportion of each land surface type connected to the dominant structure at each timestep. SP4 (uniform grassland) is shown on each panel for comparison.

6

SPATIAL DISTRIBUTION OF SHALLOW CUMULUS CLOUDS

Outline

This chapter introduces the use of persistent homology (PH) to analyze spatial patterns. Applications of PH to various concrete problems are reviewed in Section 6.1. We give the definition of the stable rank invariant, which generalizes PH, in Section 6.2. Section 6.3 illustrates the use of this invariant as a spatial statistic for regular point patterns. Section 6.4 discusses how this can be used to analyze the spatial distribution of shallow cumuli, and concrete results concerning cloud cover and the cloud size distribution are given in Section 6.5. Section 6.6 introduces a PH-based index for spatial organization, and Section 6.7 introduces the persistence contour formalism, and describes its use in obtaining a morphological classification of cloud fields. Parts of this chapter have appeared in Licón-Saláiz *et al.* [2018] and Riihimäki and Licón-Saláiz [2019].

After having dedicated Chapters 4 and 5 to the analysis of geometrical properties of atmospheric flow, we will now turn our attention to a direct manifestation of this flow, namely the formation of boundary layer clouds. More specifically, we will focus on describing the spatial distribution of shallow cumulus clouds. To this end we will introduce a new topological invariant, persistent homology, which has its roots in manifold reconstruction but has recently seen use in the study of spatial structure and distribution of diverse kinds of objects. We will use these techniques to quan-

tify and analyze properties of the spatial distribution of clouds, which is a subject of ongoing research in atmospheric science due to the fact that shallow cumuli are a significant source of uncertainty in global climate models [Bony and Dufresne, 2005]. In order to be able to parameterize these clouds correctly it is especially important to understand aspects of their spatial distribution such as how they cluster together, and what kind of characteristic spatial patterns they tend to form.

6.1 RELATED WORK

We distinguish two categories of related work: the use of PH to study the spatial properties or configurations of diverse objects, and the development of functional summaries of PH to be used as statistical descriptors.

The notion of PH itself is tied from its origins to the analysis of spatial data. The seminal paper by Edelsbrunner *et al.* [2002], where the persistence algorithm is first described, also presents its application to the extraction of topological features from molecular dynamics simulations of several structures. Central among these examples is that of the protein gramicidin-A, which has a high bactericidal potential due to its effect on bacterial cell membranes, increasing their permeability [Kelkar and Chattopadhyay, 2007]. This is possible because the protein has a physical hole or tunnel going through it, or in topological terms, a generator of 1-dimensional homology. The input to the algorithm is the time-averaged molecular dynamics simulation of a gramicidin-A molecule, which gives the locations of its atoms as points in space. A metric filtration is computed for this point set, and its corresponding persistent Betti numbers are computed. These are shown to correctly represent the existence of this tunnel in the molecule. Other features are discarded as extraneous topological noise introduced by the filtration. This idea would determine the program for the nascent field of topological data analysis (TDA) in its first years: given a set of data points embedded in some metric space, construct a filtered complex on them, and compute its persistent Betti numbers. These numbers allow us to separate

"real" features in the data from noise, if we assume that the former will have large persistence values, whereas the latter will not.

Another now-famous example is provided by Carlsson [2009], who focused on the analysis of digital images. Specifically, if we think of an image with n pixels as a vector in $\mathbb{R}^n$, we can ask whether a given set of images can be modeled as a manifold embedded in $\mathbb{R}^n$. The dataset in this case consists of a set of black-and-white photographs of the Dutch landscape taken in and around Groningen [van Hateren and van der Schaaf, 1998]. From each image, patches of 3×3 pixels are extracted and represented as vectors in $\mathbb{R}^9$, with each entry of the vector being the greyscale value in that pixel. 5000 such patches are chosen at random from each image, and the 20% with the largest spread in greyscale values out of the entire database are selected. These vectors are then mean-centered and normalized to unit length, resulting in 450 000 points lying on a 7-dimensional ellipsoid within $\mathbb{R}^8$. Analysis of the persistent Betti numbers obtained from this point set show it to be a subspace of the Klein bottle. This is interpreted as a consequence of the predominant spatial structures in the high-contrast regions of natural images: objects tend to be mostly aligned vertically or horizontally, with the less frequent diagonal positions appearing along a continuous spectrum. This can indeed be modeled as a non-orientable 2-manifold (a Klein bottle). This example would become a talking point within the TDA community, as it is not immediately clear what can be done with the knowledge that a given dataset is somehow part of a Klein bottle embedded in a high-dimensional space. This general difficulty in translating a topological fact into more concrete properties of the data under analysis would be the cause for the initially slow adoption of TDA methods in more applied areas of science.

A significant drawback of PH in a data analysis context is that its output, be it in barcode or persistence diagram form (see Section 3.6), is a multiset of intervals. It is possible to perform statistical analysis on these objects, but due to their nature as arbitrary collections of intervals or pairs of real numbers, common operations on them can be difficult, as for example the mean of a set of persistence diagrams might not be unique [Turner *et al.*,

2014]. A possible workaround is to either use a transformed version of the persistence diagram or barcode, or to look at their aggregate properties. For example, MacPherson and Schweinhart [2012] define the following transformation

$$x([b,d)) = \frac{d+b}{2}$$
$$y([b,d)) = \operatorname{arcsec}\left(\frac{d}{b}\right),$$

where $[b,d)$ is a persistence interval (Definition 3.6.4). They use the density $f(x,y)$ of points on the x,y-plane, what they call the *PH-density function*, to define the PH-dimension and the PH-self-similarity of an object. These measures are found to agree with the Hausdorff dimension of branched polymers and self-avoiding random walks, while they differ for Brownian trees. The significance of this lies in the fact that ordinary measures of shape for stochastic fractal structures are rare (these structures are not even differentiable), but PH density is readily computable in these cases. Moreover, this also gives an example where the main goal is no longer the discovery of isolated topological features. Indeed, this pursuit would be futile as these structures are all topologically trivial. The entirety of the PH information is used as a descriptive feature instead, and it is shown to detect properties of these objects not attainable by standard methods.

Another advance in this direction is the *persistence landscape* [Bubenik, 2015], which is defined as a rescaled version of the rank function

$$\lambda(b,d) = \begin{cases} \beta^{b,d} & b \leqslant d, \\ 0 & \text{otherwise.} \end{cases} \tag{6.1}$$

By casting these functions as random variables with values in a Banach space, it is possible to perform statistical inference on PH data. A similar approach is pursued by Robins and Turner [2016], where they adapt a null hypothesis testing scheme from spatial statistics to be applied on the rank functions (Equation 6.1). They show how this can then be used in extracting physical information from colloid point patterns, and in testing the hypothesis of complete spatial randomness in spatial point processes. Again, this

gives an example of the totality of the PH information being used as a descriptive feature, now in the form of the rank function.

A different kind of summary are *persistence images* [Adams *et al.*, 2017]. Given a multiset of persistence intervals

$$B = \{[b_i, d_i) \mid i = 1, \ldots, n\},$$

first transform the birth-death pairs to birth-persistence pairs by the linear map $T(b, d) = (b, d - b)$. These are used to define a scalar function $\rho_B : \mathbb{R}^2 \to \mathbb{R}$,

$$\rho_B(z) = \sum_{u \in T(B)} f(u)\phi_u(z),$$

with $f(u)$ being a weight function, and $\phi_u \sim N_2(u, \sigma^2)$ a bivariate normal density centered on u. The domain of ρ_B is discretized, and from this an array containing the integral of ρ_B on each of the discrete domain segments is obtained. This is used as a vector representation of the PH decomposition. On a similar vein, Reininghaus *et al.* [2015] use a multiscale kernel function, motivated by the fundamental solution to the diffusion equation, to discretize persistence diagrams. Kusano *et al.* [2018] propose the persistence-weighted Gaussian kernel (PWGK) function, which relies on expressing the persistence diagram decomposition of a space as a weighted measure. These approaches provide a vector representation of PH, as opposed to the functional representations described above, and are as such designed to interact with kernel-based machine learning methods.

On the applications side, Dłotko and Wanner [2016] use persistence landscapes to recover physical information from a system undergoing Cahn-Hilliard–type phase separation. They show that this topological information alone is sufficient to determine total mass in the system, as well as the precise moment during the process that a specific snapshot is obtained.

Bendich *et al.* [2016] use diverse summaries of the PH information to quantify the degree of branching and looping in brain vessels, for a population of brain artery trees, and find strong correlations between these physical shape properties and external variables such as age and sex. These correlations are also shown

to be stronger than those obtained by earlier analyses of the data, using classical statistical methods. A crucial finding of their analysis is the fact that high-persistence features alone do not yield the highest correlations. Moreover, only features with mid-range persistence values are necessary to obtain a significant correlation. This fact, seemingly at variance with the earlier interpretations of PH, has been gradually accepted by the TDA community, where the search for isolated, maximally-persistence features in the data is no longer considered the ultimate goal.

Lee *et al.* [2017] use PH to quantify geometric similarity between different porous materials. A point sample is obtained from the pore surfaces of each material, and the PH barcodes for these samples are computed. The structures that generated the samples can then be identified by computing the distances between their barcodes and the precomputed barcodes of some reference materials, which are taken as representatives of different pore structures.

Pearson *et al.* [2015] show PH provides a qualitative method to measure the degree of hexagonal order (i. e., the degree of deviation from a purely hexagonal grid) exhibited by nanodot arrays produced by ion bombardment. This method is contrasted with the qualitative use of Fourier methods, for which the 2-dimensional Fourier transform of the available data is computed and visually inspected to verify the presence of hexagonal patterns.

Robins and Turner [2016] consider the homology rank functions of a persistent homology group $H_k^{i,j}$, defined by

$$\beta_k(i,j) = \operatorname{rank} H_k^{i,j}, \tag{6.2}$$

as vectors in a Hilbert space and use functional principal component analysis to summarize their values over different realizations of a physical experiment. The setting here is that of a crystalline structure which is held in place by a magnetic field, forming a hexagonal pattern. As the magnetic force weakens, the crystal melts and undergoes two phase transitions, first into a *hexatic* phase, where its 6-fold rotational symmetry is retained, but the long-range translational order is disturbed. Second, it enters the isotropic liquid phase. The idea is that the persistent homology

rank functions can serve as a measure of spatial order which can detect both phase transitions, something no other known method can do effectively. These functions are sensitive both to local spatial patterns, and also to global topological features which emerge from higher-order spatial correlations. The different point patterns taken from the experiments can be correctly classified into their different phases by using the homology rank functions. Another application they present is the classification of spatial point patterns, as well as a parametric null-hypothesis test for complete spatial randomness using the functions $\beta_0(i,j)$ and $\beta_1(i,j)$.

The works of Pearson *et al.* [2015] and Robins and Turner [2016] motivated the use of persistent homology to study the spatial distribution of cloud fields in this dissertation. The main difference in the approach presented here is the use of the *stable rank invariant*, to be defined in the next section, in place of the homology rank function. We will show how this invariant can be used to quantify specific properties of spatial point patterns, and link these to physical characteristics of the underlying cloud fields from which the point patterns are generated. We also present a non-parametric counterpart to the test for complete spatial randomness defined by Robins and Turner, by defining an index for spatial randomness based on the stable rank. Finally, we show how a generalization of the stable rank via *persistence contours* allows the classification of point patterns based on their morphology at specific spatial scales.

6.2 STABLE RANK INVARIANT

Recent work in applied topology has focused on the problem of multiparameter persistence, which is the generalization of persistent homology when the filtered complex depends on two or more parameters. This can be the case when not only distance is important, but also local measures of density or curvature, for example. The central issue in multiparameter persistence is that there is no analogue to the decomposition theorem, Theorem 3.6.3 [Carlsson and Zomorodian, 2009]. This means that there are no topological invariants which can be computed efficiently. A current line of research is therefore the introduction of new invariants which

do not depend on the existence of such algebraic decomposition results. One such invariant is the *stable rank invariant*, which we will define in this section. A complete mathematical treatment of the subject lies beyond the scope of this dissertation, so we will simply state the basic definitions and quote the necessary results without proof. For further information, the reader is referred to Scolamiero *et al.* [2017].

First, a note on terminology. In the theory of vector spaces, the rank of a space V is the number of elements in a basis of V. Finite-dimensional vector spaces are completely characterized by this invariant. When computing persistent homology, if we assume the underlying coefficients lie in a field $\mathcal{F}$ (as is the case with the $\mathbb{Z}_2$ coefficients used here), then the persistent homology groups $H_k^{i,j}$ are actually vector spaces, and the maps induced by inclusion

$$\cdots \to H_k(K_{i-1}) \to H_k(K_i) \to H_k(K_{i+1}) \to \dots$$

are linear maps. Formally speaking, this is an $\mathbb{R}$-parameterized sequence of vector spaces. It is also possible to define the rank of this sequence: it is simply the total number of distinct homology generators that appear at each of the $H_k(K_i)$. This invariant is not, however, stable, in the sense that a small perturbation of the input data can lead to large differences in the ranks of the resulting homology sequences, as arbitrarily many generators can appear.

Stabilization of this invariant depends on being able to choose a pseudometric μ between $\mathbb{R}$-parameterized sequences of vector spaces.

Definition 6.2.1. Let $V_\bullet$ be an $\mathbb{R}$-parameterized sequence of vector spaces. Its *stable rank* is the function $S(V_\bullet) : \mathbb{R} \longrightarrow \mathbb{R}$ defined by

$$S(V_\bullet)(r) = \min\{\operatorname{rank}(U_\bullet) \mid \mu(U_\bullet, V_\bullet) \leqslant r\}, \tag{6.3}$$

where μ is a pseudometric.

This is a decreasing function of r, since as r increases the minimization happens over larger balls. Moreover, since we are computing persistent homology for finite datasets, it follows that topological changes can occur only at a finite number of steps $\{r_i\}$ in the filtration. This implies that the stable rank changes only at a

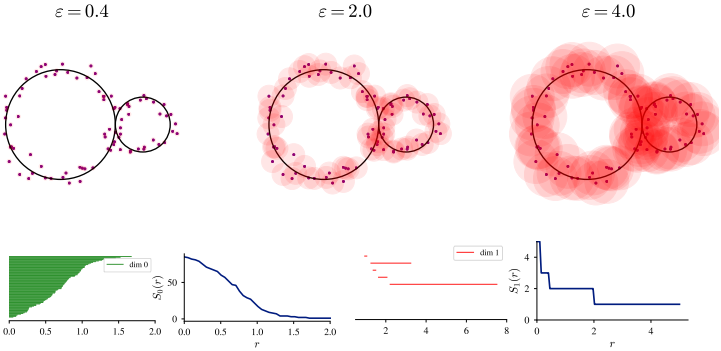


Figure 6.1: Top: Points drawn with noise from the wedge sum of two circles, and three steps in the Vietoris-Rips filtration. Bottom: The resulting persistence barcode and stable rank functions for H_0 (left) and H_1 (right).

finite number of points, hence making $S(V_\bullet)$ a piecewise constant function.

Furthermore, Scolamiero *et al.* [2017] showed the following facts about the stable rank. First, it is stable with respect to variations in the input data. Second, given the pseudometric μ, the map $V_\bullet \mapsto S(V_\bullet)$ is continuous. Third, the choice of pseudometric is equivalent to choosing a *contour* function, which is a function $C : \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R}$ satisfying

1. $v \leqslant w$ and $\varepsilon \leqslant \tau \Rightarrow C(v, \varepsilon) \leqslant C(w, \tau)$
2. $C(C(v, \varepsilon), \tau) \leqslant C(v, \varepsilon + \tau)$

for all v, w, ε, $\tau \in \mathbb{R}$. In the case of one-parameter persistence, this simplifies the computation of the stable rank to

$$S(V_\bullet)(r) = |\{[b_i, d_i) \mid C(b_i, \varepsilon) < d_i\}|. \tag{6.4}$$

For the remainder of this chapter, we will use almost exclusively the so-called *standard contour*, defined by

$$C(v, \varepsilon) = v + \varepsilon. \tag{6.5}$$

This results in the following stable rank function:

$$S(r) = |\{[b_i, d_i) \mid d_i - b_i > r\}|. \tag{6.6}$$

Applying this to the two circle example we saw before, we obtain the functions shown in Figure 6.1. This simple example shows how the stable rank encodes the structure of the persistence barcode, namely, we can see how different bars/intervals are expressed by $S_1(r)$.

The next section will illustrate the utility of this definition with a simple example, and provide some intuition on what information it expresses.

6.3 APPLICATION TO REGULAR POINT PATTERNS

We will describe the use of the stable rank invariant as a descriptor for the spatial distribution of point sets on the plane. First, recall that if γ represents a k-cycle in the persistent homology of a filtered complex K, it has an associated birth-death value pair, (b_γ, d_γ), and its persistence is defined as $d_\gamma - b_\gamma$. The stable rank invariant for persistent homology of order k, S_k, is then defined as

$$S_k(r) = |\{\gamma \in PH_k \mid d_\gamma - b_\gamma > r\}|, \tag{6.7}$$

that is, its value at r is the number of k-cycles with persistence greater than r. Here we use PH_k to denote all persistent homology classes in dimension k.

A key fact about the stable rank is, as its name implies, its stability, which we will now illustrate. For this we consider three different noisy grid patterns on the plane: a square grid, a triangular grid, and a hexagonal grid. If these patterns are perfectly regular, then the persistent homology of the filtered complex built on them is trivial: all points being pairwise equidistant, with distance r^*, we have that for any $r < r^*$ the simplicial complex K_r is the union of all points on the grid. For $r \geqslant r^*$, we obtain a single connected component that encompasses all points on the grid. This is the same fact that was used by Pearson *et al.* [2015] to define a persistent-homology-based metric for measuring spatial disorder in the case of a hexagonal grid. We will thus add random noise to the point grids by perturbing each point with a small, normally-distributed offset. Each noisy grid pattern is generated 1000 times, and for each one we compute its associated stable

ranks, $S_0(r)$ and $S_1(r)$. The results are as shown in Figure 6.2. Here we can clearly see the stability property at work, since even if each individual realization of a grid pattern is different from all others, the fact that random perturbations are small implies that the differences between the stable ranks are also small. The different functions are thus clustered together.

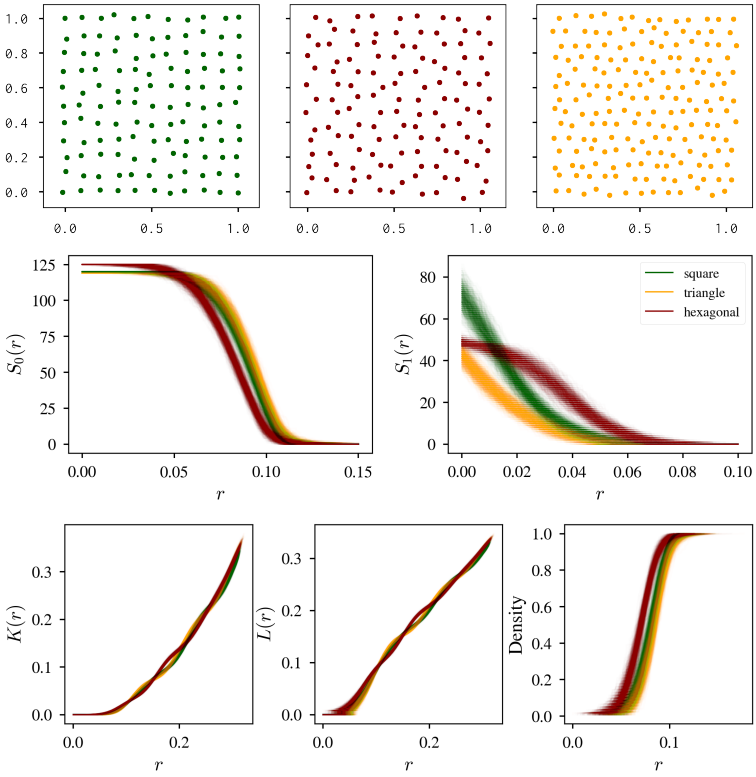


Figure 6.2: Top row: example of noisy grid point patterns. Middle row: Stable rank functions for 1000 realizations of the noisy grid patterns. Bottom row: Ripley's K and L functions, and the nearest neighbor distribution $\hat{G}(r)$.

For comparison, we also show two classical descriptive statistics for spatial patterns, Ripley's K and L functions. The K function is defined as

$$K(t) = \frac{1}{\lambda} E[\#\text{ points within circle of radius } t \text{ centered on random point}], \tag{6.8}$$

where λ is the point density. A sample estimator of this function is

$$\hat{K}(t) = \frac{1}{\lambda} \sum_{i \neq j} \frac{I(d_{ij} < t)}{n}, \tag{6.9}$$

where d_{ij} is the pairwise distance between points i and j in a spatial point pattern with a total of n points. The density is then estimated as $\lambda = n/A$, with A being the area of the spatial domain. This function measures the distribution of the pairwise distances, and summarizes second-order properties of the point process [Ripley, 1976, 1977]. The L function is a variance-normalized K function, with sample estimator given by

$$\hat{L}(t) = \left(\frac{\hat{K}(t)}{\pi} \right)^{1/2}. \tag{6.10}$$

These functions were also used by Robins and Turner [2016] in their analysis of spatial point processes. Here we compare their structure to that of the stable ranks for H_0 and H_1 homology. Also shown is the empirical nearest-neighbor distribution function, which is the cumulative distribution function of the inter-point distances in a spatial point process. Concretely, its value is given by

$$\hat{G}(r) = \frac{\#\{p \in \mathcal{P} \mid d_p \leqslant r\}}{\#\{p \in \mathcal{P}\}}. \tag{6.11}$$

There is not a significant difference between the three nearest-neighbor functions, and similarly for the H_0 stable ranks. However, the H_1 stable ranks do exhibit important qualitative differences for the three point patterns. Consider the value $S_1(0)$, which is the number of cycles in each case with a persistence greater than 0, in other words, it is simply the count of H_1 features in each case. It is significantly larger for the square grid, and smallest for the

triangular grid. Since the filtration used to compute the persistent homology uses the Vietoris-Rips complex, it is instructive to recall the following fact: given a value $r > 0$, if there are three vertices v_1, v_2, v_3 such that their pairwise distances satisfy

$$d(v_i, v_j) \leqslant r, \quad i \neq j, \tag{6.12}$$

then the 2-simplex spanned by these vertices is added to the complex. This explains why the value $S_1(0)$ is smallest for the triangular grid: since the points are laid out in an approximately triangular pattern, it is likelier that any 3 neighboring points will satisfy condition 6.12, thus resulting in the addition of a 2-simplex to the complex. As space gets filled by these simplices, it is thus also less likely that non-bounding cycles will form.

The values of $S_1(.02)$ then show that, while almost all the features appearing for the hexagonal grid have persistence at least .02, the contrary is true in the case of the square grid, with its $S_1(.02)$ value being less than half of its $S_1(0)$ value. The same is true for the triangular pattern. From this point onwards, these two patterns see a rapid decline in their feature counts with increasing r. The count for the hexagonal pattern also decreases, as it must, but at a markedly slower pace. In other words, the non-bounding cycles formed throughout its filtration have, in general, a larger persistence than those formed in the other patterns. Geometrically this relates to both the number of vertices that span a cycle, as well as to their physical size.

These differences can also be quantified by looking at the persistence barcodes. However, the stable rank representation has the distinct advantage of being a real-valued function, which is easier to operate with than an arbitrary collection of number pairs. It remains expressive enough to detect such properties of the spatial distribution of the original dataset as shown here.

6.4 GEOMETRIC REPRESENTATION OF CLOUD FIELDS

Recall that in previous chapters we needed to construct a cubical complex representation for flow structures, represented by up- and downdraft domains, and compute their associated Betti numbers. We could follow a similar strategy here and construct the

cubical complex representation of a cloud field, and obtain its first three Betti numbers: these would count the number of clouds present in the field (β_0), the number of holes or tunnels (β_1) in them, and the number of voids enclosed by them (β_2). These quantities, however, would not offer much information regarding the spatial distribution of the clouds themselves. For this reason we will use PH as the descriptive tool.

To this end, we will need to use a different type of geometric representation. In this case, the representation will be that of a *pointcloud*, that is, a finite set of point in Euclidean space $\mathbb{R}^2$ or $\mathbb{R}^3$. This has an additional advantage: statistical methods and concepts for studying spatial distributions of point-like objects are well-developed, which is not true for objects with a spatial extension (area or volume). In particular, we are interested in assessing the spatial randomness of a set of clouds, which is termed *complete spatial randomness* in the theory of spatial point processes [Ripley, 1988]. The implicit geometric construction will then be a filtered Vietoris-Rips complex (see Section 3.6) and the persistence intervals will be computed by using the C++ implementation of the persistence algorithm in Ripser [Bauer, 2017].

6.4.1 *General methodology*

The issue of variable selection in this case is simpler than when investigating convective structure: we will use the values of liquid water content (`ql`) produced by the numerical model. We also note that, given the fact that we will be dealing with shallow cumuli only, it is sensible to use a two-dimensional approximation to describe their spatial distribution. Formally, we denote the computational domain of the simulation as before by a cubical grid

$$\Omega = \{1, 2, \ldots, N_t\} \times \{1, 2, \ldots, N_x\} \times \{1, 2, \ldots, N_y\} \times \{1, 2, \ldots, N_z\}, \tag{6.13}$$

and its subsets corresponding to a constant time coordinate are denoted by

$$\begin{aligned}\Omega_t &= \{(t, a, b, c) \in \Omega \mid 1 \leqslant a \leqslant N_z,\ 1 \leqslant b \leqslant N_y,\ 1 \leqslant c \leqslant N_x\} \\ &\simeq \{1, 2, \ldots, N_x\} \times \{1, 2, \ldots, N_y\} \times \{1, 2, \ldots, N_z\}.\end{aligned}$$

Each Ω_t thus corresponds to one simulation timestep. For each such timestep, we will have a total of $N_z \times N_y \times N_x$ cells, each with its own ql value. This means we have a three-dimensional array,

$$Q_t = (ql[i, j, k])_{ijk},$$

where ql_{ijk} represents the ql value at grid coordinates $z = i$, $y = j$, $x = k$. The first step in constructing the point representation is then projecting these data to a plane, or equivalently, to construct the two-dimensional array

$$Q_{t,0} = (ql[j, k])_{jk} \tag{6.14}$$

defined by

$$ql[j, k] = \max\{ql[i, j, k] \mid 1 \leqslant i \leqslant N_z\}. \tag{6.15}$$

This array represents the cloud field as seen from above (or below). We denote the subset of $Q_{t,0}$ where $ql > 0$ (i. e.the cloudy cells) by $\mathcal{P}_t$. The next step in obtaining the geometric representation is separating the connected components of the cloud field, which is performed as before using the Union-Find (UF) algorithm (see Section 5.1). This results in a collection of subsets of $\mathcal{P}_t$, $\mathcal{C}_t \subset 2^{\mathcal{P}_t}$, each one of which represents a separate cloud. We denote the individual components by c_i, $i = 1, \ldots, n$.

Having arrived at this point, we can now consider different strategies to obtain a point representation for the cloud field $\mathcal{P}_t$:

1. Random sampling from $\mathcal{P}_t$. This does not require us to know which connected components make up $\mathcal{P}_t$.

2. Connected component sampling. From each component c_i, we sample a number of points at random. The number of sampled points is proportional to the size of each component.

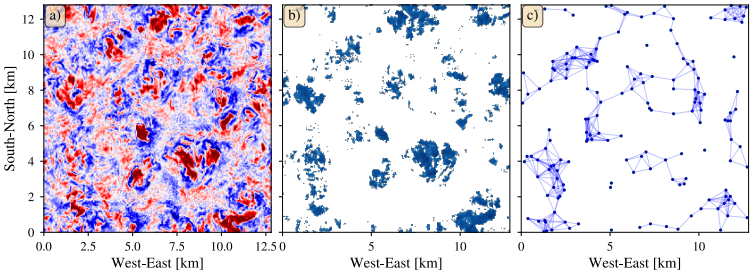


Figure 6.3: Example of the two-dimensional approximation to a shallow cumulus cloud field. Panel **a)** shows the horizontal section of vertical wind velocity w at an altitude of 1.8 km, which corresponds to the bottom of the cloud layer (red : $w > 0$, blue : $w < 0$). Panel **b)** is the column liquid water content ql, i. e.the maximum liquid water value in the vertical direction for each grid cell. Panel **c)** is the point representation of this cloud field, with each point indicating the location with the maximum ql value in each cloud (only connected components of size $\geqslant 3$ are considered). The lines are the 1-simplices in the Vietoris-Rips filtration, at a distance scale of 1.5 km. Periodic boundary conditions were not considered here for visual clarity. Figure from Riihimäki and Licón-Saláiz [2019].

3. Geometric centroid. To each component c_i we assign the point on the plane which corresponds to the center of its bounding box.

4. Maximum liquid water. To each component c_i we assign its point with the maximum ql value.

Applying any of these methods yields the desired pointcloud, which can then be used as input to the PH algorithm to obtain the corresponding persistence intervals (or equivalently, the persistent Betti numbers). An example of this is shown in Figure 6.3, which shows the vertical wind velocity at cloud base height (panel **a**) and the corresponding two dimensional cloud field $\mathcal{P}_t$ (panel **b**). The close relationship between the locations of both the individual clouds and the strong updrafts in the domain is made clear by the two figures. Panel **c** then shows both the pointcloud obtained

from $\mathcal{P}_t$ by selecting the local maxima of ql (strategy number 4 above), as well as the 1-skeleton of the Vietoris-Rips complex built on this pointcloud with a distance parameter of 1.5 km. Recall from the definition that the 1-simplex spanned by two vertices u, v is part of this complex if $d(u, v) \leqslant 1.5$, where d represents the Euclidean metric on the plane, and the 1-skeleton is the collection of all 1-simplices in the complex. What it shows, then, is the set of clouds (or, more precisely, their representative points) which are pairwise closer than a specific distance, here 1.5 km, connected by a line.

6.4.2 *Data*

For this study 10 simulation days from the DALES model (see Section 2.5) were selected which display formation of boundary layer cumulus clouds, no precipitation and small cloud cover around noon. Furthermore, we focus on the time period with shallow cumulus cloud activity, starting at 09:00h and continuing into the afternoon. This is the part of the day when the convective boundary layer dominates the system dynamics (cf. Chapter 4). Conversely, cloud formation before sunrise or at night might obey different processes (e.g larger-scale phenomena such as thunderstorms), and would give rise to spatial patterns different from the ones we are interested in here.

Figure 6.4 shows the cloud cover percentage for this time window in the 10 simulation days considered here. As can be seen, the average value for all 10 days lies somewhere in the range $[0.08, 1.5]$, and the maximum values rarely exceed 0.25. Whenever they do, this is only a transient phenomenon. Indeed, the 10 simulations have a qualitatively similar evolution: cloud cover starts at small values in the morning, steadily increasing towards noon. After reaching a maximum value, somewhere between 12:00h and 13:00h, cloud cover gradually diminishes, finally vanishing at 18:00 in most of the cases shown here. This is the typical diurnal cycle exhibited by shallow cumuli, and is connected with the diurnal evolution of the CBL as studied in Chapters 3 and 4. With this in mind, the questions we will look at in this chapter can be stated as follows:

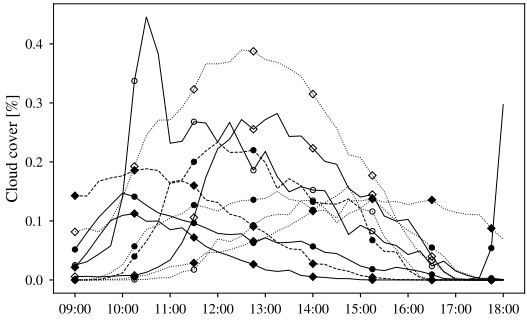


Figure 6.4: Time series showing the cloud cover for the 10 simulation days considered here, within the time period 09:00h–18:00h.

1. Can the machinery of PH be used to inform conclusions regarding the properties of a given cloud field? For example, is it possible to estimate the value of cloud cover from pointcloud data only?
2. Does PH allow for a principled assessment of the spatial structure/randomness of a cloud field?
3. Given the qualitative similarity exhibited by the 10 simulation days, and the fact that this is the normal behaviour we would expect, are there any other properties of the field that can be uncovered by using PH?

6.5 ESTIMATION OF CLOUD COVER

We begin by addressing the first of these questions, namely: what kind of physical insights about cloud populations can we obtain from the information contained in the persistent homology groups computed from model data? This is analogous to the results presented in Chapters 3 and 4, where it was shown that the (non-persistent) Betti numbers computed from the vertical wind velocity data do show a close relationship with structural properties of the CBL. In this case, following the discussion in Section 6.1 regarding the use of persistent homology in the analysis of spatial patterns, we will focus on applying it to the analysis of the spatial

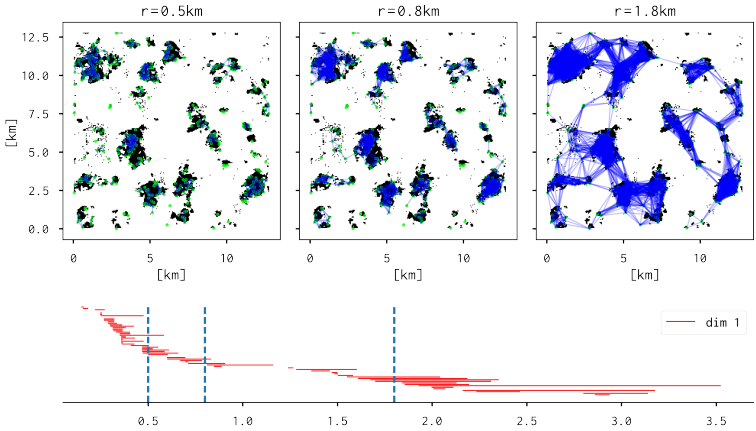


Figure 6.5: Top: Three steps in a filtration built from a sampled SCC field (only the 1-skeleta are shown, i.e. only 1-dimensional simplices; the domain C_t^+ is represented in black). Bottom: its associated barcode. Figure from Licón-Saláiz *et al.* [2018].

distribution of the clouds themselves. We will connect the values of the stable rank invariant obtained from point samples to the cloud cover of the underlying cloud field, and show how this can be used to infer the structure of the cloud size distribution.

6.5.1 *Methodology and experimental setup*

POINT REPRESENTATION Given an individual cloud field as defined in Equation 6.15, the first step is to extract from it a pointcloud representation as discussed in Section 6.4. To this end, we obtain the binary array which is 1 for all grid points (i, j), $i \in 1, \ldots, N_x$, $j \in 1, \ldots, N_y$ with positive liquid water content ($ql(i, j) > 0$), and 0 for the rest. We denote the set of grid points with positive liquid water by $\mathcal{P}_t$, but now use the physical coordinates of these points instead of their indices. That is, for each grid point (i, j) such that $ql(i, j) > 0$, the point (x_i, y_j) is added to $\mathcal{P}_t$. This set represents the cloudy cells in the two-dimensional grid. We then obtain a random sample of points $P_t \subset \mathcal{P}_t$ by drawing

random points from the connected components of $\mathcal{P}_t$, that is, from the individual clouds. The number of points sampled from each component is proportional to its size in grid cells, but always greater or equal than one. The idea behind this is to capture some of the variability in the size and shape of individual clouds in the distribution of sampled points by resampling from $\mathcal{P}_t$, thus yielding a set of samples $\{P_{(t,1)}, \ldots, P_{(t,n_s)}\}$.

BARCODE COMPUTATION We next compute the persistence barcode for each of these point clouds $P_{t,\ell}$. The necessary data for this computation is the pairwise distance matrix D, which has as its (i,j)-th element the distance between the points u_r, $u_s \in P_{t,\ell}$. As was established in Section 6.4.2, the DALES simulations from which these datasets were obtained are run with periodic lateral boundary conditions. This implies that the pointcloud $P_{t,\ell}$ is embedded in the flat torus, which is the quotient space of $\mathbb{R}^2$ under the identifications

$$(x, y) \sim (x + 1, y), \quad (x, y) \sim (x, y + 1)$$

for x, $y \in [0, 1]$. Therefore, the distance between the points $u_r = (x_r, y_r)$ and $u_s = (x_s, y_s)$ on the flat torus is

$$d((x_r, y_r), (x_s, y_s)) = \sqrt{dx^2 + dy^2},$$

where

$$dx = \min\{|x_r - x_s|, |(x_r + 1) - x_s|, |x_r - (x_s + 1)|\} \tag{6.16}$$

$$dy = \min\{|y_r - y_s|, |(y_r + 1) - y_s|, |y_r - (y_s + 1)|\}. \tag{6.17}$$

In this expression we consider the domain to be scaled to unit area, but for our computations the distances obtained were then rescaled to correspond to the true domain area of $12.8\,\mathrm{km}^2$. This set of pairwise distances is used as input for building the filtration and computing the corresponding barcode, $\mathcal{B}_{t,\ell}$.

STATISTICAL ANALYSIS The barcode $\mathcal{B}_t$ (we omit the subscript denoting the resampling for clarity) is the multiset of persistence intervals $[b_i, d_i)$ of the different topological features that appear

throughout the metric filtration on the point set P_t. Here, b_i refers to the *birth* value of a feature in the filtration, and d_i to its *death* value. Moreover, we will have two distinct multisets, the first corresponding to the H_0 features, and the second to the H_1 features. We denote these sets by $\mathcal{B}_t^0$ and $\mathcal{B}_t^1$, respectively. As discussed in Section 6.1, this dataset is not immediately amenable to statistical analysis. We will therefore use the stable rank function for the analysis, which in the case of one-dimensional persistence is equivalent to a function which for each value r counts the number of features with persistence greater than r (see Section 6.2):

$$S_k(r) = \#\{(b_i, d_i) \in \mathcal{B}_t^k \mid d_i - b_i > r\}, \tag{6.18}$$

where the subscript k refers to the dimension of the features being counted.

We will make use of a further numerical descriptor in this section, namely the *multiplicative persistence* of features appearing in a filtration. Given a persistence interval $(b_i, d_i) \in \mathcal{B}_t$, its multiplicative persistence is defined as

$$\pi_i = \frac{d_i}{b_i}. \tag{6.19}$$

This is only well-defined for persistent homology groups H_k of dimension $k \geqslant 1$, since in the case $k = 0$ all features have a birth value $b_i = 0$, at least in the present case of a Vietoris-Rips filtration. Multiplicative persistence has the property of being scale-invariant, in the sense that it does not depend on the physical size of the cycle, but only on its "shape" or stability [Bobrowski *et al.*, 2017]. The corresponding feature counting function for multiplicative persistence is then

$$S_k^\pi(r) = \#\{[b_i, d_i) \in \mathcal{B}_t^k \mid d_i/b_i > r\}, \tag{6.20}$$

and we will refer to the one defined in Equation 6.18 by $S_k^\alpha(r)$, where the superscript α represents the additive persistence $\alpha_i = d_i - b_i$. When there is no ambiguity, we will omit the superscripts and refer to the stable rank functions simply as S_0 and S_1.

We have just described a method which computes a set of descriptor functions, the stable rank invariants, for a given collection

of objects in space (in this case, clouds). Moreover, we have seen in Section 6.3 that these functions are effective in discriminating between different regular point patterns on the plane. We now look at the relationship between these functions and cloud field structure in more detail. An example is shown in Figure 6.6, where the top row shows three different cloud fields, each obtained from a different simulation, but all three at the same time of day (12:00h). Also shown in these panels are the resulting pointclouds obtained by sampling from the connected components of each field. The middle row shows the H_1 barcodes for each pointcloud, computed from the pairwise distance matrix,. The last row contains the stable ranks S_1^α corresponding to the H_1 barcodes.

The first observation we can make is that a field with larger cloud cover will result in more sampled points than a field with smaller cloud cover, hence the number of H_1 features for that sample will be larger, as will the values of $S_1^\alpha(0)$. A more interesting fact is that the three S_1^α curves shown here have a similar structure: first, we see an interval where the values of $S_1^\alpha(r)$ shrink rapidly, followed by a much more gradual decline. The rate of decrease appears similar across the three cases for the first part of the domain; the second part shows a more noticeable difference. This corresponds to the fact that the three H_1 barcodes have a large number of short bars, whereas the number of longer bars is different for each, as well as their specific lengths. The cloud fields from which these barcodes were derived give us a geometric interpretation of these facts: most of the shorter bars would correspond to small non-bounding cycles generated by the numerous sets of points clustered together in a small area. The longer bars, on the other hand, indicate larger non-bounding cycles generated by the empty spaces between the clouds (see Figure 6.5). This means that the H_1 stable rank, as shown here, should be able to signal differences in the interspersion patterns present in these clouds fields, especially if we focus on its values for larger values of the filtration parameter r.

In this section, we will not look at these different patterns, but at a question of a more general nature. Recall the observation that, as cloud cover increases, we will in general see more H_1 features. We now take a closer look at the nature of this increment. Specifically,

an increase in cloud cover can happen because the clouds are larger, because there are more of them, or some combination of these two factors. In terms of the pointclouds sampled from the cloud field, the former case should result in a more uniformly random scattering of the sampled points, whereas the latter case would show more order, or clustering, as larger clouds will have a greater likelihood of contributing several points to the random sample. These differences should be reflected in the values of the stable rank functions, specifically S_1^α and S_1^π. The number of features with a large persistence value should not decrease in this case, representing the separation of large clouds by empty space. In the uniformly random case, a denser distribution of points on the plane would preclude the appearance of such geometrical features.

In concrete terms, we will use the pointwise evaluation of the functions S_1^α, S_1^π at fixed points $\{r_i\}$ in their domains as explanatory variables in two different linear models to explain the values of cloud cover (cc) for the cloud fields under study. The models will have the form

$$\log(cc) = \beta_0 + \sum_{i=1}^{n_\alpha} \beta_i \log\left(S_\alpha^1(r_i)\right), \tag{6.21}$$

$$\log(cc) = \beta_0' + \sum_{i=1}^{n_\pi} \beta_i' \log\left(S_\pi^1(r_i)\right), \tag{6.22}$$

with n_α and n_π being the number of evaluations used in each case.

The spatial randomness of a given set of points in space can also be inferred by using the *maximal persistence values* found for the metric filtration built from the point set. These values are defined as

$$\begin{aligned} A_k &= \max\{\alpha(z) \mid z \in PH_k\}, \\ \Pi_k &= \max\{\pi(z) \mid z \in PH_k\}., \end{aligned} \tag{6.23}$$

It was shown by Bobrowski *et al.* [2017] that, when the underlying point distribution is a realization of a homogeneous Poisson

point process on the unit square, the maximal persistence has asymptotic behavior given by

$$\Pi_k(n) \sim \left(\frac{\log n}{\log\log n}\right)^{1/k},$$

with n being the number of points in a realization of the point process. A similar result holds for the Poisson point process on the flat torus, if we ignore the essential homology classes. A deviation from this behavior as the number of points n grows would thus indicate a deviation from complete spatial randomness in the process that generated the points. We would take this as an indication that the increase in cloud cover is produced by an increase in cloud size, rather than cloud number.

EXPERIMENTAL SETUP The starting point is the 10 DALES simulation days, as specified in Section 6.4.2. From each day we will use the 26 timesteps comprised between 09:00h and 15:30h, which capture most of the variability in cloud cover induced by the diurnal cycle (see Figure 6.4). The dataset thus contains 260 different cloud fields, each of which is reduced to a two-dimensional representation as described in Section 6.4. Further, we will consider 50 different values of the sampling ratio, $s = 0.001, 0.002, \ldots, 0.049, 0.05$ (recall that a sampling ratio of s gives us a sample with $s\,N$ points, where N is the number of cloudy cells in the field). For a given sampling ratio s, we generate 10 samples per cloud field. These samples are then used in the computation of the persistence barcodes and the associated stable rank functions as described above.

Having computed these, we select the values of $S_1^{\alpha}(r)$ for $r = 0.25, 0.5, 1, 1.5$, and the values of $S_1^{\pi}(r)$ at $r = 2, 3, 4, 5$ to be used as explanatory variables in the linear models given by Equation 6.21 and Equation 6.22. This is done in two separate ways:

1. For a fixed sample ratio s, use the 2600 point samples to fit the models.

2. Use the data for all sampling ratios at the same time, 130 000 point samples in total.

Additionally, for each point sample we also compute its maximal persistence values, A_1 and Π_1. An analysis of the results of this setup is given in the next subsection.

Two additional comparison datasets were generated: the first, a dataset made of diverse realizations of a homogeneous Poisson point process, and the second one made of realizations of a Thomas point process. A Poisson point process with rate parameter λ is a spatial point process defined on $\mathbb{R}^2$ which fulfills the condition that, if $E \subset \mathbb{R}^2$ is a measurable subset of the plane, then the number of points within E, $N(E)$, is a random variable with probability density function given by

$$P[N(E) = n] = \frac{(\lambda\,\mu(E))^n}{n!} e^{-\lambda\,\mu(E)}, \tag{6.24}$$

where $\mu(E)$ denotes the measure of E. In other words, the number of points within a given set has a Poisson distribution, with intensity parameter proportional to the measure of the set. This is the condition of complete spatial randomness.

The Thomas point process is a type of Poisson clustering process, and is defined as follows. Let C be a homogeneous Poisson point process with intensity parameter κ. A realization of this process, $\{c_1, \ldots, c_n\}$, gives the parent points of the clustering process. For each parent point c_i, draw m points from the bivariate normal density $N_2(c_i, \sigma^2 I_2)$, where m is Poisson with parameter μ. The Thomas point process then has three parameters: κ, which controls the number of parent points; μ, which controls the number of child points; and σ^2, which controls the dispersion of points within each cluster.

For each point sample available in the original dataset, one corresponding sample was generated for each of the Poisson and Thomas datasets. In both cases the parameters were selected so that the expected number of points would be the same as the original point sample. In the case of the Thomas point process, the number of parent points was generated from a Poisson distribution with rate parameter equal to the number of clouds times the sampling ratio used, and $\sigma = 0.05$. The edge effect was avoided by wrapping sampled points around the boundaries (flat torus). Examples of this are shown in Figure 6.7, where two samples from

a cloud field are shown, obtained with sampling ratios of 1% and 3%. The corresponding realizations of the Poisson and Thomas point processes are also shown.

6.5.2 *Results and Interpretation*

An exploratory analysis of the computed values for the stable rank functions $S_1^{\pi}(r_i)$ and $S_1^{\alpha}(r_i)$ (in the rest of this subsection we omit the subindex for simplicity) found that these have a strong positive relationship with cloud cover, or equivalently, with the number of points per sample. These relationships are shown in Figure 6.8 for $S^{\pi}(2)$ (**a**) and $S^{\alpha}(.5)$ (**b**). As expected from the discussion above, the values of $S^{\pi}(2)$ increase with the number of points in the sample, and this increase is monotonic and non-linear. For $S^{\alpha}(.5)$, there is also a positive relationship with sample size for the smaller samples, but beyond that the values of $S^{\alpha}(.5)$ actually appear to decrease. This effectively shows the scale invariance property of multiplicative persistence: as the pointclouds become denser, there will be less margin for larger, longer-lived cycles to appear. Hence, looking only at bar length as an indicator of structural importance would yield an incomplete picture in this case. This does not happen if we look instead at the death/birth ratio of multiplicative persistence. Further confirmation of this fact is provided by the corresponding histograms for the Poisson point process samples (Figure 6.8, **c, d**) and the Thomas point process samples (Figure 6.8, **e, f**). In both cases the values of $S^{\pi}(2)$ again increase monotonically with sample size, but for the case of the Poisson process $S^{\alpha}(.5)$ becomes a decreasing function of sample size for sample sizes larger than 250 points. This is further evidence for the fact that, under complete spatial randomness, an increasing number of points will be in general associated with less H_1 features.

The models given in Equations 6.21 and 6.22 were fitted to the data and evaluated as follows: for each sample ratio s, 10-fold cross validation was performed on the dataset generated with sampling ratio s, the coefficient of determination (R^2) was computed for each fold, and the 10 R^2 values were averaged. These scores are shown in Figure 6.9. These results are consistent

with the relationships described above, and we can see that the model using the multiplicative stable rank values get better as sample size increases, with an R^2 of almost 1 for sampling ratios greater than 5%. On the other hand, the model that uses additive stable rank values actually gets worse for increasing sample ratios.

A more informative experiment is to evaluate the same models given by Equations 6.21 and 6.22, but this time using the data for all sampling ratios at the same time, that is, the training set will be a subset of the 130000 point samples which were obtained from the 10 simulations. This is intended to make the size of each point sample essentially random, thus giving us a better notion of how stable these models are. The evaluation was again conducted by 10-fold cross validation and averaging the R^2 score over the different folds. The final R^2 values were 0.6645 for the model using additive persistence, and 0.4947 for the model using multiplicative persistence. For each model 50 runs of 10-fold cross-validation were performed, and the accumulated errors in % points are shown in Figure 6.10. Here a significant bias is apparent in the errors for the S^π model, which is to be understood as a consequence of the small variability seen in the relationship between pointcloud size and the values of S^π (see Figure 6.8, **a**), which makes the model very brittle. For example, a point cloud with 200 points can be the result of sampling a cloud field with 4000 cloudy cells (corresponding to a cloud cover of 6.1%) with a sampling ratio of 5%, or of sampling a cloud field with 20000 cloudy cells (with a cloud cover of 30.5%) at a ratio of 1%. In both cases, however, the values of S_π will be very similar. Conversely, the S^α model turns out to be more stable.

We now turn to the question of the maximal persistence, Π_1, observed in each pointcloud. These values are shown in Figure 6.11 for random sampling at 5% from all cloud fields, as well as the corresponding realizations of the Poisson and Thomas point processes.

As was shown in Bobrowski *et al.* [2017], for a Poisson point process we have

$$\Pi_1(n) \sim \frac{\log n}{\log\log n} = \Delta_1(n), \tag{6.25}$$

up to a constant factor, for large values of n. This result is consistent with the experimental evidence shown here in the case of a Poisson point process. The data from the Thomas point process and the cloud field samples, however, do not display this behavior. The result presented by Bobrowski *et al.* depends crucially on establishing a relationship between the persistence of a cycle in a filtered complex and the number of vertices that generate it, which in turn depends on the spatial distribution of the vertices. Such a relationship does not necessarily hold for a clustering process such as the Thomas point process, or the points sampled from a cloud field, hence the divergence shown here.

When taken together, the results in this section suggest that, despite significant morphological differences observed in the cloud fields that make up the dataset studied here, there are also underlying structural properties which are common throughout all of them. Persistent homology is expressive enough to capture these properties, which also allows for a comparison of the spatial distribution of points sampled from a cloud field with those obtained by different spatial point processes (see Figure 6.8). Moreover, this fact can be exploited to construct predictive models for cloud cover using only point data. The performance of these models and the asymptotic behavior of the maximal persistence values $\Pi(z)$ show that, for the cloud population considered here, increase in cloud cover tends to be associated with larger clouds, and not with more of them. This is consistent with the results of van Laar *et al.* [2019] on cloud size evolution for the same population.

6.6 MEASURING THE SPATIAL RANDOMNESS OF CLOUD FIELDS

6.6.1 *Methodology*

In the previous section, we have used points randomly sampled from different cloud fields to compute a persistent homology signature for these fields. These signatures have enabled us to make conclusions about the nature of the processes that generated the point samples. In particular, we could distinguish the point samples as being produced by some form of clustering process as opposed to a purely random scattering. This is an unsurprising

conclusion, because sampling several points from the same cloud will necessarily result in some form of clustering in space. In this section, by contrast, we will look at the spatial distribution of the clouds themselves. This is tied to an important issue in the study of cloud formation, namely the quantification of spatial organization, or lack thereof, in a given cloud field.

To do this, we will again need to obtain a point representation for a given cloud field, to be used in computation of its persistence barcode and the associated stable rank functions. The approach will now be different, as we want to use only one point per cloud, thus focusing the method more on the positions of the clouds and not so much on their size or shape. Two of the strategies discussed in Section 6.5.1 will be used here, namely: i) to assign to each cloud its geometric centroid as representative point; ii) to assign to it the point where its largest q_l value is found. For this analysis, only additive persistence will be used.

The stable rank function $S_k(r)$, when normalized by its maximum value $S_k(0)$, can be used to define a homological distribution function, in the following sense: denote by S_k^* the normalized stable rank, that is

$$S_k^*(r) = \frac{S_k(r)}{S_k(0)}. \tag{6.26}$$

Clearly, $S_k^*(0) = 1$, and it is a monotonically decreasing function of r. If we now define the function

$$G_k(r) = 1 - S_k^*(r), \tag{6.27}$$

this is a monotonically increasing function of r, and

$$\lim_{r\to\infty} G_k(r) = 1,$$

so we can take G_k to be an empirical cumulative distribution function (CDF), as its value at r is the relative amount of homological features that persist beyond r. This is illustrated in Figure 6.12 for one particular cloud field. Here we show the empirical CDF values obtained for the H_0 persistent homology, together with the maximum-likelihood exponential and lognormal fits to them. This interpretation of the stable rank will be used in defining a metric to quantify the randomness in a spatial point pattern.

A common metric in the assessment of spatial organization is the I_{org} index [Tompkins and Semie, 2017], defined as follows. For a two-dimensional cloud field we index the connected components (the individual clouds) as c_i, and compute their geometric centroids, $\bar{c}_i$. We are interested in how the spatial distribution of the $\bar{c}_i$ compares to what we would expect under complete spatial randomness, that is, if the centroids represent a realization of a homogeneous Poisson point process. To that end, we consider the nearest-neighbor distances d_i, which are defined as

$$d_i = \min\{d(\bar{c}_i, x) \mid x \in \bar{\mathcal{C}} \setminus \{\bar{c}_i\}\},$$

where $\bar{\mathcal{C}}$ represents the set of all centroids. The cumulative distribution function (CDF) of the d_i is

$$G_{d_i}(r) = P[d_i \leqslant r],$$

which in the case of a Poisson point process has the analytic expression

$$G_{CSR}(r) = 1 - \exp(-\lambda \pi r^2),$$

where λ is the Poisson intensity parameter. The value of I_{org} is then defined to be the area under the graph $(G_{CSR}(r), \hat{G}(r))$, where

$$\hat{G}(r) = \frac{\#\{\bar{c}_i \in \bar{\mathcal{C}} \mid d_i \leqslant r\}}{\#\{\bar{c}_i \in \bar{\mathcal{C}}\}}$$

is the empirical estimator of $G(r)$. If $\hat{G}$ matches well with G_{CSR}, the value of I_{org} will be close to 0.5. A value larger than this suggests spatial clustering, while a smaller one suggests dispersion or regularity.

For n realizations of a Poisson point process with intensity parameter λ, we find that their normalized stable ranks S_i^*, and therefore also G_{PH}^i, oscillate within a narrow band (see Figure 6.13). At this point we do not have an analytic expression for the stable rank functions obtained from a Poisson point process, but we can define persistent homology analogues to the I_{org} index via a Monte Carlo procedure by taking the area under the curves defined by $(G_{PH,CSR}^i(r), G_{PH}^i(r))$. Here we take $G_{PH,CSR}^i$ to be the mean CDF computed for n realizations of a Poisson point

process, as shown in Figure 6.13. In the case of a point process in the plane we would then get two values $I_{PH,0}$ and $I_{PH,1}$. We define the index as their arithmetic mean,

$$I_{PH} = \frac{I_{PH,0} + I_{PH,1}}{2}. \tag{6.28}$$

6.6.2 *Results and Interpretation*

We tested the performance of the index I_{PH} defined above, and compared it to the corresponding values of I_{org} in a dataset consisting of 360 distinct cloud fields: we now consider the 36 timesteps between 09:00h and 18:00h for every simulation day. The values of both indices for these cloud fields are shown in Figure 6.14. Each panel shows the 360 values of each index for all fields, computed using 4 different point representations. Panel A shows the values obtained from assigning to each connected component its point with maximum ql value (local maxima); panel B shows the indices obtained when using the local maxima but only of those components with size at least 3 grid cells (all smaller components are ignored). Panel C shows the results of using the geometric centroid of each connected component. Finally, for panel D the geometric centroids were used after discarding the smaller components. These small components can be attributed to numerical imprecisions in the underlying model, and hence are not physically meaningful.

As discussed above, if these indices have a value close to 0.5, it would indicate that the point process that they are evaluated on is close to complete spatial randomness, or a Poisson point process. In the simulations used here, we have cause to expect spatially random behavior: the domain size is too small to allow for deep convection and spatial organization to happen. Moreover, the lack of land surface features or patterns means there are no forcings at different spatial scales. Thus the spatial distribution of physical variables is dominated by the characteristic patterns present in atmospheric turbulence, itself an essentially random process. The values of the persistent homology index I_{PH} strongly support this hypothesis, while I_{org} exhibits values in general

larger than 0.5. This can be attributed to the fact that it is based on nearest-neighbor distances only, whereas the stable rank functions reflect the spatial relationships of the points throughout all spatial scales. This is confirmed by the fact that removing the smaller structures in the fields (those less than 3 grid cells in size) brings the values of I_{org} closer to 0.5 on average, whereas the average for I_{PH} is barely affected. This highlights the fact that, by virtue of using all the spatial information available, the persistent homology based method is inherently more robust than any nearest-neighbor method.

6.7 MORPHOLOGICAL CLASSIFICATION OF CLOUD FIELDS

The results from previous sections have been arrived at by using the standard contour (Equation 6.5), which makes the stable rank function S_k a feature-counting invariant, in the sense that its value $S_k(r)$ is the number of features with a persistence larger than r. This means that all k-cycles that appear throughout the filtration are given the same weight in computing this invariant, irrespective of where in the filtration they appear. In particular, the indication of spatial randomness from Section 6.6.2 is to be understood in this sense.

The stable rank invariant is, however, a very flexible tool for data analysis, since it depends on the choice of contour function. A contour can be understood in this case as a weighting function for different ranges of the filtration parameter. Figure 6.15 shows the stem plot representation of the H_1 barcode for the cloud field in Figure 6.6, as well as three different contour functions. One is the standard contour (Equation 6.5), defined as

$$C(\nu, \varepsilon) = \nu + \varepsilon.$$

This results in the condition $C(b_i, \varepsilon) < d_i$ in the definition of the stable rank, Equation 6.4, being equivalent to $\varepsilon < d_i - b_i$. Thus, a persistence interval $[b_i, d_i)$ will always count towards the value of $S_k(\varepsilon)$ if its persistence is larger than ε. As shown in the stem plot, this threshold grows uniformly with ε across the entire filtration parameter range.

Also shown in Figure 6.15 is the *exponential contour*, defined as

$$C(v, \varepsilon) = k^{\varepsilon} v, \tag{6.29}$$

for some $k \geqslant 1$. This now results in the condition of Equation 6.4 being equivalent to $d_i - b_i k^{\varepsilon} > 0$, or in $(b, d-b)$-coordinates

$$\begin{aligned} C(b_i, \varepsilon) - b_i &< d_i - b_i \\ b_i(k^{\varepsilon} - 1) &< d_i - b_i. \end{aligned}$$

Thus, a persistence interval $[b_i, d_i)$ will now count towards the value of $S_k(\varepsilon)$ only if its length is greater than $b_i(k^{\varepsilon} - 1)$, which now clearly depends on the birth value of the interval. This is the equation of a line through the origin in the $(b, d-b)$-plane, as shown in Figure 6.15, so as we see, those intervals born later in the filtration have a higher persistence threshold to be counted than those born in the first part of the filtration. It is in this sense, then, that the exponential contour gives greater weight to this part of the filtration, as the features observed here will in general have a greater weight in determining the values of $S_1(\varepsilon)$. Furthermore, the degree of this effect can be controlled with the parameter k. Here we show the contour lines for $k = 1.1$ and $k = 1.5$, and we can see that larger values of k will cause the slope of these lines to increase more rapidly for increasing values of ε, thus making the sifting out of features in the latter part of the filtration more pronounced. Using different contours thus serves a double purpose: First, it allows us to assign feature importance in a more principled manner than by simply using bar length as an absolute measure. Second, it also allows us to place focus on specific subsets of the filtration parameter range, or in other words, on specific spatial scales.

Indeed, Figure 6.15 also shows the H_1 stable rank invariants obtained by applying these three contours to the three barcodes from Figure 6.6. The three stable ranks obtained from using the standard contour are virtually indistinguishable from each other. When using the exponential contours, however, a significant difference emerges for the cloud field from 2013/04/22, especially clear for the exponential contour with $k = 1.1$. This difference is best understood by comparing the three H_1 barcodes. In particular, the 2013/04/22 barcode does show that the persistence

intervals born in the first part of the filtration, $[0, 0.25]$, have in general larger persistence values than the intervals born in this same region for the other two barcodes. More importantly, these are not the longest intervals in the entire barcode, yet here we can see how using a different contour has made them significantly more expressive.

In this case, the existence of these features also has a geometrical interpretation. When comparing the cloud field of 2013/04/22 with that of 2013/07/17, even though both have very similar cloud cover (4.7 % and 4.6%, respectively), we can see in the latter field that the smaller connected components are mostly grouped together in few clusters. Thus, there is mostly empty space between the larger components (i.e., the larger clouds), space which then allows large cycles to form in the VR filtration. In the former field, however, these small components are scattered more sparsely throughout the domain, forming spatial patterns of their own. It is these patterns, insofar as they allow for the formation of cycles in the VR filtration, that the longer persistence intervals in the early part of the filtration range represent.

We can take this idea further by using contours with more detailed sifting properties, and applying them to the full cloud field dataset. To this end, we will use the two contours C_1 and C_2 shown in Figure 6.16. Following the same logic as before, C_1 will give more weight to features in the *latter* portion of the filtration, whereas C_2 will do so for those features in the first part. These contours will be used to induce different clusterings of the cloud fields, based on the spatial distribution of clouds as characterized by the H_1 stable rank.

We again obtain 10 random samples drawn from each of the 360 cloud fields, this time using connected component sampling. That is, we draw a random sample of points from each connected component, of size equal to 5% of the original component size, while guaranteeing that at least one point from each component is always drawn. For each point sample, we compute its $H-1$ stable rank with respect to the standard contour, and the two contours C_1 and C_2. Given a contour function c, we can assign to each cloud field the mean normalized stable rank $S^*_{1,c}$, where the

averaging is performed over the 10 point samples drawn from the field.

Starting from the dataset of 360 cloud fields, and removing those without sufficient H_1 features, we are left with 254 fields. For a choice of contour c we can thus compute a similarity measure between cloud fields by computing the functional distance between their stable ranks. For this experiment we used the L_1 and L_2 metrics, defined by

$$d_{L_p}(f, g) = \left(\int_0^\infty |f(t) - g(t)|^p \, dt \right)^{1/p}, \tag{6.30}$$

and the interleaving metric, defined by

$$d_{\bowtie}(f, g) = \inf\{\varepsilon \in \mathbb{R} \mid f(t) \geqslant g(t + \varepsilon) \text{ and } g(t) \geqslant f(t + \varepsilon), t \in \mathbb{R}\}. \tag{6.31}$$

The pairwise distance matrices for all stable ranks can then be used as input to a hierarchical clustering algorithm. The final cluster assignment is then determined by visually inspecting the resulting dendrograms. We thus obtain one cluster assignment for each combination of contour function and functional metric. The best results were obtained with the interleaving distance, and the C_1, C_2 contours. These were found to produce cluster assignments which reflect different characteristics of the spatial distributions in the cloud fields contained in each cluster.

An example of diverging morphological characteristics educed from the $C_{1,2}$ clustering schemes is shown in Figure 6.17: (a) and (b) are representatives of two different clusters obtained by using contour C_1, while (c) and (d) stem from clusters in the C_2 classification. As expected from the shape of the contours, the classifications they induce are influenced by different spatial scales. Namely, despite the fact that cloud fields a) and b) have identical cloud cover, and their I_{org} values are very similar, the large-scale distribution of the individual clouds is significantly different for both. In similar fashion, both c) and d) are indistinguishable in terms of cloud cover and I_{org}, yet are distinguished by the spatial pattern of smaller structures, even if the large-scale distribution is similar in both.

This study of cloud fields shows that the use of stable rank functions as descriptors for spatial distributions can reveal morphological properties which other methods cannot. Crucially, the possibility of changing the contour enriches the scope for determining such properties. Future investigation in this direction will address questions such as: what the optimal contour is for a given problem, what these methods can reveal about the temporal evolution of cloud formation, and how the homological properties thus discovered can be related to different physical variables in the system. From a more general data analysis point of view, the optimal selection of contour functions is crucial for making our pipeline an fully end-to-end machine learning approach.

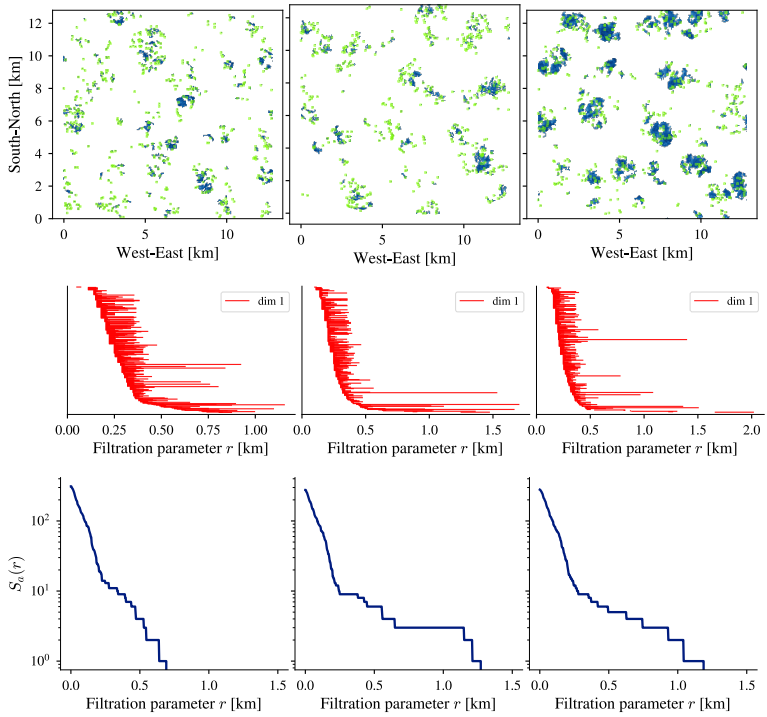


Figure 6.6: Top: Three cloud fields (in blue) with points drawn at random from them (in green). Each cloud field corresponds to a different simulation day at 12:00h. Middle: H_1 barcodes for the point samples in the cloud fields above. Bottom: stable rank function for the H_1 barcodes above.

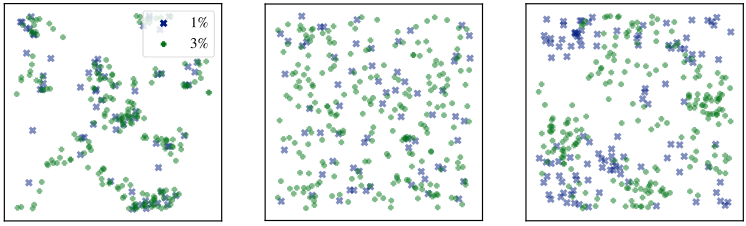


Figure 6.7: Left: Point sets sampled from a cloud field, with sampling rate of 1% (blue) and 3% (green). Middle: Realizations of a homogeneous Poisson point process, with the same expected number of points as the cloud field samples. Right: Realizations from a Thomas point process, with expected number of points also matching the cloud samples.

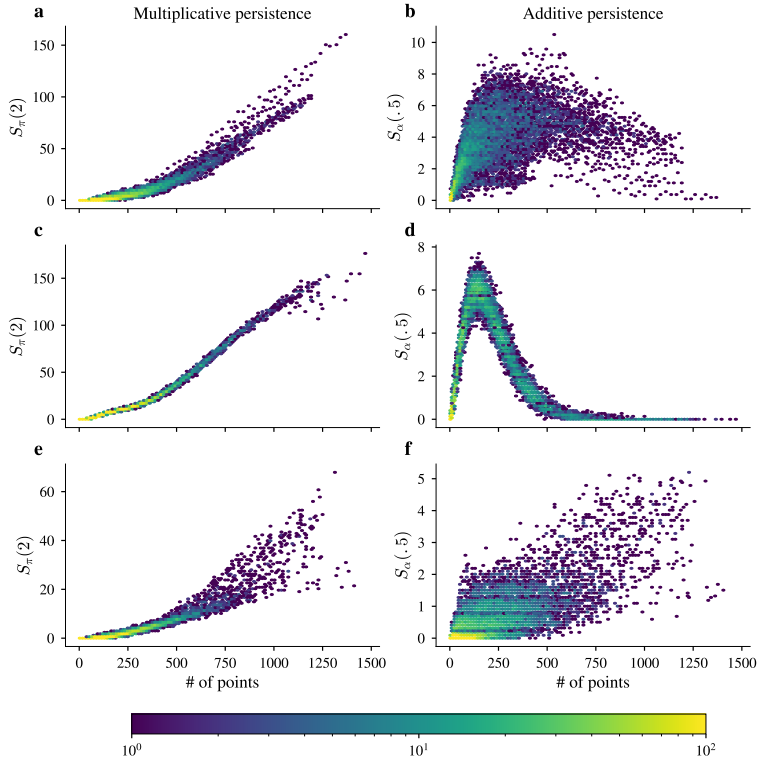


Figure 6.8: Two-dimensional histogram showing the mean number of points per sample and mean value of $S^\pi(2)$ (**a**) and $S^\alpha(0.5)$ (**b**). The average is computed over the 10 samples per cloud field and sample ratio. Also shown are the equivalent histograms for the Poisson point process samples (**c, d**) and the Thomas point process samples (**e, f**).

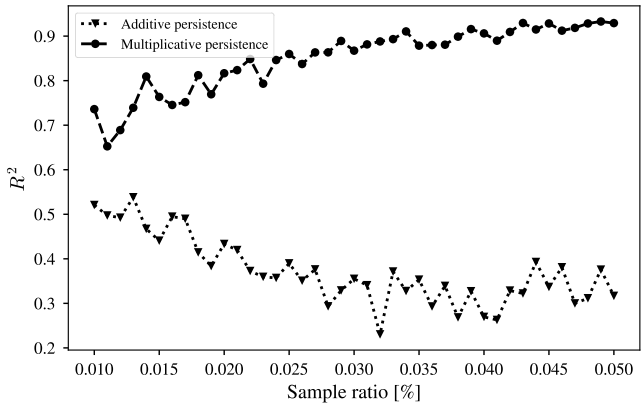


Figure 6.9: R^2 scores for the linear models of cloud cover using additive persistence features ($S^{\alpha}(r)$) and multiplicative persistence features ($S^{\pi}(r)$).

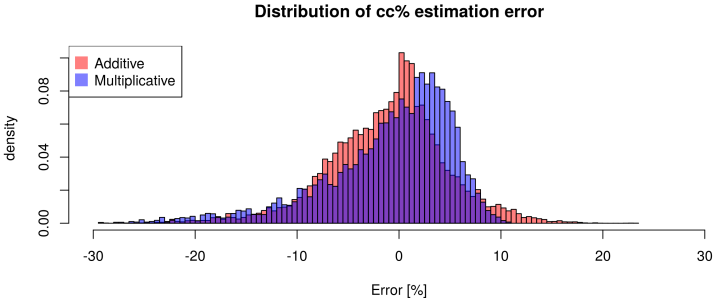


Figure 6.10: Distribution of estimation errors accumulated in 50 runs of 10-fold cross-validation, for two linear models using values of S^{α} and S^{π} as explanatory variables for cloud cover. Figure from Licón-Saláiz *et al.* [2018].

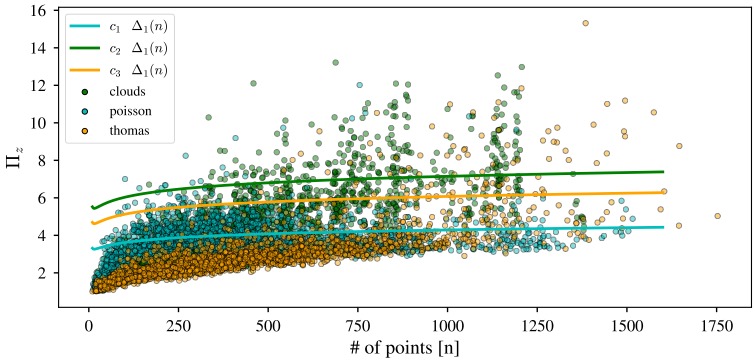


Figure 6.11: Scatterplots of Π_z against pointcloud size, for all samples with sample ratio $s = 0.05$, realizations of a Poisson point process, and a Thomas point process ($\sigma^2 = 0.05$). Also shown is the function $\Delta_1(n)$ with three different scaling constants c_i. Figure from Licón-Saláiz *et al.* [2018].

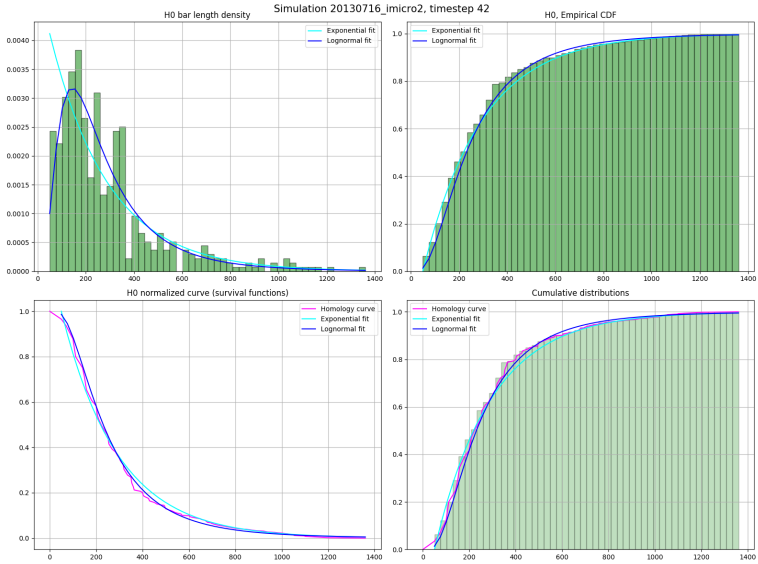


Figure 6.12: Homological density estimation for the H_0 persistent homology obtained from a point set sampled from one cloud field. Top left: empirical PDF of bar length. Top right: Empirical CDF. Bottom left: normalized stable rank for H_0. Bottom right: Empirical CDF with $1 - S_0^*(r)$.

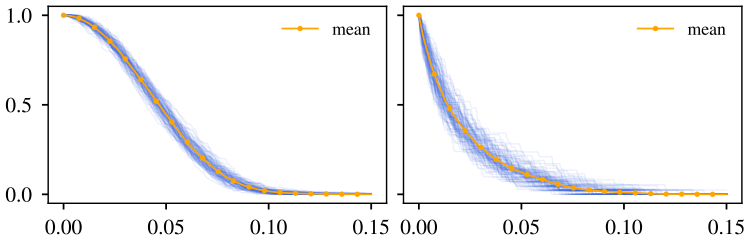


Figure 6.13: Stable rank functions obtained from 100 realizations of a homogeneous Poisson point process with $\lambda = 100$. Left: S_0^*. Right: S_1^*. Figure from Riihimäki and Licón-Saláiz [2019].

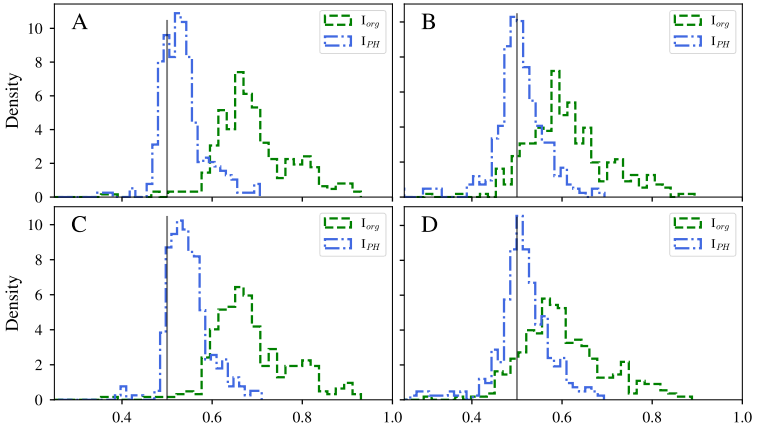


Figure 6.14: Density histograms of the I_{org} index and I_{PH} (Equation 6.28) for 360 distinct cloud fields. **A**: ql max, **B**: ql max removing cloud structures with size smaller than 3 cells, **C**: Geometric centroids, **D**: Geometric centroids removing cloud structures with size smaller than 3 cells. Figure from Riihimäki and Licón-Saláiz [2019].

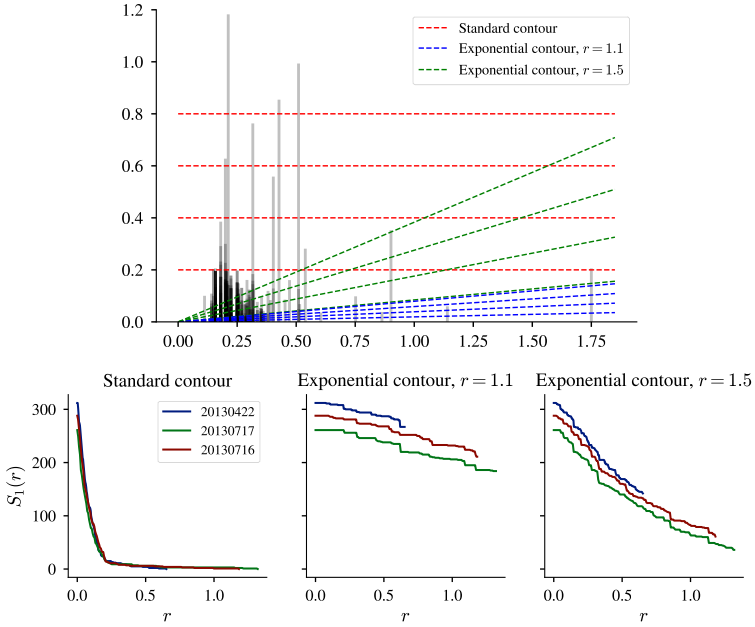


Figure 6.15: Top: H_1 stem plot representation of the barcode in Figure 6.6, bottom left. Also shown are level sets of the standard contour function, and for two different exponential contours. Bottom: Stable rank functions obtained from the standard contour (left), and from the two forms of the exponential contour illustrated in the stem plot (center, right).

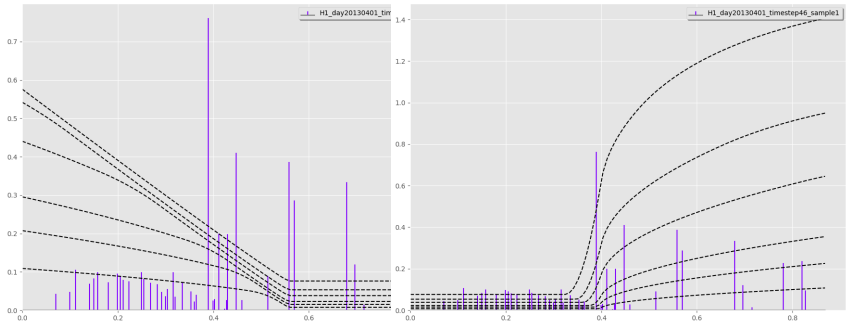


Figure 6.16: Examples of different persistence contours used in the morphological clustering of cloud fields. Left: contour 1, giving more weight to the last part of the filtration interval. Right: contour 2, giving more weight to the first part. (Figure credit: H. Riihimäki, from Riihimäki and Licón-Saláiz [2019].)

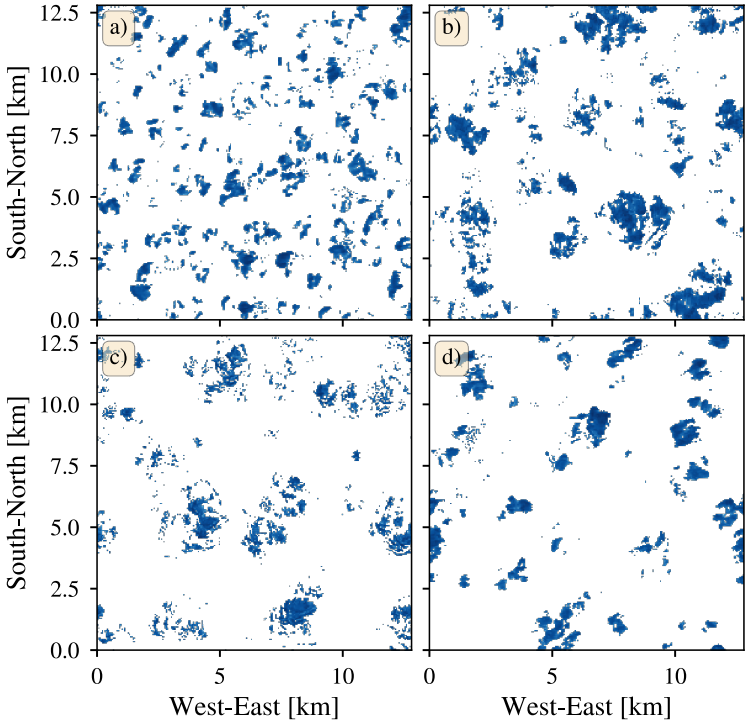


Figure 6.17: Cloud fields which are classified into different clusters, according to the methodology described in the text. We use the H_1 stable ranks and the interleaving metric to compute the distances between them. a) and b) are classified using contour C_1, and have I_{org} values of 0.45 and 0.53 respectively. Cloud cover is similar at 14% for both. c) and d) are classified with C_2, and have I_{org} values of 0.65 and 0.63 respectively, and cloud cover for both is 9.2%. Figure from Riihimäki and Licón-Saláiz [2019].

7

FINAL DISCUSSION AND OUTLOOK

7.1 DISCUSSION

The work presented in this dissertation has been focused around one central research question: to develop a mathematical representation, by means of topological descriptors, of spatial patterns in the planetary boundary layer (PBL), and their interaction with the land surface. To this end, we have used data from numerical simulations of the PBL and computed topological invariants associated to these objects by considering diverse geometrical objects constructed from the data. We have explored the relationships between the numerical values of these invariants and the underlying dynamics in the PBL, and we have found that these topological quantities carry physically meaningful information, which in some cases cannot be obtained from classical, i.e. non-topological, methods.

In Chapter 4, we have analyzed the structural properties of turbulent flow in the PBL through the Betti numbers of two-dimensional slabs of vertical wind velocity. These numbers, which in a sense capture the different patterns of interspersion between up- and downdrafts, can discriminate between data from simulations with land surface patterns that have the same relative composition of land types, but different levels of heterogeneity. As a first approach, their vertical profiles are considered, and a comparison is drawn with global measurements of the flow, such as domain averages, which cannot detect these differences (Section 4.3.1). When considering the temporal evolution of these Betti numbers, the effect of land surface heterogeneity is also dis-

cernible, as are other CBL features such as the inversion crossing time (Section 4.3.3). When considering all values of the Betti numbers as they vary across height and time, a significant relationship is found between them and the different subregions of the PBL. This is shown by training an unsupervised classifier on the Betti number data, which produces remarkable agreement with the partitioning into subregions obtained from bulk measurements of the flow (Section 4.4). We also show that this technique can reveal other model aspects which have an impact on flow morphology, such as the presence of wind shear in the surface layer, or the unresolved small-scale entrainment in the LES model used here. This approach thus allows us to capture, in a very precise and compact form, the qualitative differences observed between flow patterns that arise in the various PBL regimes.

In Chapter 5 we specialize the analysis to the spatial connectivity of turbulent flow, with the goal of representing its hierarchical organization. To this end, we use an extended version of the Union-Find data structure, which allows us to quantify the self-similar scaling of updrafts in the radiatively driven PBL. We find this pattern of self-similarity to be in agreement with the expected value from Kolmogorov's self-similar energy spectrum in the inertial subrange of turbulent motion. This is especially true for the fully-turbulent mixing layer, whereas the surface layer scaling patterns show the effect of the land surface pattern (Section 5.2.3). Comparison with the zeroth Betti number, β_0, which counts the number of connected components, shows that the effect of surface heterogeneity is greater on β_0 than on the scaling pattern (Section 5.2.5). We then extend this approach to three-dimensional space by introducing another topological invariant, the merge tree, and show how we can isolate the region of space associated to the dominant coherent structure, a convective plume (Section 5.3). Some significant advantages of this approach over the classical spectral methods are that no smoothness or periodicity assumptions are necessary, besides being able to connect these structures directly to other features in physical space. We use this fact to quantify the phenomenon of plume coalescence in the transition between surface and mixing layers. It is also possible to identify the evolution of this process throughout the PBL daily cycle, and

to quantify the effect of surface heterogeneity on the spatial depth of the plume-merging layer. Generally speaking, more heterogeneous surfaces will tend to produce larger numbers of individual, disconnected plumes in the PBL surface layer, and these will require a larger vertical space to merge into a coherent convective plume. The tree representation of this process also allows us to identify those surface cells which underlie the points of origin of this coherent structure at its base. With this we can measure the relative contribution of each land type to the formation and sustenance of the plume, and we again find that surface heterogeneity plays a role here: even if the total land area covered by a given land type is the same across simulations, the likelihood of its cells being connected to the plume strongly depends on the overall surface heterogeneity.

Finally, in Chapter 6 we consider a different topological invariant, persistent homology, and show how it can be interpreted as a multiscale descriptor of spatial patterns. To this end, we use the stable rank function as a measure of "homological density" over different spatial scales. This invariant is shown to be better at distinguishing between diverse spatial point patterns, both regular and random, than other first- and second-order spatial statistics, such as nearest-neighbor distributions and the Ripley's K and L functions (Section 6.3). After establishing this, we use the same invariant to analyze the spatial distribution of shallow cumulus clouds. The first step in this direction is to find the relationship between pointwise values of the stable rank and the cloud cover of a given field (Section 6.5). Having established this relationship, we proceed by comparing the stable rank values obtained from cloud fields to those obtained from spatial point processes, in order to asses the spatial randomness of the former. This leads us to draw conclusions regarding the cloud size distribution which agree with recent studies on the subject, which have used the same data (Section 6.5.2). After this, we then define a homology-based counterpart to the I_{org} index for spatial organization, and use it to quantify the degree of randomness or organization of cloud cores (Section 6.6). We also show the topological version to be more robust against noise than I_{org}. Finally, we use the persistence contour formalism to produce a morphological classification of cloud

fields by focusing on spatial features at different scales, since the contour functions provide a weighting scheme for these scales (Section 6.7). This provides a framework for classifying spatial patterns based on the features they present over a specific range of scales.

We set out to answer a specific research question, namely: **can topology be leveraged to represent PBL patterns, and if so, does it offer any new information not provided by classical methods?** Throughout this dissertation we have presented diverse topological invariants which can be readily and efficiently computed from numerical simulation data. We have found significant relationships between the numerical values of these invariants and the dynamical state of the PBL, and with its response in interaction with complex land surface patterns. In so doing, we have shown that these invariants are more informative when used as statistical descriptors than are bulk averaging or functional summaries of spatial distributions, such as spectral analysis or the I_{org} index. In some cases, the topological descriptors recover the information provided by spectral methods, without discarding the additional information available in physical space. With all this in mind, we can now give a positive answer to the research question, as our investigations have shown the expressiveness and relevance of topological information in analyzing and understanding the quasi-chaotic dynamics in the turbulent PBL.

7.2 OUTLOOK

The field of TDA is expanding rapidly, both in regards to the theory behind it and in concrete applications. In that sense, this dissertation has been only the first approach to using techniques from TDA in solving problems in atmospheric science. Despite (or probably, because of) the success encountered in this endeavour, many new pathways leading beyond our current vantage point stand now open. These would involve a deeper exploration of the topological aspects of boundary layer dynamics, and the relation of these with the corresponding physical processes. Here we enumerate but a few of these pathways.

1. A clear limitation in the Betti-number based methodology presented in Chapter 4 is its dependence on the selection of a threshold value. It would be possible to sidestep this limitation by using persistent homology, specifically the stable rank invariant computed for the two-dimensional scalar fields of vertical wind velocity, instead of the Betti numbers. The Betti numbers are simply the value of the corresponding stable rank functions at one point in their domains.

2. The methodology of Chapters 4 and 5 can also be extended if we consider not only connectivity in space, but also in time. Coherent structures, after all, are defined as having significant extent in both space and time.

3. Another extension of this methodology is adapting it to work on observational data. Indeed, some of the related studies recounted in Section 4.1 present results obtained from both simulations and observations.

4. Developing a method for the optimal selection of persistence contours, for example by optimizing over a space of basis functions. This would also be applicable beyond the concrete problems studied here.

5. Use persistence contours to study the effect of land-surface features, such as heterogeneity or topography, on the spatial distribution of shallow cumulus clouds.

6. Develop a null hypothesis testing framework for spatial randomness, using the persistence contour formalism.

A

LIST OF SYMBOLS

$[b_\alpha, d_\alpha)$	Persistence interval of a homology class α.	38
$B_p(K)$	Group of p-boundaries on a simplicial complex K.	30
$b(\mathbf{X}, t)$	Vector field of buoyancy.	14
$\beta_0^{\pm}$	Number of connected components of the up- and downdraft domains.	47
$\beta_1^{\pm}$	Number of loops in the up- and downdraft domains.	47
$\beta_p^{i,j}$	Persistent Betti numbers.	38
β_p	Betti number of the homology group $H_p(K)$.	32
$C_p(K)$	The group of $p-$chains defined on a simplicial complex K.	28
$C(\mathcal{P}, \varepsilon)$	Čech complex for a set of points $\mathcal{P}$ and $\varepsilon > 0$	36, 37
$\partial\sigma$	Boundary of a simplex.	26
∂_k	Boundary homomorphism defined on the k-chain group of a simplicial complex.	29
$E(k)$	Energy spectrum function.	15
ε	Rate of dissipation of kinetic energy in a fluid.	15

η	Kolmogorov length scale of a fluid, i.e. the scale of its smallest eddies.	15
$\langle Q \rangle$	Expectation or average of a variable Q.	16
$H_p^{i,j}$	p-th persistent homology groups for a filtered complex.	38
κ	Molecular diffusivity of a fluid.	14
$\mathcal{L}$	Characteristic length scale of a fluid.	15
L	Integral scale of a fluid, i.e. the scale of its largest eddies.	15
$\ell^{\pm}$	Log-quotient of Betti numbers β_1 and β_0.	60
$\mathcal{M}^-$	Set of domain points covered by downdrafts.	46
$\mathcal{M}^+$	Set of domain grid points covered by updrafts.	44
ν	Kinematic viscosity of a fluid.	14
Ω	Computational domain of a numerical simulation	94
Ω_t	Subset of Ω with constant time coordinate t	94
Re	Reynolds number of a fluid.	15
$R(\mathcal{P}, \varepsilon)$	Vietoris-Rips complex for a set of points $\mathcal{P}$ and $\varepsilon > 0$	37
S_k	Stable rank invariant for order k persistent homology.	106
S_k^*	Normalized stable rank invariant for order k persistent homology.	120

t_η	Kolmogorov time scale.	16
t_L	Large-eddy turnover time.	16
$u(\mathbf{X}, t)$	Scalar field of wind velocity in the direction of x (horizontal).	15
$[u_0, u_1, \ldots, u_k]$	$k-$simplex spanned by the points u_i.	26
$\mathcal{U}$	Characteristic velocity of a fluid.	15
$\mathbf{u}(\mathbf{X}, t)$	Vector field of wind velocity.	14
$v(\mathbf{X}, t)$	Scalar field of wind velocity in the direction of y (horizontal).	15
$\mathbb{Z}_2$	The group of integers modulo 2.	29
$z_{i,0}$	Zero-crossing height.	61
$z_{i,f}$	Flux-based height.	61
$z_{i,v}$	Variance-based height.	61
$Z_p(K)$	Group of p-cycles on a simplicial complex K.	30
z_i	Boundary layer height (depth), defined as the average height of the inversion layer base.	12

BIBLIOGRAPHY

Adams, H., Chepushtanova, S., Emerson, T., Hanson, E., Kirby, M., Motta, F., Neville, R., Peterson, C., Shipman, P., and Ziegelmeier, L. (2017). Persistence Images: A Stable Vector Representation of Persistent Homology. *Journal of Machine Learning Research*, **18**, 1–35.

Adler, R. J., Bobrowski, O., Borman, M. S., Subag, E., and Weinberger, S. (2010). Persistent homology for random fields and complexes. In I. M. Berger, James O. and Cai, T. Tony and Johnstone, editor, *Borrowing strength: theory powering applications - a Festschrift for Lawrence D. Brown*, volume 6, pages 124–143. Institute of Mathematical Statistics.

Adrian, R. J. (2007). Hairpin vortex organization in wall turbulence. *Physics of Fluids*, **19**(4), 041301.

Agee, E. M. and Chen, T. S. and Dowell, K. E. (1973). A review of mesoscale cellular convection. *Bulletin of the American Meteorological Society*, **54**(10), 1004–1012.

Alstott, J., Bullmore, E., and Plenz, D. (2014). Powerlaw: A python package for analysis of heavy-tailed distributions. *PLoS ONE*, **9**(1).

Antonia, R. A. (1980). The organized motion in a turbulent boundary layer. In *7th Australasian Conference on Hydraulics and Fluid Mechanics 1980*, pages 155–162. Institution of Engineers, Australia.

Bauer, U. (2017). Ripser: a lean C++ code for the computation of Vietoris-Rips persistence barcodes. Available at `https://github.com/Ripser/ripser`.

Bendich, P., Marron, J. S., Miller, E., Pieloch, A., and Skwerer, S. (2016). Persistent homology analysis of brain artery trees. *Annals of Applied Statistics*, **10**(1), 198–218.

Bizon, C., Werne, J., Predtechensky, a. a., Julien, K., McCormick, W. D., Swift, J. B., and Swinney, H. L. (1997). Plume dynamics in quasi-2D turbulent convection. *Chaos*, **7**(1), 107–124.

Bobrowski, O., Kahle, M., and Skraba, P. (2017). Maximally persistent cycles in random geometric complexes. *The Annals of Applied Probability*, **27**(4), 2032–2060.

Bony, S. and Dufresne, J.-L. (2005). Marine boundary layer clouds at the heart of tropical cloud feedback uncertainties in climate models. *Geophysical Research Letters*, **32**(20), 2–5.

Bubenik, P. (2015). Statistical topological data analysis using persistence landscapes. *Journal of Machine Learning Research*, **16**, 77–102.

Carlsson, G. (2009). Topology and data. *Bulletin of the American Mathematical Society*, **46**(2), 255–308.

Carlsson, G. and Zomorodian, A. (2009). The theory of multidimensional persistence. *Discrete and Computational Geometry*, **42**(1), 71–93.

Carreras, B. A., Llerena, I., Garcia, L., and Calvo, I. (2008). Topological characterization of flow structures in resistive pressure-gradient-driven turbulence. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, **78**(6), 1–9.

Cerisier, P., Rahal, S., and Rivier, N. (1996). Topological correlations in Bénard-Marangoni convective structures. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, **54**(5), 5086–5094.

Chacholski, W. and Riihimäki, H. (2020). Metrics and stabilization in one parameter persistence. *SIAM J. Appl. Algebra Geom.*, **4**(1), 69–98.

Chan, J. M., Carlsson, G., and Rabadan, R. (2013). Topology of viral evolution. *Proceedings of the National Academy of Sciences*, **110**(46), 18566–18571.

Clauset, A., Young, M., and Gleditsch, K. S. (2007). On the Frequency of Severe Terrorist Events. *Journal of Conflict Resolution*, **51**(1), 58–87.

Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, **51**(4), 661–703.

Cohen-Steiner, D., Edelsbrunner, H., and Harer, J. (2007). Stability of Persistence Diagrams. *Discrete & Computational Geometry*, **37**(1), 103–120.

de Silva, V. and Ghrist, R. (2007). Coverage in sensor networks via persistent homology. *Algebraic and Geometric Topology*, **7**(1), 339–358.

Dłotko, P. and Wanner, T. (2016). Topological microstructure analysis using persistence landscapes. *Physica D: Nonlinear Phenomena*, **334**, 60–81.

Edelsbrunner, H. and Harer, J. (2010). *Computational Topology: An Introduction*. American Mathematical Society.

Edelsbrunner, H., Letscher, D., and Zomorodian, A. (2002). Topological Persistence and Simplification. *Discrete & Computational Geometry*, **28**, 511–533.

Ellis, S. P. and Klein, A. (2014). Describing high-order statistical dependence using "concurrence topology," with application to functional MRI brain data. *Homology, Homotopy and Applications*, **16**(1), 245–264.

Escolar, E. G. and Hiraoka, Y. (2014). Computing Optimal Cycles of Homology Groups. In R. Nishii, S.-I. Ei, M. Koiso, and S. Saito, editors, *A Mathematical Approach to Research Problems of Science and Technology*, pages 101–118. Springer Japan, Tokyo.

Escolar, E. G. and Hiraoka, Y. (2016). Optimal Cycles for Persistent Homology via Linear Programming. In K. Fujisawa, Y. Shinano, and H. Waki, editors, *Optimization in the Real World*, pages 79–96. Springer Japan, Tokyo.

Euler, L. (1741). Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae*, **8**, 128–140.

Gameiro, M., Mischaikow, K., and Kalies, W. (2004). Topological characterization of spatial-temporal chaos. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, **70**(3 2), 9–12.

Gameiro, M., Mischaikow, K., and Wanner, T. (2005). Evolution of pattern complexity in the Cahn-Hilliard theory of phase separation. *Acta Materialia*, **53**(3), 693–704.

Garcia, J. R. and Mellado, J. P. (2014). The Two-Layer Structure of the Entrainment Zone in the Convective Boundary Layer. *Journal of the Atmospheric Sciences*, **71**(6), 1935–1955.

Garcia, L., Carreras, B. A., Llerena, I., and Calvo, I. (2009). Topological characterization of the transition from laminar regime to fully developed turbulence in the resistive pressure-gradient-driven turbulence model. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, **80**(4), 1–11.

Garcia-Carreras, L. and Parker, D. J. (2011). What is the Mechanism for the Modification of Convective Cloud Distributions by Land Surface–Induced Flows? *Journal of the Atmospheric Sciences*, **68**(3), 619–634.

Gelfgat, A. Y. (1999). Different Modes of Rayleigh–Bénard Instability in Two- and Three-Dimensional Rectangular Enclosures. *Journal of Computational Physics*, **156**(2), 300–324.

Gentine, P., Holtslag, A. A. M., D'Andrea, F., and Ek, M. (2013). Surface and Atmospheric Controls on the Onset of Moist Convection over Land. *Journal of Hydrometeorology*, **14**(5), 1443–1462.

Goehring, L. (2013). Pattern formation in the geosciences. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, **371**(2004).

Golovin, A., Nepomnyashchy, A., and Pismen, L. (1995). Pattern formation in large-scale Marangoni convection with deformable interface. *Physica D: Nonlinear Phenomena*, **81**(1-2), 117–147.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2nd edition.

Heinle, A., Macke, A., and Srivastav, A. (2010). Automatic cloud classification of whole sky images. *Atmospheric Measurement Techniques*, **3**(3), 557–567.

Henderson, R. D. (1997). Nonlinear dynamics and pattern formation in turbulent wake transition. *Journal of Fluid Mechanics*, **352**, 65–112.

Heus, T., van Heerwaarden, C. C., Jonker, H. J. J., Pier Siebesma, A., Axelsen, S., van den Dries, K., Geoffroy, O., Moene, A. F., Pino, D., de Roode, S. R., and Others (2010). Formulation of the Dutch Atmospheric Large-Eddy Simulation (DALES) and overview of its applications. *Geoscientific Model Development*, **3**, 415–444.

Hopcroft, J. E. and Ullman, J. D. (1973). Set merging algorithms. *SIAM Journal on Computing*, **2**(4), 294–303.

Hopf, E. (1948). A mathematical example displaying features of turbulence. *Communications on Pure and Applied Mathematics*, **1**(4), 303–322.

Jiménez, J. (2012). Cascades in Wall-Bounded Turbulence. *Annual Review of Fluid Mechanics*, **44**(1), 27–45.

Kaczynski, T., Mischaikow, K., and Mrozek, M. (2004). *Computational Homology*. Springer-Verlag New York, 1st edition.

Kaimal, J. C. and Finnigan, J. J. (1994). *Atmospheric Boundary Layer Flows - Their Structure and Measurement*. Oxford University Press, New York, NY, USA, 1st edition.

Kelkar, D. A. and Chattopadhyay, A. (2007). The gramicidin ion channel: A model membrane protein. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, **1768**(9), 2011–2025.

Koch, J., Mendiguren, G., Mariethoz, G., and Stisen, S. (2017). Spatial Sensitivity Analysis of Simulated Land Surface Patterns in a

Catchment Model Using a Set of Innovative Spatial Performance Metrics. *Journal of Hydrometeorology*, **18**(4), 1121–1142.

Kolmogorov, A. N. (1941a). Dissipation of Energy in the Locally Isotropic Turbulence. *Dokl. Akad. Nauk SSSR*, **32**(1).

Kolmogorov, A. N. (1941b). The Local Structure of Turbulence in Incompressible Viscous Fluid for Very Large Reynolds Numbers. *Dokl. Akad. Nauk SSSR*, **30**(1), 9–13.

Kondrashov, A., Sboev, I., and Dunaev, P. (2016). Evolution of convective plumes adjacent to localized heat sources of various shapes. *International Journal of Heat and Mass Transfer*, **103**, 298–304.

Krishan, K., Kurtuldu, H., Schatz, M. F., Gameiro, M., Mischaikow, K., and Madruga, S. (2007). Homology and symmetry breaking in Rayleigh-Bénard convection: Experiments and simulations. *Physics of Fluids*, **19**(11), 1–6.

Kusano, G., Fukumizu, K., and Hiraoka, Y. (2018). Kernel method for persistence diagrams via kernel embedding and weight factor. *Journal of Machine Learning Research*, **18**(189), 1–41.

Landau, L. D. (1944). On the problem of turbulence. *Dokl. Akad. Nauk SSSR*, **44**, 339–342.

Lee, N., Schultz, W., and Boyd, J. (1989). Stability of fluid in a rectangular enclosure by spectral method. *International Journal of Heat and Mass Transfer*, **32**(3), 513–520.

Lee, Y., Barthel, S. D., Dłotko, P., Moosavi, S. M., Hess, K., and Smit, B. (2017). Pore-geometry recognition: on the importance of quantifying similarity in nanoporous materials. *Nature communications*, **8**, 15396.

Licón-Saláiz, J. and Ansorge, C. (2019). Topological descriptors of spatial coherence in a convective boundary layer. In *International Workshop on Applications of Topological Data Analysis*. Available at `https://sites.google.com/view/atda2019`.

Licón-Saláiz, J., Riihimäki, H., and van Laar, T. W. (2018). Topological characterization of shallow cumulus cloud fields using persistent homology. In C. Chen, D. Cooley, J. Runge, and E. Szekely, editors, *Proceedings of the 8th International Workshop on Climate Informatics*. National Center for Atmospheric Research. Available at `https://opensky.ucar.edu/islandora/object/technotes:571`.

Licón-Saláiz, J., Ansorge, C., Shao, Y., and Kunoth, A. (2020). The structure of the convective boundary layer as deduced from topological invariants. *Boundary-Layer Meteorology*, **176**, 1–12.

Liu, S., Shao, Y., Kunoth, A., and Simmer, C. (2017). Impact of surface-heterogeneity on atmosphere and land-surface interactions. *Environmental Modelling and Software*, **88**, 35–47.

MacPherson, R. and Schweinhart, B. (2012). Measuring shape with topology. *Journal of Mathematical Physics*, **53**(7), 073516.

Máté, G., Hofmann, A., Wenzel, N., and Heermann, D. W. (2014). A topological similarity measure for proteins. *Biochimica et Biophysica Acta - Biomembranes*, **1838**(4), 1180–1190.

McCord, M. C. (1967). Homotopy Type Comparison of a Space with Complexes Associated with its Open Covers. *Proceedings of the American Mathematical Society*, **18**(4), 705–708.

Mellado, J. P., van Heerwaarden, C. C., and Garcia, J. R. (2016). Near-Surface Effects of Free Atmosphere Stratification in Free Convection. *Boundary-Layer Meteorology*, **159**(1), 69–95.

Meyer, C. W., Ahlers, G., and Cannell, D. S. (1987). Initial stages of pattern formation in Rayleigh-Bénard convection. *Physical Review Letters*, **59**(14), 1577–1580.

Mischaikow, K., Kokubu, H., Mrozek, M., and Pilarczyk, P. (2019). CHomP - Computational Homology Project. Available at `http://chomp.rutgers.edu/`.

Mitzenmacher, M. (2004). A brief history of lognormal and power law distributions. *Internet Mathematics*, **1**(2), 226–251.

Mizushima, J. (1994). Mechanism of the Pattern Formation in Rayleigh-Bénard Convection. *Journal of the Physical Society of Japan*, **63**(1), 101–110.

Mizushima, J. (1995). Onset of the Thermal Convection in a Finite Two-Dimensional Box. *Journal of the Physical Society of Japan*, **64**(7), 2420–2432.

Morris, S. W., Bodenschatz, E., Cannell, D. S., and Ahlers, G. (1993). Spiral Defect Chaos in Large Aspect Ratio Rayleigh-Bénard Convection. *Physical Review Letters*, **71**(13).

Munkres, J. R. (1984). *Elements of Algebraic Topology*. Perseus Books, 7th edition.

Munkres, J. R. (2000). *Topology*. Featured Titles for Topology Series. Prentice Hall, Incorporated, 2nd edition.

Muszynski, G., Kashinath, K., Kurlin, V., and Wehner, M. (2019). Topological Data Analysis and Machine Learning for Recognizing Atmospheric River Patterns in Large Climate Datasets. *Geoscientific Model Development Discussions*, **12**(2), 613—-628.

Nanda, V. and Sazdanović, R. (2014). Simplicial Models and Topological Inference in Biological Systems. In N. Jonoska and M. Saito, editors, *Discrete and Topological Models in Molecular Biology*, pages 109–141, Berlin, Heidelberg. Springer Berlin Heidelberg.

Obayashi, I. (2018). Volume-Optimal Cycle: Tightest Representative Cycle of a Generator in Persistent Homology. *SIAM Journal on Applied Algebra and Geometry*, **2**(4), 508–534.

Obukhov, A. M. (1941). On the distribution of energy in the spectrum of turbulent flow. *Dokl. Akad. Nauk SSSR*, **32**(1), 22–24.

Pankiewicz, G. S. (1995). Pattern recognition techniques for the identification of cloud and cloud systems. *Meteorological Applications*, **2**(3), 257–271.

Pearson, D. A., Bradley, R. M., Motta, F. C., and Shipman, P. D. (2015). Producing nanodot arrays with improved hexagonal order by patterning surfaces before ion sputtering. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, **92**(6), 062401.

Pedregosa, F., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., and Passos, A. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

Poincaré, H. (1895). Analysis situs. *Journal de l'École polytechnique*, **1**, 1–123.

Pope, S. B. (2000). *Turbulent Flows*. Cambridge University Press.

Powers, D. M. W. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *International Journal of Machine Learning Technology*, **2**(1), 37–63.

Pranav, P., Edelsbrunner, H., van de Weygaert, R., Vegter, G., Kerber, M., Jones, B. J., and Wintraecken, M. (2017). The topology of the cosmic web in terms of persistent Betti numbers. *Monthly Notices of the Royal Astronomical Society*, **465**(4), 4281–4310.

Prandtl, L. (1905). Über Flüssigkeitsbewegung bei sehr kleiner Reibung. In A. Krazer, editor, *Verhandlungen des dritten internationalen Mathematiker-Kongress*, pages 484–491. B. G. Teubner Verlag, Leipzig.

Ray, D. (1986). Variable eddy diffusivities and atmospheric cellular convection. *Boundary-Layer Meteorology*, **36**(1), 117–131.

Reininghaus, J., Huber, S., Bauer, U., and Kwitt, R. (2015). A Stable Multi-Scale Kernel for Topological Machine Learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4741–4748.

Rieck, M., Hohenegger, C., and van Heerwaarden, C. C. (2014). The Influence of Land Surface Heterogeneities on Cloud Size Development. *Monthly Weather Review*, **142**(10), 3830–3846.

Riihimäki, H. and Licón-Saláiz, J. (2019). Metrics for learning in topological persistence. In *International Workshop on Applications of Topological Data Analysis*. Available at `https://sites.google.com/view/atda2019`.

Ripley, B. D. (1976). The second-order analysis of stationary point processes. *Journal of Applied Probability*, **13**(2), 255–266.

Ripley, B. D. (1977). Modelling Spatial Patterns. *Journal of the Royal Statistical Society. Series B (Methodological)*, **39**(2), 172–212.

Ripley, B. D. (1988). *Statistical Inference for Spatial Processes*. Cambridge University Press.

Robins, V. (2002). Computational Topology for Point Data: Betti Numbers of α-Shapes. *Morphology of Condensed Matter*, **600**, 261–274.

Robins, V. and Turner, K. (2016). Principal component analysis of persistent homology rank functions with case studies of spatial point patterns, sphere packing and colloids. *Physica D: Nonlinear Phenomena*, **334**, 99–117.

Rosmond, T. E. (1973). Mesoscale Cellular Convection. *Journal of the Atmospheric Sciences*, **30**(7), 1392–1409.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, **20**, 53–65.

Ruelle, D. and Takens, F. (1971). On the nature of turbulence. *Communications in Mathematical Physics*, **20**(3), 167–192.

Scolamiero, M., Chachólski, W., Lundman, A., Ramanujam, R., and Öberg, S. (2017). Multidimensional Persistence and Noise. *Foundations of Computational Mathematics*, **17**(6), 1367–1406.

Seifert, A. and Heus, T. (2013). Large-eddy simulation of organized precipitating trade wind cumulus clouds. *Atmospheric Chemistry and Physics*, **13**, 5631–5645.

Shah, S. and Bou-Zeid, E. (2014). Very-Large-Scale Motions in the Atmospheric Boundary Layer Educed by Snapshot Proper Orthogonal Decomposition. *Boundary-Layer Meteorology*, **153**(3), 355–387.

Shao, Y., Sogalla, M., Kerschgens, M., and Brücher, W. (2001). Effects of land-surface heterogeneity upon surface fluxes and turbulent conditions. *Meteorology and Atmospheric Physics*, **78**(3), 157–181.

Shao, Y., Liu, S., Schween, J. H., and Crewell, S. (2013). Large-Eddy Atmosphere–Land-Surface Modelling over Heterogeneous Surfaces: Model Development and Comparison with Measurements. *Boundary-Layer Meteorology*, **148**, 333–356.

Strogatz, S. H. (2014). *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. Avalon Publishing, 2nd edition.

Stull, R. B. (1985). A Fair-Weather Cumulus Cloud Classification Scheme for Mixed-Layer Studies. *Journal of Climate and Applied Meteorology*, **24**(1), 49–56.

Stull, R. B. (1988). *An Introduction to Boundary Layer Meteorology*. Springer Netherlands, 1st edition.

Tompkins, A. M. and Semie, A. G. (2017). Organization of tropical convection in low vertical wind shears: Role of updraft entrainment. *Journal of Advances in Modeling Earth Systems*, **9**(2), 1046–1068.

Turner, K., Mileyko, Y., Mukherjee, S., and Harer, J. (2014). Fréchet Means for Distributions of Persistence Diagrams. *Discrete & Computational Geometry*, **52**(1), 44–70.

van Hateren, J. H. and van der Schaaf, A. (1998). Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc. R. Soc. Lond. B*, **265**(1394), 359–366.

van Heerwaarden, C. C., Mellado, J. P., and De Lozar, A. (2014). Scaling Laws for the Heterogeneously Heated Free Convective

Boundary Layer. *Journal of the Atmospheric Sciences*, **71**(11), 3975–4000.

van Laar, T. W., Schemann, V., and Neggers, R. A. J. (2019). Investigating the Diurnal Evolution of the Cloud Size Distribution of Continental Cumulus Convection Using Multiday LES. *Journal of the Atmospheric Sciences*, **76**(3), 729–747.

Vereecken, H., Pachepsky, Y., Simmer, C., Rihani, J., Kunoth, A., Korres, W., Graf, A., Franssen, H.-H., Thiele-Eich, I., and Shao, Y. (2016). On the role of patterns in understanding the functioning of soil-vegetation-atmosphere systems. *Journal of Hydrology*, **542**, 63–86.

Vuong, Q. H. (1989). Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica*, **57**(2), 307–333.

Wu, P., B, C. C., Wang, Y., Zhang, S., Yuan, C., Qian, Z., Metaxas, D., and Axel, L. (2017). Optimal Topological Cycles and Their Application in Cardiac Trabeculae Restoration. In M. Niethammer, M. Styner, S. Aylward, H. Zhu, I. Oguz, P.-T. Yap, and D. Shen, editors, *Information Processing in Medical Imaging: 25th International Conference, IPMI 2017, Boone, NC, USA, June 25-30, 2017, Proceedings*, volume 3565, pages 80–92. Springer International Publishing.

Zomorodian, A. (2012). Topological Data Analysis. In *Proceedings of Symposia in Applied Mathematics 70*, pages 1–39. American Mathematical Society.

Zomorodian, A. and Carlsson, G. (2005). Computing Persistent Homology. *Discrete & Computational Geometry*, **33**(2), 249–274.

COLOPHON

This document was typeset in LaTeX using the the typographical look-and-feel `classicthesis` developed by André Miede and Ivo Pletikosić. All figures were made by the author, unless stated otherwise, using one of: the Python library Matplotlib; Inkscape; or TikZ.

Final Version as of September 23, 2020.

DECLARATION

Ich versichere, daß ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; daß diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; daß sie – abgesehen von unten angegebenen Teilpublikationen – noch nicht veröffentlicht worden ist sowie, daß ich eine solche Veröffentlichung vor Abschluß des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen dieser Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Prof. Dr. Angela Kunoth betreut worden.

Cologne, September 2020

José Luis Licón-Saláiz

Publications

- J. Licón-Saláiz, C. Ansorge, Y. Shao, and A. Kunoth. The structure of the convective boundary layer as deduced from topological invariants. *Boundary-Layer Meteorology*, **176**, 1–12 (2020).
- H. Riihimäki and J. Licón-Saláiz. Metrics for Learning in Topological Persistence. *International Workshop on Applications of Topological Data Analysis.* University of Würzburg, September 2019.
- J. Licón-Saláiz and C. Ansorge. Topological descriptors of spatial coherence in a convective boundary layer. *International Workshop on Applications of Topological Data Analysis.* University of Würzburg, September 2019.
- J. Licón-Saláiz, H. Riihimäki, and T. van Laar. Topological Characterization of Shallow Cumulus Cloud Fields via Persistent Homology. *Proceedings of the 8th Workshop on Climate Informatics.* National Center for Atmospheric Research, October 2018.