Stefan Hergarten
Self-Organized Criticality in Earth Systems

Springer-Verlag Berlin Heidelberg GmbH

Stefan Hergarten

Self-Organized Criticality in Earth Systems

with 100 Figures

Springer

DR. STEFAN HERGARTEN
University of Bonn
Institute of Geology
Nussallee 8
53115 Bonn
Germany

ISBN 978-3-642-07790-6 ISBN 978-3-662-04390-5 (eBook)
DOI 10.1007/978-3-662-04390-5

Library of Congress Cataloging-in-Publication Data
Hergarten, Stefan:, 1964-
Self organized criticality in earth systems / Stefan Hergarten.
p.cm.
Includes bibliographical references and index.

1. Earth sciences—Mathematics. 2. Fractals. 3. Self-organizing systems. 4. Critical phenomena (Physics) I. Title.

Originally published by Springer-Verlag Berlin Heidelberg New York in 2002.
Softcover reprint of the hardcover 1st edition 2002

Camera ready by author
Cover design: E. Kirchner, Heidelberg
Printed on acid-free paper 32/3111/as 5 4 3 2 1

Preface and Acknowledgments

Self–organized criticality (SOC) has become a magic word in various scientific disciplines. In the first ten years after Per Bak and his coworkers presented their seminal idea (Bak et al. 1987), more than 2000 papers on this topic appeared. This number alone should convince the reader that Per Bak in fact started an avalanche, and that knowing at least a little about SOC should become part of education not only in physics, but also in several other disciplines.

Nine years after the idea of SOC was introduced, Per Bak published the first monograph on this topic (Bak 1996). Clearly, it cannot be my aim to compete with this stimulating and colorful book. Like most books written by pioneers in their field, it is unique. On the other hand, a book written for such a wide readership cannot go too deeply into details, so that scientists who have caught fire may ask what to do now.

Mainly focusing on theory and examples from physics, Henrik J. Jensen published a book on SOC (Jensen 1998) soon after Per Bak. As it may be expected, it goes more into details and opens the door towards a deeper understanding. But in return, it resists somewhat more against being read than Per Bak's overview, at least for an earth scientist. The book in your hands will probably do the same, although it focuses less on theory than on phenomena. As earth sciences are a fruitful field for applying the ideas of SOC, the examples are taken from this field. However, most of the methods are independent of the scientific discipline.

Parts of this book derive from lectures and courses held at the Universities of Bonn and Freiburg i. Br., Germany. The first chapters give overviews over the ingredients of SOC – fractals and deterministic chaos. In order to keep the book at a reasonable length, a restriction to a few essential aspects was inevitable. For deeper insights into these topics, the reader can revert to a variety of books (e. g. Mandelbrot 1982; Feder 1988; Turcotte 1997; Sornette 2000). The fifth chapter introduces the basic ideas behind SOC; the rest of the book (even the largest part) focuses on examples from earth sciences.

It is difficult not to forget anyone who helped me on my way towards this book. Clearly, Horst J. Neugebauer should be mentioned first, not only for his scientific and organizational support including permanently driving towards cross-disciplinary research. When I worked on landform evolution, he sent me

to a conference on fractals and dynamic systems in earth sciences in 1995, and I felt completely misplaced there. It took some time to recognize that this would become the direction of my research. Shortly after, the organizers of the conference offered to publish proceedings in a scientific journal; and Adrian E. Scheidegger reviewed my paper on erosion. His result was that all this stuff on erosion is nice, but in principle just water on always the same mill unless the results are considered in the context of SOC. So he was the one who finally gave me the direction towards SOC.

Let me refrain from mentioning the names of all my colleagues who helped me in numerous discussions, who allowed me to occupy nearly all the computers in our department, and who kept things running whenever problems occurred. I am indebted to Yves Bernabé, Jürgen Kurths, Ulrich Mebold, Markus Mendel, Martin Navarro, Michael Pullmann, Harald Schnitzler, Donald L. Turcotte, and Frank Zimmer for their reviews and for their invaluable help in proofreading the manuscript.

Finally, I have never met anyone who has been as tolerant as my wife Franziska. I am sure that she will help me on my way from a temporary workoholic back to life.

Bonn, February 2002 *Stefan Hergarten*

Contents

1. Fractals and Fractal Distributions

Scale invariance has attracted scientists from various disciplines since the early 1980's. B. B. Mandelbrot has been the pioneer on this field; he introduced first ideas in the 1960's and was the first to write a comprehensive book on *scale invariance* (Mandelbrot 1982). However, the idea of scale dependence and scale invariance is much older; D. L. Turcotte begins his book on fractals and chaos in earth sciences (Turcotte 1997) with a citation of J. Ruskin from the year 1860:

> A stone, when it is examined, will be found a mountain in miniature. The fineness of Nature's work is so great, that, into a single block, a foot or two in diameter, she can compress as many changes in form and structure, on a small scale, as she needs for her mountains on a large one; and, taking moss for forests, and grains of crystal for crags, the surface of a stone, in by far the plurality of instances, is more interesting than the surface of an ordinary hill; more fantastic in form, and incomparably richer in colour – the last quality being most noble in stones of good birth (that is to say, fallen from the crystalline mountain ranges).

J. Ruskin was not the only one who noticed some kind of scale invariance. The idea has, for instance, been present in geology for many years. Each photograph of a geological feature should include an object that determines the scale, such as a coin, a rock hammer or a person. Otherwise it might be impossible to recover whether the photograph covers 10 cm or 10 km.

But what started the avalanche in research on scale invariance in the previous decades? Apparently, scale invariance was just a phenomenon for a long time; some patterns were recognized to look similar on different scales, while the rest showed a scale dependence. Mandelbrot was the first to quantify the consequences of scale invariance and to build up a mathematical framework around it which made scale invariance attractive from a general point of view. His seminal work (Mandelbrot 1967) dealt with measuring the length of a coastline with rulers of different lengths; and he found out that the length grows with decreasing ruler length according to some power of the ruler's length. With the help of such multi-scale measurements, a non-integer dimension can be assigned to some objects; for a coastline it is between one and two. The fractional dimension led to the term *fractals.*

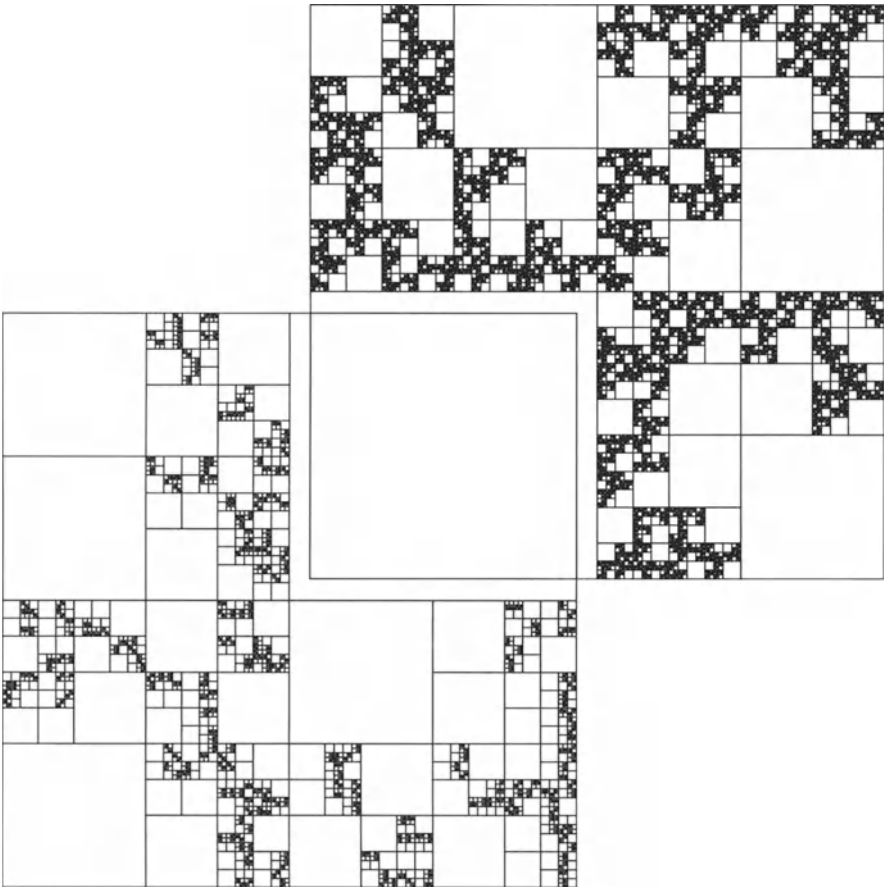

Fig. 1.1. Two computer-generated patterns. Which one is fractal?

Let us begin with an example. Figure 1.1 shows two computer-generated multi-scale patterns. We can imagine them as idealized, two-dimensional models of fractured rock. Which one is fractal? Obviously, both cover a wide range of scales, but this is not sufficient for scale invariance. If we magnify the patterns and take, e. g., a quarter as the whole, we immediately recognize that none of them looks exactly the same on all scales. Therefore, both patterns are non-fractal according to the straightforward, heuristic definition:

> A fractal is an object that looks the same on all scales.

However, things are completely different if we switch from our deterministic view to a statistical description. We can derive a set of patterns from the original pattern by considering magnified sections according to the following rules: The pattern is subdivided into quarters, and each of these quarters is

magnified by a factor two. In the next step, the same procedure is applied to the quarters, and so on. In the upper right example, all the patterns obtained from this procedure are similar; they are either empty or three quarters are partly filled, while one remains empty. Only the position of the empty square is random. Thus, we can select a section of the pattern and magnify it in such a way that we are not able to reconstruct its original size – the pattern is *statistically similar* on all scales. In contrast, the lower left pattern shows a scale dependence. Only on the largest scale, each square consists of three partly filled and one empty quarter. At smaller scales, there are some squares with only two filled quarters; their number increases towards smaller scales. Thus, we can determine the scale of a magnified part of the pattern by counting squares, provided that we know the statistical rules of the whole pattern.

Since nature is not completely regular, the definition of fractals given above is too narrow. So let us focus on *statistical fractals* in the following:

> A fractal is an object that is statistically similar on all scales.

However, there is not a straightforward definition of statistical similarity. Instead there are several definitions which are not equivalent in general; some of them are discussed in the following sections.

1.1 The Fractal Dimension

The use of the quantities length, area, and volume seems to be clear in everyday life. The size of a three-dimensional object can be characterized by its volume, measured in cubic meters. The boundary of such an object is a surface. Since an idealized surface is infinitely thin, its volume is always zero. So it makes no sense to characterize the size of a surface in cubic meters. Instead it can be characterized by its area, measured in square meters. In return, it makes no sense to characterize a volume in terms of area; trying to fill a volume with infinitely thin sheets results in an infinite area. The same arguments can be applied to lines; their size is measured in terms of length.

So there seems to be no doubt that the size of any object can be characterized by either a length, an area or a volume. Which of these measures is appropriate depends on the *dimension* of the object. Objects whose sizes are measured in terms of length are one-dimensional objects; those whose sizes are measured in terms of area are two-dimensional, and finally those whose sizes are measured in terms of volume are three-dimensional.

However, some objects fall through this grid, at least if we allow smaller and smaller structures without any limitation. Figure 1.2 gives an example based on a simple, recursive construction; it is called *Koch's island.* As illustrated in the middle of the island, the algorithm starts from a triangle by placing smaller triangles on the straight lines of the perimeter. Finally, this algorithm leads to a quite rough boundary which looks the same on all scales.

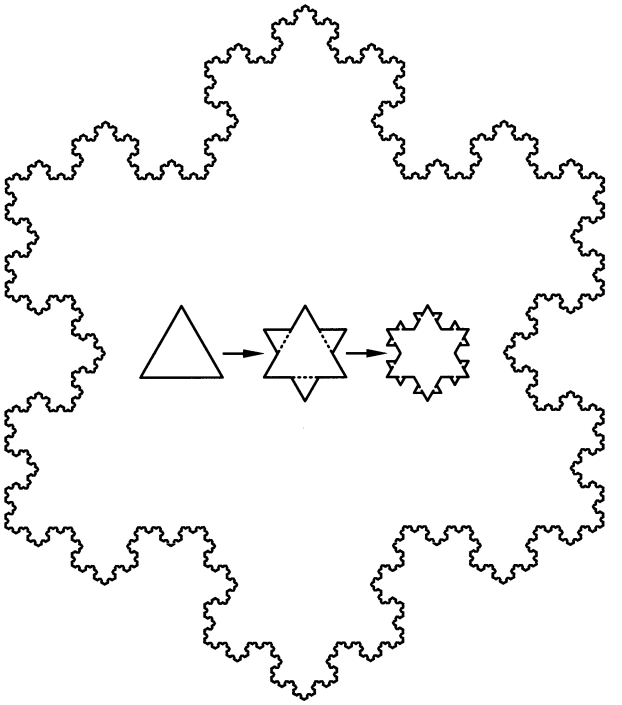

Fig. 1.2. Koch's island.

Obviously, the area of the island is well-defined; otherwise it would not fit on the page. In contrast, the perimeter is so rough that its length becomes infinite in the limit of infinite refinement: Let us assume that the edge length of the original triangle is one; then the length of the perimeter at refinement level zero is $L_0 = 3$. In the first step, each line segment is replaced with four segments of length $\frac{1}{3}$, so that $L_1 = \frac{4}{3}L_0 = 4$. The length after n steps of refinement is

$$L_n = \tfrac{4}{3}\, L_{n-1} = 3\left(\tfrac{4}{3}\right)^n \to \infty \quad \text{for} \quad n \to \infty.$$

On the other hand, the area covered by the boundary is zero. This result can be obtained by covering the perimeter with disks (solid circles) of different radii as illustrated in Fig. 1.3. In the left-hand part, the perimeter is covered by $N_0 = 3$ disks of radius $r_0 = \frac{1}{2}$. As shown in the middle and right-hand illustrations, we can also cover it with $N_1 = 12$ disks of radius $r_1 = \frac{1}{6}$ or with $N_2 = 48$ disks of radius $r_2 = \frac{1}{18}$. In general, we can cover it with $N_n = 3{\times}4^n$ disks of radius $r_n = \frac{1}{2}{\times}3^{-n}$ for any integer number n. Obviously, the area covered by the perimeter is smaller than the sum of the areas of all disks:

$$A_n = N_n \pi r_n^2 = \tfrac{3}{2}\pi \left(\tfrac{4}{9}\right)^n \to 0 \quad \text{for} \quad n \to \infty.$$

Thus, the perimeter of Koch's island is a quite strange line; its size can be neither measured in terms of length, nor in terms of area. Since the length is

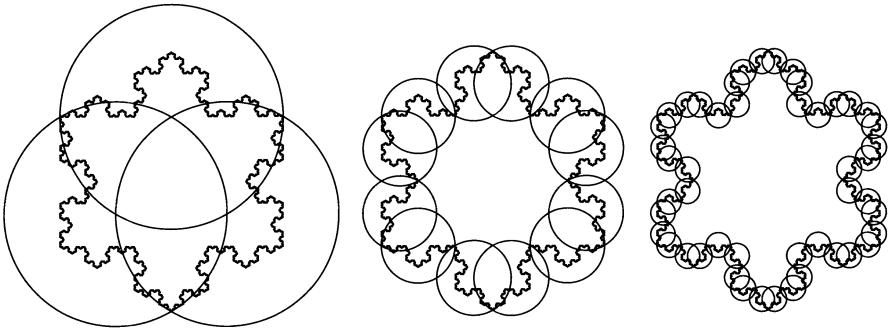

Fig. 1.3. Covering the perimeter of Koch's island with disks of different radii.

infinite, while the area is zero, the perimeter is somewhere between one- and two-dimensional objects.

In order to obtain an appropriate measure for the perimeter of Koch's island, we revisit the result from counting the number of disks needed for covering the perimeter. At least for the radii $r_n = \frac{1}{2} \times 3^{-n}$, we need

$$N(r_n) \;=\; 3 \times 4^n \;=\; 3 \times 4^{-\frac{\log(2r_n)}{\log 3}} \;=\; 3\,(2r_n)^{-\frac{\log 4}{\log 3}}$$

disks. From this we obtain the *power-law relation*

$$N(r) \;\sim\; r^{-D} \tag{1.1}$$

where $D = \frac{\log 4}{\log 3} \approx 1.26$.

The technique of covering an object with disks of different radii is closely related to the discussion on lengths and areas in the beginning of this section. If we cover a line of a well-defined length by disks, we obtain $N(r) \sim \frac{1}{r}$, so that we reproduce Eq. 1.1 with $D = 1$. Applying the same technique to a solid area leads to $N(r) \sim \frac{1}{r^2}$, so that $D = 2$. Therefore, this method suggests an extension of the term dimension towards non-integer values.

The basic idea can be transferred to three dimensions by taking balls instead of disks. It can easily be seen that the obtained exponents D are still consistent with the dimensions of the classical objects discussed in the beginning of this section. Let us in the following use the term *ball* in arbitrary dimensions as it is done in mathematics, so that balls are disks in a two-dimensional world and bars in a one-dimensional world. So let us define a *fractal set* in a n-dimensional Euclidean space:

> Let $N(r)$ be the minimum number of balls of radius r required for covering the set. The set is called *fractal* if $N(r)$ follows the power-law relation
>
> $$N(r) \;\sim\; r^{-D}$$
>
> with a non-integer exponent D. D is the *fractal dimension* of the set.

In contrast to the rather soft definition involving statistical similarity, we now speak of sets instead of objects or patterns. When considering a set, we need just one information on each point, that is, whether it belongs to the set or not. At first sight, the difference between considering sets and objects or patterns is a minor one: The patterns shown in Fig. 1.1 can be directly interpreted as a graphical representation of a set; black points belong to the set, while white points do not. This association suggests to consider the set consisting of the fractures in this example. However, this interpretation of a plot is straightforward, but not stringent. We could also define a set consisting of the fragments, i. e., the white areas in the plot. Provided that the lines are infinitely thin, the white squares fill the area almost entirely. We then obtain $D = 2$ for the sets consisting of the fragments. Although we have not yet analyzed the set consisting of the fractures, it is already clear that the fractures are far away from filling the entire area, so that the result $D = 2$ cannot hold for the sets of fractures.

Things become even more complicated if images are analyzed; and many natural data are provided in form of grayscale or color images. Then, a set must be defined by a discrimination; for each point a decision must be made whether it belongs to the set. For instance, those points which are brighter or darker than an arbitrary threshold may be selected. Obviously, the result strongly depends on this choice. Therefore, in makes no sense to speak of fractal properties of, e. g., an image with respect to the definition given above.

At first sight, the two definitions of fractals discussed so far concern different properties. The first definition refers to statistical similarity at all scales on a rather abstract level and addresses scale invariance directly. In contrast, the definition introduced in this section focuses on a fractional dimension. However, we should keep in mind that the fractal dimension is obtained from measuring the size of a set on different scales which are defined by the radii of the considered balls. Fractals are characterized by scale invariance of the measuring procedure; Eq. 1.1 implies that the number of balls increases by a certain factor if the scale (i. e., the radius of the balls) decreases by a certain factor, without regard to the absolute scale. Therefore, the definition given in this section in fact provides a specification of statistical similarity on all scales, although surely not the only one.

The fractal dimension is closely related to the *Hausdorff dimension* which was introduced in the first half of the twentieth century. Roughly speaking, the Hausdorff dimension is defined by considering the measure $N(r)r^d$ in the limit $r \to 0$ where the $N(r)$ is defined as in the definition of a fractal set. For large values of d, $N(r)r^d$ converges towards zero, while it tends towards infinity if d is small. The Hausdorff dimension is the value d where the measure switches from zero to infinity. More detailed descriptions of the formalism are given, e. g., by Feder (1988) and by Sornette (2000); a mathematical treatment is provided, e. g., in the book of Falconer (1990). From Eq. 1.1 we can easily see that the Hausdorff dimension of a fractal set coincides with its

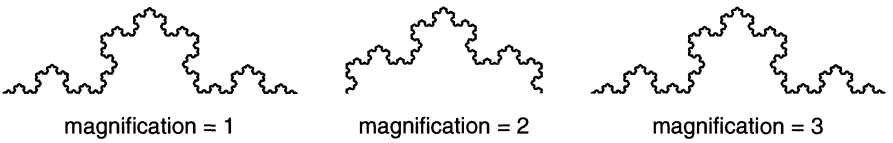

Fig. 1.4. A part of the perimeter of Koch's island at different levels of magnification.

fractal dimension. However, the Hausdorff dimension concerns only the limit of small scales, whereas the definition of a fractal set involves all scales. Thus, a fractional Hausdorff dimension does not imply any scale invariance.

Let us now come back to Koch's island. We used this example for developing a definition of a fractal set, but we did not yet investigate whether the perimeter of the island in fact meets the criteria of a fractal set. The analysis shown in Fig. 1.2 was restricted to certain radii $r_n = \frac{1}{2} \times 3^{-n}$ where n is a positive integer number. For these radii, the number of disks needed for covering the set exactly follows the power-law relation (Eq. 1.1), but we cannot tell much about the result for other radii.

In fact, the perimeter of Koch's is scale-invariant only at discrete scales. Figure 1.4 illustrates this behavior; the vicinity of the island's uppermost corner is plotted on different levels of magnification. The curve keeps its shape under magnification by a factor three, but looks different after a magnification by a factor two. Thus, scale invariance only holds for a discrete set of scaling factors which are given by 3^n where n is an integer number. This phenomenon is called *discrete scale invariance.* The upper right pattern in Fig. 1.1 is another example of this phenomenon; it keeps its statistical properties under a magnification by factors 2^n. However, the difference between continuous and discrete scale invariance is not crucial in the following, so let us refrain from going further into details. A more thorough discussion of discrete scale invariance is given by Sornette (2000).

Furthermore, the perimeter of Koch's island cannot satisfy the power-law relation (Eq. 1.1) for large radii. This problem is not restricted to Koch's island, but to any set which can be covered by a finite number of balls of fixed radius. Let us assume that we have covered the set with a finite number $N(r)$ of balls of radius r. Then there must be a (perhaps large) ball of radius $\tilde{r}$ which covers all these balls, so that this ball covers the entire set. Obviously, the number of balls needed for covering the set is one for all radii which are greater than $\tilde{r}$, in contradiction to the power-law relation (Eq. 1.1). Thus, scale-invariance according to the definition cannot hold for all scales; there must be an upper limit where scale invariance breaks down.

There are patterns such as the upper right one in Fig. 1.1 which can be extended towards infinity in a scale-invariant manner. In principle, the definition of fractal set can be extended to infinite sets, but this may be too theoretical here. Obviously, patterns in everyday life are finite, so let us

simply modify the definition of fractal sets in such a way that the power-law relation (Eq. 1.1) holds for all radii r below an upper cutoff radius $r_{\max}$.

If natural patterns are considered, there should be a lower limit $r_{\min}$ of scale invariance, too. This limitation may be introduced by either the physical process that generated the pattern or by limited observational facilities, such as a finite resolution in maps or photographs.

But even when restricted to a limited range of scales, the definition of a fractal set with the help of Eq. 1.1 causes some trouble. Obviously, the power-law relation cannot be satisfied for all values of r because $N(r)$ is an integer number. Since the resulting deviation from the power-law relation should decrease with increasing number of balls, it should be possible to get around this problem formally. However, this may lead to a theoretically consistent definition, but then we would soon start arguing about the imperfection of nature. Even within a limited range of scales, natural patterns will never exhibit perfect scale invariance. The breakdown of scale invariance at both large and small scales is not a sharp cutoff in general, but disturbs the power-law behavior within the interval between $r_{\min}$ and $r_{\max}$, too. Finding a formally correct, but feasible solution of this problem seems to be impossible. So we should perhaps soften the definition of a fractal set in the following way:

> A set is fractal if the minimum number $N(r)$ of balls of radius r required for covering the set roughly follows the power-law relation
>
> $$N(r) \sim r^{-D}$$
>
> with a non-integer exponent D within a reasonable range of scales.

However, making a sharp distinction between fractal and not fractal sets by such a soft criterion does not make much sense. Obviously, scale invariance is not absolute in nature; nature provides more or less clean fractals. Both the range of scales and the deviations from power-law behavior are criteria for assessing the quality of a fractal in the real world.

1.2 Determining Fractal Dimensions

In the previous section we have applied the definition of a fractal set only to an artificial, regular object – the perimeter of Koch's island. If we try to apply it to less regular sets, we immediately run into a problem: How can the minimum number of balls of radius r required for covering the set be computed? In principle, this leads to a quite complicated problem of optimization where the centers of the balls are the unknown variables. Since the effort becomes enormous if many small balls are involved, the original definition is hardly applied. Instead, some other definitions are used; in some cases they are equivalent to the original definition, but in some cases they are not.

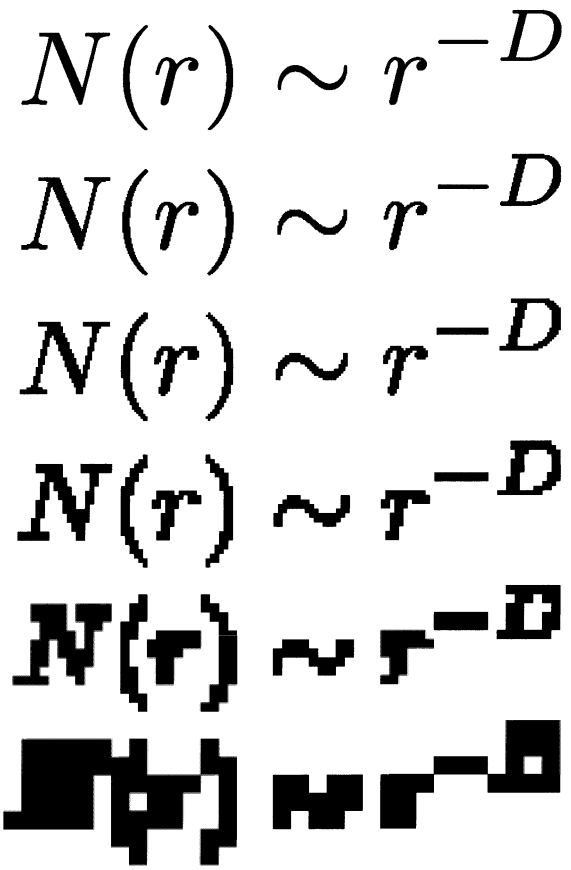

Fig. 1.5. Illustration of the box-counting algorithm, applied to an image of 4096×1024 pixels. The box size increases from 4 pixels (top) to 128 pixels (bottom).

The most widespread method is called *box counting*; it uses neatly aligned cubes instead of balls. In analogy to the term ball, the term cube is used in a generalized sense – a one-dimensional cube is a bar, and a two-dimensional cube is a square. Figure 1.5 illustrates how a set is covered by a grid of boxes of different sizes.

The set is covered by a regular grid of cubes of edge length r. If the number of boxes $N(r)$ needed for covering the set follows the power-law relation

$$N(r) \sim r^{-D}$$

with a non-integer exponent D, the set is called fractal. D is called *box-counting dimension.*

This definition is similar to the original one, but not entirely equivalent. However, we refrain from introducing an own symbol for the box-counting dimension as long as it is clear which definition is referred to.

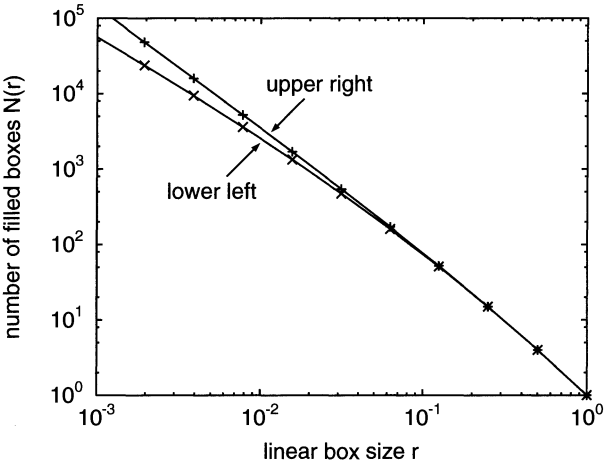

Fig. 1.6. Results of applying the box-counting method to the fractures in Fig. 1.1.

As a result of the regular lattice of cubes, the optimization required for applying the box-counting method is much simpler than that involved in the original definition. Instead of optimizing the location of each ball, only the whole grid must be shifted and rotated. However, often even this variation is omitted. Since Eq. 1.1 can be written in the form

$$\log N(r) \;=\; -D \log r + \text{const},$$

a fractal set is characterized by a straight line in a bilogarithmic plot. Except for the sign, the slope of the line gives the fractal dimension.

Figure 1.6 shows the result of applying the box-counting method to the fractures from Fig. 1.1. The lines were assumed to be infinitely thin; otherwise, the finite linewidth would disturb the result for small box sizes. The length scale was chosen in such a way that the size of the whole pattern is one length unit. Box counting was only performed for discrete box sizes $r = 2^{-n}$ where n is a positive integer number. This restriction reflects the discrete scaling properties of the patterns. Thus, the straight lines between the data points are, strictly speaking, wrong. They were only introduced for a better recognition of curvature in the plots.

Surprisingly, none of both sets yields a clean power law; both plots show a concave curvature. This result is in contradiction to the discussion given in the beginning of this chapter; the upper right pattern should be fractal. The corresponding plot follows a power law only for small box sizes below about $\frac{1}{100}$. For small box sizes, we obtain $D = 1.58$.

Such deviations from the power-law relation at large box sizes are often observed. At first sight, this may be a minor problem since we have already discussed that natural patterns are can only be scale-invariant over a limited range of scales, and that scale invariance is not perfect even within this limited range. But in sum, all these limitations make it difficult to distinguish between fractal and non-fractal patterns or to assess the quality of a fractal. Obviously,

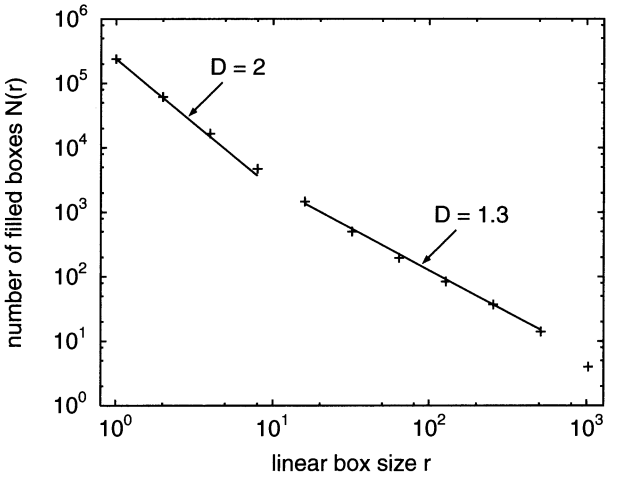

Fig. 1.7. Result of the box-counting algorithm applied to the image from Fig. 1.5.

the curves in Fig. 1.6 differ, but the difference is not as large as we might expect from the fact that one pattern was generated by a scale-invariant algorithm, while the other is clearly scale-dependent.

Many natural patterns considered to be fractal are provided in form of digitized images. In general, a resolution of $10,000 \times 10,000$ pixels is not too bad; it provides four orders of magnitude in possible box size. However, we have seen that this range may be reduced by more than one order of magnitude at large box sizes, and the discrete pixel structure may introduce a comparable loss at small sizes. Then, only two orders of magnitude are left, and distinguishing a power law over such a narrow range requires at least some belief in scale invariance. Therefore, box counting (and similar methods) should always be applied with some caution.

Figure 1.7 illustrates that box counting may in fact be misleading; it presents the data from the (clearly non-fractal) example given in Fig. 1.5. A straight line with a slope of $D = 2$ fits the data fairly well for small box sizes up to 8; the set looks like a solid area at small scales. Between $r = 16$ and $r = 512$, a straight line with $D = 1.3$ looks reasonable. So the box-counting analysis indicates some spurious scale invariance in a non-fractal pattern. Such effects are often a result of transitions between different scaling regimes. At scales smaller than the width of the thinnest lines, lines look like solid areas, while their one-dimensional character is revealed at larger scales. In combination with a variable line spacing, this transition appears as scale invariance within a narrow range of scales.

This problem is not a deficiency of the box-counting method alone; similar effects may also occur if the original definition involving balls is applied. Spurious scale invariance just reflects the general dilemma when dealing with fractals: Natural patterns can only be scale-invariant within a limited range of scales, limited observing facilities reduce the applicable range of scales

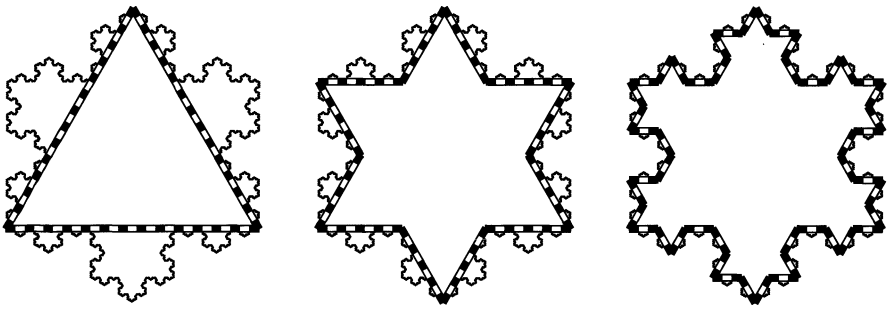

Fig. 1.8. Measuring the length of the perimeter of Koch's island with rulers of different sizes.

further, and finally noise introduces additional uncertainty. In principle, there is no way around this problem, except for being careful concerning the results.

Beside box-counting, there are several other methods for investigating scale invariance and determining fractal dimensions. The method introduced in the founding paper of Mandelbrot (1967) is based on measuring the length of a curve with rulers of different sizes. Figure 1.8 illustrates this procedure in the example of the perimeter of Koch's island. A definition of a fractal curve according to the *ruler method* is:

> A curve is fractal if the number of rulers of length r needed for measuring the length of the curve follows the power-law relation
>
> $$N(r) \sim r^{-D}$$
>
> with a non-integer exponent D. D is called *ruler dimension.*

Since the measuring procedure shown in Fig. 1.8 is very similar to the iterative construction of the island illustrated in the middle of Fig. 1.2, it is not surprising that the perimeter of Koch's island is fractal according to the ruler definition, too. In this example, the ruler dimension coincides with the fractal dimension determined by covering the curve with disks. However, this is not true in general; the ruler dimension is not necessarily the same as the fractal dimension in the original sense or the box-counting dimension.

Let us now come to the *mass method* which is widely used for analyzing sets consisting of many thin lines such as fracture patterns in planar view or river networks. The method is mainly applied in two dimensions, although the generalization is straightforward. While all methods for determining fractal dimensions discussed so far are based on scale-dependent measures, the mass method requires the existence of an absolute measure. This measure, called mass, is applied to disks of different radii. If the set consists of lines, the mass may be the total length of all lines, provided that the lines themselves are non-fractal. Otherwise, their length may not be well-defined. The mass method introduces the following definition of a fractal set:

A disk of radius r is centered on each point of the set, and the mass contained in each disk is computed. The set is fractal if the average mass per disk of radius r, $M(r)$, follows the power-law relation

$$M(r) \sim r^{-D}$$

with a non-integer exponent D. D is called *mass dimension.*

In practice, only a finite number of disks can be considered. Typically, $M(r)$ is obtained by averaging over about 100 disks for each radius. We may already guess that the mass dimension coincides with other fractal dimensions in some cases, but is not equivalent in general. Let us not go further into details here; a more thorough discussion is provided, e. g., in the review article of Bonnet et al. (2001) on fracture systems.

In summary, we have met several definitions of fractals in this section. From their fundamental ideas, they are all similar to the original definition based on covering a set with balls. In contrast to the original definition, they are feasible in practical applications; but on the other hand, they are not completely equivalent. Therefore, each method introduces a new fractal dimension – the box-counting dimension, the ruler dimension, and the mass dimension. In some cases, the different definitions lead to the same results; then these dimensions coincide with the original fractal dimension. However, there are several counterexamples where this is not true. Unfortunately, finding out whether this is the case is not easy since applying the original definition is too costly in general.

1.3 Fractal Distributions

The definitions of scale invariance discussed in the previous sections refer to sets in a n-dimensional, Euclidean space. But what if a volume of rock is fragmented by an explosion or by a volcanic eruption? We may collect the pieces, but we will not be able to put together the puzzle of millions of fragments in order to find out whether the resulting pattern is scale-invariant. But does this mean that the concept of scale invariance ends here?

Imagine that we have lost all information on the spatial alignment of the fragments in Fig. 1.1. The remaining information concerns the sizes and shapes of the fragments. Obviously, all fragments are squares in this example, so the shape does not carry any information. Let us characterize the size r of a fragment by its edge length, assuming that the length unit coincides with the size of the whole pattern. Figure 1.9 shows the cumulative size statistics of the fragments for both patterns, i. e., the number of squares $N(r)$ with linear sizes of at least r.

The plot is similar to that obtained by applying to box-counting method to the patterns (Fig. 1.6). At least for small fragment sizes, the statistics of

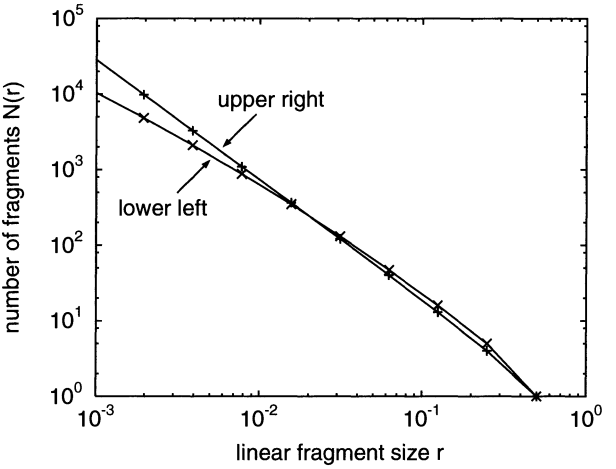

Fig. 1.9. Cumulative fragment size statistics of the examples from Fig. 1.1.

the upper right pattern roughly follow the power law $N(r) \sim r^{-D}$. At large fragment sizes, the plot shows a slight concavity. We will see in Sect. 1.6 that the concavity is an effect of the finite size of the pattern, i. e., of the absence of fragments larger than $r = \frac{1}{2}$. Beyond the qualitative coincidence, the exponent $D \approx 1.58$ is the same as the fractal dimension obtained from box counting. Apparently, the statistics of the fragment sizes reflect the properties of the original patterns concerning scale invariance.

Obviously, only fragment sizes $r = 2^{-n}$ where n is a positive integer number occur in both patterns. In Fig. 1.9, the data points are connected with lines; but this is, strictly speaking, wrong since the number of fragments is a staircase function. However, connecting the data points with lines facilitates deciding whether $N(r)$ follows a power law.

The power-law statistics of the fragment sizes suggest a definition of scale invariance which differs in its spirit from those definitions introduced in the previous sections. Let us start from a set of arbitrary objects and assign a *linear object size* r to each object. This may be any measure of the object's size with the dimension of a length, such as length, width, largest diameter, square root of area or cubic root of volume. Let us further proceed towards a statistical description and replace the number of objects $N(r)$ by the probability $P(r)$ that an arbitrarily chosen object has a size of at least r. The function $P(r)$ is called *cumulative size distribution.* Let us define:

A set of objects exhibits *fractal* or *scale-invariant size statistics* if the cumulative size distribution $P(r)$ of the linear object sizes is a *power-law distribution*:

$$P(r) \sim r^{-D}.$$

The exponent D is called *fractal dimension of the distribution.*

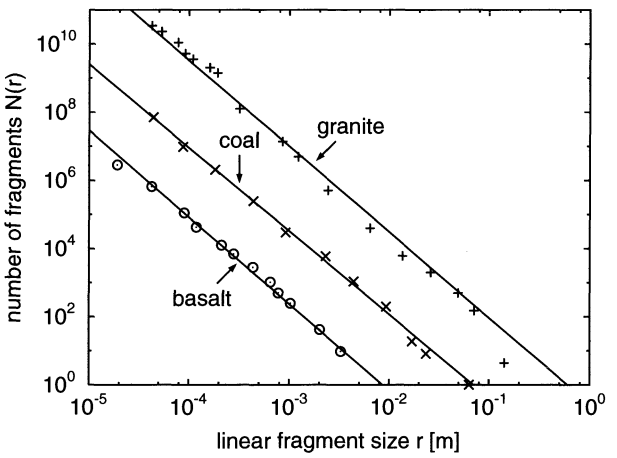

Fig. 1.10. Cumulative size statistics of rock fragments from granite broken by a nuclear explosion (Schoutens 1979), broken coal (Bennett 1936), and projectile impact on basalt (Fujiwara et al. 1977). The data were taken from Turcotte (1997).

Power-law distributions are also called *fractal distributions* or *scale-invariant distributions.* The relation between scale invariance and power-law distributions arises from the absence of characteristic scales in these distributions.

There are numerous examples of power-law distributions in earth sciences (e. g. Bak 1996; Turcotte 1997). Some of them – sizes of forest fires, rupture areas of earthquakes, landslide areas, and drainage areas of rivers – are discussed later in this book. Let us here consider the size statistics of rock fragments. In contrast to the artificial example considered above, natural fragments offer a variety of shapes. Nevertheless, it makes sense to define the linear fragment size by the cubic root of the volume as we would do if we considered cubes. Figure 1.10 shows data from granite broken by a nuclear explosion (Schoutens 1979), broken coal (Bennett 1936), and projectile impact on basalt (Fujiwara et al. 1977). The fragment sizes obey fair power-law statistics over two to four orders of magnitude with exponents $D \approx 2.5$. Although these data suggest that this value may be universal for rock fragments, this is not true in general (Turcotte 1997), and there are counterexamples where the size distribution of rock fragments is not fractal at all.

This example illustrates again that scale invariance is not absolute in nature. Beside the cutoff effects at small and large object sizes, an observed distribution will normally not exhibit perfect power-law behavior even within a limited range of scales.

In principle, there is no need for assigning a linear size to the considered objects; the power-law distribution can refer to arbitrary measures. The linear size r of two-dimensional objects is often characterized by the square root of their area A, which leads to

$$P(A) \sim A^{-b} \quad \text{where} \quad b = \tfrac{D}{2}.$$

We may even go a further step ahead and allow other quantities than length, area or volume for characterizing the size of an object. We may, e. g., consider

earthquakes or even any other events as generalized objects and use, e. g., the energy released by an earthquake for measuring its size. This leads to a more general definition of a power-law distribution of an arbitrary property s:

$$P(s) \sim s^{-b}. \tag{1.2}$$

However, it makes no sense to speak of a fractal dimension in this context, although the distribution suggests some kind of generalized scale invariance.

The probability $P(s)$ must not exceed 1. Therefore, a power-law distribution can only be valid above a minimum object size $s_{\min}$; the scale invariance of the distribution ends at this scale. However, this is mainly a mathematical problem since scale invariance of natural patterns should end at least at the atomic scale. In this sense, the power-law distribution should be replaced by the *Pareto distribution*:

$$P(s) = \begin{cases} \left(\frac{s}{s_{\min}}\right)^{-b} & \text{if} \quad s > s_{\min} \\ 1 & \text{else} \end{cases} . \tag{1.3}$$

This restriction shows that the term scale-invariant distribution should be used with caution. Obviously, the cutoff value $s_{\min}$ introduces a characteristic scale.

Scale invariance is not the only characteristic property of the power-law, respectively, Pareto distribution. It differs from many other distributions such as Gaussian (normal), log-normal, and exponential distribution concerning the occurrence of large objects. Let us consider the *exponential distribution*

$$P(s) = \exp\left(-\left(\frac{s}{s_0}\right)^{\nu}\right)$$

for $s \geq 0$ with the positive parameters s_0 and ν. This distribution is often named after Rosin and Rammler (1933). Compared to the Pareto distribution, the exponential distribution converges more rapidly towards zero in the limit of large object sizes:

$$\lim_{s\to\infty} \frac{P(s)_{\text{expon.}}}{P(s)_{\text{Pareto}}} = \lim_{s\to\infty} \exp\left(b \log\left(\frac{s}{s_{\min}}\right) - \left(\frac{s}{s_0}\right)^{\nu}\right) = 0.$$

This result holds for all values of the parameters ν, b, $s_{\min}$, and s_0. Thus, large objects are always more probable in any power-law distribution than they are in any exponential distribution, and the difference becomes stronger and stronger in the limit of infinite object sizes. The same applies to many other distributions in comparison with the power-law distribution; a detailed discussion is given in the book of Sornette (2000).

A slow power-law decay at large sizes is called *heavy tail.* The smaller the exponent b is, the heavier the tail becomes. Formally, heavy-tailed distributions can be characterized by the condition that an exponent n can be chosen in such a way that $P(s)s^n \to \infty$ in the limit $s \to \infty$. Obviously, the

occurrence of heavy tails and the "weight" of the tails is a central point in risk assessment; if a distribution has a heavy tail, even extreme events, i. e., events which are much large than the average, occur at a considerable probability. Therefore, power-law distributions (or distributions with a power-law tail) seem are often related to phenomena which may become dangerous.

1.4 Fractals or Fractal Distributions?

So far we have learned about two different definitions of scale invariance. The first definition is based on the geometric properties of a set in a n-dimensional Euclidean space, while the second definition focuses on sizes of objects and disregards any further properties such as spatial alignment. On a less formal level, the difference between both definitions can be attributed to different *notions of scale*. When the distribution of object sizes is considered, we focus on a scale defined by the object we are actually looking at. In contrast, the scale in the geometric definition is some kind of resolution; covering a set with balls of a certain radius means that we are not able to recognize smaller details than the scale defined by the size of the balls.

But after all, which definition of scale invariance is better in practice? Let us first approach this question from a technical point of view addressing the effort and the danger of being trapped by spurious scale invariance. At first sight, analyzing object size statistics seems to be preferable under both aspects. Let us revisit the artificial fracture patterns from Fig. 1.1. Compared to the results from box counting (Fig. 1.6), the cumulative size statistics (Fig. 1.9) distinguish more clearly between the fractal and the non-fractal pattern. From its basic ideas, the box-counting method comes quite close to the original definition of scale invariance involving balls. So we may conclude that fractal size distributions provide a sharper criterion for scale invariance than the original definition does, at least if only a limited range of scales is available, although we should be careful with a result obtained from a singular example.

Furthermore, the better method even seems to be the simpler one. Obviously, making a plot of the number of objects as a function of their sizes is much easier than covering a set with balls or boxes, measuring its size with rulers of different lengths or applying the mass method. However, this is only true if the objects and their sizes are already prepared. In general, three steps must precede the statistical analysis: Objects must be defined, recognized and measured. As soon as the data set consists of a map, a photograph, a digital elevation model or something similar, each of these steps may become a difficult task. Often, recognition and measuring must be done manually. Fracture patterns or river networks are good examples where even defining objects becomes difficult. What happens where rivers or cracks join? The objects must be ordered in some kind of hierarchy for deciding which one is the main river or fracture, and which one ends at a junction. So we see

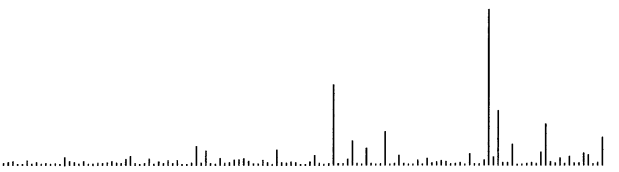

Fig. 1.11. A comb made of fractures with power-law (Pareto) distributed lengths.

that the way towards an object size statistics may be long and not always straightforward. So there is not a clear advantage of either method in general if the effort is the criterion.

However, this technical discussion misses the crucial point. Both definitions refer to different properties. There are a few examples such as the fragmentation pattern discussed above where both definitions are equivalent, but in general they are not. Even in this example we should have been suspicious from the beginning because the box-counting analysis concerns the set consisting of the fractures, while the size statistics concern the fragments. Fracture patterns are in fact one of the most prominent example where both definitions are equivalent only under certain conditions. In general, a power-law distribution of fracture lengths (in a planar view) or areas (in space) does not imply scale invariance of the fracture pattern. Beside the fracture sizes, scale invariance of the pattern involves the spacing between fractures and on their orientation. Quantitative relationships between these properties and scale invariance are discussed by Bour and Davy (1999) and by Bonnet et al. (2001). Figure 1.11 shows a simple example where fractures obeying scale-invariant length statistics (drawn from a Pareto distribution) are aligned in form of a comb. Since the spacing between the fractures is fixed, the pattern is anything but scale-invariant.

The distribution of lakes on the earth's surface is another example where both definitions of scale invariance are not necessarily equivalent. According to studies of Kent and Wong (1982) and Meybeck (1995), the sizes of lakes are power-law distributed with fractal dimensions exposed to a regional variation, but smaller than two. This fractal distribution may be the result of a fractal land surface, but this is not compelling. Studies on seafloor topography have revealed scale-invariant properties (e. g. Fox and Hayes 1985; Malinverno 1995; Turcotte 1997). But on the other hand, there is still discussion whether the land surface is fractal or whether the variety of landform evolution processes acting here destroys scale invariance (e. g. Evans and McClean 1995). Even if the land surface is fractal, the exponents of the size distributions of lakes are not necessarily correlated to the fractal dimension of the earth's surface which should be greater than two.

In summary, the general question for the better approach is ill-posed. There are examples where the analysis is a priori restricted to either of the definitions. Patterns where objects cannot be clearly distinguished fall into this class as well as data sets where the information on the spatial alignment of the objects has been lost. But even in those cases where both definitions

can be applied, both definitions may refer to essentially different properties. Thus, it is not surprising that some patterns are scale-invariant according to one definition, but scale-dependent from the other point of view. However, this is not a problem as long as we are always sure to choose the method which refers to the property we are interested in.

1.5 Are Fractals Useful?

Especially in the late 1980's, there was a great enthusiasm about fractals. Apparently, almost everything was fractal at this time. Beside a large number of patterns which were just beautiful, increasing computer power allowed the generation and visualization of artificial fractal surfaces that looked very much like the earth's surface at first sight (e. g. Voss 1985; Feder 1988).

Figure 1.12 shows three computer-generated surfaces. The upper one is the non-fractal limiting case ($D = 2$), the middle and the lower one are fractal with dimensions $D = 2.25$ respectively $D = 2.5$. For a more realistic impression, the surfaces were filled with water up to a certain level and placed on a section of a sphere in order to improve the aerial view.

Let us not worry too much about details here. The surfaces were generated using the two-dimensional generalization of fractional Brownian motion which will be discussed in Sect. 3.4. Strictly speaking, these surfaces are not fractal in the sense discussed so far; they exhibit spatially anisotropic scale invariance in the sense discussed in Chap. 3. However, assuming anisotropy is reasonable because gravity introduces a strong preferential direction in landform evolution processes.

Especially the surface with $D = 2.25$ looks quite realistic. We could support this thesis by analyzing the earth's surface with respect to scale-invariant properties. However, quantitative results are non-unique as already discussed in the previous section. But apart from this problem, can we make any use of such a fractal surface? We can, e. g., if developing a flight simulator is our task. With the help of the algorithm for generating fractal surfaces me may then get around providing detailed information on the land surface in our program. The same strategy can be applied when process models require a land surface with a high spatial resolution. Then, the large-scale information may be taken from the natural relief, while reasonable small-scale structures may be supplied by the fractal algorithm.

Fluid flow and transport phenomena in porous media are another field where fractal concepts have been applied. Hydraulic parameters of porous media with fractal pore properties can be computed (Pape et al. 1999), the fractal character of fault surfaces is used for computing hydraulic properties of individual faults (e. g. Brown 1987; Plouraboué et al. 1995), and scale-invariant properties of fracture networks in rocks can serve as a basis for modeling large-scale flow of water in fractured rocks (Kosakowski et al. 1997).

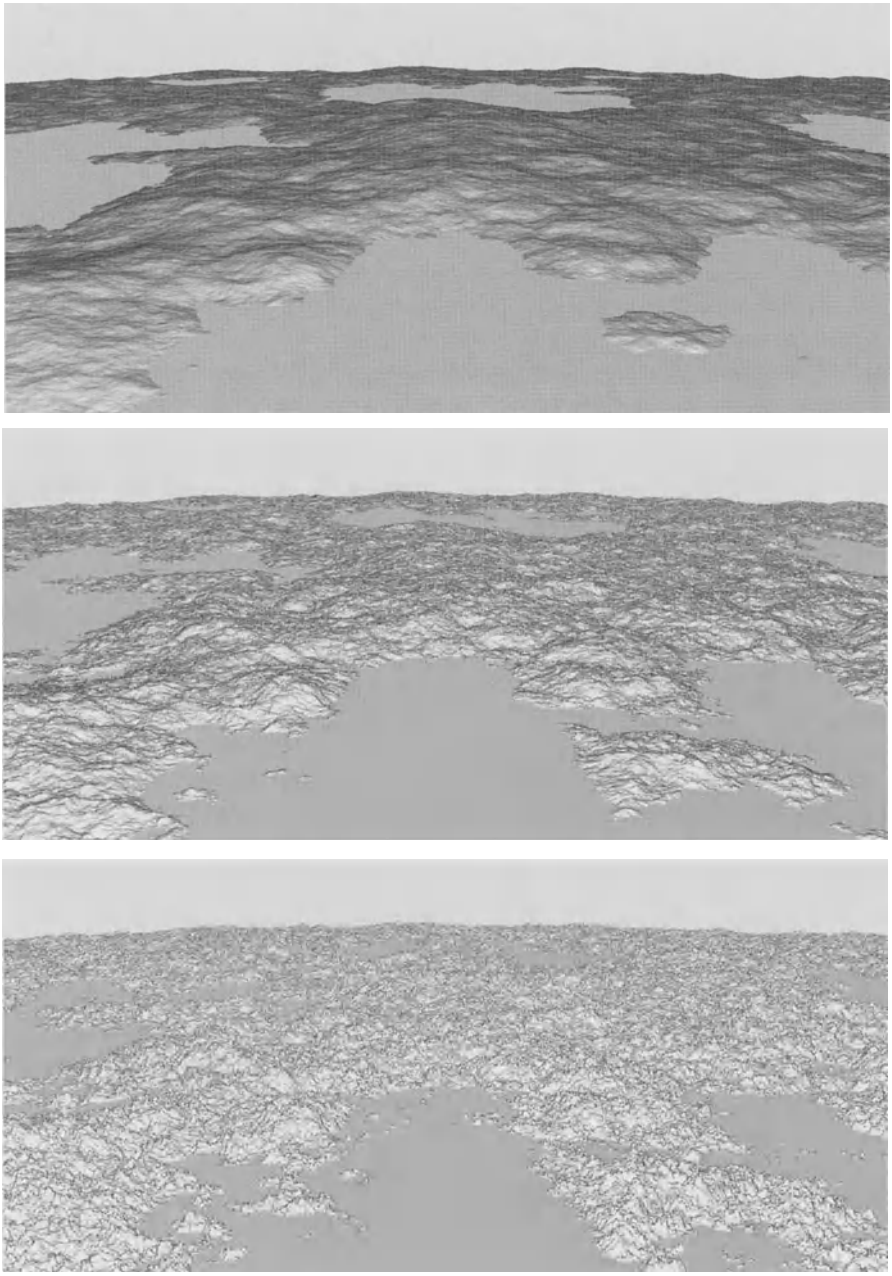

Fig. 1.12. Three computer-generated surfaces with fractal dimensions $D = 2$ (top), $D = 2.25$ (middle), and $D = 2.5$ (bottom).

Finally, mechanical properties of faults can be partly derived from fractal properties of their surfaces (e. g. Hallgass et al. 1997; Hansen et al. 2000).

Such applications of fractals are important in singular fields, but they cannot explain the enthusiasm on fractals. In fact it is the question for their origin that makes fractals attractive. Fractals and fractal distributions seem to be unifying phenomena because they occur in many phenomena which seem to be completely different at first sight. So the crucial question is whether there is a fundamental mechanism that generates scale invariance. If so, the framework of fractals greatly widens our understanding not only of specific earth processes, but also of natural phenomena in general.

1.6 Where do Fractals Come From?

Obviously, finding a unique explanation for the origin of fractals would be a great deal. However, it turned out that nature is not so simple; there are at least several mechanisms that lead to scale invariance. Nevertheless, each of these mechanisms is applicable at least to a class of phenomena, so these mechanisms retain their unifying character although none of them can explain the whole world.

The simplest mechanism explains fractals as a result of other fractals. As discussed in Sect. 1.3, the power-law size distribution of lakes may simply be a result of a fractal surface topography. In analogy, the fractal size distribution of rock fragments may be a consequence of a scale-invariant pre-design of zones of weakness in rocks. Explanations of this type will be discussed with respect to earthquakes in Chap. 7 and in the context of landslides in Chap. 8. Although an explanation of this type may be correct in some cases, its contribution to knowledge is a minor one because it just shifts the problem. At the beginning of the chain, there must be a fractal that cannot be explained by this mechanism.

In this book, we focus on two mechanisms leading to fractals or fractal size distributions, although there are several others (e. g. Sornette 2000). In the rest of this section, we discuss the idea of *nesting* – applying the same process on different scales. The second and perhaps most general concept is based on criticality; it is the main topic of this book.

Let us begin our discussion of nesting with the growth of a *fractal tree*. It starts with a stem of unit length; we assume that n branches of length λ between 0 and 1 grow out of the stem. This procedure is continued on the smaller scales, so that each branch of length r has n smaller branches of length λr. Figure 1.13 shows an example of fractal a tree, generated with $n = 4$ and $\lambda = 0.417$. Obviously, the number of k^{th}-order branches is n^k where $k \geq 0$, and their length is $r_k = \lambda^k$. Thus, the number of branches of length r_k or greater can be computed using a geometric series:

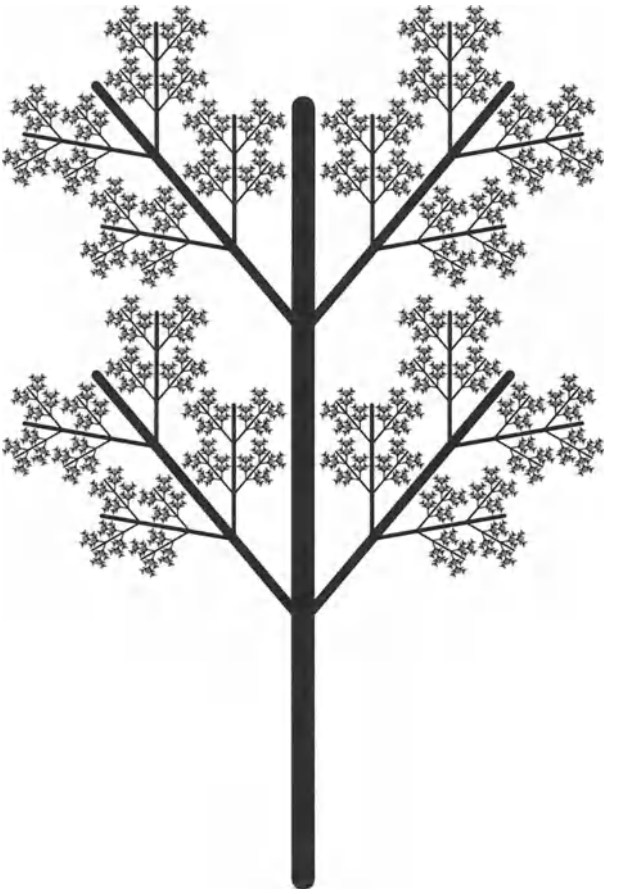

Fig. 1.13. Fractal tree. Each branch of length r has four smaller branches of length $0.417r$.

$$N(r_k) = \sum_{i=0}^{k} n^i = \frac{n^{k+1}-1}{n-1} = \frac{n\,n^k-1}{n-1}.$$

From $r_k = \lambda^k$ we obtain $k = \frac{\log r_k}{\log \lambda}$, so that

$$N(r_k) = \frac{n\,n^{\frac{\log r_k}{\log \lambda}}-1}{n-1} = \frac{n\,r_k^{\frac{\log n}{\log \lambda}}-1}{n-1} = \frac{n}{n-1}\,r_k^{-D} - \frac{1}{n-1},$$

where

$$D = -\frac{\log n}{\log \lambda}.$$

For large orders k, i. e., for small sizes, the distribution follows a power law. The deviations from the power law at large object sizes are a result of the finite size of the tree; they would vanish if we assumed that there are $\frac{1}{n-1}$ objects which are larger than the stem.

For the parameter values used in our example – $n = 4$ and $\lambda = 0.417$ – we obtain $D \approx 1.58$. This is exactly the fractal dimension of the upper right example in Fig. 1.1, although the pictures look completely different. However, this result is not surprising; it only reminds us on the necessity of keeping in mind what the objects are.

The upper right pattern from Fig. 1.1 was generated by the established *fragmentation* algorithm Turcotte (1986) which is similar to the fractal tree: We start at a block of unit volume $V_0 = 1$ and split it into f pieces of equal volumes $V_1 = \frac{V_0}{f}$. We then assume that a certain fraction p of these first-order fragments is fragmented into f smaller pieces of volume $V_2 = \frac{V_1}{f}$, and so on. The main assumption is that the parameters of the process, f and p, are arbitrary, but the same on all scales. Obviously, the number of first-order fragments is $n_1 = f(1-p)$; the number of second order fragments is $n_2 = f^2 p(1-p)$. This can be generalized to

$$n_k = f^k p^{k-1}(1-p).$$

The volume of the fragments of order k is $V_k = f^{-k}$. The relationship between volume and linear object size depends on the dimension d of the considered space; let us define

$$r_k := V_k^{\frac{1}{d}} = f^{-\frac{k}{d}}.$$

Again, the number of fragments of size r_k or greater can be computed with the help of a geometric series:

$$N(r_k) = \sum_{i=0}^{k} n_i = f(1-p) \sum_{i=1}^{k} (fp)^{i-1} = f(1-p) \frac{(fp)^k - 1}{fp-1}.$$

The definition of the linear object size leads to $k = -d\frac{\log r_k}{\log f}$, so that

$$N(r_k) = f(1-p) \frac{(fp)^{-d\frac{\log r_k}{\log f}} - 1}{fp-1} = \frac{f(1-p)}{fp-1} \left(r_k^{-D} - 1\right),$$

where

$$D = d \frac{\log(fp)}{\log f}. \tag{1.4}$$

Again, the power-law behavior is not perfect, but would be perfect if we added $\frac{f(1-p)}{fp-1}$ fragments of sizes $r > \frac{1}{2}$. The fractal pattern in Fig. 1.1 results from a two-dimensional application of this algorithm with $f = 4$ and $p = \frac{3}{4}$.

As already mentioned, the size distribution of rock fragments often follows a power law with a fractal dimension $D \approx 2.5$. Explaining the size statistics of rock fragments with this basic fragmentation algorithm is straightforward. In three dimensions, $D = 2.5$ can be achieved assuming $p = 0.89$ if $f = 2$. Strictly speaking, this parameter set is not allowed since the product fp must be an integer number in this mainly deterministic algorithm. However, the fragmentation algorithm can be generalized in a statistical way where p

is the probability that a fragment is split up into smaller pieces. Then, the limitation to integer numbers can be dropped. This generalized fragmentation model was discussed in detail by Krapivsky et al. (2000); its properties slightly differ from that discussed above because the process of fragmentation may cease after a finite number of steps.

Obviously, the fragmentation algorithm is able to explain size distributions with fractal dimensions between zero and the dimension of the embedding space. This is exactly the range of the fractal dimension which can occur according to the original definition of fractals involving balls. However, there is no reason why the exponent of a power-law size distribution should be confined to this range; it may be even larger than the spatial dimension.

Interestingly, Kaminski and Jaupart (1998) observed exactly this behavior when analyzing the sizes of fragments released by volcanic eruptions. They found power-law distributions with exponents $D > 3$ which cannot be explained by the fragmentation model. On the other hand, the model is so straightforward that it is hard to believe that a scale-invariant distribution of fragment sizes may emerge from another mechanism. Therefore, finding a reasonable explanation for the result $D > 3$ is not straightforward, provided that a bias due to the sampling procedure can be excluded. Kaminski and Jaupart (1998) suggest an extension of the fragmentation model towards a two-step process. In their model, fragments generated by a scale-invariant fragmentation process are further fragmented in a second phase. However, exponents $D > 3$ can only be achieved by assuming a scale dependence of the secondary fragmentation process; the probability of fragmentation must decrease with decreasing fragment size.

The non-fractal pattern in Fig. 1.1 was generated by such a scale-dependent fragmentation process. In this example, the probability of being split is $p = \frac{3}{4}$ for the first-order fragments and decreases with decreasing fragment size. In analogy to the model suggested by Kaminski and Jaupart (1998), a power-law relation between fragment size and probability was assumed: $p \sim r^{\frac{1}{10}}$. Although this dependence is not very strong, it leads to a significant deviation from the fractal distribution.

However, scale-dependent fragmentation processes are not only important in combination with scale-invariant fragmentation processes where they allow exponents $D > 3$. As soon as a macroscopic stress field becomes important, there should be a scale dependence. Each step of fragmentation should release stress and thus reduce the probability of further fragmentation processes. For instance, patterns resulting from thermal cracking due to cooling or desiccation should obey a size distribution similar to that of the lower left pattern in Fig. 1.1 rather than a fractal one. The deviation from the fractal distribution should reveal valuable information about the governing processes.

2. Recognizing Power-Law Distributions

Power-law distributions or, for being more general, heavy-tailed distributions (Sect. 1.3) become more and more important for assessing natural and man-made hazards. In principle, we already know how to analyze an observed distribution of object sizes with respect to scale-invariant properties. We just have to plot the cumulative size distribution with logarithmically scaled axes and look whether the data set can be approximated by a straight line. If so, the size distribution exhibits fractal properties, and the slope of the straight line is the exponent of the power-law distribution.

However, fitting a straight line graphically seems to be out of fashion in times of computer-based data analysis. A visual analysis involves some degrees of freedom, so that different persons will obtain different results from the same data set. A variety of software for determining the "best" approximation to a given data set is available. At this point, the analysis reduces to transferring the data to the computer and pressing a button.

So what is the scope of this chapter? When analyzing potentially fractal data sets, an improper statistical treatment may considerably affect the results. Often, the consequence is an under- or overestimation of the scaling exponent, but in the worst case, scale-dependent distributions are misinterpreted to be fractal or vice versa. On the other hand, the topics discussed here are rather technical; but unfortunately things become technical as soon as we look behind the pretty pictures.

In order to keep this chapter at a reasonable length, we only address a few, fundamental aspects. The line followed here is straightforward or even somewhat naive; it just describes how everyone would start from some basic knowledge on statistics. However, there are numerous other methods which go beyond our introduction; reviews are given by, e. g., Adler et al. (2000) and Sornette (2000). Among them there are methods such as rank-ordering which turned out to be very powerful when applied to small data sets where the simple methods discussed in the following fail (Sornette et al. 1996). So this chapter gives an introduction, but before performing a serious analysis of real-world data, further reading is recommended. On the other hand, the modeling aspects discussed later in this book can be understood with very little knowledge on statistical methods, so this chapter may in principle be skipped when reading this book for the first time.

2.1 Maximum Likelihood, Least Squares, and Linear Regression

Let us begin with the methods behind those tools which enable us to fit a function or its parameters to a given data set by pressing a button. Assume that we measure a quantity y which depends on a parameter x. For instance, x may be time, and y the time-dependent rainfall intensity. However, x may also be the size of any objects, and y the number of objects of at least this size. Let us further assume that n pairs $(x_1, y_1), \ldots, (x_n, y_n)$ are available, and that we know or suspect that there is a relationship between x and y, so that the measured values y_i should be $y_i = f(x_i)$ in principle. Due to inevitable errors in measurement or statistical fluctuations, the relation $y_i = f(x_i)$ is in general not exactly satisfied by the data. Instead, the measured value y_i is a superposition of the theoretical value $f(x_i)$ and a statistical error ϵ_i:

$$y_i = f(x_i) + \epsilon_i,$$

where the expected value $\overline{\epsilon_i}$ vanishes.

The basic idea of the *maximum-likelihood method* is the following: The actual data set is assumed to be randomly drawn from a statistical distribution; this distribution depends on both the function $f(x)$ and the assumptions on the errors ϵ_i. Let $P(y_1, \ldots, y_n)$ be the probability that the i^{th} measured value is at least y_i for all i from 1 to n. For convenience, we consider the *probability density*

$$p(y_1, \ldots, y_n) \coloneqq -\frac{\partial}{\partial y_1}\left(-\frac{\partial}{\partial y_2}\left(\ldots\left(-\frac{\partial}{\partial y_n}P(y_1, \ldots, y_n)\right)\ldots\right)\right)$$

instead of $P(y_1, \ldots, y_n)$. The probability that the i^{th} measured value falls into the range between y_i and $y_i + \delta y_i$ for all i from 1 to n is $p(y_1, \ldots, y_n)\delta y_1 \ldots \delta y_n$, provided that the interval lengths δy_i are sufficiently small. So, if $y_1, \ldots, y_n$ is the actual set of measured data, what does the value $p(y_1, \ldots, y_n)$ mean? If it is small, it was either bad luck in the measurement or the basic conjecture on the origin of the measured values is wrong. The latter may concern the magnitude of the errors ϵ_i as well as the function $f(x)$. In general, it is easier to believe that the function $f(x)$ is correct if $p(y_1, \ldots, y_n)$ is not too small. The maximum-likelihood method goes a step ahead by assuming that the best function $f(x)$ is that which maximizes the probability density $p(y_1, \ldots, y_n)$ for the measured values $y_1, \ldots, y_n$.

However, the maximum-likelihood concept is rather abstract; for deriving any applicable scheme from this method, a model for the errors ϵ_i is required. The *least-squares fit* is the most widespread application of the maximum-likelihood concept; it involves the simplest assumptions on the errors ϵ_i:

- The random deviations ϵ_i for $i = 1, \ldots, n$ are independent from each other, which means that $p(y_1, \ldots, y_n)$ falls into a product of individual probability densities:

$$p(y_1, \ldots, y_n) = p_1(y_1) \ldots p_n(y_n).$$

- The random values ϵ_i are Gaussian-distributed with given variances σ_i^2, so that the values y_i are Gaussian-distributed around $f(x_i)$:

$$p_i(y_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(y_i - f(x_i))^2}{2\sigma_i^2}\right).$$

These assumptions lead to

$$p(y_1, \ldots, y_n) = \frac{1}{\sqrt{2\pi}^n \sigma_1 \ldots \sigma_n} \exp\left(-\sum_{i=1}^{n} \frac{(y_i - f(x_i))^2}{2\sigma_i^2}\right).$$

If the variances σ_i^2 are independent of the function $f(x)$, the probability density becomes maximal if the expression

$$\chi^2 \coloneqq \sum_{i=1}^{n} \frac{(y_i - f(x_i))^2}{\sigma_i^2}$$

is minimal. Thus, $f(x)$ must be chosen in such a way that the sum of the squares of the deviations between the predicted values $f(x_i)$ and the observed values y_i, inversely weighted with the variances σ_i^2, is minimized.

Obviously, the best function $f(x)$ is that which exactly satisfies the condition $f(x_i) = y_i$ for all measured values. Such a function can, e. g., be obtained by interpolation. However, this is not the result we are looking for because it misinterprets all statistical fluctuations to be real. Therefore, additional assumptions on the function $f(x)$ are made in general, mostly based on a physical hypothesis. If, e. g., the same property is measured several times under the same conditions, $f(x)$ should be constant.

Apart from the example of a constant function $f(x)$, *linear regression* is the simplest application of the least-squares method. The function $f(x)$ is assumed to be linear: $f(x) = \mu x + \nu$; the parameters μ and ν must be determined to minimize χ^2. A necessary criterion for a minimum with respect to μ and ν is

$$\frac{\partial}{\partial \mu}\chi^2 = 0 \quad \text{and} \quad \frac{\partial}{\partial \nu}\chi^2 = 0.$$

From

$$\frac{\partial}{\partial \mu}\chi^2 = 2\sum_{i=1}^{n} x_i \frac{\mu x_i + \nu - y_i}{\sigma_i^2} \quad \text{and} \quad \frac{\partial}{\partial \nu}\chi^2 = 2\sum_{i=1}^{n} \frac{\mu x_i + \nu - y_i}{\sigma_i^2}$$

we obtain

$$\mu = \frac{\left(\sum \frac{1}{\sigma_i^2}\right)\left(\sum \frac{x_i y_i}{\sigma_i^2}\right) - \left(\sum \frac{x_i}{\sigma_i^2}\right)\left(\sum \frac{y_i}{\sigma_i^2}\right)}{\left(\sum \frac{1}{\sigma_i^2}\right)\left(\sum \frac{x_i^2}{\sigma_i^2}\right) - \left(\sum \frac{x_i}{\sigma_i^2}\right)^2},$$

$$\nu = \frac{\left(\sum \frac{x_i^2}{\sigma_i^2}\right)\left(\sum \frac{y_i}{\sigma_i^2}\right) - \left(\sum \frac{x_i}{\sigma_i^2}\right)\left(\sum \frac{x_i y_i}{\sigma_i^2}\right)}{\left(\sum \frac{1}{\sigma_i^2}\right)\left(\sum \frac{x_i^2}{\sigma_i^2}\right) - \left(\sum \frac{x_i}{\sigma_i^2}\right)^2}$$

where all sums extend over the range from $i = 1$ to $i = n$. These formulas look somewhat complicated, but can easily be implemented on a computer. In Sect. 2.3, we will transfer this method to hypothetical power-law distributions of object sizes.

2.2 Do Cumulative Size Distributions Tell the Truth?

So far we have focused on cumulative distributions of object sizes, i. e., on the probability $P(s)$ that a randomly chosen object has a size of at least s. Although this is convenient under theoretical aspects, it is not clear that cumulative size distributions are an appropriate tool for recognizing scale invariance or scale dependence in reality. So let us now figure out to what extent cumulative size distributions may be biased by artifical effects.

No matter whether we analyze objects in nature or run a numerical model, we always look through a finite window. Each finite data set must contain both a lower and an upper limit of object sizes. The simplest case of cutoff behavior at small object sizes was already discussed in Sect. 1.3. The Pareto distribution (Eq. 1.3) hinges on the idea of scale invariance above a certain object size $s_{\min}$, and that there are no smaller objects. It was found that the Pareto distribution is still a power-law function for all object sizes above $s_{\min}$, so that this simplest kind of cutoff behavior does not affect the analysis.

However, things are different at the upper limit of object sizes. The upper limitation may be an inherent property of the considered distribution, but may result from the process of observation, too. In the simplest case, this limitation may be the same as the lower cutoff discussed above – a sharp cutoff at a given object size $s_{\max}$ (Tebbens and Burroughs 2000). This means that all objects with larger sizes than $s_{\max}$ vanish. Due to the cumulative character of $P(s)$, this upper cutoff behavior affects the power-law behavior according to

$$P(s) \sim s^{-b} - s_{\max}^{-b} \quad \text{for} \quad s < s_{\max}.$$

For being more precise, we should regard both the upper and the lower cutoff values in an *upper-truncated Pareto distribution*

$$P(s) := \begin{cases} 1 & s \le s_{\min} \\ \dfrac{\left(\frac{s}{s_{\min}}\right)^{-b} - \left(\frac{s_{\max}}{s_{\min}}\right)^{-b}}{1 - \left(\frac{s_{\max}}{s_{\min}}\right)^{-b}} & \text{if} \quad s_{\min} < s < s_{\max} \\ 0 & s \ge s_{\max} \end{cases} . \tag{2.1}$$

Figure 2.1 illustrates this effect for $b = 1$, $s_{\min} = 1$, and different values of $s_{\max}$. The distributions follow a power law roughly up to areas that are about one order of magnitude below $s_{\max}$; so the finite-size effect costs one order of magnitude in the analysis. It can easily be seen that the loss strongly depends on the exponent b; it becomes more severe for smaller values of b. If

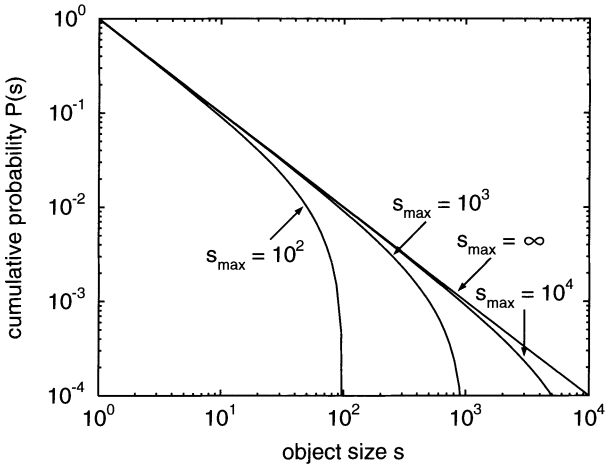

Fig. 2.1. Illustration of the cutoff effect in a cumulative size distribution with $b = 1$ and $s_{\min} = 1$.

b approaches zero – as it occurs in some of the models discussed later – the power-law behavior is completely hidden behind the cutoff effect.

In contrast to the original Pareto distribution, the truncated distribution cannot be directly recognized in a bilogarithmic plot. Instead, the function

$$\tilde{P}(s) \;:=\; P(s) + c$$

can be plotted for several, positive values of c. The function $\tilde{P}(s)$ turns into a power law if

$$c \;=\; \frac{1}{\left(\frac{s_{\max}}{s_{\min}}\right)^b - 1};$$

otherwise, the bilogarithmic plot exhibits some curvature. Too small values of c lead to concavity, whereas too large values result in convexity. As soon as the appropriate value of c has been found, the exponent b can be estimated from the slope of the approximated straight line. If desired, $s_{\min}$ and $s_{\max}$ can be estimated from $\tilde{P}(s_{\min}) = 1 + c$ and $\tilde{P}(s_{\max}) = c$.

This method is straightforward, but somewhat dangerous. Figure 2.2 shows how the size distributions of the fragments from Fig. 1.1 can be approximated by upper-truncated Pareto distributions. The distributions were obtained from the number of fragments plotted in Fig. 1.9, assuming a minimum object size $r_{\min} = \frac{1}{512}$.

In Sect. 1.6 we have already seen that the upper pattern from Fig. 1.1 is in principle scale-invariant, and that the deviation from a power-law distribution is a finite-size effect. Thus, it is not surprising that an upper-truncated Pareto distribution with $D = 1.58$ and $r_{\max} = 1$ fits the observed distribution perfectly. However, even the statistics of the non-fractal pattern can be approximated by an upper-truncated Pareto distribution. The diagram shows such a distribution with $D = 1.3$ and $r_{\max} = \frac{1}{2}$; these values were roughly estimated by visual correlation. Although not perfect, the distribution fits

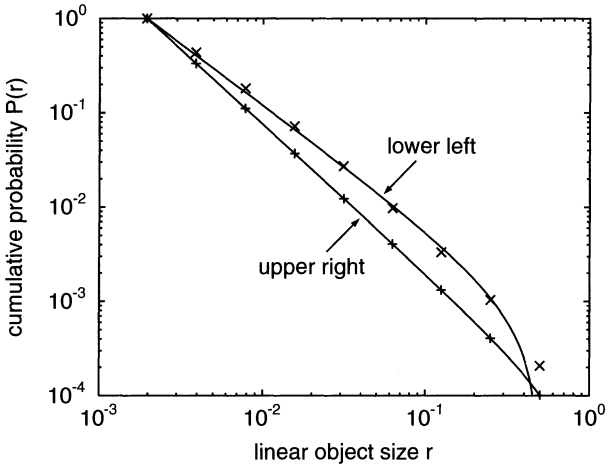

Fig. 2.2. Approximation of the size distributions of the fragments from Fig. 1.1 by upper-truncated Pareto distributions.

the data clearly better than a simple power law. A considerable deviation only occurs at the largest object sizes; but here we should remember that the corresponding data point describes just one object and cannot provide a reliable statistics at all.

Obviously, the additional degree of freedom introduced by the parameter $s_{\max}$ enables us to approximate various distributions considerably better than by a pure power law, no matter whether the distribution exhibits scale invariance or not. This problem becomes even worse if the observed range of scales is narrow. Thus, we should not simply fit both the exponent b and the cutoff value $s_{\max}$, but try to find a physically reasonable cutoff value with the help of a model for the origin of the pattern or for the process of observation. However, our example shows that even this approach does not guarantee reasonable results since our value $r_{\max} = \frac{1}{2}$ is not very far away from the correct value $r_{\max} = 1$.

After all, who knows if the finite-size effects are in fact as simple as discussed here? Why should all objects above a give size vanish? These objects might be present in principle, but may appear to be smaller than they are. In a first approach, we could assume that all objects which are originally larger than $s_{\max}$ are simply cut down to the size $s_{\max}$. The effect of this kind of truncation on the cumulative size distribution is essentially different from that discussed above because the cumulative distribution keeps its power-law shape:

$$P(s) = \begin{cases} 1 & s \le s_{\min} \\ \left(\frac{s}{s_{\min}}\right)^{-b} & \text{if } s_{\min} < s < s_{\max} \\ 0 & s_{\max} \le s \end{cases} .$$

In reality, cutoff behavior may be somewhere between these extreme cases. Thus, we need not only an estimate of $s_{\max}$, but also a model for the type of cutoff behavior in order to avoid misinterpretation.

However, the trouble arising from finite-size effects is just one facet of an inherent weakness of cumulative size distributions: Due to the cumulative character, the values of $P(s)$ for different values of s are not independent of each other, even if the considered sizes differ strongly. An additional object of size s affects $P(s')$ for all sizes $s' \leq s$. On the other hand, nearly all methods of fitting curves to measured values are based on the least-squares principle (Sect. 2.1). Statistical independence of the measured values is the essential assumption in this method. Unfortunately, this condition is really important; disregarding it leads to strongly biased results. Consequently, cumulative size distributions are in general not suitable for fitting power-law functions by means of least-squares techniques. In principle, the same limitation applies to visual correlation, i. e., drawing any function through the data points because we assume here, too, that the data points are scattered around the function symmetrically and independently of each other.

In summary, if the cumulative size distribution obtained from counting objects exhibits a clean power-law behavior over a reasonable range, there is not an argument against assuming a scale-invariant distribution and taking the slope in the bilogarithmic plot for the scaling exponent. On the other hand, if the distribution shows some curvature, fitting a power law is inappropriate, although this is frequently done. In this case, the binned probability densities considered in the next section provide a more reliable tool for examining scale-invariant properties.

2.3 Binning

When estimating a cumulative size distribution $P(s)$ from a given data set, the number of objects with size s or larger is considered. However, in the previous section we have recognized the problem that these numbers are not independent of each other. Alternatively, we can subdivide the range of object sizes into a number of intervals and count the objects in each interval. This technique is called *binning*.

Let us assume that the available range of object sizes is subdivided into n bins. Let further $s_0 < \ldots < s_n$ be the limits of the bins, so that the bin i is the interval $[s_{i-1}, s_i[$, and let n_i be the number of objects in this bin. How can we use the numbers n_i for finding out whether $P(s)$ follows a (truncated) power law and determine its exponent b? In a first step, we derive a relationship between the expected values $\overline{n}_i$ and the function $P(s)$ in case of a power-law distribution. The probability that a randomly chosen object falls into bin i is

$$P(s_{i-1}, s_i) \;\coloneqq\; P(s_{i-1}) - P(s_i),$$

so that

$$\overline{n}_i \;=\; NP(s_{i-1}, s_i) \;=\; N\left(P(s_{i-1}) - P(s_i)\right)$$

where N is the total number of considered objects. In case of a Pareto distribution (Eq. 1.3), we obtain

$$\overline{n}_i = N\, s_{\min}^b \left(s_{i-1}^{-b} - s_i^{-b}\right), \tag{2.2}$$

provided that $s_{\min} \leq s_0$. In order to recognize the behavior of $\overline{n}_i$, we choose a point $\hat{s}_i$ in each bin i according to

$$\hat{s}_i := \sqrt{s_{i-1} s_i}\, F\left(b, \tfrac{s_i}{s_{i-1}}\right) \quad \text{where} \quad F(b,x) = \left(b\, \frac{x^{\frac{1}{2}} - x^{-\frac{1}{2}}}{x^{\frac{b}{2}} - x^{-\frac{b}{2}}}\right)^{\frac{1}{b+1}}. \tag{2.3}$$

Using this definition, we can rewrite Eq. 2.2 in the form

$$\overline{n}_i = N\,(s_i - s_{i-1})\, b\, s_{\min}^b\, \hat{s}_i^{-(b+1)}. \tag{2.4}$$

This result can be related to the probability density

$$p(s) := -\tfrac{\partial}{\partial s} P(s) = \begin{cases} b\, s_{\min}^b s^{-(b+1)} & \text{if} \quad s > s_{\min} \\ 0 & \text{else} \end{cases} \tag{2.5}$$

in case of a Pareto distribution. This leads to

$$\overline{n}_i = N\,(s_i - s_{i-1})\, p(\hat{s}_i).$$

Therefore, the quantity $\frac{n_i}{N(s_i - s_{i-1})}$ provides an estimate of the probability density $p(\hat{s}_i)$ at a certain point $\hat{s}_i$ within the bin i. This result suggests to plot $\frac{n_i}{s_i - s_{i-1}}$ versus $\hat{s}_i$ in a bilogarithmic diagram, i. e., to consider

$$x_i := \log(\hat{s}_i) \quad \text{and} \quad y_i := \log\left(\frac{n_i}{s_i - s_{i-1}}\right). \tag{2.6}$$

According to Eq. 2.4, the plot should be a straight line with a slope of $-(b+1)$:

$$\overline{y}_i = \log\left(N\, b\, s_{\min}^b\right) - (b+1)\, x_i \tag{2.7}$$

for Pareto-distributed data.

Although this method is straightforward, there are a least three open questions:

1. How should the bins be chosen? We could, e. g., assume that all bins are of equal size, but it is not clear whether another subdivision is preferable.
2. How can the representative object size $\hat{s}_i$ of bin i be computed? The definition of $\hat{s}_i$ (Eq. 2.3) already involves the exponent b which is not known a priori.
3. What are the statistical properties of the random variables y_i? Are they independent of each other and Gaussian-distributed as required for applying linear regression or another least-squares technique discussed in Sect. 2.1?

Let us begin with the third, crucial question. In Sect. 2.2 we have already recognized the lack of statistical independence as a major drawback of cumulative size distributions. So the presumed advantage of the binning approach hinges on the statistical independence of the random variables y_i.

As so often, the question for statistical independence cannot be answered in general. Let us consider the statistical generalization of the fragmentation process discussed in Sect. 1.6 where each fragment is further fragmented with a given probability. Obviously, each decision whether a fragment is split up affects the number of objects at different sizes. Thus, the numbers n_i are not completely independent in this example. The same problem occurs if we, e. g., measure the sizes of N landslides in field. If N has been fixed in the beginning, the condition $\sum_{i=1}^{n} n_i = N$ disturbs the statistical independence of the values n_i and thus of the values y_i. Simply spoken, if one of the landslides had not happened, we would have analyzed another one.

However, this statistical dependence is considerably less severe than that occurring in cumulative size distributions. The problem completely vanishes if we do not assume a fixed total number of objects N, but analyze, e. g., the landslides in a certain region without knowing their number a priori. Let us assume the following model of data acquisition in the following: We make N' attempts to obtain an object, e. g., by looking at the certain point of the land surface, but are successful only with a probability q. If we obtain an object, we assume again that it falls into the bin i with the probability $P(s_{i-1}, s_i)$. Therefore, the probability that an attempt yields an object in the bin i is $qP(s_{i-1}, s_i)$. Let us now assume that the probability of success q is low, and that this is compensated by a large number of attempts: $q \to 0$ and $N' \to \infty$ while $N := N'q$ remains finite. Under this condition, the numbers n_i are in fact independent; and the probability p_{i,n_i} that the bin i contains exactly n_i objects follows a *Poisson distribution*:

$$p_{i,n_i} = \frac{(\overline{n}_i)^{n_i} \, \mathrm{e}^{-\overline{n}_i}}{n_i!}.$$

For practical purposes, the only difference towards the first model is that N is no longer the total number of observed objects, but a parameter describing the process of data acquisition.

As discussed in Sect. 2.1, the least-squares method requires Gaussian-distributed data, and the variances of the values y_i must be known. It is known from elementary statistics that the Poisson distribution converges towards a Gaussian distribution in the limit $\overline{n}_i \to \infty$. Since the variance is $\sigma_{n_i}^2 = \overline{n}_i$, the distribution becomes relatively narrow for large values of $\overline{n}_i$, so that taking the logarithm when defining y_i (Eq. 2.6) does not affect the Gaussian distribution. In the limit $\overline{n}_i \to \infty$, expected value and variance of y_i converge to

$$\overline{y}_i \to \log\left(\frac{\overline{n}_i}{s_i - s_{i-1}}\right) \quad \text{and} \quad \sigma_{y_i}^2 \to \left(\frac{\partial}{\partial n_i} y_i \Big|_{n_i = \overline{n}_i}\right)^2 \sigma_{n_i}^2 = \frac{1}{\overline{n}_i}. \quad (2.8)$$

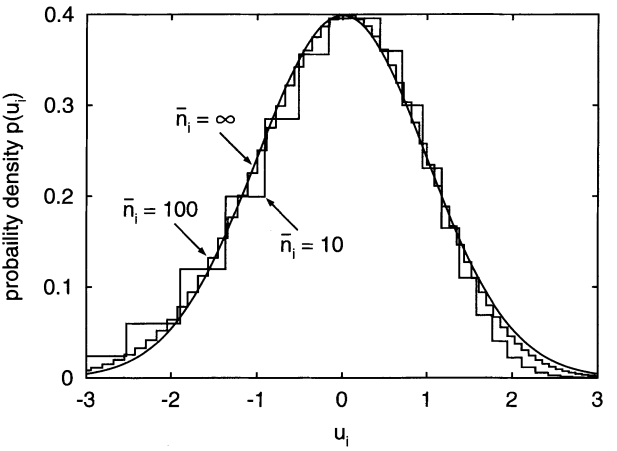

Fig. 2.3. Convergence of the logarithmically transformed Poisson distribution towards a Gaussian distribution.

However, since this result only holds in the limit $\overline{n}_i \to \infty$, it does not really help in case of finite data sets. So let us examine whether the criterion of Gaussian-distributed values y_i is satisfied at least approximately for finite values of n_i. According to Eq. 2.8, the quantity

$$u_i := \frac{y_i - \log\left(\frac{\overline{n}_i}{s_i - s_{i-1}}\right)}{\sqrt{\frac{1}{\overline{n}_i}}}$$

follows a standardized Gaussian distribution ($\overline{u}_i = 0$ and $\sigma^2_{u_i} = 1$) in the limit $\overline{n}_i \to \infty$. Figure 2.3 illustrates how the probability density of u_i converges towards a standardized Gaussian distribution if $\overline{n}_i$ increases. In principle, n_i is restricted to integer values, so that u_i is restricted to discrete values. This is illustrated by plotting staircase functions, although the use of Dirac's delta function (p. 45) would be more appropriate. Obviously, the distribution is somewhat skewed for finite values of $\overline{n}_i$, but is not very far away from a Gaussian distribution even for $\overline{n}_i = 10$. In the case $\overline{n}_i = 100$, the probability density is quite close to a standardized Gaussian distribution.

This result tells us that binning provides a sound method if the expected number of objects in each bin $\overline{n}_i$ is sufficiently large. An expected value of 100 objects in each bin is fine, but 10 objects may be sufficient, too. In combination with the result $\sigma^2_{y_i} \approx \frac{1}{\overline{n}_i}$ (Eq. 2.8), the technique of linear regression discussed in Sect. 2.1 can be directly applied to Eq. 2.7, so that the exponent b can be estimated to

$$b = -\mu - 1 = -\frac{\left(\sum \overline{n}_i\right)\left(\sum \overline{n}_i x_i y_i\right) - \left(\sum \overline{n}_i x_i\right)\left(\sum \overline{n}_i y_i\right)}{\left(\sum \overline{n}_i\right)\left(\sum \overline{n}_i x_i^2\right) - \left(\sum \overline{n}_i x_i\right)^2} - 1 . \quad (2.9)$$

In principle, this formula provides an estimate of b for arbitrarily binned data, provided that the number of objects in each bin is not too low. However, it

involves two properties which are still unknown – the expected values $\overline{n}_i$ and the representative sizes $\hat{s}_i$ defined in Eq. 2.3. Computing these properties requires knowledge of the exponent b or even of the whole distribution. So it seems that we are trapped in a circle. However, the impact of both properties is not as strong as it may seem. The expected values $\overline{n}_i$ are only used for determining the variances, so they only affect the relative weighting of the values y_i of different bins. Therefore, a rough estimate of $\overline{n}_i$ is sufficient; replacing $\overline{n}_i$ with the actual number n_i is a feasible solution. If a better approximation is desired, the following iterative scheme can be applied: In a first step, the numbers n_i are used instead of $\overline{n}_i$. Then the resulting estimate of the distribution provides a better estimate of $\overline{n}_i$, which can be used for improving the estimated distribution, and so on.

So let us now analyze to what extent the choice of the representative sizes $\hat{s}_i$ affects the result. According to Eq. 2.3, $\hat{s}_i$ is composed by the geometric mean value $\sqrt{s_{i-1}s_i}$ of the edges of the bins and a correction $F(b, \frac{s_i}{s_{i-1}})$ depending on the exponent b. Figure 2.4 gives this correction for different values of the exponent b. If $b < 1$, the representative object size $\hat{s}_i$ is larger than the geometric mean value and vice versa. The correction vanishes for $b = 1$. In general, the correction becomes small in case of small bins, i. e., if $\frac{s_i}{s_{i-1}} \approx 1$. When using the simple approximation $\hat{s}_i = \sqrt{s_{i-1}s_i}$, problems may only arise for relatively large bins at small object sizes.

Finally, the question for an appropriate choice of the bins should be addressed. In principle, Eq. 2.9 can be applied for arbitrary bins, provided that the number of objects in each bin is high enough. However, in most applications one of the following, simple approaches are used:

Linear binning means than all bins are equally sized, so that $s_i - s_{i-1}$ is constant. In this case, the bin width can be disregarded when defining y_i (Eq. 2.6), so that we can simply consider $y_i := \log n_i$. This straightforward strategy is widely used, but suffers from two problems. The first problem

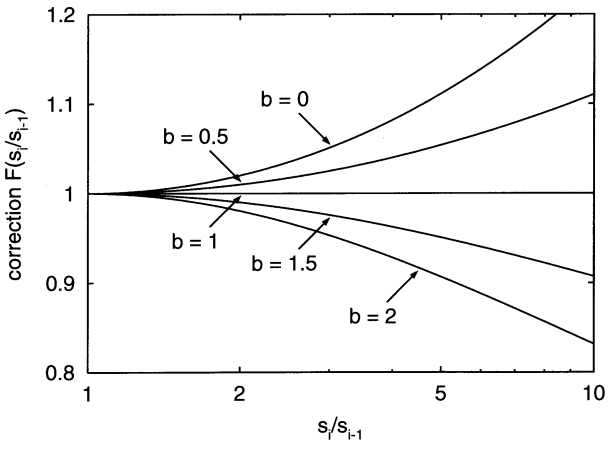

Fig. 2.4. Correction of the representative object size $\hat{s}_i$ compared to the geometric mean value for different values of the exponent b.

concerns the small object sizes because the condition $\frac{s_i}{s_{i-1}} \approx 1$ cannot be met then. Thus, taking the geometric mean for $\hat{s}_i$ is only appropriate if $b \approx 1$, otherwise the correction suggested by Eq. 2.3 should be applied. However, we will learn in the next section that the Pareto distribution is often inappropriate at small object sizes; mostly, the breakdown of scale invariance is more complicated than a sharp cutoff. Consequently, the bins at small object sizes are often useless, independent from any corrections of $\hat{s}_i$.

The second problem arising from linear binning is more severe. According to Eq. 2.4, $\overline{n}_i$ decreases like $\hat{s}_i^{-(b+1)}$. Even if the exponent b is small, this decrease is quite rapid, so that the number of objects in each bin is likely to become too small at large object sizes. Therefore, linear binning requires an extensive object statistics in order to perform a sound analysis.

Logarithmic binning avoids both problems at least partly by assuming that the ratio $\frac{s_i}{s_{i-1}}$ is the same for all bins. The nomenclature refers to the fact that the difference $\log s_i - \log s_{i-1}$ is constant, which means that the bin widths are equal on a logarithmic axis. Since the bin widths grow with increasing object sizes, the decrease of $\overline{n}_i$ at large object sizes is less severe than under linear binning. Let us select an arbitrary point $\hat{s}_i$ in each bin, not necessarily according to Eq. 2.3, but in such a way that its relative position in each bin is the same, i. e., $\frac{\hat{s}_i}{s_i}$ is constant. From Eq. 2.2 we immediately obtain

$$\overline{n}_i \;=\; N s_{\min}^b \left(1 - \left(\frac{s_{i-1}}{s_i}\right)^b\right) \left(\frac{\hat{s}_i}{s_{i-1}}\right)^b \hat{s}_i^{-b} \;\sim\; \hat{s}_i^{-b}.$$

Thus, the number of objects per bin only decreases like $\hat{s}_i^{-b}$, which coincides with the behavior of the cumulative size distribution, but is slower than in case of linear binning. As a result, the problem of low numbers of objects per bin at large sizes is less severe for logarithmic binning than for linear binning. Moreover, the question for an appropriate choice of $\hat{s}_i$ becomes unimportant under logarithmic binning since the power-law decrease holds for any choice of $\hat{s}_i$, provided that its relative position is the same in all bins.

The advantages of logarithmic binning compared to linear binning are obvious so far. But since every rose has its thorns, there should be some disadvantages, too. In principle, there is just one drawback; it arises if the data are already binned in a linear way. Transforming from linear to logarithmic bins is nearly impossible at small object sizes; the new bins are either non-logarithmic or so large that there is a severe loss in resolution. One may argue whether this is a disadvantage of the binning method: If we are stupid enough to put our data into linear bins first and lose the original data, it is up to us to get along with this problem. Unfortunately, things are not so easy as digital data become more and more important. As soon as the data are obtained from digitized maps or photographs or from discrete numerical models, the object sizes are discrete because they are multiples of the pixel size or the size of the basic cell in the model. Since this discrete structure is some kind of inherent linear binning, the problem is inevitable here.

As so often, the optimum is to be somewhere in the middle. The algorithm developed above can be applied to arbitrary bins. So it makes sense to keep the linear binning at small object sizes in digitally sampled data sets because the number of small objects is often large enough. In contrast, larger bins should be chosen at large object sizes in order to guarantee a sufficient number of objects in each bin.

2.4 Censoring

Up to now we have made a very simple assumption on the breakdown of scale invariance at small object sizes; the Pareto distribution is based on a sharp cutoff at a given minimum object size $s_{\min}$. But even if nature provides a perfect Pareto distribution, we cannot be sure that our limited facilities of observation enable us to recognize all events. The distribution of landslide sizes shown in Fig. 8.1 (p. 165) is quite illustrative in this sense. In a logarithmically binned plot, the power-law distribution even seems to be reverted at small sizes. Such a behavior may be an inherent property of the original distribution; it may be just a blurred cutoff at small sizes. On the other hand, we do in general not simply recognize all objects above a certain size; instead, the probability of recognizing an object decreases with decreasing object size. In other words, small objects tend to be undersampled; this effect is called *censoring*.

Censoring is ubiquitous in nearly all size distributions observed in nature. The effect can be described with the help of a function $c(s)$ with values in the range between 0 and 1; $c(s)$ shall be the probability that an object of size s is in fact recorded in the measurement. If $p(s)$ is the probability density of the original size distribution, the probability density $p_c(s)$ of the observed, censored distribution is proportional to the product $c(s)p(s)$. The condition

$$\int_0^\infty p_c(s)\,ds \;=\; P_c(0) - P_c(\infty) \;=\; 1$$

leads to

$$p_c(s) \;=\; \frac{c(s)\,p(s)}{\int_0^\infty c(s)\,p(s)\,ds},$$

or in terms of the cumulative size distribution:

$$P_c(s) \;=\; \int_s^\infty p_c(s')\,ds' \;=\; \frac{\int_s^\infty c(s')\,p(s')\,ds'}{\int_0^\infty c(s')\,p(s')\,ds'} \;=\; \frac{\int_s^\infty c(s')\,\frac{\partial}{\partial s'}P(s')\,ds'}{\int_0^\infty c(s')\,\frac{\partial}{\partial s'}P(s')\,ds'}.$$

In some cases, it may be possible to find a physical model for the effect of censoring. For instance, the probability of not detecting an earthquake is mainly determined by its strength and location and by the density and spatial distribution of the seismic stations. In this case, the original distribution $P(s)$ can in principle be reconstructed from the measured distribution $P_c(s)$.

However, estimating the effect quantitatively is often difficult. If we, e. g., consider the sizes of landslides, the finite resolution of maps and photos obviously leads to censoring. However, landform evolution processes themselves introduce a second source of censoring which may be even more important: The scars of landslides may be healed through time due to erosion, and large landslides may erase the scars of smaller ones.

If the effect of censoring cannot be quantified, two methods can be applied. The established method is hoping that the effect of censoring vanishes above a certain object size s_c, so that $c(s) = 1$ (or at least $c(s) = \text{const}$) for $s \geq s_c$. In this case, the part of the data where $s \geq s_c$ can be used for estimating the probability density. However, this method has two disadvantages (Stark and Hovius 2001): First, only a small part of the data set may meet the condition, so that the statistics may become too small. Second, $c(s)$ is often not exactly constant even for large object sizes; and this leads to a concavity in the bilogarithmic plot. If a power law is fitted then, the exponent b is underestimated.

Let us focus on the example of landslides because the size distributions of landslides will be discussed later in this book (Chap. 8). Stark and Hovius (2001) introduced a method which is based on a parametric approach. This approach is just an example of treating censoring effects, but a quite illustrative one. They describe the effect of censoring by a function with two adjustable parameters. Roughly speaking, one of these parameters describes the object size where a transition between power-law behavior and censoring-dominated behavior occurs; and the second parameter describes how smooth the crossover is. In order to keep the equations as simple as possible, we assume that the original (uncensored) distribution is a Pareto distribution without any upper cutoff, and that the lower cutoff value $s_{\min}$ is so small that censoring is very strong here. In other words, the lack of small objects is rather caused by censoring than by the cutoff in the original distribution. Under this simplification, the censored distribution suggested by Stark and Hovius (2001) reads

$$P_c(s) = \left(1 + \left(\frac{s}{s_c}\right)^{-b}\right)^{-\gamma},$$

where s_c and γ are the new parameters. The corresponding censored probability density reads

$$p_c(s) = -\frac{\partial}{\partial s} P_c(s) = \gamma \left(1 + \left(\frac{s}{s_c}\right)^{-b}\right)^{-(1+\gamma)} b\, s_c^b\, s^{-(b+1)}. \qquad (2.10)$$

As expected, the probability density decreases like $s^{-(b+1)}$ at large object sizes in coincidence with the uncensored Pareto probability density. This result just reflects the assumption that the censoring effect vanishes asymptotically for

large object sizes. For small object sizes, $p_c(s)$ behaves like a power law, too, but with a positive exponent:

$$p_c(s) \sim s^{b\gamma-1} \quad \text{for} \quad s \to 0.$$

For this reason, Stark and Hovius (2001) introduced the term *double Pareto distribution.* However, the power-law behavior at small object sizes is not characteristic with respect to most data sets obtained from nature; they normally do not reach far into this region. Thus, the interesting range is that of the decreasing power law and the region of crossover between increasing and decreasing power law.

Figure 2.5 shows the probability densities and the cumulative size distributions for $b = 1$, $s_c = 1$, and different values of γ. The curves for other values of s_c have the same shape. Compared to the curves for $s_c = 1$, the cumulative size distributions are just shifted by a factor s_c to the right, and

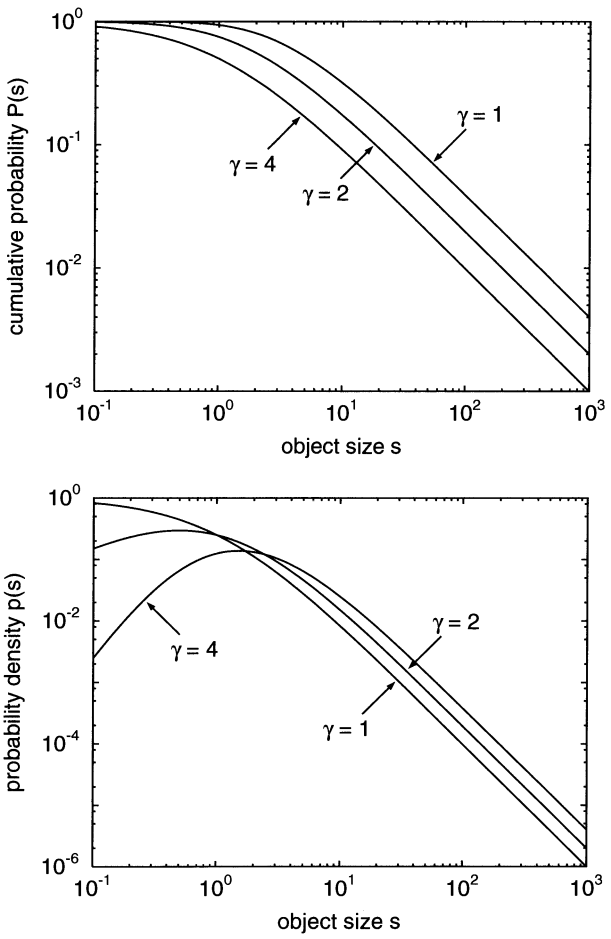

Fig. 2.5. Double Pareto distributions for $b = 1$, $s_c = 1$, and different values of γ. Upper diagram: cumulative size distribution; lower diagram: probability density.

the probability densities by a factor s_c to the right, and by the same factor downwards. Therefore, the parameter s_c determines at which object sizes the crossover between both power-law regimes occurs, although the location of the maximum in the probability density depends on b and γ, too.

The double Pareto distribution has been applied to landslide statistics (Stark and Hovius 2001), resulting in considerable modifications of the estimated exponents compared to power-law fits made earlier. Despite the improvement in approximation, some cautioning remarks should be made. The behavior of the censoring function $c(s)$ can be reconstructed from Eq. 2.10:

$$c(s) \sim \left(1 + \left(\frac{s}{s_c}\right)^{-b}\right)^{-(1+\gamma)} .$$

This result reveals a problem of the approach: Since $c(s)$ depends on the exponent b of the original size distribution, the double Pareto approach does not comply with the original view on censoring since we assumed that the probability that an object is recorded only depends on its size. From a technical point of view, the problem can be fixed by replacing the exponent b in the censoring function by an exponent b' which is independent from b. Then, the distribution contains four instead of three parameters which must be adjusted.

In summary, the double Pareto distribution is a function for fitting data sets rather than a physically reasonable model for censoring. Fitting a double Pareto distribution may therefore be misleading. Since it contains three (or even four) adjustable parameters, we should not be surprised if a double Pareto distribution approximates a given data set much better than a simple Pareto distribution does. The problem is similar to that occurring when approximating size distributions by upper-truncated power laws (Sect. 2.2); the method is hardly suitable for deciding whether the original (uncensored) distribution is scale-invariant or not. Consequently, the double Pareto distribution should only be applied if the data set shows a clean power-law behavior over a reasonable range; otherwise, a scale-dependent distribution may be misinterpreted to be scale-invariant, and the obtained exponent has no meaning at all. On the other hand, this method may be suitable for obtaining more reliable statistics and thus a better estimate of the scaling exponent if scale invariance can be guaranteed. Similar arguments apply to other methods of treating censoring effects unless they are based on a physical model.

3. Self-Affine Time Series

So far we have focused on spatial scales. However, when dealing with processes, time introduces an additional axis in coordinate space. The simplest point of view considers time just as a parameter. We may analyze the state of the system at any time and apply all we have learned about scale invariance and scale dependence to these *time slices.*

But what keeps us from considering spatial and temporal scaling properties simultaneously? Assume that we go to a lake with a model of the Titanic and a video camera. Even apart from the missing icebergs, the result will be disappointing. The audience of our film will immediately recognize that it is just a model. On the other hand, downsized models were state of the art in film production for a long time. How did they make their downscaled scenarios realistic? In principle, they did not only rescale the sizes of their objects, but rescaled time, too. If we use a high-speed camera for recording and choose a lower speed for playback, the result may be much better.

If we have found a playback speed which leads to a realistic impression of the sinking ship, we have recognized spatio-temporal scale invariance – the characteristic properties of the process persist under a simultaneous rescaling of the spatial and temporal coordinates. However, this does not mean that we must rescale time by the same factor as the spatial coordinates. If the scaling factors concerning different coordinates differ, we speak of *anisotropic scaling.* Objects which are statistically similar under anisotropic scaling are called *self-affine* fractals in order to distinguish them from the isotropic, *self-similar* fractals introduced in Chap. 1. Like many ideas on fractals, the idea of self-affine scale invariance was introduced by Mandelbrot (1985); it is discussed in detail in most of the books on fractals which have appeared since then (e. g. Feder 1988; Turcotte 1997).

Self-affine scaling is likely to occur in spaces with physically different coordinate axes, e. g., if time is involved as discussed above. However, self-affine scale invariance may also be found in two- or three-dimensional Euclidean space if the pattern-forming processes are anisotropic. The earth's surface is an example for an anisotropy that is introduced by the direction of gravity. Since gravity is the major component in landform evolution processes, we may expect self-affine scale invariance rather than self-similar scale invariance here. The surfaces shown in Fig. 1.12 (p. 20) are in fact self-affine fractals.

In the following, we focus on *time series*, i. e., on functions describing a time-dependent, scalar quantity $f(t)$ such as the position of a particle in a one-dimensional space, the water table in a lake, the amount of rainfall per time or a temperature. In order to keep the chapter at a reasonable length, we only address some central aspects. More thorough descriptions are given in the books mentioned above and in a series of two papers (Malamud and Turcotte 1999; Pelletier and Turcotte 1999) presenting both theory and applications.

3.1 Brownian Motion

The motion of atoms or molecules in a gas is governed by a huge number of collisions. For simplicity, we assume the atoms or molecules to be particles in the sense of classical mechanics instead of a quantum-mechanical treatment. As a result of the huge number of collisions, a statistical description is preferable to a deterministic one, although each collision is a deterministic process in principle. The simplest statistical description of this process is called *Brownian motion*; it belongs to the class of *random-walk* processes. Let us focus on the one-dimensional case; the generalization is straightforward. The position of a particle is a time-dependent random variable $x(t)$ starting at $x(0) = 0$ with a random velocity v_1. This velocity persists until $t = \delta t$; then it is replaced by another random velocity v_2 which is statistically independent from v_1. This velocity persists until $t = 2\delta t$, and so on. All velocities v_i are Gaussian-distributed random variables with expected values $\overline{v_i} = 0$ and identical variances $\sigma_v^2 \coloneqq \overline{v_i^2}$. Compared to the physical process, Brownian motion introduces three idealizations:

- Further degrees of freedom such as rotation are neglected.
- The time span δt between two collisions is assumed to be constant, while it is exposed to a statistical variation in reality.
- The velocities before and after a collision are assumed to be statistically independent, while there is some correlation in reality.

The assumption of Gaussian-distributed velocities is not mentioned in this list because it is justified by statistical physics. In thermodynamic equilibrium, the kinetic energies of the particles follow a Boltzmann distribution which is equivalent to a Gaussian distribution of the velocities.

The upper part of Fig. 3.1 shows five examples of Brownian motion, generated with $\sigma_v^2 = 1$ and $\delta t = 1$. The lower diagram illustrates that Brownian motion is not self-similar with respect to simultaneous spatial and temporal scaling. It shows the dashed section of the upper diagram; it is a quarter of the original plot in width and height. Obviously, the curves are not similar to the original traces; they are steeper and tend to leave the section. In contrast, the right-hand diagram shows an anisotropically scaled section – the solid one in the upper plot. This section is a quarter of the original diagram in length, but one half in height. The traces in the anisotropically magnified section

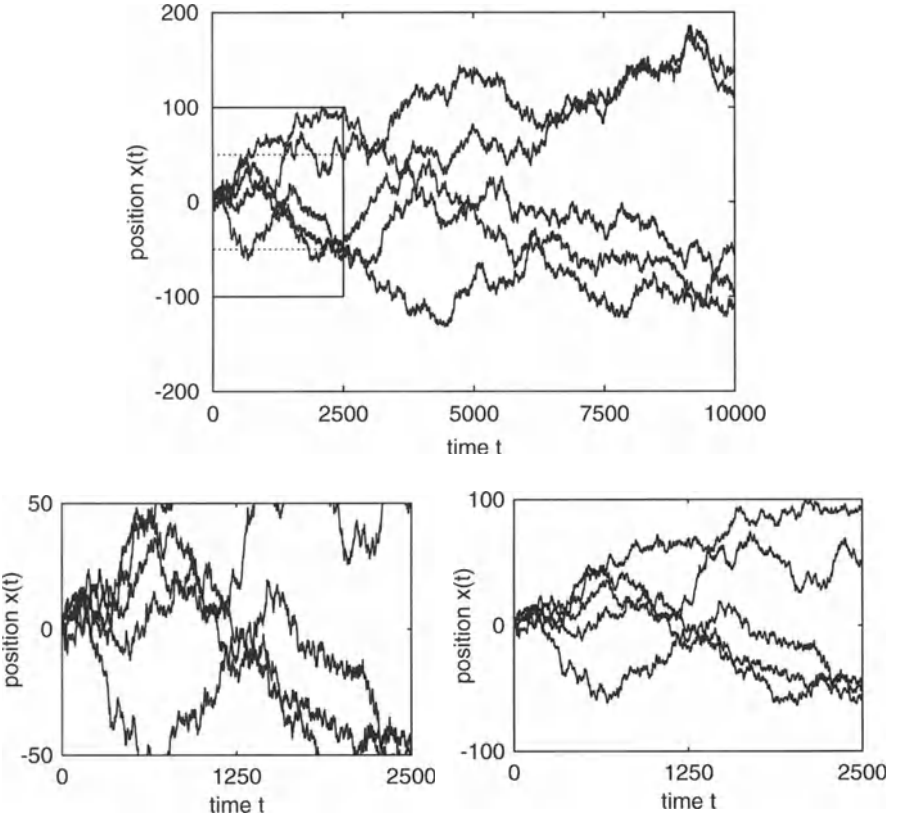

Fig. 3.1. Five examples of Brownian motion (top) and magnified sections (below). Left: isotropic magnification (dashed section of the upper diagram). Right: anisotropic magnification (solid section of the upper diagram).

look similar to those in the upper diagram. This result suggests self-affine scale invariance. The following analysis confirms this conjecture. At the time $t = n\delta t$, the position of the particle is

$$x(t) = \sum_{i=1}^{n} \delta t \, v_i .$$

Thus, $x(t)$ is a linear combination of independent, Gaussian-distributed random variables. As a result of elementary statistics, such a combination is Gaussian-distributed, too; expected value $\overline{x(t)}$ and variance $\sigma(t)^2$ are

$$\overline{x(t)} = \sum_{i=1}^{n} \delta t \, \overline{v_i} = 0,$$

$$\sigma(t)^2 = \overline{(x(t) - \overline{x(t)})^2} = \sum_{i=1}^{n} \sum_{j=1}^{n} \delta t^2 \, \overline{v_i v_j} = n \, \sigma_v^2 \, \delta t^2 = \sigma_v^2 \, \delta t \, t.$$

Therefore, $x(t)$ follows a Gaussian distribution with an expected value of zero and a standard deviation $\sigma(t)$ that increases through time like $\sqrt{t}$. This result explains the anisotropic scaling behavior of Brownian motion; if we rescale time by a factor λ, we have to rescale the spatial coordinate by a factor $\sqrt{\lambda}$ in order to preserve the statistical properties of the process.

In principle, this is only true for the discrete stages $t = n\delta t$. Self-affine scale invariance in the strict sense only arises in the limit $\delta t \to 0$. However, the transition may not be performed for a constant value σ_v^2 because $\sigma(t)^2 \to 0$ then. A reasonable result in the limit $\delta t \to 0$ can only be obtained if the velocities increase according to $\sigma_v^2 = \frac{c}{\delta t}$ where c is a constant. This leads to $\sigma(t)^2 = ct$ for all times t.

The relationship $\sigma(t) \sim \sqrt{t}$ is often called *root-t law*; it is ubiquitous in diffusive transport phenomena such as heat transport and spreading of pollutants in the ground. The diameter of a cloud of particles or solutes does not increase linearly through time as it might be expected first, but only proportionally to the square root of the time. In general, diffusive transport of particles or solutes is described by the diffusion equation

$$\frac{\partial}{\partial t} c(x,t) \;=\; \kappa \, \frac{\partial^2}{\partial x^2} c(x,t)$$

for the concentration $c(x,t)$ where κ is the diffusivity. If we consider the motion of a large number of particles starting at the location $x = 0$, we see that the diffusion equation is closely related to the rules of Brownian motion. Since the position of each particle is a Gaussian-distributed random function with $\overline{x(t)} = 0$ and $\sigma(t)^2 = ct$, the resulting concentration is

$$c(x,t) \;=\; \frac{1}{\sqrt{2\pi}\,\sigma(t)} \, \mathrm{e}^{-\frac{x^2}{2\sigma(t)^2}} \;=\; \frac{1}{\sqrt{4\pi ct}} \, \mathrm{e}^{-\frac{x^2}{4ct}}.$$

This function satisfies the diffusion equation if $\kappa = \frac{1}{2}c$. Thus, the diffusion equation provides an alternative description of Brownian motion, and diffusion exhibits the same self-affine scaling properties as Brownian motion.

3.2 White Noise

We now consider the velocities involved in Brownian motion as a statistical process. Their properties are even simpler than those of the particle traces since the velocities at different times are just independent, Gaussian-distributed random variables. This statistical process is called *white noise*. In analogy to Brownian motion, we distinguish between *physical* white noise (based on finite time intervals δt) and *idealized* white noise (in the limit $\delta t \to 0$). For defining physical white noise, we just have to recapitulate the assumptions on the velocities in Brownian motion. Since white noise is not necessarily associated with velocities, we use $f(t)$ of instead of v_i and σ instead of σ_v in the following. White noise is a statistical process where a random variable $f(t)$ with the following properties is assigned to each time t:

- The random variables $f(t)$ are Gaussian-distributed with expected values $\overline{f(t)} = 0$ and constant variances $\sigma^2 = \overline{f(t)^2}$.
- The random variables $f(t)$ and $f(t')$ coincide if t and t' are within the same time interval. Otherwise, $f(t)$ and $f(t')$ are statistically independent.

In order to preserve the analogy to Brownian motion, we assume that σ^2 depends on δt in the limit $\delta t \to 0$ according to the relationship $\sigma^2 = \frac{c}{\delta t}$ derived in the previous section. The properties of the Gaussian-distributed random variables $f(t)$ can be summarized in the form

$$\overline{f(t)} = 0 \quad \text{and} \quad \overline{f(t)f(t')} = \begin{cases} \frac{c}{\delta t} & \text{if } \left[\frac{t}{\delta t}\right] = \left[\frac{t'}{\delta t}\right] \\ 0 & \text{else} \end{cases} .$$

The brackets [] denote the largest integer number which is smaller than the argument within the brackets. However, the function $\overline{f(t)f(t')}$ is not well-defined in the limit $\delta t \to 0$. It converges towards zero for $t \neq t'$, while it diverges for $t = t'$. In order to quantify this behavior, we consider an arbitrary, continuous function $\psi(t)$ and the integral

$$\int_{-\infty}^{\infty} \overline{f(t)f(t')}\,\psi(t')\,dt' = \frac{c}{\delta t} \int_{\delta t\left[\frac{t}{\delta t}\right]}^{\delta t\left[\frac{t}{\delta t}\right]+\delta t} \psi(t')\,dt'.$$

The integrand at the right-hand side converges towards $\psi(t)$ in the limit $\delta t \to 0$, so that

$$\lim_{\delta t \to 0} \int_{-\infty}^{\infty} \overline{f(t)f(t')}\,\psi(t')\,dt' = \frac{c}{\delta t}\,\psi(t)\,\delta t = c\,\psi(t). \tag{3.1}$$

This result can be used for characterizing $\overline{f(t)f(t')}$ in the limit $\delta t \to 0$. The formalism can be streamlined by introducing the following notation:

> Consider a class of functions $\delta_\epsilon(t,t')$ depending on a parameter ϵ. If
>
> $$\lim_{\epsilon \to 0} \int_{-\infty}^{\infty} \delta_\epsilon(t,t')\,\psi(t')\,dt' = \psi(t)$$
>
> for all continuous functions $\psi(t)$, the formal limit
>
> $$\delta(t-t') := \lim_{\epsilon \to 0} \delta_\epsilon(t,t')$$
>
> is called *Dirac's delta function.*

In this sense, we can write

$$\lim_{\delta t \to 0} \overline{f(t)f(t')} = c\,\delta(t-t').$$

This formalism leads to a formal definition of (idealized) white noise:

White noise is a statistical process where a Gaussian-distributed random variable $f(t)$ with

$$\overline{f(t)} = 0 \quad \text{and} \quad \overline{f(t)f(t')} = c\,\delta(t-t')$$

is assigned to each time t.

3.3 Fourier Transforms

Properties of time series are often examined in the frequency representation with the help of *Fourier transforms* rather than in the original signal. Since Fourier transforms are a standard topic in calculus, we only recapitulate the most important results of the mathematical theory here.

Let $f(t)$ be a (in general complex) function. If $\int_{-\infty}^{\infty} |f(t)|^2\,dt$ exists, $f(t)$ can be composed of complex exponential functions:

$$f(t) = \int_{-\infty}^{\infty} \phi(\nu)\,\mathrm{e}^{2\pi\mathrm{i}\nu t}\,d\nu. \tag{3.2}$$

This procedure is called inverse Fourier transform. The (complex) *Fourier amplitude* $\phi(\nu)$ quantifies the contribution of the frequency ν; it can be computed by the forward Fourier transform

$$\phi(\nu) = \int_{-\infty}^{\infty} f(t)\,\mathrm{e}^{-2\pi\mathrm{i}\nu t}\,dt. \tag{3.3}$$

However, data analysis mainly concerns real functions $f(t)$. In this case, $\phi(\nu)$ remains a complex number, but we can easily see that $\phi(-\nu) = \phi(\nu)^\star$ where the star denotes the conjugated number.

The Fourier transform is an established tool in many fields, e. g., in solving differential equations. Let us consider a function $f(t)$ and its derivative $f'(t) \coloneqq \frac{\partial}{\partial t} f(t)$. The Fourier representation of $f'(t)$ is

$$f'(t) = \frac{\partial}{\partial t} \int_{-\infty}^{\infty} \phi(\nu)\,\mathrm{e}^{2\pi\mathrm{i}\nu t}\,d\nu = \int_{-\infty}^{\infty} \phi(\nu)\,2\pi\mathrm{i}\nu\,\mathrm{e}^{2\pi\mathrm{i}\nu t}\,d\nu,$$

so that its Fourier amplitude is $\phi'(\nu) = 2\pi\mathrm{i}\nu\phi(\nu)$. Thus, computing the derivative becomes even simpler after switching to the Fourier representation; it reduces to multiplying the Fourier amplitudes by $2\pi\mathrm{i}\nu$. In return, integrating a function corresponds to dividing the Fourier amplitudes by $2\pi\mathrm{i}\nu$.

In the previous sections we discussed the statistical properties of Brownian motion and white noise. Since a function $f(t)$ can be uniquely characterized by its Fourier amplitudes, we may expect that both processes can be characterized by certain properties of their Fourier amplitudes $\phi(\nu)$, too.

Let us begin with white noise. As white noise consists of Gaussian-distributed random variables $f(t)$ and the Fourier transform (Eq. 3.3) is some kind of linear combination of the values $f(t)$, both the real and the imaginary parts of the Fourier amplitudes $\phi(\nu)$ are Gaussian-distributed random variables, too. From Eq. 3.3 and the definition of white noise we obtain

$$\overline{\phi(\nu)} \;=\; \int_{-\infty}^{\infty} \overline{f(t)}\, \mathrm{e}^{-2\pi \mathrm{i}\nu t}\, dt \;=\; 0.$$

For computing $\overline{\phi(\nu)\phi(\nu')^\star}$, we again need Dirac's delta function (p. 45). So let us consider an arbitrary, continuous function $\psi(\nu)$ and the integral

$$\begin{aligned}\int_{-\infty}^{\infty} \overline{\phi(\nu)\phi(\nu')^\star}\, \psi(\nu')\, d\nu' \;&=\; \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \overline{f(t)f(t')}\, \mathrm{e}^{-2\pi \mathrm{i}\nu t} \mathrm{e}^{2\pi \mathrm{i}\nu' t'} \psi(\nu') \\ &\qquad\qquad dt\, dt'\, d\nu' \\ &=\; c\int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} \psi(\nu')\, \mathrm{e}^{2\pi \mathrm{i}\nu' t} d\nu' \right) \mathrm{e}^{-2\pi \mathrm{i}\nu t} dt\end{aligned}$$

according to Eq. 3.1. Since last line is nothing but inverse and forward Fourier transform, we obtain

$$\int_{-\infty}^{\infty} \overline{\phi(\nu)\phi(\nu')^\star}\, \psi(\nu')\, d\nu' \;=\; c\,\psi(\nu),$$

or in short notation:

$$\overline{\phi(\nu)\phi(\nu')^\star} \;=\; c\,\delta(\nu - \nu').$$

In combination with the relationship $\phi(-\nu) = \phi(\nu)^\star$ for real functions $f(t)$, this result determines the statistical properties of both the real and the imaginary parts of $\phi(\nu)$:

$$\begin{aligned}\overline{\mathrm{Re}(\phi(\nu))\mathrm{Re}(\phi(\nu'))} \;&=\; \tfrac{1}{2}\,\mathrm{Re}\left(\overline{\phi(\nu)\phi(\nu')^\star} + \overline{\phi(\nu)\phi(-\nu')^\star}\right) \\ &=\; \tfrac{1}{2}\,c\,(\delta(\nu-\nu') + \delta(\nu+\nu')), \\ \overline{\mathrm{Im}(\phi(\nu))\mathrm{Im}(\phi(\nu'))} \;&=\; \tfrac{1}{2}\,\mathrm{Re}\left(\overline{\phi(\nu)\phi(\nu')^\star} - \overline{\phi(\nu)\phi(-\nu')^\star}\right) \\ &=\; \tfrac{1}{2}\,c\,(\delta(\nu-\nu') - \delta(\nu+\nu')), \\ \overline{\mathrm{Re}(\phi(\nu))\mathrm{Im}(\phi(\nu'))} \;&=\; \tfrac{1}{2}\,\mathrm{Im}\left(\overline{\phi(\nu)\phi(-\nu')^\star} - \overline{\phi(\nu)\phi(\nu')^\star}\right) \;=\; 0.\end{aligned}$$

Thus, white noise can be uniquely characterized by the properties of its Fourier amplitudes $\phi(\nu)$ instead of the original representation $f(t)$. So we can give an alternative, but equivalent definition of white noise:

> White noise is a statistical process where both the real and the imaginary parts of the Fourier amplitudes $\phi(\nu)$ are Gaussian-distributed random variables with
>
> $$\overline{\phi(\nu)} = 0 \quad \text{and} \quad \overline{\phi(\nu)\phi(\nu')^\star} \;=\; c\,\delta(\nu-\nu').$$

Both characterizations of white noise are not only equivalent, but also formally the same, except for the fact that $f(t)$ is a real random function, while $\phi(\nu)$ is a complex random function. In other words, the Fourier transform of white noise is white noise, too.

The properties of white noise concerning the Fourier amplitudes motivate its name from the analogy to white light. In the mean, all frequencies contribute equally to white noise; so white noise is a homogeneous mixture of the whole spectrum of frequencies from zero to infinity. However, this analogy should not be overinterpreted because white light is a mixture of a wide spectrum of frequencies, but not a homogeneous mixture of all frequencies.

Characterizing Brownian motion by means of its Fourier representation is straightforward now. Brownian motion arises from integrating white noise, and we have already seen that integrating is equivalent to dividing the Fourier amplitudes by $2\pi\mathrm{i}\nu$. This leads to

$$\overline{\phi(\nu)\phi(\nu')^\star} = \frac{c}{2\pi\mathrm{i}\nu\,(2\pi\mathrm{i}\nu')^\star}\,\delta(\nu-\nu') = \frac{c}{(2\pi\nu)^2}\,\delta(\nu-\nu').$$

So we can give an alternative definition of Brownian motion:

> Brownian motion is a statistical process where both the real and the imaginary parts of the Fourier amplitudes $\phi(\nu)$ are Gaussian-distributed random variables with
>
> $$\overline{\phi(\nu)} = 0 \quad \text{and} \quad \overline{\phi(\nu)\phi(\nu')^\star} \sim \nu^{-2}\,\delta(\nu-\nu').$$

Strictly speaking, this definition of Brownian motion is not completely equivalent to the one given in Sect. 3.1. According to the original definition, Brownian motion starts at $f(0) = 0$ and is undefined for all times $t < 0$. In contrast, the alternative definition refers to a function whose derivative is white noise for all times t. Roughly speaking, the new definition describes Brownian motion starting at $t = -\infty$. But since the essential properties are the same, the difference is not crucial.

3.4 Fractional Brownian Motion

Both white noise and Brownian motion are processes where a Gaussian-distributed random variable $f(t)$ is assigned to each time t. The characteristic difference between both concerns the temporal correlations. In white noise, the values $f(t)$ and $f(t')$ are totally uncorrelated if $t \neq t'$, while there is some correlation in Brownian motion. In the previous section we characterized this difference in the frequency representation. We have seen that the temporal correlation is not reflected by a correlation of the Fourier amplitudes; $\phi(\nu)$ and $\phi(\nu')$ are uncorrelated if $\nu \neq \nu'$ in both white noise and Brownian

motion. The temporal correlation only affects the magnitude of the Fourier amplitudes as a function of the frequency; we obtained

$$\overline{\phi(\nu)\phi(\nu')^\star} \;=\; P(\nu)\,\delta(\nu-\nu')$$

where $P(\nu)$ is constant for white noise and decreases like ν^{-2} for Brownian motion. The function $P(\nu)$ is denoted *power spectrum* of the process. Obviously, $P(\nu)$ quantifies the contribution of the frequency ν in some sense, but let us refrain from going further into detail.

The power spectra of white noise and Brownian motion suggest a generalization towards power-law functions $P(\nu)$ with arbitrary exponents:

Fractional Brownian motion (FBM) is a statistical process where both the real and the imaginary parts of the Fourier amplitudes $\phi(\nu)$ are Gaussian-distributed random variables with

$$\overline{\phi(\nu)} \;=\; 0 \quad \text{and} \quad \overline{\phi(\nu)\phi(\nu')^\star} \;=\; P(\nu)\,\delta(\nu-\nu')$$

with

$$P(\nu) \;\sim\; |\nu|^{-\beta}.$$

The exponent β is called *spectral exponent* of the process.

White noise is FBM with $\beta = 0$, while Brownian motion is FBM with $\beta = 2$. From the argument which showed that the Fourier amplitudes of white noise are Gaussian-distributed random variables with zero mean value, we can see that the random variables $f(t)$ of FBM are Gaussian-distributed around zero.

The term FBM can be motivated from its relationship to fractional derivatives or fractional integration. In Sect. 3.3 we have learned that deriving a function $f(t)$ with respect to the time corresponds to multiplying its Fourier amplitudes $\phi(\nu)$ by $2\pi\mathrm{i}\nu$. So we can define a *fractional derivative*

$$\frac{\partial^\alpha}{\partial t^\alpha} f(t) \;=\; \int_{-\infty}^{\infty} \phi(\nu)\,(2\pi\mathrm{i}\nu)^\alpha\,\mathrm{e}^{2\pi\mathrm{i}\nu t}\,d\nu$$

for arbitrary real numbers α. Applying this definition for negative values α leads to *fractional integration*. In this sense, FBM with a spectral exponent β arises from applying the fractional integration with $|\alpha| = \frac{1}{2}\beta$ to white noise.

Fig. 3.2 gives examples of FBM for different spectral exponents β between 0 and 3, generated according to the algorithm described in the following section. The signal with $\beta = 1$ just in the middle between white noise ($\beta = 0$) and Brownian motion ($\beta = 2$) is called *pink noise*, *1/f noise* or *flicker noise*. In terms of fractional derivatives, pink noise arises from applying half of a derivative to Brownian motion or half of an integration to white noise. Pink noise is interesting not only because many signals occurring in nature are rather pink noise than white noise or Brownian motion (e. g. Bak 1996; Jensen 1998). We will see in Sect. 3.10 that pink noise defines the limit of predictability.

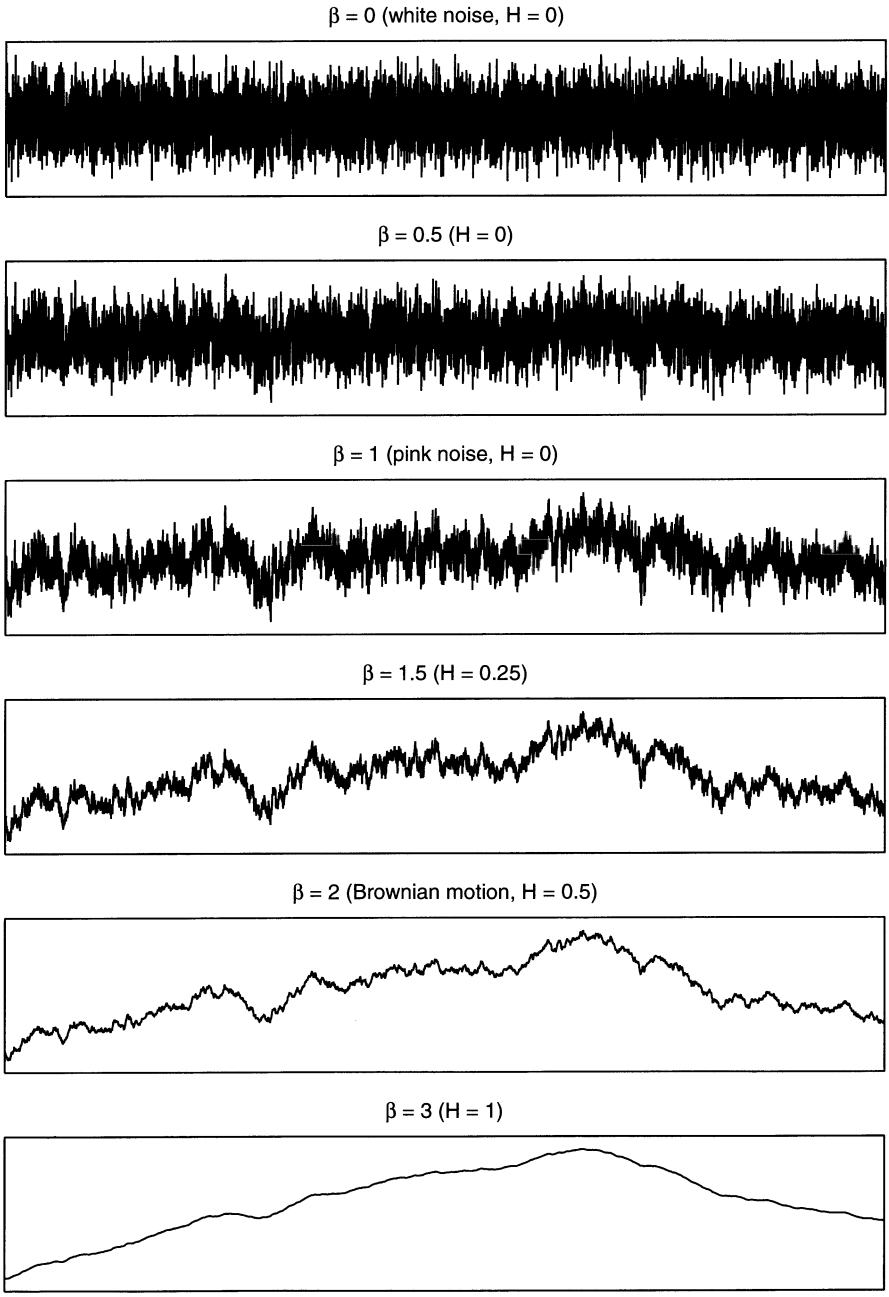

Fig. 3.2. Examples of FBM with different spectral exponents β between 0 and 3. The signal $f(t)$ is plotted versus t in arbitrary units. The meaning of the variable H is explained in Sect. 3.6.

3.5 Generating FBM

Obviously, the computer-generated curves shown in Fig. 3.2 cannot be examples of FBM in the strict sense. On a computer, only discrete approximations to the Fourier transform involving a finite number of frequencies can be handled. Therefore, the power spectrum of FBM must be approximated by a finite number of frequencies ν_j in order to obtain an approximation to FBM. It is convenient to choose equally spaced frequencies $\nu_j = \frac{j}{2n\delta t}$ for integer numbers j from $-n$ to n. The parameter δt determines the resolution of the resulting time series $f(t)$; it is half of the period of the highest frequency ν_n. Therefore, it makes no sense to look at $f(t)$ in shorter time intervals than δt. Let us consider a random function with the Fourier amplitudes

$$\phi(\nu) = \sum_{j=-n}^{n} \phi_j \, \delta(\nu - \nu_j),$$

where both the real and imaginary parts of the *Fourier coefficients* ϕ_j are Gaussian-distributed random variables with

$$\overline{\phi_j} = 0 \quad \text{and} \quad \overline{\phi_j \phi_k^\star} = \begin{cases} c\,|\nu_j|^{-\beta} & \text{if} \quad j = k \\ 0 & \text{else} \end{cases} . \tag{3.4}$$

In order to avoid the problem of an undefined variance at $j = 0$ we define $\phi_0 := 0$. It can easily be shown that the coefficient ϕ_0 just shifts the function $f(t)$ vertically, so that this choice does not affect the shape of $f(t)$. The Fourier amplitudes $\phi(\nu)$ defined above satisfy the relations $\overline{\phi(\nu)} = 0$ and

$$\begin{aligned} \overline{\phi(\nu)\phi(\nu')^\star} &= \sum_{j=-n}^{n} \sum_{k=-n}^{n} \overline{\phi_j \phi_k^\star} \, \delta(\nu - \nu_j) \, \delta(\nu' - \nu_k) \\ &= \sum_{j=-n}^{n} c\,|\nu_j|^{-\beta} \, \delta(\nu - \nu_j) \, \delta(\nu' - \nu_j). \end{aligned}$$

Using the definition of Dirac's delta function (p. 45), this result can be written in the form

$$\overline{\phi(\nu)\phi(\nu')^\star} = P(\nu) \, \delta(\nu - \nu')$$

where

$$P(\nu) = c\,|\nu|^{-\beta} \sum_{j=-n}^{n} \delta(\nu - \nu_j).$$

Comparing this result with the definition of FBM (p. 49) shows that we have in fact arrived at an approximation to FBM by a finite number of frequencies.

This result suggest the following algorithm for generating a function $f(t)$ whose properties come close to those of FBM: First, random Fourier coefficients ϕ_j are drawn from Gaussian distributions obeying Eq. 3.4. Then, $f(t)$ is computed from

$$f(t) = \int_{-\infty}^{\infty} \phi(\nu)\, \mathrm{e}^{2\pi \mathrm{i} \nu t}\, d\nu = \sum_{j=-n}^{n} \phi_j\, \mathrm{e}^{2\pi \mathrm{i} \nu_j t}.$$

If only the times $t_k = k\delta t$ are considered, this expression turns into an inverse *discrete Fourier transform*

$$f(t_k) = \sum_{j=-n}^{n} \phi_j\, \mathrm{e}^{2\pi \mathrm{i} \frac{jk}{2n}}. \tag{3.5}$$

Let us call the series of the values $f(t_k)$ *discrete* FBM.

It is convenient to choose n as a power of two; in this case the *fast Fourier transform* (FFT) algorithm can be applied. This algorithm was introduced by Cooley and Tukey (1965) and should be described in any textbook on elementary numerics. If n is large, the FFT algorithm is much faster than the naive algorithm for computing the sum; it allows to compute discrete FBM with some millions of frequencies.

Discrete FBM suffers from some artefacts resulting from the limited spectrum of frequencies. Obviously, $f(t_{k+2n}) = f(t_k)$, so that the generated time series is periodic with a period of $2n$ points. This problem arises from the existence of a lower limit of the frequencies. In order to get around the problem, $2n$ should be chosen much larger than the number of points finally desired. In other words, a large part of the time series should be skipped. In the examples shown in Fig. 3.2, a total of $2n = 2^{16} = 65,536$ values were computed, but only 10,000 values were plotted.

3.6 Scaling Properties of FBM

In Sect. 3.1 we have observed self-affine scaling properties of Brownian motion. The statistical properties of Brownian motion are unaffected by simultaneously rescaling time by a factor λ and the spatial coordinate by the factor $\sqrt{\lambda}$.

Before investigating whether FBM exhibits a self-affine scaling properties, we should develop a formal definition of self-affine scale invariance for random functions. Scaling the time axis by a factor λ corresponds to considering the function $f_\lambda(t) := f(\frac{t}{\lambda})$ instead of the original function $f(t)$. Since $f_\lambda(0) = f(0)$, this procedure does not affect the origin of the axis, so it corresponds to magnifying the graph of the function around $t = 0$ as illustrated in Fig. 3.1 for Brownian motion. Let us go a step ahead and allow scaling with an arbitrary fixed point T, which means that we consider $f_{\lambda,T}(t) := f(\frac{t-T}{\lambda} + T)$. This is entirely equivalent to considering $f_{\lambda,\tau}(t) := f(\frac{t-\tau}{\lambda})$ where $\tau = (1-\lambda)T$. The next step is rescaling the values of the function. If we take a look again at Fig. 3.2, we may guess that the scaling factor $\sqrt{\lambda}$ is a certain property of Brownian motion and does not hold for FBM in general. So let us assume

scaling by an arbitrary factor $s(\lambda)$. We can easily see that s may not be any arbitrary function, but only $s(\lambda) = \lambda^H$ where H is a real number. This result can be obtained from the fact that rescaling time a factor λ_1 first and then by a factor λ_2 must lead to the same result as scaling it once by the factor $\lambda_1\lambda_2$. This leads the condition $s(\lambda_1\lambda_2) = s(\lambda_1)s(\lambda_2)$, and $s(\lambda) = \lambda^H$ is the only function which meets this criterion together with the conditions $s(1) = 1$ and $s(\lambda) > 0$. So let us define:

A random function $f(t)$ is a *self-affine fractal* if all random functions $f_{\lambda,\tau}(t)$ which arise from the transform

$$f_{\lambda,\tau}(t) \;:=\; \lambda^H f(\tfrac{t-\tau}{\lambda})$$

are characterized by the same statistical properties. The exponent H is called *Hausdorff exponent.*

Since we have characterized the statistical properties of FBM with the help of its Fourier amplitudes, we use the latter to examine whether FBM is a self-affine fractal. According to Eq. 3.3, the Fourier amplitudes $\phi_{\lambda,\tau}(\nu)$ of $f_{\lambda,\tau}$ are related to the original Fourier amplitudes $\phi(\nu)$ through the relation

$$\begin{aligned}\phi_{\lambda,\tau}(\nu) &= \int_{-\infty}^{\infty} \lambda^H f(\tfrac{t-\tau}{\lambda})\, \mathrm{e}^{-2\pi \mathrm{i}\nu t}\, dt \;=\; \lambda^H \int_{-\infty}^{\infty} f(t')\, \mathrm{e}^{-2\pi \mathrm{i}\lambda(\nu t'+\tau)}\, \lambda\, dt' \\ &= \lambda^{H+1}\, \mathrm{e}^{-2\pi \mathrm{i}\nu\tau} \int_{-\infty}^{\infty} f(t')\, \mathrm{e}^{-2\pi \mathrm{i}\lambda\nu t'}\, dt' \;=\; \lambda^{H+1}\, \mathrm{e}^{-2\pi \mathrm{i}\nu\tau}\, \phi(\lambda\nu).\end{aligned}$$

Therefore, the Fourier amplitudes $\phi_{\lambda,\tau}(\nu)$ are Gaussian-distributed random variables with

$$\overline{\phi_{\lambda,\tau}(\nu)} \;=\; \lambda^{H+1}\, \mathrm{e}^{-2\pi \mathrm{i}\nu\tau}\, \overline{\phi(\lambda\nu)} \;=\; 0 \;=\; \overline{\phi(\nu)}$$

and

$$\begin{aligned}\overline{\phi_{\lambda,\tau}(\nu)\phi_{\lambda,\tau}(\nu')^\star} &= \lambda^{2H+2}\, \mathrm{e}^{-2\pi \mathrm{i}(\nu-\nu')\tau}\, \overline{\phi(\lambda\nu)\phi(\lambda\nu')} \\ &= \lambda^{2H+2}\, P(\lambda\nu)\, \delta(\lambda\,(\nu-\nu')) \\ &= \lambda^{2H+2}\, \lambda^{-\beta}\, P(\nu)\, \delta(\lambda\,(\nu-\nu'))\end{aligned}$$

because $P(\nu)$ follows a power law. From the definition of Dirac's delta function (p. 45) we immediately obtain

$$\delta(\lambda\,(\nu-\nu')) \;=\; \tfrac{1}{\lambda}\delta(\nu-\nu'),$$

so that

$$\overline{\phi_{\lambda,\tau}(\nu)\phi_{\lambda,\tau}(\nu')^\star} \;=\; \lambda^{2H+1}\, \lambda^{-\beta}\, P(\nu)\, \delta(\nu-\nu') \;=\; \lambda^{2H+1-\beta}\, \overline{\phi(\nu)\phi(\nu')}.$$

Thus, the statistical properties of FBM persist under anisotropic scaling exactly if $2H + 1 = \beta$. According to the definition given above, FBM is a self-affine fractal with

$$H = \tfrac{1}{2}(\beta - 1). \tag{3.6}$$

However, this result should be interpreted with some caution. Let us consider the *range* of $f(t)$ in an interval $[t_1, t_2]$:

$$R(t_1, t_2) \coloneqq \max_{t\in[t_1,t_2]}\{f(t)\} - \min_{t\in[t_1,t_2]}\{f(t)\}. \tag{3.7}$$

The self-affine scaling properties of FBM result in

$$\begin{aligned}\overline{R(t_1,t_2)} &= \overline{\max_{t\in[t_1,t_2]}\{f(t)\} - \min_{t\in[t_1,t_2]}\{f(t)\}} \\ &= \lambda^H \overline{\max_{t\in[t_1,t_2]}\left\{f\left(\tfrac{t-\tau}{\lambda}\right)\right\} - \min_{t\in[t_1,t_2]}\left\{f\left(\tfrac{t-\tau}{\lambda}\right)\right\}}.\end{aligned}$$

With $\tau = t_1$, $\lambda = t_2 - t_1$, and $t' = \frac{t-t_1}{t_2-t_1}$, this expression turns into

$$\overline{R(t_1,t_2)} = (t_2-t_1)^H \overline{\max_{t'\in[0,1]}\{f(t')\} - \min_{t'\in[0,1]}\{f(t')\}} = (t_2-t_1)^H \overline{R(0,1)}. \tag{3.8}$$

Obviously, the range of any function must increase if the interval is extended, but Eq. 3.8 satisfies this condition only if $H \geq 0$. On the other hand, Eq. 3.6 yields negative exponents H for $\beta < 1$. But where is the mistake? The only reasonable explanation is that $\overline{R(t_1,t_2)}$ is not well-defined if $\beta < 1$. A similar problem occurs if $H > 1$: Obviously, each function must satisfy the relation $R(t_1,t_2) \leq R(t_1,t) + R(t,t_2)$ for all points $t \in [t_1,t_2]$, but this is consistent with Eq. 3.8 only if $H \leq 1$. Therefore, $\overline{R(t_1,t_2)}$ cannot be well-defined for FBM with $\beta > 3$. We will come back to this problem when discussing the variogram analysis in Sect. 3.9.

More than 50 years ago, H. E. Hurst (Hurst 1951, 1957; Hurst et al. 1965) introduced the range $R(t_1,t_2)$ as a tool for measuring correlations in time series. His method is called *rescaled range analysis* or *R/S analysis*; it is still widely used. Since the R/S analysis is much older than the idea of self-affine scale invariance, it is not surprising that it does not directly refer to the Hausdorff exponent H. However, the method hinges on the power-law increase of $R(t_1,t_2)$ found above (Eq. 3.8) in principle, although it involves a second property $S(t_1,t_2)$ which describes an estimated standard deviation. By analyzing both properties in intervals of different lengths, Hurst found that many time series are characterized by a power-law increase of the ratio between $R(t_1,t_2)$ and $S(t_1,t_2)$. Today, the resulting exponent is called *Hurst exponent*. However, we will prefer other methods that directly determine the spectral exponent β or the Hausdorff exponent H in the following, although numerical experiments with FBM (Bassingthwaite and Raymond 1994) have shown that the Hurst exponent converges towards the Hausdorff exponent in the limit of infinite data sets under certain conditions. So let us refrain from going further into details of the method; detailed descriptions, including Hurst's original example of the flow of water through a reservoir, are given in the literature mentioned in the introduction of this chapter.

3.7 Self-Affine Scale Invariance and Fractal Dimensions

In Chap. 1, scale-invariant properties were quantified by the fractal dimension D. Can we assign a fractal dimension to self-affine random functions, too? Let us, for simplicity, focus on the definition of scale invariance according to the box-counting method (Sect. 1.2). However, this definition immediately raises the question for the aspect ratio of the boxes. Box counting is defined for quadratic boxes of length r, but here the axes of the coordinate system are physically different. Since we cannot define quadratic boxes here, the definition must be enlarged to boxes with arbitrary aspect ratios. Let r be the box size in direction of t, and μr be the box size in direction of $f(t)$. Let us cover the graph of the function $f(t)$ in an interval $[T_1, T_2]$ with a regular lattice of boxes of size $r \times \mu r$. We assume that the box sizes are chosen in such a way that $\frac{T_2-T_1}{r}$ is a integer number. The number of filled boxes per unit time interval, $N(r)$, can be computed according to

$$N(r) = \frac{1}{T_2-T_1} \sum_{k=1}^{\frac{T_2-T_1}{r}} \left(\left[\frac{\max\limits_{t\in[(k-1)r,kr]} \{f(t)\}}{\mu r} + 1 \right] - \left[\frac{\min\limits_{t\in[(k-1)r,kr]} \{f(t)\}}{\mu r} \right] \right)$$

where the brackets [] denote the integer part of the argument. The brackets can be omitted in the mean over all functions $f(t)$, so that

$$\begin{aligned} \overline{N(r)} &= \frac{1}{T_2-T_1} \sum_{k=1}^{\frac{T_2-T_1}{r}} \left(\frac{\overline{\max\limits_{t\in[(k-1)r,kr]} \{f(t)\} - \min\limits_{t\in[(k-1)r,kr]} \{f(t)\}}}{\mu r} + 1 \right) \\ &= \frac{1}{T_2-T_1} \sum_{k=1}^{\frac{T_2-T_1}{r}} \left(\frac{\overline{R((k-1)r, kr)}}{\mu r} + 1 \right) \\ &= \frac{1}{T_2-T_1} \sum_{k=1}^{\frac{T_2-T_1}{r}} \left(\frac{r^H \overline{R(0,1)}}{\mu r} + 1 \right) = \frac{\overline{R(0,1)}}{\mu} r^{H-2} + \frac{1}{r} \end{aligned}$$

according to Eqs. 3.7 and 3.8. Fig. 3.3 illustrates this result for Brownian motion ($H = \frac{1}{2}$) and boxes of different aspect ratios μ. Obviously, the behavior is not consistent with the definition of scale invariance that hinges on the relation $N(r) \sim r^{-D}$, but we obtain

$$N(r) \sim r^{-D_l} \quad \text{for} \quad r \to 0 \quad \text{and} \quad N(r) \sim r^{-D_g} \quad \text{for} \quad r \to \infty.$$

$D_l = 2 - H$ is denoted *local* fractal dimension, while $D_g = 1$ is the *global* fractal dimension. Local and global dimension only coincide in case of isotropic (self-similar) scale invariance ($H = 1$). The region of crossover in the plot of $N(r)$ versus r is characterized by a non-fractal relationship between the box size and the number of boxes; it can be arbitrarily shifted towards smaller or larger box sizes by adjusting the aspect ratio of the boxes.

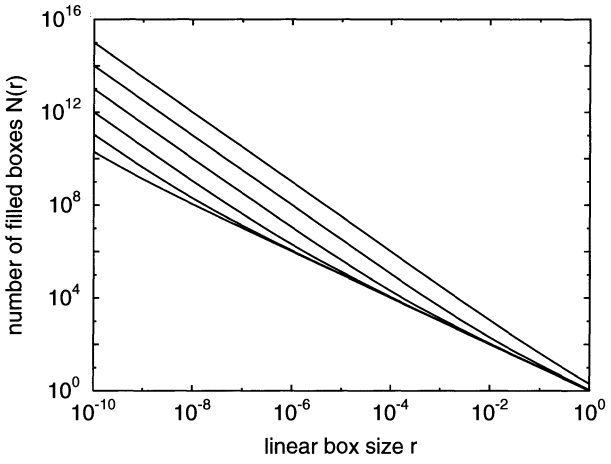

Fig. 3.3. Result of box counting for Brownian motion. The different curves correspond to aspect ratios of the boxes from $\mu = 10^5 \overline{R(0,1)}$ (left) to $\mu = \overline{R(0,1)}$ (right) in steps of a factor 10.

At this point we may wonder why the formalism introduces two fractal dimensions while only the local dimension D_1 is non-trivial. The reason is that the result $D_g = 1$ is only valid for graphs of functions, i. e., if exactly one value $f(t)$ is assigned to each time t. In contrast, self-affine scale invariance is a more general concept. Imagine a *braided river* in planar view. In contrast to meandering rivers, flow in braided rivers is not confined to a single channel, but is distributed among several channels. These channels branch and join, so the drainage pattern consists of many river segments, forks, and junctions. Since the channels roughly follow the main direction of the valley, the pattern is strongly anisotropic. If there is any scale invariance in braided rivers, it should be self-affine rather than self-similar. In this case, scale invariance can be quantified by two non-trivial fractal dimensions D_1 and D_g where $D_g < D_1$. In fact, studies on the scaling properties of braided rivers indicate self-affine scale invariance (Sapozhnikov and Foufoula-Georgiou 1996b), although the available data sets are limited with respect to the range of scales.

3.8 Recognizing FBM

In Sect. 3.5 we have learned how to generate discrete FBM on a computer. However, the opposite task will be waiting for us as soon as we turn towards applications. No matter whether we analyze data obtained from nature or from a numerical model, we start from a time series $f(t)$ given at discrete points t_k. Then, the idea of FBM provides a tool for characterizing the time series. So the question is whether $f(t)$ is consistent with FBM, and if so, what the spectral exponent β is.

Let us assume that the signal $f(t)$ is recorded at a constant *sampling rate*, i. e., that the values $f(t_k)$ are given at the times $t_k = k\delta t$. We then have to invert Eq. 3.5 in order to determine the Fourier coefficients ϕ_j. In analogy to

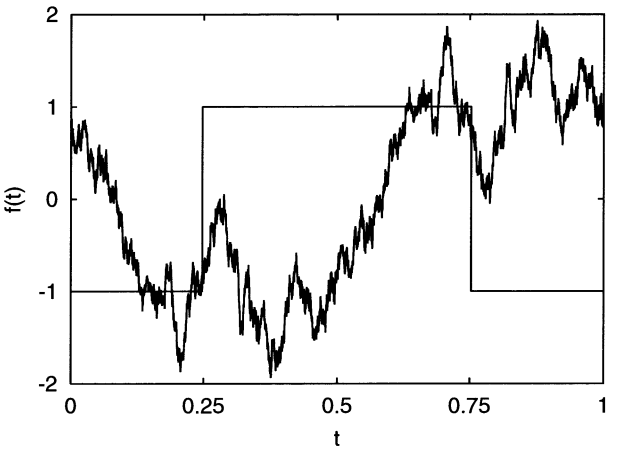

Fig. 3.4. Two functions with the same power spectrum.

the continuous Fourier transform, the forward discrete Fourier transform

$$\phi_j = \frac{1}{2n} \sum_{k=0}^{2n-1} f(t_k)\, \mathrm{e}^{-2\pi \mathrm{i} \frac{jk}{2n}}$$

takes this part. Thus, recognizing FBM appears to be straightforward. The Fourier coefficients ϕ_j are computed from the time series, and it is examined whether they are consistent with the assumptions introduced when discussing discrete FBM: They should be Gaussian-distributed and satisfy Eq. 3.4.

The analysis is often restricted to a plot of $|\phi_j|^2$ versus ν_j. Since the expected value of $|\phi_j|^2$ is proportional to the power spectrum $P(\nu)$, the plot should follow a power law with an exponent $-\beta$, except for statistical fluctuations. However, this analysis may be misleading for several reasons. First, the property $|\phi_j|^2$ does not characterize $f(t)$ uniquely because it only determines the absolute values of the complex Fourier amplitudes and does not refer to their further statistical properties such as correlations. Fig. 3.4 shows two obviously different functions with the same power spectrum, i. e., with the same values $|\phi_j|^2$ for all frequencies. The first function is a deterministic function that switches between -1 and 1. The second function was obtained by rotating its Fourier coefficients randomly in the complex plane; the absolute values $|\phi_j|$ are not affected by a rotation.

Thus, the analysis may be strongly biased by deterministic components in the original signal. The assumed periodicity of the function $f(t)$ may result in a similar problem. As mentioned in the previous section, the function $f(t)$ composed of a finite number of frequencies is periodic with a period $2n\delta t$. When generating FBM on a computer, the problem could easily be fixed by considering only a part of the computed data. In contrast, we would have to prolongate the observed data in such a way that it becomes periodic here, but this is impossible without any knowledge on its statistical properties. If we disregard this phenomenon, we analyze a function with a step between t_{2n-1}

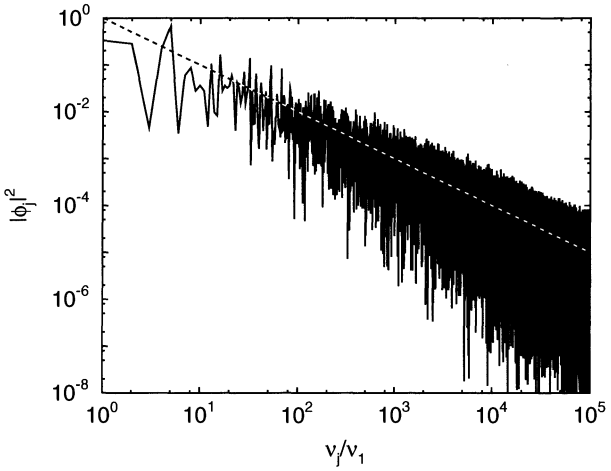

Fig. 3.5. Power spectrum of an example of discrete pink noise. The dashed line shows the expected value.

and t_{2n} or, in other words, the superposition of the original signal with a sawtooth function. In the worst case, we mainly analyze the power spectrum of the sawtooth function. This problem can only be avoided by preparing the signal carefully by means of filtering algorithms.

The third problem arises from the asymmetric distribution of the values $|\phi_j|^2$. If both the real and the imaginary parts of the Fourier coefficients ϕ_j are Gaussian-distributed random variables with zero mean value, the squared values $|\phi_j|^2$ follow a strongly asymmetric distribution. Fig. 3.5 illustrates this phenomenon; it shows the power spectrum of an example of discrete pink noise. Random Fourier amplitudes ϕ_j were drawn from Gaussian distributions according to Eq. 3.4 with $\beta = 1$ and $c = \nu_1$. The values of $|\phi_j|^2$ are plotted versus the frequency; the dashed line shows the expected value. Obviously, the data obtained from the individual function are asymmetrically distributed around the expected power law in the bilogarithmic plot. The data even indicate a spurious curvature in the spectrum. If we fit a power law to the data by visual correlation, we will be misled towards higher exponents β. The same problem will occur if we apply a least-square fit (Sect. 2.1) to the logarithmized data. Thus, established methods for fitting power laws are likely to fail when applied to determining the spectral exponent β. In principle, the problem can be avoided by applying the general maximum-likelihood method, but this is rarely done because it is somewhat complicated and not part of established software tools.

In summary, analyses based on analyzing power spectra by means of the squared Fourier amplitudes are often biased by artifical effects and therefore not very reliable. In the next section we will consider the variogram analysis as an alternative method that suffers less from artefacts than the direct analysis of the power spectrum.

3.9 The Variogram Analysis

When discussing white noise and Brownian motion, we have already seen that these processes differ with respect to temporal correlations. In white noise, $f(t)$ and $f(t')$ are entirely uncorrelated if $t \neq t'$, while there is some correlation in Brownian motion. Temporal correlation can be quantified by the *variogram*

$$\Gamma(t,\tau) \coloneqq \overline{(f(t+\tau) - f(t))^2} \tag{3.9}$$

which describes the spreading of the probability density through time. The variogram is a generalization of the time-dependent variance $\sigma(t)^2 = \overline{f(t)^2}$ towards functions which do not satisfy the condition $f(0) = 0$ assumed when introducing Brownian motion in Sect. 3.1. The variogram was introduced by Matheron (1963). In geostatistics, the *semivariogram* $\gamma(t,\tau) \coloneqq \frac{1}{2}\Gamma(t,\tau)$ is often preferred to the variogram, but the difference between both is just a factor two.

The variogram of FBM can be directly derived from the self-affine scaling behavior found in Sect. 3.6:

$$\Gamma(t,\tau) = \overline{(f(t+\tau) - f(t))^2} = \tau^{2H}\,\overline{(f(1) - f(0))^2} = \tau^{2H}\Gamma(0,1).$$

However, when computing the range $\overline{R(t_1,t_2)}$ of FBM we have already learned that this result may be misleading because $\Gamma(t,\tau)$ is well-defined only for certain values of the Hausdorff exponent H. We found that the range is not well-defined at least for $H < 0$ and for $H > 1$. So let us investigate the variogram of FBM more thoroughly. From Eq. 3.2 we obtain

$$\begin{aligned}
\Gamma(t,\tau) &= \overline{|f(t+\tau) - f(t)|^2} = \overline{\left|\int_{-\infty}^{\infty} \phi(\nu)\left(\mathrm{e}^{2\pi\mathrm{i}\nu(t+\tau)} - \mathrm{e}^{2\pi\mathrm{i}\nu t}\right) d\nu\right|^2} \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \overline{\phi(\nu)\phi(\nu')^\star}\left(\mathrm{e}^{2\pi\mathrm{i}\nu(t+\tau)} - \mathrm{e}^{2\pi\mathrm{i}\nu t}\right)\left(\mathrm{e}^{2\pi\mathrm{i}\nu'(t+\tau)} - \mathrm{e}^{2\pi\mathrm{i}\nu' t}\right)^\star d\nu' d\nu \\
&= \int_{-\infty}^{\infty} P(\nu)\left|\mathrm{e}^{2\pi\mathrm{i}\nu(t+\tau)} - \mathrm{e}^{2\pi\mathrm{i}\nu t}\right|^2 d\nu \\
&= 2\int_{-\infty}^{\infty} P(\nu)\,(1 - \cos(2\pi\nu\tau))\,d\nu.
\end{aligned}$$

Obviously, $\Gamma(t,\tau)$ depends only on the time lag τ, but not on the absolute time t. So we can omit the argument t in the following.

If the power spectrum $P(\nu)$ follows a power law for all frequencies ν, the integral does not converge for all spectral exponents β. Let us assume that $P(\nu)$ follows a power law within a limited range of frequencies:

$$P(\nu) = \begin{cases} c\,|\nu|^{-\beta} & \text{if} \quad \nu_1 < |\nu| < \nu_2 \\ 0 & \text{else} \end{cases}.$$

The restriction of the power spectrum results in

$$\Gamma(\tau) \;=\; 4c\int_{\nu_1}^{\nu_2}\nu^{-\beta}\,(1-\cos(2\pi\nu\tau))\,d\nu \;=\; 4c\,\tau^{\beta-1}\int_{\nu_1\tau}^{\nu_2\tau}u^{-\beta}\,(1-\cos(2\pi u))\,du. \tag{3.10}$$

The integral cannot be solved analytically, but its properties in the double limit $\nu_1 \to 0$ and $\nu_2 \to \infty$ can be determined. From lengthy, but not very difficult calculations we obtain

$$\Gamma(\tau) \;\sim\; \begin{cases} \nu_2^{1-\beta}\left(1-\left(\frac{\nu_1}{\nu_2}\right)^{1-\beta}A_\beta(\nu_1\tau)^2-\frac{B_\beta}{(\nu_2\tau)^{1-\beta}}\right) & 0<\beta<1 \\ 1-A_\beta\,(\nu_1\tau)^2-B_\beta\log\left(\frac{1}{\nu_2\tau}\right) & \beta=1 \\ \tau^{\beta-1}\left(1-A_\beta\,(\nu_1\tau)^{3-\beta}-\frac{B_\beta}{(\nu_2\tau)^{\beta-1}}\right) & \text{for}\quad 1<\beta<3 \\ \tau^2\left(1-A_\beta\,\log(\nu_1\tau)-\frac{B_\beta}{(\nu_2\tau)^2}\right) & \beta=3 \\ \frac{\tau^2}{\nu_1^{\beta-3}}\left(1-A_\beta\,(\nu_1\tau)^{\beta-3}-\left(\frac{\nu_1}{\nu_2}\right)^{\beta-3}\frac{B_\beta}{(\nu_2\tau)^2}\right) & \beta>3 \end{cases} \tag{3.11}$$

if τ satisfies the relation $\frac{1}{\nu_2} \ll \tau \ll \frac{1}{\nu_1}$. A_β and B_β are positive constants.

Fig. 3.6 illustrates this result, although it does not exactly refer to FBM with a limited power spectrum, but to discrete FBM introduced in Sect. 3.5. However, both processes are similar, except for the periodicity of discrete FBM which results in a decrease of the variogram at large time lags. The total number of points is $2n = 2^{20} = 1,048,576$; the factor c in the power spectrum (Eq. 3.4) has been adjusted in such a way that $\overline{f(t)^2} = 1$. As expected, the variogram follows a power law with an exponent $\beta-1$ for $1 < \beta < 3$, but deviates from this behavior at large and small time lags. The deviation at small time lags is most significant if β comes close to unity. In contrast, the deviation at large time lags becomes strong if β approaches three, but this effect is partly hidden by the artifical periodicity of discrete FBM. For $0 < \beta < 1$, the variogram becomes constant at large time lags, but deviates from this behavior on short time scales. Again, this effect is most significant for β close to unity. The boundary between both regimes is defined by pink noise ($\beta = 1$) where the variogram increases logarithmically.

Let us now come back to FBM in the original sense. In Eq. 3.11, the double limit $\nu_1 \to 0$ and $\nu_2 \to \infty$ can only be performed if $1 < \beta < 3$. This result confirms our conjecture that the variogram of FBM is only well-defined for spectral exponents β in a certain range. For $\beta \leq 1$, only the limit $\nu_1 \to 0$ can be performed. As a result of the slow decay of $P(\nu)$ at large frequencies, the variogram diverges in the limit $\nu_2 \to \infty$. Finally, the limit $\nu_2 \to \infty$ can be performed for $\beta \geq 3$, while the $\Gamma(\tau)$ diverges for $\nu_1 \to 0$.

This results suggest a modified definition of FBM where only those limits leading to well-defined variograms are performed. In order to distinguish this modified FBM from FBM in the original sense, we use the terms *physical* and *idealized* FBM in the following.

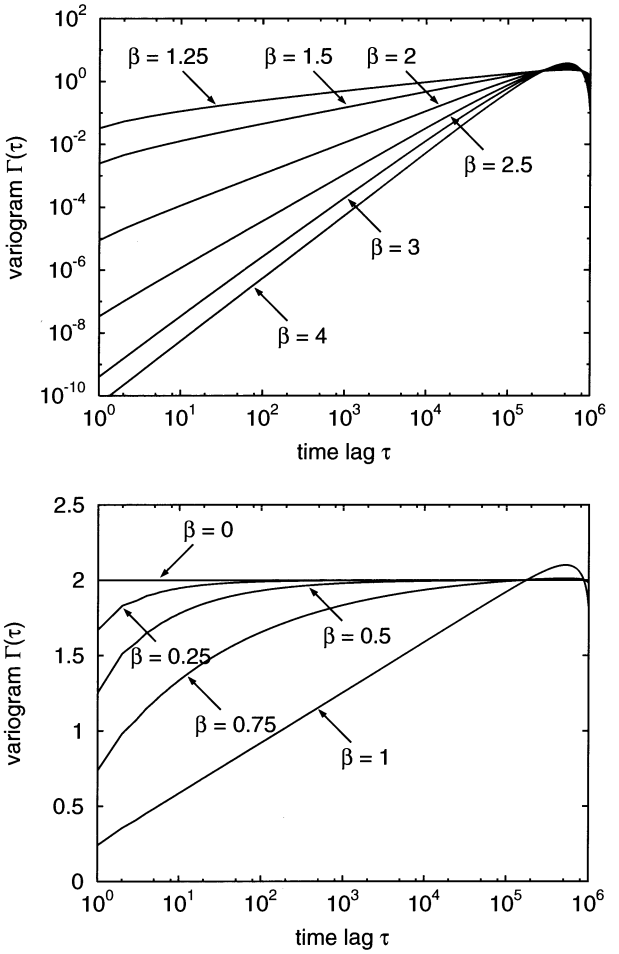

Fig. 3.6. Variograms of discrete FBM for different spectral exponents β.

Physical FBM is a statistical process where both the real and the imaginary parts of the Fourier amplitudes $\phi(\nu)$ are Gaussian-distributed random variables with

$$\overline{\phi(\nu)} = 0 \quad \text{and} \quad \overline{\phi(\nu)\phi(\nu')^\star} = P(\nu)\,\delta(\nu-\nu').$$

with

$$P(\nu) \sim \begin{cases} |\nu|^{-\beta} & \text{if} \quad \nu_1 < |\nu| < \nu_2 \\ 0 & \text{else} \end{cases}$$

and $\nu_1 := 0$ if $\beta < 3$, respectively, $\nu_2 := \infty$ if $\beta > 1$.

Physical and idealized FBM coincide for $1 < \beta < 3$. Equation 3.11 can directly be transferred to the variogram of physical FBM:

$$\Gamma(\tau) \sim \begin{cases} 1 - B_\beta\,(\nu_2\tau)^{-(1-\beta)} & 0<\beta<1 \\ 1 + B_\beta \log(\nu_2\tau) & \beta = 1 \\ \tau^{\beta-1} \quad \text{for} & 1<\beta<3 \\ \tau^2\,(1 - A_\beta \log(\nu_1\tau)) & \beta = 3 \\ \tau^2\left(1 - A_\beta\,(\nu_1\tau)^{\beta-3}\right) & \beta > 3 \end{cases} \quad \begin{array}{l} \text{if } \tau \gg \frac{1}{\nu_2} \\ \\ \\ \\ \text{if } \tau \ll \frac{1}{\nu_1} \end{array} . \tag{3.12}$$

In principle, this result can be used for determining the spectral exponent of a time series obtained from nature or from a numerical model, provided that the time series is an example of FBM. However, the definition of the variogram involves expected values in the statistical sense; so it cannot be computed directly from a given time series. Instead, the expected values must be estimated from the values $f(t)$. The best estimate is obtained from averaging over the available range in time. If the time series is observed within an interval $[T_1, T_2]$, this leads to:

$$\Gamma(\tau) \approx \frac{1}{T_2 - T_1 - \tau} \int_{T_1}^{T_2-\tau} (f(t+\tau) - f(t))^2\, dt. \tag{3.13}$$

If the available data are restricted to discrete times $t_1, \ldots, t_n$, the integral must be approximated by a discrete sum. For a constant sampling rate, i. e., if $\delta t = t_{k+1} - t_k$ is constant, Eq. 3.13 can be approximated for time lags $\tau = k\delta t$:

$$\Gamma(k\delta t) \approx \frac{1}{n-k} \sum_{i=1}^{n-k} (f(t_{i+k}) - f(t_i))^2. \tag{3.14}$$

The method seems to be clear now: After the variogram has been estimated by either Eq. 3.13 or Eq. 3.14, the result must be compared to Eq. 3.12. If $\Gamma(\tau) \sim \tau^\alpha$ where $0 < \alpha < 2$, Eq. 3.12 yields $\beta = 1+\alpha$. Deviations from power-law behavior may occur at small and at large time lags τ, but they should not be too severe if both the resolution and the length of the time series are sufficiently large. Thus, the method provides a reliable estimate of spectral exponents in the range $1 < \beta < 3$.

In contrast, the method is less promising if β is outside this range. The variogram is roughly constant if $\beta < 1$, while it increases like τ^2 if $\beta > 3$. The exact value of β does not affect this behavior strongly; it just controls the deviations occurring at small, respectively, large time lags. Theoretically, these deviations can be used for estimating β, but we should be aware of the following limitations: First, Eq. 3.12 is only valid for not too small time lags if $\beta < 1$, respectively, for not too large time lags if $\beta > 3$. In other words, it only applies to the region where the deviations are small. Therefore, the result will probably be affected by statistical variations. Second, Eq. 3.12 has been derived for physical FBM, i. e., under the assumption of a sharp cutoff in the power spectrum. If processes with different cutoff behavior are considered, the overall behavior of the variogram may be the same as for physical FBM, but the deviations may be quantitatively different.

In summary, estimating spectral exponents from a variogram analysis is only advisable for FBM with spectral exponents between one and three. The entire range between white noise ($\beta = 0$) and pink noise ($\beta = 1$) cannot be distinguished clearly by this method. This problem can be fixed if we remember the result that integrating the function increases the spectral exponent by two, while deriving the function with respect has the opposite effect. So, if the variogram increases like τ^2, we should not try to quantify the deviations from this power law, but apply the variogram analysis to the function $\frac{\partial}{\partial t} f(t)$. In extreme cases where $\beta > 5$, it is necessary to derive the signal more than once. However, FBM with $\beta > 3$ is very smooth and thus not very interesting; so we will hardly be requested to estimate spectral exponents in this range. The case $0 < \beta < 1$ is more interesting. Here we should analyze the variogram of the *cumulated* signal $f_c(t) := \int_{-\infty}^{t} f(t')\,dt'$:

$$\Gamma_c(\tau) = \overline{(f_c(t+\tau) - f_c(t))^2} = \overline{\left(\int_t^{t+\tau} f(t')\,dt'\right)^2}. \tag{3.15}$$

In analogy to Eqs. 3.13 and 3.14, $\Gamma_c(\tau)$ can be estimated according to

$$\Gamma_c(\tau) = \frac{1}{T_2 - T_1 - \tau} \int_{T_1}^{T_2-\tau} \left(\int_t^{t+\tau} f(t')\,dt'\right)^2 dt, \tag{3.16}$$

respectively,

$$\Gamma_c(k\delta t) = \frac{1}{n-k+1} \sum_{i=1}^{n-k+1} \left(\sum_{j=i}^{i+k-1} f(t_j)\right)^2. \tag{3.17}$$

With these extensions, the variogram analysis provides a reliable method for estimating spectral exponents of FBM. However, many signals occurring in nature or obtained from models are not FBM. Often, $f(t)$ is not Gaussian-distributed or the expected value $\overline{f(t)}$ is not zero. Precipitation rates are a good example where the distribution is strongly skewed since only positive values occur. In principle, the assumption of Gaussian-distributed values is not crucial; when deriving the relationship between variogram and power spectrum (Eq. 3.12), we did not use this property explicitly. Thus, it makes sense to generalize FBM towards non-Gaussian random process.

In contrast, we should be careful if $\overline{f(t)} \neq 0$. Let us compare the processes $f(t)$ and $f'(t) := f(t) - \overline{f(t)}$. Both are entirely equivalent, except for the fact that $f(t)$ has a non-vanishing mean value, while $\overline{f'(t)} = 0$. From Eq. 3.9 we immediately obtain $\Gamma(\tau) = \Gamma'(\tau)$; so the variogram analysis is not affected by a non-vanishing mean value. However, this result does not hold for the variogram of the cumulated process (Eq. 3.15) because

$$\Gamma_c(\tau) = \overline{\left(\int_t^{t+\tau} \left(f'(t) + \overline{f(t)}\right) dt'\right)^2} = \Gamma'_c(\tau) + \overline{f(t)}^2 \tau^2.$$

So the variogram of the cumulated signal is biased by a term which grows quadratically with τ in case of a non-vanishing mean value. Therefore, the mean value of the time series must be removed as described above before analyzing the variogram of the cumulated signal.

3.10 Predictability

Will the next summer be hot and dry or rather rainy? We should know that nobody is able to give a clear answer to this question. Many phenomena in earth sciences elude an exact prediction, and among them there are several phenomena such as earthquakes where even a short-term prediction might save lives. In the previous decades, the question for *predictability* has become one of the most challenging topics not only in natural sciences, but in economy and social sciences, too.

When discussing Brownian motion (Sect. 3.1) and white noise (Sect. 3.2), we have already experienced essentially different behavior concerning predictability. Since the random variables $f(t)$ are completely uncorrelated in white noise, the latter is entirely unpredictable. In contrast, assuming that $f(t+\tau) \approx f(t)$ should not be too bad in Brownian motion at least if a short-term prediction is the goal. So, if climate is Brownian motion, it makes sense to predict that the next summer will be similar to the recent summer, but if it is white noise, this prediction is built on sand. We may even go a step ahead and include the previous summer into the prediction. If the recent summer is better than the previous, will the next one still be better or will the trend be reverted? In Brownian motion, neither of those assumptions holds in the mean because the increments are independent here, but the result may be different for FBM in general.

Let us not go into climate dynamics here, but discuss general strategies of prediction and assess their value with respect to FBM. We assume that the recent value $f(t)$ and perhaps an older value $f(t-\tau)$ are known, and that we aim at predicting $f(t+\tau)$. Obviously, the quality of the prediction must be assessed statistically; the mean quadratic error $\sigma_p(\tau)$ defined by

$$\sigma_p(\tau)^2 \;\coloneqq\; \overline{(f_p(t+\tau) - f(t+\tau))^2}$$

where $f_p(t+\tau)$ is the predicted value provides a straightforward measure. Let us construct the prediction linearly from the recent value $f(t)$ and the trend between previous and recent value, $f(t) - f(t-\tau)$:

$$f_p(t+\tau) \;\coloneqq\; \lambda f(t) + \mu\,(f(t) - f(t-\tau))$$

where λ and μ are constants. Computing $\sigma_p(\tau)^2$ is a little lengthy, but simple; we finally obtain

$$\sigma_p(\tau)^2 = (1-\lambda)\left(\mu\,\overline{f(t-\tau)^2} - (\lambda+\mu)\,\overline{f(t)^2} + \overline{f(t+\tau)^2}\right) + (\mu+1)\,(\lambda+\mu)\,\Gamma(\tau) - \mu\,\Gamma(2\tau)$$

with the variogram $\Gamma(\tau)$ defined in Eq. 3.9. If $P(\nu) = c|\nu|^{-\beta}$ for $\nu_1 < |\nu| < \nu_2$, we obtain

$$\begin{aligned}\overline{f(t)^2} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \overline{\phi(\nu)\phi(\nu')^\star}\, \mathrm{e}^{2\pi \mathrm{i}\nu t}\, \mathrm{e}^{-2\pi \mathrm{i}\nu' t}\, d\nu' d\nu \\ &= \int_{-\infty}^{\infty} P(\nu)\, d\nu = \frac{2c}{1-\beta}\left(\nu_2^{1-\beta} - \nu_1^{1-\beta}\right). \end{aligned} \tag{3.18}$$

Thus, $\overline{f(t-\tau)^2}$, $\overline{f(t)^2}$, and $\overline{f(t+\tau)^2}$ coincide, which leads to

$$\sigma_p(\tau)^2 = (1-\lambda)^2\,\overline{f(t)^2} + (\mu+1)\,(\lambda+\mu)\,\Gamma(\tau) - \mu\,\Gamma(2\tau). \tag{3.19}$$

Now the parameters λ and μ must be adjusted to minimize $\sigma_p(\tau)^2$ since this leads to the best prediction. According to Eq. 3.11, $\Gamma(\tau)$ remains finite in the double limit $\nu_1 \to 0$ and $\nu_2 \to \infty$ in the case $1 < \beta < 3$. In contrast, $\overline{f(t)^2}$ diverges for $\nu_1 \to 0$. Thus, $\sigma_p(\tau)^2$ diverges if $\lambda \neq 1$, so that any reasonable prediction must be based on $\lambda = 1$. The same applies for $\beta > 3$ because $\overline{f(t)^2}$ diverges like $\nu_1^{-(\beta-1)}$, while $\Gamma(\tau)$ only diverges like $\nu_1^{-(\beta-3)}$. With $\lambda = 1$, Eq. 3.19 reduces to

$$\sigma_p(\tau)^2 = (\mu+1)^2\,\Gamma(\tau) - \mu\,\Gamma(2\tau).$$

The mean error $\sigma_p(\tau)$ becomes minimal if

$$\mu = \frac{\Gamma(2\tau)}{2\,\Gamma(\tau)} - 1 = \begin{cases} 2^{\beta-2} - 1 & \text{for } 1 < \beta < 3 \\ 1 & \text{for } \beta > 3 \end{cases}$$

according to Eq. 3.12. Before discussing this result in detail, we consider the case $0 < \beta < 1$. Here, both $\overline{f(t)^2}$ and $\Gamma(\tau)$ diverge like $\nu_2^{1-\beta}$ in the limit $\nu_2 \to \infty$. We therefore need a quantitative relationship between both instead of the asymptotic behavior. From Eqs. 3.10 and 3.18 we obtain

$$\Gamma(\tau) = 2\,\overline{f(t)^2}\left(1 - (1-\beta)\,(\nu_2\tau)^{-(1-\beta)}\int_0^{\nu_2\tau} u^{-\beta}\cos(2\pi u)\, du\right)$$

for $\nu_1 = 0$ as assumed in the definition of physical FBM (p. 61). Inserting this result into Eq. 3.19 leads to

$$\begin{aligned}\frac{\sigma_p(\tau)^2}{\overline{f(t)^2}} &= 1 + \mu^2 + (\lambda+\mu)^2 \\ &\quad -2\,(\mu+1)\,(\lambda+\mu)\,(1-\beta)\,(\nu_2\tau)^{-(1-\beta)}\int_0^{\nu_2\tau} u^{-\beta}\cos(2\pi u)\, du \\ &\quad +2\,\mu\,(1-\beta)\,(2\nu_2\tau)^{-(1-\beta)}\int_0^{2\nu_2\tau} u^{-\beta}\cos(2\pi u)\, du.\end{aligned}$$

Since the integrals remain finite in the limit $\nu_2\tau \to \infty$, we obtain

$$\lim_{\nu_2\tau\to\infty} \frac{\sigma_p(\tau)^2}{\overline{f(t)^2}} = 1 + \mu^2 + (\lambda+\mu)^2.$$

This expression is minimized by $\lambda = \mu = 0$. Let us summarize the results on the best prediction of $f(t+\tau)$:

$$f_p(t+\tau) = \begin{cases} 0 & 0 < \beta < 1 \\ f(t) + \left(2^{\beta-2} - 1\right)\left(f(t) - f(t-\tau)\right) & \text{for} \quad 1 < \beta < 3 \\ f(t) + \left(f(t) - f(t-\tau)\right) & \beta > 3 \end{cases}.$$

The result obtained for $\beta > 3$ is nothing but linear extrapolation from the values $f(t-\tau)$ and $f(t)$ towards $f(t+\tau)$. This kind of prediction is straightforward for smooth functions. Since FBM is quite smooth for $\beta > 3$, this result is not surprising.

In the range $1 < \beta < 3$, the actual value $f(t)$ is still a good basis for predicting $f(t+\tau)$. However, the trend from the past must be regarded in a different way. If $\beta > 2$, the best prediction $f_p(t+\tau)$ is still larger than $f(t)$ if the signal increased in the past. This means that the trend of the signal tends to persist. In other words, the signal is likely to increase in the future if it has increased so far, and vice versa. This property is called *persistence*. If $\beta < 2$, the behavior is opposite. If the signal has increased, it is likely to decrease in the future, and vice versa. This property is called *anti-persistence*. As expected, Brownian motion ($\beta = 2$) is just at the edge between persistence and anti-persistence because the increments are completely uncorrelated here.

The result is completely different in the range $0 < \beta < 1$. The best prediction is simply zero, i. e., the expected value of the signal. In other words, all information on the recent or past status of the system is useless for predicting into the future; the process is essentially *unpredictable*. Transferred to the question for the next summer, this result means that the next summer will be like all summers are in the mean, independently from the recent or the previous summer. However, we must not forget that this result was obtained from physical FBM (p. 49) in the limit $\nu_2\tau \to \infty$. So, if the upper cutoff frequency ν_2 is finite, the process is unpredictable only over time spans τ which are much longer that the period $\frac{1}{\nu_2}$ of the upper cutoff frequency. In physical FBM, there is always some degree of predictability over very short time spans. This result can be immediately recognized when looking at the examples of FBM shown in Fig. 3.2.

Interestingly, pink noise ($\beta = 1$) is just at the point where the behavior concerning predictability changes drastically. Obviously, FBM becomes more regular if the spectral exponent β increases, but the regularity is too weak to allow a prediction if $\beta < 1$. In the following chapter, the term *chaos* will be introduced for entirely unpredictable behavior. In this context, pink noise defines the *edge of chaos*.

4. Deterministic Chaos

In the previous chapters, we mainly dealt with fractal patterns, distributions, and time series. The approach was a rather descriptive one; the ideas on the origin of fractals were rather mathematical algorithms than physical processes. In the following we will focus on non-linear processes in order to see whether they contribute anything to our understanding of fractals.

Our way of analyzing phenomena is often based on linear approaches. When applying a disturbance to a system, we expect a more or less significant effect. Understanding phenomena mostly consists of attributing effects to disturbances. Linear systems are characterized by a linear relationship between disturbance and effect; if we apply a disturbance which is twice as strong, the effect will be twice as strong, too. Moreover, the effects of different disturbances can be separated in linear systems; if we apply two different disturbances, the effect of both is nothing but the sum of the effects resulting from applying both disturbances separately.

A few phenomena are linear, while the majority is not. The sources of non-linearity are manifold. Some systems are roughly linear up to a threshold where the behavior changes drastically. Failure of materials is one of the most prominent examples of threshold behavior, but threshold behavior plays an important part, e. g., in social sciences, too. Many of our decisions are not continuous, but simply "yes" or "no". Interaction of different system components is another source of non-linearity. In many cases, the effect of interactions is less drastic than threshold behavior, but in the following we will meet examples where the system's behavior is governed by interactions, while the behavior of the individual components becomes quite unimportant. The Lorenz equations which will accompany us on our way through this chapter will serve as a first example, but the effect becomes even more dominant when we turn to self-organized critical systems in the following chapters.

Although non-linearity seems to be the key to nearly any kind of interesting system behavior, non-linear dynamics is a quite new field. During the last centuries, much theory was developed for solving linear differential equations, but most of these methods cannot be applied to highly non-linear systems. Basic properties of non-linear systems were discovered using computer models. Consequently, progress in non-linear dynamics has been closely related to the rapidly increasing computer power since the 1960's.

4.1 The Lorenz Equations

In the following we consider the Lorenz equations (Lorenz 1963), the perhaps most famous example of non-linear dynamics. The Lorenz equations are a system of three non-linear, ordinary differential equations:

$$\begin{aligned} \frac{\partial}{\partial t} A(t) &= Pr\,(-A(t) + B(t))\,, \\ \frac{\partial}{\partial t} B(t) &= -B(t) + rA(t) - A(t)\,C(t), \\ \frac{\partial}{\partial t} C(t) &= -\tfrac{8}{3}\,C(t) + A(t)\,B(t). \end{aligned} \tag{4.1}$$

They describe the evolution of three functions $A(t)$, $B(t)$, and $C(t)$ through time. The Prandtl number Pr and the relative Rayleigh number r are the model parameters. Non-linearity arises from the product terms in the second and third equation. Despite their simplicity, the Lorenz equations offer nearly all types of system behavior – steady states, periodic behavior, and total irregularity, called *chaos*. This variety even fills books (e. g. Sparrow 1982) and makes the Lorenz equations still fascinating after nearly 40 years.

The Lorenz equations may be seen as a rather abstract system of equations, but in fact they have some relevance to physics and earth sciences as they describe a simplified model of thermal *convection*. Thermal convection arises from the fact that the mass density of most materials decreases if the temperature increases. If a fluid is heated from below, an unstable situation occurs: Heavy fluid is above, and light fluid is below. The light fluid tends to rise, while the heavy fluid tends to sink. Under certain conditions, this instability results in the formation of convection cells; they consist of regions where a hot, upward flow occurs and regions with a cool, downward flow. In this way, a net upward heat transport takes place; in general it is more efficient than the transport by heat conduction. There is strong evidence for the occurrence of thermal convection in the earth's interior; anomalies in the surface heat flow, seafloor spreading, seismic activity at continental margins, and plate motion can only be explained in a consistent model if there is convection below the earth's crust.

Figure 4.1 illustrates the flow and temperature distribution in a rectangular cell in two dimensions; this is exactly the situation addressed by the Lorenz equations. The upper and lower boundaries are held at constant temperatures; while there is no flux of heat across the left and right boundaries. Concerning the fluid, all boundaries are closed, but the fluid slips along the boundaries without any friction. The physical parameters of the system, namely density, viscosity, thermal diffusivity, coefficient of thermal expansion, temperature difference between lower and upper boundary, and dimensions of the considered box can be summarized with the help of the two non-dimensional parameters mentioned above. The aspect ratio of the cell is $\sqrt{2}$. We will see in the following section that this is the preferred geometry if the model parameters are close to the point where convection starts. However, the Lorenz equations can be written for any other aspect ratio, too.

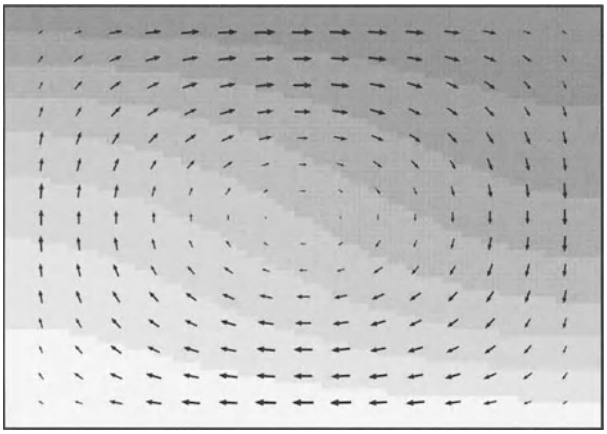

Fig. 4.1. Example of a simulated convecting system in a rectangular, two-dimensional domain. Bright areas indicate high temperatures, dark areas low temperatures.

4.2 The Physics Behind the Lorenz Equations

In this section, the Lorenz equations are derived from the basic equations of continuum mechanics, namely, from the conservation of mass, momentum, and energy. However, understanding the physical background of the Lorenz equations is not essential for understanding the rest of this book, so the reader should feel free to skip this section.

Thermal convection is a result of an interaction between fluid flow and heat transport. A minimum set of variables required for modeling thermal convection consists of the flow velocity vector $\boldsymbol{v}$, the fluid pressure p, and the temperature T. All of them depend on the spatial coordinate vector $\boldsymbol{x}$ and on time t. Determining the evolution of these variables involves the following set of partial differential equations:

- The *mass balance*:
$$\frac{\partial}{\partial t}\rho \;=\; -\operatorname{div}(\rho \boldsymbol{v}),$$
where ρ is the density of the fluid, and div denotes the divergence operator.
- The *Navier-Stokes equations* which constitute the conservation of momentum in a Newtonian fluid:
$$\rho\left(\frac{\partial}{\partial t}\boldsymbol{v} + (\boldsymbol{v}\cdot\nabla)\boldsymbol{v}\right) \;=\; -\rho g\,\boldsymbol{e} - \nabla p + \rho\nu\Delta\boldsymbol{v},$$
where g is the gravitational acceleration, $\boldsymbol{e}$ a unit vector pointing in upward direction, ∇ the gradient operator, ν the kinematic viscosity of the fluid, and Δ the Laplace operator.
- The *heat equation* describing of conductive and advective heat transport:
$$\rho c\,\frac{\partial}{\partial t}T \;=\; \operatorname{div}\left(\rho c\kappa\nabla T - \rho c T\boldsymbol{v}\right),$$
where c denotes the specific heat capacity and κ the thermal diffusivity of the fluid. Heat production by viscous friction is neglected.

Deriving these basic equations would be beyond the scope of this book; derivations can be found in any textbook on continuum physics.

Thermal convection arises from the fact that the mass density ρ decreases with increasing temperature. As long as the variations are small, the Boussinesq approximation can be applied. This means that ρ is constant everywhere except for the first term at the right-hand side of the Navier-Stokes equations which concerns the gravitational force and thus describes effects of buoyancy. Here, the linear approximation

$$\rho \;=\; \rho_0 \left(1 - \alpha \left(T - T_0\right)\right)$$

is used, where ρ_0 denotes the density at a reference temperature T_0, and α is the coefficient of thermal expansion. In this approximation, the governing equations turn into

$$\begin{aligned} \operatorname{div} \boldsymbol{v} &= 0, \\ \tfrac{\partial}{\partial t}\boldsymbol{v} + (\boldsymbol{v}\cdot\nabla)\boldsymbol{v} &= -\left(1 - \alpha\left(T - T_0\right)\right) g\, \boldsymbol{e} - \tfrac{1}{\rho_0}\nabla p + \nu \Delta \boldsymbol{v}, \\ \tfrac{\partial}{\partial t} T &= \kappa\, \Delta T - \boldsymbol{v}\cdot\nabla T, \end{aligned}$$

where all parameters are assumed to be constant now.

The next step is introducing non-dimensional coordinates and variables. Let us consider a rectangular domain of width w ($0 \le x_1 \le w$) and height h ($0 \le x_2 \le h$) in two dimensions. Rescaling the variables and coordinates according to

$$\boldsymbol{x} \;:=\; \tfrac{1}{h}\,\boldsymbol{x}, \quad t \;:=\; \tfrac{\kappa}{h^2}\,t, \quad \boldsymbol{v} \;:=\; \tfrac{h}{\kappa}\,\boldsymbol{v}, \quad \text{and} \quad w \;:=\; \tfrac{1}{h}\,w$$

leaves the equation of mass conservation unaffected, while the Navier-Stokes equations and the heat equation turn into

$$\begin{aligned} \tfrac{\partial}{\partial t}\boldsymbol{v} + (\boldsymbol{v}\cdot\nabla)\boldsymbol{v} &= Pr\left(-\tfrac{h^2}{\rho_0 \nu \kappa}\nabla p - \tfrac{g h^3 (1-\alpha(T-T_0))}{\nu\kappa}\,\boldsymbol{e} + \Delta \boldsymbol{v}\right), \\ \tfrac{\partial}{\partial t} T &= \Delta T - \boldsymbol{v}\cdot\nabla T \end{aligned}$$

with the *Prandtl number* $Pr := \frac{\nu}{\kappa}$.

We assume that the bottom and the top of the model domain are held at constant temperatures T_1 and T_2, respectively. This is expressed by the *Dirichlet boundary condition*

$$T \;=\; T_1 \quad \text{if} \quad x_2 = 0 \quad \text{and} \quad T \;=\; T_2 \quad \text{if} \quad x_2 = 1.$$

There shall be no heat flux across the boundary at the sides of the domain, the corresponding *Neumann boundary condition* reads:

$$\tfrac{\partial}{\partial x_1} T \;=\; 0 \quad \text{if} \quad x_1 \in \{0, w\}.$$

In case of purely conductive heat transport ($\boldsymbol{v} = 0$), a linear vertical temperature profile solves the heat equation under the given boundary conditions.

Since we are interested in the effect of convective transport, we consider the deviation of the actual temperature distribution (resulting from conduction and convection) from the linear profile. So let us define the non-dimensional temperature deviation

$$\vartheta \;:=\; \tfrac{gh^3\alpha}{\nu\kappa}\,(T-(T_1+(T_2-T_1)\,x_2))$$

which must satisfy the boundary conditions

$$\tfrac{\partial}{\partial x_1}\vartheta \;=\; 0 \quad \text{if} \quad x_1 \in \{0,w\} \quad \text{and} \quad \vartheta \;=\; 0 \quad \text{if} \quad x_2 \in \{0,1\}.$$

Finally, it is convenient to shift and rescale the pressure p according to

$$p \;:=\; \tfrac{h^2}{\rho_0\nu\kappa}\left(p+\rho_0 g h\, x_2\left(1-\alpha\left(T_1-T_0+\tfrac{1}{2}\,(T_2-T_1)\,x_2\right)\right)\right),$$

which alters the Navier-Stokes equations and the heat equation to

$$\begin{aligned}\tfrac{\partial}{\partial t}\boldsymbol{v}+(\boldsymbol{v}\cdot\nabla)\boldsymbol{v} \;&=\; Pr\,(-\nabla p+\vartheta\,\boldsymbol{e}+\Delta\boldsymbol{v}),\\ \tfrac{\partial}{\partial t}\vartheta \;&=\; \Delta\vartheta+Ra\,v_2-\boldsymbol{v}\cdot\nabla\vartheta\end{aligned}$$

with the *Rayleigh number* $Ra := \frac{gh^3\alpha(T_1-T_2)}{\nu\kappa}$.

The next step is expressing the flow velocity vector $\boldsymbol{v}$ by a scalar function. As long as the boundary conditions do not concern p, we can eliminate p by applying the curl operator to the Navier-Stokes equations; this leads to

$$\tfrac{\partial}{\partial t}\mathrm{curl}\boldsymbol{v}+(\boldsymbol{v}\cdot\nabla)\mathrm{curl}\boldsymbol{v} \;=\; Pr\left(\tfrac{\partial}{\partial x_1}\vartheta+\Delta\mathrm{curl}\boldsymbol{v}\right),$$

where

$$\mathrm{curl}\boldsymbol{u} \;=\; \tfrac{\partial}{\partial x_1}u_2-\tfrac{\partial}{\partial x_2}u_1$$

for an arbitrary vector function $\boldsymbol{u}$ in two dimensions. Then, $\boldsymbol{v}$ is written in the form

$$v_1 \;=\; -\tfrac{\partial}{\partial x_2}\psi \quad \text{and} \quad v_2 \;=\; \tfrac{\partial}{\partial x_1}\psi.$$

The scalar function ψ is called *stream function.* As a major advantage of this approach, the mass balance $\mathrm{div}\boldsymbol{v}=0$ is automatically satisfied.

At this point we should be aware that these two steps – applying the curl operator and introducing the stream function – are non-trivial. In general, applying a differential operator to a differential equation causes a loss of information, so that the new equation does not determine the solution uniquely. Moreover, it is not clear that each velocity field can be represented by a stream function. However, with some theory of elliptic differential equations it can be shown that the two steps are correct if there is no flux across the boundaries:

$$v_1 \;=\; 0 \quad \text{if} \quad x_1 \in \{0,w\} \quad \text{and} \quad v_2 \;=\; 0 \quad \text{if} \quad x_2 \in \{0,1\}.$$

As a result of this boundary condition, ψ must be constant along the entire boundary, so that we can assume $\psi = 0$ since ψ can be shifted by a constant without affecting $\boldsymbol{v}$. However, theory of the Navier-Stokes equations shows that the posed boundary condition is not sufficient; we need a second boundary condition that describes the interaction of the fluid with the boundary. Let us assume that the fluid can slip freely along the boundaries without any friction; this *free-slip boundary condition* can be written in the form

$$\tfrac{\partial}{\partial x_1} v_2 \;=\; 0 \quad \text{if} \quad x_1 \in \{0, w\} \quad \text{and} \quad \tfrac{\partial}{\partial x_2} v_1 \;=\; 0 \quad \text{if} \quad x_2 \in \{0, 1\},$$

which finally leads to the following boundary conditions for the stream function:

$$\psi \;=\; \Delta\psi \;=\; 0 \quad \text{if} \quad x_1 \in \{0, w\} \quad \text{or} \quad x_2 \in \{0, 1\}.$$

Inserting the stream function approach into the governing equations leads to

$$\tfrac{\partial}{\partial t}\Delta\psi + \tfrac{\partial}{\partial x_1}\psi \tfrac{\partial}{\partial x_2}\Delta\psi - \tfrac{\partial}{\partial x_2}\psi \tfrac{\partial}{\partial x_1}\Delta\psi \;=\; Pr\left(\tfrac{\partial}{\partial x_1}\vartheta + \Delta\Delta\psi\right)$$

and

$$\tfrac{\partial}{\partial t}\vartheta \;=\; \Delta\vartheta + Ra\,\tfrac{\partial}{\partial x_1}\psi + \tfrac{\partial}{\partial x_1}\vartheta \tfrac{\partial}{\partial x_2}\psi - \tfrac{\partial}{\partial x_2}\vartheta \tfrac{\partial}{\partial x_1}\psi.$$

The simplest non-trivial functions ψ and ϑ which satisfy the boundary conditions are

$$\psi \;=\; A \sin(\tfrac{\pi}{w} x_1) \sin(\pi x_2) \quad \text{and} \quad \vartheta \;=\; B \cos(\tfrac{\pi}{w} x_1) \sin(\pi x_2),$$

where A and B are arbitrary functions of time. The approach describes a simple convection cell as illustrated in Fig. 4.1 if $A > 0$. For $A < 0$, a counterclockwise circulation occurs. The absolute value of A corresponds to the intensity of convection. If $B > 0$, the temperature in the left-hand part of the domain is higher than in the right-hand part, and vice versa.

Inserting these approaches into the governing equations and comparing the coefficients in front of the trigonometric functions yields

$$\begin{aligned} \tfrac{\partial}{\partial t} A &= Pr\left(-\tfrac{\pi^2(1+w^2)}{w^2} A + \tfrac{w}{\pi\,(1+w^2)} B\right), \\ \tfrac{\partial}{\partial t} B &= Ra\,\tfrac{\pi}{w} A - \tfrac{\pi^2(1+w^2)}{w^2} B. \end{aligned}$$

However, this approach only provides an exact solution of the Navier-Stokes equations, whereas the heat equation is only solved approximately because the term

$$\tfrac{\partial}{\partial x_1}\vartheta \tfrac{\partial}{\partial x_2}\psi - \tfrac{\partial}{\partial x_2}\vartheta \tfrac{\partial}{\partial x_1}\psi \;=\; -\tfrac{\pi^2}{2\sqrt{2}}\, A\, B \sin(2\pi x_2)$$

remains.

This linear system was introduced by Rayleigh (1916) for his analysis of the onset of convection. Such linear systems where the coefficients are constant through time can be solved analytically; the solution is a linear

combination of (in general complex) exponential functions. Basic properties of the solutions can be directly derived by writing the system in matrix form:

$$\frac{\partial}{\partial t}\begin{pmatrix} A \\ B \end{pmatrix} = \begin{pmatrix} -\frac{Pr\,\pi^2(1+w^2)}{w^2} & \frac{Pr\,w}{\pi(1+w^2)} \\ \frac{Ra\,\pi}{w} & -\frac{\pi^2(1+w^2)}{w^2} \end{pmatrix}\begin{pmatrix} A \\ B \end{pmatrix}.$$

An exponentially growing solution exists if at least one of the (complex) eigenvalues of the matrix has a positive real part, otherwise the solutions decay exponentially. The eigenvalues λ of the matrix are

$$\lambda = -\frac{(1+Pr)\pi^2(1+w^2)}{2w^2} \pm \sqrt{\left(\frac{(1+Pr)\pi^2(1+w^2)}{2w^2}\right)^2 + \frac{Ra\,Pr}{1+w^2} - \frac{Pr\,\pi^4(1+w^2)}{w^4}}.$$

A positive solution exists if the sum of the second and third term in the square root is positive, which is the case if Ra exceeds the *critical Rayleigh number*

$$Ra_c := \frac{\pi^4(1+w^2)^3}{w^4},$$

Consequently, each disturbance introduced in the purely conductive case ($A = B = 0$) results in (infinitely growing) convection if $Ra > Ra_c$. On the other hand, convection ceases if $Ra < Ra_c$. The critical Rayleigh number depends on the aspect ratio w; its minimum value of about 657.5 is achieved for $w = \sqrt{2}$. Thus, convection cells with an aspect ratio of $w = \sqrt{2}$ are preferred in case of low Rayleigh numbers, so let us assume this aspect ratio in the following.

However, infinitely growing convection is not realistic; there should be a limitation arising from the non-linear term in the heat equation. The structure of the remaining terms suggests to add a term to ϑ which is proportional to $\sin(2\pi x_2)$, so that

$$\vartheta = B\cos(\tfrac{\pi}{\sqrt{2}}x_1)\sin(\pi x_2) - C\sin(2\pi x_2),$$

where C is a function of time. This additional term describes thermal layering. If $C > 0$, the major temperature gradients are confined to the lower and upper boundaries, while the temperature differences are concentrated in the middle if $C < 0$. The choice of the sign reflects the fact that the latter case rarely occurs. This modified approach still satisfies the Navier-Stokes equations, so that the first of Rayleigh's equations persists. Inserting the approach into the heat equation leads to

$$\begin{aligned} &\tfrac{\partial}{\partial t}B\cos(\tfrac{\pi}{\sqrt{2}}x_1)\sin(\pi x_2) - \tfrac{\partial}{\partial t}C\sin(2\pi x_2) \\ &\quad = -\tfrac{3\pi^2}{2}B\cos(\tfrac{\pi}{\sqrt{2}}x_1)\sin(\pi x_2) + 4\pi^2 C\sin(2\pi x_2) \\ &\qquad + Ra\,\tfrac{\pi}{\sqrt{2}}A\cos(\tfrac{\pi}{\sqrt{2}}x_1)\sin(\pi x_2) - \tfrac{\pi^2}{2\sqrt{2}}AB\sin(2\pi x_2) \\ &\qquad - \tfrac{\pi^2}{\sqrt{2}}AC\cos(\tfrac{\pi}{\sqrt{2}}x_1)\left(\sin(\pi x_2) - \sin(3\pi x_2)\right). \end{aligned}$$

As a consequence of the last term, this approach does still not solve the heat equation exactly; but it provides a better approximation than Rayleigh's approach. If we compare the factors in front of the sine and cosine expressions, we obtain a first version of the Lorenz equations:

$$\begin{aligned}
\frac{\partial}{\partial t}A &= Pr\left(-\frac{3\pi^2}{2}A + \frac{\sqrt{2}}{3\pi}B\right), \\
\frac{\partial}{\partial t}B &= -\frac{3\pi^2}{2}B + Ra\,\frac{\pi}{\sqrt{2}}A - \frac{\pi^2}{\sqrt{2}}AC, \\
\frac{\partial}{\partial t}C &= -4\pi^2 C + \frac{\pi^2}{2\sqrt{2}}AB.
\end{aligned}$$

Rescaling the variables A, B, and C and the time t according to

$$A \coloneqq \tfrac{1}{3}A, \quad B \coloneqq \tfrac{2\sqrt{2}}{27\pi^3}B, \quad C \coloneqq \tfrac{4}{27\pi^3}C, t \coloneqq \tfrac{3\pi^2}{2}t$$

and introducing the relative Rayleigh number $r \coloneqq \frac{Ra}{Ra_c}$ leads to the established form of the Lorenz equations (Eq. 4.1).

4.3 Phase Space, Attractors, and Bifurcations

The Lorenz equations are a low-dimensional approximation of thermal convection processes. The low dimension refers to the dimension of the *phase space* of the system. The phase space is spanned by all the variables which are required for characterizing the state of a system uniquely. Let us begin with the simple example of a ball moving around in space. If there are no limitations, the ball can be everywhere, so we need a three-dimensional vector for characterizing its position. However, this is not enough; even if the force acting on the ball is given, predicting the motion of the ball requires keeping track of the velocity, too. Thus we need a second three-dimensional vector, so that we can characterize the state of the ball by a point in an abstract, six-dimensional space spanned by location and velocity. Formally, the result that the dimension of phase space is twice the spatial dimension in this example reflects the fact that the equations of motion (Newton's laws) are differential equations of second order. Introducing further degrees of freedom, such as rotation and deformation, requires a further extension of phase space.

In continuum physics, the dimension of phase space is infinite because degrees of freedom are assigned to each point in space. In our example of thermal convection, a temperature, a fluid pressure, and a fluid velocity are assigned to each point of the model domain. In this sense, the Lorenz equations are an attempt to approximate this infinite variety with few degrees of freedom, namely, with three variables $A(t)$, $B(t)$, and $C(t)$. Obviously, such a reduction can only be an approximation with a limited range of validity. Transferred to our example, describing thermal convection by Lorenz equations is only appropriate for small Rayleigh numbers. However, this limitation

only concerns the relevance of the Lorenz equations to nature, but not the following general aspects of non-linear system behavior.

The Lorenz equations only contain first-order derivatives; if the values of the three variables are known at a time, the equations define the state of the system completely. Thus, explicitly keeping track of the derivatives is not necessary, so that the phase space of this system is three-dimensional. As explained in Sect. 4.2, $A(t)$ quantifies the direction and the intensity of the fluid's motion, while $B(t)$ and $C(t)$ describe the temperature field.

Let us first consider the stationary solutions of the Lorenz equations, i. e., solutions where $A(t)$, $B(t)$, and $C(t)$ are constant. These solutions are called *fixed points* in phase space; if the system is at a fixed point at any time, it will stay there forever. The purely conductive case which is characterized by $A(t) = B(t) = C(t) = 0$ is a fixed point of the Lorenz equations, but it is not the only one. If the relative Rayleigh number r exceeds unity, there are two additional fixed points which are characterized by

$$A(t) \;=\; B(t) \;=\; \pm\sqrt{\tfrac{8}{3}(r-1)} \quad \text{and} \quad C(t) \;=\; r-1.$$

Both solutions only differ concerning the direction of circulation.

Figure 4.2 shows the velocity and temperature fields at the fixed points (positive solutions) for different Rayleigh numbers r. For small Rayleigh num-

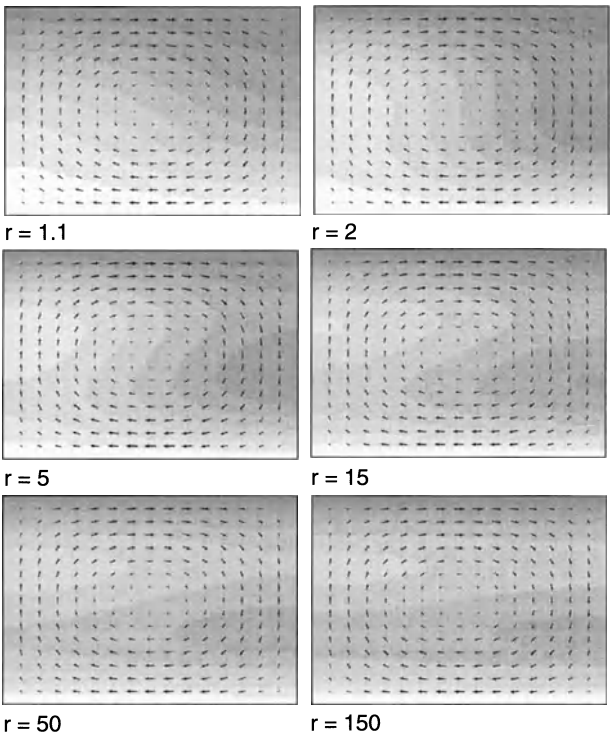

Fig. 4.2. Velocity and temperature fields at the fixed points of the Lorenz equations for different relative Rayleigh numbers r. Bright areas indicate high temperatures, dark areas low temperatures. The velocity vectors are scaled individually for each diagram and thus cannot be compared.

bers, the temperature field is deformed by the flow as expected. For higher Rayleigh numbers, thermal boundary layers evolve; the highest temperature gradients are concentrated near the upper and lower boundaries. This phenomenon can easily be understood: If convection is strong, nearly the whole domain is mixed efficiently, so that the temperature differences are small there. Mixing cannot take place near the upper and lower boundaries since the boundary conditions require that the vertical component of the velocity vanishes there. Thus, heat is transferred between the boundary and the mixed region only by conduction which requires high temperature gradients.

Beside layering, a temperature maximum near the left boundary and a minimum near the right boundary occur at high Rayleigh numbers. From the theory of the heat equation it is known that minima and maxima cannot occur in an exact, stationary solution. Thus, they are an artefact of the approximation leading to the Lorenz equations and give evidence that the Lorenz equations are not a good approximation at high Rayleigh numbers.

The practical relevance of a fixed point depends on its stability. A fixed point is *stable* if the system approaches the fixed point once it is close to it. Stable the fixed points are called *attractors*. If, on the other hand, a fixed point is repulsive (unstable), the system's state evolves away from the fixed point. Thus, a repulsive fixed point will not be approached in a real system, so that it has not any practical importance. We can, e. g., put a ball on top of a needle theoretically, but we will normally not succeed in keeping it there.

In the example of the Lorenz equations, the stability of the fixed points can be examined analytically by linearizing the equations around the fixed points. The analysis provided in Sect. 4.2 can easily be enlarged to the purely conductive fixed point; it is an attractor for $r < 1$. For $r > 1$, this fixed point is repulsive, which only means that convection occurs then. Analyzing the stability of the two non-trivial fixed points occurring if $r > 1$ is a little more complicated, although the method of linearizing the equations around the fixed point is essentially the same. It can be shown that the convective fixed points are stable within the range $1 < r < r_\mathrm{T}$ where

$$r_\mathrm{T} = \frac{Pr\,(3\,Pr + 17)}{3\,Pr - 11}.$$

Thus, the system's behavior changes drastically at the points $r = 1$ and $r = r_\mathrm{T}$. These points are called *bifurcations*. The first bifurcation (at $r = 1$) is a *pitchfork bifurcation*; one stable fixed point is split up into two (or more, in general) stable fixed points. The bifurcation at $r = r_\mathrm{T}$ is more interesting: The two fixed points describing stationary convection cells become unstable; the system's behavior turns into a unsteady one. This kind of bifurcation is called *Hopf bifurcation*. Figure 4.3 illustrates the bifurcations for $Pr = 10$.

Figure 4.4 shows how these bifurcations divide the parameter space into three distinct regions. The transition to unsteady behavior takes place at the lowest Rayleigh numbers for $Pr \approx 10$, which is slightly larger than the Prandtl number of water ($Pr \approx 7$ at a temperature of $20°C$), but, e. g., several

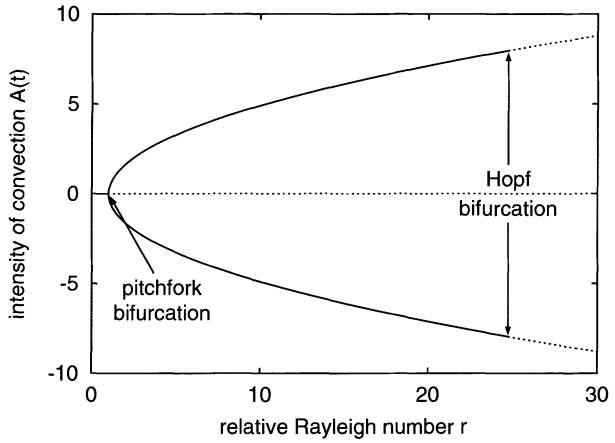

Fig. 4.3. Fixed points of the Lorenz equations and their stability at $Pr = 10$. Stable fixed points (attractors) are drawn with solid lines, unstable fixed points are dashed.

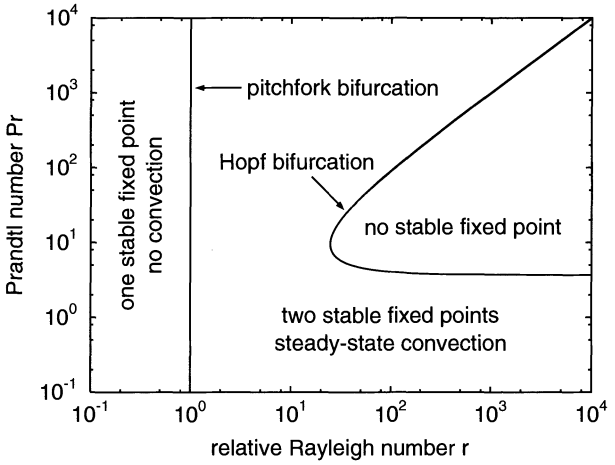

Fig. 4.4. Bifurcations and system behavior of the Lorenz equations.

orders of magnitude smaller than that of the earth's mantle. Let us assume $Pr = 10$ in the following.

4.4 Limit Cycles and Strange Attractors

Let us now consider the situation where steady-state behavior is impossible, which means that r must be larger than about 24.74 for $Pr = 10$. This parameter range is the domain of numerical experiments. Since the Lorenz equations are a quite simple set of ordinary differential equations, numerous methods for solving them approximately on a computer are available. A brief overview over the simplest and most widespread methods is given in Appendix A. The results presented in this section were obtained using the established fourth-order Runge-Kutta scheme.

The system's behavior can be visualized with the help of phase space diagrams where the trace of the solution through the phase space is plotted. Since following a curve in three-dimensional space is somewhat difficult, it is better to consider two-dimensional cross sections through phase space. Let us plot the projection of the curve onto the A-B plane into one diagram and the projection onto the C-B plane into a second one.

Figure 4.5 shows 10 traces in phase space for $r = 100$ and $Pr = 10$, starting at random initial conditions. In the beginning, the behavior appears to be quite irregular; the system fills a certain region in phase space. But over long times, the preferred region contracts until it collapses to two lines. This means that the system's behavior finally becomes periodic; the solution follows a closed curve in phase space. These curves are called *limit cycles.* There are two limit cycles for this parameter set; for nearly all initial conditions, the system finally approaches one of these cycles. Thus, a limit cycle can be seen as some kind of generalized attractor (a line instead of a point).

A less regular and thus more interesting behavior occurs if we come closer to the Hopf bifurcation. Figure 4.6 shows phase space diagrams for $r = 28$ and $Pr = 10$; this parameter set is often chosen for numerical experiments. For clarity, only a single trace starting at a randomly chosen initial condition is plotted. Obviously, the system experiences a bounded region around the fixed points; but in contrast to the limit-cycle behavior, this region does not contract to a line through time. The curve moves around one of the unstable fixed points; after some spiral-shaped loops it changes over to the vicinity of the other fixed point, and so on. There seems to be no regularity in the time spans between these transitions. Thus, the system prefers a certain region in phase space; it turns out that this region is mainly independent of the initial condition, in analogy to attractors and limit cycles. This preferred region in phase space is called *strange attractor.*

Strictly speaking, numerical experiments are not an adequate tool for distinguishing between limit cycles and strange attractors, as the following argument shows: If we use double-precision floating-point variables (64 bit), each variable is constrained to a maximum of 2^{64} discrete values. Thus, phase space of the numerical model consists of a maximum of $2^{3\times64} \approx 6\times10^{57}$ distinct points, so that the system must finally revisit a point where it was before. This simply means that numerical simulations always end at either a stable fixed point or a limit cycle. However, if the limit cycle is long enough, it is easy to believe that it is just an artefact of the discrete model, and that the behavior is chaotic in reality.

4.5 The Lyapunov Exponent

The irregular behavior observed in the previous section provided a first insight into what chaotic behavior could be, but for a deeper understanding, we need a more quantitative concept.

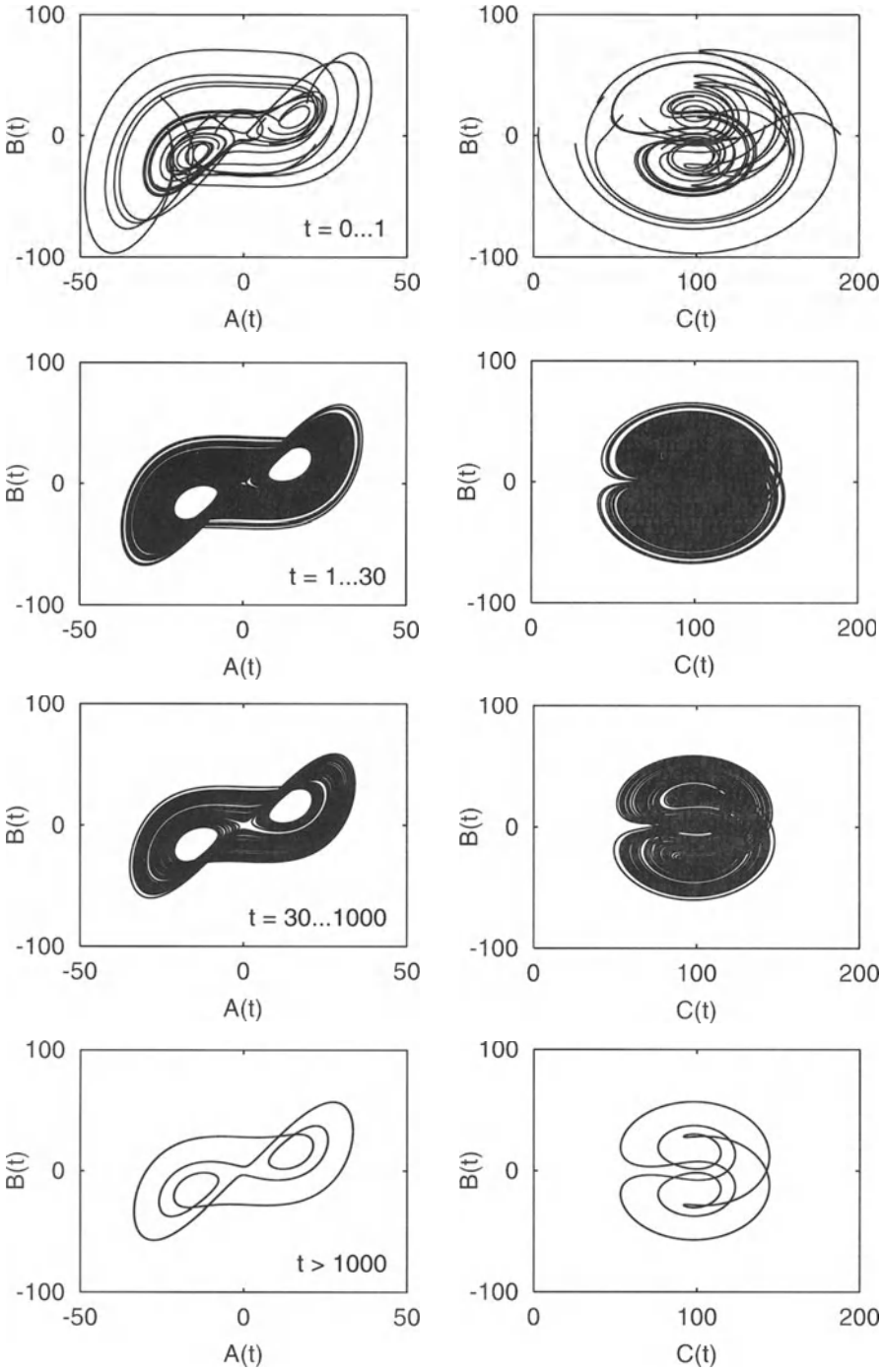

Fig. 4.5. Phase space diagrams for $r = 100$ and $Pr = 10$.

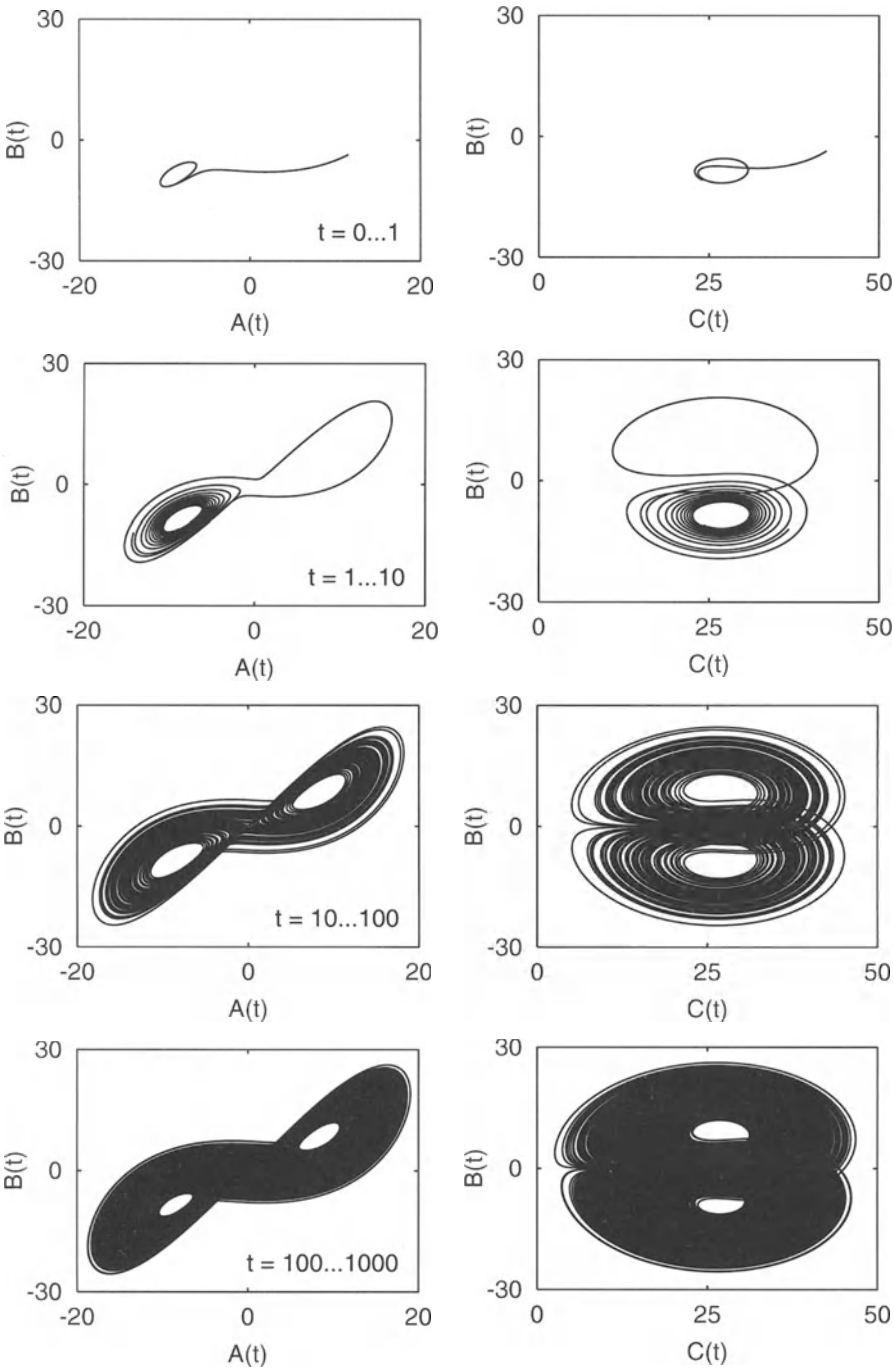

Fig. 4.6. Phase space diagrams for $r = 28$ and $Pr = 10$.

The framework of deterministic chaos hinges on considering traces in phase space which start close to each other; their convergence or divergence through time is the criterion for distinguishing between regular and chaotic behavior. Let $f(t)$ be a curve in phase space of an arbitrary system, no matter whether $f(t)$ is a scalar function or a vector, such as a solution of the Lorenz equations which consists of three components $A(t)$, $B(t)$, and $C(t)$. Let us now consider all these curves $\tilde{f}(t)$ in phase space which are sufficiently close to the original one at a given time t_0:

$$||\tilde{f}(t_0) - f(t_0)|| < \epsilon, \tag{4.2}$$

where $||\ldots||$ denotes a measure of distance in phase space, and ϵ is a small, positive number. Figure 4.7 shows what may happen to these curves in the example from Fig. 4.6. A total of 10,000 initial conditions were randomly chosen close to the initial condition from Fig. 4.6; each of their components deviates from those of the original initial condition by less than $\epsilon = 10^{-3}$. While they stay close together for a time, they tend to diverge later. In a first step, they spread on lines, but later they fill the strange attractor entirely.

Roughly speaking, the initial condition becomes irrelevant through time. The example shows that the initial condition does not help us to predict the system's state at the time $t \approx 30$, at least if we cannot specify it more precisely than $\epsilon = 10^{-3}$. For quantifying this behavior, let $\Delta_\epsilon(t)$ be the lowest number which satisfies the condition

$$||\tilde{f}(t) - f(t)|| < \Delta_\epsilon(t)$$

for all curves $\tilde{f}$ in phase space which satisfy Eq. 4.2. Figure 4.8 shows the evolution of $\Delta_\epsilon(t)$ through time for three examples with $r = 22$ (two stable fixed points exist), $r = 28$ (the data set shown in Fig. 4.7) and $r = 100$ (with initial conditions close to that of one of the traces from Fig. 4.5). Again, the maximum deviation of the initial conditions is $\epsilon = 10^{-3}$.

In all examples, $\Delta_\epsilon(t)$ increases in the beginning; the curves diverge in phase space. For $r = 22$, divergence ceases, and the curves recontract. For $t > 10$, $\Delta_\epsilon(t)$ decreases exponentially, superposed by an oscillation. The increase in the beginning indicates that the way towards one of the fixed points may be long, and the transition may look irregular in the beginning. From this we learn that we must not consider $\Delta_\epsilon(t)$ for small times, but rather in the limit $t \to \infty$.

Unfortunately, the behavior of $\Delta_\epsilon(t)$ is similar for $r = 28$ (irregular behavior) and for $r = 100$ (regular limit-cycle behavior). Still worse, the latter one even diverges faster. The problem arises from the fact that the system experiences a limited region in phase space, and this limitation finally applies to $\Delta_\epsilon(t)$, too. For $r = 100$, the way to the limit cycle is too long; during this transition the traces diverge so strongly that they finally fill the limit cycle entirely. However, the solution of this problem is straightforward: If we

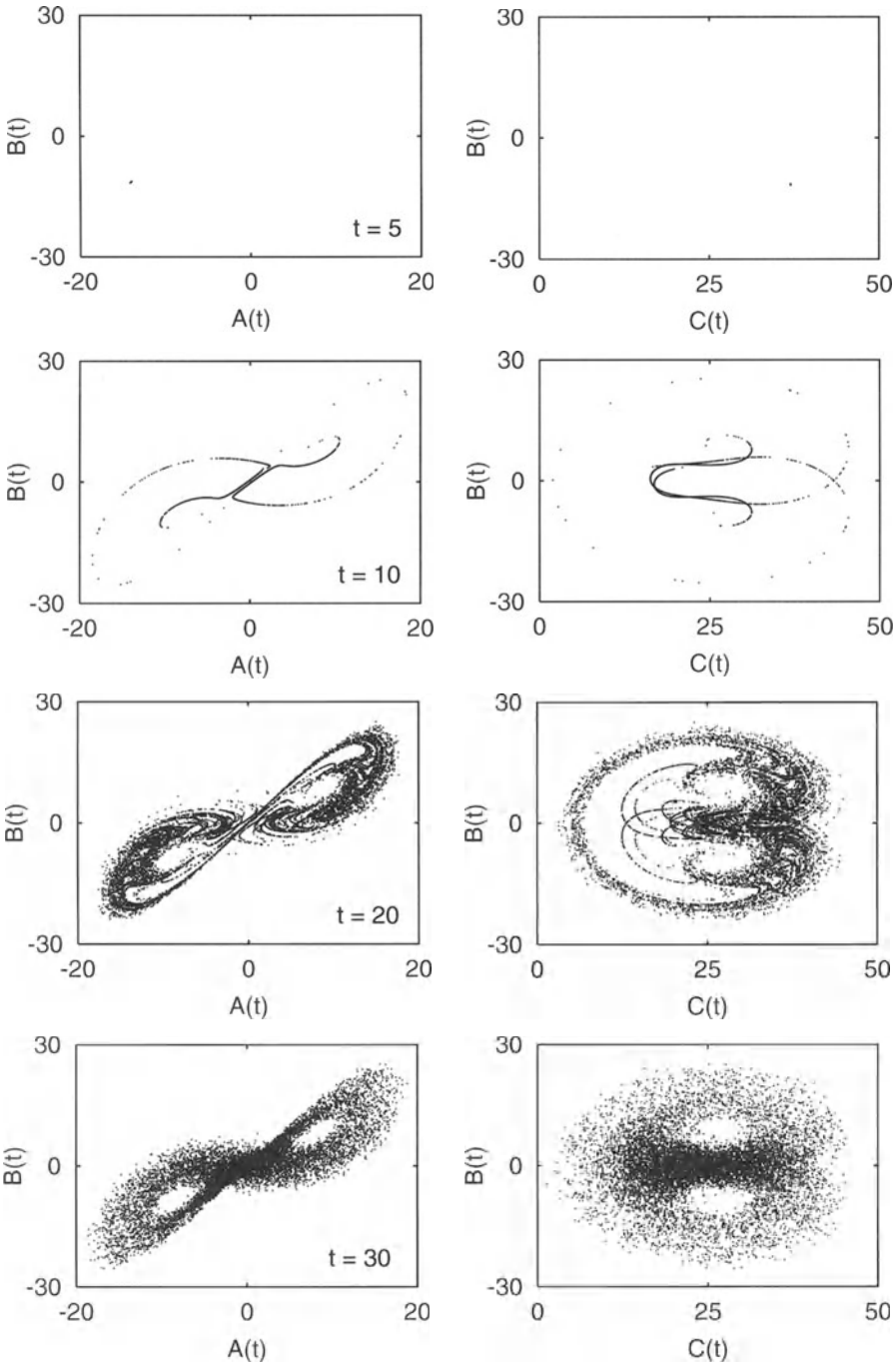

Fig. 4.7. Snapshots of 10,000 curves in phase space for $r = 28$ and $Pr = 10$, all starting close to the initial condition from Fig. 4.6.

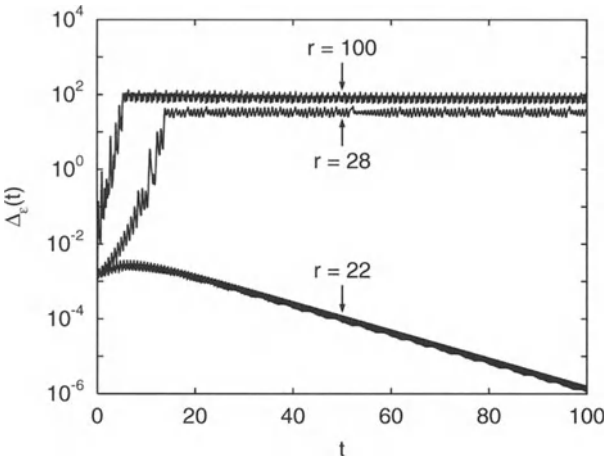

Fig. 4.8. Spreading and contracting of curves in phase space for three examples with different Rayleigh numbers.

choose smaller values of ϵ, it should take a longer time until the traces diverge. Finally, if ϵ is small enough, the limit cycle should be reached without diverging too much. Then the divergence should cease, provided that our interpretation of regular behavior at the limit cycle is correct. Figure 4.9 shows what happens for $\epsilon = 10^{-3}$, 10^{-6}, 10^{-9}, and 10^{-12}. As expected, $\Delta_\epsilon(t)$ grows up to the same value for large times in the example $r = 28$ (upper diagram), while growth ceases at $t \approx 5$ for $r = 100$ (lower diagram). For quantifying this difference, we consider the function

$$\Delta(t) = \lim_{\epsilon \to 0} \frac{\Delta_\epsilon(t)}{\epsilon}.$$

If we extrapolate from Fig. 4.9, we see that $\Delta(t)$ becomes constant (except for oscillations) in the limit $t \to \infty$ for $r = 100$, while it grows exponentially for $r = 28$. It can easily be seen that $\Delta(t)$ decreases exponentially if the system tends towards a stable fixed point. The parameter λ of exponential growth or decay: $\Delta(t) \sim e^{\lambda t}$ for $t \to \infty$ is called *Lyapunov exponent*; it can be formally defined by

$$\lambda = \lim_{t \to \infty} \frac{\log(\Delta(t))}{t}.$$

The system's behavior is *chaotic* if $\lambda > 0$.

Strictly speaking, each point in phase space (being the initial condition of the considered trace) and each time t_0 has its own Lyapunov exponent. Even in the range $1 < r < r_T$, where we have two stable attractors, some points have positive Lyapunov exponents and are thus chaotic. Let us, e. g., consider initial conditions where $A(0) = B(0) = 0$, while $C(0)$ is arbitrary. For $t \to \infty$, the system tends towards the (unstable) fixed point $A(t) = B(t) = C(t) = 0$. If we apply a little disturbance to $A(t)$ or $B(t)$, the system approaches one of the stable fixed points, no matter how small the disturbance is. This divergence results in a positive Lyapunov exponent, so that there is at least

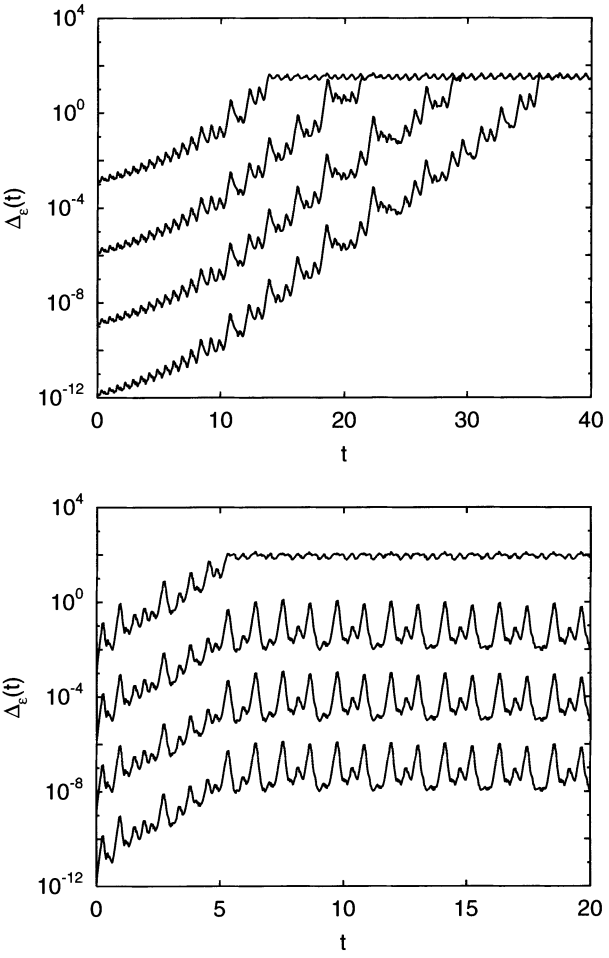

Fig. 4.9. Spreading of curves in phase space for $r = 28$ (upper diagram) and $r = 100$ (lower diagram) and $\epsilon = 10^{-3}, 10^{-6}, 10^{-9}$, and 10^{-12}.

a chaotic line in phase space. The same is true in case of limit-cycle behavior. However, as long as set of chaotic points is a low-dimensional region in phase space, it has no practical relevance. Chaotic systems are those where nearly all points in phase space have positive Lyapunov exponents.

4.6 Does it Matter whether God Plays Dice?

In Chap. 3, random processes with self-affine scaling behavior (FBM) were discussed. While these processes involve randomness on an elementary level, the Lorenz equations are entirely deterministic. Here, randomness may only concern the initial conditions unless random terms are explicitly added. For this reason, the behavior of the Lorenz equations is predictable, but only

theoretically. If once the initial conditions are fixed, the system's state at any time can be predicted in principle, but practically this is impossible due to the limited accuracy. This behavior is denoted *deterministic chaos.*

Quantum mechanics was the first field in physics where randomness plays a central part. Instead of describing objects (electrons, nuclei etc.) by a trace in space (or better by one in a six-dimensional phase space which includes velocity, too), only probability densities for the occurrence of an object in a certain state (perhaps at a certain location, momentum or energy) are considered. Everything that goes beyond such a probability distribution is interpreted to be random. Although quantum mechanics turned out to be one of the most successful concepts in twentieth-century physics, this randomness at an elementary level did not fit into the established, deterministic view of the universe. Hence early discussion on quantum mechanics often focused on the question whether God plays dice or not.

However, the difference is not really important for us. We have learned that we cannot predict the result of the Lorenz equations over long times in the parameter region of deterministic chaos. Thus, writing down the Lorenz equations may be a good theory for explaining the behavior over short times, but over long times it may be better just to give the probability density of the Lorenz attractor in phase space and say that the state of the system is a random function following this distribution. We know that this theory is strictly speaking not correct since there is no randomness at all; but for someone who cannot look at short time intervals and recognize the regularity, there is no reason to raise doubts against it.

From this point of view, it is not important whether something is in fact random or just looks as if it was. Random numbers generated on a computer are everything but random; they are computed using a strictly deterministic algorithm which may include several hidden parameters such as time and process identification number. This is not a problem as long as we do not apply an analysis which is able to reveal the system behind the random numbers; but in this case we may run into problems because basic laws of statistics are violated. Thus, a good random number generator is one which hides the system behind its output well. From this point of view, even the Lorenz equations may provide a good, but perhaps a little slow random number generator.

4.7 Deterministic Chaos and Self-Affine Fractals

The random processes considered in Chap. 3 (FBM) were generated by rather abstract algorithms without any straightforward physical interpretation, except for white noise and Brownian motion. In the previous section we have learned that the difference between real randomness and apparent randomness as it emerges from deterministic chaos is not important under many

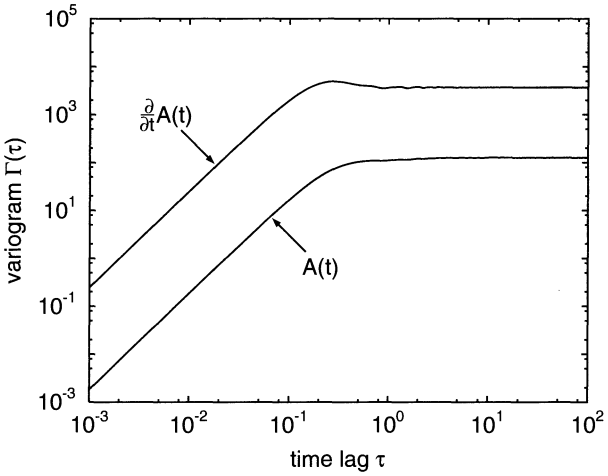

Fig. 4.10. Variogram analysis of $A(t)$ and $\frac{\partial}{\partial t}A(t)$.

aspects. So let us just apply the methods for analyzing self-affine scaling behavior and predictability from Chap. 3 to solutions of the Lorenz equations and figure out whether they exhibit self-affine scaling properties and can thus provide insights into the origin of self affinity in physical systems.

The lower curve plotted in Fig. 4.10 shows the variogram (Eq. 3.9) of the intensity of convection, described by the variable $A(t)$. For time lags $\tau < 0.1$, $\Gamma(\tau)$ increases like τ^2. According to the results obtained in Sect. 3.9, this behavior indicates that the variation of the solution is quite regular. Transferred to FBM, it corresponds to a spectral exponent $\beta \geq 3$, respectively, to a the non-fractal limiting case $D = 1$. This finding is not really surprising since the Lorenz equations are a set of quite regular differential equations. As $A(t)$, $B(t)$, and $C(t)$ are bounded, their derivatives are bounded according to the differential equations, too. Thus,

$$|A(t+\tau) - A(t)| \approx \frac{\partial}{\partial t}A(t)\,\tau \quad \text{for} \quad \tau \to 0, \tag{4.3}$$

which implies that the variogram increases like τ^2. The derivative $\frac{\partial}{\partial t}A(t)$ shows the same behavior, so that taking the derivative does not affect the scaling behavior in this example.

For $\tau > 1$, the variogram becomes constant, which is just opposite to the regular behavior at short times. In this case, the system's behavior is essentially unpredictable, and the fractal dimension is $D = 2$. The region of transition between both extreme behaviors is narrow; there is no range where a power law with a scaling exponent between those found above occurs. In other words, the Lorenz equations switch between the two non-fractal limiting cases without showing any non-trivial self-affine scaling properties. So they help us to understand what chaotic behavior is, but cannot contribute to understanding of self-affine scaling behavior in time series.

5. Self-Organized Criticality

In the previous chapter we have learned about non-linear systems, especially about deterministic chaos. Irregular fluctuations turned out to be the fingerprint of deterministic chaos. However, although fluctuations are ubiquitous in many natural phenomena, nature is obviously not completely irregular. Otherwise, phenomena involving a certain degree of order, such as life, would be impossible. From this point of view, nature often seems to be somewhere between chaos and order. *Complexity* has become the magic word in this context, but let us refrain from trying giving a precise definition of complexity. Let us, instead, recapitulate which of the topics discussed in the previous chapters may fit into this context.

When discussing fractional Brownian motion (FBM) in Chap. 3, we observed different behavior concerning predictability. White noise was found to be completely irregular and unpredictable, while Brownian motion turned out to be more regular and showed at least a limited predictability. In the middle between both there is pink noise: Although looking more regular than white noise, it is still unpredictable – just at the edge of predictability. Under these aspects, pink noise may be the temporal fingerprint of complexity. However, when examining deterministic chaos in the example of the Lorenz equations (Chap. 4), we did not observe such a behavior. Depending on the considered time scale, the Lorenz equations switch from regular to completely irregular behavior. The region of transition between both is narrow, so that it cannot be interpreted as pink noise. Consequently, at least this type of chaos is not complexity.

But what is the spatial counterpart to pink noise in complexity? Since FBM is a (self-affine) fractal in time, the idea of spatial scale invariance being the spatial fingerprint of complexity is tempting. As already mentioned, nature provides several examples of fractal size distributions. Some of them, e. g., the size distributions of rock fragments discussed in Chap. 1 can be satisfactorily explained as the final result of a process. On the other hand, there are several examples of power-law distributions which arise from counting events during a process; we will consider forest fires, earthquakes, and landslides as the most prominent examples from earth sciences later.

In their two seminal papers, Bak et al. (1987, 1988) introduced a completely new mechanism resulting in fractals concerning both space and time

– *self-organized criticality*. In the following ten years, more than 2000 papers on self-organized criticality were published, making the 1987 paper the most-cited paper in physics during this period. However, everyone should know that even the best idea is worthless without an acronym such as SOC. This acronym may stand for both the noun *self-organized criticality* and the adjective *self-organized critical*.

The framework of SOC was originally developed introducing a simple cellular automaton model, often referred to as *Per Bak's sandpile* or *Bak-Tang-Wiesenfeld model*. But before discussing this model in detail, we start with an introduction to the physics of criticality in the classical sense.

5.1 Critical-Point Phenomena

The term *critical* frequently occurs not only in physics, but also in everyday life. Roughly speaking, things are critical if they tend to run out of control. Let us develop a more precise definition of criticality using the example of a *nuclear chain reaction*, which can be seen as a prototype of an *avalanche*. Light nuclei tend to consist half of protons and half of neutrons, while the number of neutrons exceeds the number of protons in heavy nuclei. Thus, fissionating a heavy nucleus into two smaller parts often results in a relative excess of neutrons; these neutrons are emitted until the smaller nuclei become stable. For instance, fissionating a nucleus of uranium ^{235}U yields about $2\frac{1}{2}$ neutrons in the mean. A part of these neutrons is able to destroy other nuclei, leading to the emission of further neutrons, and so on. Roughly speaking, this is the principle of a nuclear chain reaction.

The probability that an emitted neutron causes further fissions depends on the concentration of fissionable nuclei in the sample. If this concentration is low, the emitted neutrons lose their energy in collisions with non-fissionable atoms. Let us assume that each fission releases two neutrons, and that each of them fissionates another nucleus with a probability q. This model can easily be simulated on a computer. Figure 5.1 shows the resulting cumulative size distribution of the avalanches, i. e., the probability P_n that a chain reaction involves at least n nuclei, for different values of the probability q.

Obviously, the behavior changes drastically at $q = \frac{1}{2}$; this point is a bifurcation (Sect. 4.3). This is not surprising since each fissionated nucleus destroys one more atom in the mean then. For $q < \frac{1}{2}$, the probability of large avalanches decreases rapidly, so that the avalanches are in fact limited in size; this situation is called *subcritical*. On the other hand, P_n does not tend towards zero in the limit $n \to \infty$ if $q > \frac{1}{2}$. In this case, a certain fraction of fissions causes chain reactions of infinite size, although this is, strictly speaking, only true for infinite samples. This unstable situation is called *overcritical*; if the chain reaction takes place instantaneously, it results in a nuclear explosion. As Fig. 5.1 illustrates, the fraction of infinite avalanches increases if q increases; it is about 8 % for $q = 0.51$ and 56 % for $q = 0.6$.

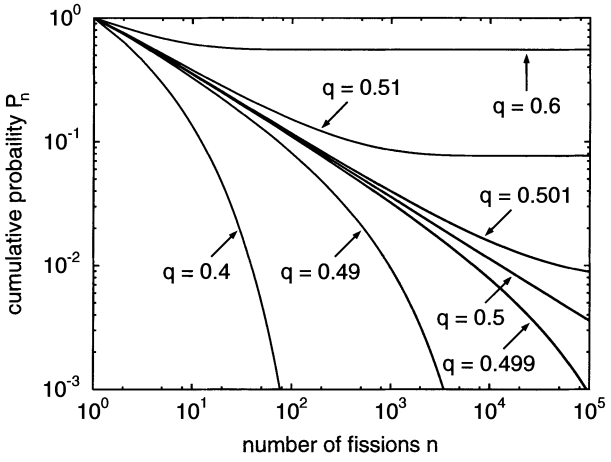

Fig. 5.1. Cumulative size distribution of a nuclear chain reaction releasing two neutrons per fission for different values of the fissionating probability q.

The situation where $q = \frac{1}{2}$ is called *critical point* or *critical state*; it is characterized by a power-law decay of the event size distribution which indicates scale-invariant properties. In the critical state, events of all sizes occur, but large events are less likely than small events.

The amount of energy released by a single fission is quite low; about 4×10^{10} uranium nuclei must be fissionated for an energy of 1 Joule (which is in fact not much). Thus, even the largest avalanches occurring at $q = 0.499$ are negligible from a macroscopic point of view. As a consequence, building a nuclear reactor is not as easy as it may seem; at least it is hopeless to obtain considerable amounts of energy per initial fission, but to avoid a nuclear explosion just by choosing an appropriate density of fissionable material.

This example reveals a general property of many critical systems. Criticality arises at a microscopic scale, and macroscopic effects can only be observed if the system is precisely *tuned* to its critical point; even slightly subcritical states are far away from criticality if macroscopic properties are considered. So, why should nature tune its phenomena exactly to be critical?

Phase transitions in equilibrium thermodynamics are well-understood critical-point phenomena (e. g. Binney et al. 1992). Below a critical temperature which is characteristic for the fluid, perturbations in the system only affect the local neighborhood; their effect decays exponentially with the distance. In contrast, the exponential decay turns into a power law at the critical temperature. Then, a local distortion may propagate through the entire system. Finally, the situation becomes completely unstable above the critical temperature, so that liquid and gaseous state cannot be distinguished any longer for overcritical fluids.

Starting from this example, one may argue that parameters are not constant in nature, but vary spatially and temporally. Let us, e. g., consider the properties of water in the earth's crust. Since the temperature increases with depth, there should be a depth where the critical temperature of about

647.4 K is achieved. Thus, water must be critical somewhere in the earth's crust. However, this is strictly speaking, a layer of zero thickness. Above this layer, water is subcritical, while it is overcritical in larger depths. So there is no reason why a considerable amount of water should be critical in the earth's crust. The same argument holds for temporally varying parameters; if, e. g., the temperature slowly increases, the system may evolve from an subcritical to an overcritical state. However, criticality is only achieved for a short time then; and one may doubt that this short time is of any importance compared to the overcritical situation occurring immediately afterwards. From this point of view, criticality seems to be important for technical applications rather than for understanding natural phenomena unless there are at least some systems which evolve towards their critical state without any tuning.

5.2 The Bak-Tang-Wiesenfeld Model

In 1987, Per Bak and his coworkers presented an incredibly simple model that evolves towards a critical state without any external tuning. This model is often called *Per Bak's sandpile*, *Bak-Tang-Wiesenfeld model* or BTW model. In the seminal paper (Bak et al. 1987), the BTW model was derived from a model for the dynamics of an array of coupled pendulums. Just a few months later, the same model was interpreted in terms of sandpile dynamics (Bak et al. 1988) by the same authors.

This way of developing models seems to be somewhat strange, at least an earth scientist's point of view, and perhaps from the view of many other natural sciences, too. When developing models, we are used to start from a phenomenon or from a class of phenomena. Then, the model is developed in order to capture as much as possible of the phenomenon. In the BTW model, it seems to be just the other way round; a nice model was developed, and then people started looking for what the model could be used for.

Clearly, it was not like this, but even if it was, new results would justify the approach. Per Bak and his coworkers had a certain phenomenon in mind when they developed their model. But in contrast to those who work on just one, rather narrow topic their whole lives, is was a more general question from physics. Let us not worry whether it was the question how some systems manage to become critical without any tuning or whether it was the search for the origin of pink noise; it must have been something on this level and neither pendulums nor sandpiles. The latter are just interpretations of this fundamental model, perhaps developed for selling the model. So let us begin with discussing the BTW model without any serious process-based background and come back to its relationship to physical processes later.

The BTW model is discrete concerning space and time; it is defined on a quadratic, two-dimensional lattice. Let us number the sites with a pair of indices (i, j). The state of each site is characterized by a non-negative integer

variable $u_{i,j}$. In every discrete time step, a site (i,j) is randomly chosen, and $u_{i,j}$ is incremented:

$$u_{i,j} \;:=\; u_{i,j} + 1.$$

In the models discussed in this book, many assignments look like this example; the value of a variable is increased or reduced by a certain amount. So let us for convenience introduce the symbols $+\!\!=$ for increasing a value and $-\!\!=$ for reducing a value in analogy to the combined operators in some programming languages. In this short notation, the rule for incrementing $u_{i,j}$ reads:

$$u_{i,j} \;+\!\!=\; 1. \tag{5.1}$$

The BTW model assumes that nothing happens as long as $u_{i,j} < 4$, so that we proceed with the next time step then. If, on the other hand, $u_{i,j} \geq 4$, the site (i,j) becomes unstable and relaxes according to the rule

$$u_{i\pm1,j} \;+\!\!=\; 1, \quad u_{i,j\pm1} \;+\!\!=\; 1, \quad \text{and} \quad u_{i,j} \;-\!\!=\; 4. \tag{5.2}$$

In other words, an amount of four units of the model variable is uniformly transferred from the unstable site to its four nearest neighbors. This relaxation may cause some of the neighbors to become unstable, which may result in some kind of avalanches. Since up to four neighbors may become unstable, we must include the order of unstable sites being relaxed into the model rules. The straightforward order is relaxing all sites which have become unstable simultaneously, and then relax those adjacent sites which have become unstable (this may be even more than four), and so on. Figure 5.2 shows a flow chart of a *cellular automaton* based on this order of relaxations.

At this point one may wonder what a cellular automaton is. Roughly speaking, a cellular automaton is a system which is discrete concerning space and time, and whose evolution through time is defined by some rules. Since these rules seem to be a little arbitrary in some cellular automata, scientists are often more suspicious towards results obtained from cellular automata than they are towards results of apparently well-established differential equations. However, let us not go into detail here; a more thorough discussion on the relation between differential equations and cellular automata is given in Sect. 7.4. So let us for the moment be satisfied with the statement that cellular automata are discontinuous concerning space and time and thus easier to implement and numerically less demanding than differential equations.

Figure 5.3 shows an example of an avalanche in the BTW model. The values $u_{i,j}$ are represented by a number of dots at each site; unstable sites are marked with grey. During the avalanche, 28 sites became unstable; and it took 11 relaxation cycles until all sites became stable again. A total of 27 sites participated in the avalanche; so one cell became unstable twice.

The BTW model was originally motivated with the help of an array of coupled pendulums (Bak et al. 1987). Shortly after, the model was interpreted in terms of sandpile dynamics (Bak et al. 1988), too. However, some caution

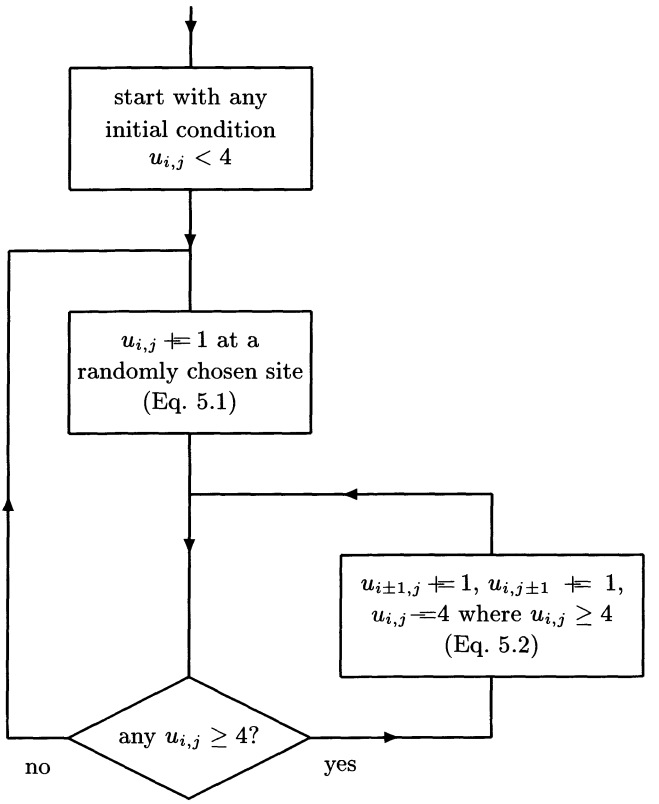

Fig. 5.2. Flow chart of the BTW model.

is necessary here because there are two different sandpile interpretations of the BTW model. The original one can be derived from simplified sandpile mechanics (Sect. 5.5), but is somewhat cumbersome and not consistent in detail. Therefore, a much simpler sandpile analogy is often referred to (e. g. Bak 1996): The integer variable $u_{i,j}$ corresponds directly to a number of grains which are neatly stacked on top of each other at the site (i, j). Then, the driving rule (Eq. 5.1) corresponds to adding a grain to a randomly chosen site. According to the criterion of stability, $u_{i,j} < 4$, a stack of grains remains stable as it consists of no more than three grains. If it becomes unstable, four grains are uniformly redistributed among the four nearest neighbors (Eq. 5.2).

It should be clear that this kind of sandpile model only works with "theoretical physicist's sand" (Bak 1996). The problem is not just the shape of the grains; perhaps we can try it with some building bricks. The crucial point is the criterion of stability: Why should a stack become unstable if its height exceeds a critical value, independently from the heights of adjacent stacks? Nevertheless, this sandpile analogy provides at least a simple nomenclature; so let us simply speak of adding and redistributing grains.

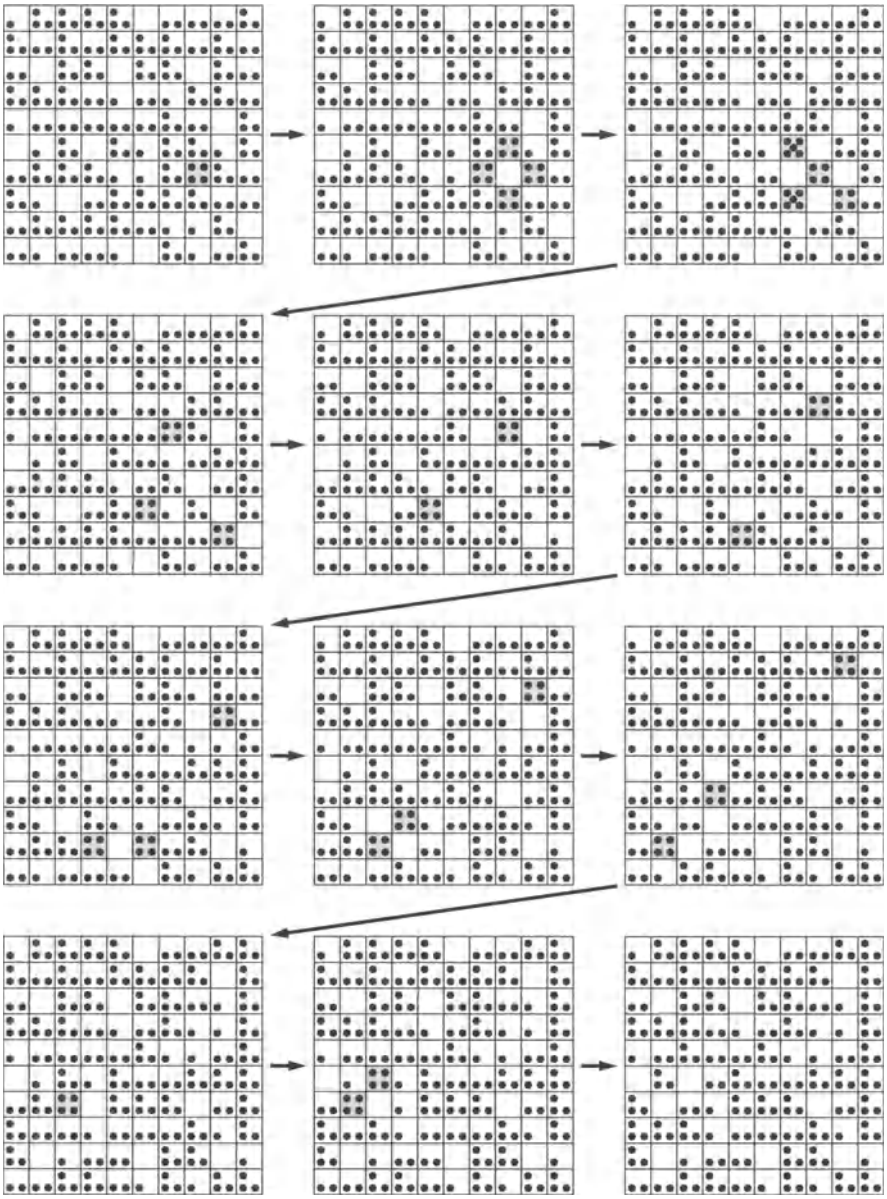

Fig. 5.3. Example of an avalanche in the BTW model. The dots represent the values $u_{i,j}$; unstable sites are marked with grey.

There is at least one physically consistent interpretation of the BTW model; it was developed by Peter Grassberger (e. g. Bak 1996). In this interpretation, the lattice corresponds to an office where each site is a desk with a bureaucrat. In each step, a sheet of paper representing some work to be done is put on a randomly chosen desk. However, in this rather unrealistic office nobody really works. As long as less than four sheets of paper are on a desk, there is no need to wake up. On the other hand, if there are four (or more) sheets, the bureaucrat may find that this is quite a lot of work and that it is better to redistribute it among his neighbors.

The question whether anyone works in the office or not leads us to a central property of the BTW model: The relaxation rule (Eq. 5.2) is *conservative.* This means that the sum of all values $u_{i,j}$ (the total number of grains or sheets of paper) remains constant during an avalanche. It turned out that conservation, respectively, dissipation during the avalanches is a fundamental property for distinguishing apparently similar models. The difference was discovered first by Feder and Feder (1991) who replaced the part $u_{i,j} -= 4$ of relaxation rule (Eq. 5.2) with $u_{i,j} := 0$ as a result of a misprint. Both versions only differ if $u_{i,j} > 4$ temporarily. The third stage shown in Fig. 5.3 shows that this case may occur during an avalanche. Then, grains are lost; in this example, one grain is lost during the avalanche. Thus, the modified version is not completely conservative. The fact that the results of Feder and Feder (1991) differed slightly from those of the original BTW model showed that conservation is a crucial point; and non-conservative models have become the key to understanding the dynamics of earthquakes (Chap. 7).

Up to now we have disregarded the fact that the BTW model is defined on a finite grid. In order to complete the model, we must introduce appropriate boundary conditions. In the BTW model, different sites only interact through the relaxation rule (Eq. 5.2), while both the driving rule (Eq. 5.1) and the stability criterion are local. Thus, boundary conditions shall only concern the relaxation rule, while the other rules are the same in the bulk and at the boundaries. In the original BTW model, the relaxation rule is transferred to boundary sites by simply omitting those parts of the rule which concern sites outside the domain. In other words, grains or sheets of paper passing the boundary are simply lost; this is called *open boundary condition.* Obviously, conservation is lost at the boundaries then; but assuming such a dissipation at the boundaries is necessary here since the driving rule (Eq. 5.1) is non-conservative. In each time step, the sum of all values $u_{i,j}$ is incremented; together with an entirely conservative relaxation rule this would lead to a permanent increase in the sum of all values $u_{i,j}$. We would finally end up at an everlasting avalanche because an avalanche can only cease if all values $u_{i,j}$ are smaller than four, but this is not what we are looking for.

Despite its simplicity, the BTW model has not yet been solved analytically. Analytical approximations have only been found under further simplifications which affect the results considerably (e. g. Jensen 1998). Even

the one-dimensional version has not been fully solved, although an enormous effort was spent (Chabra et al. 1993). So it seems that the fundamental properties of the BTW model can only be obtained from computer simulations.

Following the flow chart (Fig. 5.2), programming the BTW model is quite easy. Only finding out which cells are actually unstable requires some more thoughts. If all sites are checked in each relaxation cycle, even a small avalanche requires a numerical effort that is proportional to the total number of sites which is mostly be much larger than the number of sites participating in the avalanche. Thus, one should make use of the property that only the neighbors of an actually unstable site can become unstable in the next step by using one of the two following strategies:

1. Keeping track of the unstable sites with the help of a list. Strictly speaking, two lists are necessary: one for the sites which are unstable now, and another for those which will become unstable in the next step.
2. A recursive implementation of the relaxation. This means that there is a function that relaxes a site according to Eq. 5.2 and then invokes itself up to four times in order to relax those of the neighbors which have become unstable.

Writing a code for the second version is easier, but may consume a little more CPU time and memory. Still more important, the second version does not comply with the rule of simultaneous relaxation, but leads to a quite strange order: If, e. g., a single instability causes two neighbors to become unstable, all the instabilities resulting from that of the first neighbor are considered first, before the second neighbor is relaxed.

However, we may perhaps not be interested in simulating avalanches in detail, but only in the result of the avalanche after all sites have become stable again. In this case we can take advantage of the property that the BTW model is an *Abelian* model, which means that the result of an avalanche does finally not depend on the order of performing the relaxations. Let us refrain from giving a formal proof of this result; the reader may easily recognize that it is correct. So, if we are only interested in the statistics of the avalanche sizes, both versions of the model are equivalent. Then, the recursive strategy can be recommended for the BTW model if a simple realization is the goal, but one should keep in mind that the majority of all cellular automata is non-Abelian.

Figure 5.4 shows the number of grains that are present on the lattice versus the total number of supplied grains, starting at an empty grid of 128×128 cells. In the beginning, only few relaxations take place because the number of grains per site is small. Thus, only few grains are lost through the boundary, which results in a linear increase of the curve with unity slope. After about 35,000 grains have been added, a long-term equilibrium between the number of supplied grains and the number of those being lost through the boundaries has been achieved. In this state, the average number of grains per site is about 2.1. The lower part of Fig. 5.4 shows that this state is not a steady state in the strict sense; the number of grains is not constant but

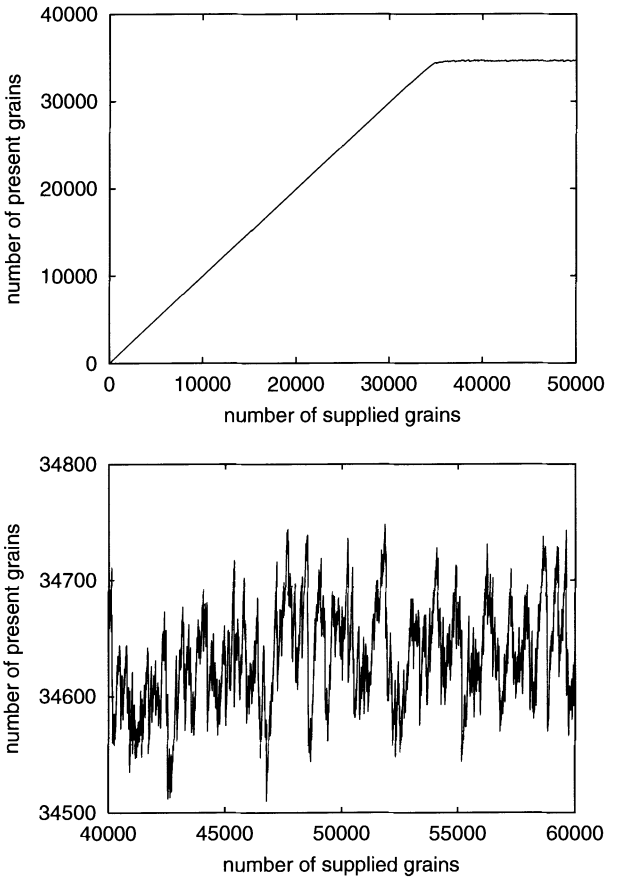

Fig. 5.4. Number of grains on the lattice versus the total number of grains supplied on a 128×128 grid.

fluctuates in an irregular way. This indicates the occurrence of avalanches of various sizes: While individual grains are added, a large number of grains (some hundreds) can be lost during one avalanche. In terms of non-linear dynamics, this quasi-steady state is a strange attractor in phase space.

Figure 5.5 shows the distribution of the avalanche sizes in the BTW model, computed on a 64×64 and on a 128×128 grid. Let us measure the size of an avalanche in terms of the number of sites participating in the avalanche. This property is called *cluster size*; it can be seen as a measure of the area affected by the avalanche. When measuring event sizes in the BTW model or in similar models, one should take care which quantity is considered. Terms like avalanche size are non-unique. We could, e. g., also analyze the number of relaxations taking place during an avalanche; both quantities may follow different distributions due to the occurrence of multiple relaxations.

The statistics shown in Fig. 5.5 include 10^6 avalanches each; the first 10^5 avalanches were skipped in order to avoid effects of the initial state. The cumulative distribution does not look like a power law at all. Surprisingly, the

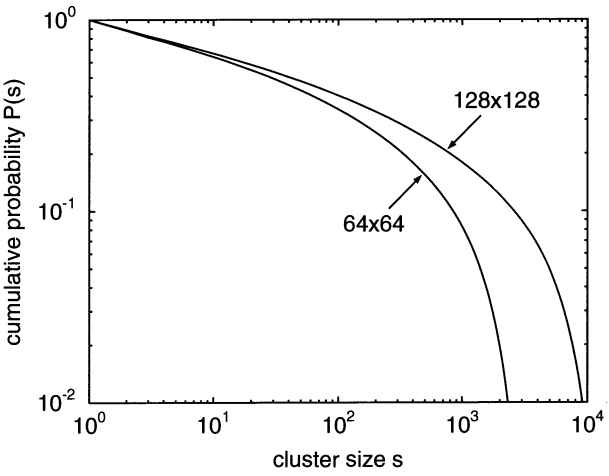

Fig. 5.5. Cumulative distribution of the cluster sizes in the BTW model on two different grids. The statistics include 10^6 avalanches each; the first 10^5 avalanches were skipped in order to avoid effects of the initial state.

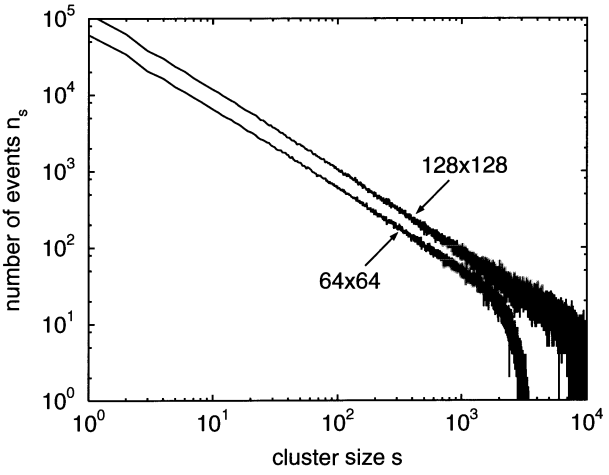

Fig. 5.6. Non-cumulative number of events corresponding to the cumulative cluster size distribution shown in Fig. 5.5. The curve belonging to the 64×64 grid has been shifted by a factor two downwards in order to make it distinguishable from that obtained from the 128×128 grid at small cluster sizes.

observed deviation from power-law behavior is an artificial effect of the finite model size, as the plot of the non-cumulative number of events in Fig. 5.6 illustrates. The non-cumulative number decreases like a power law up to a cutoff size which depends on the size of the lattice. The problem with the cumulative distribution is quite simple: The exponent of the probability density (estimated from the non-cumulative number of events) is about 1.05 and thus close to one. According to Eq. 2.5, this corresponds to a cumulative power-law distribution with a quite low exponent of $b \approx 0.05$. As discussed in Sect. 2.2, cutoff effects due to the finite system size become very strong if b is small and may thus shadow the power-law behavior.

5.3 The Critical State

In the previous section we have seen that the BTW model evolves towards a state where events of various sizes occur. The size distribution of these events follows a power law, so that this state is critical in the sense discussed in Sect. 5.1. However, the use of the term state in this context requires some explanation. Clearly, the BTW model does not approach a steady state in the strict sense; events in a system of this type must result in fluctuations in its state. Thus, the system's properties can only be roughly constant through time; the phase space of the system contains a preferred region which is approached through time. We have already observed this behavior when discussing the Lorenz equations in the previous chapter; and we have learned that such a preferred region is called strange attractor.

The phase space of the BTW model is spanned by all variables $u_{i,j}$; so its dimension is quite high. Consequently, the strange attractor of the BTW model cannot be visualized as easily as that of the Lorenz equations. Only a few statistical properties of the strange attractor can be specified; e. g., the sum of all variables $u_{i,j}$ shown in Fig. 5.4.

One may argue whether the difference between steady-state behavior in the strict sense and the existence of a strange attractor is of any relevance for us. At first sight, this might just be a matter of the definition what the state of a system is: If we say that a gas is in a certain state, we do normally not refer to individual atoms or molecules, but only to macroscopic properties such as density, pressure and temperature. Macroscopic properties are mostly derived from the microscopic properties by averaging. Since the number of microscopic components is very large in general, fluctuations on the microscopic level are hardly visible in the macroscopic variables. However, the critical state is an example where this argument cannot be applied. Here fluctuations on the microscopic level may affect a considerable part of the system or even the entire system, and then the macroscopic variables are exposed to fluctuations, too. In the BTW model, the sum of all values $u_{i,j}$ is such a macroscopic property, and we have seen in Fig. 5.4 that this property is in fact not constant through time. So let us in the following use the term *critical state* in the sense of a strange attractor with critical properties. Sometimes, this state is called *quasi-steady* or *quasi-stationary*, which means something like steady except for fluctuations. However, defining this property formally is difficult because the fluctuations are not limited in size.

A second point that should be clarified concerns the quality of the resulting power-law distribution. Everything we have learned about power-law distributions in the first two chapters applies here, too. Concerning finite-size effects, the situation in modeling is even more lucky than it is when data sets from nature are considered. Models of different sizes can be run, and the results should look like that of the BTW model shown in Fig. 5.5. So let us extend the definition of criticality given in Sect. 5.1 in the following way:

> A finite system is critical if the size distribution of fluctuations tends towards a power law in the limit of infinite system size.

However, this criterion should still be used with caution if applied to the results of computer simulations. We cannot investigate the limit of infinite model size then, but can only compute distributions on grids of different, but still finite sizes. Thus, we are free to believe in an apparent convergence towards a power law or not. Figure 5.5 makes it easy to believe in asymptotic power-law behavior, but reveals some minor deviations at small event sizes. They can be attributed to the discrete model structure. However, we have already learned in Sect. 2.4 that the breakdown of scale invariance at small object sizes is not necessarily a sharp cutoff as expressed by a Pareto distribution, but may be superposed by various effects. There is at least no reason why the lower limit of scale invariance should coincide with the size of the smallest unit in the model (one cell), so that deviations from power-law behavior at small object sizes are tolerable, provided that there is a reasonable range where a power law can be recognized.

This discussion shows that there cannot be an absolute criterion for scale invariance in modeling; but this is not surprising since scale invariance is not absolute in nature, too. In general, we will not obtain perfect power laws even within a limited range. As discussed in Sect. 1.3, the quality of a power law is determined by both its range of validity and its "cleanness". As a consequence, the definition of a critical state, and thus any definition of SOC, must be soft with respect to the quality of power laws; and there is not an absolute criterion to what extent deviations from scale invariance are tolerable.

5.4 What is SOC?

After having discussed the properties of the critical state in detail, we can now give a preliminary definition of SOC:

> A system exhibits SOC if its phase space contains a strange attractor where events of all sizes occur, and where the size distribution of these events follows a power law.

When using this definition we should keep in mind all the limitations and ambiguities discussed in the previous section, especially that a power-law distribution may only be approached in the limit of infinite system size, and that there is not an absolute criterion whether an observed distribution is an acceptable power law or not.

Although this definition is widely used, it captures only half of the original definition suggested by Bak et al. (1987). As already mentioned in the introduction of this chapter, pink noise may be the temporal fingerprint of

complexity. So let us now go one step ahead and include the occurrence of pink noise into a complete definition of SOC:

> A system exhibits SOC if its phase space contains a strange attractor with the following scale-invariant properties:
>
> 1. The attractor is critical; fluctuations (events) of various sizes occur, and the distribution of the event sizes follows a power law or tends towards a power law in the limit of infinite system size.
> 2. The temporal signal of the system is pink noise or tends at least towards pink noise in the limit of infinite system size.

Let us now find out whether the BTW model meets the criteria of SOC since the occurrence of pink noise is still unclear, although the fluctuations of various sizes shown in Fig. 5.4 are promising.

But what is time in the BTW automaton? The evolution of the system takes place in discrete steps without any relation to time in the physical sense. We can, e. g., assume that each cycle of relaxations, i. e., each execution of the inner loop in the flow chart (Fig. 5.2) takes place within a time interval of given length, and that a new grain is added shortly (again, by a given time interval) after the avalanche has ceased. On the other hand, assuming that there is no correlation between the end of an avalanche and supplying a new grain is more reasonable. This requires a *separation* of the time scales of the avalanches and of driving; otherwise avalanches may overlap in contradiction to the rules. We must therefore assume that there is a short time scale which describes the dynamics of individual avalanches, and a long time scale which describes the system's behavior over many avalanches. Viewed from the long time scale, each avalanche takes place and ceases instantaneously, and time can be identified with the number of supplied grains.

The temporal fingerprint of the system may strongly depend on the considered time scale. On the short time scale, the lifetimes of avalanches, i. e., the number of relaxation cycles needed until the avalanche ceases, can be measured (Jensen et al. 1989; Jensen 1998). One can imagine that this property may follow a power-law distribution which is closely related to the distribution of the (spatial) event sizes. This may result in spatio-temporal scaling properties on the short time scale, but this kind of scale invariance cannot be related to pink noise or FBM in general. However, the behavior on the long time scale may be completely different since it is determined by a sequence of many avalanches rather than by properties of individual events.

But even after deciding for the long time scale, some freedom in defining the temporal signal of a system remains. Let us here take the number of grains being lost through the boundaries during the recent avalanche. However, we cannot expect that this signal is pink noise in the strict sense. Since the avalanche sizes are power-law distributed, the number of lost grains will not be Gaussian-distributed. Thus, we should not be too strict with the

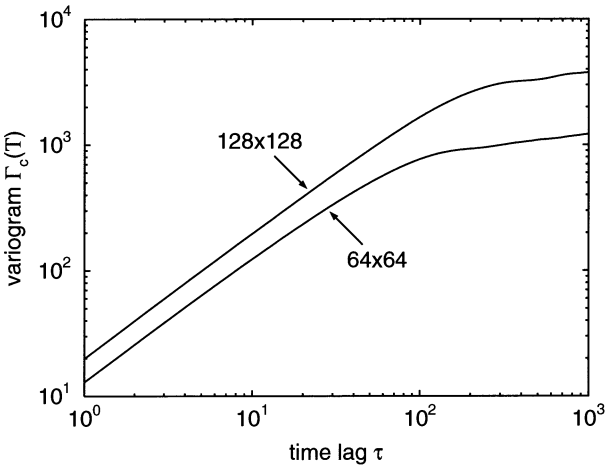

Fig. 5.7. Variogram analysis of the BTW model (cumulated signal).

definition of pink noise, but allow non-Gaussian signals with appropriate scaling properties, too.

As discussed in Sections 3.3 and 3.9, distinguishing pink noise with the help of a variogram analysis is often preferable to computing the spectral exponent directly from a Fourier transform. Depending on the system's character, either the variogram of the original or that of the cumulated signal must be considered. If the number of grains being lost through the boundaries is the original signal, the cumulated signal is the total number of grains that are actually present on the lattice (except for the sign).

Figure 5.7 gives the variogram analysis of the cumulated signal, computed according to Eq. 3.17. Except for a rollout at time lags, the variogram shows power-law behavior. Comparing the curves for the 64×64 and 128×128 grids suggests that the rollout at large time lags is an effect of the finite model size. The variogram increases linearly with the time lag, so the cumulated signal behaves like Brownian motion. Thus, the original signal is not pink noise as requested in the definition of SOC, but rather white noise. In other words, avalanches seem to be completely uncorrelated in time.

This little flaw in the first paper on SOC (Bak et al. 1987) was soon discovered (Jensen et al. 1989). Interestingly, the title of the paper focused on pink noise rather than on fractal size distributions. Apart form this, the framework of SOC is built on sand unless there is at least one system which meets the criteria of SOC. However, if the BTW model is driven in a specific way from an edge, certain properties turn out to have 1/f spectra (Jensen 1991, 1998). So this little accident can be fixed, but the model differs from the original BTW model then. Still more important, we see that some properties of the system may be pink noise, while others are not. The idea of SOC does not provide any rules which properties are to be chosen; this may be seen as an inherent weakness of the concept in its present state.

Under these aspects, thinking a little about the role of pink noise in the definition of SOC is not wasted time. There is no doubt that explaining both fractal size distributions and pink noise with just one concept is a great deal. However, a bird in the hand is worth two in the bush. So, is the original BTW model of minor value just because it meets the criteria of SOC only partly? Obviously, introducing more criteria into a definition sharpens the concept, but reduces its applicability since the number of systems which meet the criteria decreases.

Several authors disregard the criterion concerning pink noise completely and focus on fractal size distributions instead. Although not correct with respect to the original spirit of SOC, this is not necessarily wrong. Perhaps we should define three different levels of SOC:

- The lowest level should at least include a strange attractor with critical properties, regardless of the system's temporal fingerprint. However, a discussion concerning landform evolution models (Sapozhnikov and Foufoula-Georgiou 1996a) showed that even models which result in steady states with static fractal properties were claimed to exhibit SOC, although they do not even reach this lowest level.
- Obviously, there is an asymmetry concerning the spatial and temporal characteristics in the original definition of SOC. While a scale-invariant distribution of spatial event sizes with an arbitrary exponent (fractal dimension) is sufficient, the spectral exponent of the temporal signal is confined to unity. Thus, an intermediate level of SOC could require a self-affine scaling behavior in time with an arbitrary spectral exponent. The original BTW model falls into this category.
- Finally, the original criteria could define the highest level of SOC.

5.5 Sandpile Dynamics and the BTW Model

Let us now return to the sandpile interpretation of the BTW model. We begin with a simplified, one-dimensional model as illustrated in Fig. 5.8. Two

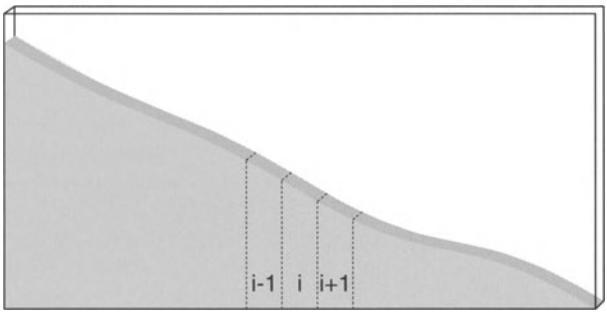

Fig. 5.8. Illustration of a sandpile between two plates.

vertical plates are aligned parallel to each other; their distance is not much larger than the diameter of a sand grain. Let us assume that there is a third plate that keeps the left-hand boundary closed, while that at the opposite side is open. Let us now fill some sand between the plates.

One may wonder why we call this a one-dimensional sandpile and argue that it is rather a two-dimensional one. However, activity in sandpiles is mainly confined to a thin layer at the surface. Thus, the evolution of this two-dimensional sandpile is determined by a moving-boundary problem, and the evolution of the boundary can be reduced to a one-dimensional problem.

In order to develop a one-dimensional formalism, we divide the sandpile into N cells, consecutively numbered with an index i (Fig. 5.8). In each cell there is a number n_i of sand grains. The sandpile remains stable as long as the local slope

$$\Delta_i \;:=\; \max\{n_i - n_{i-1}, n_i - n_{i+1}\}$$

is below the critical slope Δ_c that corresponds to the angle of repose of sand. If the sites are larger than the area covered by a grain, there should be an additional factor in the definition of the slope, but this can also be incorporated by an appropriate choice of Δ_c. Since the left-hand boundary is closed and the right-hand boundary is open, the surface height will mostly decrease towards the right-hand side. Hence $n_i - n_{i-1}$ will in general not reach Δ_c, so that we use a simplified definition of the slope:

$$\Delta_i \;:=\; n_i - n_{i+1}.$$

As soon as the slope Δ_i at any site reaches Δ_c, this site becomes unstable. Let us assume that one grain topples from the site i to the site $i+1$:

$$n_i \;-\!=\; 1 \quad \text{and} \quad n_{i+1} \;+\!=\; 1.$$

As a result of this relaxation, the slopes at the sites $i-1$, i, and $i+1$ have changed in the following way:

$$\Delta_i \;-\!=\; 2 \quad \text{and} \quad \Delta_{i\pm1} \;+\!=\; 1.$$

This relaxation rule is exactly the one-dimensional version of that of the BTW model (Eq. 5.2). However, a one-dimensional version of the BTW model's stability criterion should be something like $u_i < 2$, while we have $\Delta_i < \Delta_c$ here. This little discrepancy can be fixed by introducing the variables

$$u_i \;:=\; \Delta_i - \Delta_c + 2 \;=\; n_i - n_{i+1} - \Delta^c + 2.$$

Then the stability criterion in fact turns into $u_i < 2$, while the relaxation rule remains the same:

$$u_i \;-\!=\; 2 \quad \text{and} \quad u_{i\pm1} \;+\!=\; 1. \tag{5.3}$$

However, deriving the driving rule of the BTW model (Eq. 5.1) from this kind of sandpile dynamics is not straightforward. Adding a grain to a randomly chosen site corresponds to $n_i \mathrel{+}= 1$ and thus leads to

$$u_i \mathrel{+}= 1 \quad \text{and} \quad u_{i-1} \mathrel{-}= 1. \tag{5.4}$$

In contrast to the driving rule of the BTW model, this rule is conservative. As a result, the sum of all values u_i remains constant through time, except for possible boundary effects.

Let us now take a look at the boundary conditions of the sandpile model. Since u_i only depends on n_i and n_{i+1}, but not on n_{i-1}, the boundary condition at the left-hand edge ($i = 1$) is trivial. Both the driving rule (Eq. 5.4) and the relaxation rule (Eq. 5.3) can be applied without any changes; only those parts of the rules which concern u_0 must be omitted. In contrast, further assumptions are needed to specify the behavior at the right-hand boundary. We have already assumed that grains may leave the model domain here, but this is not sufficient because we cannot determine the gradient Δ_i and thus u_N from this. Let us assume that all grains which leave the model domain are immediately removed, as it happens if we place the sandpile at the edge of a table. For this, we assume that the condition $n_{N+1} = 0$ persists. The driving rule (Eq. 5.4) remains the same, while the relaxation rule (Eq. 5.3) is slightly modified:

$$u_N \mathrel{-}= 1 \quad \text{and} \quad u_{N-1} \mathrel{+}= 1. \tag{5.5}$$

Obviously, this boundary condition differs slightly from that we would expect in a one-dimensional version of the BTW model which should be

$$u_N \mathrel{-}= 2 \quad \text{and} \quad u_{N-1} \mathrel{+}= 1. \tag{5.6}$$

At this point it is noteworthy that terms such as *open* and *closed* boundary conditions are often ambiguous. As already mentioned, grains may leave to model domain at the right-hand boundary, so it is open. On the other hand, the relaxation rule at the right-hand boundary (Eq. 5.5) is still conservative. This means that not any amount of the variable u_i passes this boundary, so it is a closed boundary with respect to the abstract property u_i. At the left-hand boundary, things are just opposite. Particles cannot pass this boundary, so it is closed. On the other hand, the relaxation rule is non-conservative, so this boundary is open with respect to the property u_i.

Unfortunately, this one-dimensional sandpile model does not achieve the complexity of the BTW model. In contrast to the latter, the fundamental properties of this model can be examined without numerical simulations. There is exactly one configuration which is stable under all conditions, namely that where all sites are just below the angle of repose ($u_i = 1$ everywhere). No matter where a grain is added, it simply topples down the sandpile without any further effect until it leaves the model domain. In other words, each

avalanche involves exactly one grain. The reader may easily proof that any initial condition will finally lead to this state. Therefore, the phase space of this sandpile model contains an attractor in the classical sense; all fluctuations in the system will cease through time.

But why do avalanches on real sandpiles often involve several or even many grains? In contrast to our idealized grains, real grains picks up kinetic energy while toppling downslope. If this energy is sufficiently high, further grains may be destabilized and topple downslope, too. This results in a propagation of avalanches in downward direction. However, everyone who has ever played with sand knows that sandpile avalanches proceed in upslope direction, too. This phenomenon can easily be understood: If a grain has been destabilized by the impact of a toppling grain, it is no longer able to stabilize its upslope neighbor, so that this grain may become unstable, too.

This discussion shows that a realistic sandpile model must include effects of *inertia*; otherwise it is not possible to determine whether a grain is able to destabilize further grains or not. Unfortunately, effects of inertia make sandpile dynamics complicated. Several experiments on sandpile dynamics were performed, detailed discussions are, e. g., given by Bak (1996) and Jensen (1998). As a major result of these studies, small sandpiles roughly show SOC behavior. If, in contrast, the sandpile becomes so large that effects of inertia become important, the behavior changes drastically: Instead of irregularly occurring avalanches of various sizes, large avalanches dominate, and they occur quite regularly. For normal beach sand, the transition to this nearly periodic behavior takes place at sandpile diameters of a few centimeters, which makes observation difficult. Frette et al. (1995) conducted experiments with rice instead of sand; by choosing grains of different shapes they were able to recognize SOC behavior even in quite large ricepiles. A simple model approach regarding effects of inertia was proposed by Bouchaud et al. (1995). But let us refrain from going into details because this model leads us further away from the BTW model. At this point we must accept that the BTW model can only be derived from a perhaps oversimplified sandpile analogy and keep in mind that it was not our aim to develop a realistic model of sandpile dynamics.

In order to derive a one-dimensional version of the BTW model we must change the driving rule (Eq. 5.4) and the right-hand boundary condition. The driving rule should be $u_i \mathrel{+}= 1$; this rule can still interpreted in terms of adding grains to the sandpile. However, if we add a grain to a randomly chosen site i, but want u_{i-1} to remain constant, we must add a grain at the site $i-1$, too. Then we must add a grain at the site $i-2$ in order to keep u_{i-2}, constant, and so on. In summary, we choose a site i randomly, and then add one grain each to the site i and to all sites which are left of this site. So the driving rule of the BTW model can be interpreted in terms of a sandpile model at least in the one-dimensional case, although the reader may find it somewhat strange.

In contrast to the original driving rule, the distribution of supplied grains is spatially inhomogeneous then; the number of supplied grains decreases from the left to the right-hand boundary. This inhomogeneous distribution even allows closed boundaries (with respect to the grains) at both edges because it maintains the slope. In contrast, the uniform distribution arising from the original driving rule (Eq. 5.4) in combination with a closed boundary finally results in a flat sandpile, and avalanching would cease.

The original boundary relaxation rule of the BTW model (Eq. 5.6) can be derived from a closed boundary condition. For this it is convenient to add an additional site $N+1$ and assume that this site never becomes unstable and that no grains are added here. So there is no need to keep track of the variable u_{N+1}, and the relaxation rule of the site N turns into Eq. 5.6.

When proceeding towards the original, two-dimensional BTW model, things become still more complicated. Let us divide the ground area into square tiles which are numbered with two indices i and j in such a way that increasing indices correspond to the downslope direction, i. e.,

$$n_{i,j} \geq n_{i+1,j} \quad \text{and} \quad n_{i,j} \geq n_{i,j+1},$$

where $n_{i,j}$ is the number of grains in the cell (i,j). Then the slope is

$$\Delta_{i,j} = \sqrt{(n_{i,j} - n_{i+1,j})^2 + (n_{i,j} - n_{i,j+1})^2}.$$

Since this expression is somewhat cumbersome with respect to driving and relaxation rules, it is replaced by a linear approximation. Let us characterize the slope direction by an angle $\alpha_{i,j}$ between 0 and $\frac{\pi}{2}$ according to

$$n_{i,j} - n_{i+1,j} = \Delta_{i,j} \cos\alpha_{i,j} \quad \text{and} \quad n_{i,j} - n_{i,j+1} = \Delta_{i,j} \sin\alpha_{i,j}.$$

In a linear approximation, we obtain the following rule for changing $\Delta_{i,j}$ if grains are added:

$$\Delta_{i,j} \mathrel{+}= \begin{cases} \frac{\partial}{\partial n_{i,j}} \Delta_{i,j} = \cos\alpha_{i,j} + \sin\alpha_{i,j} & n_{i,j} \mathrel{+}= 1 \\ \frac{\partial}{\partial n_{i+1,j}} \Delta_{i,j} = -\cos\alpha_{i,j} & \text{if} \quad n_{i+1,j} \mathrel{+}= 1 \\ \frac{\partial}{\partial n_{i,j+1}} \Delta_{i,j} = -\sin\alpha_{i,j} & n_{i,j+1} \mathrel{+}= 1 \end{cases}. \tag{5.7}$$

As soon as $\Delta_{i,j}$ reaches the critical slope Δ_c, the site becomes unstable. In contrast to the one-dimensional model, the destination of a toppling grain is non-unique; we can at least choose between the two nearest downslope neighbors $(i+1,j)$ and $(i,j+1)$. It makes sense to assume that the direction of toppling depends on the direction of the local slope; we can perhaps assume that an unstable grain topples towards the lower one of the sites $(i+1,j)$ and $(i,j+1)$. Obviously, the relaxation rule of the BTW model (Eq. 5.2) does not include such a decision. So let us assume that a toppling grain does not decide for either of the two lower neighbors, but is redistributed among them

according to the direction of the slope gradient. A fraction proportional to $\cos\alpha_{i,j}$ arrives at the site $(i+1,j)$, while a fraction proportional to $\sin\alpha_{i,j}$ arrives at the site $(i,j+1)$. If we do not worry about fractions of grains, this redistribution leads to the relaxation rule

$$n_{i,j} \mathrel{-}= 1, \quad n_{i+1,j} \mathrel{+}= \frac{\cos\alpha_{i,j}}{\cos\alpha_{i,j}+\sin\alpha_{i,j}}, \quad \text{and} \quad n_{i,j+1} \mathrel{+}= \frac{\sin\alpha_{i,j}}{\cos\alpha_{i,j}+\sin\alpha_{i,j}}.$$

Combining this rule with Eq. 5.7, written for the sites (i,j), $(i\pm1,j)$, $(i,j\pm1)$, and $(i\pm1,j\mp1)$ yields a relaxation rule for the slopes. However, this rule is still somewhat cumbersome as it involves the angles $\alpha_{i,j}$ of different sites simultaneously. We therefore assume that the angles $\alpha_{i,j}$ are the same for all sites involved in the relaxation. This leads to the following relaxation rules:

$$\begin{aligned} \Delta_{i,j} &\mathrel{-}= \frac{1+(\cos\alpha+\sin\alpha)^2}{\cos\alpha+\sin\alpha}, & \Delta_{i\pm1,j\mp1} &\mathrel{-}= \frac{\cos\alpha\,\sin\alpha}{\cos\alpha+\sin\alpha}, \\ \Delta_{i\pm1,j} &\mathrel{+}= \cos\alpha, & \Delta_{i,j\pm1} &\mathrel{+}= \sin\alpha. \end{aligned}$$

In contrast to the relaxation rule of the BTW model, this rule concerns six neighbors; the relaxation of the site (i,j) destabilizes its four nearest neighbors; but in addition, it stabilizes two of its diagonal neighbors.

Obviously, the relaxation rule for the slopes strongly depends on the orientation of the lattice in relation to the main slope direction, i. e., on the angle α. One may hope that this artificial effect does not affect the results strongly, but this is not the case: If the grid is aligned parallel to the main slope direction, only two neighbors are involved in each relaxation. The sandpile is decomposed into a series of independent, one-dimensional sandpiles then; and this is certainly unrealistic. For other grid orientations, the model retains its two-dimensional structure. Unfortunately, the relaxation rule cannot be transformed into that of the BTW model (Eq. 5.2) by simply choosing an appropriate angle α and a suitable definition of $u_{i,j}$. However, let us assume $\alpha = \frac{\pi}{4}$, leading to the relaxation rule

$$\Delta_{i,j} \mathrel{-}= \tfrac{3}{\sqrt{2}}, \quad \Delta_{i\pm1,j\mp1} \mathrel{-}= \tfrac{1}{2\sqrt{2}}, \quad \Delta_{i\pm1,j} \mathrel{+}= \tfrac{1}{\sqrt{2}}, \quad \Delta_{i,j\pm1} \mathrel{+}= \tfrac{1}{\sqrt{2}}. \tag{5.8}$$

If we define

$$u_{i,j} := \sqrt{2}\,(\Delta_{i,j} - \Delta_c) + 4,$$

we obtain at least a rough coincidence with the BTW model. The criterion of stability, $u_{i,j} < 4$, exactly coincides with that of the BTW model, while the relaxation rule (Eq. 5.8) turns into

$$u_{i,j} \mathrel{-}= 3, \quad u_{i\pm1,j\mp1} \mathrel{-}= \tfrac{1}{2}, \quad u_{i\pm1,j} \mathrel{+}= 1, \quad \text{and} \quad u_{i,j\pm1} \mathrel{+}= 1. \tag{5.9}$$

At this point, our rather formal way towards the BTW model ends; we cannot proceed further by making straightforward assumptions and approximations. The relationship between the relaxation rule derived here and that of the original BTW model can only be understood by considering some fundamental properties of both relaxation rules which are the same:

- The relaxed point itself becomes more stable, while its neighbors are mainly destabilized.
- The relaxation rule is conservative, which means that the sum of all variables $u_{i,j}$ remains constant during the relaxation.
- The effect on the diagonal neighbors is smaller (respectively not present at all in the BTW model) than that on the nearest neighbors.

So, if we simply neglect the impact of the relaxation on the diagonal neighbors, but keep these fundamental properties, we should replace the rule $u_{i,j} \mathrel{-}= 3$ with $u_{i,j} \mathrel{-}= 4$. Then we have arrived at the relaxation rule of the BTW model. The reasoning can be strengthened by the argument that nature does not know anything about regular grid topology; so Eq. 5.9 is clearly affected by artificial grid effects. Under this aspect, the transition to the relaxation rule of the BTW model is not a severe limitation of the model, at least not worse than all the other approximations and simplifications introduced earlier. Nevertheless, we should keep in mind that the arguments have become quite soft in the end.

Concerning the driving rule and the boundary conditions, the problems are mainly the same as those encountered when discussing the one-dimensional case. Adding a grain to a randomly chosen site (i,j), i. e., $n_{i,j} \mathrel{+}= 1$, leads to

$$u_{i,j} \mathrel{+}= 2, \quad u_{i-1,j} \mathrel{-}= 1, \quad \text{and} \quad u_{i,j-1} \mathrel{-}= 1$$

if the linear approximation of the slope (Eq. 5.7) is used. This conservative rule is the straightforward generalization of the one-dimensional case, but the driving rule of the BTW model (Eq. 5.1) can only be obtained by adding several grains or even fractions of grains. Increasing $u_{i,j}$ by one corresponds to adding a fraction of a grain to the site (i,j). However, keeping the values $u_{i-1,j}$ and $u_{i,j-1}$ constant requires that fractions of grains are added to these sites, too. Then, keeping $u_{i-2,j}$, $u_{i-1,j-1}$, and $u_{i,j-2}$ constant requires that something is added at these sites, too, and so on. Eventually, parts of grains must be added to all sites which are located upslope from the originally driven cell. This sounds strange, but we will see in Chap. 8 that such a driving rule is not too bad at least for landslides which are driven by slowly tilting a slope.

So let us summarize the results of this section: Starting from a simplified model of sandpile dynamics, we have roughly derived the BTW model. However, we had to introduce some special assumptions concerning gradients in two dimensions, boundary conditions, and long-term driving. The boundaries must be closed, corresponding to a sandpile in a box, and grains must a added according to a certain, spatially inhomogeneous distribution. Under these aspects, we may now understand why the BTW model is often interpreted in terms of sandpile dynamics, but we should take care not to take this analogy too seriously.

6. The Forest-Fire Model – Tuning and Universality

We continue our introduction into the ideas of SOC with the forest-fire model and a discussion on tuning parameters and SOC systems in non-equilibrium. A first version of this model was published by Bak et al. (1990), but Drossel and Schwabl (1992) were the first to discover SOC uniquely in a modified approach. Although it is even simpler than the BTW model (Sect. 5.2), the forest-fire model has some relevance to nature as the sizes of real forest fires exhibit power-law statistics to some degree (Malamud et al. 1998).

6.1 The Forest-Fire Model

In analogy to the BTW model, the forest-fire model is defined on a quadratic lattice. We focus on the two-dimensional version, although the model can be generalized to arbitrary dimensions. The original forest-fire model (Bak et al. 1990) is a cellular automaton where each site can be either empty or occupied by a tree which may be green or burning. The model does not distinguish between small and large or young and old trees, although this may be important for the propagation of fires in reality. In each step, the lattice is updated according to the following rules:

1. A green tree catches fire if any of its four nearest neighbors is burning.
2. A burning tree turns into an empty site.
3. At an empty site, a green tree grows with a given probability p.

These rules are applied simultaneously to all sites.

The results presented in the original paper pointed towards SOC. At least under certain conditions, the model evolves towards a quasi-steady state where the number of burnt trees in each step is power-law distributed. However, it was soon recognized that this model suffers from some problems. In principle, it does not simulate sequences of fires, but only one fire which extends over the whole simulation and is kept alive by the re-growth of trees. Apart from the problem that the fire might die, this behavior is not very realistic with respect to forest fires in reality. Further simulations (Grassberger and Kantz 1991; Moßner et al. 1992) revealed that the original forest-fire model is not SOC, but exhibits a mainly regular behavior with spiral-shaped fire fronts.

Drossel and Schwabl (1992) proposed an extension of the forest-fire model which fixes the shortcoming that there is just one fire. However, their model was not new at this time; Henley (1989) published essentially the same approach in a smaller journal. It seems that it was the context of SOC which focused interest on the forest-fire model; Drossel and Schwabl were the first to investigate the model with respect to SOC. In addition to the original rules, the modification introduces a spontaneous ignition of green trees as it takes place in reality by lightning or by human impact with the following rule:

4. A green tree becomes a burning tree with a probability f even if none of its neighbors is burning.

The second modification introduces a separation of time scales. In analogy to real forest fires, burning down a cluster of trees takes place much more rapidly than raising new trees. Therefore, it is assumed that growth and spontaneous ignition stop as long as any trees are burning. Formally, this separation of time scales can be achieved by performing the limit $p \to 0$ and $f \to 0$, while the ratio $r := \frac{p}{f}$ is kept constant. In a computer model, the separation of time scales can be realized by burning down the whole cluster of trees connected to a tree that has been ignited within one model step. This leads to a modified set of rules for updating the lattice in each step:

1. Each tree is ignited with a probability f. The cluster of trees connected to an ignited tree is burnt down immediately, i. e., the corresponding sites become empty.
2. At each empty site, a green tree grows with a probability p.

These rules come quite close to the simplest ideas on the propagation of forest fires in reality, but the numerical realization is cumbersome. In each step, all sites must be checked for either the growth of a new tree or for ignition. Therefore, the numerical effort per step is proportional to the total number of sites, even if not any site is finally ignited. Therefore, a further modification was suggested (Grassberger 1993; Clar et al. 1994) in order to make the model feasible on large grids. This modification is based on the following interpretation: In each step, nf sparks are thrown in the mean, where n is the total number of sites. A fire occurs if a match hits a tree. In analogy, np sites are randomly selected for the growth of new trees in the mean; new trees grow where the selected cells are empty. The time span between two sparking events consists of $\frac{1}{nf}$ steps in the mean; to that $\frac{np}{nf} = r$ cells are selected for the growth of new trees between two sparking events in the mean. The modification consists of assuming that exactly r attempts are made to grow new trees between two sparking events; this removes a part of the model's randomness. So the rules of the model turn into:

1. A randomly chosen site is ignited. If it is occupied by a tree, this tree and all trees connected to it are immediately burnt down.
2. A total of r new trees is randomly placed on the grid. If a site is already occupied by a tree, the new tree is ignored.

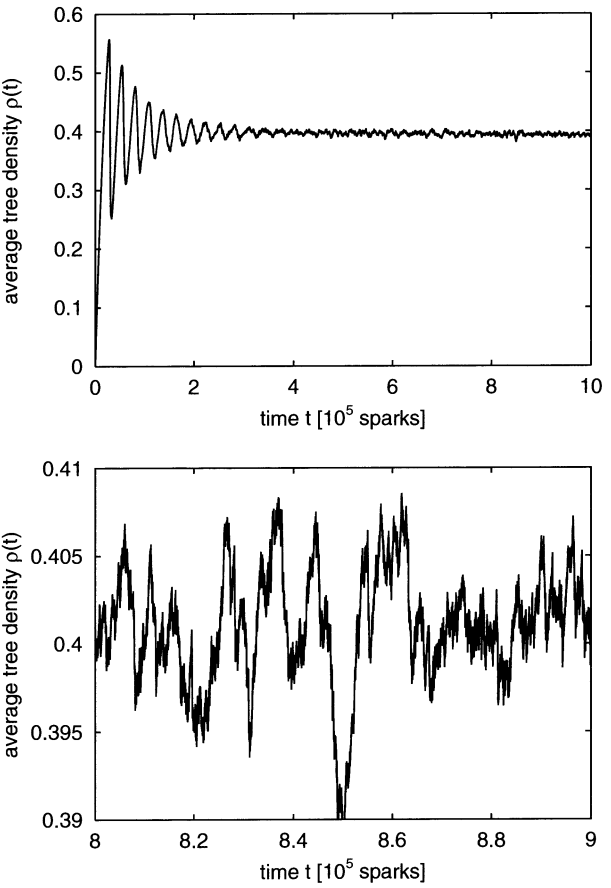

Fig. 6.1. Evolution of the mean tree density for $r = 2048$ on a 8192×8192 grid.

Although this modified forest-fire model is not entirely equivalent to that of Henley (1989), respectively, Drossel and Schwabl (1992), numerical simulations of the forest-fire model are mainly based on this set of rules.

Introducing appropriate boundary conditions is straightforward and needs no further explanation if we assume that the forest is bounded by roads, fire breaks, water or other areas that inhibit the propagation of fires. Alternatively, periodic boundary conditions can be posed.

Figure 6.1 shows the average tree density $\rho(t)$ (number of present trees divided by the total number of sites) as a function of time, which is identified with the number of sparks. The growth rate is $r = 2048$, and the simulation started with an empty grid of 8192×8192 sites. In the beginning, the tree density oscillates roughly periodically with a period length of several thousand sparking events. But after some time, the regular oscillations cease, and irregular fluctuations become visible. As the BTW model did, the forest-fire model evolves towards a quasi-steady state; although this state is approached

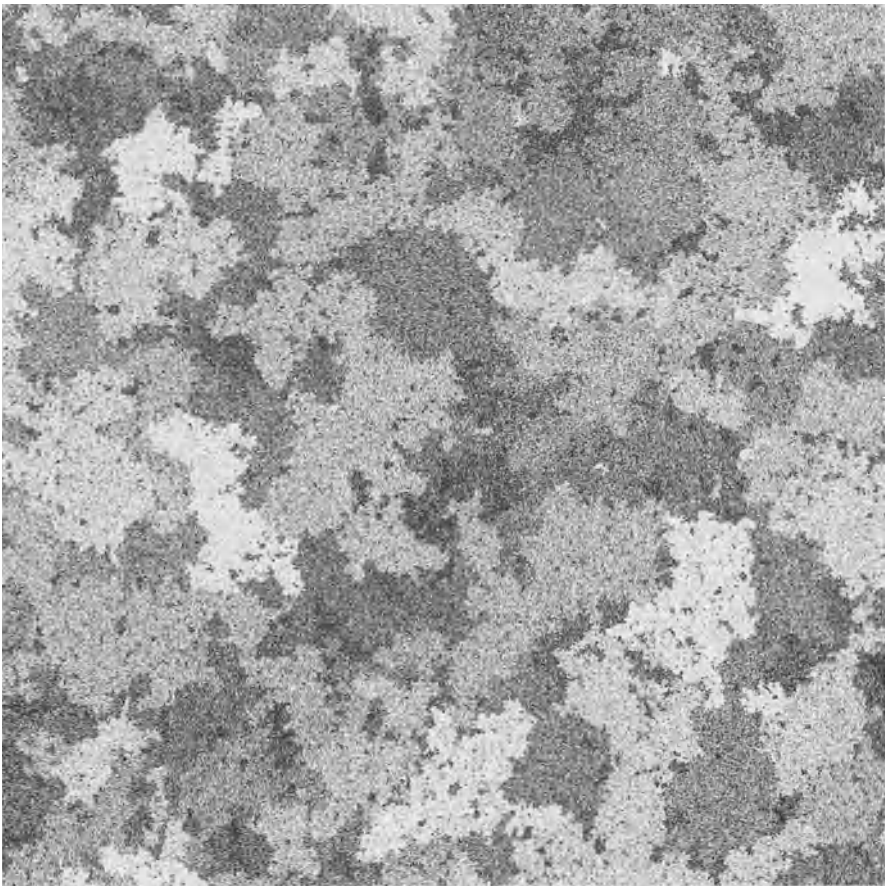

Fig. 6.2. Section of the model forest for $r = 2048$. Black points correspond to trees, while white areas are empty.

less directly in the forest-fire model. In the quasi-steady state, the tree density fluctuates around an average of about 40 %. A section of the forest in displayed in Fig. 6.2; it is a snapshot taken randomly in the quasi-steady state. Black points correspond to trees, while white areas are empty. The scars of several fires are clearly visible.

Figure 6.3 shows the non-cumulative event statistics of the simulation described above. The cluster size s refers to the number of burnt trees. However, some caution is necessary when speaking of the cluster-size distribution in the forest-fire model. The statistics presented here concern the size distribution of the burnt clusters which differs from the size distribution of the clusters present during the simulation. The difference arises from the fact that large clusters are more likely to be hit by a spark than small clusters; the probability of being ignited is proportional to the cluster size. If $p(s)$ is the probability

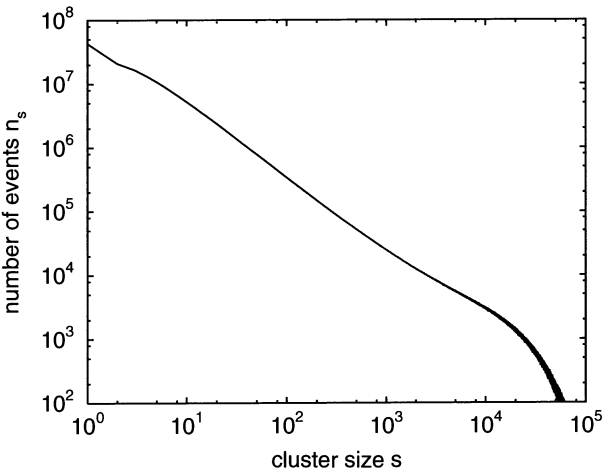

Fig. 6.3. Non-cumulative event size statistics for $r = 2048$ on a 8192×8192 grid. The statistics include 10^9 sparking events; the first 10^8 events were skipped in order to avoid effects of the initial state.

density of the burnt clusters (as analyzed here), the probability density of the present clusters is proportional to $\frac{p(s)}{s}$. Therefore, a power-law distribution of the fire sizes is equivalent to a power-law size distribution of the clusters being present in the forest, but their exponents differ by one.

The simulated statistics extend over 10^9 sparking events. In the beginning, 10^8 events were skipped in order to avoid effects of the initial state, although Fig. 6.1 suggests that much less are sufficient. All simulations shown later include the same number of sparking events. The statistics of the small and mid-sized fires looks similar to the size statistics of the avalanches in the BTW model shown in Fig. 5.5 (p. 97). Apart from some deviations at small sizes, the size statistics follow a power law. However, the statistics differ strongly from those of the BTW model at large event sizes. In the BTW model, we have observed a simple cutoff behavior. Above a certain event size, the number of events rapidly decreases; the cutoff size depends on the size of the lattice. In contrast, the behavior at large event sizes is more complex in the simulation of the forest-fire model discussed here. The number of events increases first compared to a power law at event sizes of about 10^3 trees, resulting in an excess of large fires. The relative excess becomes maximal at sizes of about 10^4 trees; then the number of events decreases rapidly.

In order to understand this behavior, we consider some further results. Figure 6.4 shows the size statistics obtained from a much smaller grid of 128×128 sites for different values of the growth rate r. Obviously, the probability of large fires increases if r increases. For very large growth rates above about 2048, the number of events does not decrease monotonically, but increases again at large event size. This phenomenon can easily be recognized to be an effect of the finite grid size. The 128×128 lattice consists of $16,384$ sites, so that the average tree density increases considerably in each step of growth. This increased density facilitates the propagation of large fires. Fig-

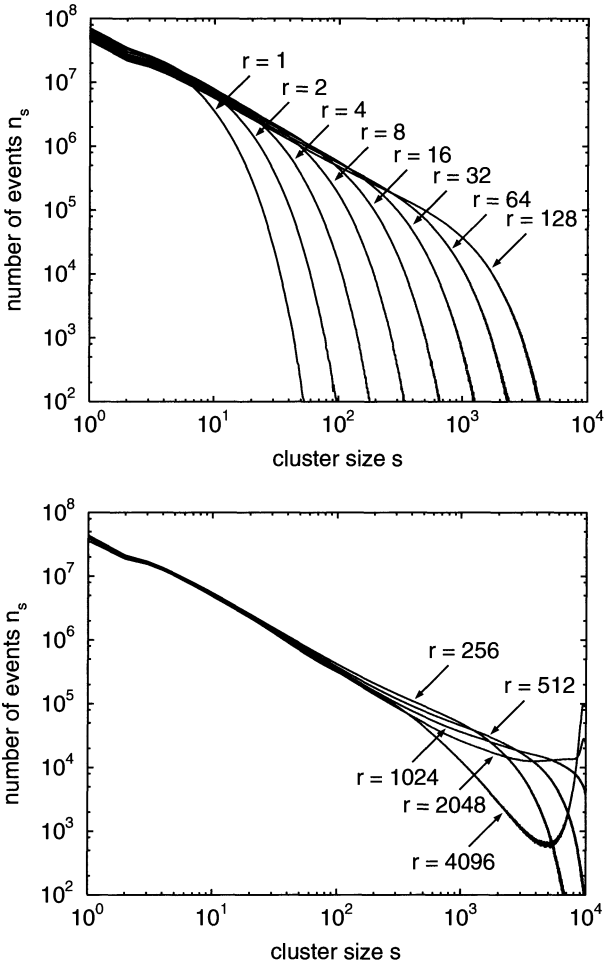

Fig. 6.4. Non-cumulative event size statistics for different values of r on a 128×128 grid.

ure 6.5 confirms this suspicion; the mean tree density strongly increases for large growth rates. For comparison, the average tree density obtained from simulations on a quite large grid of $32,768 \times 32,768$ sites was included in the plot. In contrast to that obtained from the smaller lattice, the density seems to converge towards a fixed value of about 0.4 for large growth rates. This result clearly identifies the finite-size effect on the small lattice.

This finite-size effect is not very interesting since it can be avoided by choosing sufficiently large grids. In contrast to the BTW model, the forest-fire model requires only one bit per site in a computer simulation; 128 MByte RAM are sufficient for storing the state of the $32,768 \times 32,768$ grid.

So let us come back to the increase of the mean event size $\overline{s}$ with r. In the quasi-steady state, there must be a long-term equilibrium between growth and destruction; the mean number of trees destroyed by a sparking event

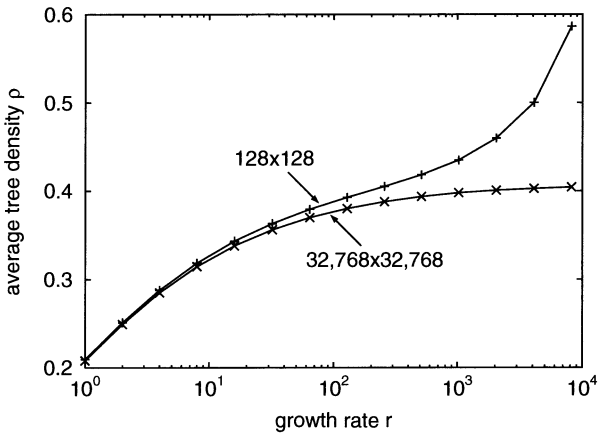

Fig. 6.5. Average tree density as a function of the growth rate, obtained on grids of 128×128 and $32,768\times32,768$ sites.

must coincide with the mean number of trees growing between two events:

$$\rho\,\overline{s} \;=\; (1-\rho)\,r. \tag{6.1}$$

Therefore, either the density must approach unity for large growth rates r or $\overline{s}$ must increase linearly with r. Since the density approaches a finite value if the lattice is sufficiently large, $\overline{s}$ increases. From this argument it becomes clear that a power-law distribution can only arise in the limit $r \to \infty$. Furthermore, the exponent b must not exceed unity. Let us, for proving the latter result, assume an upper-truncated Pareto distribution (Eq. 2.1) with a lower cutoff value $s_{\min} = 1$ and an upper cutoff value $s_{\max}$ that depends on r. If $p(s)$ denotes the probability density, the expected value of the cluster size is

$$\overline{s} \;=\; \int_0^\infty p(s)\,s\,ds \;=\; \frac{b}{1-b}\,\frac{s_{\max}^{1-b}-1}{1-s_{\max}^{-b}} \;\approx\; \begin{cases} \frac{b}{1-b}\,s_{\max}^{1-b} & 0<b<1 \\ \log(s_{\max}) \quad \text{for} & b=1 \\ \frac{b}{b-1} & b>1 \end{cases},$$

provided that $s_{\max} \gg 1$. Thus, Eq. 6.1 can only be satisfied for $b \leq 1$ in the limit $r \to \infty$.

The results found above suggest that the forest-fire model evolves towards a critical state in the limit $r \to \infty$ or, in the original notation, in the limit $\frac{f}{p} \to 0$, provided that the grid is sufficiently large with respect to the actual growth rate r. In this sense, interpreting the deviation from power-law behavior in Fig. 6.3 is straightforward; a lattice of 8192×8192 sites is not large enough for $r = 2048$. However, a more detailed analysis shows that this is wrong. If we plotted the size statistics for larger grids, we would recognize not any difference; the deviation from the power law at large event sizes persists even in the limit of infinite grid size. Finite-size effects are present, but too small to be recognized visually. Figure 6.6 shows the size statistics obtained from simulations on a $32,768\times32,768$ lattice which turns out to be sufficient

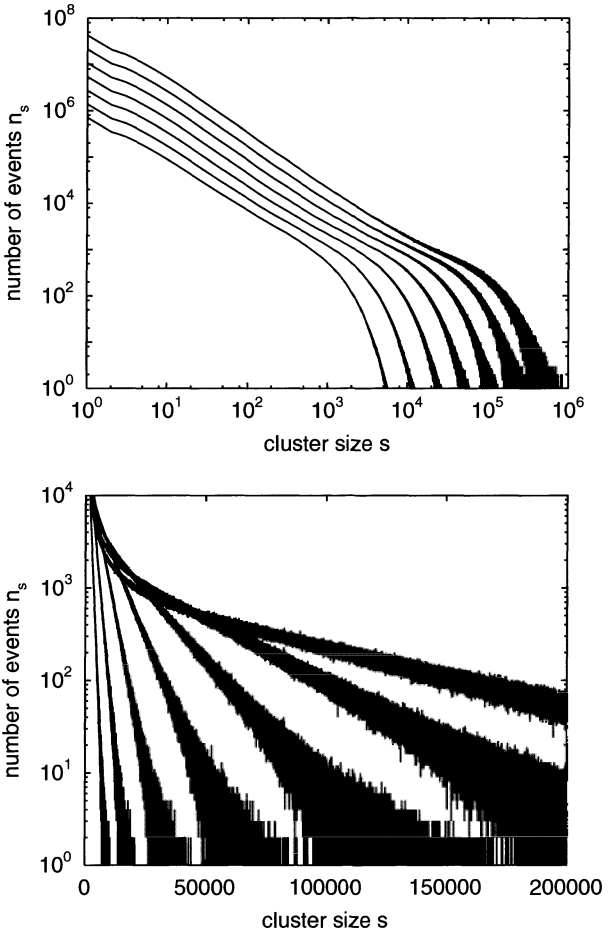

Fig. 6.6. Non-cumulative event statistics for $r = 128$, 256, 512, 1024, 2048, 4096, 8192 (from left to right) on a $32,768 \times 32,768$ grid. Both diagrams represent the data, but the ordinates are differently scaled. In the upper diagram, the curves for $r \leq 4096$ have been shifted downwards in order to allow a distinction at small event sizes.

at least for a visual analysis in the range of r considered here. The range where the fire sizes are power-law distributed extends if r increases, but the deviations from the power law become even stronger.

The lower diagram in Fig. 6.6 represents the same data, but plotted with a linear ordinate. The finding that all curves decrease linearly is surprising; it indicates an exponential decay of the number of fires with their size which is in contradiction to the assumption of a power-law distribution. So the bilogarithmic plot suggests power-law behavior, while the lower diagram suggests an exponential distribution. The explanation of this phenomenon is simple: Small fires are power-law distributed, while large fires follow an exponential distribution. The range of small events is stretched under logarithmic scaling of the ordinate, while it is small under linear scaling. Therefore, the linearly scaled axis focuses on large event sizes, while the logarithmically scaled axis focuses on small event sizes.

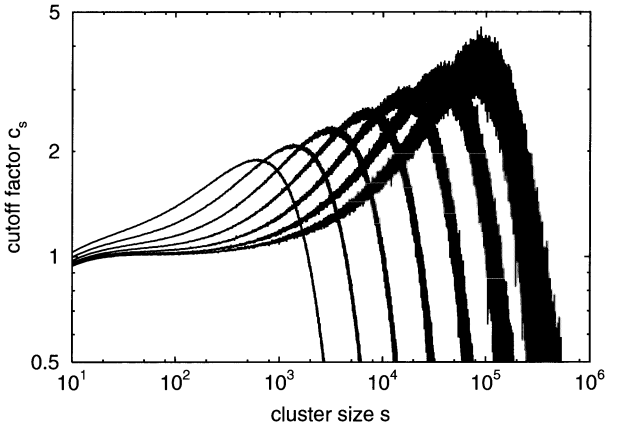

Fig. 6.7. Cutoff factor for $r = 128, 256, 512, 1024, 2048, 4096, 8192$ (from left to right) on a $32,768 \times 32,768$ grid.

Fitting a power-law function to the curves is difficult because even that for $r = 8192$ exhibits a considerable curvature. We therefore perform another analysis by writing the number of fires n_s in the form

$$n_s = \alpha c_c s^{-b-1},$$

where b is the exponent of the corresponding cumulative power-law distribution, c_s is a cutoff factor describing the effect of finite values r, and α is a constant. The factor c_s is computed from the statistical data n_s for various values of b and α; these values are iteratively altered until $c_s \approx 1$ within a reasonable range of object sizes in the limit $r \to \infty$. Figure 6.7 shows the result for $b = 0.23$ and $\alpha = 9.4 \times 10^7$. The plot suggests that these values are appropriate, so that we estimate the exponent of the cumulative size distribution to $b = 0.23$ in the limit $r \to \infty$. In addition, the analysis confirms the finding that the deviation from a power-law distribution at large event sizes becomes even stronger if the growth rate r increases.

Our value $b = 0.23$ considerably differs from earlier estimates. Drossel and Schwabl (1992) found $b = 0$ first, i. e., $n_s \sim \frac{1}{s}$. However, this result turned out to be wrong; later studies resulted in $b = 0.15$ (Grassberger 1993; Henley 1993), respectively, $b = 0.14$ (Clar et al. 1994). Recent estimates (Pastor-Satorras and Vespignani 2000) yielded an even smaller value $b = 0.08$. However, none of these studies takes the local maximum of c_s which even increases if r increases into account; this maximum seems to have a considerable impact on the estimated exponent.

The new result $b = 0.23$ brings the model closer to the statistical properties of forest fires observed in nature. Malamud et al. (1998) found power-law distributions for the sizes of forest fires in different regions with exponents between 0.31 and 0.49. Our result is still outside this range, but the discrepancy is smaller than the variations observed in nature.

One may argue whether the variation in the observed values and the deviation of the simulated results is significant or not. Clearly, the occurrence

of power-law distributions in both nature and model is the central point. But how important is a misfit in the exponents? In the following chapters we will meet examples where power-law distributions with exponents $b > 1$ occur – earthquakes and landslides. Compared to these exponents, both the observed and the simulated exponents of forest-fire distributions are small, and the absolute variation is not very strong. However, if we measure the exponents relatively to each other, the variation is considerable: The highest exponent observed in nature is more than twice as high as that obtained from the model, and even the observed exponents nearly vary by a factor two.

So, shall the variation of the exponents be assessed absolutely or relatively? As so often, it depends on the question. Let us assume two Pareto distributions with exponents b_1 and b_2, but with the same lower limit $A_{\min} = 1$:

$$P_1(A) = A^{-b_1} \quad \text{and} \quad P_2(A) = A^{-b_2} \quad \text{for} \quad A \geq 1.$$

If we are interested in the frequency of events of a given size A or larger, we obtain

$$\frac{P_1(A)}{P_2(A)} = A^{b_2 - b_1}.$$

The absolute difference between the exponents is important here; and a difference of about 0.2 may not be crucial. However, when assessing hazard, we may also ask for the largest event which occurs with a given probability. We then we have to solve the equation $P_1(A_1) = P_2(A_2)$, which leads to

$$A_2 = A_1^{\frac{b_1}{b_2}}.$$

Thus, the ratio of the exponents is important here, so that a factor two between both exponents has a strong effect on this analysis.

After all, is the forest-fire model SOC or rather critical in the classical sense? The major difference towards the BTW model (Sect. 5.2) is the existence of a *tuning parameter* r. When defining SOC in Sect. 5.4, tuning was the crucial point. In many systems, the critical point is only achieved by precisely tuning the system, while SOC systems evolve towards a critical state without any tuning. Formally, the forest-fire model becomes critical only in the limit $r \to \infty$, so it is not SOC in the strict sense.

But from a practical point of view, things are different. If the tuning parameters in a classical system slightly deviate from their critical values, the system is far away from being critical. In contrast, the forest-fire model is approximately critical if r is sufficiently high. Thus, the forest-fire model comes close to its critical state within a wide parameter range, and the assumption $r \gg 1$ can be motivated physically. We could even have introduced the limit $r \to \infty$ directly into the model rules, and then there would be no tuning parameter at all. If we, in contrast, develop a model for water in the earth's crust and only consider critical conditions, nobody would believe in its relevance to nature. Thus, the forest-fire model is much closer to SOC than to tuned criticality, so the existence of a tuning parameter is not crucial here.

6.2 Universality

In the previous section we have learned that the forest-fire model approaches a critical state only in the limit $r \to \infty$, but that the size distribution of the events roughly follows a power law over a limited range of sizes for finite values of r, too. Let us revisit the results presented in Fig. 6.6. If r is large, the statistics coincide at small and intermediate event sizes; remember that the vertical shift was introduced artifically in order to distinguish the curves. This finding can be interpreted in such a way that the exponent b of the power law is roughly independent of r; only the deviations from the power law at large sizes strongly depend on r. Based on this finding, we define:

> If a SOC system involves a parameter, but the scaling exponent b of the event-size distribution is independent of the parameter, the exponent is *universal.*

Strictly speaking, the forest-fire model is not a good example of universality since it only shows SOC in the limit $r \to \infty$. In order to investigate whether the value $b = 0.23$ is universal, we now consider some variations of the original model and find out whether the exponent is affected by modifying the model.

A straightforward modification concerns the mechanism for spreading fires; the rule that a fire propagates to the nearest neighbors on a quadratic lattice is somewhat arbitrary. We could, e. g., choose a hexagonal grid or allow diagonal connections on a quadratic lattice. Let us consider three different modifications:

- We allow diagonal connections, i. e., we assume that each site has eight neighbors. This modification corresponds, e. g., to a drier climate where fires can propagate more easily.
- The propagation of fires is dominated by wind in a pre-defined direction. We assume that fire can only propagate towards the southern and the eastern neighbor, but not towards the northern and the western neighbor.
- The propagation of fires is dominated by wind, but the direction of the wind changes through time. We assume that a new direction is assigned to each fire, but does not change during a fire. The direction is randomly drawn from the set south-east, south-west, north-east, and north-west.

In order to keep the numerical effort at a reasonable level, we choose $r = 2048$ on a grid of 8192×8192 sites, although this is not really close to the limit $r \to \infty$. Figure 6.8 shows the results of these simulations, compared to the simulation based on the original rules. Except for very small and very large events, allowing diagonal connections hardly affects the statistics. The curve obtained from the eight-neighbor lattice is only shifted vertically in downward direction. This result suggests that the exponent $b = 0.23$ does not depend on the choice whether diagonal connections are allowed or not.

This result is surprising at first sight. One may expect that the eight-neighbor lattice facilitates the propagation of fires and should thus lead to

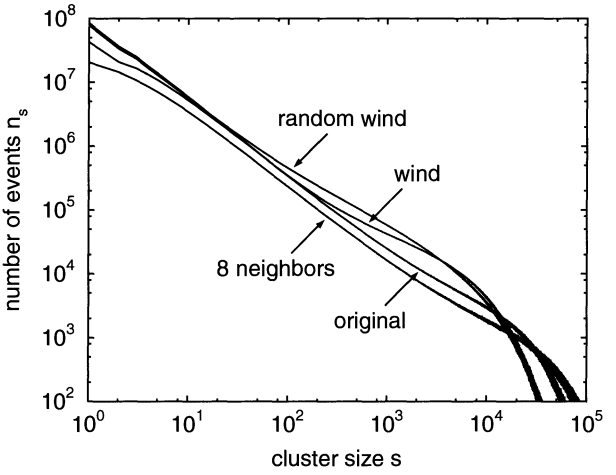

Fig. 6.8. Size statistics obtained under modified rules for the propagation of fires.

a predominance of large fires. Figure 6.9 illustrates that this effect is compensated by a lower tree density. Under the original rules, the mean density approaches about 40 %, while it is only 27 % if diagonal connections are allowed. Thus, the properties of the critical state depend on the model rules in this example, but the scaling exponent of the distribution is universal.

This result even persists in the anisotropic, wind-driven cases. As shown in the lower pictures in Fig. 6.9, the shape of the fires is entirely different from that obtained under the original rules and clearly reveals the direction of the wind. The mean tree density is considerably higher in the anisotropic case; it is about 54 % under constant wind direction and even increases to about 58 % if the wind changes randomly. In both cases, the deviations from a power-law distribution are quite strong, but the plot suggests that they can still be attributed to the finite growth rate r.

Universality means that we cannot influence the exponent b of the event-size distribution at all. In other words, all measures for reducing the number of large fires are useless. However, the exponent cannot be universal against arbitrary changes of the model rules. If we, e. g., allow the propagation of fires only in northern and southern direction, the model falls into a set of independent, one-dimensional forest-fire models. The one-dimensional version was found to show SOC, but the number of fires follows the relation $n_s \sim \frac{1}{s}$, so that $b = 0$ (Clar et al. 1994).

Let us finally investigate whether the number of large fires can be reduced by extinguishing fires. We assume that an attempt is made to extinguish every site which has caught fire. This attempt fails with a given probability q; in this case, those adjacent sites which are occupied by trees are ignited. Otherwise, the extinguished site is burnt down, but nothing else happens. Compared to the established forest-fire model, the probability q introduces an additional parameter. For $q = 1$, the original model is reproduced. In

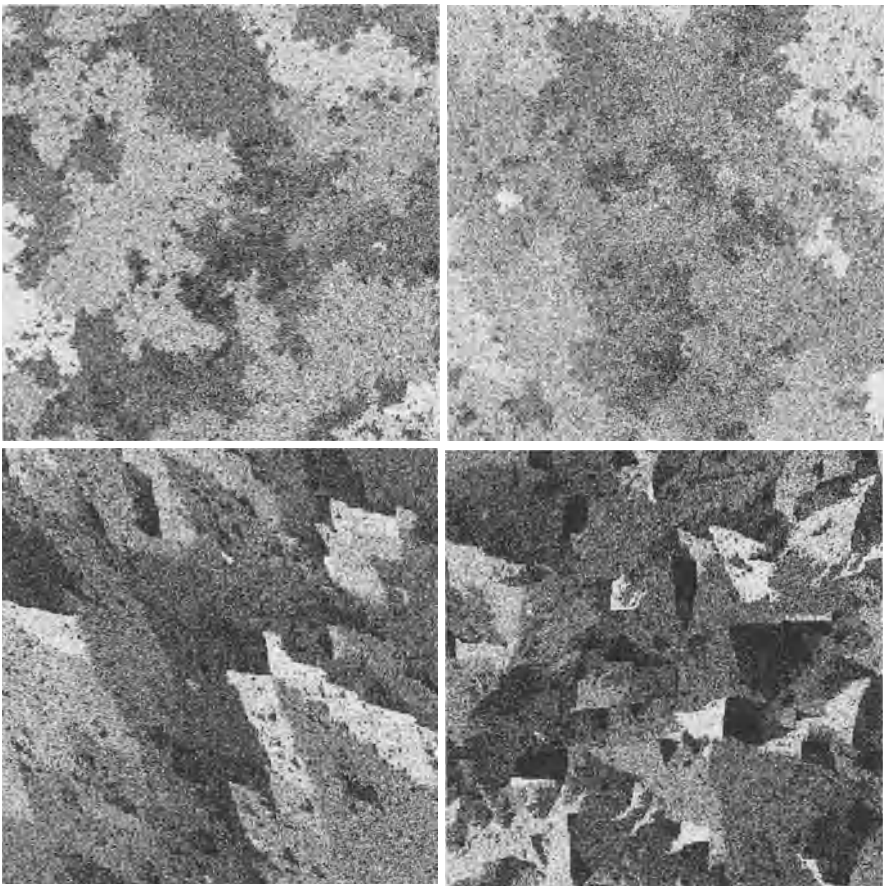

Fig. 6.9. Section of the model forests under different rules for the propagation of fires. Upper left: original rules. Upper right: with diagonal connections. Lower left: wind from north-western direction. Lower right: random wind direction.

contrast, the model becomes trivial for $q = 0$ because all fires are restricted to a single site then. Figure 6.10 shows the statistics obtained for different values of q. For $q = 0.75$, the statistics hardly differs from the original model. The number of large fires gradually decreases, while the number of mid-sized fires increases. More effort must be spent for extinguishing fires in order to achieve a considerable reduction of the frequency of large fires. The curves for $q = 0.5$ and $q = 0.25$ illustrate that the power-law distribution is entirely lost if we spent enough amount for fighting fires. This behavior is reflected by the mean tree density. It increases from 40 % for $q = 1$ to 56 % for $q = 0.75$; then it rapidly grows to 97.5 % for $q = 0.5$ and reaches 99.8 % for $q = 0.25$.

This result allows two conclusions. First, universality is not absolute in general; in some cases it breaks down if strong variations are introduced.

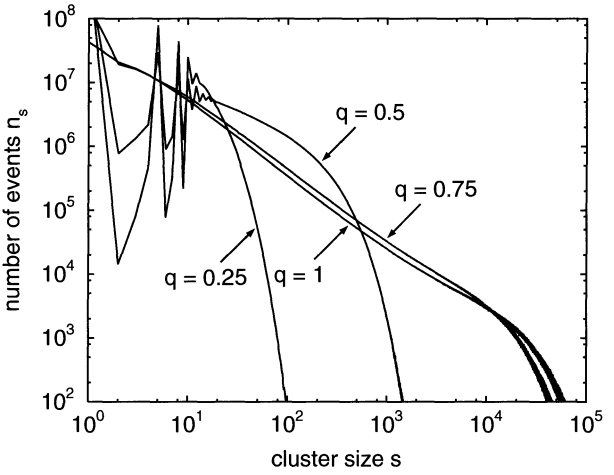

Fig. 6.10. Size statistics of the fires for different values of the spreading probability q.

Second, extinguishing fires may in fact be a good strategy for reducing the danger of large fires. However, we should be careful with this result as facilities for fighting fires are limited in reality. So the probability q will increase as soon as a fire once becomes large. If we extend the model by such a rule, the result may be entirely different. But no matter whether we end up at a realistic model or not, simple models such as the forest-fire model offer at least a wide field for playing different scenarios.

6.3 Non-Equilibrium States in SOC Systems

We have seen so far that the behavior of SOC systems may be universal under variations of the model parameters or not. However, this does not imply that temporal changes in the model parameters, e. g., variations in the direction of wind, do not affect the size distribution of the events. When investigating universality, we always gave the system enough time to approach the critical state with respect to the new conditions. The critical state may strongly depend on the model rules, although the size distribution of the events may finally be the same. If we, e. g., allow diagonal connections, the mean tree density eventually decreases from about 40 % to 27 %. Obviously, this *transition* takes some time. But what happens during this time?

Let us begin with another scenario. We start at a forest in its critical state with respect to the rule that fire propagates towards diagonal neighbors, too. We then cut down the branches of the trees, so that the propagation of fires is aggravated. Let us assume that this leads back to the original rules where diagonal connections are not allowed. So we start at a tree density of 27 % which increases through time to 40 %. The simulation is performed in the following manner: In the beginning of each step, we analyze what happens

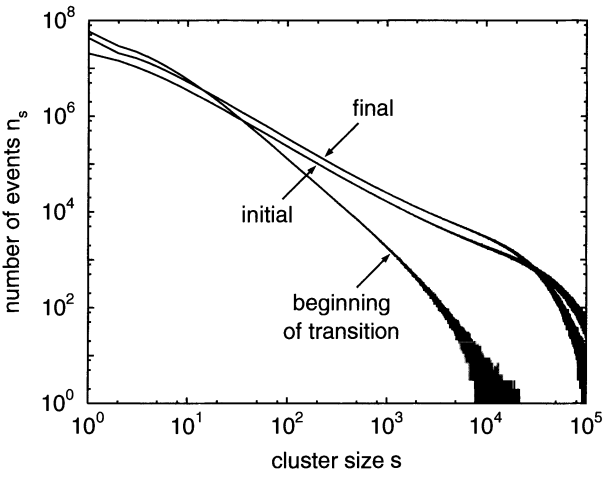

Fig. 6.11. Size statistics of the fires during the transition from next-neighbor and diagonal to next-neighbor connection.

if we ignite a randomly chosen site and burn it down without allowing diagonal connections. After determining the size of this fire, we go back to the state where the site has just been ignited and burn it down using diagonal connections, too. Finally, $r = 2048$ trees are seeded as before. So the forest remains in its critical state according to the rules allowing diagonal connections, but we can make statistics of the fires which occur immediately after the branches have been cut. Figure 6.11 shows that even permanently cutting trees has no effect in the long-term run, but immediately after doing so, the number of large fires is reduced by some orders of magnitude. The mean damage per spark (the number of burnt trees) is about 1500 in the original state, decreases to about 10 immediately after altering the rules, and finally recovers to about 1220 trees. In terms of criticality in the classical sense (Sect. 5.1), the state with a mean tree density of 27 % is *subcritical* if diagonal connections are not allowed.

Figure 6.12 refers to the opposite transition. Starting from the original rules, we investigate what happens immediately after allowing diagonal connections. The reader may easily recognize that the scheme introduced above may lead to series of fires which affect nearly the same cluster, so that consecutive events are not independent. In order to avoid this problem, we apply 10,000 steps according to the original rules between investigating fires with diagonal connections. The price of this modification is a smaller statistics. However, the data clearly show that the tail of the distribution exceeds the power law by far; it looks like the size distribution obtained in Sect. 5.1 under *overcritical* conditions. The mean number of destroyed trees per spark is about 1.8×10^6; this is more than 1000 times larger than the mean damage in the original model and even larger than the largest fire that occurred when simulating 10^9 events in the original model. Obviously, the effect of such a sudden change in the conditions would be a disaster in nature.

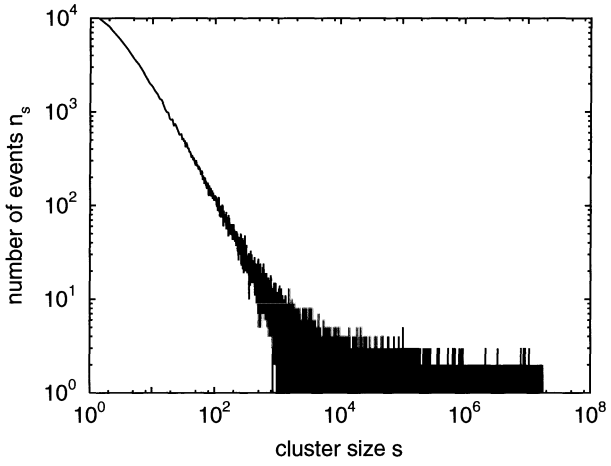

Fig. 6.12. Size statistics of the fires immediately after diagonal connections have been allowed.

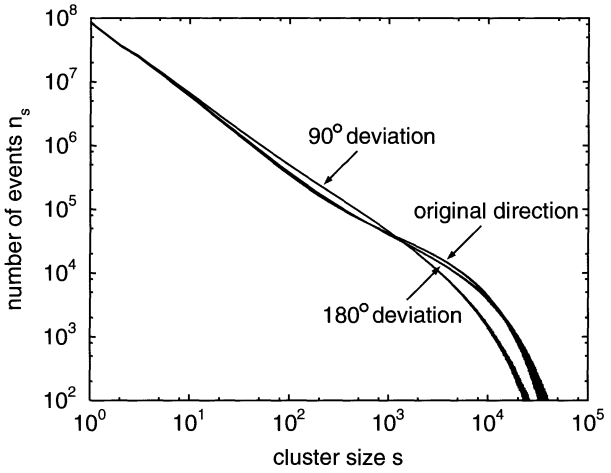

Fig. 6.13. Size statistics of the fires immediately after the direction of the wind has changed.

However, suddenly allowing diagonal connections is a strong interference. Changing, e. g., the direction of the wind in the wind-driven case is more moderate. The results shown in Fig. 6.13 suggest that switching towards the opposite wind direction hardly affects the size distribution, while changing the direction by 90° introduces a moderate change. The number of fires of intermediate sizes increases, whereas the number of large events decreases.

In summary, research on SOC mainly addresses properties of the critical state. However, transitions resulting from changes in parameters or boundary conditions offer a variety of different behaviors which are at least as interesting as the properties of the critical state itself. When assessing natural hazards regarding changing climatic conditions or human impact, investigating these transitions may be the key to understanding the often highly non-linear responses of natural systems.

7. Earthquakes and Stick-Slip Motion

Earthquake research has been one of the most challenging fields of geophysics. Apart from the disastrous effects of large earthquakes, their occurrence is the most striking evidence for the earth's crust not being a static object, but exposed to permanent deformation.

In the twentieth century, plate tectonics became the central concept in understanding crustal dynamics. From a simplified point of view, the lithosphere is subdivided into several plates; boundaries between the plates are subdivided into spreading centers (ocean ridges), subduction zones (ocean trenches), and transform margins. The motion of the plates is mainly driven by thermal convection in the earth's mantle. However, this is only half of the story; at least the oceanic crust is not just pushed by mantle convection, but rather a part of the convecting system. Ocean ridges are located where a strong upward flow of heat and mantle's material occurs. Due to the cooling of the material, parts of it form new crust; this process is called seafloor spreading. As a consequence of seafloor spreading and convection below the crust, oceanic crust is driven away from the ridges. Since oceanic crust is heavier than continental crust, it is mainly subducted under the continental crust at ocean trenches; this process closes the cycle of convection. Finally, plates neither converge nor diverge considerably at transform margins where displacement mainly takes place along the margin.

However, we should not imagine the plates to be isolated objects with smooth edges sliding along each other continuously. Instead, the deformation at subduction zones and transform faults is governed by a *brittle rheology.* If stress is applied, the material mainly behaves elastically until the stress exceeds a certain threshold. Then, a displacement occurs instantaneously, while the stress drops to a lower value – the material breaks. Simply spoken, this is an earthquake.

In some cases, the displacement arising from large earthquakes is even visible at the earth's surface. However, the consequences of such a displacement are negligible compared to those of the *seismic waves* released by the earthquake. Since the propagation of seismic waves is a standard topic in geophysics, their theory should be found in any book on seismology (e. g. Aki and Richards 1980; Bolt 1999; Lay and Wallace 1995; Scholz 1990). So let us here be satisfied with the result that a major part of the energy released

by an earthquake is carried away by seismic waves, and that these waves are responsible for the damage.

As seismic waves contain long-range components, the region of the earth's surface affected by an earthquake is much larger than the region of the fault or subduction zone where the earthquake took place. Obviously, it is mainly this property which makes earthquakes so hazardous. On the other hand, this can be seen as a lucky situation from an academic point of view if we take the freedom of looking for positive aspects in phenomena with such a high death toll. Since the twentieth century, an extensive network of seismic stations for monitoring earthquake activity has been built up, providing an excellent data basis. From these data, not only global and regional earthquake catalogs can de derived, but also valuable information on the structure of the earth's interior and the crust.

Since earthquakes cannot be avoided, applied earthquake research has focused on prediction both in a deterministic and in a statistical sense. Statistical prediction concerns the probability of earthquakes of certain sizes, mainly at a regional scale; it is an essential part of, e. g., defining standards for the safety of buildings. However, statistical prediction may not only address average rates of seismic activity, but also the temporal behavior. Here, the fundamental question is whether the temporal distribution of large earthquakes is completely random or involves some regularity. In other words: Is there a significantly increased danger at San Francisco because there has been about one disastrous earthquake there per 100 years in the past, and the last one took place nearly 100 years ago?

However, for avoiding loss of life, a deterministic, short-term prediction would be desirable. Much effort has been spent on analyzing *precursor phenomena* such as foreshocks, emission of gases from the crust or anomalies in the magnetic field. Unfortunately, precursor phenomena are rarely unique. In some cases, a series of foreshocks after a period of quiescence announced a large earthquake, but several disastrous earthquakes came as a complete surprise. Thus, precursor phenomena can often be recognized afterwards, but it seems that we are still far away from a reliable prediction.

As earthquakes are in fact not restricted to subduction zones and large transform faults, the idea of plate boundaries being the source of earthquakes is oversimplified. Crustal deformation is more complex and usually distributed over relatively broad zones. This deformation is mainly concentrated at a huge number of large and small faults, and each of them is in principle able to generate earthquakes. Figure 7.1 shows the two major types of faults, apart from transform faults discussed above. Normal faults occur under extension tectonics where volume is supplied to allow the hanging wall to sink. Thrust faults, also called reverse faults behave the opposite way. Due to a compression, the hanging wall is pushed upon the foot wall. However, from their basic mechanics with respect to earthquakes, both are similar to subduction zones and transform faults.

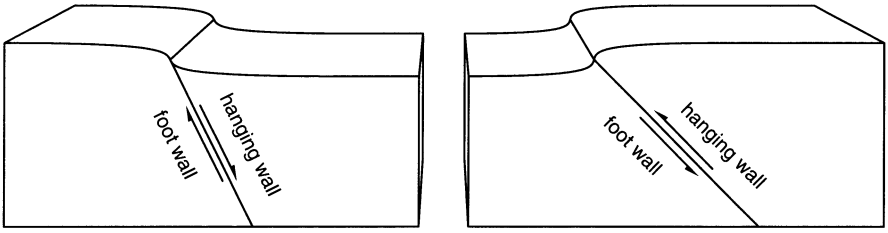

Fig. 7.1. Normal (left) and thrust or reverse (right) faults.

7.1 The Fractal Character of Earthquakes

Monitoring seismic activity has led to an extensive statistics concerning the frequency of earthquake occurrence. While early measures of earthquake sizes were based on the damage caused by an earthquake, the *magnitude* was the first quantitative measure of the source strength of an earthquake. It was introduced by the seismologist C. F. Richter in the 1930's. Since it is based on an instrumental record, it is more objective than purely empirical measures, although there are still some ambiguities. Originally, the logarithm of the amplitudes of the seismic waves was used for defining the magnitude.

Richter's setup was based on a particular type of seismometer, placed in a distance of 100 kilometers from the earthquake source. Since this setup is not realizable in nature, corrections must be applied to the measured oscillation of the seismometer. These corrections depend on the location of the seismometer with respect to the source, the type of waves considered, and the geological setting. Thus, there is not a unique definition of the magnitude. Existing definitions overlap in certain ranges of magnitude, but may considerably differ in some cases. Gutenberg and Richter (1954) were first able to state the relation

$$\log_{10} N(m) \;=\; a - b\,m$$

for the number $N(m)$ of earthquakes per unit time interval with a magnitude of at least m in a given region. The *Gutenberg-Richter law* (GR law in the following) contains two parameters a and b. It is found to be applicable over a wide range of earthquake magnitudes globally as well as locally. The parameter b is often denoted *Richter's b-value*; it slightly varies from region to region, but is generally between 0.8 and 1.2. The world-wide earthquake catalog consist of about 50,000 earthquakes per year with magnitudes between 3 and 4, 6000 with magnitudes between 4 and 5, 800 with magnitudes between 5 and 6, 100 with magnitudes between 6 and 7, and 15 earthquakes with magnitudes between 7 and 8. This result corresponds to a global b-value of about 0.9. Figure 7.2 shows an example of local earthquake statistics consistent with the GR law.

The information supplied by the GR law goes beyond a statistical distribution of the earthquake sizes. Obviously, it implies that the cumulative probability that the magnitude of an earthquake is at least m is $P(m) \sim 10^{-bm}$,

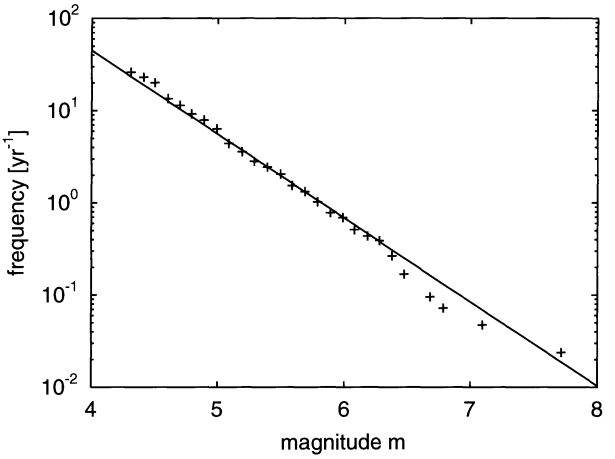

Fig. 7.2. Frequency of earthquake occurrence in Southern California between 1932 and 1972 (Main and Burton 1986). The data were taken from Turcotte (1997).

but in addition, it quantifies the absolute rate of earthquake occurrence. The parameter a quantifies the absolute number of earthquakes with positive magnitudes; so it is a measure of the regional seismic activity. Therefore, a is exposed to a strong regional variation, in contrast to the b-value which quantifies the relationship between earthquakes of different sizes. Relationships describing the frequency of events in terms of their size are called *frequency-magnitude relations*; they are an essential part of hazard assessment. We will meet another example of a frequency-magnitude relation when discussing landslides in Chap. 8; further prominent examples are the frequency-magnitude relations of rainstorms and floods.

The distribution of the magnitudes found above, $P(m) \sim 10^{-bm}$, is not a power law with respect to m. However, we have seen that m is a logarithmic measure of a physical property. Originally, this property was the amplitude of a seismometer, but the magnitude can be related to several other properties such as the total amount of energy released by the earthquake. Roughly speaking, the energy increases proportionally to $10^{\frac{3}{2}m}$; an earthquake of magnitude $m+1$ releases about 30 times more energy than an earthquake of magnitude m. If we consider the property $E := 10^{\frac{3}{2}m}$ instead of m, the size distribution immediately turns into a power law: $P(E) \sim E^{-\frac{2}{3}b}$. Under this aspect, the GR law is a power-law distribution which is hidden behind the logarithmic definition of the magnitude.

In theory, another property is often preferred to amplitudes of seismometers or energies – the *seismic moment* M. Let us for simplicity assume a planar fault where the locations along the fault plane are characterized by two-dimensional vectors $\boldsymbol{x}$. Then, a displacement vector $\boldsymbol{u}(\boldsymbol{x})$ can be assigned to each point at the fault plane; it describes the displacement of the material at both sides of the fault relative to each other occurring as a consequence of an earthquake. The set of points where $\boldsymbol{u}(\boldsymbol{x}) \neq 0$ is called *rupture area*; let A denote its size. The seismic moment is defined by

$$M \coloneqq \mu \left| \int \boldsymbol{u}(\boldsymbol{x})\, dx_1 dx_2 \right|, \tag{7.1}$$

where μ is the shear modulus of the rock. In this equation, it does not matter whether we integrate over the whole fault or only over the rupture area. If we introduce the mean displacement along the rupture area:

$$\overline{u} \coloneqq \frac{\left| \int \boldsymbol{u}(\boldsymbol{x})\, dx_1 dx_2 \right|}{A},$$

Eq. 7.1 can be written in the form

$$M = \mu A \overline{u}. \tag{7.2}$$

Empirically, the relationship

$$\log_{10}\left(\tfrac{M}{1\mathrm{Nm}}\right) \approx \tfrac{3}{2} m + 16$$

was found to be valid for not too large earthquakes. A theoretical basis of this relationship was given by Kanamori and Anderson (1975). Since the relationship between moment and magnitude is essentially the same as that between energy and magnitude, the statistical distribution of the moments is similar to the distribution of the energies:

$$P(M) \sim M^{-\frac{2}{3}b}. \tag{7.3}$$

The power-law distribution of the moments can be transformed to a power-law distribution of the sizes of the rupture areas A, which is a scale-invariant distribution with respect to spatial sizes. Again, Kanamori and Anderson (1975) showed theoretically that the mean displacement $\overline{u}$ increases like the linear extension r of the rupture area, i. e., like $\sqrt{A}$. This leads to

$$M \sim A^{\frac{3}{2}}, \tag{7.4}$$

and thus to a fractal distribution of the sizes of the rupture areas:

$$P(A) \sim A^{-b}. \tag{7.5}$$

In Sect. 1.3, the symbol b was already used for the scaling exponent of power-law distributions. At this point we see that this duplicate meaning is even a fortunate choice since the b-value in the GR law coincides with the exponent of the size distribution of the rupture areas at least theoretically.

But where does the fractal distribution of the earthquake sizes come from? The distribution of fault sizes in the earth's crust may be fractal, and each fault may have its own characteristic earthquakes whose sizes are related to the fault size. In this model, the GR distribution is nothing but a reflection of the fractal fault size distribution, combined with the average recurrence time of fault activity.

This theory is partly supported by observations. Fault sizes seem to be fractally distributed (e. g. Turcotte 1997), although the quality of the available data sets cannot compete with earthquake statistics. Furthermore, seismic activity is not completely random in time. Observations suggest some kind of regularity in the occurrence of large earthquakes, so that they can be interpreted as events being characteristic for certain faults. This behavior is expressed in the term *seismic cycles*. However, this cannot be the whole story. First, there is no evidence for a relationship between the observed GR law and the fault-size distribution. And second, seismic cycles are far away from being periodic in the strict sense; and only a part of the total seismic activity fits into this view. We will come back to this topic in Chap. 10.

Afterwards, it seems that SOC must have been exactly the concept seismologists were waiting for. Since earthquakes are much more important than sandpiles or forest fires from an applied point of view, it is not surprising that the discovery of SOC introduced a completely new direction in seismology. Shortly after SOC was discovered, the original BTW model was transferred to earthquake dynamics (Bak and Tang 1989; Sornette and Sornette 1989; Ito and Matsuzaki 1990). Although the relationship of this model to fault mechanics is as vague as its relationship to real sandpile dynamics, and the exponent b is poorly reproduced, this was a major step towards understanding the fractal character of earthquakes. Under this aspect, reading in Per Bak's book (Bak 1996) about his problems in publishing on this topic is amazing.

If the SOC concept can be applied to earthquakes, each fault may be able to generate earthquakes of various sizes with a fractal size distribution. A modification of the BTW model with a more concrete relation to fault mechanics was published soon after the first papers on SOC in earthquakes had appeared (Chen et al. 1991). However, the majority of the papers on SOC in earthquakes can be seen as a revival of a spring-block model which was more than 20 years old at this time – the Burridge-Knopoff model.

7.2 The Burridge-Knopoff Model

In order to understand earthquake dynamics, we must first come back to the basics of fault mechanics. The fundamental idea leading to distinct events in contrast to continuous sliding is a *threshold behavior* where rupture occurs if the stress exceeds a critical value. Burridge and Knopoff (1967) introduced a simplified, discrete representation of this *stick-slip motion*. Let us start with the one-dimensional version (Fig. 7.3); it can be interpreted as an analog model of a transform fault that is long compared to the thickness of the crust. This model can be seen as the prototype of a *spring-block model*; let us denote it BK model in the following.

For simplicity, one side of the fault is replaced with a rigid plate. The other side is represented by a number of blocks that stick on the rigid plate

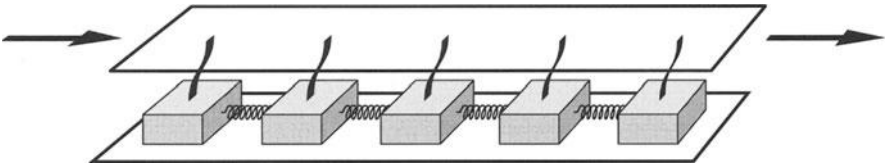

Fig. 7.3. One-dimensional Burridge-Knopoff model.

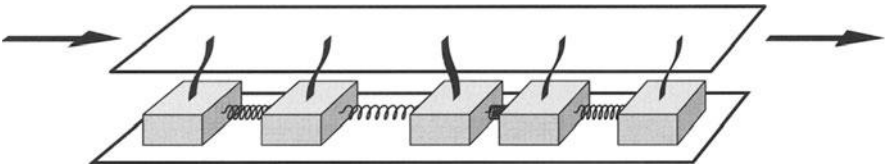

Fig. 7.4. Example of a displaced block in the BK model.

due to friction. A simplified elastic behavior of the rock is regarded by connecting adjacent blocks with horizontal springs. The permanent driving due to tectonic forces which causes the long-term displacement along the fault is incorporated by coupling the blocks with a second rigid plate through leaf springs. This plate is assumed to move at a constant velocity. The force acting on each block is the sum of the forces transmitted through the two horizontal springs and the force transmitted through the leaf spring. If the total force acting on any block exceeds the maximum friction force F, rupture occurs and the block is displaced (Fig. 7.4). We number the blocks with an index i. Let $u_i(t)$ be the displacement of the block i at any time t relative to its initial position, and let v be the constant velocity of the upper plate. According to Hooke's law, the force acting on the block i is

$$F_i(t) = \underbrace{k\left(vt - u_i(t)\right)}_{\text{leaf spring}} + \underbrace{\kappa\left(u_{i-1}(t) - u_i(t)\right)}_{\text{left-hand spring}} + \underbrace{\kappa\left(u_{i+1}(t) - u_i(t)\right)}_{\text{right-hand spring}}. \quad (7.6)$$

The symbol κ denotes the elastic constant of the horizontal springs, while k is that of the leaf springs. We have assumed that all springs are relaxed at $t = 0$. We now assume that a block i becomes unstable at some time t_0. In this case, the static friction is replaced by a (smaller) dynamic friction. If we neglect the dynamic friction for the moment, the displacement of the block is determined by Newton's principle:

$$m \frac{\partial^2}{\partial t^2} u_i(t) = F_i(t),$$

where m now denotes the mass of the block; but there is no danger of confusion with the seismic magnitude in the following.

This is a good point for simplifying the model and getting rid of some of the parameters by introducing the non-dimensional variables

$$t \coloneqq \frac{kv}{F} t, \quad u_i(t) \coloneqq \frac{\kappa}{F} u_i(t), \quad F_i(t) \coloneqq \frac{1}{F} F_i(t)$$

and parameters

$$k \;:=\; \tfrac{1}{\kappa}\,k, \quad m \;:=\; \tfrac{k^2 v^2}{\kappa F^2}\,m.$$

Then, Eq. 7.6 turns into

$$F_i(t) \;=\; t + u_{i-1}(t) - (2+k)\,u_i(t) + u_{i+1}(t). \tag{7.7}$$

The block i becomes unstable if $F_i(t) \geq 1$; its motion is defined by

$$m\,\tfrac{\partial^2}{\partial t^2}u_i(t) \;=\; F_i(t) \;=\; t + u_{i-1}(t) - (2+k)\,u_i(t) + u_{i+1}(t). \tag{7.8}$$

At this point, the reader may prefer variables with clear physical units instead of the rather abstract, non-dimensional quantities. Especially for the time, this seems to be favorable, but the non-dimensional time can easily be interpreted because it is directly related to the velocity of loading the system by moving the upper plate. From Eq. 7.7 we immediately see that a time span of unit length is exactly the time needed to make a completely relaxed block ($F_i(t) = 0$) unstable ($F_i(t+1) = 1$). So a unit time span can be identified with the longest possible time interval between two earthquakes at any part of the fault.

The linear equation of motion (Eq. 7.8) can be solved analytically; the result is a superposition of a motion with constant velocity $\frac{1}{2+k}$ and a harmonic oscillation. Chain reactions are the major feature of this model; each change of the displacement $u_i(t)$ affects not only the force $F_i(t)$, but also the forces acting on the neighbors, $F_{i-1}(t)$ and $F_{i+1}(t)$. As a result, these neighbors may be destabilized during the motion, too. In this case, a system of coupled ordinary differential equations for a set of displacements $u_i(t)$ has to be solved. In principle, such a piecewise linear system can be solved analytically, too. The analytical solution remains valid until the next block becomes unstable; then a system consisting of one more block must be considered. From this point of view, the analytical solution is not as feasible as it may seem, and it may be better to use one of the numerical schemes reviewed in Appendix A.

No matter whether we prefer an analytical or a numerical solution, Eq. 7.8 does not yield the behavior we are looking for. More and more blocks will become unstable through time, but none of them will ever come to rest again. In contrast, instability spreads over a part of a fault during an earthquake, but the motion ceases soon. Obviously, this discrepancy arises at least partly from neglecting the dynamic friction which leads to a dissipation of kinetic energy. Thus, regarding dynamic friction is one way towards more realistic behavior. In the simplest approach, a constant dynamic friction is introduced in Eq. 7.8; and it is assumed that the static friction recovers as soon as the block comes to rest. Thus, the motion stops after half a period of oscillation. This model was investigated in detail for small systems consisting of two blocks. While Nussbaum and Ruina (1987) found periodic behavior, introducing inhomogeneity in the physical properties resulted in deterministic

chaos (Huang and Turcotte 1990, 1992; McCloskey and Bean 1992). However, a model with such a low number of degrees of freedom is not able to reproduce a wide spectrum of event sizes as requested by the GR law.

Still more important, a point or a region within an elastic medium cannot oscillate independently from its vicinity. The oscillation will activate the neighborhood; this results in the propagation of seismic waves which are the major effect of earthquakes at the earth's surface. The dissipation of energy arising from this effect can be roughly incorporated by introducing a radiation term that increases linearly with the velocity into Eq. 7.8:

$$m \frac{\partial^2}{\partial t^2} u_i(t) = F_i(t) = t + u_{i-1}(t) - (2+k)\, u_i(t) + u_{i+1}(t) - \gamma \frac{\partial}{\partial t} u_i(t), \quad (7.9)$$

where γ is a positive parameter. Again, the solution of this system is a superposition of a linear motion and oscillations, but now these oscillations cease through time.

Increasing computer power made spring-block models attractive again in the late 1980's. Carlson and Langer (1989a,b) performed simulations with some hundred blocks and non-constant dynamic friction. With this modified spring-block model, they obtained size distributions which were consistent with the GR law. However, it turned out later that this result only holds for small events, while there is an excess of large earthquakes (Carlson 1991; Carlson et al. 1991, 1994; Knopoff et al. 1992; de Sousa Vieira et al. 1993; Lin and Taylor 1994). For further insights into friction laws for rocks, the review article of Scholz (1998) is recommended.

7.3 Separation of Time Scales

The motion of unstable blocks described by Eq. 7.9 contains three potentially different time scales. These scales can be recognized by writing the equation in the form

$$\tau_o^2 \frac{\partial^2}{\partial t^2} u_i(t) + \tau_d \frac{\partial}{\partial t} u_i(t) + u_i(t) = \frac{t + u_{i-1}(t) + u_{i+1}(t)}{2+k}$$

where

$$\tau_o = \sqrt{\frac{m}{2+k}} \quad \text{and} \quad \tau_d = \frac{\gamma}{2+k}.$$

The first time scale, defined by τ_o, is characteristic for the oscillations of unstable blocks. The second time scale is given by τ_d and arises from the damping term. Both time scales are characteristic properties of the relaxation process; they are therefore called *intrinsic* time scales. Apart from the coupling with adjacent blocks, the right-hand side of the equation introduces the third time scale. It describes slowly loading the system by moving the upper plate. When introducing non-dimensional variables in the previous section, time was rescaled in such a way that this time scale has the order of

magnitude of unity. With respect to the relaxation process, this time scale is an *external* time scale.

With respect to earthquake dynamics, the external time scale is much longer than the internal scales, i. e., $\tau_o \ll 1$ and $\tau_d \ll 1$. So the idea of separating these scales, i. e., performming the double limit $\tau_o \to 0$ and $\tau_d \to 0$, is straightforward. In our example, this means that the motion of the upper plate stops as soon as a block becomes unstable, i. e., that the time t in Eq. 7.9 is replaced by the time of the initial instability t_0. Then, the motion of the blocks is computed, where the potential destabilization of adjacent blocks must be taken into account. Nevertheless, it can be shown that all blocks will come to rest again in the limit $t \to \infty$. Afterwards, the time is reset to t_0, and the simulation of slowly loading continues. We already know the separation of time scales from the sandpile models and the forest-fire model. In the BK model, it was introduced by Rundle and Jackson (1977).

However, this first step of separation does not really make things easier in the BK model. The point is that we are not interested in the short time scale, i. e., in the motion of the blocks during an earthquake. Only the result is important, especially the number of blocks which have been moved, and their displacements. Thus, obtaining this information without simulating the motion during the earthquake should be the goal now. The position of an unstable block immediately after an earthquake is characterized by

$$\frac{\partial}{\partial t} u_i(t) \;=\; 0 \quad \text{and} \quad m \frac{\partial^2}{\partial t^2} u_i(t) \;=\; F_i(t) \;=\; 0.$$

Posing this equation for all these blocks which have become unstable during the earthquake constitutes a system of coupled linear equations for the new positions $\tilde{u}_i(t)$ of these blocks:

$$\tilde{F}_i(t) \;=\; t + \tilde{u}_{i-1}(t) - (2+k)\,\tilde{u}_i(t) + \tilde{u}_{i+1}(t) \;=\; 0. \tag{7.10}$$

In general, a separation of time scales is a major progress in treating a multi-scale system in time. Applying this technique, e. g., to a partial differential equation involving spatial coordinates and time reduces the problem to a steady-state equation involving only spatial coordinates. However, the BK model is a threshold system, which means that the right-hand sides of the equations of motion are discontinuous. As soon as a threshold (here the static friction) is exceeded, the behavior changes drastically. Under this aspect, threshold behavior is the worst kind of non-linearity. As long as we know which blocks are involved in an earthquake, the system's behavior is linear, and we can easily solve the set of equilibrium equations (Eq. 7.10) for these blocks. Unfortunately, the information which blocks become unstable is hidden behind the motion of the blocks. Thus, we only know that each block must satisfy either of the conditions

$$\tilde{F}_i(t) \;=\; 0 \quad \text{or} \quad \tilde{u}_i(t) \;=\; u_i(t) \quad \text{and} \quad F_i(t) \;<\; 1.$$

This system in non-unique; it allows several solutions. One of them should be that resulting from solving the equations of motion completely and performing the limit $t \to \infty$. We can try to obtain this solution by the following iterative procedure: We first solve Eq. 7.10 for the block i which is initially unstable. Afterwards, either of the blocks $i-1$ and $i+1$ (or both) may be unstable, too. In this case, Eq. 7.10 is solved for all unstable blocks simultaneously. As a result of displacing these blocks, some more neighbors may become unstable, and the procedure is repeated. Finally, we end up at a minimum set of blocks which became unstable.

Two problems arise from this procedure. First, the iterative scheme may be time-consuming, so that we may lose the gain of separating time scales partly. Still more important, this minimum set of unstable blocks is not necessarily the one we are looking for. Since a block oscillates before it finally comes to rest, its displacement may exceed its equilibrium displacement during the motion, and thus the force acting on the neighbor may behave similarly. Consequently, the number of blocks becoming unstable during an earthquake may be underestimated.

This unfortunate situation often occurs in modeling. The simplification not just leaves more problems than expected, but also takes us away from the original problem. So there are two ways; going back to the original problem or going a further step ahead – live with rough approximation and try to simplify the model really. For the BK model, this is the transition to a cellular automaton discussed in the following.

7.4 Cellular Automata

In the approach developed in the previous section, simulating an earthquake still requires to solve a system of linear equations or, even worse, several systems until the appropriate solution is found. We can get around this effort by introducing a further simplification: In a first step, we relax that block which is initially unstable according to Eq. 7.10, which means that we replace its displacement $u_i(t)$ with

$$u_i(t) \;:=\; \frac{t + u_{i-1}(t) + u_{i+1}(t)}{2+k} \;=\; u_i(t) + \frac{F_i(t)}{2+k}.$$

In the short notation introduced in Sect. 5.2, this relaxation rule reads

$$u_i(t) \;+\!=\; \frac{F_i(t)}{2+k}. \tag{7.11}$$

Up to this point, this is the same as before; but now we assume that this block immediately becomes stable again at its new position. We then check the stability of its neighbors as before, relax those which have become unstable according to the same rule (Eq. 7.11), and assume that they become stable

immediately. If any of their neighbors have become unstable, the procedure is repeated until all blocks have become stable again.

We thus have replaced a system of equations which determine the equilibrium position of several blocks simultaneously with an explicit rule for each individual block. Such explicit rules are characteristic for cellular automata models. Obviously, the modification simplifies the problem, but takes us further away from the original BK model. In the latter, all blocks being involved in an earthquake are equilibrated afterwards. Transferred to fault dynamics, this means that the stress along the whole rupture area drops to zero. In contrast, the majority of the blocks arrives at a non-zero force finally in the cellular automaton, so that the drop of stress is underestimated.

Figure 7.5 shows a flow chart of this model; it is not optimized in any way. From its basic structure, it looks similar to that of the BTW model given in Fig. 5.2 (p. 92), although the origin of these models is essentially different. This cellular automaton version of the BK model was originally introduced

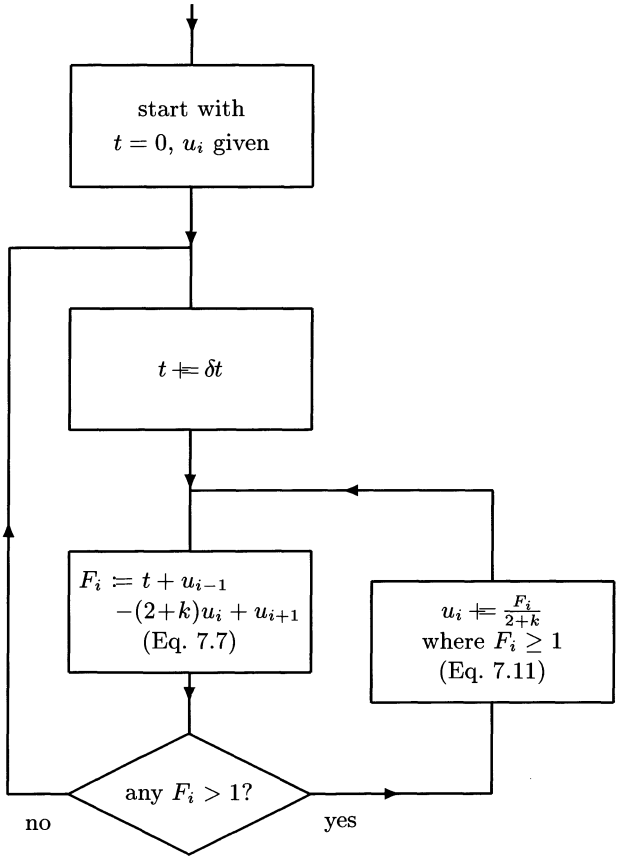

Fig. 7.5. Flow chart of a cellular automaton based on the BK model.

by Nakanishi (1990, 1991). His numerical studies showed that the statistical properties of the cellular automaton are in fact similar to those of the original model which requires a much higher numerical effort.

This model is a good example for understanding what cellular automata are, and how they differ from the majority of models which are based on differential equations:

- The term *cellular* refers to the fact that the model is discrete concerning space. In contrast, partial differential equations (PDEs) provide a continuous description. However, solving a PDE numerically on a computer requires a discretization, too. At this point, the difference reduces to the aspect that a discretized PDE involves physical parameters as well as the mesh width of the discretization, and that it must lead to a well-defined limit if the mesh width tends towards zero.
- The term *automaton* means that the evolution of the system takes place in discrete steps which are not necessarily linked to time in the physical sense. In each step, a set of explicit rules is applied to the cells without solving equations. In contrast, discretizing a PDE involves a time step length and must lead to a well-defined limit if the length of the time steps tends towards zero.

Strictly speaking, the model discussed here is not a cellular automaton in this sense since it involves a continuous time at least on the large scale. This property is often expressed by the terms *continuous automaton* or *continuously driven automaton* (Olami et al. 1992).

Under the aspects discussed above, differential equations are closer to a mathematically correct description of a physical process than cellular automata. At least, posing rules for a cellular automaton leaves more freedom than writing down PDEs, so that cellular automata provide better chances of doing nonsense. Thus, it is not surprising that some scientists decline cellular automata as serious modeling tools.

However, models are not intended to be an exact copy of something, although scientists sometimes seem to look for the truth by putting everything they can find into a model. In contrast, models are invaluable tools for understanding nature by simplifying phenomena and looking at natural phenomena from a more abstract point of view. In this sense, a simple model is not necessarily weaker than a comprehensive one, and this argument can be applied to the comparison of differential equations and cellular automata, too. In principle, cellular automata are just going one step ahead towards a simplification or taking a more abstract point of view. Regardless of its type, a model is good if it captures al least some of the essential properties of a phenomenon. The transition from differential equations to a cellular automaton is just one chance of losing the phenomenon's fundamental properties or introducing artefacts, but it is clearly not the only one. In this sense, good cellular automata are those whose rules are not much further away from the phenomena than a PDE.

After all, how good are the cellular automata models discussed so far? The sandpile model from Sect. 5.5 is a very simple approach to sandpile dynamics. While slope stability was incorporated in a somehow reasonable way, effects of inertia during the motion of unstable grains were completely neglected. Since these effects are quite important in sandpile dynamics, is it not surprising that the behavior of real sandpiles is not described very well by these models. Clearly, toppling of grains can be described by assigning an ordinary differential equation (based on Newton's principle) to each grain. Thus, we can blame the loss of effects of inertia to the transition to a cellular automaton, so that we should stay at differential equations or at least be more careful when deriving a cellular automaton if we aim at a realistic description of sandpile dynamics.

The forest-fire model (Sect. 6.1) is different in this sense. Would describing the propagation of fires by a PDE necessarily be better? In fact, such a model could provide a tool for simulating the migration of the fire front through time; and it could perhaps help to coordinate activities for fighting fires. However, fire will either spread from one tree to its neighbors or not; so we cannot expect that the result obtained from the PDE really differs from that of the cellular automaton in the end. Thus, if we focus on a certain aspect, e. g., the distribution of the fire sizes, the cellular automaton may be as good as a PDE, but is simpler and numerically less demanding.

The cellular automaton realization of the BK model is somewhere between these extreme cases. We have learned that the transition to the cellular automaton alters the properties of the model considerably, e. g., with respect to the drop of stress during an earthquake. However, even the original BK model provides only a very rough description of fault dynamics. The crucial question is which of both steps is worse; under this aspect the loss due to the transition to the cellular automaton may be a minor problem.

7.5 The Olami-Feder-Christensen Model

The one-dimensional spring-block model can easily be enlarged towards two dimensions. This leads to a more realistic, although still abstract, representation of a fault. Figure 7.6 shows the setup; additional leaf springs have been introduced for connecting the blocks in direction perpendicular to the motion of the upper plate. For simplicity, the blocks are still allowed to move only in direction of the motion of the upper plate; introducing a two-dimensional motion would increase the theoretical and numerical effort without leading to significant new results.

The horizontal leaf springs introduce an additional model parameter because their elastic constants are in principle independent of the other parameters. However, they can be related to the properties of the considered material. The force needed for compressing the system in longitudinal direction is proportional to the elastic constant of the longitudinal springs which

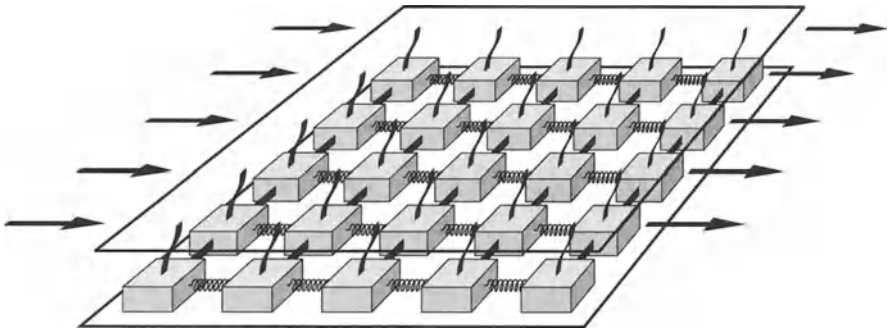

Fig. 7.6. Setup of a two-dimensional spring-block model.

was scaled to unity when introducing non-dimensional variables in Sect. 7.2. In contrast, the force needed for shearing the block is proportional to the elastic constant of the new leaf springs. Thus, their elastic constant s (in the non-dimensional formalism) is the ratio of shear modulus and linear elastic modulus of the rock. In general, s is between 0 and $\frac{3}{4}$; for most rocks it is close to $\frac{1}{2}$. But let us, for simplicity, make the somewhat unrealistic assumption $s = 1$. In other words, the elastic constants of the horizontal leaf springs coincide with those of the longitudinal springs. This will finally lead to an isotropic cellular automaton and keep the equations shorter, although all steps can be performed for the anisotropic case as well.

Let us number the blocks with a pair of indices (i, j). In extension of Eq. 7.7, the force acting on the block (i, j) is

$$F_{i,j}(t) \;=\; t+u_{i-1,j}(t)+u_{i+1,j}(t)+u_{i,j-1}(t)+u_{i,j+1}(t)-(4+k)\,u_{i,j}(t). \quad (7.12)$$

We again assume that the block (i, j) becomes unstable if $F_{i,j}(t) \geq 1$ and relaxes to its equilibrium position where $F_{i,j}(t) = 0$. This leads to the relaxation rule

$$\begin{aligned} u_{i,j}(t) \;&:=\; \frac{t + u_{i-1,j}(t) + u_{i+1,j}(t) + u_{i,j-1}(t) + u_{i,j+1}(t)}{4 + k} \\ &=\; u_{i,j}(t) + \frac{F_{i,j}(t)}{4 + k}. \end{aligned} \quad (7.13)$$

With respect to efficient simulations of large systems, some minor modifications should be applied to the model. The first point concerns the time step length δt. If it is too large, we run into conflict with the assumption that a block becomes unstable immediately if its force exceeds the static friction. Still more important, separation of earthquakes becomes difficult if they start at the same time. On the other hand, numerical effort is wasted if δt is too small. The question for the optimum time step length can easily be answered: From Eq. 7.12 we obtain

$$F_{i,j}(t+\delta t) \;=\; F_{i,j}(t) + \delta t \quad (7.14)$$

during a period of quiescence between two earthquakes. Thus, the block with the highest force will be the first to become unstable; it reaches the static friction after

$$\delta t \;=\; 1 - \max_{i,j}\{F_{i,j}(t)\}. \tag{7.15}$$

Another point for optimizing the model is that the stability is governed by the forces, while we describe the earthquakes themselves by changes in the displacements. It is quite easy to get rid of the displacements: If the block (i,j) is unstable, its force $F_{i,j}$ decreases to zero, while the forces acting on its neighbors increase. This increase can be computed by inserting Eq. 7.13 into Eq. 7.12, written for the blocks $(i-1,j)$, $(i+1,j)$, $(i,j-1)$, respectively $(i,j-1)$ instead of (i,j). Then the rule for updating the forces reads:

$$F_{i\pm1,j}(t) \;+\!\!=\; \alpha\, F_{i,j}(t), \quad F_{i,j\pm1}(t) \;+\!\!=\; \alpha\, F_{i,j}(t), \quad \text{and} \quad F_{i,j}(t) \;:=\; 0, \tag{7.16}$$

where

$$\alpha \;=\; \frac{1}{4+k}.$$

The cellular automaton defined by Eqs. 7.14 and 7.16 is the Olami-Feder-Christensen (OFC) model (Olami et al. 1992); it is the perhaps most

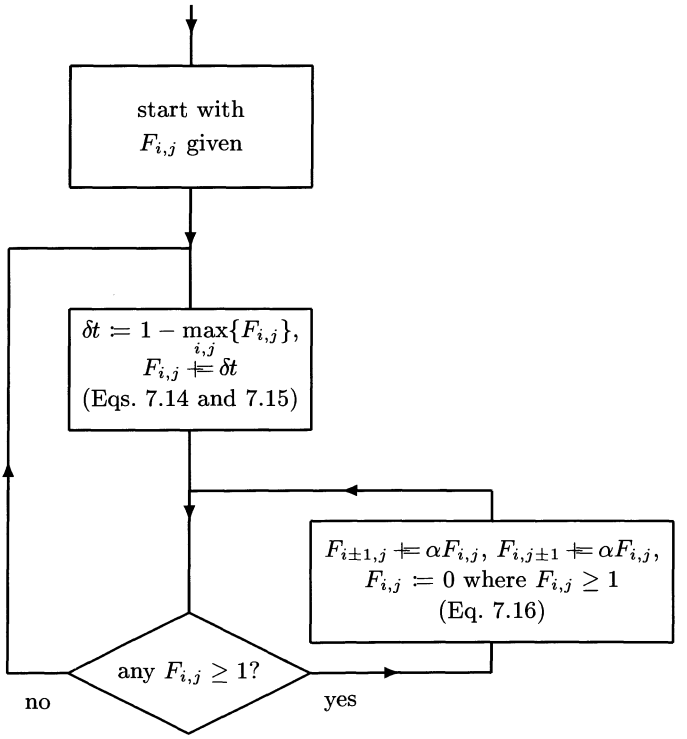

Fig. 7.7. Flow chart of the OFC model.

widespread spring-block model beside the original, one-dimensional BK model. The flow chart sketched in Fig. 7.7 shows a high degree of similarity to the BTW model (Sect. 5.2). The variables $u_{i,j}$ in the BTW model correspond to the forces $F_{i,j}$. Both models are driven by increasing their variables through time. If the variable reaches a given threshold, a relaxation takes place. In both models, the value of the variable at the unstable site decreases; and the difference is redistributed among the adjacent sites at least partly. However, taking a closer look at the rules reveals some more or less important differences:

- The variables $u_{i,j}$ of the BTW model are discrete, while the variables $F_{i,j}$ of the OFC model are continuous.
- The relaxation rule of the BTW model (Eq. 5.2) is conservative; the sum of all values $u_{i,j}$ on the lattice remains constant. In contrast, Eq. 7.16 implies that a total amount of
$$4\alpha F_{i,j} = \tfrac{4}{4+k} F_{i,j}$$
is transferred to the neighbors, while $F_{i,j}$ drops to zero. Thus, the relaxation rule of the OFC model is *non-conservative*; each relaxation leads to a loss in the sum of the forces over the whole lattice. The relative *level of conservation* is $\frac{4}{4+k}$. The OFC model becomes conservative in the limit of weak coupling with the moving plate, i. e., in the limit $k \to 0$.
- The number of grains transferred by a relaxation of an unstable site is fixed in the BTW model; it is four, even if more than four grains are present at the moment. In contrast, the actual amount of force $F_{i,j}$ is relaxed in the OFC model. As a consequence, the BTW model is an *Abelian* automaton, while the OFC model is *non-Abelian*. In contrast to the BTW model, the result of an avalanche depends on the order of performing the relaxations in the OFC model. This result can easily be recognized by considering two adjacent sites both participating in an avalanche. Only the site being relaxed later than the other can end up at zero force; the other one will receive a certain amount of force from the later relaxation. When programming the OFC automaton, one should be aware of this difference. A recursive implementation as discussed in Sect. 5.2 may lead to unwanted preferential directions. Mostly, the order suggested by the flow chart (Fig. 7.7) is used. In the first step, that site which has just become unstable is relaxed. Then, these neighbors of the first site which have become unstable are relaxed simultaneously. In the third step, these neighbors of the sites relaxed in the second step which have become unstable are relaxed simultaneously, and so on.
- The BTW model is driven by adding grains at randomly chosen sites, while the variables $F_{i,j}$ of the OFC model increase uniformly through time. Thus, the OFC model is entirely deterministic and does not involve any randomness. However, we have seen in Sect. 4.6 that the randomness in the BTW model is a pseudo-randomness since random numbers on a computer are

not really random. Therefore, this difference may be a minor one. Nevertheless, we should take care that the deterministic OFC model does not get stuck at an unstable fixed point or run into another artificial behavior by an unfortunate choice of the initial condition. If we, e. g., start with a completely relaxed system ($F_{i,j} = 0$ everywhere), all sites will become unstable simultaneously, which may result in an unrealistic periodic behavior. Starting the OFC model with random values of the forces is the simplest strategy for avoiding this problem.

7.6 Boundary Conditions in the OFC Model

So far we have disregarded the fact that we cannot simulate infinite systems on a computer. As in the models discussed earlier, we have to introduce appropriate boundary conditions for the blocks at the edges of the system. Three types of boundary conditions are reasonable here:

- *Free boundaries* correspond to the original setup shown in Fig. 7.6 where the boundary blocks have less neighbors than those in the bulk. This type of boundary conditions leads to a modified rule for updating the forces when a block becomes unstable (Eq. 7.16). It can easily be shown that the transmission parameter α is no longer the same for all blocks; it has to be replaced by an expression that depends on the actual number of neighbors $n_{i,j}$ according to $\alpha_{i,j} = \frac{1}{n_{i,j}+k}$. With respect to the forces, this boundary condition describes *reflecting boundaries.*
- *Rigid-frame boundary conditions* hinge on the idea that the blocks at the boundaries are connected with a rigid frame in the same way as they are connected with the other blocks (Fig. 7.8). In this case, the rule for updating the forces when a block becomes unstable (Eq. 7.16) remains the same; the forces transmitted to the frame are simply lost. For this reason, this boundary condition is an *open boundary condition* with respect to the

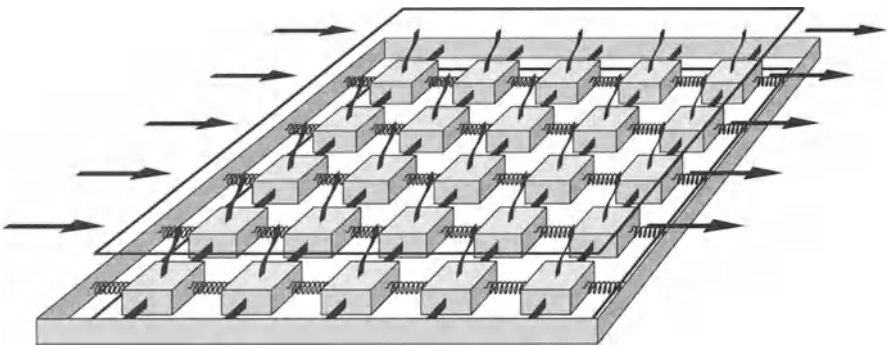

Fig. 7.8. Rigid-frame boundary conditions in the two-dimensional OFC model.

forces; it is very similar to the boundary condition of the BTW model. The long-term behavior of the frame must be specified, too. If its position remains constant through time, nothing has to be changed in the updating of forces between two earthquakes (Eqs. 7.14 and 7.15), but the blocks will hit the frame after some time. Surprisingly, the rules of the cellular automaton do not encounter any problem with this as a result of the linear elastic approach. All the blocks will finally be outside the frame, but this is rather a matter of consistency with the original model than a real problem in a simulation. This little inconsistency can be avoided by connecting the frame with the upper plate, so that it moves with the same velocity as the blocks in the mean. In this case, the rule for updating the forces between two earthquakes must be modified according to the fact that the forces at the boundaries grow faster than in the bulk due to the motion of the frame. The driving rule (Eqs. 7.14 and 7.15) has to be replaced with

$$F_{i,j}(t+\delta t) \;=\; F_{i,j}(t) + (1+k\,(4-n_{i,j}))\,\delta t$$

where

$$\delta t \;=\; \min_{i,j}\left\{\frac{1-F_{i,j}(t)}{1+k\,(4-n_{i,j})}\right\}.$$

- *Periodic boundary conditions* seem to be the most elegant way of getting around boundary effects. The blocks at the boundary are connected with those at the opposite boundary. Thus, the force leaving the region at one boundary comes in again at the opposite edge.

7.7 Efficient Simulation of the OFC Model

Let us now come back to the numerical realization of the flow chart shown in Fig. 7.7. Since the inner loop is similar to that of the BTW model, it should be treated with the help of a list as discussed in Sect. 5.2. With this simple strategy, the numerical effort for computing an avalanche is proportional to the number of blocks involved in the avalanche. In principle, this is the optimum.

In contrast, the outer loop is crucial for the model's performance on large grids. After each earthquake, the block with the maximum force must be determined in order to find out when and where the next earthquake is initiated. Implementing the search for the maximum force in the naive way leads to a numerical effort which is proportional to the total number of blocks, even for an earthquake which involves just one block. Consequently, avoiding this problem is the key to efficient simulations of the OFC model on large grids.

Grassberger (1994) introduced a scheme for determining the time and location of the next event. Let us here discuss a slightly modified version which can easily be transferred to similar models, too. The algorithm hinges

on the fact that we can compute the time $t_{i,j}$ when the block (i,j) will become unstable in advance from Eqs. 7.14 and 7.15:

$$t_{i,j} = t + 1 - F_{i,j}(t).$$

The scheme starts with computing $t_{i,j}$ for all blocks. The time axis is subdivided into a certain number of equally-sized bins $[\nu\tau, (\nu+1)\tau[$, where $\nu \geq 0$ is the number of the bin and τ the (arbitrary) bin width. All blocks are ordered into the system of bins according to their values $t_{i,j}$. But what is the advantage of this binning? Imagine that the simulation has arrived at the time t. For proceeding, we must only search the bin with the number ν where ν is the integer part of $\frac{t}{\tau}$ for the block which becomes unstable next. Only if this bin is empty, the subsequent bin must be searched, and so on. So, if the bin width τ is neither too large nor too small, only a few operations are required for determining the block which becomes unstable next.

However, the binning scheme must be permanently maintained throughout the simulation, but this is much less demanding than searching the whole lattice for each earthquake. This result arises from the fact that only the values $t_{i,j}$ of those blocks which have become unstable and of their neighbors are altered as a consequence of an earthquake. Thus, maintaining the binning requires updating the values $t_{i,j}$ of those blocks which are involved in an earthquake and their neighbors and filling them into the bins. This effort is roughly proportional to the size of the earthquake, which is mostly much smaller than the total number of blocks. Therefore, this scheme makes the numerical effort per earthquake nearly independent from the model size, except for the fact that larger grids may generate larger earthquakes.

In its present form, this scheme requires more bins than necessary. At the time t, all blocks fall into the bins with numbers between $\frac{t}{\tau}$ and $\frac{t+1}{\tau}$. Thus, $\frac{1}{\tau}$ bins are in fact necessary, but the lowest bins become empty through time, while new bins must be supplied at the other edge. Originally, Grassberger (1994) fixed this problem by writing the OFC model in a modified form, but there are several other ways which can be transferred to other models such as the two-variable model discussed in Sect. 8.4 more easily. In principle, a periodic bin structure with a period of $1+\tau$ can be used, but we can also let all those blocks where $t_{i,j}$ leaves the range of the bins fall out of the binning system. After some time, all bins have been spent, and all blocks are outside the binning system. Then, a new set of bins is built up and used until it has been spent, too. Compared to the periodic binning system, this procedure looks somewhat crude, but in fact it is advantageous: The whole algorithm works well only if the bin width τ is appropriate. If τ is too large, there are several blocks in each bin, so that too many searches must be performed within the bins. If, in contrast, τ is too small, we must search several empty bins. The highest efficiency is achieved if the number of searches within the bins and the number of searched bins are of the same order of magnitude. From this point of view, the state when all bins have been spent is a good opportunity for optimizing the bin width adaptively.

7.8 Is the OFC Model Self-Organized Critical?

Let us now investigate the OFC model with respect to SOC. The procedure is the same as that applied to the BTW model and the forest-fire model discussed in the previous chapter. A simulation is started with an arbitrary initial condition or better, with several different initial conditions. All events are skipped until the statistical distribution of the event sizes has become stationary; the following events are analyzed.

In the following, results obtained from simulations on a grid of 512×512 blocks with free, rigid-frame (non-moving), and periodic boundary conditions are analyzed. We consider the cases $k = \frac{1}{2}$ (level of conservation $= \frac{8}{9}$), $k = 1$ (level of conservation $= \frac{4}{5}$), and $k = 2$ (level of conservation $= \frac{2}{3}$). The conservative limiting case ($k \to 0$) will be discussed in Chap. 8. The initial condition consists of forces which are randomly drawn from a uniform distribution between 0 and $\frac{1}{2}$.

Figure 7.9 shows the non-cumulative number of events for $k = 1$. The size of the rupture area A is identified with the number of blocks participating in the earthquake. The analysis was started after 10^9 events; the statistics include 10^8 earthquakes each. For all three types of boundary conditions, the plots show some kind of power-law behavior. However, the results obtained under periodic boundary conditions are strange at first sight. Compared to the upper two diagrams, there is a strong variation in the number of events, but some regularity, too. The number of events of a certain size is either zero, 1098, 1099 or larger than 2000. The only reasonable explanation of this strange statistics is periodic behavior in time. Obviously, 1098 identical cycles are passed through during the simulation of 10^8 earthquakes; the number 1099 results from incomplete cycles at the beginning and the end.

The transition towards periodic behavior in coupled systems is called *synchronization*; it was first described by C. Huygens in the seventeenth century. He noted that the pendulums of two clocks hanging on a wall always ended up moving in phase and attributed this behavior to the slight coupling through the wall. Synchronization occurs in many coupled systems; a detailed discussion is provided in the book of Pikovsky et al. (2002). In the OFC model with periodic boundary conditions, the tendency to synchronize was discovered by Grassberger (1994). However, it was already observed in the original BK model (Xu and Knopoff 1994). Obviously, periodic behavior does not meet the criteria of SOC, but we should have learned that these things are not so clear in numerical models. First, periodic behavior may not be recognized if the simulations are too short. In our example, 10^8 earthquakes constitute about 1000 periods, so that the length of the period is about 10^5 earthquakes. Thus, statistics will be suspicious only if the simulation covers several periods, perhaps about 10^6 events. Even worse, we have seen when discussing strange attractors in Chap. 4 that each numerical simulation is in principle periodic. Consequently, the border between periodic and (apparently) irregular behavior is soft. Clearly, a period of about 10^5 earthquakes is

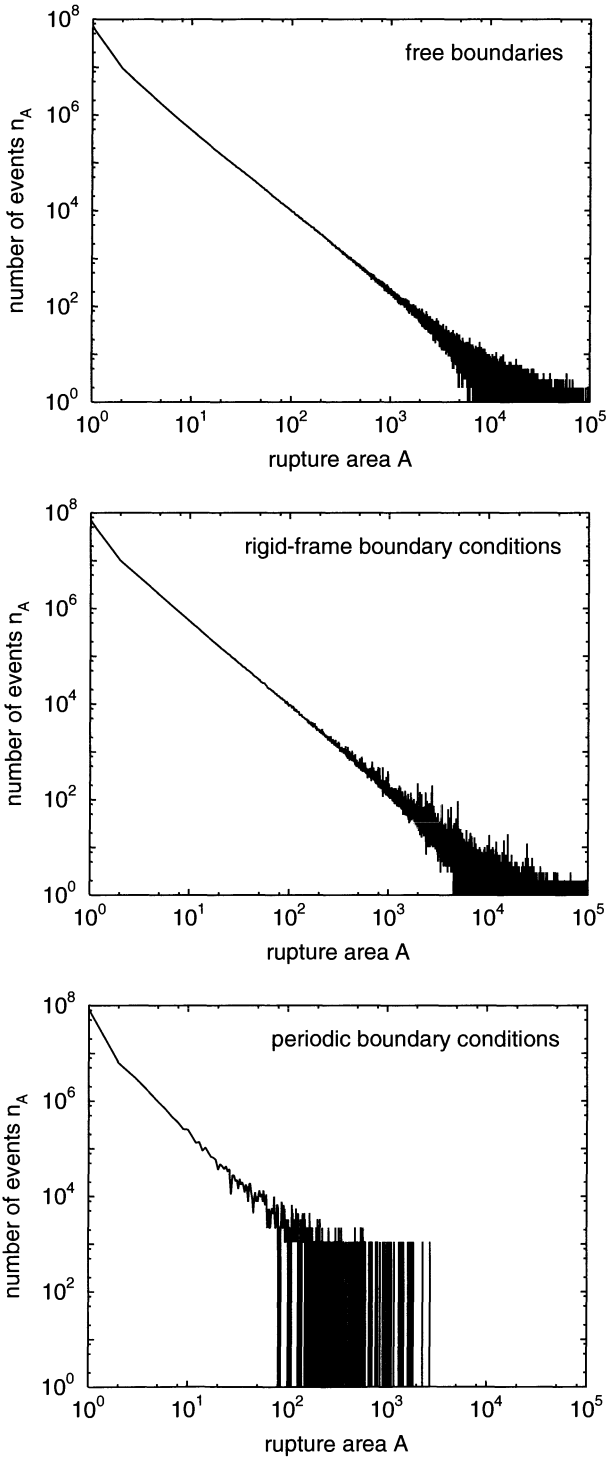

Fig. 7.9. Non-cumulative number of events for $k = 1$, obtained from analyzing 10^8 earthquakes on a 512×512 grid after the first 10^9 events were skipped.

too short; but what if the length of the period tends towards infinity in the limit of infinite system size? But after all, periodic boundary conditions are at least more complicated than expected; so let us refrain from considering them further.

Since we have seen that we can easily be trapped by an insufficient statistics, we should take a closer look at the results obtained under free and rigid-frame boundary conditions, too. Both curves exhibit similar power-law behavior. However, the distribution obtained under rigid-frame boundary conditions is less smooth than that obtained assuming free boundaries. Is this just a matter of statistics or does it hide any deeper difference? Let us, e. g., consider the spikes occurring at sizes of about 2000 blocks. The plot suggests that the average number of events of a certain size is about 20 here, while the peaks reach up to about 200 events. According to the theory of binning presented in Sect. 2.3, the observed number of events should be Gaussian-distributed with a standard deviation of about $\sqrt{20}$. Thus, the peaks deviate from the expected value by up to 40 times the standard deviation.

In general, statistical tests based on such arguments should be treated carefully since we have to decide whether an observation arises just from bad luck or gives evidence that the assumed distribution is not correct. However, this is an example where things are clear since the probability that a random number drawn from a Gaussian distribution deviates from the expected value by more than 40 times the standard deviation is smaller than the smallest floating-point number in double-precision arithmetics on a computer. Therefore, the spikes are in clear contradiction to the assumption of a power-law probability density; the underlying probability density must be spiky, too. Again, this behavior indicates some kind of periodicity. However, analyzing the next 10^8 events shows that the behavior is not strictly periodic. A similar behavior was already observed when investigating the strange attractor in the Lorenz equations (Chap. 4), but things are slightly different here. Simulations of longer sequences reveal that the regularity ceases through time. The time required until transient effects have vanished depends on both the model size and on the parameter k. For $k = 1$, 10^8 events are sufficient on a 128×128 grid, while at least 5×10^8 events should be skipped on a 256×256 lattice for getting rid of periodic components. So the number of events to be skipped strongly increases with the size of the lattice. Under this aspect, our result that skipping 10^9 events is not sufficient on a 512×512 grid is not surprising. The dependence on k is similar; the effect becomes more severe if k increases.

In summary, the OFC model with both free and rigid-frame boundary conditions evolves towards a critical state, but the critical state is approached faster with free boundaries than with rigid-frame boundaries. For free boundaries, 10^8 events are sufficient at least for $k \leq 1$ and up to model sizes of 512×512 blocks. Since free boundaries are theoretically slightly simpler than those with a frame, this is a good reason for focusing on free boundary con-

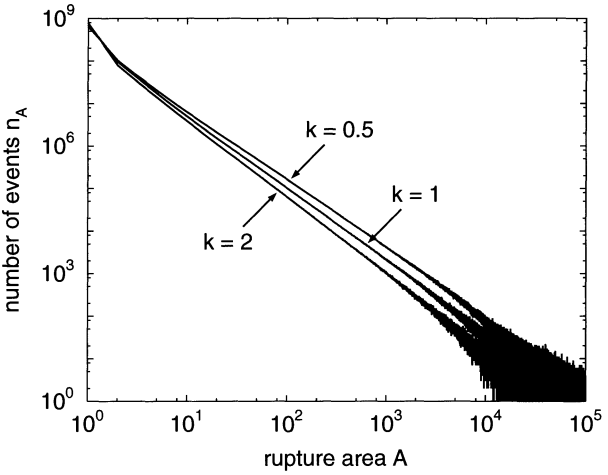

Fig. 7.10. Non-cumulative size statistics of the rupture areas (number of displaced blocks) in the OFC model with free boundaries. Each curve is the result of 10^9 events.

ditions in the following. Nevertheless, when comparing our results with the literature, we should be aware that rigid-frame boundary conditions were posed in the original version of the OFC model (Olami et al. 1992), and that the majority of authors follows this line.

Figure 7.10 gives the non-cumulative size statistics resulting from simulations with free boundary conditions. Apart from a finite-size effect being visible for $k = 0.5$, the distributions follow power laws for all considered values of the parameter k. The scaling exponent b increases with decreasing parameter k, i. e., with decreasing level of conservation. Fitting power laws as described in Sect. 2.3 yields the following values of the exponent: $b = 0.58$ for $k = 0.5$, $b = 0.68$ for $k = 1$, and $b = 0.78$ for $k = 2$. So it seems that the b-values between about 0.8 and 1.2 which are observed in nature are approached for $k > 2$, but we will see in the following section that this result should be treated carefully.

In contrast to the BTW model and the forest-fire model, the exponent b can be tuned with the help of a model parameter. This loss of universality (Sect. 6.2) was recognized by Olami, Feder, and Christensen (Olami et al. 1992; Christensen and Olami 1992; Olami and Christensen 1992); it is the most important difference towards former SOC models.

Interestingly, the OFC model was not essentially new at this time. Apart from the very similar one-dimensional cellular automaton already mentioned (Nakanishi 1990, 1991), at least two nearly identical models were published a somewhat earlier (Brown et al. 1991; Matsuzaki and Takayasu 1991). However, these authors only recognized criticality in the conservative limiting case where the model behaves similarly to the BTW model. So Olami, Feder, and Christensen were at least the first to recognize that the model becomes critical in the non-conservative case, too, and that the resulting non-universality was

the major step towards understanding the fractal character of earthquakes quantitatively.

However, it is not clear whether SOC occurs for all values of the parameter k. In the limit $k \to \infty$, the coupling between two blocks vanishes, so that they move independently of each other. Thus, only earthquakes involving a single block occur in this limiting case, so that criticality is lost then. There is still discussion whether the OFC model evolves towards a critical state for all finite values of k, i. e., if the level of conservation is larger than zero, or if the transition between SOC and non-critical behavior takes place at a certain level of conservation. A discussion of different findings is given by Jensen (1998). Again, the quasi-periodic components are the major problem in answering this question. If the level of conservation decreases, the number of events which are needed until the regularity vanishes increases rapidly, so that simulations at a low level of conservation become costly.

The discussion on SOC in the non-conservative OFC model has recently been renewed. In Sect. 5.2 we have already recognized a fundamental problem concerning SOC in numerical models. We can perform simulations on lattices of different sizes, but are finally free to believe in a convergence towards a power law in the limit of infinite system size or not. In a recent study, de Carvalho and Prado (2000) analyzed another statistical property in addition to the event-size distribution – the *branching rate*. The idea of branching can be applied to all models where some kind of instability propagates from one site to another; it was introduced in the context of SOC by Zapperi et al. (1995). The branching rate is the mean number of sites which are destabilized by an unstable site. The idea hinges on the assumption that the branching rate is the same throughout the whole avalanche process. Under this assumption, it can be shown that a power-law distribution of the event sizes can only occur if the branching rate is unity (Sect. 5.1). The analysis revealed that the branching rate is lower than unity in the non-conservative OFC model. The conclusion is that the OFC model is SOC only in the conservative limiting case, but that its behavior is close to SOC in the non-conservative regime. This behavior is described by the terms *almost critical* or *quasi-critical* (Kinouchi and Prado 1999). However, the analysis crucially depends on the assumption that the system can be mapped on a branching process, and the difference between criticality and quasi-criticality is still a matter of controversy (Christensen et al. 2001; de Carvalho and Prado 2001).

7.9 Rupture Area and Seismic Moment

In the previous section, we used the size of the rupture area as a measure of earthquake strength. The distributions obtained from the model were compared with Eq. 7.5. This distribution is not directly observed in nature, but is derived from the GR law which itself is observed in nature. However, the derivation is mainly based on a relationship between the seismic moment M

and the size of the rupture area A (Eq. 7.4): $M \sim A^{\frac{3}{2}}$. Thus, we should at least check whether the OFC model is able to reproduce this relationship before deciding which value of k is appropriate with respect to the GR law.

Let Q be the set of all sites (i,j) which are displaced during an earthquake. From Eq. 7.13 we see that each block is displaced by an amount $\frac{F_{i,j}}{4+k}$, where $F_{i,j}$ is the force acting on the block immediately before it becomes unstable. The seismic moment was defined in Eq. 7.1 to be the integral of these displacements over the entire rupture area, multiplied with the shear modulus of the rock. Since the latter is constant, we neglect it when defining a non-dimensional moment; so let us just add the displacements of all blocks involved in the earthquake:

$$M := \sum_{(i,j)\in Q} \frac{F_{i,j}}{4+k}.$$

Strictly speaking, there are two minor flaws in this definition: First, the displacement slightly differs at free boundaries. However, since the number of boundary sites is small compared to the number of sites in the bulk for large grids, we neglect this discrepancy for convenience. But more important, one block may be displaced several times during an earthquake. In this case, each displacement contributes to the seismic moment. This problem can be avoided by including these blocks more than once into the set Q, although Q is no longer a set in the mathematical sense then.

Figure 7.11 gives the relation between seismic moment and rupture area in the OFC model. The data were obtained from the simulations already used for computing the size distributions of the rupture areas. Events with equally sized rupture areas may have different seismic moments, except for those involving just one block. Thus, the relationship between moment and rupture area can only be of statistical character; it describes the average moment of

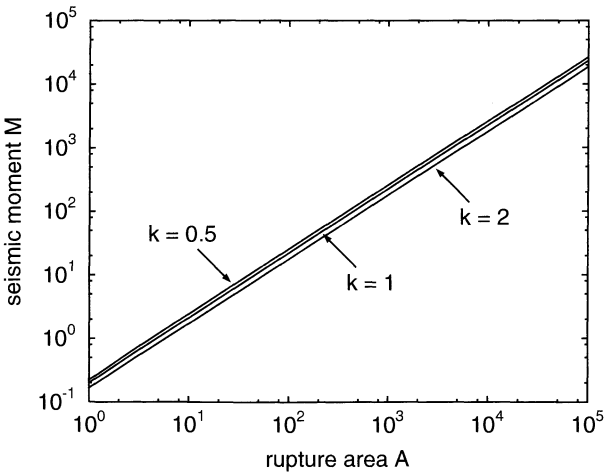

Fig. 7.11. Relation between seismic moment and rupture area in the OFC model for different values of k.

all earthquakes with a given rupture area. This average is plotted in Fig. 7.11; for large earthquakes where the statistics are small, several rupture areas are pooled in bins of at least 10^4 events.

The plot reveals a weakness of the OFC model. In contradiction to the relation $M \sim A^{\frac{3}{2}}$, the seismic moment increases linearly with the size of the rupture area. However, the established relationship is not universal in fact; e. g., Sammis et al. (1999) found that the mean displacement only increases as $M^{0.17}$ for repeating earthquakes on the San Andreas fault in central California. Inserting this result into Eq. 7.2 leads to $M \sim A^{1.2}$; so the exponent in the area-moment relation may vary. Nevertheless, the results of the OFC model suggest $M \sim A$, which means that the mean displacement of all earthquakes is the same. This result is quite unrealistic.

Obviously, the linear relation between moment and rupture area shows that multiple relaxations are rare if they occur at all. In most cases, each block involved in an earthquake is displaced only once. But where does this problem come from? Is it already present in the original BK spring-block model, or did we raise it when switching to the OFC cellular automaton? Let us start at the forces (Eq. 7.12) without assuming anything about the process of relaxation. If we add the forces acting on all blocks, the contributions of the horizontal springs compensate each other in sum, so that

$$\sum_{i,j} F_{i,j} \;=\; \sum_{i,j} (t - k\, u_{i,j})\,.$$

Let us now assume that the blocks are displaced to new positions $\tilde{u}_{i,j}$, where $\tilde{u}_{i,j} \coloneqq u_{i,j}$ for all blocks which do not participate in the earthquake. Clearly, the forces $\tilde{F}_{i,j}$ after the earthquake must satisfy the same relationship

$$\sum_{i,j} \tilde{F}_{i,j} \;=\; \sum_{i,j} (t - k\, \tilde{u}_{i,j})\,,$$

so that

$$M \;=\; \sum_{i,j} (\tilde{u}_{i,j} - u_{i,j}) \;=\; \frac{1}{k} \sum_{i,j} \left(F_{i,j} - \tilde{F}_{i,j}\right).$$

The forces acting on those blocks which remain stable either increase as a result of an earthquake (one of the neighbors becomes unstable) or remain unaffected, so that $F_{i,j} - \tilde{F}_{i,j} \leq 0$ for these blocks. In contrast, $F_{i,j} \leq 1$ and $|\tilde{F}_{i,j}| \leq 1$ for those blocks which are displaced, so that $F_{i,j} - \tilde{F}_{i,j} \leq 2$ here. Since the rupture area A is just the number of displaced blocks, we immediately obtain

$$M \;\leq\; \frac{2A}{k}.$$

Consequently, the seismic moment cannot grow stronger than linearly with the rupture area in the limit of large earthquake sizes. So this weakness is not a property of the realization of the spring-block model in form of a cellular

automaton, but already present in the original model. More precisely, it arises from the coupling of the blocks with the upper rigid plate since the limitation of the moment becomes less strict if the coupling strength k decreases. Finally, the limitation vanishes in the conservative limiting case ($k \to 0$), so that a spring-block model may reproduce a realistic area-moment relation only in the conservative limiting case.

Even if we can live with this problem, it affects at least the choice of the parameter k with respect to the GR law. The size distribution of the rupture areas (Eq. 7.5) was derived from the frequency-magnitude relation of the seismic moments (Eq. 7.3) assuming $M \sim A^{\frac{3}{2}}$. Hence we must decide whether we want to reproduce the distribution of the seismic moments or that of the rupture areas. Since the seismic moment is more directly related to the observed earthquake properties than the rupture area, it is preferable. As we have observed $M \sim A$ in the OFC model, rupture areas should still obey a fractal size distribution, but with an exponent b that differs from the b-value of the GR law. From Eq. 7.3 we see that the exponent b is two thirds of the b-value of the GR law, so that $b \approx \frac{2}{3}$. In this sense, $k = 1$ is a quite realistic parameter value as it leads to $b \approx 0.68$ under free boundary conditions.

7.10 The Temporal Fingerprint of the OFC Model

In Sect. 5.4 we suggested to consider different levels of SOC, depending on the temporal characteristics of the system in addition to the distribution of the event sizes. So let us first define what the output of the model shall be. The amount of seismic moment released per time could, e. g., be a meaningful definition of the output with respect to the seismic properties measured at the earth's surface. However, we have learned in the previous section that the seismic moment increases linearly with the size of the rupture area; so let us, for convenience, characterize earthquakes by the sizes of their rupture areas.

Since the time scales of earthquakes and motion of the plates are decoupled in the OFC model, the duration of each earthquake is zero. Consequently, a time series $f(t)$ composed of the sizes of the rupture areas is zero nearly all the time. This problem can be avoided by dividing the time axis into intervals and summarize all the earthquakes occurring within the same interval. This value is divided by the length of the intervals in order to obtain a rate (sum of rupture areas per time). In the limit of zero interval length, the output can be expressed with the help of Dirac's delta function (p. 45). Let us consider a (sufficiently long) time span from 0 to T containing $n \gg 1$ earthquakes of sizes $A_1, \ldots, A_n$ occurring at the times $t_1, \ldots, t_n$. Then, the output is

$$f(t) \;=\; \sum_{i=1}^{n} A_i \, \delta(t - t_i).$$

Figure 7.12 provides an example, obtained for $k = 1$. The three diagrams represent the same data set, but refer to different time scales. The longest

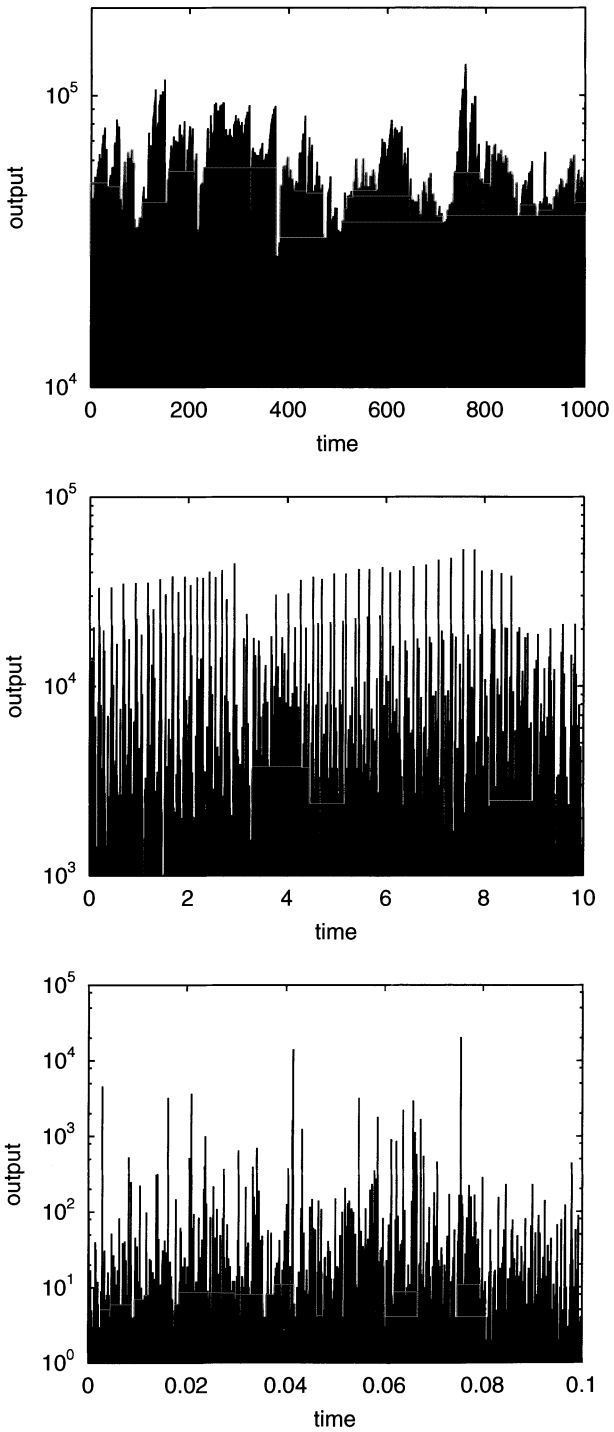

Fig. 7.12. Output of the OFC model with free boundaries, obtained for $k = 1$ in the quasi-steady state.

time interval of length 1000 covers about 5×10^7 events. Strictly speaking, the impulses do not represent the output, but only the factors in front of Dirac's delta function.

Let us now quantify the temporal characteristics of the OFC model with the help of variograms (Sect. 3.9). Since $f(t)$ is not a function in the strict sense, its variogram cannot be computed directly. However, considering the variogram of the cumulated (integrated) signal $f_c(t)$ avoids this problem. As discussed in Sect. 3.9, the expected value $\overline{f}$ must be estimated and removed from the time series, so that the output has to be replaced with

$$f(t) \;:=\; f(t) - \overline{f} \;=\; f(t) - \frac{1}{T}\int_0^T f(t)\,dt \;=\; \sum_{i=1}^{n} A_i\,\delta(t-t_i) - \frac{1}{T}\sum_{i=1}^{n} A_i.$$

The variogram of the cumulated output can be computed from Eq. 3.16 which turns into

$$\Gamma_c(\tau) \;=\; \frac{1}{T-\tau}\int_0^{T-\tau}\left(\sum_{\substack{i=1\\ t\le t_i<t+\tau}}^{n} A_i - \frac{\tau}{T}\sum_{i=1}^{n} A_i\right)^2 dt.$$

The variogram of the simulation with $k = 1$ is plotted in Fig. 7.13. Up to time lags τ of about 0.1 (corresponding to about 5000 earthquakes), the variogram follows a power law with an exponent of one. This is exactly the result obtained from the BTW model; the cumulated output can be described by Brownian motion. So the non-cumulative output is (non-Gaussian) white noise; we do not need to compute its variogram explicitly.

If we follow the idea of three different levels of SOC introduced in Sect. 5.4, the OFC model is on the intermediate level. In analogy to the BTW model, the output of the OFC model exhibits self-affine scaling properties to some

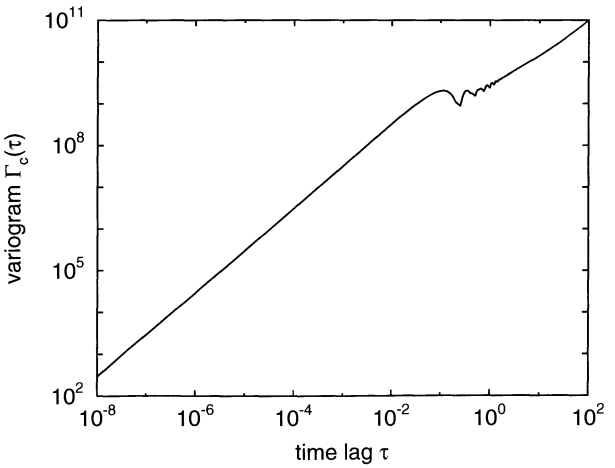

Fig. 7.13. Variogram of the cumulated output for $k = 1$.

extent, but seems to be uncorrelated, white noise instead of pink noise as requested by the original definition of SOC.

The preliminary result that earthquakes are completely uncorrelated in the OFC model is in clear contradiction to the observed properties of real earthquakes. In nature, earthquakes show a more complex behavior; the most important deviations from white noise are the following:

- As already mentioned in Sect. 7.1, there is often a correlation between large earthquakes (seismic cycles).
- Earthquakes exhibit some *spatio-temporal clustering* on the short time scale. Many large earthquakes are accompanied by a certain pattern which begins with a period of *seismic quiescence* and turns into an increasing rate of mainly small events, called *foreshocks*, until the *mainshock* occurs. Then, the most characteristic part of the temporal pattern follows; seismic activity ceases through an often large number of *aftershocks*.

Figure 7.13 suggests that the temporal behavior of the OFC model may in fact be more complex than just white noise. The observed self-affine scaling behavior breaks down at $\tau \approx 0.1$, and this breakdown looks somewhat strange. Local minima occur at time lags $\tau = 0.25$, 0.5, 0.75, 1, 1.25, and so on; they indicate some regularity with a period length of 0.25. If we take a look again at Fig. 7.12, we recognize this regularity in the diagram in the middle; there are sequences of peaks with a time lag of 0.25. If we remember that $\tau = 1$ is the maximum time span until a completely relaxed block becomes unstable, $\tau = 0.25$ is a rather long time.

Thus, the OFC model shows some kind of seismic cycles. One may suspect that these seismic cycles could be a transient effect of the initial conditions which vanishes through time. However, simulations with skipping more than 10^9 events in the beginning show that the results persist as well as all the results presented in the following.

The occurrence of seismic cycles was one of the first results obtained from the OFC model (Olami and Christensen 1992). It was originally recognized by analyzing the recurrence times of large earthquakes. Obviously, the time τ_r between two consecutive earthquakes is a statistical property. What kind of distribution shall be expected for τ_r? Let $P(\tau_r)$ be the cumulative distribution of the recurrence times, i. e., the probability that there is quiescence within a randomly chosen time interval $[t, t+\tau_r[$ of length τ_r. If earthquakes are entirely uncorrelated, $P(\tau_r)$ must satisfy the relationship

$$P(\tau_r+\tau_r') \;=\; P(\tau_r)\,P(\tau_r').$$

The only function which meets this criterion is $P(\tau_r) = \mathrm{e}^{-r\tau_r}$, where r is an arbitrary constant. Thus, the cumulative distribution of the recurrence times decreases exponentially, provided that earthquakes are uncorrelated.

Let us begin with the recurrence times of large earthquakes, say, those events which involve at least 30,000 blocks. The simulation of 10^9 events

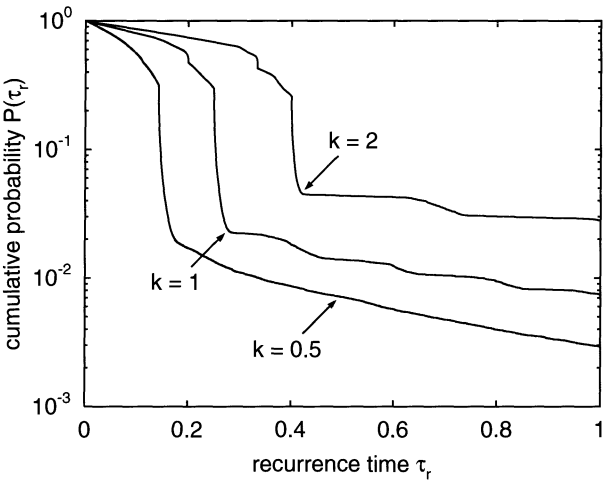

Fig. 7.14. Cumulative distribution of the recurrence times for earthquakes with $A \geq 30,000$.

contains about 10^5 earthquakes of this category. Figure 7.14 illustrates that the resulting distribution of the recurrence times strongly differs from an exponential distribution which would be a straight line in the half-logarithmic diagram. Thus, large earthquakes do not occur randomly. The step-like shape of the cumulative distribution implies that recurrence times in a certain interval are strongly preferred. As already suggested by the variogram (Fig. 7.13), the preferred recurrence time is $\tau_r \approx 0.25$ for $k = 1$. The distribution shows that the recurrence time of these large earthquakes is between 0.2499 and 0.2538 with a probability of 20 %. If we, e. g., identify the preferred recurrence time with 100 years, we can predict the subsequent large earthquake with an accuracy of one and a half year, and we will be right with a probability of 20 %. In reality, a long-term prediction of this quality would be great. However, we must not forget that the OFC model is an idealized approach; it is based on very simple physical assumptions and does not include any kind of inhomogeneity that is ubiquitous in geology. Under this aspect, the result obtained here may define an upper limit of the long-term predictability of earthquakes.

One may be surprised at first sight that the observed regularity does not affect the size distribution of the earthquakes (Fig. 7.10); we have at least explained the spikes occurring under rigid-frame and periodic boundary conditions by periodic components in the signal. Figure 7.12 shows the reason for this result; the corresponding large earthquakes are nearly periodic in time, but of different sizes.

Let us now focus on shorter time scales. As discussed earlier, the variogram (Fig. 7.13) suggests that earthquakes are entirely uncorrelated on the short time scale. So it seems that the OFC model cannot reproduce the spatio-temporal clustering observed in nature, consisting of foreshocks, aftershocks, and seismic quiescence. Several extensions of the OFC model were suggested

in order to obtain a more realistic behavior; a review is given by Pelletier (2000). Hainzl et al. (1999) introduced a viscous crust relaxation process; their results show a quite realistic short-term clustering. However, let us refrain from going deeper into the physical details of these more complicated models, but re-analyze the results of the OFC model.

Figure 7.15 shows the distribution of the recurrence times for earthquakes of different classes. The upper diagram refers to all earthquakes, no matter whether large or small. The graphs in the half-logarithmic plots exhibit a convex curvature, which means that short recurrence times are preferred. In other words, earthquakes occur preferably in swarms. However, this analysis does not reveal any information on the character of this clustering. The swarms may be series of foreshocks or aftershocks related to large earthquakes, but may also consist of events of similar sizes. The effect of temporal clustering becomes stronger if k increases. The lower diagrams show that the effect vanishes for earthquakes with intermediate sizes.

The statistics of *aftershocks* have been investigated since the late nineteenth century (Omori 1894). Omori's empirical law (e. g. Utsu 1961) states that the number of aftershocks per time (regardless of their size) decreases like $(t-t_{\mathrm{m}})^{-p}$, where t_{m} is the time where the mainshock occurred. The exponent p was found to be close to unity.

Obviously, identifying aftershocks requires the definition and the recognition of mainshocks first. A mainshock is an earthquake of considerable size which is the largest event within a given time interval. Unfortunately, there is not a straightforward, unique criterion for the minimum size of a mainshock and the length of the time interval, so that every quantitative definition is arbitrary to some degree. Let us define the minimum size (rupture area) of a mainshock to be $A = 1000$. Then, the simulation of 10^9 events includes about 3×10^6 potential mainshocks for $k = 1$. Clearly, the dynamics of aftershocks should take place on a considerably smaller time scale than the mean recurrence time of the mainshocks which is about 7×10^{-3} for $k = 1$. So let us define that a mainshock is an earthquake with $A \geq 1000$ which is the largest event within the interval $[t_{\mathrm{m}}-10^{-4}, t_{\mathrm{m}}+10^{-4}]$.

The further analysis is straightforward: After each mainshock, the number of earthquakes occurring within the intervals $[t_{\mathrm{m}}, t_{\mathrm{m}}+\delta t[$, $[t_{\mathrm{m}}+\delta t, t_{\mathrm{m}}+2\delta t[$, $[t_{\mathrm{m}}+2\delta t, t_{\mathrm{m}}+3\delta t[$, and so on, is recorded, where $\delta t \ll 10^{-4}$, but large enough to guarantee a sufficient number of events per bin. Let us choose $\delta t = 2\times10^{-9}$ in the following. If we again identify a time interval of length 0.25 with 100 years, δt is about 25 seconds. The results of all mainshocks are stacked, so that the first bin contains all events which occur within a time interval of length δt after any mainshock, and so on.

Figure 7.16 shows the result of this analysis. Obviously, the seismic activity is high immediately after mainshocks and decreases through time. This behavior can be interpreted in terms of aftershocks, although all graphs exhibit some convex curvature that seems not to be consistent with Omori's

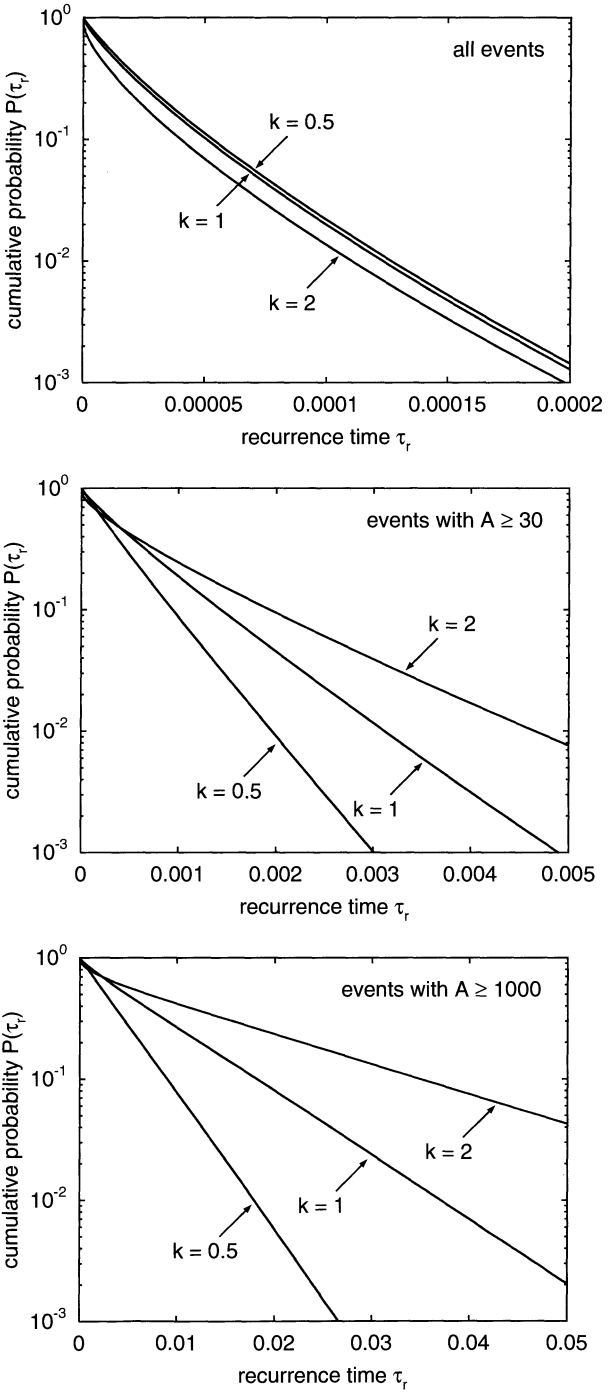

Fig. 7.15. Cumulative distribution of the recurrence times for different classes of earthquakes.

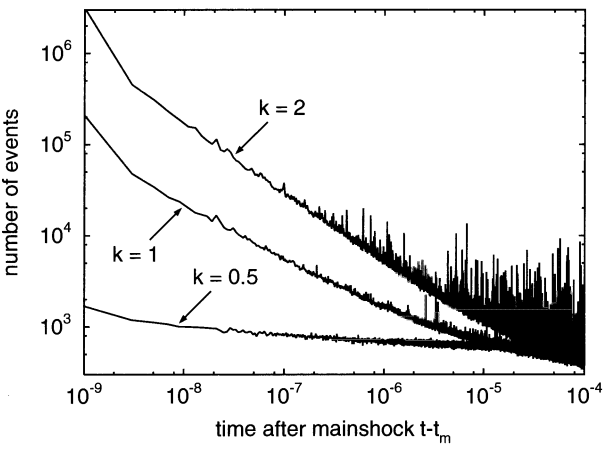

Fig. 7.16. Number of earthquakes within certain time intervals after mainshocks.

law. However, at this point we should be aware that our analysis does not distinguish aftershocks uniquely. In reality, aftershocks are not only temporally assigned to the mainshock, but spatially, too. Consequently, we record both aftershocks and the background seismicity which is not necessarily related to any mainshock, so that we should remove the latter from the analysis in order to recognize Omori's law. In a first step, the results shown in Fig. 7.16 must be transformed into a rate $R_a(t-t_m)$ (events per time). We then consider the quantity

$$A_a(t-t_m) \;:=\; \frac{R_a(t-t_m)-R}{R},$$

where R is the average activity, measured throughout the whole simulation. This quantity can be considered as the relative excess activity; e. g., a relative excess activity of one means that the total activity is twice the average activity.

The relative excess activity is plotted in Fig. 7.17. Shortly after the mainshocks, the seismic activity exceeds the average activity by nearly a factor 1000 for $k = 1$ and even more for $k = 2$. Compared to the raw data shown in Fig. 7.16, the relative excess activity comes closer to a power law as predicted by Omori's law. For estimating the exponents p, we use the range between $t-t_m = 10^{-8}$ and $t-t_m = 10^{-6}$. We obtain $p = 0.15$ for $k = \frac{1}{2}$, $p = 0.62$ for $k = 1$, and $p = 0.76$ for $k = 2$.

Thus, even the original OFC model reproduces Omori's law at least qualitatively. In analogy to the exponent b of the GR law, the exponent p depends on the parameter k, i. e., on the level of conservation. The data suggest that the aftershock activity vanishes in the conservative limiting case ($k \to 0$). However, two concerns may be raised against our findings: First, the exponents p are lower than observed in nature. In other words, the decay of aftershock activity predicted by the OFC model is too slow. Realistic values $p \approx 1$ can only be achieved for $k > 2$, while realistic exponents b are

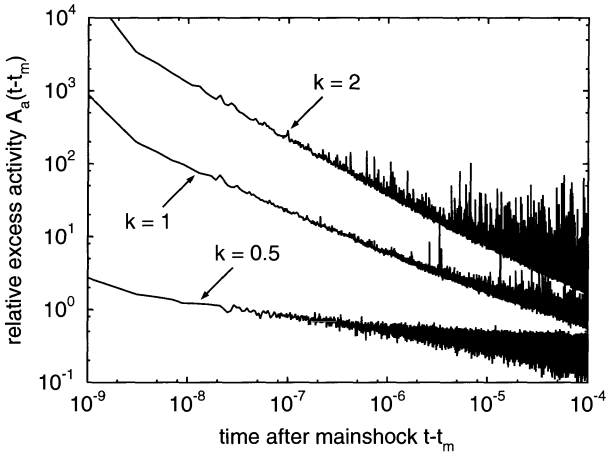

Fig. 7.17. Relative excess activity after mainshocks.

achieved for $k \approx 1$. The second objection concerns the physical basis of the OFC model. The aftershocks occurring in the OFC model are initiated in the same way as all other earthquakes – by stress transfer through the upper place. In contrast, other explanations of aftershocks hinge on different mechanisms, e. g., fluids in the crust or changes in the threshold of failure as a result of a large earthquake. So there is a tendency to believe that only these earthquakes which would take place if stress transfer ceases, i. e., if the upper plate stops, are real aftershocks. As a consequence, it is different to tell whether the discovery of aftershocks in the OFC model is in fact a new result or whether someone found it earlier, but did not succeed in publishing this result in the geophysics literature, too.

The analysis applied to the aftershock activity can be transferred to the seismic activity before large earthquakes, too. Figure 7.18 shows the activity

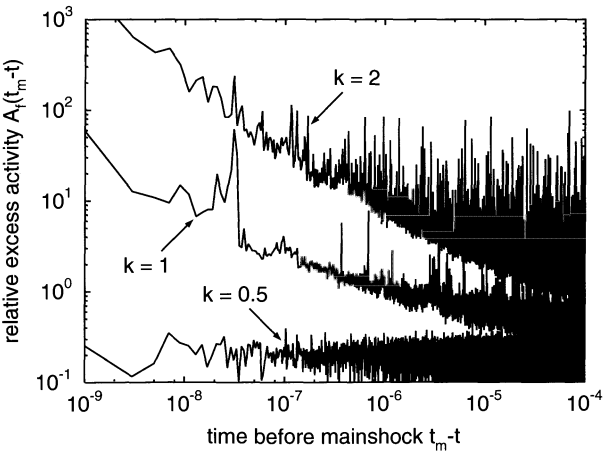

Fig. 7.18. Relative excess activity within certain time intervals before mainshocks.

of *foreshocks* $A_f(t_m - t)$. Compared to the activity of aftershocks, $A_f(t_m - t)$ is about one order of magnitude smaller. Since the data are somewhat noisy, determining the exponents is more difficult than in the aftershock statistics, but they roughly coincide with the exponents of the aftershock sequences for identical values of k. Both findings are in agreement with empirical results (Papazachos 1975; Kagan and Knopoff 1978; Jones and Molnar 1979) and results obtained from the more complicated model of Hainzl et al. (1999). However, reasons why these events are not foreshocks may still be found.

Finally, the reader may wonder about the several spikes in the plots of the foreshock and aftershock activities which are clearly outside the statistical variation. These spikes reflect a well-known phenomenon – *secondary aftershocks*. Aftershocks are not necessarily small, and if they are large, they may be accompanied by further series of aftershocks. However, secondary aftershocks are just one side of the medal. Each foreshock or aftershock may be accompanied by both foreshocks and aftershocks. Since the rate of foreshocks is lower than the rate of aftershocks in the mean, the spikes in the foreshock activity are even stronger than in the aftershock activity.

7.11 How Complex is the OFC Model?

The results of this chapter show that spring-block models, even the simple OFC model, are powerful tools for understanding the dynamics of earthquakes. The OFC model accounts for the GR law, although it fails to explain the empirical relationship between rupture area and seismic moment. At first sight, the temporal signature of the model seems not to be more interesting than that of the BTW model (Chap. 5). Instead of pink noise as required by the original definition of SOC, it is white noise. However, the more thorough analysis presented in the previous section has revealed a more complex behavior: The scale-invariant behavior concerning time breaks down at both small and large time scales, and the OFC model yields both series of fore- and aftershocks and seismic cycles in an at least qualitatively reasonable way.

At this point, the question how far we can go arises. Figure 7.19 shows three sequences of events around large earthquakes in the OFC model. The first diagram represents a large earthquake without any foreshocks and aftershocks. In the middle, the mainshock comes as a surprise, too, but is followed by a large number of aftershocks which decay concerning both frequency and strength. Finally, the lower diagram presents a sequence consisting of both foreshocks and aftershocks. This example indicates a high degree of clustering: The window of length 4×10^{-6} contains more than 2500 events, while the average recurrence time of all events is about 2×10^{-5}, i. e., 20 times larger than the window size. Again, all these types of sequences are observed in nature. So, no matter whether we believe that a model of this simplicity can be realistic, the OFC model accounts at least for a large part of the enormous complexity observed in real earthquakes.

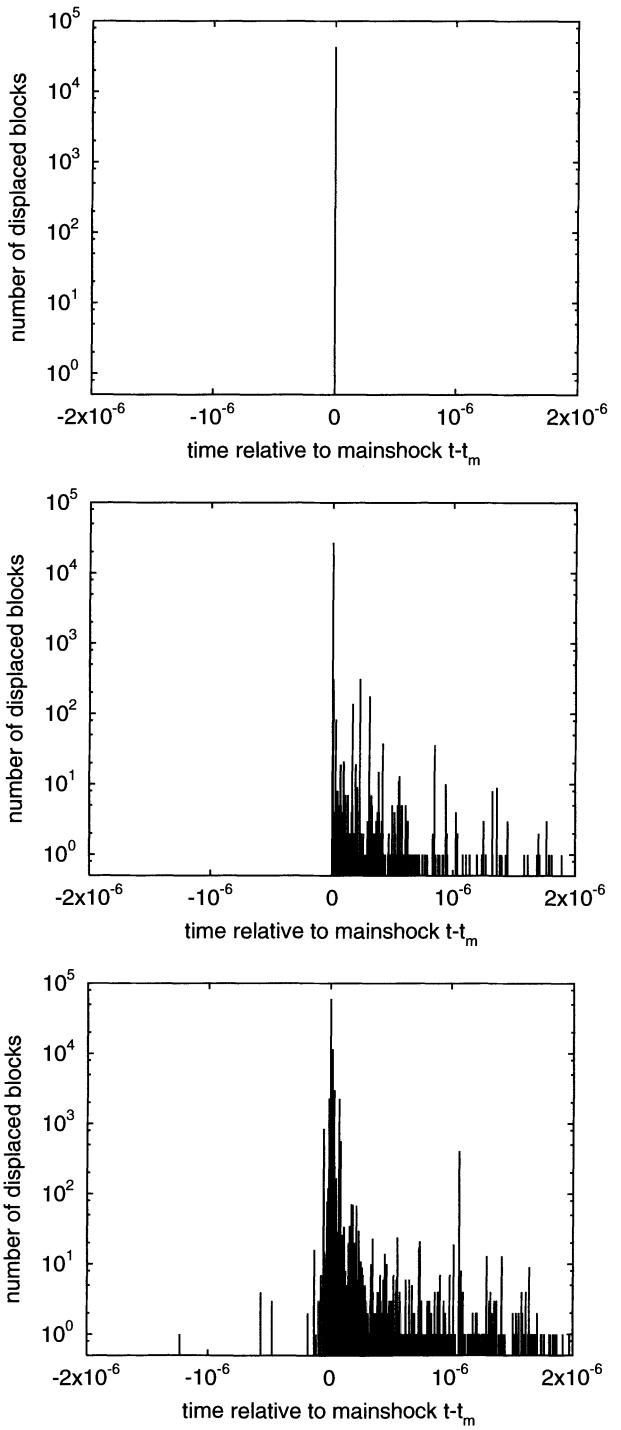

Fig. 7.19. Some sequences of earthquakes preceding and following large earthquakes, obtained for $k = 1$ on a 512×512 grid.

8. Landslides

Earthquakes, floods, storms, volcanic eruptions, and landslides are the major natural hazards on earth. Concerning death toll, earthquakes must be ranked first; however, this does not mean that earthquakes are in principle more dangerous than landslides. The difference in hazard arises mainly from the fact that the seismic waves released by an earthquake cause damage on a regional scale, while the impact of landslides is often limited to smaller areas. In mountainous regions, landslides that are able to wipe out whole villages occur quite frequently; but fortunately, only few of them affect civilization. In contrast, nearly each large earthquake occurring in populated regions causes damage. In some cases, even a clear separation of disasters is impossible; earthquakes in mountainous regions often trigger hundreds of landslides which may in sum be as disastrous as the earthquake itself.

There have been disastrous landslides killing thousands of people. Sometimes the danger is not just being buried by rock or mud, but arises from secondary effects. The Vaiont reservoir disaster (e. g. Kiersch 1964) in the Dolomite Region of the Italian Alps is one of the most terrible examples. In 1963, a block of more than a quarter cubic kilometer detached from one wall and slid into the lake at velocities of up to 30 meters per second. As a result, a wave of water overtopped the dam and swept onto the valley below, with the loss of about 2500 lives.

Strictly speaking, the term landslides is too narrow; *gravity-driven mass movements* is the established, more general term. Mass movements are classified into several groups; the coarsest scheme concerns falls, slides, and flows. These groups can be further subdivided according to the material (rock, soil) and to the role of water. Let us not go into detail, but just review the basic properties of these three classes. Rockfalls occur on extremely steep slopes; the material moves downslope at high velocities and temporarily loses contact to the ground. Slides and flows are often slower, but may cover a wide range of velocities from less than one millimeter per month up to several meters per second. In a slide, deformation is concentrated along a distinct slip surface, while a flow involves strong deformation within the entire moving mass. Nevertheless, we will use the term landslides as a synonym for gravity-driven mass movements in the following.

Since the 1940's, a large effort has been spent on developing landslide models. As a result, engineers can revert to a variety of tools from simple equations to comprehensive, three-dimensional numerical models. Most of these approaches aim at either assessing the static stability of slopes or at predicting the runout of fast mass movements. Modeling slope stability mainly relates the driving forces at a potential area of failure, e. g., the edge of a clay layer, to the maximum retaining forces. By considering several areas of different geometries, the most probable area of failure is determined. However, none of the established slope stability models accounts for scale-invariant properties of landslides which will be the main topic of this chapter. So let us refrain from discussing the established approaches and recommend, e. g., the books on slope stability of Brunsden and Prior (1984) and Bromhead (1992), respectively, the overviews concerning debris flows given by Iverson (1997) and by Jan and Shen (1997).

Landslides are mainly caused by heavy rainfall or by earthquakes; the impact of both water and earthquakes on slope stability is obvious. Earthquakes introduce a strong acceleration and thus amplify the driving forces. The influence of increasing soil water content is twofold: First, the driving force in downslope direction increases due to the weight of the water. But still more important, the pore water pressure at a potential slip surface often reduces the shear strength and thus the maximum retaining forces drastically.

However, this is only half of the truth since all gravity-driven mass movements are dissipative phenomena. During the motion, kinetic energy is dissipated, so that potential energy is spent. Thus, the average height of the material involved in a landslide decreases, although parts of it may even move upslope. As a consequence of the dissipative character, landsliding activity ceases through time in absence of processes which supply potential energy. These processes may be tectonic uplift on the large scale or fluvial erosion which steepens the slopes from their toes. Therefore, any reasonable model for predicting the landsliding activity over long times should include a long-term driving process.

8.1 Fractal Properties of Landslides

Several studies on scale invariance in landslides addressing different fractal properties have been carried out. Box-counting analyses (Sect. 1.2) were performed by Goltz (1996) for landslides in Japan, finally referring to multifractal properties (e. g. Mandelbrot 1985; Feder 1988; Turcotte 1997). Yokoi et al. (1995) addressed a much smaller scale by analyzing the internal structure of individual landslides.

However, the majority of these studies concerns frequency-magnitude relations (Sect. 7.1), mainly for the size of the areas affected by landslides in certain regions. In a quite comprehensive study (Hovius et al. 1997), 7691

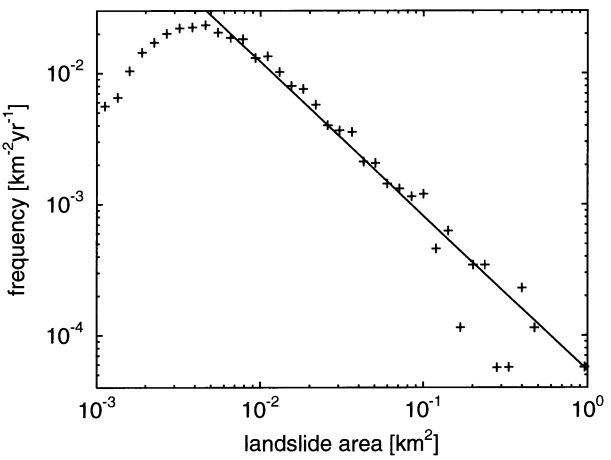

Fig. 8.1. Frequency-magnitude relation obtained from landslide mapping in the central western Southern Alps of New Zealand (Hovius et al. 1997). The non-cumulative were binned logarithmically (Sect. 2.3). The straight line shows a power law with an exponent of 1.16.

landslides in the western Southern Alps of New Zealand were mapped. Figure 8.1 shows the resulting frequency-magnitude relation, obtained from those 4984 landslides located in the montane zone. The data suggest a power-law decrease of the logarithmically binned data with an exponent of about 1.16. Since the exponent obtained under logarithmic binning coincides with the exponent of the cumulative size distribution for power-law distributions (Sect. 2.3), the data suggest a fractal size distribution of the areas with an exponent $b = 1.16$ in the cumulative sense. A similar study performed in Taiwan (Hovius et al. 2000) resulted in a fractal distribution with $b = 0.70$.

Later, the same authors (Stark and Hovius 2001) investigated the effect of censoring (Sect. 2.4) in the process of observation. As a result, they corrected the exponents to $b = 1.46$ (New Zealand) respectively $b = 1.11$ (Taiwan). However, the discussion in Sect. 2.4 has shown that the physical basis of their method is not clear, and that it may be biased as well as their original analysis. Thus, the modified exponents should not be overinterpreted; and the discrepancy shows that the uncertainty in determining the exponents is still large. Under this aspect, their results are in agreement with earlier studies. Fuyii (1969) found a power-law distribution with $b = 0.96$ in 650 events induced by heavy rainfall in Japan. Even measuring the areas of landslide deposits instead of those of the landslide scars leads to similar results; Sugai et al. (1994) found a power-law distribution with $b \approx 1$. Whitehouse and Griffiths (1983) obtained a power-law distribution for the volumes of rock avalanche deposits, too.

However, the value $b \approx 1$ may not be as universal as it may seem. Pelletier et al. (1997) compiled and analyzed landslide data from Japan (Ohmori and Sugai 1995), California (Harp and Jibson 1995, 1996), and Bolivia. They obtained power-law distributions over a quite narrow range (not much more than one order of magnitude in area) with exponents b between 1.6 and 2. Again, the quality of the power-laws is not sufficient for determining the

exponent precisely; e. g., the exponent from the California data was estimated to $b = 1.6$ first, but later to $b = 1.3$ (Malamud et al. 2001). Recently, rather comprehensive analyses of 4369 (recent) and 21,433 (older) landslides in Italy resulted in exponents of about 1.5 (Malamud et al. 2001).

In summary, there is evidence for fractal statistics in landslides. Nevertheless, being cautious concerning this result is not a mistake. The available documentation of landslides cannot compete with the extensive earthquake catalogs. In all studies mentioned above, both the number of events and the range where the distribution looks like a power-law is low compared to earthquake statistics. The danger is obvious: Fractals are sold much better than doubts against scale invariance. Who knows how many data sets suggesting non-fractal statistics are rotting away since they seem not to be interesting? From this point of view, modeling landslide statistics may be even more important and interesting than modeling earthquake statistics. In the latter case, modeling is always one step behind the data obtained from nature. Available data from seismology provide several criteria for assessing and comparing different models. These criteria are hardly met by recent models; so they drive model development and deepen our understanding of the process. But on the other hand, they prevent the models from providing reliable information which goes beyond the available data. In contrast, the sparse knowledge on scale-invariant properties of landslides can perhaps be considerably enlarged by physically based models.

Since landslide dynamics is determined by a variety of processes, there are several candidates for being responsible for the fractal size distribution of landslides. However, it should be clear from the title of this book that we will search for SOC in a landform evolution process in the following, perhaps in analogy to sandpiles. Nevertheless, we must take care that we are not too blind and run into a dead-end street; we should at least think about alternative theories which might lead to fractal size distributions.

The distribution of the slope sizes in the landscape is not a good candidate. On each slope, mass movements with a wide range of sizes occur; there is not a characteristic landslide size for a slope. In contrast, assuming that the size distribution of landslides is governed by the size distribution of the triggering events may be more reasonable. As discussed in Chap. 7, earthquakes obey a fractal size distribution. In hydrology, the ideas of multifractals (Mandelbrot 1985; Feder 1988; Turcotte 1997) have been applied to heavy rainfall events; precipitation dynamics turned out to be heavy-tailed, which means that the size distribution of large events follows a power law. However, there are two arguments against this theory: First, the size distribution of the landslides seems to be mainly independent from the triggering mechanism; there is not a significant difference between rainfall-triggered and earthquake-triggered landslides (Pelletier et al. 1997). But still more important, a single earthquake may trigger a large number of landslides, and they obey roughly the same size distribution as those collected over long times (Harp and Jibson 1995, 1996).

The same result was found for landslides triggered by heavy rainfall events and by rapid snow melt (Malamud et al. 2001). So there is no clear correlation between the fractal properties of landslides and those of the triggering events.

An interesting relation between hydrology and landslides was proposed by Pelletier et al. (1997). In an experimental watershed, Rodriguez-Iturbe et al. (1995) and Jackson and Vine (1996) discovered scale-invariant properties in the spatial distribution of . Scale invariance was observed with the help of spectral methods (Sect. 3.4) as well as in the size distribution of soil patches with a moisture larger than a given threshold. Apparently, the variations in soil moisture can be at least partly attributed to variations in soil properties such as porosity (Rodriguez-Iturbe et al. 1995). Since soil moisture has a strong impact on slope stability, the idea of variations in soil moisture being responsible for the fractal size distribution of landslides is straightforward. Pelletier et al. (1997) generated self-affine soil moisture fields and self-affine surface topographies, both consistent with observed fractal properties. In a second step, they assumed that landsliding is likely to occur if the product of both properties exceeds a given threshold. Analyzing the sizes of the patches where this criterion is met resulted in power-law distributions with realistic exponents. However, confirming this promising idea further with the help of field data is difficult. After a landslide has occurred, it is mostly not possible to reconstruct the original distribution of soil moisture. Consequently, nobody knows if landslide areas do in fact coincide with patches of high soil moisture.

Still more important, the approach ends at the old question for the chicken and the egg. The spatial distribution of soil moisture is governed by soil properties, geology, and landform. Therefore, it is the result of a long-term evolution. Landslides are an important component in this evolution; porosity may considerably change if material is moved, and sometimes landslides even cause ponds which may persist for a long time. So there may be a relationship between the fractal properties of landslides and soil moisture; but then, understanding the fractal character of either of these phenomena requires understanding the other, too. From this point of view, it may be better to start with a minimum set of phenomena. So let us examine whether landslides may be fractally distributed even in absence of scale invariance in the hydrological components, perhaps only as a result of landform evolution.

8.2 Are Landslides like Sandpile Avalanches?

The importance of the slope gradient in slope stability is obvious. The driving force in downslope direction increases with increasing gradient, while the normal force, and thus the maximum retaining force, decreases. Starting from this knowledge, the simplest landslide model is similar to the sandpile model introduced in Sect. 5.5: We assume that the slope becomes unstable as soon as its gradient Δ reaches or exceeds a threshold Δ_c somewhere; in this case, material moves from the unstable site in downslope direction.

In Sect. 5.5 we have already derived the BTW model from the sandpile model. As large parts of the derivation can be transferred, this work pays off now. Again, the ground area is divided into square tiles, numbered by two indices i and j. Let $H_{i,j}$ be the surface height at the site (i,j), measured in units of the linear size of the tiles. If the grid is aligned in such a way that $H_{i,j} > H_{i+1,j}$ and $H_{i,j} > H_{i,j+1}$, the slope gradient is

$$\Delta_{i,j} = \sqrt{(H_{i,j} - H_{i+1,j})^2 + (H_{i,j} - H_{i,j+1})^2}. \tag{8.1}$$

Introducing some approximations, we obtained a relaxation rule for the slopes (Eq. 5.8). Let us, for convenience, make the same step that led from the sandpile model to the BTW model and disregard the effect of diagonal neighbors, but keep the conservative character of the relaxation rule. Under this simplification, Eq. 5.8 turns into

$$\Delta_{i,j} \mathrel{-}= \tfrac{4}{\sqrt{2}}, \quad \Delta_{i\pm1,j} \mathrel{+}= \tfrac{1}{\sqrt{2}}, \quad \text{and} \quad \Delta_{i,j\pm1} \mathrel{+}= \tfrac{1}{\sqrt{2}}.$$

However, the original relaxation rule was based on the assumption that a fixed amount of material (a grain) is displaced towards the lower neighbors. In the context of landslides, there is no reason for this assumption. Let us, instead, assume that such an amount of material is moved that the slope gradient at the unstable site decreases to a given residual value Δ_r. In order to introduce this modification into the relaxation rule, we must multiply the changes applied to the slopes by a factor $\frac{1}{4}\sqrt{2}(\Delta_{i,j} - \Delta_r)$, which leads to

$$\Delta_{i\pm1,j} \mathrel{+}= \tfrac{1}{4}(\Delta_{i,j} - \Delta_r), \quad \Delta_{i,j\pm1} \mathrel{+}= \tfrac{1}{4}(\Delta_{i,j} - \Delta_r), \quad \text{and} \quad \Delta_{i,j} := \Delta_r.$$

Both the values Δ_r and Δ_c can be eliminated from the relaxation rule and from the stability criterion by introducing the variables

$$u_{i,j} := \frac{\Delta_{i,j} - \Delta_r}{\Delta_c - \Delta_r}. \tag{8.2}$$

In terms of $u_{i,j}$, instability occurs if $u_{i,j} \geq 1$, and the relaxation rule turns into

$$u_{i\pm1,j} \mathrel{+}= \tfrac{1}{4} u_{i,j}, \quad u_{i,j\pm1} \mathrel{+}= \tfrac{1}{4} u_{i,j}, \quad \text{and} \quad u_{i,j} := 0. \tag{8.3}$$

This relaxation rule differs from that of the BTW model (Eq. 5.2) gradually. Both are conservative and isotropic. The only major difference arises from the fact that the amount of displaced material is not pre-defined in the landslide model, but depends on the recent value of $u_{i,j}$. Although this difference only occurs if $u_{i,j} > 1$ temporarily, we need floating-point variables $u_{i,j}$ here, whereas the BTW model can be run with integer variables.

But what about the boundary conditions and about long-term driving? Let us first consider a slope within a closed box where no material can leave the model domain in analogy to the sandpile model. In analogy to the latter, it is convenient to imagine an additional row of sites which cannot become

unstable at the lower boundaries. The relaxation rule (Eq. 8.3) can be applied at boundary sites, too, while those parts of the rule concerning sites outside the model region are neglected.

In contrast to the sandpile model, the landslide model should not be driven by adding discrete portions of material, but by continuously changing the surface heights. From Eqs. 8.1 and 8.2 we obtain the driving rule

$$\frac{\partial}{\partial t} u_{i,j} = \frac{(H_{i,j} - H_{i+1,j}) \frac{\partial}{\partial t}(H_{i,j} - H_{i+1,j}) + (H_{i,j} - H_{i,j+1}) \frac{\partial}{\partial t}(H_{i,j} - H_{i,j+1})}{(\Delta_c - \Delta_r)\, \Delta_{i,j}}.$$

If we assume

$$H_{i,j} - H_{i+1,j} = \Delta_{i,j} \cos\alpha \quad \text{and} \quad H_{i,j} - H_{i,j+1} = \Delta_{i,j} \sin\alpha,$$

where the angle α describes the main slope direction in analogy to the sandpile model, the driving rule turns into

$$\frac{\partial}{\partial t} u_{i,j} = \frac{\cos\alpha \frac{\partial}{\partial t}(H_{i,j} - H_{i+1,j}) + \sin\alpha \frac{\partial}{\partial t}(H_{i,j} - H_{i,j+1})}{\Delta_c - \Delta_r}. \tag{8.4}$$

The simplest way of driving consists of homogeneously tilting the slope in main slope direction, which can be introduced by the rule

$$\frac{\partial}{\partial t} H_{i,j} = c - r\,(i \cos\alpha + j \sin\alpha).$$

The parameter r determines the rate of tilting, and the constant c may be chosen in such a way that the average height of the slope remains constant through time. The left-hand part of Fig. 8.2 illustrates how a slope profile would evolve under this driving rule if it was made of a viscous fluid. After some time, the slope approaches a steady state. As a result of the closed boundaries, the maximum gradient occurs in the middle of the slope, while the surface is flat near the boundaries. Inserting this approach into Eq. 8.4 leads to

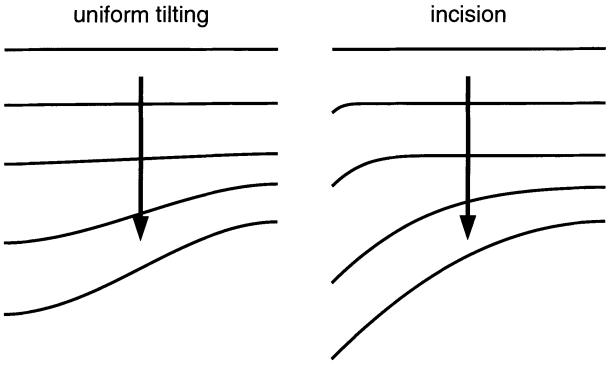

Fig. 8.2. Sketch of two different ways of driving slope evolution: homogeneous tilting (left) and fluvial incision (right).

$$\frac{\partial}{\partial t} u_{i,j} = \frac{r}{\Delta_c - \Delta_r}.$$

The right-hand side only determines the time scale of the model, so we can rescale time in such a way that

$$\frac{\partial}{\partial t} u_{i,j} = 1.$$

At this point, we have arrived at the conservative limiting case of the OFC model (Sect. 7.5). The only difference concerns the meaning of the variables. While the OFC model concerns forces acting on blocks in stick-slip motion, the landslide model derived here concerns slope gradients. Even the boundary conditions of the landslide model can be related to those of the OFC model; they are equivalent to rigid-frame boundary conditions (Sect. 7.6). The coincidence with the OFC model is surprising since the physics behind both approaches is essentially different. However, this result shows that earthquakes and landslides may be similar at an abstract level, even if we do not describe landslides directly with the help of forces at a slip surface.

However, the results of the OFC model presented in Fig. 7.10 (p. 148) are not very promising with respect to the size distribution of landslides. With increasing level of conservation (i. e., if k decreases), the exponent b in the size distribution decreases. This result is not restricted to free boundaries as considered in Fig. 7.10, but holds for rigid-frame boundary conditions, too. The upper diagram in Fig. 8.3 shows the cumulative size distribution in the conservative case, obtained from simulations on grids of 128×128, 256×256, and 512×512 sites. Compared to the non-conservative version, the model approaches the quasi-steady state rapidly. In coincidence with the BTW model, skipping the first 10^5 avalanches turns out to be sufficient; the following 10^6 landslides were analyzed. As already observed in the BTW model, the power-law behavior is strongly affected by finite-size effects. However, the non-cumulative plot of the raw data obtained on the 512×512 grid shows a clean power-law behavior with an exponent of about 1.23, leading to $b = 0.23$ in the cumulative distribution in the limit of infinite system size.

As discussed in the previous section, the lowest exponents of the distributions observed in nature are close to unity. Compared to the exponent $b = 0.05$ obtained from the BTW model, the value $b = 0.23$ is gradually better, but the difference between 0.23 and one is still large. Again, one may argue that the occurrence of a power law itself is the important property; but the deviations in the exponents are too strong for explaining the fractal size distribution of landslides with the sandpile analogy.

But what is wrong with the analogy between landslides and sandpile avalanches? Is it the neglect of inertia, the way of driving the model by uniformly tilting the slope or something else? The effects of inertia on sandpile dynamics are well-known. As soon as a sandpile becomes so large that toppling grains pick up a considerable amount of kinetic energy, the SOC behavior turns into a rather periodic behavior where large avalanches dominate. There have been attempts to simulate this behavior (e. g. Bouchaud

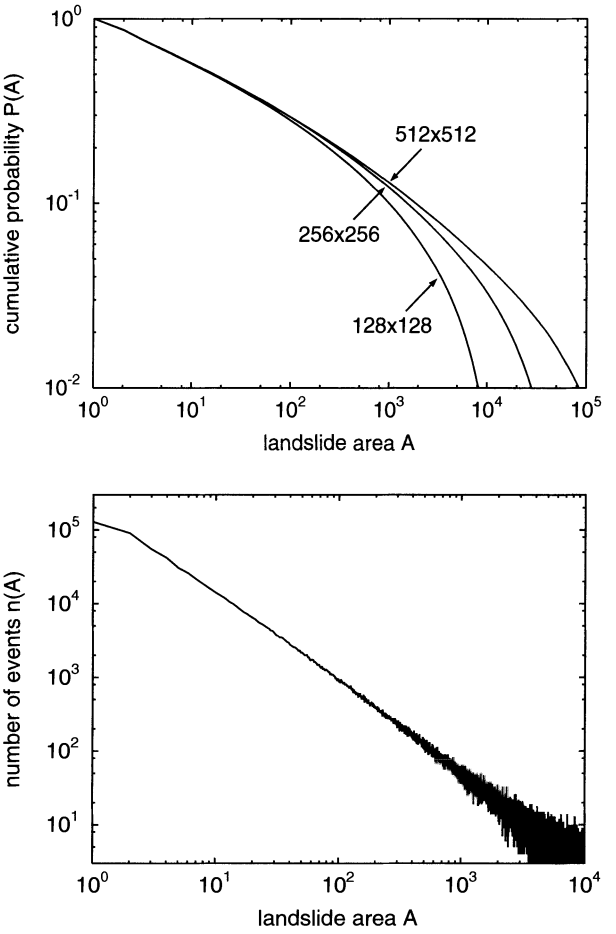

Fig. 8.3. Cumulative size distribution of the avalanches in the conservative limiting case of the OFC model for different grid sizes and non-cumulative size statistics simulated on a 512×512 grid.

et al. 1995); but obviously effects of inertia take us further away from a realistic distribution of the landslide sizes.

Concerning long-term driving, it may be more realistic to assume incision of a river at the toe of the slope instead of tilting the whole slope. Let us still assume that no material can leave the model domain as a result of a landslide, but that a given amount of material per unit time is removed at the lower boundaries. In other words, material falls into a river and is carried away. The right-hand sketch in Fig. 8.2 illustrates this kind of driving, again assuming that the slope was made of a viscous fluid. According to our assumptions on the slope direction, it makes sense to assume that fluvial incision acts at the boundaries where $i = N$ or $j = N$ on a $N \times N$ lattice. It is again convenient to add a row of sites which cannot become unstable at these two boundaries; let us assume that those additional sites are eroded at a given rate r, while all other sites are unaffected. From Eq. 8.4 we obtain the driving rule

$$\frac{\partial}{\partial t} u_{i,j} = \frac{r\,(\delta_{i,N}\,\cos\alpha + \delta_{j,N}\,\sin\alpha)}{\Delta_c - \Delta_r},$$

where $\delta_{i,j}$ denotes *Kronecker's delta* which is one if the indices coincide, and zero else. Let us again assume $\alpha = \frac{\pi}{4}$ and rescale time so that

$$\frac{\partial}{\partial t} u_{i,j} = \delta_{i,N} + \delta_{j,N}.$$

Thus, the rate of driving is one at two adjacent edges, two at their common corner, and zero elsewhere. Consequently, landslides can only be initiated immediately at the river. This obviously unrealistic behavior arises at least partly from neglecting the triggering mechanisms completely. However, including triggering mechanisms would make the model more complicated and does not comply with our aim of including as few components as possible.

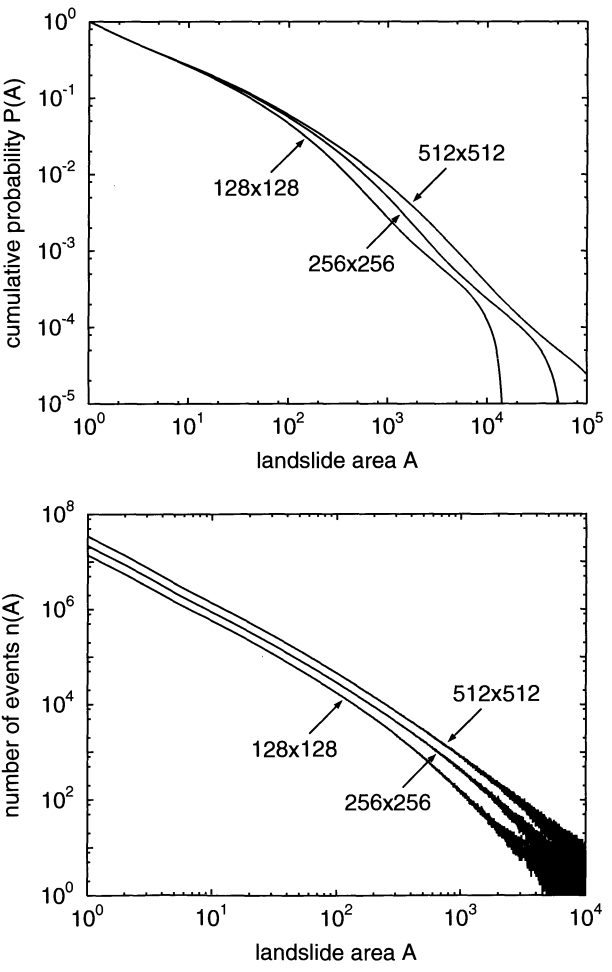

Fig. 8.4. Size distribution of the avalanches in the conservative, edge-driven OFC model. Upper diagram: cumulative size distribution. Lower diagram: non-cumulative size statistics. The non-cumulative data obtained from the 128×128 and 256×256 lattices have been shifted vertically in order to distinguish them from the data obtained from the 512×512 lattice; the original data coincide at small event sizes.

Figure 8.4 displays the size distribution of the landslide areas obtained under this driving rule. Obviously, the scale-invariant distribution is disturbed; the cumulative size distribution only follows a power law for small event sizes, i. e., only for those landslides which are confined to a narrow region close to the river. However, the non-cumulative statistics reveal a surprising result: The probability density comes closer to a power law with increasing model size. Thus, the edge-driven model may be critical, although the grid sizes analyzed here are not sufficient for assessing whether the distribution converges towards a power law in the limit of infinite system size or not. But even if it does, the exponent b is not much larger than that resulting from homogeneously tilting the slope; extrapolating the power law from small sizes shows that b is not larger than about 0.3.

The cumulative data confirm the doubts against analyzing cumulative size distributions raised in Sect. 2.2. A power law with an exponent $b = 1.25$ fits the data fairly well over about two orders of magnitude at intermediate event sizes; e. g., in the range between 10^2 and 10^4. This exponent would be in perfect agreement with landslide distributions obtained from nature, but the non-cumulative data reveal that it is just an artificial effect.

In summary, the dynamics of landslides are far away from that the BTW model, respectively, the conservative limiting case of the OFC model – landslides are neither idealized not real sandpile avalanches.

8.3 Data, Models, and Reality

The results of the previous section are a good basis for a discussion on models and reality. What is a good model? On a general level, we can give a definition which may satisfy both earth scientists and physicists: A model is good if it captures at least some of the important properties of a phenomenon in reality. However, we run into problems as soon as we try to define what reality is. From the view of an earth scientist in the classical sense, this may not be a question at all. We have just one earth, and its recent state is a result of its history. So, if we focus our interest on a region, an individual slope or a mountain belt, the landslides occurring there within a certain time interval define the reality concerning landslides. Clearly, the history of these landslides is infinitely more complex than any sandpile model can ever describe; so it is not a surprise that such a model reproduces the observed landslide distributions poorly.

At this point, we are not far away from reviving the old war between earth sciences and physics. Earth scientists may be tempted to say that all those physical models are useless, and physicists may argue that earth scientists do not recognize what is really important. But what distinguishes the *reductionist approach* of physics from the approach of earth sciences? Roughly speaking, physicists tend to interpret each observation as an example and cannot cease from searching for a basic principle behind all these examples.

This subdivision of sciences into two groups is as old as science. At one side, there are the hard sciences where knowledge is obtained from reproducible experiments and is collected in form of universal laws; on the other side, there are the soft sciences which deal with phenomena with a strong inherent variability, so that only a narrative account is possible. In his book on SOC, Per Bak discusses these essentially different approaches under the topic "storytelling versus science" (Bak 1996). However, his discussion is not polemic at all, although the headline might suggest the opposite. Clearly, it would be a great deal if we could explain everything by fundamental physical laws, but we are not able to do this. For instance, the evolution of life is a long story of events of which each was unlikely, but happened since there has been an infinite number of unlikely events which might have happened. From this point of view, soft sciences are mainly storytelling, but often storytelling is the only way of going ahead.

From this discussion, we can understand why physicists might be satisfied with the analogy between landslides and idealized sandpiles while earth scientists may not be, but things are different here. Obviously, we have taken the reductionist approach in this chapter or even throughout the whole book. When discussing ideas on the origin of the fractal size distribution of landslides in Sect. 8.1, we searched for reasons for throwing away everything which may affect the distribution of landslides in a certain environment; and we found such reasons since similar distributions were observed under various conditions. Interestingly, we have arrived at the result that landslides are far away from idealized sandpile avalanches from the reductionist view, too. Obviously, idealized sandpiles comply with the reality of landslides neither from the view of earth sciences, nor from the view of physics.

But what is reality from the view of physics? As long as physics is restricted to reproducible experiments, things are clear. But what happens if we can only observe a small part of what we call reality? At this point, physics becomes quite soft, too. In principle, we start to build our own reality then; this reality is composed from both observed data and from what we believe in. We tend to make this reality as simple as possible, but close enough to the observed data for blaming the occurring differences on the imperfection of nature. As soon as we have found a physical model which explains our self-made reality, we tend to believe that all the deviations occurring in the real world can be explained by this mechanism, superposed by some perturbations which regard the peculiarities of a given situation.

There is no doubt that this is a way towards deepening our understanding of nature, but we should be aware that we may build a nice theory on sand. In our example of landslides, the self-made reality consists of a fractal size distribution where the exponent is within a certain range. However, we have already mentioned in Sect. 8.1 that observed size distributions of landslides mostly follow a power law only within a narrow range, and that the coincidence is often not very clear even within this limited range. So let us

recall the result that the cumulative size distribution of the edge-driven OFC model (Fig. 8.4) is consistent with a power law with an exponent of $b = 1.25$ within a limited range of scales. If we only consider cumulative distributions; this result is in perfect agreement with the observations. Even the rollout at smaller landslide sizes may be fine since observed landslide distributions often behave similarly. As already mentioned in Sect. 2.4, this rollout is often attributed to the process of observation, but in some cases it seems to be a real property of the distribution (Malamud et al. 2001). Except for the result that the model generates a quite clean power-law distribution in the limit of small landslide sizes which is not observed in nature, the observed data do not uniquely falsify the edge-driven OFC model with respect to landslides.

In summary, physics may become quite soft, too. The reasoning often hinges on the idea that power-law distributions are something fundamental since they are observed in various natural phenomena. Only for this reason we tend to believe that an idealized landslide distribution must be scale-invariant, although available data sets are obviously not sufficient.

8.4 The Role of Time-Dependent Weakening

In the previous sections we have discussed the applicability of a modified sandpile automaton to landslide dynamics. The model reproduces the observed power-law distributions qualitatively, but the exponent is significantly too low. Although we have seen that the available field data are not sufficient to falsify this model clearly, it would at least be easier to believe in a model which generates power-law distributions with a realistic exponent over a reasonable range of landslide sizes. So let us now examine whether the model can be modified in this way.

When discussing the OFC model in Chap. 7, we have seen that the level of conservation is the crucial parameter. If the model is non-conservative, i. e., if a certain amount of the variable $u_{i,j}$ respectively $F_{i,j}$ is lost during a relaxation, the exponent b increases. By adjusting the level of conservation, the OFC model can be tuned to be consistent with the Gutenberg-Richter law. So, why do we not introduce a non-conservative relaxation rule? At this point, physics becomes hard again. When deriving the BTW model from the idealized sandpile model in Sect. 5.5 and deriving the landslide model in Sect. 8.2, we have seen that the relaxation rule must be conservative as long as the variable $u_{i,j}$ describes the gradient of the surface. Thus, assuming a non-conservative relaxation rule would considerably improve the results, but destroy the model's relationship to landform evolution.

So it seems that the surface gradient alone cannot be responsible for the fractal size statistics of landslides; there must be another component. In principle, this result is not surprising since slope stability depends on a variety of influences beside the slope gradient. Skipping everything except for the slope gradient was a first approach which has turned out to be too

simple; so we must introduce a second component in the model. The most straightforward idea is assigning the second component to the mechanical properties of the soil or rock which may change through time.

Early approaches in this direction were developed for different topics in landform evolution. Bouchaud et al. (1995) introduced a two-state approximation for modeling sandpile dynamics. Grains are assumed to be either at rest or rolling with a pre-defined velocity; so the evolution of the sandpile is governed by both the slope gradient and the number of rolling grains. As mentioned earlier, this approach was introduced for explaining the deviations of real sandpiles from the BTW model due to effects of inertia. On a larger scale, a similar approach was introduced for modeling erosion processes (Hergarten and Neugebauer 1996). Here, grains are assumed to be either tightly connected to their neighbors or mobile; the number of mobile grains is the second variable beside the slope gradient. Later, this approach was transferred to landslide dynamics by replacing the number of loose grains with an amount of mobile material (Hergarten and Neugebauer 1998, 1999). Although this approach is based on partial differential equations, which makes it numerically demanding, it results in a fractal size distribution over a reasonable range of scales with realistic exponents. However, criticality occurs only within a quite narrow range of parameter combinations, so that it is somewhere between SOC and (tuned) criticality in the classical sense.

At the same time, Densmore et al. (1998) introduced an approach for landsliding as a component of a rather comprehensive landform evolution model. Similar to the approaches discussed above, slope stability is governed by two components; one of them describes the surface geometry, while the other uniformly increases through time and introduces some kind of weakening. Instability occurs as soon as the sum of both components exceeds a given threshold. However, the landslides do not propagate in this model like the avalanches in the sandpile model; instead an explicit rule for their size has been introduced. Justified by experiments (Densmore et al. 1997), it is assumed that the size of a landslide being initiated at a certain location depends on the time span since the previous landslide at the location. The model yields power-law distributions with realistic exponents for the landslide volumes, but clean power laws can be recognized only over about one order of magnitude in landslide volume, which is in fact a very narrow range.

Let us now transfer the idea of time-dependent weakening to the sandpile-like model derived in Sect. 8.2. In addition to the variable $u_{i,j}$ which behaves as before, a second variable $v_{i,j}$ is introduced. Since $v_{i,j}$ describes time-dependent weakening, it increases through time at a pre-defined rate r between landslides. Let us assume that this rate is constant concerning space and time, so that the driving rule of the model reads:

$$\begin{aligned} \frac{\partial}{\partial t} u_{i,j} &= \begin{cases} 1 \\ \delta_{i,N} + \delta_{j,N} \end{cases} \text{ under } \begin{matrix} \text{homogeneous tilting} \\ \text{fluvial incision} \end{matrix}, \\ \frac{\partial}{\partial t} v_{i,j} &= r. \end{aligned}$$

Defining relaxation rules for unstable sites is straightforward; that of $u_{i,j}$ should remain the same, while $v_{i,j}$ is reset to zero without any further effect. The latter rule reflects the idea that $v_{i,j}$ describes some kind of memory of the material, i. e., it simply measures the time since the last landslide at the site. So the relaxation rule of the two-variable model reads:

$$u_{i\pm1,j} \mathrel{+}= \tfrac{1}{4}\,u_{i,j}, \quad u_{i,j\pm1} \mathrel{+}= \tfrac{1}{4}\,u_{i,j}, \quad u_{i,j} := 0, \quad \text{and} \quad v_{i,j} := 0. \tag{8.5}$$

The model combines a conservative with a completely dissipative variable. There are several ways of combining the variables $u_{i,j}$ and $v_{i,j}$ to a criterion for the stability of the site (i,j). The simplest combinations are adding a multiple of $v_{i,j}$ to $u_{i,j}$, so that the slope remains stable as long as

$$u_{i,j} + \lambda\, v_{i,j} \;<\; 1$$

or multiplying $u_{i,j}$ and $v_{i,j}$, so that the slope remains stable as long as

$$\mu\, u_{i,j}\, v_{i,j} \;<\; 1.$$

In these expressions, λ and μ are positive parameters. If we reconsider the definition of $u_{i,j}$ in Eq. 8.2, we see that both criteria differ with respect to the interpretation of the critical slope gradient Δ_c and the residual gradient Δ_r. In the sum approach, the slope remains stable as long as

$$\Delta_{i,j} \;<\; \Delta_c - (\Delta_c - \Delta_r)\,\lambda\, r\, \tau,$$

where τ is the time since the last landslide at the site (i,j). Consequently, the critical slope angle starts at Δ_c immediately after a landslide and decreases linearly then. After a sufficiently long time, even a flat surface ($\Delta_{i,j} = 0$) will become unstable. In contrast, the slope remains stable as long as

$$\Delta_{i,j} \;<\; \Delta_r + \frac{\Delta_c - \Delta_r}{\mu\, r\, \tau}$$

in the product approach. This means that even a very steep slope would be stable immediately after a landslide; however, this does not happen in general since the gradient decreases during a landslide. After a sufficiently long time, each slope which is steeper than the residual slope Δ_r becomes unstable, while shallower slopes will remain stable forever.

Under these aspects, both approaches are not completely realistic. The problems can be fixed by combining sum and product in a criterion such as

$$\mu\, u_{i,j}\,(1 + \lambda\, v_{i,j}) \;<\; 1.$$

In this case, the maximum gradient where the slope remains stable decreases from Δ_c for $\tau = 0$ to Δ_r for $\tau \to \infty$. On the other hand, the little inconsistencies occurring in the original approaches are not severe since the slopes will be between Δ_r and Δ_c in general.

From a numerical point of view, both approaches are similar to the OFC model. Especially, the algorithm discussed in Sect. 7.7 can be applied in both cases since the time when a site will become unstable next can be computed analytically. In the sum approach, this is as simple as in the OFC model, whereas a quadratic equation has to be solved in the product approach.

A fundamental difference between sum and product approach concerns the model parameters. In the product approach, both parameters μ and r can be eliminated by introducing rescaled variables

$$u_{i,j} \coloneqq \sqrt{\mu r}\, u_{i,j}\,, \quad v_{i,j} \coloneqq \sqrt{\frac{\mu}{r}}\, v_{i,j}, \quad \text{and} \quad t \coloneqq \sqrt{\mu r}\, t.$$

The relaxation rule (Eq. 8.5) remains the same, while the driving rule turns into

$$\begin{aligned} \frac{\partial}{\partial t} u_{i,j} &= \begin{cases} 1 \\ \delta_{i,N} + \delta_{j,N} \end{cases} \text{under} \begin{matrix} \text{homogeneous tilting} \\ \text{fluvial incision} \end{matrix}, \\ \frac{\partial}{\partial t} v_{i,j} &= 1. \end{aligned}$$

Finally, the criterion of stability is simplified to

$$u_{i,j}\, v_{i,j} < 1.$$

In contrast, only one of the two parameters can be eliminated in the sum approach, e. g., by introducing

$$v_{i,j} \coloneqq \tfrac{1}{r}\, v_{i,j} \quad \text{and} \quad \lambda \coloneqq r\,\lambda.$$

Then, both the relaxation rule (Eq. 8.5) and the driving rule coincide with those of the product approach, while the criterion of stability remains and still contains a parameter:

$$u_{i,j} + \lambda\, v_{i,j} < 1.$$

Thus, there is not any degree of freedom in the product approach, except for the choice between driving by tilting the slope or by fluvial incision. In contrast, the sum approach involves a parameter λ which quantifies the importance of the dissipative variable $v_{i,j}$ in relation to the conservative variable $u_{i,j}$. If λ is small, the model is governed by the conservative variable $u_{i,j}$, so that it turns into the sandpile-like model discussed in Sect. 8.2. On the other hand, the sites are weakly coupled if λ is large, so that only small landslides occur then. So we may expect that the sum approach behaves similarly to the OFC model (Sect. 7.5) for different levels of conservation, especially, that it shows SOC with a non-universal exponent. In contrast, we cannot say much about the product approach in advance; it may either exhibit SOC with a universal scaling exponent or may not be critical at all.

Both approaches were investigated in the case of uniformly tilting the slope (Hergarten and Neugebauer 2000). As illustrated in Fig. 8.5, the sum

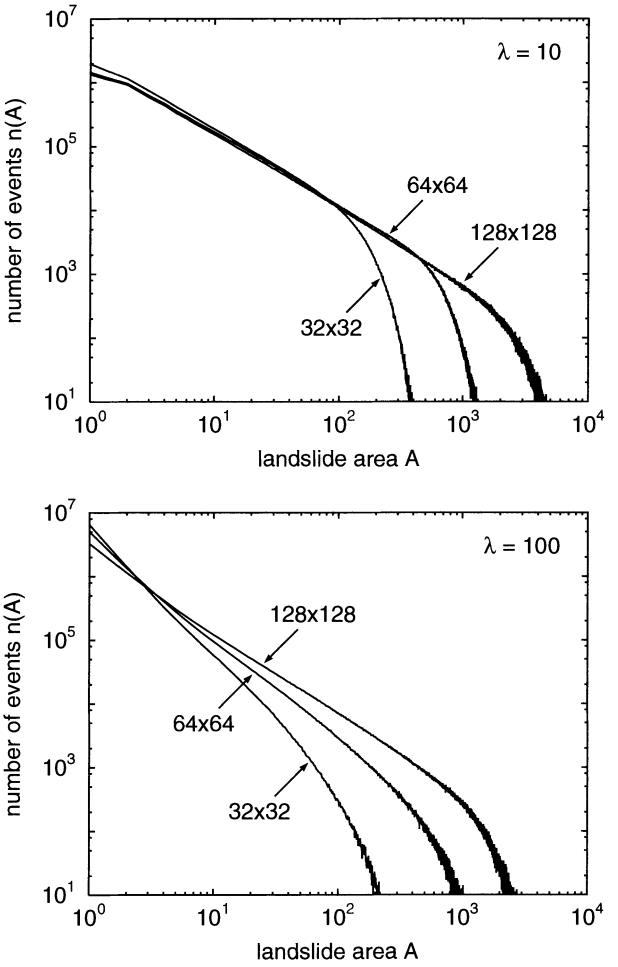

Fig. 8.5. Non-cumulative size statistics of the landslides in the sum approach under homogeneous tilting for different values of the parameter λ and different grid sizes. The statistics include 10^7 events after the quasi-steady state has been approached.

approach leads to power-law distributions where the exponent b depends on the parameter λ. The exponent increases if λ increases; time-dependent weakening becomes more important compared to the slope gradient. However, the diagrams show that the exponent strongly depends on the size of the grid, too; it decreases with increasing model size. For $\lambda = 10$, the exponent is independent of the grid size, at least for the sizes investigated here. However, the resulting distribution is very similar to that obtained from the sandpile-like model discussed in Sect. 8.2 then; its exponent is $b = 0.23$ in the cumulative sense. Thus, the effect of time-dependent weakening vanishes not only for $\lambda \to 0$, but also for arbitrary values of λ in the limit of infinite system size.

The origin of this phenomenon can easily be understood. Let us assume that the grid consists of $N \times N$ sites, and that a quasi-steady state where the values $u_{i,j}$ and $v_{i,j}$ are roughly constant in the mean has been approached.

When a site is relaxed, the amount $v_{i,j}$ is lost according to the dissipative relaxation rule. In contrast, a loss in the sum of all values $u_{i,j}$ only occurs if the relaxed site is at the boundary of the domain. Since the number of boundary sites is roughly $4N$, the average loss in the sum of all values $u_{i,j}$ per relaxation is not greater than $\frac{4N}{N^2} u_{\max}$, where $u_{\max}$ is the maximum value which may occur during an avalanche. Since the rates of driving are the same for $u_{i,j}$ and $v_{i,j}$, the average loss in both variables must be the same in the quasi-steady state. This leads to $v_{i,j} \leq \frac{4}{N} u_{\max}$ in the mean, so that $v_{i,j} \to 0$ in the limit $N \to \infty$; the variables $v_{i,j}$ become unimportant for large grids.

But what is the meaning of the limit of infinite system size? When dealing with discretized partial differential equations, focus is mainly on this limit, although it is in general not directly accessible by numerical simulations. Numerical simulations are restricted to lattices of a finite resolution, but the mesh width should be so small that it does not affect the results. In other words, numerical simulations should come as close as possible to the continuous limit which corresponds to the exact solution of the differential equation. Obviously, decreasing mesh width corresponds to an increasing number of sites if the physical size of the whole domain is given, so that the continuous limit can only be approached in the limit of infinite system size. So, if our model was based on a partial differential equation, the observed deviations of the exponent at finite system sizes would be an artificial effect of the discretization; the sum approach would be nothing new compared to the conservative limiting case of the OFC model.

However, we have already learned about the differences between differential equations and cellular automata in Sect. 7.4. In contrast to discretized differential equations, the mesh width may be a physical property in a cellular automaton, e. g., the size of a grain in the sandpile model. In the landslide model, the slope decreases from Δ_c (or a slightly larger value) to Δ_r if a site becomes unstable. If d is the mesh width, the surface height decreases by $d(\Delta_c - \Delta_r)$. This finding provides at least a rough interpretation of the mesh width in terms of typical depths of landslides; so it makes sense to assume that the mesh width is not arbitrary, but a physical property of the landsliding process. Under the assumption of a fixed cell size, the system size corresponds to the size of the considered slope. The sum approach predicts that power-law distributed landslides occur on slopes of arbitrary sizes, but that the exponents of the distributions depend on the slope sizes. Comparing this result with field observations suggests a way of assessing whether the sum approach is realistic or not. However, there are no reports on correlations between slope sizes and the exponents of the observed distributions; field evidence seems to falsify the sum approach. But on the other hand, there is no proof that both are independent, so that at least available data sets must be re-analyzed with focus on correlations in order to obtain a reliable result. But let us not go further into details because long-term driving by tilting the slope may be realistic in laboratory experiments, but not with respect to

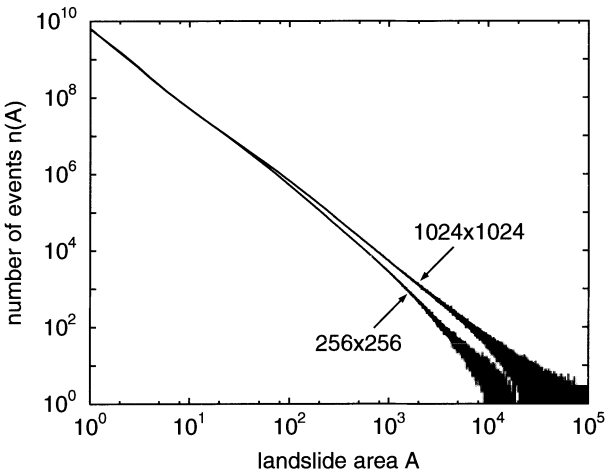

Fig. 8.6. Size statistics of the landslides in the product approach under homogeneous tilting. The statistics include 10^{10} events after the quasi-steady state has been approached.

landform evolution in nature. Here, fluvial incision is more realistic, and it is not yet clear whether the same phenomenon occurs in this case, too.

In the product approach, the size distribution of the landslides is hardly affected by the system size. Compared to the original results (Hergarten and Neugebauer 2000), the data presented in Fig. 8.6 provide more extensive on larger grids. The effect of the grid size on the distribution is not very strong. This result is somewhat surprising because the relationship $v_{i,j} \leq \frac{4}{N} u_{\max}$ holds for both the sum and the product approach. Thus, the values $v_{i,j}$ must be much smaller than the values $u_{i,j}$ in the mean for large grids. However, the major difference between product and sum is that both factors in a product are of the same relative importance, no matter whether they are large or small. In contrast, a sum is mainly determined by the larger contribution.

The probability density shows a slight kink at landslide sizes of about 200 sites which can be interpreted as a finite-size effect, although its origin is unclear. The location of the kink does not depend on the system's size, but it vanishes in the limit of infinite size. In this limit, the size distribution approaches a power law with $b = 1.05$. This finding is consistent with the results obtained from landslide mapping discussed in Sect. 8.1, although it cannot explain the apparent non-universality in the observed exponents.

As already observed in the OFC model, it takes a long time until the critical state is approached; the number of events required for this rapidly increases with the grid size. On a 256×256 lattice, 10^8 events are sufficient; this number roughly increases by a factor 10 if the number of sites in each direction is doubled.

The results of the product approach are hardly affected by switching to another way of long-term driving; Fig. 8.7 shows the results obtained under fluvial incision. Compared to the case of tilting the slope, the decrease in the

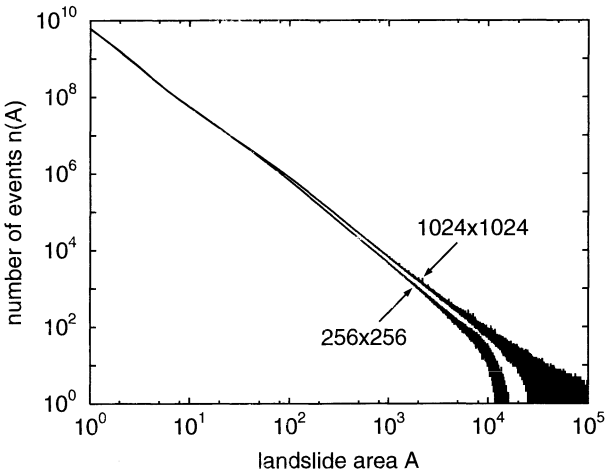

Fig. 8.7. Non-cumulative size statistics of the landslides in the product approach under fluvial incision. The statistics include 10^{10} events after the quasi-steady state has been approached.

number of events at large sizes compared to the power law is slightly less pronounced.

Let us now investigate whether the more realistic case of long-term driving by fluvial incision improves the results in the sum approach. Under driving by tilting the slope, we have observed a dependence of the results on the system size which appears to be unrealistic. In principle, driving the system by fluvial incision may fix this problem. As already mentioned, the property $u_{i,j}$ is dissipated only at the boundaries, while $v_{i,j}$ is dissipated in the whole domain. Under fluvial incision, the same is valid for the driving rule, but the property $u_{i,j}$ increases only at two of the four edges. Thus, the mean values of $u_{i,j}$ and $v_{i,j}$ may be independent from the grid size now.

Driving the system by fluvial incision makes the sum approach more complex. Figure 8.8 shows results for $\lambda = 1$, $\lambda = 10$, and $\lambda = 100$. In contrast to the simulations presented before, a fixed number of events (10^{10}) was skipped in the beginning. The spikes occurring at large grids for $\lambda = 1$ indicate regular behavior. The analogy to the results of the OFC model (Sect. 7.5) suggests that the regularity is a transient effect of the initial state, so that skipping 10^{10} events may not be enough on large grids. However, longer simulations do not provide evidence that the effect ceases through time. The same applies to the local maxima occurring at large sizes. The distribution is strongly time-dependent here; the maxima occur and vanish through time. Thus, interpreting the model's behavior in the context of SOC is difficult.

Nevertheless, the events are roughly power-law distributed with non-universal exponents. For $\lambda = 1$, we obtain $b \approx 0.7$; the exponent increases to $b \approx 1$ for $\lambda = 10$ and $b \approx 1.1$ for $\lambda = 100$. However, the latter value again depends on the model size; it was determined for the 256×256 grid and decreases with increasing number of sites. Further simulations have shown that this result persists for $\lambda > 100$. The parameter λ quantifies the relative importance

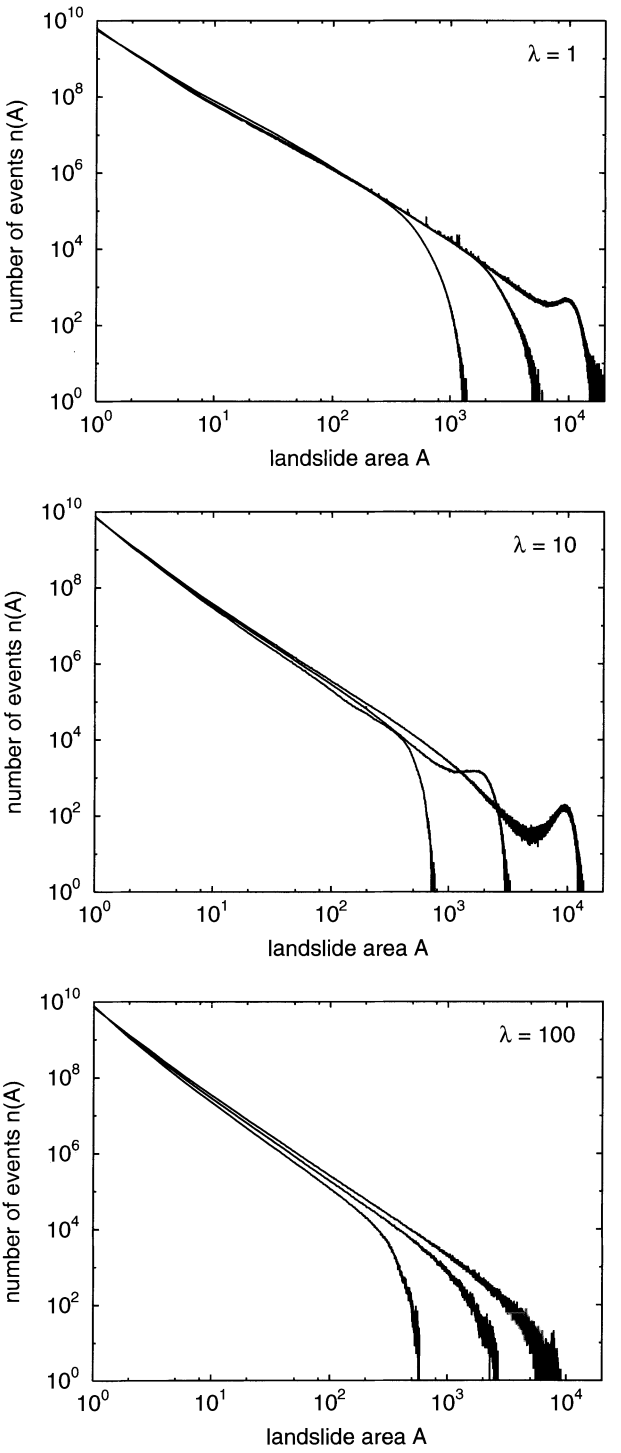

Fig. 8.8. Non-cumulative size statistics of the landslides in the sum approach under fluvial incision for different values of the parameter λ. The statistics include 10^{10} events after the same number of events was skipped; the curves correspond to grid sizes of 64×64, 128×128, and 256×256 sites (from left to right).

of time-dependent weakening compared to that of the slope gradient. Since it was a result of rescaling the model parameters, it summarizes several influences such as material properties, climatic conditions, and the rate of fluvial incision. Thus, the edge-driven sum approach is able to explain a variation of the exponent b, depending on material properties, climatic conditions, the rate of incision, and partly on the slope size.

So what have we learned about landslides in this section? The major result is that considering time-dependent weakening in addition to the slope gradient considerably improves the results. This result supports the idea that time-dependent weakening is an important component in landform evolution by landsliding. But on the other hand, we should be aware that this is still just one theory.

Another question remains open. By simply combining two variables in different ways, we have derived two models. One of them, the product approach, predicts SOC with a universal exponent which fits into the range observed in nature quite well. In contrast, the sum approach exhibits a more complicated behavior where power-law distributions with parameter-dependent exponents occur. Since the results of the product approach are robust and can be immediately be identified with SOC, the product approach seems to be more promising than the sum approach. If we follow these arguments, we might attribute the variation of the exponents observed in nature to a bias in the process of measuring or to statistically insufficient data. But here we have exactly arrived at the discussion started in Sect. 8.3. If a model generates clean power laws or even shows SOC, we tend to believe that there is some truth behind it. This is a good point for giving the problem back to those scientists who analyze data from nature in order to analyze potential correlations between physical parameters and the exponents in the landslide size distributions. So these models are clearly not able to explain everything about landslides, but at least to sharpen the questions.

8.5 On Predicting Slope Stability

Although the results of the previous section account for scale-invariant properties of size distributions of landslides and pose questions for further research, a geologist or an engineer may still ask what all this is good for. Obviously, the models were designed with respect to long-term behavior, but not for predicting slope stability under certain conditions. However, they can explain an effect occurring frequently in modeling slope stability – instability of slopes which are assessed to be stable by standard methods.

As mentioned in the introduction of this chapter, slope stability analysis mainly relates driving forces and maximum retaining forces. The ratio between maximum retaining forces and driving forces is called *factor of safety*; it is mostly abbreviated by the symbol F. There are several methods of determining the factor of safety, but it seems to be clear from its definition that

a slope becomes unstable exactly if $F < 1$. However, it is observed that slope failure often occurs at $F > 1$; and this result is explained by the phenomenon of *progressive slope failure*. At least on a qualitative level, progressive slope failure is similar to the propagation of avalanches in SOC models such as the BTW model, the OFC model or the landslide models discussed in this chapter. Instability occurs at a small part of the slip surface, and since load is transferred to the neighborhood, instability may spread over a large area.

Some sophisticated slope stability models already regard effects of progressive failure, but the limitation of these approaches is obvious: Propagation of instabilities crucially depends on the spatial distributions of the driving forces and of the strength of the material at the potential failure area, but these distributions are not accessible in general. Thus, progressive failure can be simulated if assumptions on these distributions are made, e. g., nearly homogeneous distributions, but the results depend on these assumptions. The spatial distribution of forces and mechanical parameters is a result of the history of the slope, and thus dominated by processes such as weathering as well as by small and large landslides. In general, this history is too complex for being described in detail by "storytelling" (Sect. 8.3).

Replacing the complex history with a result of self-organization seems to be the only way out of this problem. SOC systems evolve towards a state where their variables obey a certain statistical distribution with certain spatial correlations. So let us examine whether our landslide model can help us to understand anything concerning the factor of safety. As it has turned out to be most promising with respect to the size distribution of landslides, we focus on the product approach in the following.

In a first approximation, the driving force increases linearly with the slope gradient. Since the variable $u_{i,j}$ is a linear function of the slope gradient, it makes sense to identify $u_{i,j}$ with the driving force. Strictly speaking, this is only appropriate if the driving force vanishes if the slope approaches the residual gradient Δ_r; otherwise a constant term should be added to $u_{i,j}$ before it can be identified with the driving force. According to the product criterion, instability occurs if $u_{i,j} \geq \frac{1}{v_{i,j}}$. Thus, $\frac{1}{v_{i,j}}$ can be identified with the maximum retaining force. From this, a factor of safety over an arbitrary set of sites can be defined. The sum of the driving forces over the sites is $\sum_{i,j} u_{i,j}$, while the sum of the maximum retaining forces is $\sum_{i,j} \frac{1}{v_{i,j}}$. Then, the overall factor of safety is

$$F = \frac{\sum_{i,j} \frac{1}{v_{i,j}}}{\sum_{i,j} u_{i,j}}.$$

Obviously, the factor of safety is unity for all landslides which involve only one site. However, even the next-smaller landslides (those with $A = 2$) have a larger factor of safety since only one of these two sites is unstable in the beginning of the landslide. It can easily be shown that landslides of size $A = 2$ occur at factors of safety between 1 and $\frac{5}{4}$ in the product approach, while the factors of safety of larger landslides may be even larger than $\frac{5}{4}$.

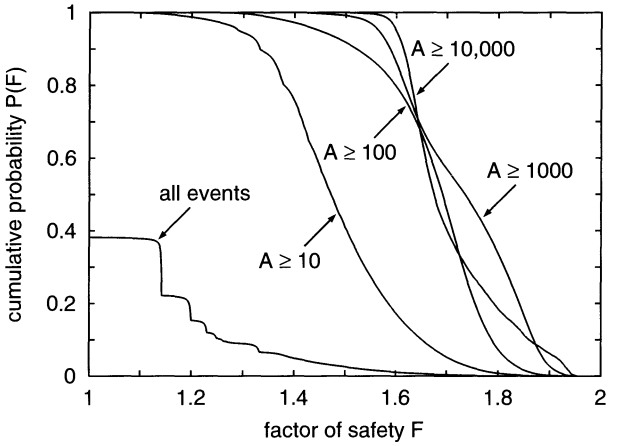

Fig. 8.9. Probability distribution of the factor of safety for landslides of different sizes in the product approach.

Figure 8.9 gives the statistical distribution of the factor of safety for landslides of different sizes. The distribution is plotted separately for different classes of landslides: all landslides, those with $A \geq 10$, $A \geq 100$, $A \geq 1000$, and $A \geq 10,000$. The statistics were obtained from a simulation on a grid of 1024×1024 sites; long-term driving was realized by uniformly titling the slope. The plot illustrates that large landslides occur at values of F which are considerably larger than unity. For instance, about 90 % of all landslides with $A \geq 100$ take place at factors of safety greater than 1.5, and nearly 10 % at $F \geq 1.8$. The distribution becomes narrower for larger landslide sizes, so that F is between 1.6 and 1.9 for about 90 % of the landslides with $A \geq 10,000$.

Figure 8.10 shows the average factor of safety as a function of the landslide size. The average factor of safety increases with increasing landslide size up to sizes of about 1000, but then the increase ceases. For large landslides, the

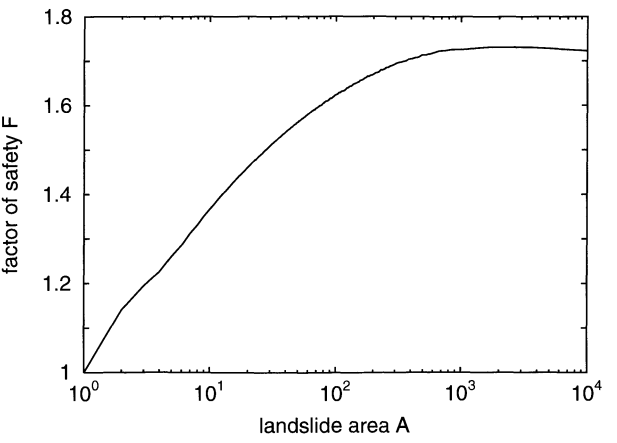

Fig. 8.10. Average factor of safety as a function of the landslide size in the product approach. In order to eliminate statistical noise, landslides of similar sizes have been grouped in bins.

average factor of safety approaches a constant value between 1.7 and 1.75 or even decreases gradually.

In summary, the landslide model that was originally designed for modeling size distributions of landslides also provides a prediction of the factor of safety of landslides. As expected, it predicts that the factor of safety where instability occurs is not an absolute criterion, but a statistical property which depends on the landslide size, too. This result is reasonable at least from a qualitative point of view. Under quantitative aspects, the critical values predicted by the model appear to be quite high; but we have to keep in mind that our model does not consider any triggering mechanisms. Thus, the predicted critical factors of safety are not the effective values which are approached during or after a heavy rainfall event or during an earthquake, but rather average values (under normal conditions) of slopes which will become unstable within a reasonable time span.

8.6 Are SOC and Universality Important in Landform Evolution?

From the view of geomorphology, it may be hard to believe in the findings of the previous sections. If the distribution of landslide sizes is fractal at all, can there be a universal exponent? Gravity-driven mass movements offer such a variety of facets that universality seems to be misplaced here.

However, universality does not mean that the frequency-magnitude relation of landslides is the same everywhere and under all conditions. The absolute number of landslides clearly depends on climate and geology; a universal exponent only implies that the relation between the number of small and large landslides is always the same. Earthquakes roughly behave the same way, although different faults may also differ strongly concerning their mechanical properties. From this point of view, field experience concerning individual landslides clearly deepens the understanding of the process, but we should beware from declining ideas such as SOC and universality just from this experience.

On the other hand, the number of events needed until the critical state is approached in the models discussed here is not very promising. In Sect. 8.4 we have learned the we need about 10^8 landslides on an individual slope on a 256×256 grid; this large number may considerably affect the applicability of the SOC concept to nature. However, most of these events are very small. Let us assume that the landslide sizes obey a Pareto distribution (Eq. 1.3) with $A_{\min} = 1$ and $b = 1.05$ as obtained from the product approach. Using the probability density $p(A)$ (Eq. 2.5), we obtain for the average landslide area

$$\overline{A} = \int_1^\infty p(A)\, A\, dA = \int_1^\infty b\, A^{-b}\, dA = \frac{b}{b-1} \approx 20.$$

On a grid of 256×256 sites, 10^8 landslides require that each site has become unstable about 30,000 times. So, if we assume that each site is affected by a landslide once per 100 years, it takes some millions of years until the critical state has been approached. On this time scale, at least glacial periods disturb the evolution towards the critical state; assuming constant conditions over such a long time span in unrealistic. But as already mentioned, a rough power-law distribution emerges much earlier; and the question for the transient periodic components which disturb the power law in the model may be a secondary question in nature. First, available data are not sufficient for distinguishing such effects, and second, we cannot tell whether they are real or an artificial effect of the model. Under these aspects, the question for the time required until a rough power-law distribution of the landslide sizes occurs becomes more important than the question for the time needed until the critical state in the strict sense is approached.

Nevertheless, the adjustment of the system to changing conditions may become a crucial question concerning SOC in landsliding processes. If the exponent of the size distribution is universal, the system will finally adjust to any altered conditions without a significant effect. If we, for instance, increase the rate of incision by a certain factor, the frequency-magnitude relation of the landslides will finally increase by the same factor, but nothing else happens. On the other hand, we have already observed in Sect. 6.3 that the behavior during the phase of transition may be entirely different; the power-law distribution may even be lost temporarily.

Research on SOC mainly concerns the properties of the critical state; everything that happens before the critical state has been approached is mostly disregarded. Consequently, knowledge on transitions resulting from changes in the model parameters is sparse. However, at least in the example of landslides, studying such transitions will be necessary for obtaining a deeper understanding of the phenomena and assessing natural and man-made hazards.

9. Drainage Networks

The earth's surface is shaped by a variety of processes; beside gravity-driven mass movements, erosion by water, wind, and glaciers are the most important examples. Fluvial erosion is ubiquitous in many climates; river networks can be considered as the backbone of the landscape. Thus, understanding the evolution of drainage networks has been a major challenge in geomorphology.

Although one may argue whether fluvial erosion is the most important landform evolution process or not, it is in fact the one that covers the widest range of scales. If we consider small erosion rills as the non-permanent part of the drainage pattern, the process of network formation reaches from the centimeter scale up to the continental scale; and these scales differ by more than seven orders of magnitude.

9.1 Fractal Properties of Drainage Networks

In Chap. 1, several definitions of scale invariance were discussed. Some of them address geometric properties of sets, while the approach preferred in the previous chapters refers to the sizes of objects. When analyzing the earth's surface, areas covered by water can easily be detected, e. g., from satellite images. In contrast, distinguishing individual rivers is more difficult and often requires a manual processing. Therefore, analyzing the pattern of water-covered areas geometrically is simpler than investigating rivers with respect to scale-invariant size distributions. So let us first apply the box-counting method (Sect. 1.2) to a map of rivers and lakes. Figure 9.1 shows the rivers and lakes above a certain size in Germany; the data were exported from a Geographic Information System (GIS).

However, the results presented in Fig. 9.2 are not very promising; they are strongly influenced by the properties of the digital map. The a step in the number of filled boxes at box sizes of about 500 m is a minor problem; it turns out to be an effect of the finite line width: In this map, nearly all rivers are represented by lines of a fixed width of about 500 m; only very large rivers and lakes are represented by filled areas. Thus, most rivers appear to be one-dimensional objects for box sizes larger than about 500 m, while they turn into areas for smaller box sizes. This artefact can be removed by making the lines infinitely thin; the lower graph shows the result.

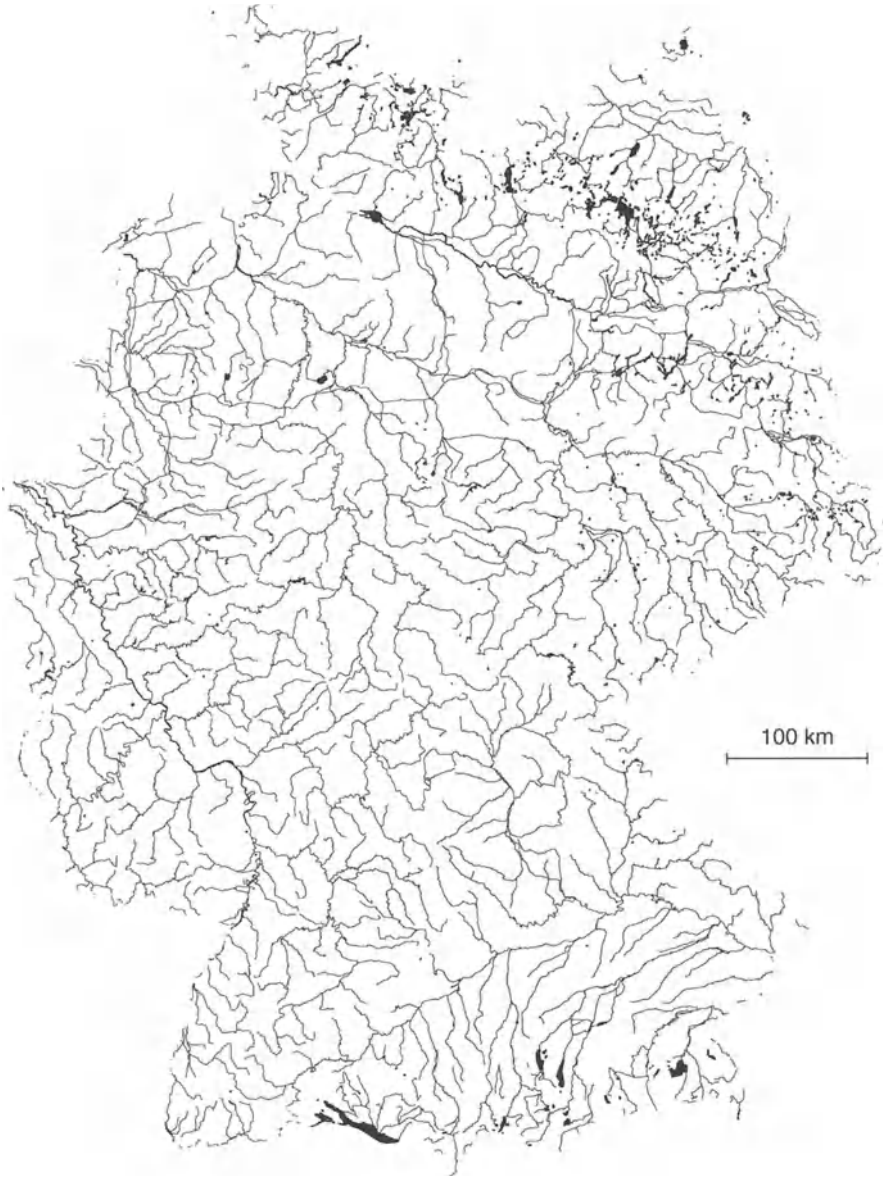

Fig. 9.1. Map of some rivers and lakes in Germany.

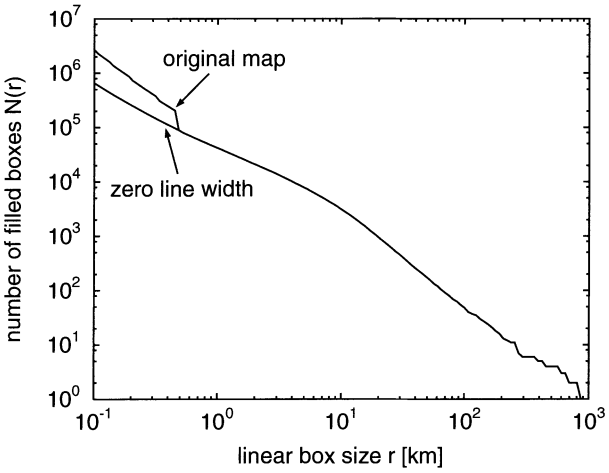

Fig. 9.2. Result of applying the box-counting method to the map from Fig. 9.1.

After this correction, the result looks like results from box counting often do. For large box sizes, here for $r \geq 10\,\text{km}$, the plot shows a power-law dependence with a fractal dimension $D = 2$. This only means that nearly all boxes are filled for larger box sizes. Below this size, a transition to the opposite non-fractal limiting case which is characterized by $D = 1$ takes place. This behavior arises from the limited resolution of the data set; only rivers above a certain size are displayed. For this reason, the box-counting algorithm recognizes a pattern of mainly independent lines at small box sizes, leading to a power law with $D = 1$. Between these boundary regions, a crossover occurs where the graph is curved. In some analyses found in literature, only the region of crossover is plotted, and a straight line with a non-trivial fractal dimension is fitted there. In this case, the transition is misinterpreted to indicate scale invariance. At small box sizes, the result obtained from our map is slightly more complicated because of the large rivers and lakes. Since the box-counting algorithm recognizes them as areas in the limit of small box sizes, they finally govern the statistics. So the exponent increases to $D = 2$ again in the limit $r \to 0$.

In summary, applying the box-counting method to the map does not indicate any scale invariance. Blaming the lack of scale invariance on the limited resolution of the map seems to be straightforward. But what would happen if the map included smaller rivers, too? Obviously, the range of the non-fractal power law at small box sizes would decrease. In return, all boxes would tend to be filled even for $r < 10\,\text{km}$, so that the range of the non-fractal power law at large box sizes would increase. Consequently, the crossover between the two non-fractal regimes would be shifted towards smaller scales, but probably still without showing scale invariance.

Even if box-counting revealed scale invariance of drainage networks, what would be the interpretation of the fractal dimension? Interpreting fractal dis-

tributions of object sizes (perhaps lengths or widths of rivers) is much easier than interpreting a rather abstract fractal dimension obtained geometrically. In fact, scale-invariant properties of drainage networks were discovered long before the framework of fractals was established in earth sciences. In analogy to the examples discussed in the previous chapters, they are scale-invariant in the sense of the object-based definition rather than in the geometric sense.

The earliest results, known as *Horton's laws*, hinge on a classification of rivers in hierarchical schemes, such as the ordering schemes of Horton (1945) and Strahler (1952). In Horton's original scheme, the rivers which deliver their discharge directly into an ocean are first-order streams; their tributaries are second-order rivers, and so on. In contrast, Strahler's scheme considers river segments and starts at the sources. River segments emerging directly from a source are of first order. Segments of higher orders may arise where river segments join: If two river segments of equal orders n join, they constitute a segment of order $n+1$ in downstream direction. On the other hand, a river segment of order n simply captures all segments of lower order, so that the order n persists in this case.

Horton (1945) discovered relations between the numbers of rivers of different orders and between their sizes. Let us use Strahler's ordering scheme, and let ν_n be the number of river segments of order n, Horton's first law states that the *bifurcation ratio* $R_\mathrm{b} \coloneqq \frac{\nu_i}{\nu_{i+1}}$ is roughly the same for all orders n. If L_n is the mean length of all river segments of order n, the *length ratio* is $R_\mathrm{l} \coloneqq \frac{L_{i+1}}{L_i}$. According to Horton's second law, R_l is roughly the same for all orders, too.

The fractal tree discussed in Sect. 1.6 is a simple pattern obeying Horton's laws. In the terms used there, n is the bifurcation ratio R_b, and λ is the inverse of the length ratio, $\frac{1}{R_\mathrm{l}}$. The size distribution of the branches led to a power law with a fractal dimension $D = \frac{\log n}{\log \lambda}$. Transferred to the nomenclature of Horton's laws, we obtain

$$D = \frac{\log R_\mathrm{b}}{\log R_\mathrm{l}}.$$

Strictly speaking, this relation is only valid if all river segments of a given order have the same lengths. In contrast, the length ratio R_l involves the mean lengths. Therefore, Horton's laws do not imply a power-law distribution of river lengths in the strict sense. However, the fractal dimension determined above is widely used for characterizing scale-invariant properties of drainage networks; it is called *network similarity dimension.* Typical values of Horton's ratios are $R_\mathrm{n} \approx 4.6$ and $R_\mathrm{l} \approx 2.3$, leading to $D \approx 1.8$ (e. g. Turcotte 1997).

The simple fractal tree is not the only tree-like pattern that satisfies Horton's laws. Figure 9.3 illustrates that the same Horton's ratios can be achieved by networks of different topologies. All of them have $R_\mathrm{b} = 4$ and $R_\mathrm{l} = 2.3$, leading to a network similarity dimension $D = 1.66$. The upper left network is topologically equivalent to the simple tree shown in Fig. 1.13. Two river segments of order $n-1$ join and constitute a segment of order n. This segment

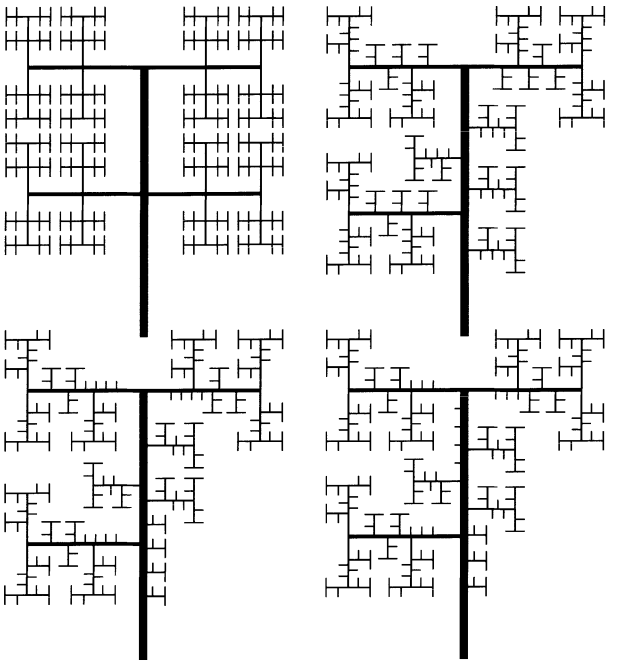

Fig. 9.3. Networks of different topologies leading to the same Horton's ratios.

captures two further segments of order $n-1$ before it joins another segment of order n to constitute a segment of order $n+1$. The upper right network is similar, but the segment of order n captures one segment of order $n-1$ and three segments of order $n-2$. In the lower networks, even segments of order $n-3$, respectively, $n-4$ deliver their discharge directly to segments of order n. The topology of drainage networks was studied in detail by Tokunaga (1978, 1984); a review is given by Turcotte (1997).

Apart from Horton's laws, the relation found by Hack (1957) is the most established fractal relation for drainage networks; it relates the sizes of drainage areas with river lengths. As long as braided rivers are not considered, a unique *drainage area* can be assigned to each point of a river network; it consists of that part of the basin which delivers its discharge to the considered point. Strictly speaking, cross sections through rivers must be considered instead of points. Let A be the size of the drainage area of an arbitrary point, and L be the length of the longest river within this area. *Hack's law* relates these quantities for different locations in a basin:

$$L \sim A^h. \tag{9.1}$$

Hack's law is only a statistical relation; neither natural drainage basins nor the artificial networks obtained by the models discussed later obey Hack's law exactly. Observed values of the exponent h are between 0.52 and 0.6 (Maritan et al. 1996b; Rigon et al. 1996). Figure 9.4 shows the original data obtained

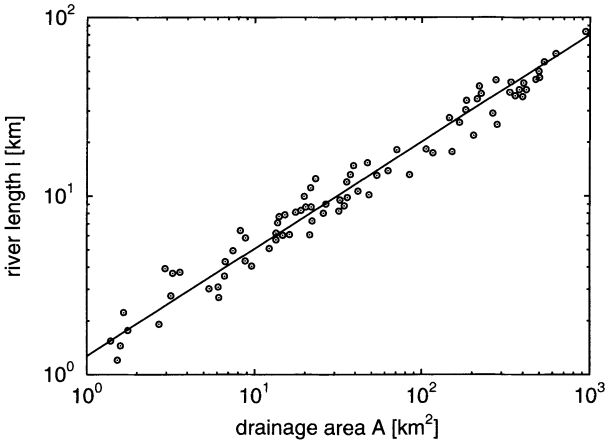

Fig. 9.4. Relationship between drainage area and river length obtained by Hack (1957) from some drainage networks in the north-eastern USA.

by Hack (1957) from some drainage networks in the USA. His result was a power-law relation with an exponent $h = 0.6$. When comparing Fig. 9.4 with the diagram in the original paper and converting miles into kilometers, the reader will notice that the axes are incorrectly scaled in the original paper.

Hack's law does not immediately refer to any of the definitions of scale invariance discussed in Chap. 1. It concerns objects (rivers and drainage areas), but not does provide a size distribution. The fractal character of Hack's law arises from the deviation of the exponent h from $\frac{1}{2}$, which is the expected exponent of non-fractal length-area relations since areas grow quadratically with linear sizes in Euclidean geometry. Hack's law allows two different interpretations:

1. Rivers are convoluted on all scales, so that their lengths grow faster than the distances between source and end point. This effect is expressed by a length-scaling exponent ϕ_L through the relation

$$L \sim d^{\phi_L}, \tag{9.2}$$

   where d is the distance between the considered point and the river's source. For natural rivers, ϕ_L is between 1.02 and 1.12 (Maritan et al. 1996b; Rigon et al. 1996).
2. Small and large drainage areas differ in shape, their lengths grow faster than their widths. So, if there is any scale invariance in the drainage areas, it must be self-affine, i. e., anisotropic (Chap. 3).

Figure 9.5 illustrates these explanations. In the left-hand part, fractal rivers are placed circular areas. The rivers are represented by *Koch's curves*; they are parts of the perimeter of Koch's island (Sect. 1.1). In the other sketch, straight rivers are placed in elliptic areas with anisotropic scaling properties. Both examples satisfy Hack's law with a realistic exponent $h = 0.56$.

At first sight, the fractal length scaling of the rivers provides a more realistic interpretation of Hack's law than the idea of self-affine drainage areas

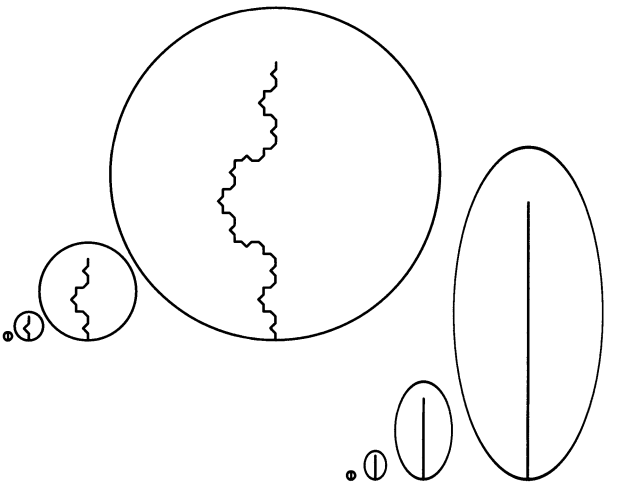

Fig. 9.5. Two different explanations of Hack's law. Left: fractal rivers. Right: elongated drainage areas.

does; but this alone leads to $h = \frac{1}{2}\phi_{\mathrm{L}}$. Thus, the fractal length scaling alone only accounts for a range of h between 0.51 and 0.56, which is roughly half of the deviation of h from $\frac{1}{2}$. Thus, both fractal length scaling of the rivers and anisotropic scaling properties of the drainage areas contribute significantly to Hack's law; the convolution of the rivers and the elongation of the drainage areas are roughly half as strong as shown in Fig. 9.5.

The power-law distribution of the discharges in a basin was discovered later (Rodriguez-Iturbe et al. 1992a):

$$P(q) \sim q^{-\beta},$$

where $P(q)$ is the probability that an arbitrarily chosen point (strictly speaking, a cross section through a river) has a discharge of at least q. The exponent β is between 0.41 and 0.46 for real basins (Maritan et al. 1996b).

As long as the assumption of stationary flow holds, the discharge at a point is the rate of precipitation integrated over its drainage area. Thus, if precipitation acts uniformly over the whole basin, the discharge q is proportional to the size of the drainage area A. So the sizes of the drainage areas obey the same distribution as the discharges:

$$P(A) \sim A^{-\beta}, \tag{9.3}$$

where $P(A)$ is the probability that an arbitrarily chosen point has a drainage area of at least A. At this point, we have arrived at a fractal distribution as discussed in Sect. 1.3. The symbol β has been chosen instead of b in order to maintain the established nomenclature. In principle, the scale-invariant size distribution of the drainage areas is a combination of the fractal length distribution arising from Horton's laws and the area-length relation constituted by Hack's law.

Beside the fractal properties discussed here, there are still some more, but they turned out not to be independent from those mentioned above (Maritan et al. 1996b; Rinaldo et al. 1998). Understanding the fractal properties of river networks has been a challenge since they were discovered. In the following, some general aspects and some ideas on their origin are discussed.

9.2 Discharge, Drainage Areas, and Water Balance

In a natural environment, a river or a lake collects water from several sources:

- the amount of precipitation falling directly on its surface,
- the runoff from adjacent areas (surface runoff during heavy rainfall events and subsurface flow), and
- the contribution of smaller rivers.

Thus, each model for understanding relations between rivers of different sizes should start at a water balance. In general, the model domain is divided up into (mostly regular) cells. In each cell, a point is chosen to be representative for the cell; this point is called node. In contrast to the models discussed in the previous chapters, the nodes and cells are numbered with a single index i here. In many cases, there is no need to distinguish between the node i and the cell i; so we will often speak of the site i when referring to nodes or cells.

The discharge of each cell consists of the amount of precipitation acting on it and the contribution from adjacent cells. This discharge must be distributed among the adjacent cells, and the question how this distribution is performed is a central point. If a surface height H_i is assigned to each node, hydrodynamic models based on the Navier-Stokes equations can in principle answer this question, but even simplified models such as the diffusive wave approach based on a modification of Manning's equation (Paul et al. 1999; Hergarten et al. 2000) are likely to cause trouble here. The problem is that these approaches require resolving all structures of the surface which are significant for the flow field. Consequently, a model of this type must cover the regional scale of several kilometers with a resolution of less than a few meters; so running such a model is likely to become costly.

Under certain conditions, it makes sense to assume that each site has a unique drainage direction, i. e., that it delivers its discharge completely to one of its neighbors. This assumption seems to be reasonable if flow is channelized, but may be wrong in case of sheet-like surface runoff or subsurface flow. However, braided rivers are an example where this assumption is not appropriate even for channelized flow, but on the other hand, drainage areas and river lengths can only be consistently defined under this assumption.

Let d_i be the index of that site where the site i delivers its discharge q_i (volume per time) to, and let N_i be the neighborhood of the site i. If r_i is the precipitation captured by the cell i (volume per time), the water balance for the site i reads

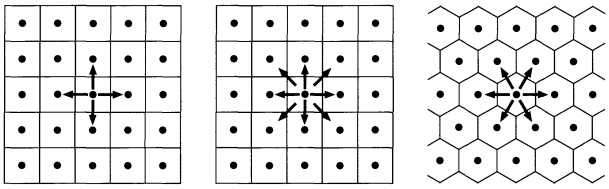

Fig. 9.6. Three regular grid topologies: quadratic with four flow directions, quadratic with eight flow directions, and hexagonal.

$$q_i = r_i + \sum_{\substack{j \in N_i \\ d_j = i}} q_j .$$

The line between the nodes i and d_i is a river segment; let us denote it with the index i of the source cell in the following. However, we should only speak of a river segment if its discharge is high enough to justify the assumption of channelized flow; otherwise a segment just defines a discharge and a drainage direction.

In this approach, river segments are restricted to discrete directions which depend on the topology of the lattice. Figure 9.6 shows three widely used, regular grid topologies. Two of them are quadratic; they only differ with respect to the number of allowed flow directions. In the first case, the neighborhood includes only the nearest neighbors, so that four different flow directions are allowed. However, the diagonal neighbors can be included in the neighborhood, too, resulting in eight possible flow directions. This extension reduces the anisotropy of the grid. In the right-hand part of Fig. 9.6, a hexagonal grid is shown. Here, six different flow directions are allowed, which is less than on the eight-neighbor quadratic grid. In return, these directions are equivalent, while diagonal and non-diagonal neighbors are different in the quadratic case.

Apart from the topology of the mesh, established models differ concerning the assumptions on the drainage direction d_i. The simplest models discussed in the following sections are based on regular topologies or on randomly chosen flow directions. In contrast, the more complex models discussed later are based on a surface topography which may change through time and determines the flow directions. In most models, the precipitation rate is assumed to be constant through time and distributed homogeneously over the entire domain. In this case, it is convenient to transform the model to non-dimensional variables; the spatial coordinates are rescaled in such a way that the area of the cells is unity, and time is rescaled so that the precipitation captured by each site is one, too. After this, the water balance reduces to

$$q_i = 1 + \sum_{\substack{j \in N_i \\ d_j = i}} q_j . \tag{9.4}$$

Still more important, the difference between discharges and drainage areas vanishes then: The size A_i of the drainage area of any site i coincides with its discharge q_i.

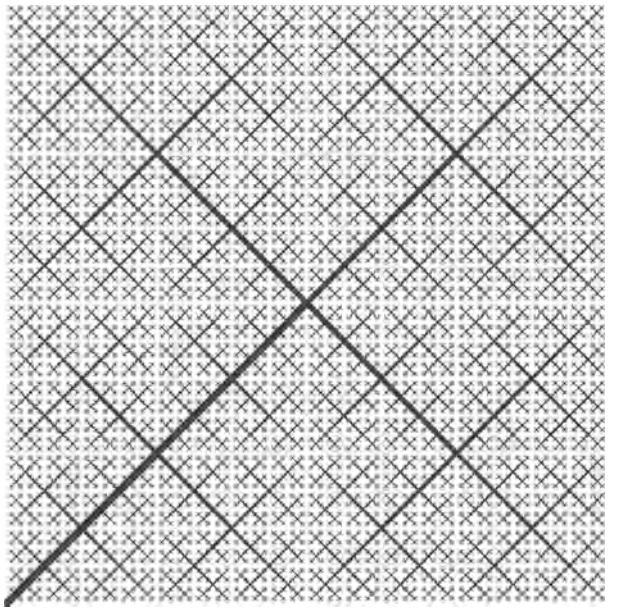

Fig. 9.7. Peano's basin, computed on a 1024×1024 lattice. Only sites with drainage areas of at least 20 tiles are plotted. The line widths are proportional to the fourth root of the discharge.

9.3 Peano's Basin

In Sect. 1.6 we obtained scale-invariant patterns by nesting objects of different sizes in a regular way. Can this principle help us to understand the fractal properties of drainage networks?

Figure 9.7 shows a network resulting from a nesting algorithm; it is called *Peano's basin*. We start at a square region with a single outlet at a corner of the domain, using a quadratic grid topology where diagonal flow directions are allowed. The domain is subdivided into four patches which are topologically the same as the whole domain. However, in order to obtain a consistent network, they must be oriented differently. Those three patches which are not directly connected to the outlet are oriented in such a way that their outlets join in the middle of the domain. The fourth patch is aligned towards the outlet of the whole domain, so that it additionally carries the discharge of the three other tiles. Consequently, three patches are identical concerning the discharges, while the main river of the fourth patch has a higher discharge than the main rivers of the others.

The realization given in Fig. 9.7 was computed on a grid of 1024×1024 sites. Only sites with a drainage area of at least 20 tiles are assumed to be channelized, smaller flows are not plotted. The line widths indicate the discharges (and thus the drainage areas). According to an empirical rule for natural rivers (Leopold and Maddock 1953), the widths of the rivers in a drainage network increase roughly as the square roots of their discharges, but this choice is not appropriate for a graphic representation. The large

rivers would be quite thick lines, while small rivers were hardly visible. Thus, the line widths are weighted with the fourth root of the discharge for clarity, although the differences between large and small rivers are less pronounced than in nature then.

Obviously, the longest river within the entire domain is as long as the diagonal of the domain; strictly speaking, there are many rivers of equal lengths. As already mentioned, the network is composed of four patches; three of them are identical with the whole domain, except for their size. Thus, the longest rivers within these tiles are again as long as the diagonal of the tile. The same is valid for the smaller tiles which are $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$, ... of the whole domain in linear size. Consequently, Peano's basin satisfies Hack's law (Eq. 9.1) qualitatively, but with a non-fractal scaling exponent $h = \frac{1}{2}$.

Since the network was generated by nesting identical patterns, the drainage areas have the same shape, so that they exhibit isotropic scaling properties instead of reproducing the elongation of natural drainage areas. As a consequence of this property and the result $h = \frac{1}{2}$, the length-scaling relation of the rivers (Eq. 9.2) turns into a non-fractal relationship with $\phi_L = 1$. Thus, neither the self-affine scaling properties of the drainage areas nor the fractal length-scaling properties of the rivers are reproduced by Peano's basin.

Figure 9.8 shows the cumulative size distribution of the drainage areas for Peano's basin on a 1024×1024 grid. The staircase-shaped distribution indicates that scale invariance is restricted to a discrete set of scales as already observed in Sect. 1.6. Apart from this effect, the distribution follows a power law with an exponent $\beta = \frac{1}{2}$. Thus, Peanos's basin reproduces the fractal size distribution of natural drainage areas (Eq. 9.3) qualitatively. However, the value of the exponent is outside the range observed in nature.

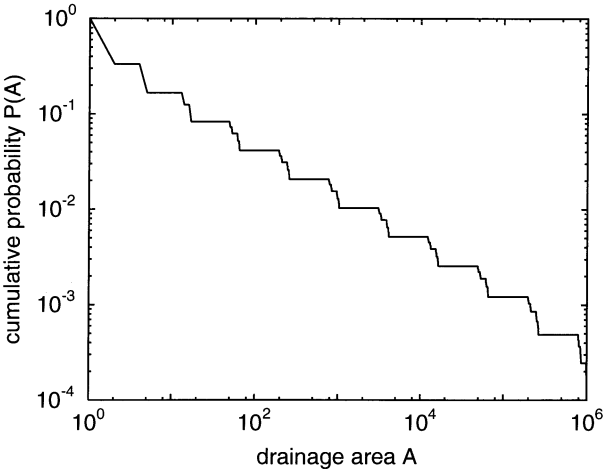

Fig. 9.8. Cumulative size distribution of the drainage areas for Peano's basin on a 1024×1024 grid.

9.4 Random-Walk Approaches

Leopold and Langbein (1962) were the first to reproduce fractal properties of drainage networks without explicitly introducing scale invariance by nesting. In their approach, a rectangular region is divided regularly into quadratic tiles. Precipitation acts uniformly on all tiles. In the first step, the source of the first river is randomly chosen. This river performs some kind of *random walk*: The drainage direction from the source is randomly chosen among the connections to the four nearest neighbors; diagonal flow directions are not allowed. Then a random flow direction is chosen for this site under the restriction that backward flow is not allowed. The process ends when the river leaves the model area. Afterwards, a second river is generated by the same procedure; it ends if it leaves the area or if it joins an existing river. Rivers are generated until the whole area is drained.

These rules are straightforward, but may result in loops in the network. Additional rules must be posed in order to prevent a river from looping back on itself. A simple way of avoiding loops is rejecting a river immediately as soon as a loop occurs, which means that it is removed from the network. Figure 9.9 shows an example of a Leopold/Langbein network, generated on a 256×256 grid using the loop-avoiding rule. Only sites with a drainage area of at least 10 tiles are assumed to be channelized, smaller flows are not plotted. The line widths are proportional to the fourth root of the discharge.

The properties of Leopold/Langbein networks can only be derived from numerical simulations. In order to provide a sufficient statistics, either a very large network must be computed or an ensemble average of several networks must be performed. Rinaldo et al. (1998) report that they partly yield reasonable fractal properties; Hack's law (Eq. 9.1) is reproduced with an exponent $h = 0.57$ in nearly perfect coincidence with the mean value of real basins. However, Leopold/Langbein rivers are considerably more convoluted than natural rivers, their length scaling exponent (Eq. 9.2) is significantly too high: ϕ_{L} is about 1.4 instead of being between 1.02 and 1.12. In return, the elongation of the drainage areas (their self-affinity) goes into the wrong direction. So the good value of h is misleading. Furthermore, this model is yields a fractal distribution of the drainage areas (Eq. 9.3), but the exponent is too high. The value $\beta = 0.52$ is even further away from the observed range than the exponent obtained from Peano's basin.

Several scientists have followed the random-walk idea, e. g., Scheidegger (1967). His model aims at simulating the drainage in alpine valleys, so he introduced a strong preferential flow direction. The simplest realization is based on a hexagonal lattice where only flow into two out of six directions is allowed, perhaps into these two pointing roughly southwards in Fig. 9.6. The resulting networks are called *Scheidegger's trees*; an example is given in Fig. 9.10. The restriction of flow directions immediately avoids the problem of loops; and as a further advantage, fractal properties of Scheidegger's trees can be computed analytically in the limit of infinite system size (Huber 1991;

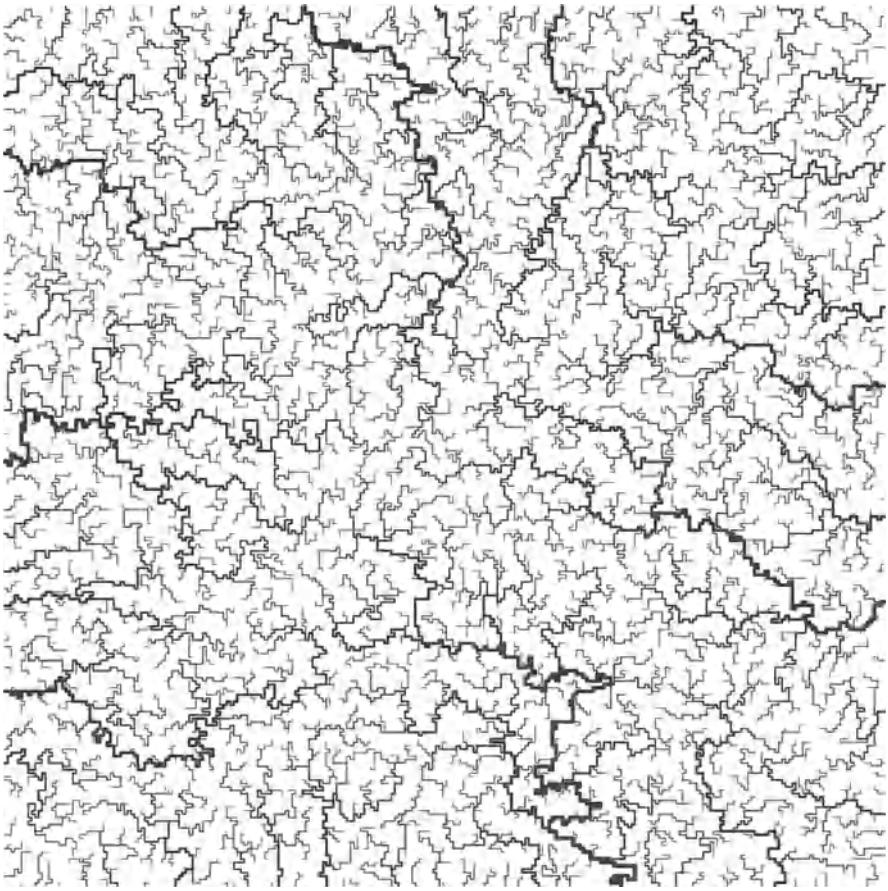

Fig. 9.9. An example of a Leopold/Langbein network computed on a 256×256 grid. Only sites with a drainage area of at least 10 tiles are plotted. The line widths are proportional to the fourth root of the discharge.

Maritan et al. 1996a). Their scaling exponents are $h = \frac{2}{3}$, $\phi_L = 1$, and $\beta = \frac{1}{3}$. Thus, these networks are quantitatively further away from natural drainage networks than those of Leopold and Langbein.

In summary, these early random-walk approaches provide some stimulating ideas on the origin of the fractal properties of drainage networks. However, the quantitative results are not so good that these models can really be a satisfying explanation for these properties.

9.5 Drainage Networks and Landform Evolution

Although nested geometric models (Sect. 9.3) and chance-dominated algorithms that describe the growth of a network (Sect. 9.4) may provide qual-

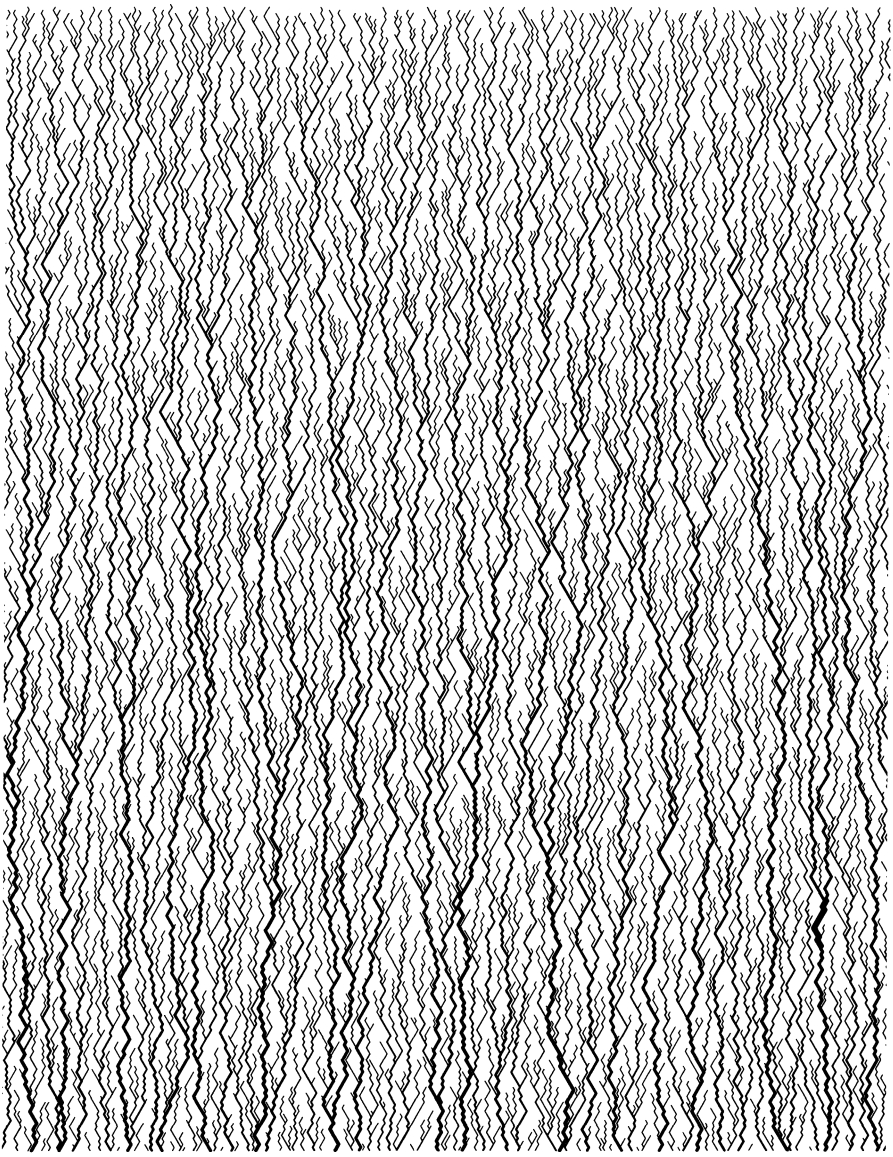

Fig. 9.10. An example of Scheidegger's trees computed on a hexagonal grid with 256×384 sites. The line widths are proportional to the fourth root of the discharge. Only river segments with a drainage area of at least 50 sites are plotted.

itatively reasonable results, their value for deepening our understanding of network organization is limited. The major problem of these approaches is their rather vague relation to more or less established process-based concepts. Believing that rivers grow from their source until they reach an ocean or another river is difficult. Thus, the random-walk approaches generate patterns by a certain algorithm without being able to explain why nature should act the same way.

Obviously, drainage patterns observed in nature are inseparably linked to the landform. Water mainly flows into direction of steepest descent, so that the pattern of valleys defines the drainage pattern in general. However, the relationship is twofold because fluvial erosion is a major landform evolution process in many climates. From this point of view, modeling drainage network evolution is modeling landform evolution by fluvial erosion, although we may run into problems if we, e. g., consider drainage patterns on landforms shaped during glacial periods.

Especially since the 1990's, a large number of landform evolution models has been developed. Among them are rather comprehensive models that distinguish between fluvial and hillslope processes (Howard 1994; Kooi and Beaumont 1996; Coulthard et al. 1998; Densmore et al. 1998; Tucker and Bras 1998), but also simplified models that focus on fractal properties of the drainage pattern (Kramer and Marder 1992; Takayasu and Inaoka 1992; Inaoka and Takayasu 1993; Rinaldo et al. 1993).

According to the ideas presented in the previous section, flow routing on a given surface is straightforward if we assume that water flows from each site to that neighbor where the steepest downslope gradient occurs. However, local depressions of the surface, which mostly result in lakes in reality, cause some trouble as the idea of flow routing in direction of the steepest descent is not applicable here. One could think of allowing a flow in uphill direction in this case, but this may result in loops in the drainage network. In a realistic model, these lakes must be treated in a more sophisticated manner, e. g., by filling them up with water (Kramer and Marder 1992; Takayasu and Inaoka 1992). However, simulations show that all lakes vanish through time under suitable boundary conditions; so we simply assume that the water vanishes at local minima of the surface without further effects.

Erosion and deposition in fluvial systems are governed by the their capability of detaching particles from the river bed and by their transport capacity. Depending on discharge, slope, and sediment properties, either or both may become the limiting factor to erosive action. Let us, for simplicity, focus on the detachment-limited case; this means that all particles that are once detached leave the basin without being deposited anywhere. In this case, balancing the downstream sediment transport can be replaced with a *local erosion rule* for the surface height H_i at the node i. Mostly, detachment of particles from the river bed is assumed to be governed by the *shear stress* τ acting on the sediment surface, so that the local erosion rule can be written

in the form

$$\frac{\partial}{\partial t} H_i \;=\; -\,f(\tau_i).$$

Obviously, this approach leads to permanently decreasing surface heights, whereas the land surface rather tends to approach a long-term equilibrium between erosion and tectonic uplift in nature. So let us enlarge the approach by an uplift rate R:

$$\frac{\partial}{\partial t} H_i \;=\; R - f(\tau_i),$$

where we assume that the uplift rate R is constant within the whole basin for convenience. For many sediments, erosion is negligible if τ is lower than a critical shear stress τ_c, so that $f(\tau) = 0$ then. Above this threshold, a power-law dependence $f(\tau) \sim (\tau - \tau_c)^p$ where $p \geq 1$ is often assumed.

We aim at a simple landform evolution model where the surface heights H_i are the variables. If the heights H_i are known at a time, the gradients Δ_i (in direction of steepest downslope gradient) can be computed. According to the water balance, the discharges q_i can be computed at all sites, too. One may suspect that discharges and slopes are the governing parameters to shear stress. Under certain assumptions on the geometries of the cross sections of the rivers and with some further knowledge on channel hydraulics, we could derive a relationship between shear stress, discharge, and slope. However, the assumptions on the geometry become crucial here because the cross sections of small and large rivers differ strongly, so that we cannot avoid using empirical relationships. The most important empirical relationship in this context was introduced by Leopold and Maddock (1953):

$$\Delta_i \;\sim\; q_i^{-m} \quad \text{where} \quad m \;\approx\; 0.5 \tag{9.5}$$

for various river segments within the same environment.

Let us assume that most natural landforms are roughly in equilibrium, so that the erosion rates $f(\tau_i)$ nearly coincide for different river segments. This implies that $f(\tau_i)$ is constant if the product $q_i^m \Delta_i$ is constant, so that $f(\tau_i)$ can be written as a function of this product. Unfortunately, determining $f(\tau_i)$ completely is difficult. The problem arises from the fact that the product is constant for many rivers, so that reliable values of $f(\tau_i)$ over a wide range of the product $q_i^m \Delta_i$ are sparse. Under this aspect, it is not surprising that various approaches are established. Rinaldo et al. (1993) assumed a staircase-shaped function where $f(\tau_i)$ is constant if $q_i^m \Delta_i$ exceeds a given threshold, and zero else. In contrast, Takayasu and Inaoka (1992) assumed that erosion acts even for low discharges and slopes according to

$$f(\tau_i) \;=\; \frac{C\, q_i\, \Delta_i^2}{1 + D\, q_i\, \Delta_i}. \tag{9.6}$$

Strictly speaking, their approach does not comply with the empirical slope-discharge relation if the denominator becomes large. The main effect of the denominator is that the erosion rate becomes constant in the limit $q_i \to \infty$;

this avoids potential numerical instabilities. On the other hand, a physical reasoning is not given; so the denominator may in fact have been introduced for numerical rather than for physical reasons. However, we will introduce a more elegant way of avoiding instabilities later, so that we do not need the denominator. So let us use the approach

$$\frac{\partial}{\partial t} H_i \;=\; R - C\,(q_i^m \Delta_i)^2$$

in the following. In order to examine to what extent the exponent m affects the results, we will not only use the empirically based value $m = 0.5$ where the approach coincides with that of Takayasu and Inaoka if the denominator is neglected, but also $m = 0.25$ and $m = 1$ for comparison.

The parameters R and C can be eliminated by rescaling the variables. Let us assume a quadratic grid topology with eight flow directions, and that we have already applied a scaling to the spatial coordinates and the discharges in such a way that the grid spacing is unity, and that the discharge of each river segment coincides with its drainage area. One can easily see that such a scaling only affects the parameter C. If we then replace the variables with

$$t \;\coloneqq\; \sqrt{CR}\,t \quad \text{and} \quad H \;\coloneqq\; \sqrt{\frac{C}{R}}\,H,$$

we obtain

$$\frac{\partial}{\partial t} H_i \;=\; 1 - (q_i^m \Delta_i)^2 \,. \tag{9.7}$$

Although our procedure was straightforward, we should keep in mind that all relationships used here can only be applied for completely channelized flow or, strictly speaking, only if each site is drained by a single river. This is reasonable on large scales, but on small scales, hillslope processes may be the governing part. Since surface runoff on hillslopes is a major process in soil erosion, it has been addressed in several studies of both empirical and theoretical character (Emmett 1970; Govers 1992; Abrahams et al. 1996; Hergarten and Neugebauer 1997). Most of the relationships obtained from these studies are qualitatively similar to these used here for both sheet-like and channelized flow. However, the problem is that the parameters in the relations strongly depend on the type of flow and on the rill geometry in case of preferential flow in rill systems. Due to the strong variation of these parameters concerning space and time, including a simple, but quantitatively reasonable approach for hillslope processes is difficult and costs much of the model's simplicity and elegance. So let us apply the equations for channelized flow on the hillslope scale, although this is not appropriate in detail.

As mentioned above, local depressions of the surface (lakes) are treated in a simplified way by assuming that the water vanishes at the deepest site without any erosive action. Thus, Eq. 9.7 has to be replaced with $\frac{\partial}{\partial t} H_i = 1$ for these locations. In addition, there should be at least one outlet o at the boundary of the model domain where water can leave the basin. Let us in the following focus on single-outlet networks. At the outlet, the slope gradient in

stream direction is undefined, so that the erosion rate cannot be determined from Eq. 9.7 there. Instead, a boundary condition for the surface height at the outlet must be introduced; we assume an equilibrium between erosion and uplift here, i. e., $\frac{\partial}{\partial t} H_o = 0$.

The implementation of this model on a computer is straightforward. The evolution of the land surface is simulated in discrete time steps of length δt. Each time step is subdivided into two tasks: First, the drainage direction, i. e., the index of the site d_i where the site i delivers its discharge to, and the discharge q_i are computed for each site i. Computing the drainage direction just requires selecting that of the eight (or less, at the boundaries) neighbors where the steepest downslope gradient occurs. For computing the discharge q_i according to Eq. 9.4, we need the discharges of all sites which deliver their discharge to the site i. However, this sounds more complicated than it is and can easily be implemented recursively.

Finally, the discharges are used to update the surface heights according to Eq. 9.7. The downslope gradient is approximated by

$$\Delta_i \approx \frac{H_i - H_{d_i}}{\mathrm{dist}(i, d_i)}$$

where the distance $\mathrm{dist}(i, d_i)$ is either 1 or $\sqrt{2}$. The simplest algorithm for computing the heights at the time $t+\delta t$ is based on the explicit Euler scheme (Appendix A):

$$H_i(t+\delta t) = H_i(t) + \delta t \left(1 - \left(q_i(t)^m \frac{H_i(t) - H_{d_i}(t)}{\mathrm{dist}(i, d_i)} \right)^2 \right).$$

As nearly all explicit methods, this scheme becomes unstable if δt is too large. A rough analysis of stability can be performed by assuming that an equilibrium between erosion and deposition has already been achieved, so that all the surface heights remain constant through time. Then, a little disturbance δH is applied to the surface height at a single site i. If we assume $\mathrm{dist}(i, d_i) = 1$ for simplicity, inserting the disturbed height $H_i(t)+\delta H$ into Eq. 9.7 leads to

$$H_i(t+\delta t) \approx H_i(t) + \delta H \, (1 - 2\, \delta t \, q^m) .$$

Thus, a little perturbation grows in each time step if $\delta t > q^{-m}$. This growth continues until the deviation is so strong that the flow directions change, so that finally the solution is far off from the equilibrium solution. Thus, the time step length must be limited; and this limitation depends on the grid size: Since the maximum discharge equals the total number of sites, larger grids require shorter time steps. As a consequence, the numerical effort grows as the model's size increases not only because more sites must be treated, but also because more time steps are needed until a given time interval is simulated. This problem is a common feature of explicit schemes.

In this model, the problem can be avoided in two ways. Modifying the erosion rate in such a way that it remains finite even in the limit $q_i \to \infty$ is one approach, e. g., by introducing a denominator as shown in Eq. 9.6. But then, the empirical relation between slope and discharge (Eq. 9.5) is no longer satisfied for large rivers. Alternatively, stability can be achieved without affecting the physics of the model with the help of a partly implicit scheme (Appendix A). For this, we discretize Eq. 9.7 in the following way:

$$\begin{aligned} \frac{H_i(t+\delta t) - H_i(t)}{\delta t} &= \left(1 - (q_i(t)^m \, \Delta_i(t+\delta t))^2\right) \\ &= \left(1 - \left(q_i(t)^m \, \frac{H_i(t+\delta t) - H_{d_i}(t+\delta t)}{\mathrm{dist}(i, d_i)}\right)^2\right). \end{aligned}$$

Thus, we have kept the explicit discretization of the discharges, so that they can still be computed in a first step, but have introduced an implicit discretization of the slope gradients. The solution of the quadratic equation is

$$\begin{aligned} H_i(t+\delta t) &= H_{d_i}(t+\delta t) - \frac{\mathrm{dist}(i, d_i)^2}{2\, q_i^{2m}\, \delta t} \\ &+ \sqrt{\left(\frac{\mathrm{dist}(i, d_i)^2}{2\, q_i^{2m}\, \delta t}\right)^2 + \frac{\mathrm{dist}(i, d_i)^2 \, (H_i(t) + \delta t - H_{d_i}(t+\delta t))}{q_i^{2m}\, \delta t}}. \end{aligned} \tag{9.8}$$

The second solution of the quadratic equation differs by a minus sign in front of the square root. But since this solution does not converge towards $H_i(t)$ for $\delta t \to 0$, Eq. 9.8 is the solution we are looking for.

Equation 9.8, written for all sites i, constitutes a triangular system of nonlinear equations; the triangular structure refers to the fact that each equation only involves the site itself and its downstream neighbor. Such a system can easily be solved by starting at the outlet and follow the branches of the tree in upstream direction. This procedure is opposite to that used for computing the discharges; however, a recursive implementation is possible here, too.

The partly implicit scheme avoids the instability discussed above; Eq. 9.8 yields the correct asymptotic behavior

$$\lim_{\delta t \to \infty} q_i^m \, \frac{H_i(t+\delta t) - H_{d_i}(t+\delta t)}{\mathrm{dist}(i, d_i)} = 1.$$

However, this result does not imply that arbitrary time steps can be used; the accuracy of each scheme, no matter whether explicit or implicit, decreases if δt increases. Here, the discharges q_i are the crucial point because their variation through time is discontinuous. If a single site changes its drainage direction due to a small variation in the surface heights, the discharge may change significantly in large parts of the basin. The proper solution of this problem is clear: The time steps should be so short that flow routing persists during the whole time step, while it changes exactly at the end of a step. Unfortunately, performing an adaptive variation of δt according to this criterion is

cumbersome. For this reason, we introduce a simpler adaptive variation by accepting the error resulting from changing the flow direction at one site during a time step. If two or more sites change their drainage direction, the time step is rejected, and we switch to time steps of half length in the following. In return, we must introduce a criterion for coarsening the time steps if nothing happens; we increase δt by a factor two if the entire drainage pattern persists through a time step.

Simulations show that the surface (and consequently the network) approaches a steady state after some time. According to Eq. 9.7, the slope gradients are uniquely determined by the discharges then: $\Delta_i = q_i^{-m}$. Consequently, the topology of the drainage network, defined by the flow direction d_i at each node i, characterizes the land surface uniquely in case of stationarity. In a first step, the discharges can be computed from the flow directions, and then the surface heights are given by

$$H_i \;=\; H_{d_i} + \mathrm{dist}(i, d_i)\, \Delta_i \;=\; H_{d_i} + \mathrm{dist}(i, d_i)\, q_i^{-m}\,. \tag{9.9}$$

This procedure can be applied to any single-outlet network which is topologically consistent. The latter means that there are no loops or lakes, so that the outlet captures the whole precipitation. However, it is not clear that the resulting surface complies with the pre-defined flow directions of the network, i. e., that each site is drained in direction of steepest descent. The left-hand part of Fig. 9.11 illustrates that, e. g., Peano's basin (Sect. 9.3) does not meet this criterion. The network was generated on a 128×128 lattice; the drainage direction of 3122 out of 16,384 sites do not coincide with the direction of steepest descent. These sites are grey-shaded in the plot. Inconsistencies between drainage direction and surface gradient preferably occur at drainage

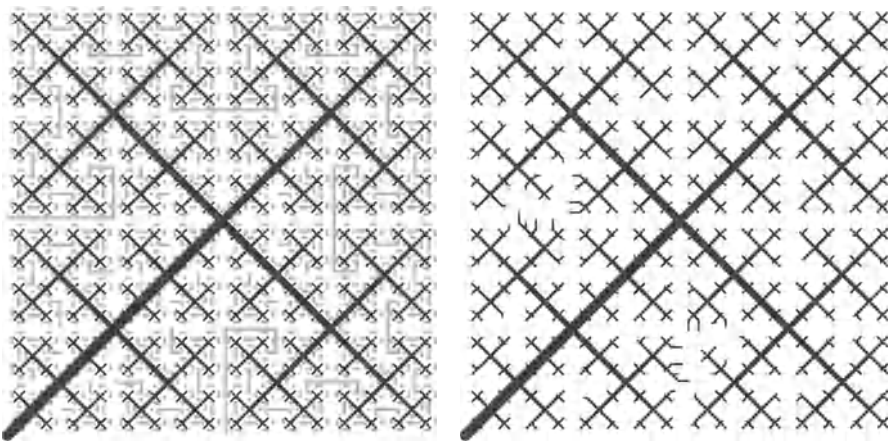

Fig. 9.11. Left: Peano's basin on a 128×128 lattice. Those sites where the drainage direction is not consistent with the surface gradient are grey-shaded. Right: ECN based on Peano's basin.

divides where the lengths of the paths towards the outlet differ between the two watersheds. Let us use this property for a definition which is not established, but useful in the following:

> A drainage network is called *equilibrated channel network* (ECN) with respect to a local erosion rule if it can be placed on a stationary surface in such a way that the drainage direction coincides with the direction of steepest descent everywhere.

As already stated, Peano's basin is not an ECN, at least not with respect to the erosion rule with $m = 0.5$. However, we can use the (inconsistent) surface as an initial condition for a simulation and wait until a steady state is approached. In some cases, the resulting ECN is not far away from the original network; the right-hand diagram in Fig. 9.11 shows an ECN which is similar to Peano's basin. Obviously, the fractal properties of this ECN are determined by the original Peano basin rather than by the landform evolution model. Thus, we should look for a certain class of ECNs here – those where the initial condition does not affect the final network strongly. Under this aspect, a flat and smooth initial surface seems to be ideal, but this may induce some unwanted symmetry. So it is better to start at a nearly flat surface where the initial heights are sufficiently small random values. Let us call ECNs resulting from a simulation with such an initial condition *unstructured* ECNs.

Figure 9.12 gives the evolution of an unstructured ECN on a grid with 128×128 nodes for $m = 0.5$. The initial surface heights were uncorrelated random numbers drawn for a uniform distribution between -0.01 and 0.01. The outlet location o was randomly chosen. The network grows like a tree, starting at the outlet. At $t \approx 17$, the network covers the entire model area for the first time. Afterwards, the network is partly reorganized; the lower plots show that a large river vanishes completely. This river is in the middle between two other large rivers, and three large rivers being nearly parallel are obviously too many. The two other rivers have larger drainage areas, and thus larger discharges. Therefore, the erosion of the river in the middle cannot keep track with that of the others until finally its valley is higher than the others. In return, the higher valley loses drainage area to the adjacent valleys until it finally vanishes and turns into a crest. At $t \approx 136$, network reorganization is finished, so that we finally arrive at an ECN.

Due to the small statistics, a single ECN like that shown in Fig. 9.12 is not sufficient for recognizing or predicting fractal network properties such as Hack's law (Eq. 9.1) or the distribution of drainage areas (Eq. 9.3). For this reason, we must either simulate ECNs on a much larger grid or perform an ensemble average over several ECNs. Obviously, effects of the finite grid size which may hinder recognizing power-law behavior are less significant in a single large-grid simulation than in an ensemble average. However, the numerical effort increases faster than linearly with increasing model size, so that a single simulation with $k \times n$ nodes is more expensive than k simulations

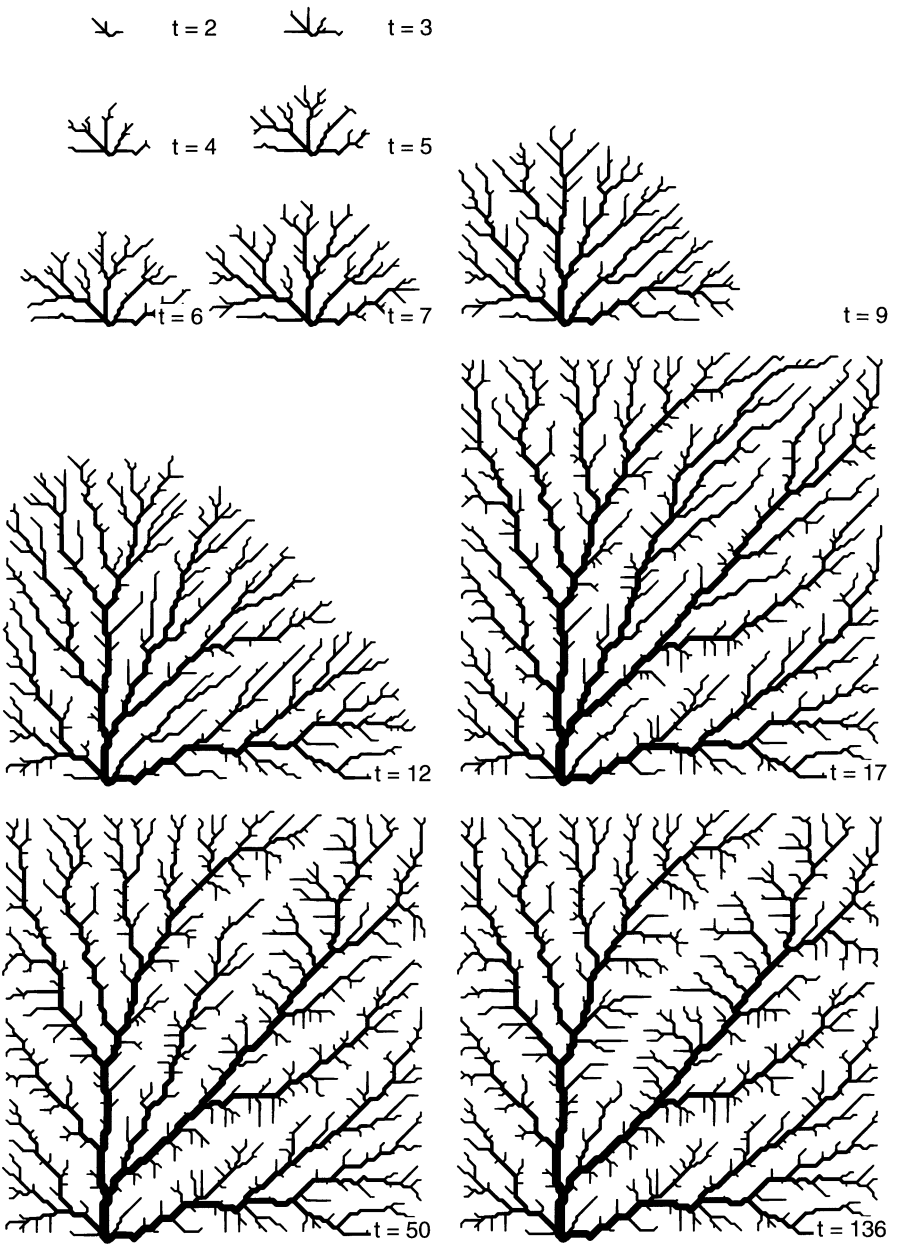

Fig. 9.12. Evolution of an unstructured ECN. Only sites which deliver their discharge to the outlet and whose drainage area is at least 10 tiles are plotted. The line widths are proportional to the fourth root of the discharge. The network in the middle on the right-hand side ($t = 17$) describes the stage where the network first drains the whole area, i. e., immediately after the last lake has vanished; the lower right one shows the final state.

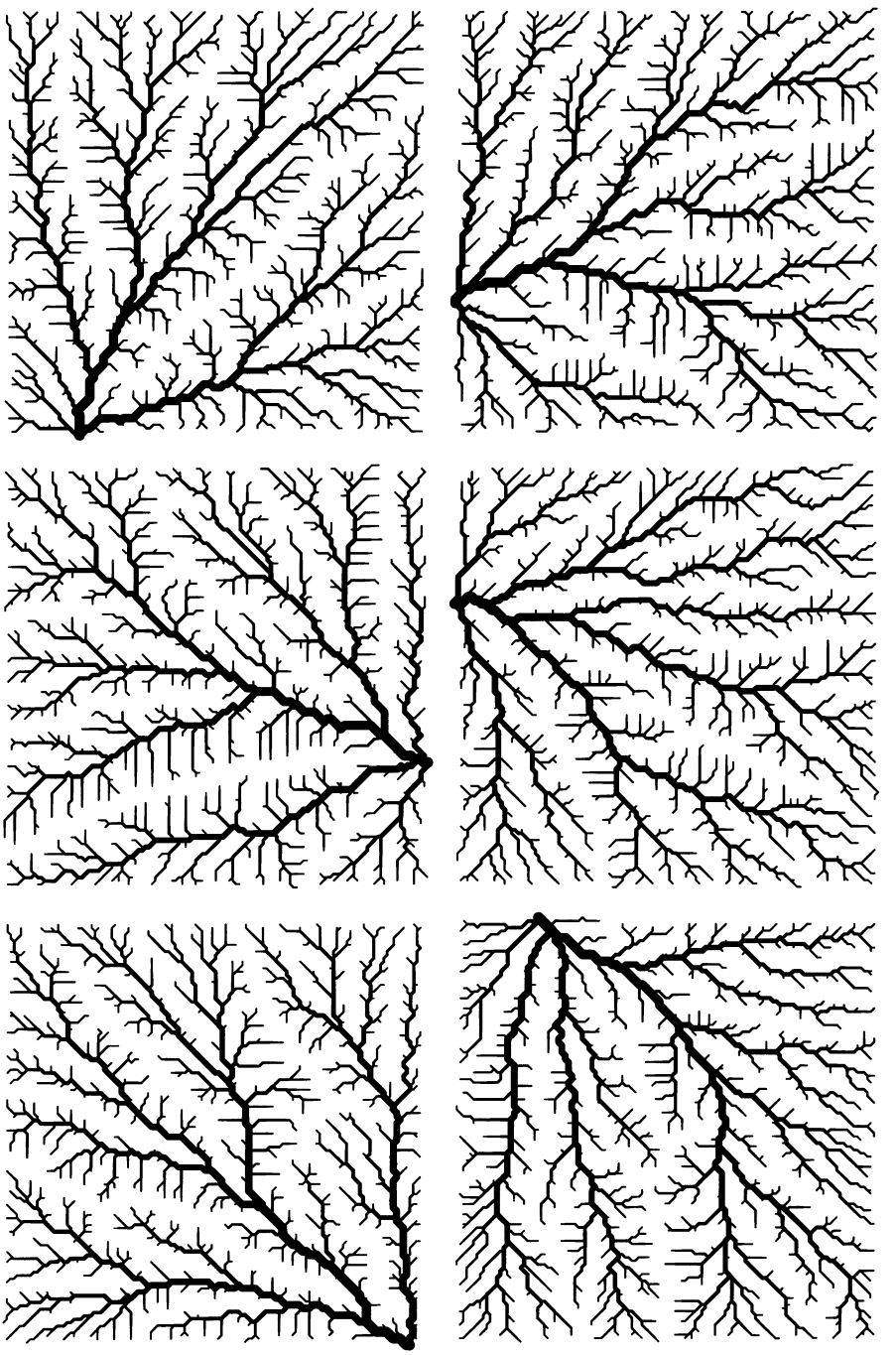

Fig. 9.13. Six examples of unstructured ECNs with random initial conditions and randomly chosen outlet locations.

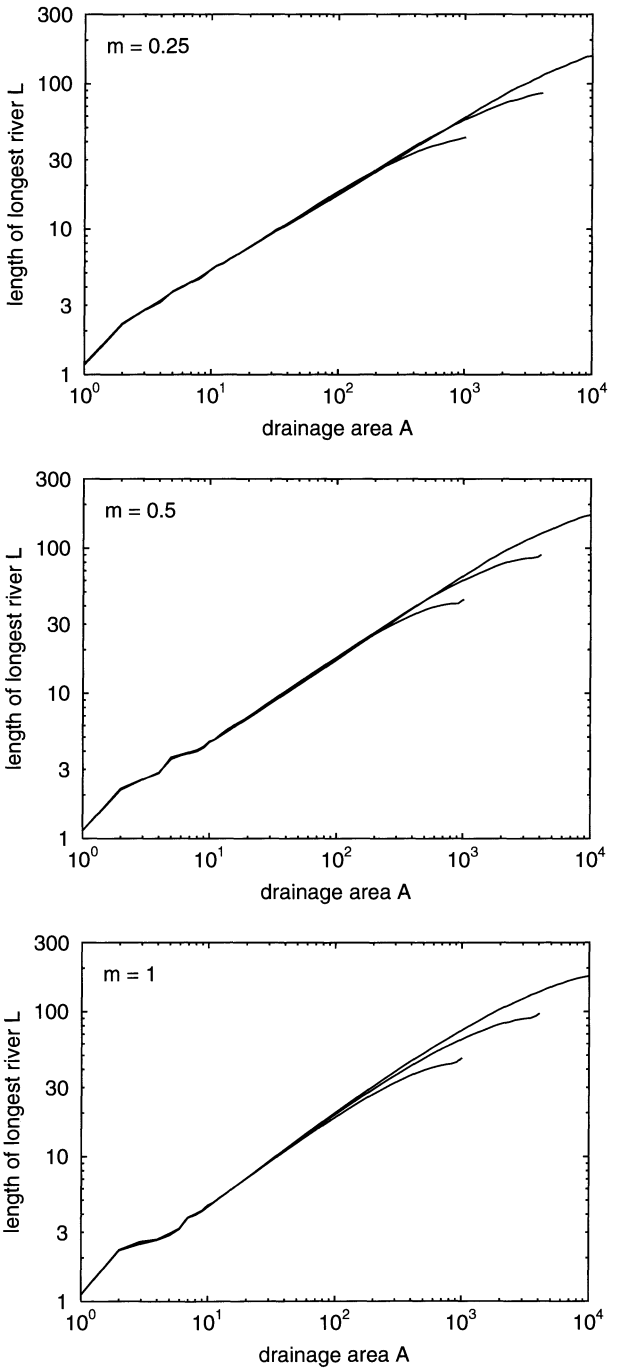

Fig. 9.14. Lengths of the longest rivers in a drainage area versus the size of the drainage area (Hack's law), obtained for different values of m on grids of different sizes (32×32, 64×64, and 128×128) sites.

with n nodes each, while both provide comparable statistics in sum. For this reason, we prefer an ensemble average over several ECNs on lattices of 128×128 nodes with randomly chosen outlet locations. For assessing finite-size effects, we compare the results obtained from 1500 ECNs on a 128×128 lattice with those from 6000 ECNs on a 64×64 grid, and 24,000 ECNs on a 32×32 grid. Figure 9.13 shows six ECNs from the simulated ensemble.

The analysis of the properties related to Hack's law (Eq. 9.1) is given in Fig. 9.14. For each site of each ECN, the size of the drainage area A and the length of the longest river L within the drainage area are computed. In order to avoid statistical noise, the axis of the areas is subdivided into small bins, and the lengths are averaged over these bins. Except for a finite-size effect, the plots show power-law behavior and thus reproduce Hack's law at least qualitatively. The deviations at small areas ($A < 10$) seem to be an effect of the discretization and can be ignored. The exponents are $h = 0.52$ for $m = 0.25$, $h = 0.58$ for $m = 0.5$, and $h = 0.64$ for $m = 1$. If we remember that realistic values for the parameter m are about 0.5, this result is in perfect coincidence with the observed range of h between 0.52 and 0.6.

Unfortunately, the perfect coincidence with nature ends at this point. Figure 9.15 shows the relationship between river length and distance between source and reference point (Eq. 9.2), again analyzed for each site of the ECNs and averaged over small bins. Finite-size effects are significant here, so that plotting the data of the largest grid is sufficient. For all three parameter values, a power-law relation emerges, except for short rivers where artefacts of the discretization occur. However, the exponent ϕ_L is always between 1 and 1.01, and this cannot be considered as a fractal scaling relation. As mentioned in Sect. 9.1, ϕ_L is not necessarily far away from unity, but compared to the observed range between 1.02 and 1.12, the obtained result is not satisfying.

Similar problems were already encountered in Peano's basin and in the random-walk model of Leopold and Langbein (1962); although they are more

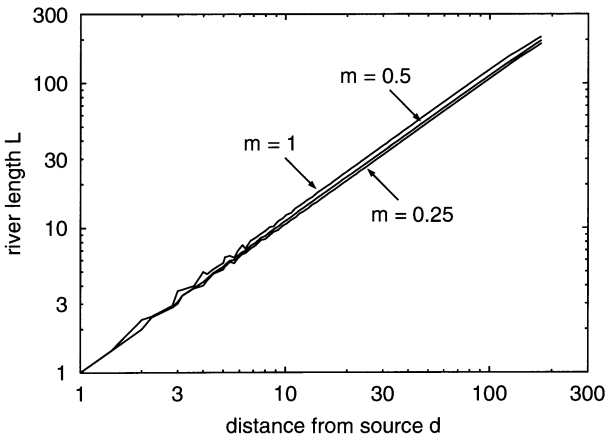

Fig. 9.15. Length scaling of the rivers for different values of m.

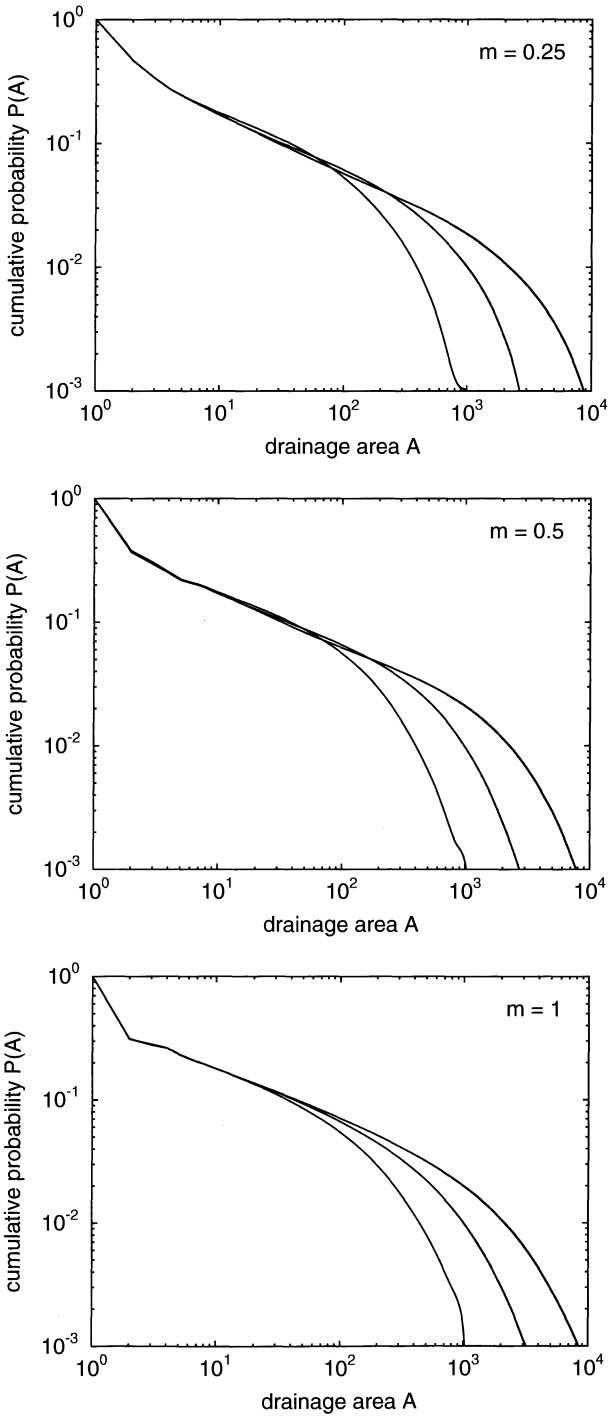

Fig. 9.16. Cumulative size distribution of the drainage areas, obtained for different values of m on grids of different sizes (32×32, 64×64, and 128×128 sites).

severe there. Hack's law is reproduced well, but the role of the two contributions discussed in Sect. 9.1 is not correct. Here, the contribution of fractal length scaling of the rivers is underestimated, while the elongation of the drainage areas is exaggerated. Both errors compensate each other in sum, so that finally Hack's law is reproduced well, but the coincidence is misleading.

Reproducing the power-law size distribution of the drainage areas (Eq. 9.3) causes some problems, too. Figure 9.16 shows the distribution, obtained from the simulated ECNs on grids from 32×32 to 128×128 nodes. The plots suggest that the distributions approach power laws for large lattices, but the behavior strongly depends on the parameter m. For $m = 1$, extrapolating towards infinite model size is straightforward and leads to an exponent $\beta = 0.39$, which is slightly too low compared to the observed range between 0.41 and 0.46. However, $m = 1$ is neither realistic, nor reproduces Hack's law well.

The plots for $m = 0.25$ and $m = 0.5$ have larger slopes, leading to larger and thus more realistic exponents β. However, the curves exhibit a convex curvature at intermediate areas. As a consequence, the power laws are not very clean. From the data shown here, it is not clear whether this is a finite-size effect or a principal deviation from power-law behavior. The distribution obtained by Takayasu and Inaoka (1992) is very similar to that shown in the middle, and they fitted a power law with $\beta = 0.43$. This is easy as long as we know that this value is in perfect coincidence with the exponent observed in nature. Figure 9.17 shows the raw data for $m = 0.5$, i. e., the non-cumulative number of sites with a given drainage area. The peaks at the right-hand end of the curves look strange, but are not a problem. They arise from the fact that each ECN has exactly one site (the outlet) which drains the whole basin. However, none of the curves exhibits a clean power-law behavior over a reasonable range. Depending on the range of fitting, even exponents lower than 0.4 or greater than 0.55 (transferred to the cumulative distribution) are

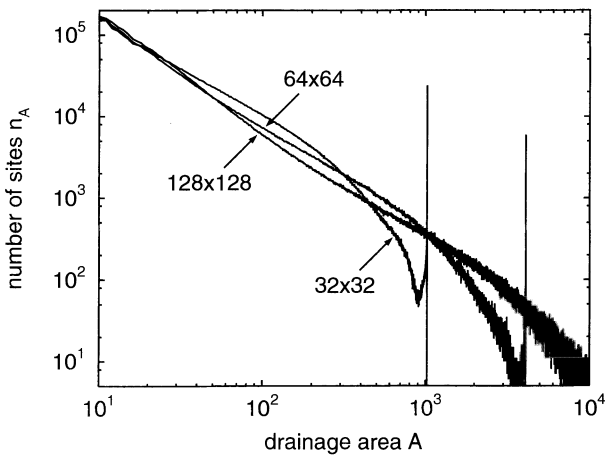

Fig. 9.17. Non-cumulative number of sites with a given drainage area for $m = 0.5$, obtained on grids of different sizes.

obtained. With respect to this poor power-law behavior, the result $\beta = 0.43$ of Takayasu and Inaoka (1992) should not be overinterpreted.

In summary, unstructured ECNs provide a better explanation of fractal properties of drainage networks than the purely geometric approach of Peano's basin (Sect. 9.3) and the random-walk approaches (Sect. 9.4) do. Since they are, in addition, much closer to existing knowledge on landform evolution and explain networks as the result of a physically reasonable process, they considerably deepen understanding of this phenomenon. However, the deficiencies revealed by the quantitative analysis concerning the length scaling relation and the size distribution of the drainage areas show that there is still room for improvements.

9.6 Optimal Channel Networks

In the previous section we have learned that simple landform evolution models result in drainage networks with partly realistic fractal properties, but reproduce the fractal size distribution of drainage areas not very well. May this be a result of an incomplete organization? Let us revisit the network evolution shown in Fig. 9.12. After the network once drained the whole model area, some further reorganization took place. Obviously, the network emerging from the evolution up to this point was not good enough in some sense, so that one large river had to be removed later. But how do we know that this reorganization was sufficient? Maybe, the surface achieved a steady state before the network became perfect, although we cannot tell what this exactly means.

Rodriguez-Iturbe et al. (1992b) proposed the total energy dissipation of the water flowing through the network as a criterion for the quality of a drainage network. Under the assumption of steady-state flow (not necessarily equilibrium between erosion and uplift), the energy expenditure of a river segment is the product of discharge, height difference between the end points, and specific weight of water. If we use the non-dimensional variables introduced earlier and rescale the energy in such a way that the specific weight becomes unity, the energy expenditure of the whole network is

$$E = \sum_{i \neq o} q_i \left(H_i - H_{d_i}\right) = \sum_{i \neq o} \left(H_i - H_o\right),$$

provided that the outlet o captures the precipitation of the entire basin. For ECNs, E can be computed from the discharges and flow directions without explicitly knowing the surface heights; from Eq. 9.9 we obtain

$$E = \sum_{i \neq o} \mathrm{dist}(i, d_i)\, q_i^{1-m}. \tag{9.10}$$

Figure 9.18 shows the cumulative distribution of the energy expenditure; each curve represents an ensemble of 1500 ECNs, computed on a 128×128 lattice.

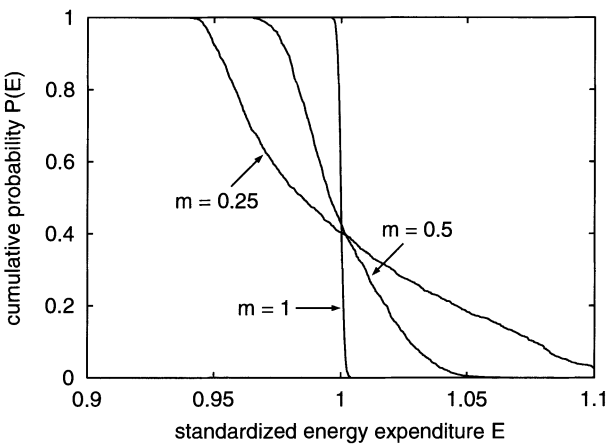

Fig. 9.18. Cumulative distribution of the energy expenditure for unstructured ECNs.

The expenditures were standardized so that the average expenditure of each ensemble is unity. Rescaling is necessary in order to obtain comparable values which can be plotted into a single diagram because we can easily see from Eq. 9.10 that lower values of m lead to higher values of E. The distribution becomes narrower if m increases; the relative standard deviation is about 5 % for $m = 0.25$, 2 % for $m = 0.5$, and 0.1 % for $m = 1$. The narrow distribution for $m = 1$ is not surprising; from Eq. 9.10 one can see that the result only depends on the numbers of diagonal and non-diagonal river segments, but not on the discharges in this case.

Let us now return to the parameter value $m = 0.5$ and examine whether there is any relationship between the energy expenditure and the fractal properties of the networks. It turns out that both Hack's law (Eq. 9.1) and the length-scaling relation (Eq. 9.2) are mainly independent from the energy expenditure, while there is a relation to the size distribution of the drainage areas (Eq. 9.3). For a quantitative analysis, we analyze those 10 % with the lowest energy expenditure and those 10 % with the highest energy expenditure of the 1500 simulated ECNs separately. Figure 9.19 shows that ECNs with a low energy expenditure come closer to a clean power-law distribution than those with a high energy expenditure. From this point of view, ECNs with a lower energy expenditure are better organized than those with a higher energy dissipation. Rinaldo et al. (1992) coined the term *optimal channel networks* (OCNs) for those ECNs which drain a given area with the minimum energy expenditure; this concept turned out to be quite fruitful.

However, even the unstructured ECN with the lowest energy dissipation is not necessarily an OCN. Since the process of evolution from a nearly plain surface is a quite strict constraint, all these networks may be far away from the state of minimum energy expenditure. For this reason, an optimization procedure must be applied in order to obtain OCNs in the strict sense. Since only steady-state networks are considered in the process of optimization, Eq. 9.10

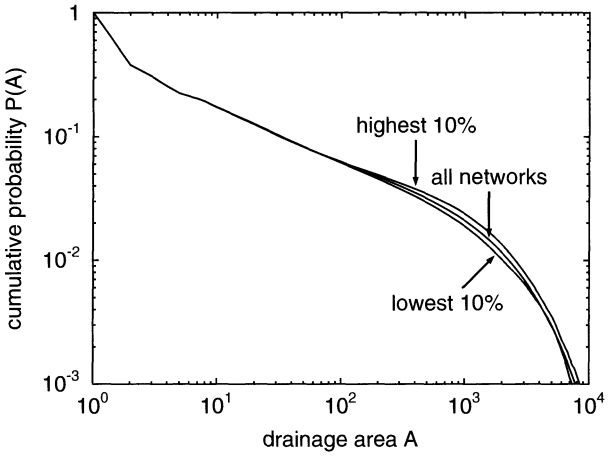

Fig. 9.19. Cumulative size distribution of the drainage areas for ECNs on a 128×128 grid, distinguished by energy expenditure.

is the best basis for the minimization. This equation allows an optimization of the networks without keeping track of the surface.

Since the number of network topologies on a lattice of finite size is limited, the optimization problem is a discrete one. However, the number of possible networks is huge; if we can choose among eight flow directions for each site, there are 8^{N^2} combinations on a $N \times N$ grid. Even if we can exclude those combinations which lead to loops in the network, the number of combinations remains high; e. g., Scheidegger's trees (Sect. 9.4) which have only two possible flow directions per node, lead to $2^{256} \approx 10^{77}$ allowed networks on a (very small) 16×16 grid. Thus, just comparing all allowed network topologies is hopeless. Instead, iterative methods which start at an initial network and optimize it successively must be applied. An unstructured ECN may be the starting point for the optimization, but networks generated by artifical algorithms such as Peano's basin may be used, too. A straightforward algorithm for network optimization is randomly changing the flow direction at an individual site and computing the resulting change in energy expenditure. The change in network topology is accepted if energy dissipation has decreased, otherwise it is rejected.

Several studies were performed in order to elucidate fractal properties of OCNs. Early results (Rinaldo et al. 1992, 1993; Rigon et al. 1993) suggested a power-law distribution for the sizes of the drainage areas (Eq. 9.3) with exponents β between 0.41 and 0.45. However, the power laws turned out to be not much cleaner than those obtained from unstructured ECNs, at least not cleaner than those obtained by selecting the best networks from an ensemble of ECNs. Thus, the perfect coincidence of the obtained exponents with those observed in nature should not be overinterpreted here, too.

Further studies on OCNs (Sun et al. 1994, 1995; Colaiori et al. 1997; Rinaldo et al. 1998) revealed significant differences between globally optimal networks (where the energy expenditure achieves a global minimum) and

those which are only optimal in a local sense. The latter means that energy expenditure cannot be further reduced by applying local changes to the network topology. Obviously, the simple optimization procedure illustrated above only guides us towards local minima, and getting stuck at local minima is a common problem in many minimization algorithms. Several strategies for getting around this problem have been applied in order to obtain globally optimal networks. The major result is that globally optimal networks exhibit considerably cleaner power-law distributions than those which are only locally optimal, especially for the size distribution of the drainage areas. On the other hand, the exponents themselves are disappointing; globally optimal configurations exhibit the same scaling properties as Peano's basin: $\beta = \frac{1}{2}$, $h = \frac{1}{2}$, and $\phi_L = 1$ (Maritan et al. 1996a; Colaiori et al. 1997; Rodriguez-Iturbe and Rinaldo 1997). Thus, β is outside the range observed in nature, and the fractal character of Hack's law (Eq. 9.1), i. e., the deviation of h from $\frac{1}{2}$ is not reproduced at all. Consequently, Maritan et al. (1996a) conclude that globally optimal channel networks do not describe the properties of real river basins appropriately.

Apart from these aspects, there is another problem with the OCN concept. How does nature manage to minimize energy expenditure, and what does nature know about energy expenditure at all? Let us not sink into philosophic questions here, but just emphasize that the principle of minimum energy expenditure is still unproven. There is no doubt that deriving properties of a system from extremum principles is tempting and well-established in physics. The Euler-Lagrange equations are the bridge between variational principles and partial differential equations. However, the principle of minimum energy expenditure is not an inherent property of the erosion models discussed here, so it cannot be derived formally from a local erosion rule such as Eq. 9.7. There were attempts to derive the condition of stationary networks from a variational principle such as that of Sinclair and Ball (1996), but their approach suffers from two problems: First, their ideas on the transfer from steady-state networks to a condition for the surface are in clear contradiction to established landform evolution equations as well as to the steady-state version of Eq. 9.7. Moreover, they derive a function which becomes minimal or maximal (strictly speaking, stationary) in the steady state of network evolution according to their interpretation of the steady state, but this property is not related to energy expenditure in general. Thus, the relevance of the principle of minimum energy expenditure can be reasoned theoretically only by introducing additional physics instead of manipulating the existing equations mathematically.

As we cannot proof the principle of minimum energy expenditure, we have exactly arrived at the question discussed in Sect. 8.3. Are OCNs better than unstructured ECNs? If so, it is just the cleaner power-law distribution of the drainage area's sizes which makes them better. However, everyone who has ever seen a power law obtained from real-world data knows about their

often poor quality. In fact, observed power-law distributions of drainage areas (Rodriguez-Iturbe et al. 1992a; Maritan et al. 1996b) are often less clean than the worst obtained from ECNs.

Consequently, data obtained from nature cannot really help us to decide whether OCNs are better than unstructured ECNs or not. This question can only be answered if we believe that fractal distributions are something fundamental, and that drainage networks found on earth are not perfect in this sense. Rinaldo et al. (1998) give a discussion why natural drainage networks deviate from OCNs; finding reasons for the imperfection of nature is not a problem because nature provides a variety of disturbances and inhomogeneities. Nevertheless, since both nature and models are exposed to fluctuations, a level between both, some kind of artificial reality, is necessary for combining them to a deeper understanding of the processes behind the patterns. However, we should be aware that we may be walking on thin ice.

9.7 Drainage Networks and Self-Organized Criticality

Since SOC is the main topic of this book, we should now close the loop and discuss the models from this chapter with respect to SOC. We have learned about several different concepts. Some of them are not linked to a surface at all, while the others focus on equilibrium landforms. It was soon recognized that this behavior does not meet the criteria of SOC. As discussed in Chap. 5, SOC requires that the system evolves towards a strange attractor where fluctuations of all sizes occur. The fractal size distribution of these events is the spatial fingerprint of SOC systems, whereas the fractal properties of the drainage networks discussed here are static.

Obviously inspired by the ideas of SOC, both Takayasu and Inaoka (1992) and Rinaldo et al. (1993) interpreted the behavior of their networks as a new type of SOC, entitled *spatial SOC*. However, discussion (e. g. Sapozhnikov and Foufoula-Georgiou 1996a) has confirmed that this behavior is far away from SOC in the original sense, and that considering it a new type of SOC does not provide new insights.

In principle, this would be a good conclusion of this chapter. Why should all fractals occurring in nature be a result of SOC? From this point of view, this negative example could at least caution the reader not to believe that SOC explains the whole world.

On the other hand, the earth's surface is in fact not in a steady state, but exposed to permanent changes. Climate varies on several time scales, tectonic uplift may change through time, and even variations in erodibility due to different layers of bedrock may prevent the land surface from becoming stationary. However, we can pick up the arguments from the previous section and define our own, idealized reality where everything is homogeneous and constant through time, and finally discuss why nature is not perfect. Imperfection has two facets here, a spatial and a temporal. Spatial inhomogeneity

is ubiquitous in both the climatic conditions and the geological properties of the earth's crust; effects of spatially inhomogeneous model parameters in OCNs were already investigated by Maritan et al. (1996a) and by Caldarelli et al. (1997). However, temporal variations may be still more interesting. Can we assume that steady states are the perfect situation, and that nature just runs behind such states? If so, we can expect that temporal variability is just another source of imperfection.

Let us play a little with the landform evolution model discussed in Sect. 9.5. As soon as a steady state (an unstructured ECN) has been reached, the boundary conditions are changed drastically by introducing a randomly chosen, second outlet somewhere at the boundary. Since we do not just want to introduce more and more outlets, we assume that the second outlet is eroded at the same pre-defined rate as the original one was, while erosion at the original outlet ceases. For being more precise, we assume that water may still leave the basin at the old outlet as long as it is a local minimum of the surface, but without any erosive action. As a result, a new drainage network will grow from the new outlet, while erosion within the drainage area of the old outlet decreases as a consequence of decreasing slope gradients. After some time, the growing drainage area of the new outlet will reach the old one, so that the old outlet will not be a local minimum of the surface any longer. At this time, we assume that water does not leave the basin here any more, but flows into direction of steepest descent according to the regular flow routing scheme. Finally, the surface becomes stationary again; then a new outlet is chosen and the procedure is repeated. Figure 9.20 shows an example of the transition from one ECN to another. The black part of the network corresponds to the drainage area of the new outlet, while the grey part corresponds to that of the old one.

From the view of an earth scientist one may doubt whether this kind of game provides an appropriate representation of temporal variability in landform evolution. Surely, effects induced by changing tectonic conditions or variable erodibility due to different layers of bedrock are not as drastic as realized here. On the other hand, changing the outlet location is one of the simplest approaches regarding the fact that a drainage basin is not an isolated object, but part of an environment which may change through time. At least, this approach is not more unrealistic than assuming that drainage networks evolve from a nearly flat surface or that they know anything about energy expenditure.

Let us call ECNs resulting from changing the location of the outlet *sequential* ECNs. Figs. 9.21 and 9.22 show a series of 12 sequential ECNs, simulated on a 128×128 lattice. As in the previous sections, the parameter m in Eq. 9.7 is set to $m = 0.5$. The transition from the fifth to the sixth ECN was already shown in detail in Fig. 9.20. Those parts of the networks having changed since the previous steady state are marked black; grey segments have persisted. Despite the drastic changes applied through the variable boundary

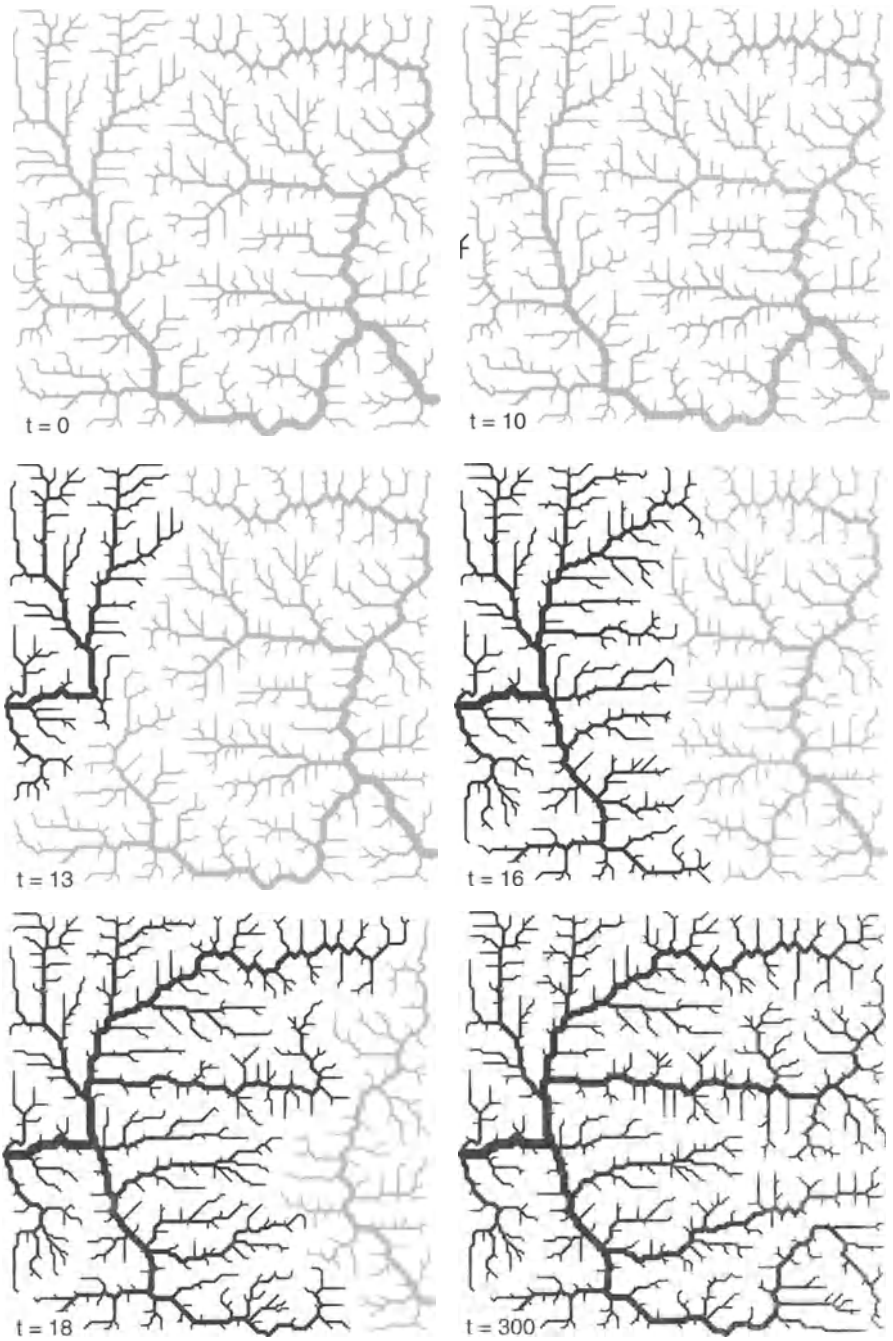

Fig. 9.20. Transition from one ECN to another as a result of changing the location of the outlet.

condition, only parts of the network are reorganized as a response to introducing a new outlet. Parts of these alterations simply result from changing the flow direction, while the valley remains. However, this does not mean that large parts of the network persist over long times; more than 90 % of the network are reorganized during the 12 steps shown here.

Since our series of sequential ECNs starts at an unstructured ECN, the first networks of a series may be biased by this initial condition and hence not be representative. Thus, we should skip all results until all relicts of the initial condition are wiped out, as we did in the self-organizing models before. After about 1300 equilibrated networks were computed, each node had changed its drainage direction at least 25 times; this should be more than enough for wiping out any initial structure. Then, 5000 sequential ECNs were computed and analyzed. During this simulation, each site changed its drainage direction at least 144 times, and in the mean 394 times. This ensures that the simulation provides a sufficient statistics which is not biased by persisting structures, although consecutive ECNs are not independent.

Figure 9.23 gives an analysis of the resulting size distribution of the drainage areas. The statistics include 5000 networks on the 128×128 grid, 20,000 networks on the 64×64 grid, and 80,000 networks on the 32×32 grid. Surprisingly, the plot shows a quite clean power-law behavior; and the deviations from power-law behavior look like a finite-size effect. Figure 9.24 confirms that the power-law distribution emerging from sequential ECNs is in fact considerably cleaner than that obtained from unstructured ECNs. The power law is even cleaner than that obtained by selecting those 10 % of the unstructured ECNs with the lowest energy expenditure, which were used as a first step towards OCNs in Sect. 9.6. The exponent β is determined to $\beta = 0.46$. In contrast to the reference models, the clean power law allows a quite precise estimation of the exponent. Its value is at the upper edge of the observed range between 0.41 and 0.46 and thus much more realistic than the value $\beta = 0.5$ for globally optimal channel networks.

If we take both the quality of the power-law distribution and the value of the exponent as criteria for the quality of the model, this model is the best one among these discussed in this chapter. However, before being too enthusiastic, let us check whether these promising results persist for Hack's law (Eq. 9.1) and for the fractal length-scaling relation of rivers (Eq. 9.2). Figure 9.25 shows the average length of the longest rivers in the corresponding drainage areas. Again, sequential ECNs provide a cleaner power-law relationship than unstructured ECNs. The exponent $h = 0.56$ fits perfectly into the range observed in nature.

Figure 9.26 shows the length scaling of rivers, i. e., the relation between the upstream length L of the convoluted rivers and the distance d between the considered point and the river's source. Except for the Leopold/Langbein model introduced in Sect. 9.4, all models discussed earlier in this chapter obey a non-fractal scaling relation ($\phi_L = 1$), while observations in nature suggest

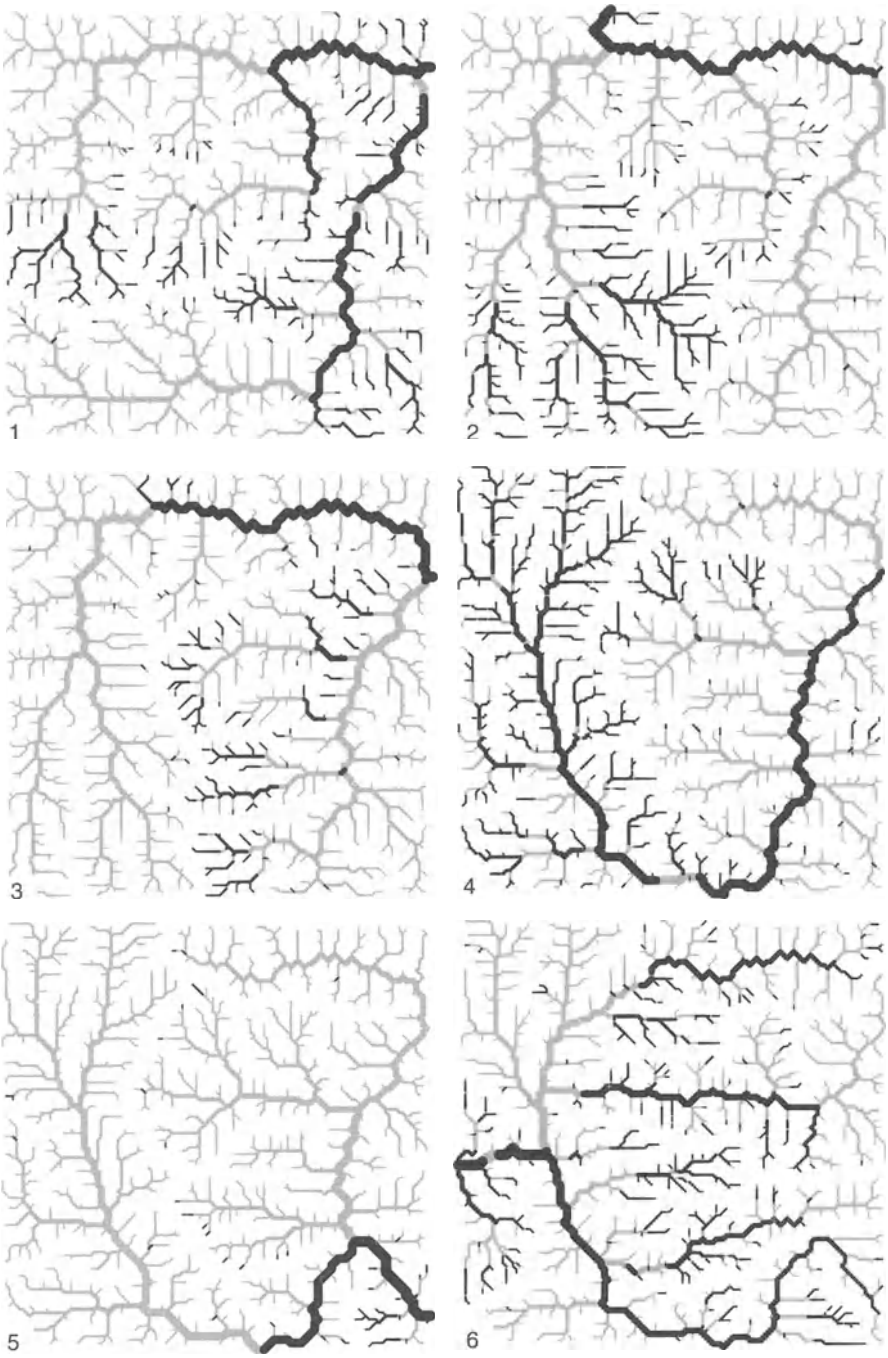

Fig. 9.21. Sequential ECNs as a result of randomly moved outlet locations. The next six networks are shown in Fig. 9.22.

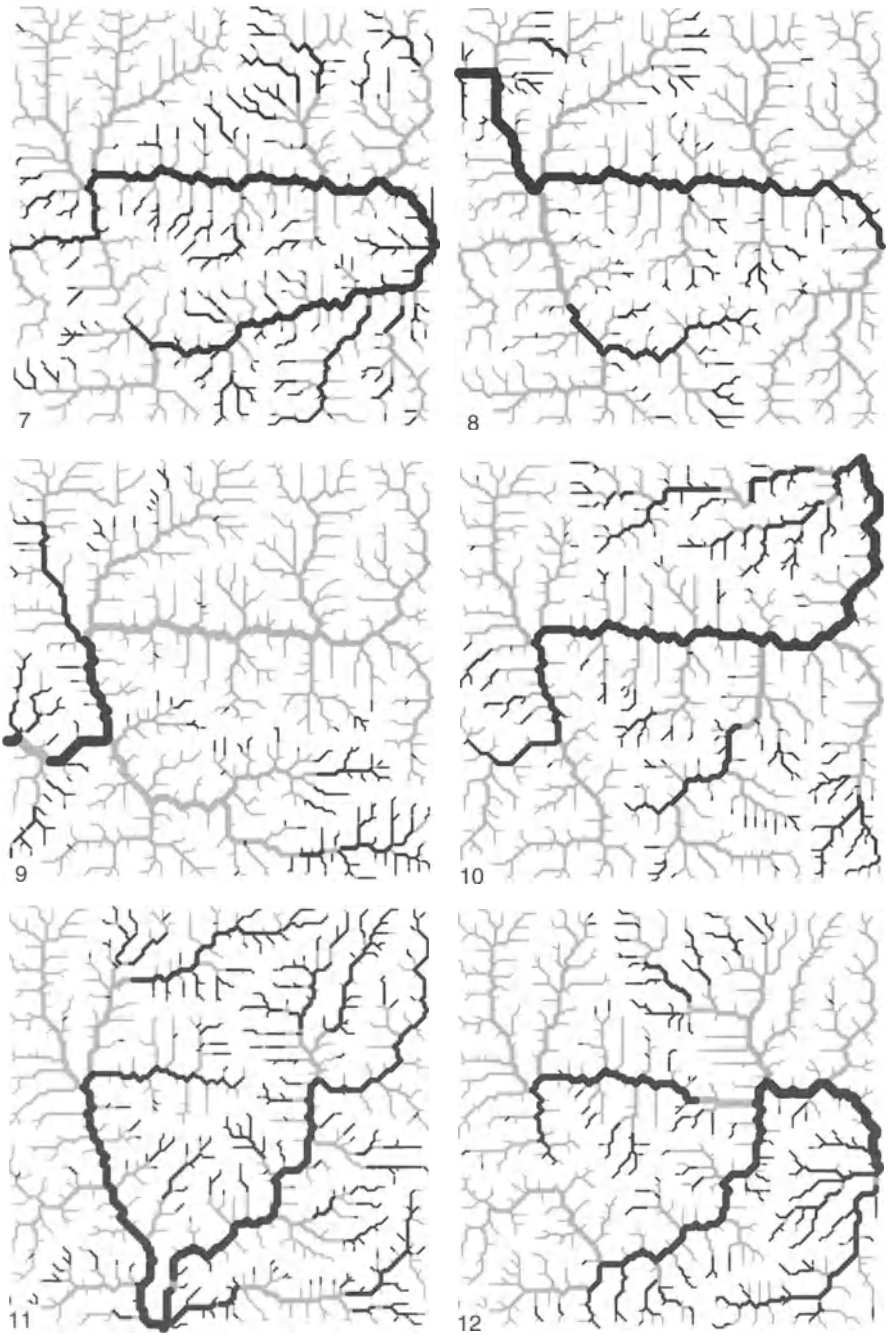

Fig. 9.22. Sequential ECNs as a result of randomly moved outlet locations. The previous six networks are shown in Fig. 9.21.

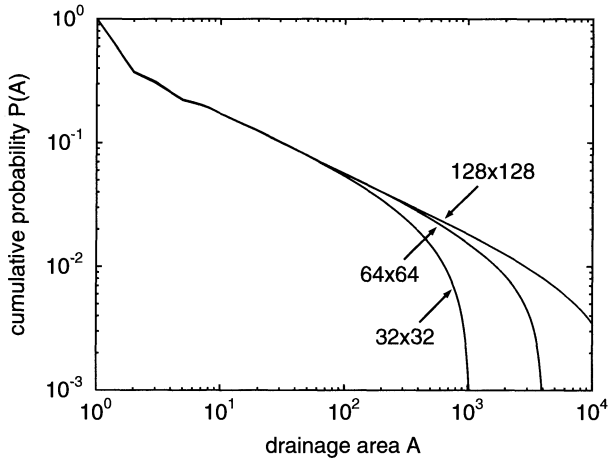

Fig. 9.23. Cumulative size distribution of the drainage areas of sequential ECNs for model sizes from 32×32 to 128×128.

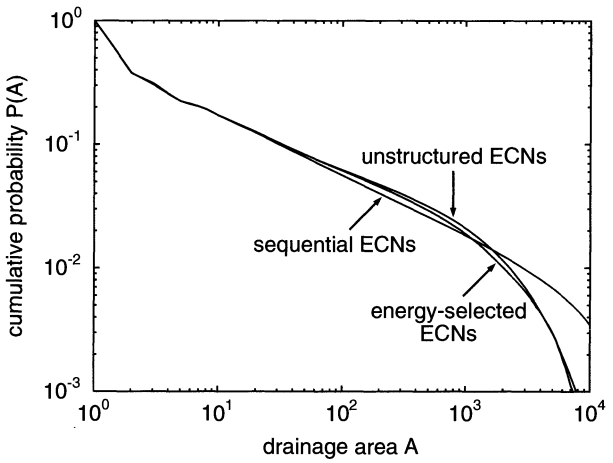

Fig. 9.24. Comparison of the cumulative size distribution of the drainage areas of unstructured ECNs, energy-selected unstructured ECNs (the lowest 10%), and sequential ECNs, computed on a 128×128 grid.

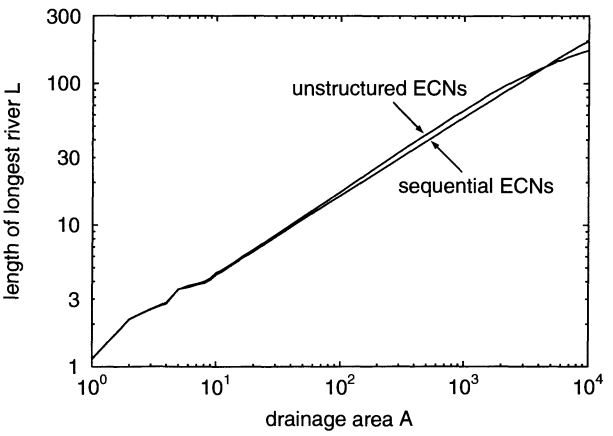

Fig. 9.25. Length of the longest river in a drainage area versus the size of the drainage area (Hack's law).

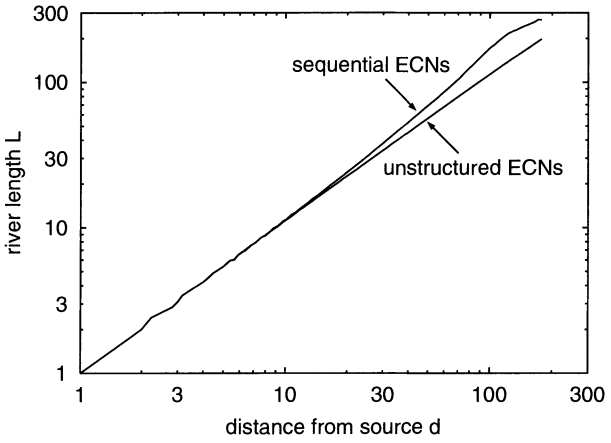

Fig. 9.26. Length scaling of the rivers. Upper diagram: sequential ECNs compared to unstructured ECNs. Lower diagram: illustration of the finite-size effect for sequential ECNs by comparing results of simulations on grids of different sizes.

$\phi_L \in [1.02, 1.12]$ as mentioned in Sect. 9.1. Obviously, large rivers are convoluted more strongly in sequential ECNs than they are in unstructured ECNs; this indicates $\phi_L > 1$. On the other hand, the resulting power law is not very clean; so fitting a power law with $\phi_L = 1.06$ which is in perfect coincidence with natural basins it is not a problem. However, when we discussed the size distribution of the drainage areas in unstructured ECNs and in OCNs, we criticized just this point, and we should not turn the arguments around now.

In summary, the idea of episodically disturbing the model in order to prevent the network from becoming stationary led to a surprising result: In contrast to the expected effect of a disturbance, this kind of disturbance considerably improves the results. Under this aspect, we should think about the idealized reality which seemed to consist of unstructured ECNs or OCNs at first sight. The improvement of the results arising from the permanent reorganization suggests that reorganization plays an important or even the governing part with respect to fractal properties of drainage networks.

But is this SOC? All the fractal properties analyzed yet are static. In contrast, SOC hinges on a fractal distribution of event sizes. And, does the system approach a quasi-steady state with certain statistical properties for nearly all initial conditions? The latter question was addressed in a quite sloppy way when investigating SOC models in the previous chapters. In general, we just performed one simulation with a deterministic or random initial condition and showed that the system evolved towards a critical state then. However, from the model rules it was quite easy to believe that the system evolves towards the same preferred region in phase space for arbitrary initial conditions as long as they are not too regular. In principle, this is exactly what we did when computing a long series of sequential ECNs, starting from an unstructured ECN.

Since the network evolution model differs considerably from the models discussed in the previous chapters, we go a little deeper into detail here.

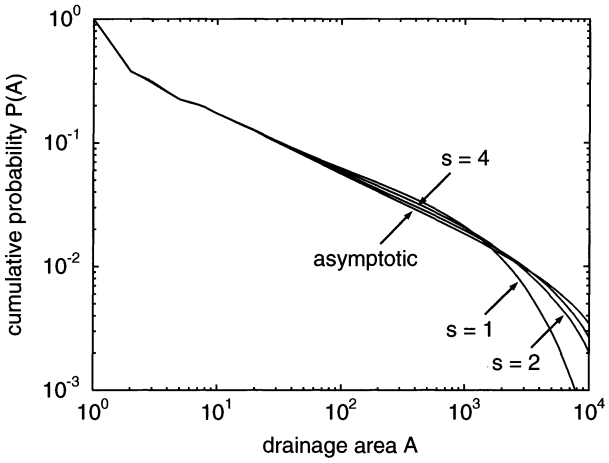

Fig. 9.27. Cumulative size distributions of the drainage areas for different generations s of sequential ECNs.

We first compute a ensemble of 1000 unstructured ECNs on a 128×128 lattice; they differ concerning the location of the outlet and the small initial surface heights. These unstructured ECNs are used as initial conditions for computing 1000 series of sequential ECNs. Finally, the statistical properties of the 1000 networks of the first generation (the unstructured ECNs), of the 1000 networks of the second generation (those obtained from changing the outlet location once), and of the following generations of networks are analyzed. Figure 9.27 shows the size distribution of the drainage areas for different generations s. Since the first generation ($s = 1$) consists of unstructured ECNs, the power-law distribution is quite poor here; it becomes cleaner with increasing generation. The distribution approaches a power law rapidly within a few generations of sequential ECNs; so there is no need for skipping as many ECNs in the beginning as we did when simulating a series of sequential ECNs.

Thus, at least simulations starting from different unstructured ECNs result in sequential ECNs with identical statistical properties. In principle, we should show that this is the case not only for series of sequential ECNs starting at unstructured ECNs, but at arbitrary initial conditions. However, let us be satisfied with one example here; Fig. 9.28 illustrates how the regular structure of the ECN similar to Peano's basin from Fig. 9.11 is destroyed after changing the outlet location about 100 times.

But after all, what are the events obeying a fractal size distribution in our model? In comparison to the models discussed earlier, such as the BTW model, the forest-fire model, the OFC model or the landslide model, the definition of events and their sizes is not straightforward in the network evolution model. As already mentioned, only parts of the drainage network are reorganized from one steady state to another, while other parts persist. According to this observation, interpreting each transition from one ECN to another as an event, and defining its size to be the number of sites which have changed their drainage direction, is straightforward. Figure 9.29 shows the cumulative

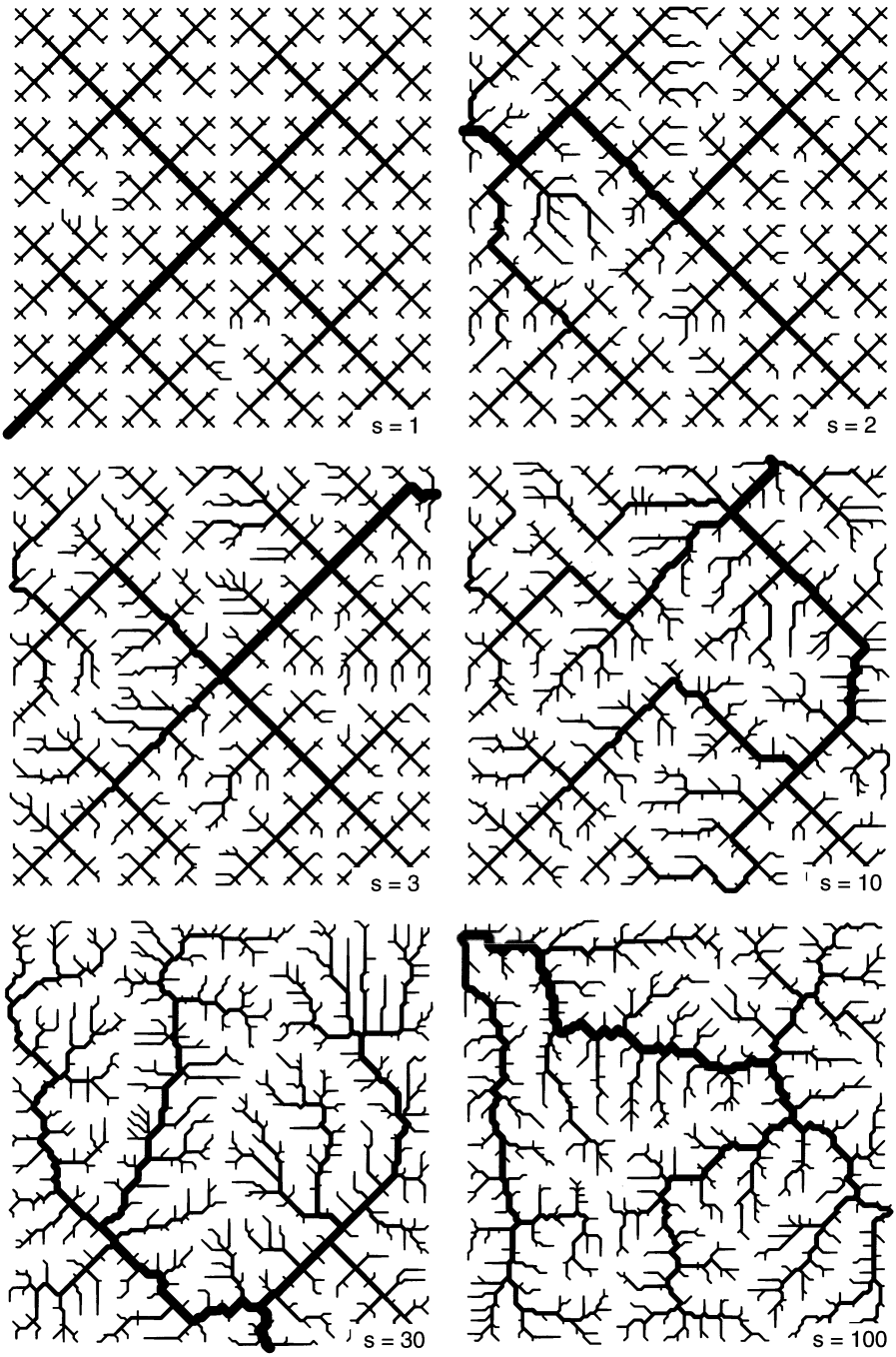

Fig. 9.28. Evolution of a series of sequential ECNs starting from an ECN similar to Peano's basin.

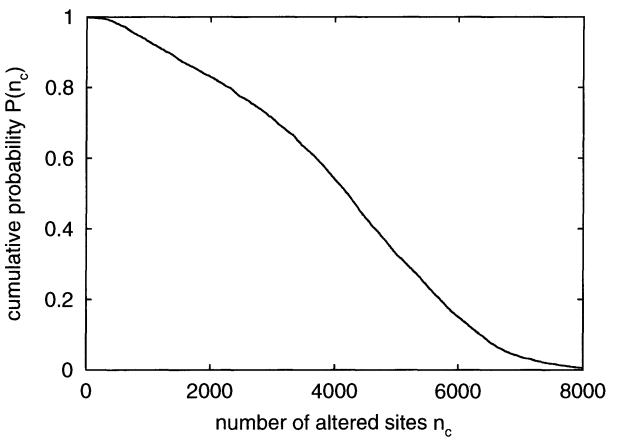

Fig. 9.29. Cumulative distribution of the number of sites which change their drainage direction from one ECN to the next.

distribution of the event sizes in this sense, obtained from the 5000 transitions monitored during the simulation on the 128×128 lattice. Obviously, the distribution is far away from power-law behavior; so these events do not provide any bridge between our model and SOC.

However, our model covers at least two time scales. The long scale is that already discussed which concerns the ECNs. On the other hand, why do we not look at the smaller scale, i. e., analyze what happens during the transition from one ECN to the next? While the surface evolves continuously during these transitions, the evolution of the drainage network, especially of the drainage areas, is discontinuous. Whenever a site changes its flow direction, its whole drainage area may switch from one outlet to the other. Thus, the drainage divide between old and new outlet migrates in discrete steps during the transition from one ECN to another. In this context, we can interpret every change in network topology as an event. We can define the size of an event to be the resulting fluctuation in drainage area, i. e., the size A_c of the area which switches from the old outlet to the new one as a result of changing the drainage direction at a single node. Figure 9.30 illustrates these changes in drainage areas; the three pairs of networks were taken from the transition shown in Fig. 9.20. The drainage direction of one node changed between the pairs; the gray-shaded areas mark the resulting change in drainage areas.

The upper diagram in Fig. 9.31 shows the size distribution of the fluctuations in drainage area, monitored over all 5000 transitions. Apart from finite-size effects, they obey power-law statistics:

$$P(A_c) \sim A_c^{-b}$$

for $A_c > 10$. Similarly to the static size distribution of the drainage areas (Fig. 9.27), the power-law distribution of the fluctuations becomes cleaner from generation to generation. The lower diagram in Fig. 9.31 shows the evolution on a 128×128 grid. The value of s denotes the generation the networks evolve to, so that the evolution starts with $s = 2$.

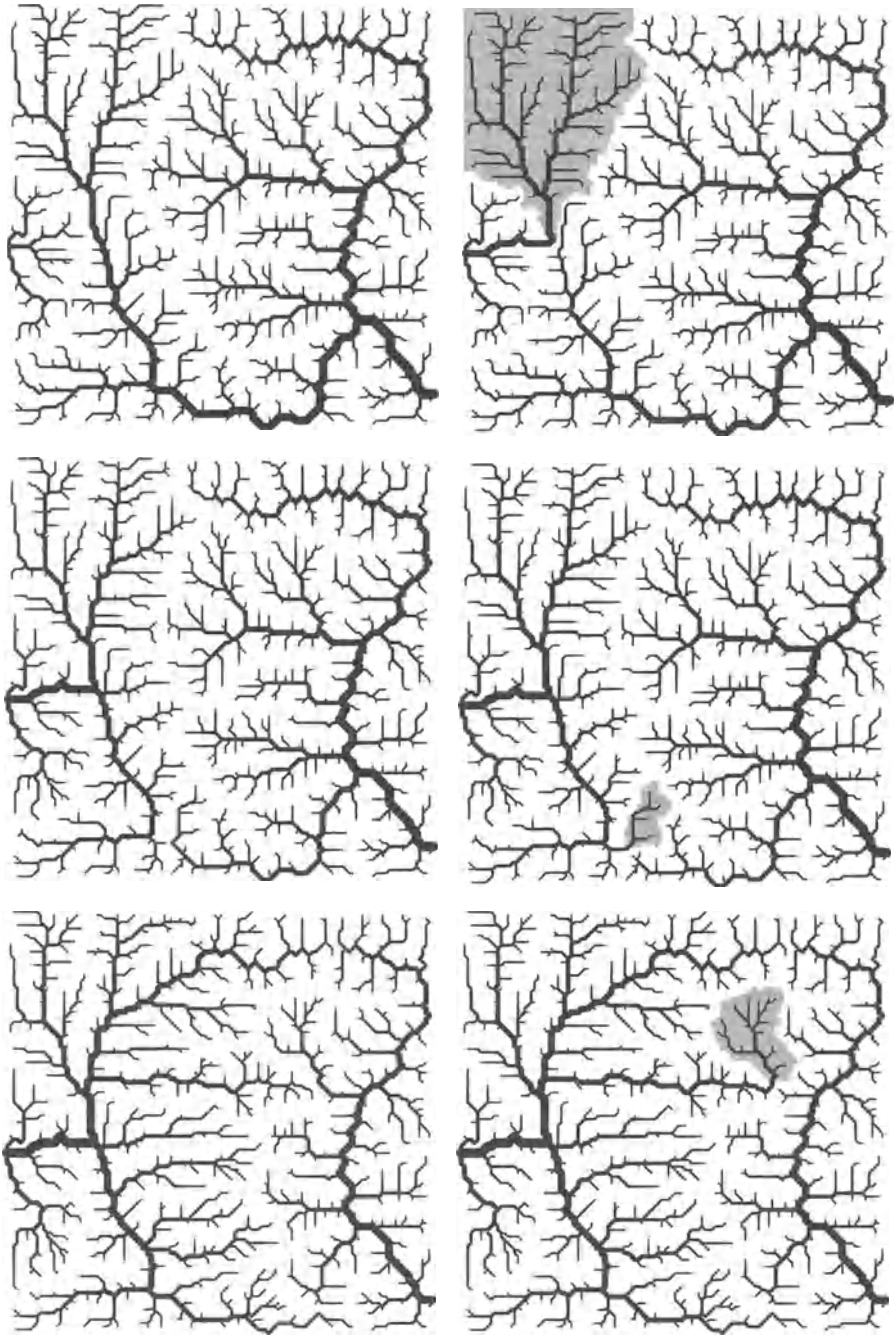

Fig. 9.30. Three examples of fluctuations in drainage areas. The drainage direction of one node changed between the left and right-hand networks; the gray-shaded areas mark the resulting change in drainage areas.

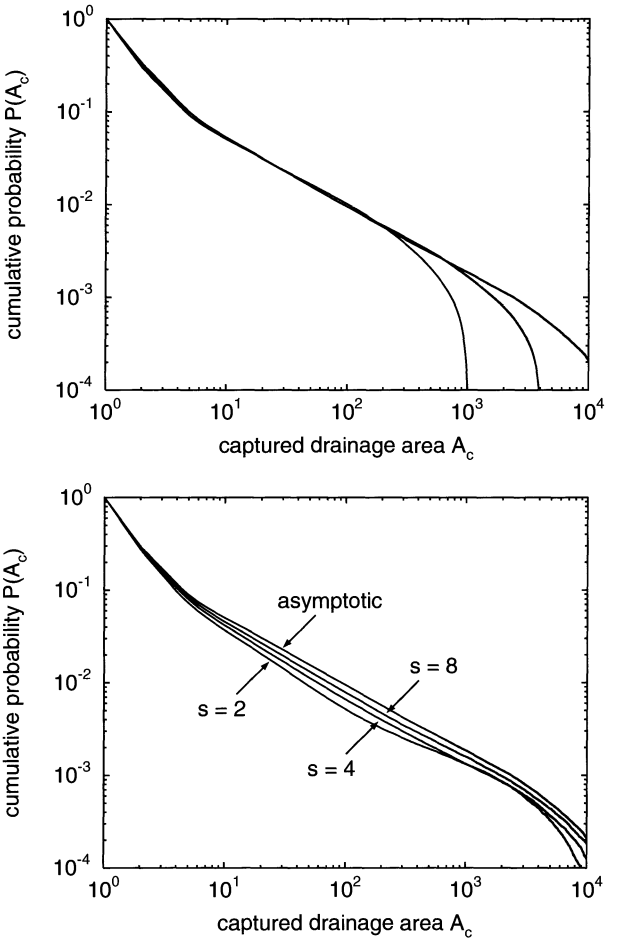

Fig. 9.31. Cumulative size distribution of the fluctuations in drainage areas for lattices of 32×32, 64×64, and 128×128 sites (top) and for different generations of sequential ECNs.

However, is the occurrence of a power-law distribution in the fluctuations in drainage areas surprising? Could it not be just a facet of the fractal size distribution of the drainage areas (Eq. 9.3) in the equilibrated networks? In principle, we just select points from a drainage pattern and make statistics of their sizes, so why should the distribution of the fluctuations differ from the static distribution of the areas? The more or less synchronous evolution of both properties towards power-law distributions supports this hypothesis. On the other hand, both distributions differ quantitatively. While the static exponent is $\beta = 0.46$, that of the fluctuations is $b = 0.70$ and thus considerably larger. However, the static drainage areas A and the fluctuations A_c are essentially different properties. The drainage areas themselves are nested objects; each drainage area of size A contains several smaller drainage areas. In contrast, the fluctuations in the drainage areas occurring during the transition from one ECN to the next are not nested, but fill the whole basin in

sum. Therefore, the difference between β and b is not an argument against a relationship between both; but on the other hand, deriving a relationship consistent with the observed data seems to be a non-trivial problem.

Let us summarize the achievements of this section. Regardless of the origin of the fractal size distribution of the fluctuations, they constitute fractally distributed events as required for SOC. Thus, sequential ECNs evolve towards a quasi-steady state with critical properties; the model exhibits SOC at least on the lowest level discussed in Sect. 5.4. In addition, the static scaling properties of sequential ECNs are cleaner and at least partly more realistic than those emerging from the models discussed in the previous sections, so that the idea of permanently forcing the landform towards non-equilibrium may deepen our understanding of drainage network evolution. However, it is too early to tell whether this concept will become successful because it has just been published (Hergarten and Neugebauer 2001).

Nevertheless, we should be aware that this model differs from the SOC models discussed in the previous chapters under some important aspects. In contrast to these models, defining the events is not straightforward here. We have seen that some quantities exhibit scale invariance, while others do not. In this sense, the SOC concept is still fuzzy because it does not provide any rules how to define the events. So this may be rather an inherent weakness of the SOC concept in its present state than a deficiency of this specific model.

The second difference concerns the relevance of the events to nature, or better to the properties observed in nature. In the models discussed earlier, the events are closely linked to observable properties, such as sizes of forest fires, rupture areas of earthquakes or landslide sizes. In contrast, analyzing drainage networks in nature is restricted to static properties, except for some small-scale laboratory experiments. These experiments mainly address braided rivers where the permanent reorganization is visible even on a short time scale. Sapozhnikov and Foufoula-Georgiou (1997) obtained fractal distributions of fluctuations in the pattern indicating SOC, although the results are not sufficient for recognizing SOC uniquely. However, established models for the evolution of braided rivers (e. g. Murray and Paola 1994) are self-organizing, but do not evolve towards a critical state. Finally, the processes behind braided rivers are far away from the detachment-limited case where we have observed SOC, so the physics behind both phenomena strongly differs. Since we are not able to monitor drainage network evolution in field over long time scales, we may unfortunately not be able to verify of falsify the distribution of the event sizes in field found in our model.

9.8 Optimization by Permanent Reorganization?

In the previous section we introduced sequential ECNs is an alternative theory to established models such as unstructured ECNs or OCNs. When discussing unstructured ECNs in Sect. 9.5, we have already suspected that these

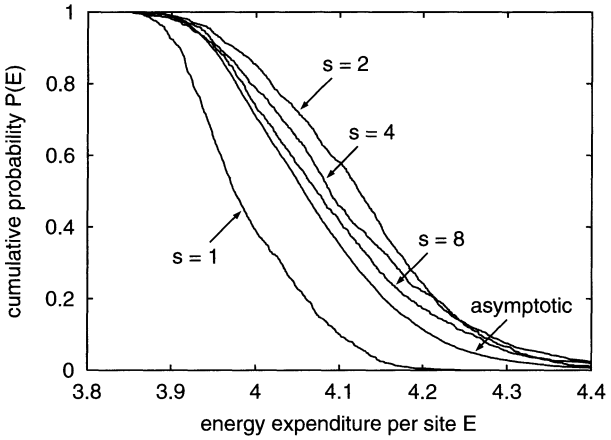

Fig. 9.32. Cumulative distribution of the energy expenditure per node, computed for different generations s of sequential ECNs.

networks may not have enough chances to organize towards a state with reasonable fractal properties; so we blamed the not very clear fractal properties of these networks on this perhaps incomplete organization.

The fractal properties of OCNs (Sect. 9.6) are cleaner than those of unstructured ECNs; but on the other hand, it is not clear how nature should know about energy expenditure and manage to minimize this property. If we now remember that sequential ECNs are just ECNs which are forced to reorganize permanently, it seems that we are close to understanding the whole story. Do sequential ECNs evolve towards a state of minimum energy expenditure? In other words: Are sequential ECNs the physical realization of the unproven optimization principle?

Figure 9.32 shows the cumulative distribution of the energy expenditure per site, computed for different generations of sequential ECNs on a 128×128 grid. Obviously, the behavior is more complex than expected. Compared to the first generation (unstructured ECNs), the energy expenditure of the second generation ($s = 2$) is considerably higher. So it seems first that sequential ECNs evolve away from the state of minimum energy expenditure. In contrast, the energy expenditure of the further generations of sequential ECNs decreases, so that they evolve towards the state of minimum energy expenditure then. However, this decrease in energy expenditure ceases soon; in the limit $s \to \infty$, the energy expenditure approaches a distribution which is roughly in the middle between the first and the second generation of sequential ECNs.

In summary, sequential ECNs do not just provide a physical realization of the principle of minimum energy expenditure. The relation between sequential ECNs and OCNs is still unclear; at the moment there seems to be no relationship between SOC in drainage network evolution and energy expenditure.

10. SOC and Nothing Else?

In the last 15 years, SOC turned out to be a powerful concept for elucidating the occurrence of fractal distributions in dynamic systems. Therefore, we may be tempted to look for SOC whenever fractals or fractal distributions occur. However, we should be aware that SOC is just one mechanism for explaining scale invariance. It is perhaps the most interesting one because it is still poorly understood, and perhaps even that with the widest field of applications. Nevertheless, there are numerous other mechanisms for generating fractals; in Sect. 1.6 we have already discussed fragmentation being one of the most prominent examples; several other concepts are discussed in the book of Sornette (2000).

In this chapter, we focus on a few interesting developments in this direction. We have already discussed scale invariance as a result of optimization in the context of drainage networks (Chap. 9). The interest in optimization principles has been renewed by introducing an idea that hides behind the acronym HOT: highly optimized tolerance (Carlson and Doyle 1999, 2000). The basics of this concept are reviewed in Sect. 10.3.

But first we discuss what happens if additional scale invariance is introduced into SOC systems. At first sight, this seems to be a strange idea. Scale invariance emerges from SOC systems without any need for tuning, so why should we put in additional scale invariance artificially? However, in the previous chapters we have met several examples where SOC might be just a part of the story. The forest-fire model discussed in Chap. 6 explains the fractal size distribution of forest fires observed in nature on a qualitative level, but the predicted exponents are too low. So the physics of this model may be oversimplified, but there might also be another impact which is not related to the propagation of fires in any way. Fractal properties of the earth's surface or vegetation are candidates for such an effect that might be responsible for the observed distributions in combination with SOC in the process itself.

Earthquakes are the most prominent example where the observed statistics appear to be a result of SOC in combination with further fractal properties. In Chap. 7 we have discussed a simplified model of stick-slip motion at an individual fault. This model is able to explain the Gutenberg-Richter law and even some further seismic properties within the framework of SOC. On the other hand, we have also reported that the size distribution of faults

in the earth's crust is fractal, and that the Gutenberg-Richter law might in principle emerge from this property, too. So where does the Gutenberg-Richter law finally come from? At least earth scientists should be familiar with the idea that only a few phenomena can be explained uniquely by just one process. Each fault may generate its own Gutenberg-Richter law, but the observed statistics may arise from a combination of the Gutenberg-Richter law of individual faults with a fractal fault size distribution. A first idea in the direction was proposed by Sornette et al. (1991) a few years after the concept of SOC had been introduced in seismology.

Different concepts of combining SOC with pre-defined scale invariance are reviewed in the following two sections. The first idea starts from an ensemble of independent SOC systems of different sizes. The finite system sizes introduce cutoff effects, and we will see that the scaling properties of the ensemble depend on both the properties of the individuals and on their size distribution. Transferred to the example of earthquakes, this idea corresponds to a set of non-interacting faults. Although this assumption may not be very realistic in this example, it is quite illustrative in a general context.

The second way of introducing additional scale invariance into SOC systems assumes a pre-defined structure behind a SOC system. In Sect. 10.2 we illustrate the effect of such a structure in a model which is similar to the forest-fire model.

10.1 Ensembles of SOC systems

Let us start from a hypothetic SOC system with an idealized cutoff behavior at large event sizes as considered in Sect. 2.2. We assumed there that the object sizes s follow a power-law distribution between a minimum size $s_{\min}$ and a maximum size $s_{\max}$, and that there are no objects outside this range.

In the following, we consider an *ensemble* of independent systems of this type which only differ in size. Let us further assume that the minimum event size is a property of the physical process, so that $s_{\min}$ is the same within the whole ensemble. Then we can choose the units of event sizes in such a way that $s_{\min} = 1$. In the example of earthquakes, $s_{\min}$ may correspond to the smallest rupture that can occur due to the microscopic structure of the rock. In contrast, the maximum event size shall arise from the finite system size; in the example of earthquakes, $s_{\max}$ may correspond to an event where rupture affects the entire fault. According to Eq. 2.1, the cumulative size distribution of the events occurring in a member of the ensemble is

$$P_{s_{\max}}(s) = \begin{cases} 1 & s \leq 1 \\ \frac{s^{-b} - s_{\max}^{-b}}{1 - s_{\max}^{-b}} & \text{if} \quad 1 < s < s_{\max} \\ 0 & s_{\max} \leq s \end{cases} . \tag{10.1}$$

The index $s_{\max}$ in $P_{s_{\max}}(s)$ shall remind us on the fact that the maximum event size $s_{\max}$ is a property of the individual system and varies within the ensemble. Let $Q(s_{\max})$ be the cumulative distribution of the upper cutoff values, i. e., the probability that a system arbitrarily picked out of the ensemble has an upper cutoff value of at least $s_{\max}$.

At this point we may think that the size distribution of the events occurring in the whole ensemble can be derived from combining the size distributions $P_{s_{\max}}(s)$ of the individual systems with the distribution of the cutoff values $Q(s_{\max})$, but this is not possible in general. The problem arises from the fact that the activity of the individual systems is not necessarily the same. Therefore, the activity $A(s_{\max})$, i. e., the total number of events per unit time occurring in a system with an upper cutoff value $s_{\max}$, must be regarded, too. Then the number of events per unit time with a size of s or larger is

$$\dot{N}_{s_{\max}}(s) \;=\; A(s_{\max})\, P_{s_{\max}}(s).$$

From this we can compute the number of events per unit time with a size of at least s for the whole ensemble:

$$\dot{N}(s) \;=\; \int_0^\infty q(s_{\max})\; \dot{N}_{s_{\max}}(s)\, d_{s_{\max}} \;=\; \int_0^\infty q(s_{\max})\, A(s_{\max})\, P_{s_{\max}}(s)\, d_{s_{\max}},$$

where $q(s_{\max}) = -\frac{\partial}{\partial s_{\max}} Q(s_{\max})$ is the probability density of $Q(s_{\max})$. The cumulative size distribution of all events occurring in the ensemble is

$$P(s) \;=\; \frac{\dot{N}(s)}{\dot{N}(0)} \;=\; \frac{\int_0^\infty q(s_{\max})\, A(s_{\max})\, P_{s_{\max}}(s)\, d_{s_{\max}}}{\int_0^\infty q(s_{\max})\, A(s_{\max})\, d_{s_{\max}}}$$

because $P_{s_{\max}}(0) = 1$. Let us in the following assume that the activity of all members of the ensemble is the same, although, in the example of earthquakes, constraints such as a constant rate of deformation at all faults may lead to a different condition. Since

$$\int_0^\infty q(s_{\max})\, ds_{\max} \;=\; Q(0) - Q(\infty) \;=\; 1,$$

this simplification leads to

$$P(s) \;=\; \int_0^\infty q(s_{\max})\, P_{s_{\max}}(s)\, ds_{\max}. \tag{10.2}$$

As mentioned in the introduction to this chapter, we assume that the sizes of the ensemble's members obey a scale-invariant distribution in order to superpose the fingerprint of an SOC system by a an additional scale-invariant structure. If we assume a linear relationship between the system size and the maximum event size $s_{\max}$, we obtain a scale-invariant distribution for $s_{\max}$, too. The exponent β of this distribution coincides with that of the size

distribution of the ensemble's members. According to Eq. 2.5, the probability density of the corresponding Pareto distribution is

$$q(s_{\max}) = \begin{cases} \beta\, s_{\max}^{-\beta-1} & \text{if} \quad s_{\max} > 1 \\ 0 & \text{else} \end{cases} .$$

Here we have assumed that the smallest members of the ensemble are so small that they only allow the smallest physically possible events ($s_{\min} = 1$). Inserting $P_{s_{\max}}(s)$ from Eq. 10.1 and $q(s_{\max})$ into Eq. 10.2 leads to

$$P(s) = \int_s^\infty \beta\, s_{\max}^{-\beta-1} \frac{s^{-b} - s_{\max}^{-b}}{1 - s_{\max}^{-b}}\, ds_{\max} \tag{10.3}$$

for $s \geq 1$. This integral cannot be solved analytically, but its behavior in the limit of large event sizes s can be examined. Since $s_{\max} > s$ in the integrand, the denominator $1 - s_{\max}^{-b}$ converges towards unity then, so that

$$P(s) \to \int_s^\infty \beta\, s_{\max}^{-\beta-1} \left(s^{-b} - s_{\max}^{-b}\right) ds_{\max} = \frac{b}{b+\beta}\, s^{-(b+\beta)} \tag{10.4}$$

for $s \to \infty$. Thus, the resulting distribution is not fractal in the strict sense, but behaves like a power law in the limit of large event sizes. The exponent of the power-law tail is the sum of the exponent β which characterizes the ensemble and the exponent b which characterizes the SOC behavior of the individual members of the ensemble. Therefore, the combination of SOC and fractally distributed finite-size behavior may shed new light on a phenomenon we have already observed when discussing forest fires (Chap. 6) and landslides (Chap. 8): Event statistics observed in nature exhibit scale invariance, so that the system's behavior looks like SOC, but at least the established SOC models fail on a quantitative level. These models predict size distributions whose exponents are significantly too low, which means that the frequency of large events is overestimated. In the example of landslides, we found a solution by introducing time-dependent weakening, but in principle the observed landslide size distribution might be a result of the original sandpile model in combination with a fractal distribution of slope sizes in a region. In analogy, a fractal size distribution of forests may make the results of the simple forest-fire model more realistic.

As already mentioned, this idea is able to unify two apparently contrary theories on the origin of the Gutenberg-Richter law in seismology. The Gutenberg-Richter law may arise from a combination of SOC in stick-slip motion at individual faults and a fractal distribution of the fault sizes. Then, the exponent of the observed size distribution of the rupture areas which is close to unity is the sum of the exponent b of the stick-slip motion and the exponent β of the fault-size distribution. Obviously, b may be smaller than unity then, and the observed dispersion in the Gutenberg-Richter exponent may be explained by a regional variation in the fault-size distribution.

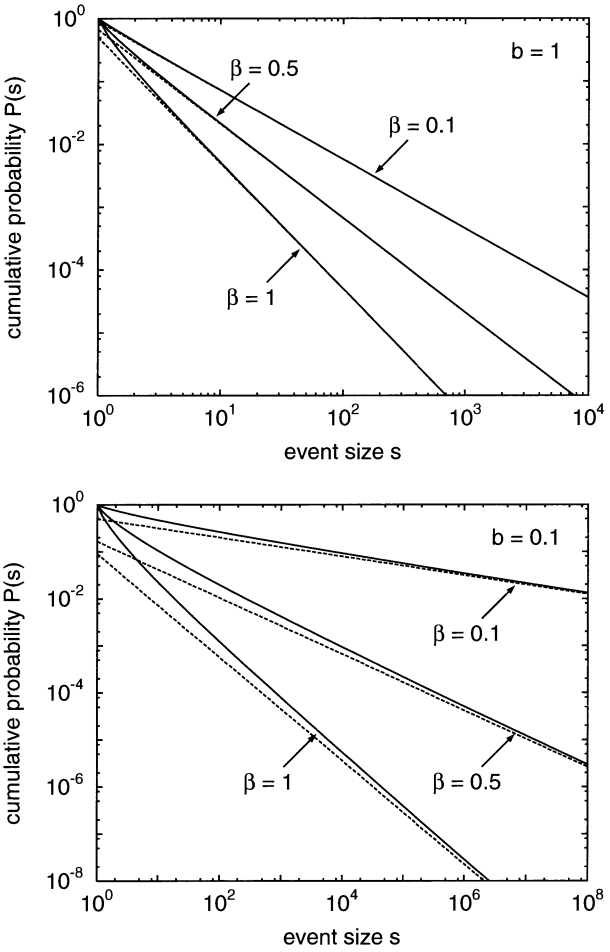

Fig. 10.1. Cumulative size distributions of the events occurring in ensembles of independent SOC systems with fractally distributed upper cutoff values. The dashed lines represent the asymptotic power laws.

However, we should keep in mind that an ensemble of independent SOC systems with a fractal distribution of the cutoff values does not generate Pareto-distributed event statistics in the strict sense, but only in the limit of large event sizes. Figure 10.1 shows some examples, obtained numerically by developing Eq. 10.3 in a Taylor series concerning the variable s^{-b}. The dashed lines represent the asymptotic power laws according to Eq. 10.4. The convex curvature of the distributions in the bilogarithmic plots indicates a relative excess of small events. This result holds for all combinations of the exponents b and β because the factor $\frac{b}{b+\beta}$ in front of the asymptotic power law (Eq. 10.4) is always smaller than unity. The relative excess of small events is strong if β is much larger than b and vice versa. Comparing the distributions shown in Fig. 10.1 confirms this result. The distributions for $b = 1$ converge rapidly towards power laws, even the distribution obtained for $\beta = b = 1$ can

hardly be distinguished from a power law for $s \geq 10$. In contrast, convergence is slow for $b = 0.1$, especially if β is significantly larger than b.

But what does this result mean with respect to the examples considered in the previous chapters? A Gutenberg-Richter exponent close to unity may be consistently explained by a combination such as $\beta \approx b \approx \frac{1}{2}$. However, when considering the conservative limiting case of the OFC model for the individual faults in Sect. 8.2 (Fig. 8.3), we obtained $b = 0.23$. Combining this result with $\beta = 0.77$ should lead to a considerable excess of small earthquakes. However, we should be careful when concluding that the conservative OFC model in combination with a fractal fault-size distribution is not consistent with the Gutenberg-Richter law. The smallest physically possible earthquakes may be far outside the range of seismic monitoring, so that we will probably not recognize the resulting deviation from the Gutenberg-Richter law. But on the other hand, the complexity found in Sect. 7.10 provides clear arguments in favor of the non-conservation in the OFC model.

The forest-fire model discussed in Chap. 6 yields exponents $b \approx 0.23$, while observed exponents are in the range between 0.3 and 0.5. In principle, this range can be explained by applying the forest-fire model to an ensemble of distinct forests obeying a fractal size distribution with an exponent β in the same order of magnitude as b. However, there are two arguments that raise doubts against this explanation: First, there are no reports on a fractal distribution of forest sizes in nature so far, in contrast to the example of earthquakes where fractal fault-size distributions have been observed. But still more important, the effect of the finite grid size in the forest-fire model strongly differs from the simple cutoff behavior assumed in this section. As found in Sect. 6.1, the frequency of large fires even increases if the lattice is too small. Thus, the theory developed here cannot be applied directly to the forest-fire model so that the question for the origin of the exponents in observed distributions of forest fires is still open.

The lack of field evidence remains a problem when applying this theory to landslides. As discussed in Sect. 8.2, the sandpile model is qualitatively reasonable, but the predicted exponent $b < 0.3$ is much lower than those observed in nature. Explaining the range between $b \approx 0.7$ and $b \approx 1.6$ found in nature would require a fractal distribution of the slope sizes with quite large exponents β. However, neither a fractal slope-size distribution nor such a strong variation in the fractal properties of the land surface have been observed so far. Thus, the idea of superposing SOC by additional scale invariance is not very promising with respect to landslide dynamics; the other approaches discussed in Chap. 8 appear to be more reasonable.

10.2 SOC in Pre-Structured Systems

In this section we review a model recently introduced by Tebbens et al. (2001). In its original form, this model accounts for the statistical distribution of

hotspot seamount volumes, but let us refrain from discussing the physical background in detail. On a more abstract level, the basic question is the same as in the previous section. A fractal event-size distribution observed in nature indicates SOC, but the exponents predicted by some established SOC models such as sandpile and forest-fire model are significantly too low.

The model is defined on a two-dimensional, quadratic lattice. In contrast to the models discussed before, it contains two different types of sites. Some sites behave exactly as the sites in the forest-fire model; trees can grow and be burnt down. If any site in a cluster of trees catches fire, the whole cluster is burned down immediately. In contrast, sites of the second type cannot be occupied by trees, but are able to ignite adjacent trees. Let us call these sites *trigger cells*. In the beginning, the lattice is once divided up into trigger cells and regular cells; trigger cells will always remain trigger cells and vice versa. In each step, a site is randomly chosen. If this site is a regular, but empty cell, a tree grows there. If it is already occupied by a tree, nothing happens. If the selected site is a trigger cell, those of its nearest neighbors which are occupied by trees are ignited, and the corresponding cluster is burnt down. Therefore, the trigger cells may be interpreted as fireplaces in a forest; a fire does not cause any damage there, but may ignite the surrounding forest.

The original model introduced by Tebbens et al. (2001) gradually differs from that introduced above. The first difference only concerns the nomenclature; trigger cells are called critical cells in the original paper. However, this term does not refer to criticality in the sense of scale-invariant fluctuations, so it may be a little misleading in this context. For this reason, we prefer the term trigger cells which comes closer to their physical meaning. The second difference concerns the number of trees per site. In terms of the forest-fire model, the original model allows more than one tree per site, and the size of an event is characterized by the number of destroyed trees. In contrast, we prefer the number of burnt sites in analogy to the forest-fire model.

The statistical properties of this model strongly depend on the spatial distribution of the trigger cells. Let us begin with a completely random pattern. We consider two examples where 4096 respectively 531,441 trigger cells are randomly distributed on a lattice of 4096×4096 sites. The reason for choosing these apparently strange numbers will become clear later. Figure 10.2 shows the non-cumulative event statistics obtained from analyzing 10^9 fires in each simulation. In analogy to the SOC models considered earlier, a sufficiently large number of events (10^8) was skipped in the beginning in order to avoid artificial effects of the initial state.

The results obtained from this model can be directly compared with those of the original forest-fire model. In the simulation with 531,441 trigger cells, about 30.6 trees grow between two sparking events in the mean. Thus, this simulation should be compared to the case $r = 32$ in the simulations performed in Chap. 6. For the same reason, the simulation with 4096 trigger cells should be compared to the case $r = 4096$ in the original model. In both

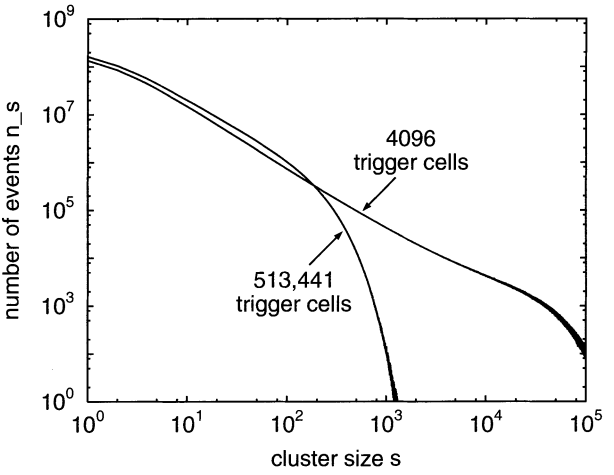

Fig. 10.2. Non-cumulative event statistics in the modified forest-fire model with randomly distributed trigger cells.

cases, the event-size statistics obtained from the modified model are similar to those of the original forest-fire model. Therefore, the modification achieved by introducing randomly distributed trigger cells does not provide significant new insights. So far, this model is just an example for universality in the forest-fire model.

However, things turn around if we follow the original paper of Tebbens et al. (2001) and introduce a fractal pattern of trigger cells. They generated a simple, fractal pattern according to the following algorithm: The lattice is subdivided into $n \times n$ identical tiles. Then, k tiles are randomly selected. The procedure is recursively applied to the selected tiles until it arrives at individual sites. The selected tiles on the finest level become trigger cells, while all other sites are regular cells. This recursive procedure is essentially the same as the fragmentation algorithm introduced in Sect. 1.6. The only difference towards the latter is that we consider those fragments which are smaller than a given size instead of the fractures. Figure 10.3 shows the upper right pattern from Fig. 1.1, transformed into this representation. Those fragments which are not larger than $\frac{1}{512}$ of the domain's size are plotted, so this pattern defines a spatial distribution of trigger cells according to the rules introduced above on a 512×512 lattice with $n = 2$ and $k = 3$.

It can easily be seen that these patterns of trigger cells show discrete scale invariance in the range between the size of one site and the size of the whole lattice. The fractal dimension coincides with the exponent of the size distribution of the fragments in the corresponding fragment pattern (Eq. 1.4):

$$D = 2\frac{\log k}{\log n^2} = \frac{\log k}{\log n}.$$

Obviously, the same fractal dimension D can be achieved by choosing different combinations of n and k. So we are in principle free to generate different patterns of trigger cells by selecting a fixed value of n and variable values of

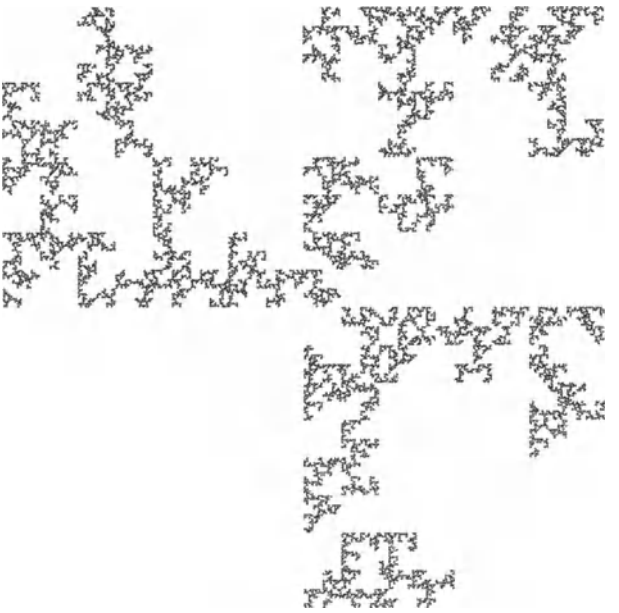

Fig. 10.3. Fractal pattern of trigger cells on a 512×512 lattice, obtained with $n = 2$ and $k = 3$.

k. The value of n defines the steps of discrete scale invariance. Therefore, small values n are preferable in order to come as close as possible to continuous scale invariance. However, this choice allows only a few fractal dimensions as k is restricted to integer numbers in the range between 2 and $n^2 - 1$. The cases $k = 1$ and $k = n^2$ are trivial because all cells are trigger cells for $k = n^2$, while there is just one trigger cell for $k = 1$. Thus, choosing $n = 2$ allows only fractal patterns of trigger cells with $D = 1$ ($k = 2$) and $D = 1.585$ ($k = 3$).

In the original paper, $n = 5$ was chosen. This allows 23 different fractal dimensions in the range between 0.43 and 1.975. However, if we consider a lattice of 125×125 sites as done in the original paper, the pattern covers only three different scales. Let us, in the following, use $n = 2$ and accept the restriction to $k = 2$ and $k = 3$. On a 4096×4096 lattice, the resulting pattern consists of 4096, respectively, 513, 441 trigger cells; this explains the strange number of trigger cells in the model with randomly distributed trigger cells.

Figure 10.4 shows the statistics obtained from simulating this model. Obviously, the fractal pattern of trigger cells strongly affects the properties of the model. In both cases, the number of events follows a power-law distribution, so this model approaches a critical state at least for the two parameter values considered here. In the original paper, further patterns of trigger cells with different fractal dimensions were used. The simulations resulted in power-law distributions, too, although grid sizes and statistics were much smaller than in the simulations discussed here. So there is evidence that the model exhibits SOC at least for fractal patterns with dimensions in a certain range.

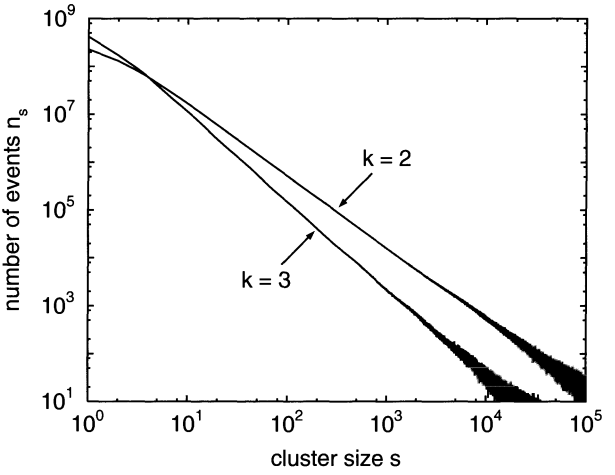

Fig. 10.4. Non-cumulative event statistics in the modified forest-fire model based on fractal patterns of trigger cells with $D = 1$ and $D = 1.585$. The statistics include 10^9 events.

This behavior distinguishes the modified model from the original forest-fire model; the latter becomes critical only in the limit $r \to \infty$ and shows more or less significant deviations from power-law behavior for any finite values of the parameter r. In return, universality (Sect. 6.2) has been lost by introducing a fractal pattern of trigger cells; the exponent of the power-law distribution depends on the fractal dimension of the pattern. Transferred to a cumulative distribution, we obtain $b = 0.48$ for $D = 1$ and $b = 0.77$ for $D = 1.585$. This result suggests that the model is able to generate power-law distributions with a certain range of exponents.

Tebbens et al. (2001) introduced a new term including a new acronym for the observed behavior: *self-similar criticality* (SSC). However, the reader may wonder whether this is necessary because we may run out of three-character acronyms some time. But still more important, this new term may be misleading. The term criticality addresses the occurrence of scale-invariant fluctuations, so what shall the term self-similar express in this context? This term would make sense if it was introduced for distinguishing isotropic (self-similar) from anisotropic (self-affine) scale invariance, but this is clearly not the point addressed by this model.

So what distinguishes the behavior of this model from SOC? Obviously, it self-organizes towards a state with critical properties. Depending on the pattern of the trigger cells, the model yields fractal distributions with significantly higher exponents than the original forest-fire model and the BTW model do. Since these models are still the most widespread examples of SOC, one may be tempted to associate SOC with exponents b close to zero. However, the definition of SOC discussed in Chap. 5 does not pose any restrictions with respect to the exponent b, and we have already met SOC systems with exponents close to unity in Chap. 7 and in Chap. 8.

In summary, there is no reason why this model should not be SOC. Compared to the original forest-fire model, introducing additional scale invariance has destroyed universality, but it is well-known that the exponents are universal in some SOC models, while they depend on the model parameters in other models. So there is no doubt that the model considered in this section can deepen our understanding of certain phenomena, but there is no need for a new nomenclature.

So far we have learned about two different ways of forcing changes in the scaling properties of SOC systems. In the previous section, ensembles of independent SOC systems were considered; additional scale invariance was introduced by assuming a fractal distribution of the system sizes. In contrast, a fractal pattern was merged into a single SOC system here. At this point we may wonder whether the difference between both approaches is as large as it seems first. Obviously, the results are similar: Compared to the reference models, the exponents of the size distributions increase; the larger the fractal dimension of the additional pattern or distribution is, the larger the increase in the exponent becomes. Maybe the fractal pattern of trigger cells just splits up large clusters of trees, so that it roughly subdivides the model area into more or less independent parts. However, let us not go deeper into details here, but turn to a completely different mechanism for generating power-law distributions in the following section.

10.3 Highly Optimized Tolerance

In Sect. 9.6 we have discussed the role of optimization in drainage network evolution. Recently, a new formalism for deriving power-law distributions from optimization principles was introduced by Carlson and Doyle (1999, 2000), called *highly optimized tolerance* (HOT). Let us follow the original paper and explain the basic ideas with the help of the forest-fire model. In Chap. 6 we have seen that this model shows SOC in the limit of a low sparking frequency, provided that the model area is sufficiently large.

Economic interests are often opposite to the behavior of undisturbed, natural systems. We may therefore expect that a SOC forest is not necessarily the best system if we are interested in exploiting the forest in order to obtain wood. In this case, a high density of trees ρ is desirable. As shown in Sect. 6.1, the SOC forest-fire model generates a relatively high number of large fires; these fires prevent the average tree density from becoming much larger than 0.4. According to Eq. 6.1, the mean loss, i. e., the mean number of burned trees per sparking event, is $L = r(1-\rho)$, where r is the number of growing trees between two sparking events. Thus, density and loss are closely related; we should reduce the mean loss L in order to maintain a high tree density.

Introducing fire breaks is a straightforward strategy for reducing the loss. Let us assume that a large forest is subdivided into equal tiles of area A, and

that these tiles are separated by fire breaks that completely inhibit the propagation of fires. If we measure areas in numbers of model sites, we immediately obtain $L \leq A$. Thus, fire breaks reduce the average damage if the forest is subdivided into sufficiently small areas. Fig. 6.5 (p. 115) illustrates this effect. Let us start from a forest of $32,768 \times 32,768$ sites and a given growth rate r. This forest can be split up into 256×256 forests of size 128×128. Since the growth rate r is the ratio of two probabilities per unit time (p and f), each of the small forests is described by the same rules as the original forest with the same parameter r. Thus, the increase in density resulting from splitting up the forest can be directly determined by comparing the two curves in Fig. 6.5; the gain becomes drastic for large growth rates, i. e., if the original forest is close to the SOC state.

Under this aspect, building more and more fire breaks is the best solution. However, we cannot expect to get fire breaks for free. First, fire breaks occupy some area, so that the tree density decreases if there are too many fire breaks. In the example discussed above, this effect was disregarded. If the width of the fire breaks is one site, we obtain only 254×254 forests of size 128×128, but this reduces the tree density by less than two percent. More important, fire breaks must be maintained; this requires additional effort and thus deteriorates the economic balance. Therefore, the aim cannot be just to reduce the mean loss as much as possible, but to reduce it under the *side condition* of a given amount of fire breaks. Let us measure the effort in terms of the total length of all fire breaks, i. e., assume that the sum of the lengths of all fire breaks in the forest is fixed. Finding the best spatial distribution of the fire breaks leads to the question addressed in the concept of HOT: Minimizing the mean loss induced by an event or maximizing the average output of a system under the condition that the effort to be spent is limited.

Let us now apply this idea to the forest-fire model. If the forest is subdivided into many small pieces, the model switches from SOC to a much simpler behavior where the individual patches are strongly affected by finite-size effects. Between two fires, the tree density approaches a high value, and the next fire will probably destroy nearly all trees in the patch. Therefore, the size distribution of the fires in an individual patch is no longer a power law, but rather governed by fires whose sizes are roughly the size of the patch. Under this aspect, we can switch to a simplified forest-fire model. Instead of a discrete lattice, we consider a continuous model area. This area is subdivided by fire breaks consisting of lines. Sparks are uniformly distributed over the model area, but now each spark causes a fire that destroys the corresponding patch bounded by fire breaks. The damage caused by the fire is characterized by the area of the burnt patch.

But what is the best distribution of fire breaks in this model? Let us start with a rectangular lattice of fire breaks that subdivides the model area into rectangles of size $l_1 \times l_1$. The effort per area, i. e., the total length of the fire breaks per unit area is then

$$E = \frac{l_1 + l_2}{l_1 l_2}. \tag{10.5}$$

The loss L caused by a fire is given by the area of the rectangles, no matter where the fire takes place. Therefore, the function $L = l_1 l_2$ must be minimized under the side condition defined by Eq. 10.5 for a given effort E. The minimization can be performed using the method of Lagrangian multipliers, but here it is easier to write l_1 and l_2 in the form

$$l_1 := \frac{1}{E \cos^2\alpha} \quad \text{and} \quad l_2 := \frac{1}{E \sin^2\alpha} \quad \text{where} \quad 0 < \alpha < \tfrac{\pi}{2}$$

which automatically satisfies Eq. 10.5. In this parameterization, the loss is

$$L = l_1 l_2 = \frac{1}{E^2 \cos^2\alpha \sin^2\alpha} = \left(\frac{2}{E \sin(2\alpha)}\right)^2.$$

Obviously, L becomes minimal if $\alpha = \frac{\pi}{4}$, so that $l_1 = l_2$ then. Thus, dividing the forest into square tiles is better than dividing it into any other rectangles. This result is not really surprising; it just reflects the fact that the square is the rectangle with the best ratio of area to perimeter length. However, this result does not mean that square tiles are in fact the best solution of our optimization problem. We could, e. g., also divide the forest into triangles or hexagons; and the latter provide a better solution of the optimization problem than squares do.

But how can this a concept be an alternative to SOC? Obviously, all fires are equally sized if the fire breaks form a regular pattern, and this behavior is far away from any kind of scale invariance. However, the idea of minimizing the loss becomes more interesting in case of a spatially inhomogeneous distribution of the sparks. Inhomogeneity may arise, e. g., from surface topography. We may expect that the rate of lightning is higher on top of hills than in valleys. The self-organization in the original forest-fire model may compensate the inhomogeneity, but in contrast, inhomogeneity obviously affects the strategy of minimizing loss: At those locations where fires frequently occur, it is advisable to spend more effort for limiting the sizes of the fires. Figure 10.5 shows how an optimized pattern of fire breaks may look under spatially inhomogeneous conditions. Here we have assumed that the fires are caused by careless people. At one edge, the forest is bounded by a road, and we have assumed that the danger of fires is high in the vicinity of this road.

Let us now derive a quantitative criterion for minimizing the average loss per fire under spatially inhomogeneous conditions. Let $q(\boldsymbol{x})$ be the probability density of the spatial distribution of the sparks at the location $\boldsymbol{x}$. It is defined in such a way that $\int_\Omega q(\boldsymbol{x})\, dx_1 dx_2$ is the probability that a randomly chosen spark hits the subdomain Ω for each subdomain Ω. We further assume that the model domain is separated by fire breaks into quadratic areas whose sizes are not necessarily the same. A value $A(\boldsymbol{x})$ is assigned to each point $\boldsymbol{x}$ of the

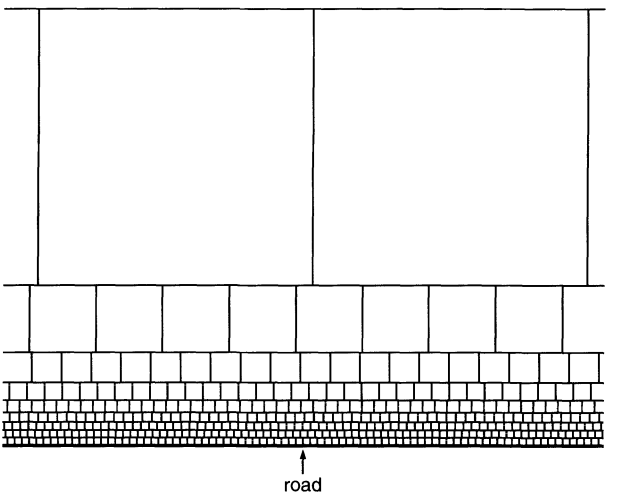

Fig. 10.5. Sketch of an optimized pattern of fire breaks under spatially inhomogeneous conditions. The sparking frequency is assumed to be high in the vicinity of the road.

model domain; it describes the size of the patch that includes the location $\boldsymbol{x}$. Then, the average loss induced by a spark is

$$L = \int q(\boldsymbol{x})\, A(\boldsymbol{x})\, dx_1 dx_2, \tag{10.6}$$

where the integration is performed over the whole domain. The total effort E – the sum of the lengths of all fire breaks – can be expressed with the help of $A(\boldsymbol{x})$, too. The effort attributed to a quadratic patch of area A is half the length of its perimeter, i. e., $2\sqrt{A}$. Thus, the function

$$e(\boldsymbol{x}) := \frac{2\sqrt{A(\boldsymbol{x})}}{A(\boldsymbol{x})} = 2\, A(\boldsymbol{x})^{-\frac{1}{2}} \tag{10.7}$$

quantifies the effort per unit area within the patch, so that the total effort required for subdividing the whole domain is

$$E = \int e(\boldsymbol{x})\, dx_1 dx_2. \tag{10.8}$$

The average loss (Eq. 10.6) can easily be expressed by $e(\boldsymbol{x})$ instead of $A(\boldsymbol{x})$ with the help of Eq. 10.7:

$$L = \int \frac{q(\boldsymbol{x})}{4\, e(\boldsymbol{x})^2}\, dx_1 dx_2.$$

This expression must be minimized under the side condition given by Eq. 10.8.

Apart from this condition, only a limited set of functions may be used for minimizing L: According to the procedure introduced above, $e(\boldsymbol{x})$ must be constant within quadratic areas. At this point, the concept of HOT introduces a simplification by allowing arbitrary functions. Then the technique of

Lagrangian multipliers can be applied to the minimization problem; otherwise a quite costly numerical minimization would we necessary. As above, we can get around the theoretical effort of introducing Lagrangian multipliers by choosing an appropriate parameterization of $e(\boldsymbol{x})$. Let us write $e(\boldsymbol{x})$ in the form

$$e(\boldsymbol{x}) \;=\; \phi(\boldsymbol{x}) - \overline{\phi} + E$$

for an arbitrary function $\phi(\boldsymbol{x})$ where

$$\overline{\phi} \;\coloneqq\; \frac{\int \phi(\boldsymbol{x}) dx_1 dx_2}{\int dx_1 dx_2}$$

denotes the mean value of $\phi(\boldsymbol{x})$ over the whole domain. Then, $e(\boldsymbol{x})$ automatically satisfies the side condition defined by Eq. 10.8, so that our task has turned into minimizing

$$L(\phi) \;=\; \int \frac{q(\boldsymbol{x})}{4\left(\phi(\boldsymbol{x}) - \overline{\phi} + E\right)^2}\, dx_1 dx_2$$

among all functions $\phi(\boldsymbol{x})$. Let $\phi(\boldsymbol{x})$ be a function which minimizes $L(\phi)$. We now consider $L(\phi + \epsilon\psi)$ where $\psi(\boldsymbol{x})$ is an arbitrary function and ϵ is a real number. For each function $\psi(\boldsymbol{x})$, $L(\phi + \epsilon\psi)$ becomes minimal for $\epsilon = 0$, so that

$$\tfrac{\partial}{\partial\epsilon} L(\phi + \epsilon\psi)\big|_{\epsilon=0} \;=\; 0.$$

Form straightforward calculations we obtain

$$\begin{aligned}
\tfrac{\partial}{\partial\epsilon} L(\phi + \epsilon\psi)\big|_{\epsilon=0} \;&=\; -\tfrac{1}{2} \int \frac{q(\boldsymbol{x})}{e(\boldsymbol{x})^3} \left(\psi(\boldsymbol{x}) - \overline{\psi}\right) dx_1 dx_2 \\
&=\; -\tfrac{1}{2} \int \left(\frac{q(\boldsymbol{x})}{e(\boldsymbol{x})^3} - \frac{\int \frac{q(\boldsymbol{x}')}{e(\boldsymbol{x}')^3} dx_1' dx_2'}{\int dx_1' dx_2'} \right) \psi(\boldsymbol{x})\, dx_1 dx_2.
\end{aligned}$$

Since this term vanishes for all functions $\psi(\boldsymbol{x})$, the term within the brackets must vanish for all locations $\boldsymbol{x}$. This means that the function $q(\boldsymbol{x})/e(\boldsymbol{x})^3$ is constant because the second term in the brackets is independent of $\boldsymbol{x}$. We thus obtain

$$e(\boldsymbol{x}) \;\sim\; q(\boldsymbol{x})^{\frac{1}{3}}.$$

This result confirms our hypothesis that much effort should be spent at these locations where the probability of sparking is high and vice versa. With the help of Eq. 10.7, this relationship can be transformed into an expression for the sizes of the patches:

$$A(\boldsymbol{x}) \;\sim\; q(\boldsymbol{x})^{-\frac{2}{3}}.$$

This may be an interesting result when applied to optimizing certain systems. But after all, we have still not seen any power-law distribution of event sizes

in this concept. So let us now find out under which conditions and to what extent HOT results in Pareto-distributed event sizes.

For obtaining a more general result, we leave the forest-fire model here and assume an arbitrary system where $q(\boldsymbol{x})$ is still the probability density of the spatial event distribution. We assume that the sizes $s(\boldsymbol{x})$ (not necessarily measured in terms of area) of the events occurring at the location $\boldsymbol{x}$ are related to $q(\boldsymbol{x})$ by

$$s(\boldsymbol{x}) \;=\; c\,q(\boldsymbol{x})^{-\alpha}. \tag{10.9}$$

Other exponents than the value $\alpha = \frac{2}{3}$ obtained above may arise from different criteria for measuring the loss induced by an event, different criteria for measuring the effort or different spatial dimensions. The constant c specifies the total effort in some way. Let Ω_q be the part of the model domain where $q(\boldsymbol{x}) \geq q$, and let $|\Omega_q| := \int_{\Omega_q} dx_1 dx_2$ denote the size of Ω_q. According to Eq. 10.9, Ω_q is exactly that part of the domain where the event sizes are not larger than $cq^{-\alpha}$; so the cumulative distribution of the event sizes follows the relationship

$$1 - P(s) \;=\; \int_{\Omega_q} q(\boldsymbol{x})\, dx_1 dx_2 \quad \text{for} \quad s \;=\; c\,q^{-\alpha}. \tag{10.10}$$

According to Eq. 1.3, the Pareto distribution is $P(s) = 1$ for all sizes below a minimum event size $s_{\min}$. Consequently, there must be an upper limit $q_{\max}$ of $q(\boldsymbol{x})$ that is related to $s_{\min}$ by $s_{\min} = cq_{\max}^{-\alpha}$. From Eq. 1.3 we obtain

$$P(s) \;=\; \left(\frac{s}{s_{\min}}\right)^{-b} \;=\; \left(\frac{q}{q_{\max}}\right)^{\alpha b} \quad \text{for} \quad q \;\leq\; q_{\max}.$$

Inserting this result into Eq. 10.10 and deriving both sides of the equation with respect to q yields

$$-\frac{\alpha b\, q^{\alpha b - 1}}{q_{\max}^{\alpha b}} \;=\; \frac{\partial}{\partial q} \int_{\Omega_q} q(\boldsymbol{x})\, dx_1 dx_2 \quad \text{for} \quad q \;<\; q_{\max}. \tag{10.11}$$

The derivative of the integral can be computed according to

$$\begin{aligned} \frac{\partial}{\partial q} \int_{\Omega_q} q(\boldsymbol{x})\, dx_1 dx_2 \;&=\; \lim_{\epsilon \to 0} \frac{\int_{\Omega_{q+\epsilon}} q(\boldsymbol{x})\, dx_1 dx_2 - \int_{\Omega_q} q(\boldsymbol{x})\, dx_1 dx_2}{\epsilon} \\ &=\; -\lim_{\epsilon \to 0} \frac{\int_{\Omega_q \setminus \Omega_{q+\epsilon}} q(\boldsymbol{x})\, dx_1 dx_2}{\epsilon} \end{aligned}$$

where $\Omega_q \setminus \Omega_{q+\epsilon}$ is the difference between Ω_p and $\Omega_{q+\epsilon}$, i.e., the set where $q(\boldsymbol{x})$ is between q and $q+\epsilon$. Thus, the term $q(\boldsymbol{x})$ in the integral converges towards q in the limit $\epsilon \to 0$, so that

$$\frac{\partial}{\partial q} \int_{\Omega_q} q(\boldsymbol{x})\, dx_1 dx_2 \;=\; -q \lim_{\epsilon \to 0} \frac{\int_{\Omega_q \setminus \Omega_{q+\epsilon}} dx_1 dx_2}{\epsilon}$$

$$= q \lim_{\epsilon \to 0} \frac{\int_{\Omega_{q+\epsilon}} dx_1 dx_2 - \int_{\Omega_q} dx_1 dx_2}{\epsilon}$$

$$= q \lim_{\epsilon \to 0} \frac{|\Omega_{q+\epsilon}| - |\Omega_q|}{\epsilon} = q \frac{\partial}{\partial q} |\Omega_q|.$$

Combining this result with Eq. 10.11 leads to

$$\frac{\partial}{\partial q} |\Omega_q| = - \frac{\alpha b \, q^{\alpha b - 2}}{q_{\max}^{\alpha b}},$$

so that

$$|\Omega_q| = \int_q^{q_{\max}} \frac{\alpha b \, q'^{\alpha b - 2}}{q_{\max}^{\alpha b}} dq' = \begin{cases} \frac{\log\left(\frac{q_{\max}}{q}\right)}{q_{\max}} & \text{if} \quad \alpha b = 1 \\ \frac{\alpha b}{\alpha b - 1} \frac{1}{q_{\max}} \left(1 - \left(\frac{q}{q_{\max}}\right)^{\alpha b - 1}\right) & \text{else} \end{cases} . \tag{10.12}$$

Obviously, $|\Omega_q|$ is proportional to the probability $P(q)$ that $q(\boldsymbol{x})$ is at least q at an randomly chosen site $\boldsymbol{x}$. Thus, Pareto distributions of the event sizes with arbitrary exponents b can be achieved by choosing an appropriate distribution $P(q)$. Three different cases can be distinguished:

1. Pareto distributions with exponents $b < \frac{1}{\alpha}$ emerge from distributions of the form
$$P(q) \sim q^{-\gamma} - q_{\max}^{-\gamma}$$
where $\gamma = 1 - \alpha b$. As discussed in Sect. 2.2, this is an upper-truncated power-law distribution.
2. A Pareto distribution with an exponent $b = \frac{1}{\alpha}$ emerges from a distribution of the form
$$P(q) \sim \log\left(\frac{q_{\max}}{q}\right).$$
This is just a limiting case of the upper-truncated power-law distribution.
3. Pareto distributions with exponents $b > \frac{1}{\alpha}$ emerge from distributions of the form
$$P(q) \sim q_{\max}^{\gamma} - q^{\gamma}$$
where $\gamma = \alpha b - 1$. In principle, this is an upper-truncated power-law distribution with a positive exponent.

Therefore, HOT transforms power-law distributions (including the logarithmic limiting case and distributions with positive exponents) into power-law distributions of event sizes.

Let us now come back to the forest-fire model ($\alpha = \frac{2}{3}$). We consider a forest that is characterized by the conditions $x_1 > 0$ and $0 < x_2 < w$. As illustrated in Fig. 10.5, the line defined by $x_1 = 0$ shall be a road. The sparking probability $q(\boldsymbol{x})$ is highest immediately at the road and decreases monotonically with increasing distance x_1 from the road. Let us further assume that $q(\boldsymbol{x})$ is independent of x_2. Under these conditions, we obtain $\Omega_{q(\boldsymbol{x})} = w x_1$.

This result can be inserted into Eq. 10.12 in order to determine $q(\boldsymbol{x})$. Again, three cases can be distinguished:

1. Power-law distributed fires with exponents $b < \frac{3}{2}$ arise from spatial distributions of the sparking probability according to

$$q(\boldsymbol{x}) = q_{\max}\left(1 + \frac{1-\frac{2}{3}b}{\frac{2}{3}b}\, w\, q_{\max}\, x_1\right)^{-\frac{1}{1-\frac{2}{3}b}}.$$

2. Power-law distributed fires with an exponent $b = \frac{3}{2}$ arise from a spatial distribution of the sparking probability according to

$$q(\boldsymbol{x}) = q_{\max}\, \mathrm{e}^{-w q_{\max} x_1}.$$

3. Power-law distributed fires with exponents $b > \frac{3}{2}$ arise from spatial distributions of the sparking probability according to

$$q(\boldsymbol{x}) = \begin{cases} q_{\max}\left(1 - \frac{\frac{2}{3}b-1}{\frac{2}{3}b}\, w\, q_{\max}\, x_1\right)^{\frac{1}{\frac{2}{3}b-1}} & \text{if} \quad x_1 < \frac{\frac{2}{3}b}{\left(\frac{2}{3}b-1\right) w\, q_{\max}} \\ 0 & \text{else} \end{cases}.$$

This example shows that scale invariance of the input data may be hidden well if we only consider the spatial distribution of the sparking probability $q(\boldsymbol{x})$. In none of the tree cases, $q(\boldsymbol{x})$ suggests any scale invariance. Thus, it may look first as if HOT was a concept that derives scale invariance from a principle of optimization.

However, we have learned that HOT requires a fractal size distribution in the input data. We have already seen in Chap. 1 that a fractal size distribution does not imply any scale invariance in the spatial pattern of the input data. Thus, it is not surprising that the fractal properties of the input data required for generating power laws from HOT can be hidden behind a non-fractal spatial pattern. But in principle, HOT only reproduces these hidden scale invariance and alters the exponents.

11. Where do we Stand?

The concept of SOC is about 15 years old now, and there is no doubt that it has been a seminal idea. Considering phenomena from various scientific disciplines under aspects of SOC has been tempting and often successful. In this book, we have focused on examples from earth sciences – forest fires, earthquakes, landslides, and drainage networks. Forest fires and earthquakes were already placed in the framework of SOC in the early 1990's. Surprisingly, field evidence for a fractal size distribution of forest fires was found much later. Although modeling is still one step behind the available data concerning seismic activity, SOC has considerably deepened the understanding of the physics of earthquakes. Simple spring-block models do not only account for the size distribution of earthquakes, but also capture a large part of the complexity of earthquakes in nature.

The idea of explaining the observed scale-invariant properties of landslides with the help of SOC is rather new. It seems that SOC has made the frequency-magnitude relation of landslides interesting again; more and more extensive studies on the distribution of landslide sizes in nature have been carried out in the last years. Both modeling and field studies are just at the beginning, so only both branches together may be able to answer the open questions, e. g., whether the exponent of the distribution is universal or not.

In the example of drainage networks, SOC has recently been proposed as an alternative idea to established optimization approaches. Although the ideas are promising, it is too soon to assess their contribution to our understanding of landform evolution by fluvial erosion. Verifying of falsifying the results in field is difficult if not impossible at all.

The idea of SOC brings the reductionist approach of physics closer to earth sciences. Clearly, the apparent simplicity polarizes as fractals did in the 1980's. However, we should beware from deciding for either of the sides too soon; SOC has some relevance in earth sciences, but not everything is SOC. Deeper knowledge on the capabilites and on the limitations of SOC shall guide us towards a less polemic discussion.

The approach followed in this book is a quite phenomenological one, starting from models that were designed to understand certain phenomena occurring in nature. This approach is common in various disciplines, but the reader may miss the closure of the loop. Finally, the phenomena where SOC occurs

should be unified by a theory. But even after nearly 15 years, SOC is still far away from being fully understood. Some analytical theories have been applied to SOC, e. g., the mean field theory and the renormalization group method (e. g. Jensen 1998). However, these methods capture the complex spatial structure of most SOC systems only partly; so they can only be applied to even simpler models than those discussed in this book.

Due to the lack of an comprehensive analytical theory of SOC, theoretical physics is pulled back to the descriptive approach, too. From their basic structure, most of the models showing SOC are similar. The striking similarity can be characterized by the term *slowly driven, interaction-dominated threshold systems* (Jensen 1998). The first term expresses the need for incorporating a long-term driving force; otherwise a quasi-steady state which is characterized by fluctuations of all sizes cannot be maintained. Thresholds are also present in most SOC models, although threshold behavior is only one facet of non-linearity. However, thresholds at least make distinguishing events easier. The term interaction-dominated may be the most interesting aspect. In general, SOC models consist of a large number of interacting elements, and the interaction seems to be much more important than the behavior of the individual. However, everyone who as ever developed a model and searched for SOC knows that these criteria are not sufficient. It often depends on apparently minor points such as boundary conditions or a small randomness whether a model evolves towards a critical state or not. These phenomena are poorly understood. Finally, models such as the network evolution model exhibit SOC, but do not fit into these criteria at all.

As SOC systems evolve towards a critical state, elucidating the properties of the critical state is the goal in nearly all studies on SOC. However, the example of forest fires has shown that sudden changes in the conditions, i. e., model parameters or boundary conditions, have a strong impact on the system's properties. Although the system may finally adapt to the altered conditions and return to a critical state, the characteristics may be entirely different during the phase of transition. Landslides are an example where transitions may be even more important than the critical state itself. The long time required until transient components have vanished suggests that at least parts of the landscape are still on their way towards a critical state. While slow climatic changes may result in moderate deviations from the critical state, human impact is often strong enough to induce a sudden change in the conditions. Much more research on this field is needed, but it will probably improve our understanding of natural and man-made hazards.

So, what is SOC in its present state? It seems to be a phenomenon rather than a theory. Nevertheless, it unifies or at least categorizes a variety of phenomena which seem to be completely different from classical field studies, laboratory experiments or process models. In this sense, the framework of SOC gives common ground for multi-disciplinary research and will hopefully help to tear down walls between specialists from various disciplines.

A. Numerics of Ordinary Differential Equations

Many physical systems can be described by *ordinary differential equations* (ODEs) or by systems of ODEs. ODEs concern a function $f(t)$ of a single variable which is often identified with the time. In contrast, *partial differential equations* (PDEs) concern functions of more than one variable. The general form of an ODE of first order is

$$\frac{\partial}{\partial t} f(t) \;=\; \phi(t, f(t)), \tag{A.1}$$

where ϕ is a given function. The function $f(t)$ may be a vector; in this case we speak of a system of ODEs. The Lorenz equations discussed in Sect. 4.1 are a system where $f(t)$ consists of three components $A(t)$, $B(t)$, and $C(t)$. In the landform evolution models introduced in Chap. 9, $f(t)$ has as many components as the lattice has nodes, and it consists of the surface heights $H_i(t)$ of the nodes. In the Burridge-Knopoff model from Sect. 7.2, the situation is slightly different because second-order derivatives of the displacements $u_{i,j}(t)$ occur. This system can be transformed by introducing additional variables in form of the derivatives of the displacements (the velocities of the blocks). Then the number of components of $f(t)$ is twice the number of blocks.

Provided that ϕ satisfies some conditions concerning regularity, Eq. A.1 determines the evolution of $f(t)$ through time uniquely if $f(t_0)$ is given at an arbitrary time t_0. Solving this *initial-value problem* consists of computing $f(t)$ for times $t > t_0$.

A few systems of ODEs can be solved analytically, but the majority can only be solved approximately by means of numerical simulations. Although theory of ODEs is less complicated than that of PDEs, the variety of methods makes numerical treatment of ODEs a wide field. In this short introduction, we can just review briefly the most widespread approaches and discuss their properties in the simple example

$$\frac{\partial}{\partial t} f(t) \;=\; -\, f(t)$$

with the initial condition $f(0) = 1$. This ODE describes a decay as it occurs in nuclear physics; its solution is $f(t) = \mathrm{e}^{-t}$. General theory is presented in several books (e. g. Hairer et al. 1987, 1990; Stetter 1973).

Most of the numerical methods compute the solution at discrete times; the transition to such a discrete system is called *discretization*. In solving initial-value problems, the *time step* is the fundamental part of the algorithm; its

task is computing an approximate solution at the time $t+\delta t$ if that at the time t (and perhaps at earlier times) is known.

In general, the accuracy of a numerical scheme decreases if the time step length δt increases. Different algorithms may lead to approaches of different accuracies, too. Depending on the problem, there are various criteria for assessing the quality of a numerical scheme, such as

- high accuracy in the limit $\delta t \to 0$;
- high accuracy at a given, finite time step length δt;
- reasonable approximation even if δt is large.

The first criterion can be quantified by the *order* of a scheme. A scheme is of order p if the error, i. e., the difference between exact solution and numerical approximation, decreases like δt^p in the limit $\delta t \to 0$. One can easily see that a scheme of a higher order is better than any scheme of lower order if δt is sufficiently small. However, schemes of high order require a high degree of smoothness of the function ϕ; if, for instance, ϕ is discontinuous, the advantage of a high-order scheme vanishes. Moreover, there is in principle not any relation between the order and the two other criteria, so that one must take care that δt is in fact small enough to take advantage of a high-order scheme. Finally, the quality of the scheme must be compared to the effort in programming and in computing time.

Let us only consider schemes which involve information from the times t and $t+\delta t$, but not from earlier times. A straightforward approximation of the derivative is

$$\frac{\partial}{\partial t} f(t) \approx \frac{\tilde{f}(t+\delta t) - \tilde{f}(t)}{\delta t}, \tag{A.2}$$

where $\tilde{f}(t)$, respectively $\tilde{f}(t+\delta t)$ is the approximate solution. Inserting this approximation into Eq. A.1 leads to *Euler's scheme*:

$$\tilde{f}(t+\delta t) = \tilde{f}(t) + \delta t\, \phi(t, \tilde{f}(t)).$$

Applying one step of Euler's scheme to our example leads to $\tilde{f}(\delta t) = 1 - \delta t$. Comparing this result with the Taylor series of exact solution yields

$$\tilde{f}(\delta t) - f(\delta t) = 1 - \delta t - \mathrm{e}^{-t} = 1 - \delta t - \sum_{n=0}^{\infty} \frac{(-\delta t)^n}{n!} = -\tfrac{1}{2}\delta t^2 + \ldots$$

Thus, the numerical solution is exact up to the linear term in δt after the first time step; the error decreases quadratically with δt in the limit $\delta t \to 0$. For computing the approximate solution at a given time t, we need several time steps; the number of required steps is $\frac{t-t_0}{\delta t}$. Since the error is systematic (not random), the total error a given time t is the error of a single step multiplied by the number of steps. Therefore, the error at a given time t only decreases linearly with δt in our example, so that Euler's scheme is a first-order scheme.

Euler's method is an *explicit scheme.* This means that only information from the time t (where everything is already known) is used. However, when inserting the approximation of the derivative (Eq. A.2) into Eq. A.1, it is not clear at all that the right-hand side must be evaluated at the time t; we could also have taken the time $t+\delta t$, which leads to the *backward Euler scheme*:

$$\tilde{f}(t+\delta t) - \delta t\, \phi(t+\delta t, \tilde{f}(t+\delta t)) \;=\; \tilde{f}(t). \tag{A.3}$$

A scheme where the time $t+\delta t$ occurs at the right-hand side of the discretized equation is denoted *implicit scheme.* If the right-hand side is simply evaluated at the time $t+\delta t$, the scheme is *fully implicit* (like the backward Euler method).

In our simple example, a fully implicit time step can be performed as easily as an explicit step: $\tilde{f}(\delta t) = \frac{1}{1+\delta t}$. With the help of a geometric series we can quantify the error of this approximation:

$$\tilde{f}(\delta t) - f(\delta t) \;=\; \frac{1}{1+\delta t} - \mathrm{e}^{-t} \;=\; \sum_{n=0}^{\infty} (-\delta t)^n - \sum_{n=0}^{\infty} \frac{(-\delta t)^n}{n!} \;=\; \tfrac{1}{2}\delta t^2 + \ldots$$

So this scheme is of first order, too, but a major difference to the explicit scheme occurs if δt becomes large. While the solution obtained from Euler's scheme diverges in the limit $\delta t \to \infty$ and thus becomes totally wrong, that of the backward Euler scheme approaches the correct value 0. Thus, the fully implicit scheme provides reasonable results in case of too large time steps in contrast to the explicit scheme.

This result is not restricted to our example, but holds for many equations; this is often a clear advantage of implicit schemes. However, implicit schemes result in a system of (perhaps non-linear) equations in case of a system of ODEs. This system has to be solved in each time step, which requires a higher numerical effort than simply applying an explicit scheme. Consequently, it depends in the problem which scheme is preferable.

Forward and backward Euler scheme can be combined to a *symmetric* scheme, the *Crank-Nicholson scheme.* Here, the mean value of $\phi(t, \tilde{f}(t))$ and $\phi(t+\delta t, \tilde{f}(t+\delta t))$ is inserted into the right-hand side of Eq. A.1. In combination with Eq. A.2 this leads to

$$\tilde{f}(t+\delta t) - \tfrac{1}{2}\delta t\, \phi(t+\delta t, \tilde{f}(t+\delta t)) \;=\; \tilde{f}(t) + \tfrac{1}{2}\delta t\, \phi(t, \tilde{f}(t)).$$

This seems to be the most straightforward choice because (Eq. A.2) is the best approximation of the derivative in the middle between t and $t+\delta t$. So it is not surprising that this scheme is of higher accuracy than forward and backward Euler schemes in the limit $\delta t \to 0$. In our example we obtain

$$\tilde{f}(\delta t) - f(\delta t) \;=\; \frac{1 - \frac{1}{2}\delta t}{1 + \frac{1}{2}\delta t} - \mathrm{e}^{-t} \;=\; -\tfrac{1}{12}\delta t^3 + \ldots$$

Thus, the approximation converges towards the exact solution faster than those obtained from forward and backward Euler methods in the limit $\delta t \to 0$; the Crank-Nicholson scheme is of second order. However, it does not converge towards the correct value for $\delta t \to \infty$, so that the backward Euler scheme is still better for large time steps.

Second-order accuracy cannot only be achieved by combining explicit and fully implicit method; the implicit term $\phi(t+\delta t, \tilde{f}(t+\delta t))$ can also be estimated from the first-order Euler scheme:

$$\phi(t+\delta t, \tilde{f}(t+\delta t)) \approx \phi(t+\delta t, \tilde{f}(t) + \delta t\, \phi(t, \tilde{f}(t))),$$

which leads to *Heun's method*:

$$\tilde{f}(t+\delta t) = \tilde{f}(t) + \tfrac{1}{2}\delta t \left(\phi(t, \tilde{f}(t)) + \phi(t+\delta t, \tilde{f}(t) + \delta t\, \phi(t, \tilde{f}(t)))\right).$$

Alternatively, we can estimate ϕ at the time $t+\frac{1}{2}\delta t$ directly according to

$$\phi\left(t+\tfrac{1}{2}\delta t, \tilde{f}\left(t+\tfrac{1}{2}\delta t\right)\right) \approx \phi\left(t+\tfrac{1}{2}\delta t, \tilde{f}(t)+\tfrac{1}{2}\delta t\, \phi(t, \tilde{f}(t))\right)$$

which results in the modified Euler scheme:

$$\tilde{f}(t+\delta t) = \tilde{f}(t) + \delta t\, \phi\left(t + \tfrac{1}{2}\delta t, \tilde{f}(t) + \tfrac{1}{2}\delta t\, \phi(t, \tilde{f}(t))\right).$$

If ϕ is sufficiently smooth and a high accuracy is desired, even schemes of higher orders are applied. The most widespread high-order schemes are the *Runge-Kutta methods*. Among them, the fourth-order scheme is widely used:

$$\tilde{f}(t+\delta t) = \tilde{f}(t) + \tfrac{1}{2}\delta t\,(k_1 + 2k_2 + 2k_3 + k_4),$$

where

$$k_1 := \phi\left(t, \tilde{f}(t)\right), \qquad k_2 := \phi\left(t+\tfrac{1}{2}\delta t, \tilde{f}(t)+\tfrac{1}{2}\delta t\, k_1\right),$$
$$k_3 := \phi\left(t+\tfrac{1}{2}\delta t, \tilde{f}(t)+\tfrac{1}{2}\delta t\, k_2\right), \quad \text{and} \quad k_4 := \phi\left(t+\delta t, \tilde{f}(t)+\delta t\, k_3\right).$$

In this scheme, reducing δt by one order of magnitude enhances the accuracy of the result by four orders of magnitude, provided that ϕ is sufficiently smooth, and that δt is small enough. This property makes the scheme (and other high-order schemes) attractive if a high accuracy is required. However, one must be careful to ensure that δt is not too large; if it is, a high-order scheme may be even worse than a method of low order. Schemes of higher order than four are rarely used because the error tends to reach the numerical accuracy of floating point numbers in the computer before δt is small enough to take advantage of the rapid convergence.

References

Abrahams AD, Li G, Parsons AJ (1996) Rill hydraulics on a semiarid hillslope, Southern Arizona. Earth Surf Process Landforms 21:35–47

Adler A, Feldman R, Taqqu MS (eds.) (2000) A Practical Guide to Heavy Tails: Statistical Techniques for Analyzing Heavy Tailed Distributions. Birkhäuser, Basel, Berlin, Boston

Aki K, Richards PG (1980) Quantitative Seismology. W. H. Freeman & Co., San Francisco

Bak P (1996) How Nature Works – the Science of Self-Organized Criticality. Copernicus, Springer, Berlin, Heidelberg, New York

Bak P, Chen K, Tang C (1990) A forest-fire model and some thoughts on turbulence. Phys Lett A 147:297–300

Bak P, Tang C (1989) Earthquakes as a self-organized critical phenomenon. J Geophys Res 94:15,635–15,637

Bak P, Tang C, Wiesenfeld K (1987) Self-organized criticality. An explanation of 1/f noise. Phys Rev Lett 59:381–384

Bak P, Tang C, Wiesenfeld K (1988) Self-organized criticality. Phys Rev A 38:364–374

Bassingthwaite JB, Raymond GM (1994) Evaluating rescaled range analysis for time series. Ann Biomed Eng 22:432–444

Bennett JG (1936) Broken coal. J Inst Fuel 10:22–39

Binney JJ, Dowrick NJ, Fisher AJ, Newman MEJ (1992) The Theory of Critical Phenomena – an Introduction to the Renormalization Group. Oxford University Press, Oxford

Bolt BA (1999) Earthquakes. W. H. Freeman & Co., New York

Bonnet E, Bour O, Odling NE, Davy P, Main I, Cowie P, Berkowitz B (2001) Scaling of fracture systems in geological media. Rev Geophys 39:347–383

Bouchaud JP, Cates ME, Prakash JR, Edwards SF (1995) Hysteresis and metastability in a continuum sandpile model. Phys Rev Lett 74:1982–1985

Bour O, Davy P (1999) Clustering and size distribution of fault patterns: theory and measurements. Geophys Res Lett 26:2001–2004

Bromhead EN (1992) The Stability of Slopes. Blackie Academic & Professional, Chapman & Hall, Glasgow, 2nd edn.

Brown SR (1987) Fluid flow through rock joints: the effect of surface roughness. J Geophys Res Solid Earth 92:1337–1347

Brown SR, Scholz CH, Rundle JB (1991) A simplified spring-block model of earthquakes. Geophys Res Lett 18:215–218

Brunsden D, Prior BD (1984) Slope Instability. John Wiley & Sons, New York, Chichester, Brisbane

Burridge R, Knopoff L (1967) Model and theoretical seismicity. Bull Seismol Soc Am 57:341–371

Caldarelli G, Giacometti A, Maritan A, Rodriguez-Iturbe I, Rinaldo A (1997) Randomly pinned landform evolution. Phys Rev E 55:4865–4868

Carlson JM (1991) Time intervals between characteristic earthquakes and correlations with smaller events: an analysis based on a mechanical model of a fault. J Geophys Res 96:4255–4267

Carlson JM, Doyle J (1999) Highly optimized tolerance: a mechanism for power laws in designed systems. Phys Rev E 60:1412–1427

Carlson JM, Doyle J (2000) Highly optimized tolerance: robustness and design in complex systems. Phys Rev Lett 84:2529–2532

Carlson JM, Langer JS (1989a) Mechanical model of an earthquake fault. Phys Rev A 40:6470–6484

Carlson JM, Langer JS (1989b) Properties of earthquakes generated by fault dynamics. Phys Rev Lett 62:2632–2635

Carlson JM, Langer JS, Shaw BE (1994) Dynamics of earthquake faults. Rev Mod Phys 66:657–670

Carlson JM, Langer JS, Shaw BE, Tang C (1991) Intrinsic properties of a Burridge-Knopoff model of an earthquake fault. Phys Rev A 44:884–897

Chabra AB, Feigenbaum MJ, Kadanoff LP, Kolan AJ, Procaccia I (1993) Sandpiles, avalanches, and the statistical mechanics of non-equilibrium stationary states. Phys Rev E 47:3099–3121

Chen K, Bak P, Obukhov SP (1991) Self-organized criticality in a crack-propagation model of earthquakes. Phys Rev A 43:625–630

Christensen K, Hamon D, Jensen HJ, Lise S (2001) Comment on "Self-organized criticality in the Olami-Feder-Christensen model. Phys Rev Lett 87:039081

Christensen K, Olami Z (1992) Scaling, phase transitions, and nonuniversality in a self-organized critical cellular-automaton model. Phys Rev A 46:1829–1838

Clar S, Drossel B, Schwabl F (1994) Scaling laws and simulation results for the self-organized critical forest-fire model. Phys Rev E 50:1009–1018

Colaiori F, Flammini A, Maritan A, Banavar JR (1997) Analytical and numerical study of optimal channel networks. Phys Rev E 55:1298–1310

Cooley JW, Tukey JW (1965) An algorithm for the machine calculation of complex Fourier series. Math Comput 19:297–301

Coulthard TJ, Kirkby MJ, Macklin MG (1998) Non-linearity and spatial resolution in a cellular automaton model of a small upland basin. Hydrology and Earth System Sciences 2:257–264

de Carvalho JX, Prado CPC (2000) Self-organized criticality in the Olami-Feder-Christensen model. Phys Rev Lett 84:4006–4009
de Carvalho JX, Prado CPC (2001) Reply to comment on "Self-organized criticality in the Olami-Feder-Christensen model. Phys Rev Lett 87:039082
de Sousa Vieira M, Vasconcelos GL, Nagel SR (1993) Dynamics of spring-blocks models: tuning to criticality. Phys Rev E 47:2221–2224
Densmore AL, Anderson RS, McAdoo B, Ellis MA (1997) Hillslope evolution by bedrock landslides. Science 275:369–372
Densmore AL, Ellis MA, Anderson RS (1998) Landsliding and the evolution of normal-fault-bounded mountains. J Geophys Res 103:15203–15219
Drossel B, Schwabl F (1992) Self-organized critical forest-fire model. Phys Rev Lett 69:1629–1632
Emmett WW (1970) The hydraulics of overland flow on hillslopes. No. 662-A in US Geol. Survey Prof. Papers. US Government Printing Office, Washington D.C.
Evans IS, McClean CJ (1995) The land surface is not unifractal: variograms, cirque scale and allometry. Z Geomorph N F Suppl. 101:127–147
Falconer K (1990) Fractal Geometry – Mathematical Foundations and Applications. John Wiley & Sons, New York, Chichester, Brisbane
Feder HJS, Feder J (1991) Self-organized criticality in a stick-slip process. Phys Rev Lett 66:2669–2672
Feder J (1988) Fractals. Plenum Press, New York
Fox CG, Hayes DE (1985) Quantitative methods for analyzing the roughness of the seafloor. Rev Geophys 23:1–48
Frette V, Christensen K, Malthe-Sørenssen A, Feder J, Jøssang T, Meakin P (1995) Avalanche dynamics in a pile of rice. Nature 379:49
Fujiwara A, Kamimoto G, Tsukamoto A (1977) Destruction of basaltic bodies by high-velocity impact. Icarus 32:277–288
Fuyii Y (1969) Frequency distribution of the magnitude of landslides caused by heavy rainfall. Seismol Soc Japan J 22:244–247
Goltz C (1996) Multifractal and entropic properties of landslides in Japan. Geol Rundsch 85:71–84
Govers G (1992) Relationship between discharge, velocity, and flow area for rills eroding in loose, non-layered material. Earth Surf Process Landforms 17:515–528
Grassberger P (1993) On a self-organized critical forest fire model. J Phys A 26:2081–2089
Grassberger P (1994) Efficient large-scale simulations of a uniformly driven system. Phys Rev E 49:2436–2444
Grassberger P, Kantz H (1991) On a forest fire model with supposed self-organized criticality. J Stat Phys 63:685–700
Gutenberg B, Richter CF (1954) Seismicity of the Earth and Associated Phenomenon. Princeton University Press, Princeton, 2nd edn.

Hack JT (1957) Studies of longitudinal profiles in Virginia and Maryland. No. 294-B in US Geol. Survey Prof. Papers. US Government Printing Office, Washington D.C.

Hainzl S, Zöller G, Kurths J (1999) Similar power laws for foreshock and aftershock sequences in a spring-block model for earthquakes. J Geophys Res 104:7243–7253

Hairer E, Nørsett SP, Wanner G (1987) Solving Ordinary Differential Equations I. Nonstiff Problems. Springer, Berlin, Heidelberg, New York

Hairer E, Nørsett SP, Wanner G (1990) Solving Ordinary Differential Equations II. Stiff Problems. Springer, Berlin, Heidelberg, New York

Hallgass R, Loreto V, Mazzella O, Paladin G, Pietronero L (1997) Earthquake statistics and fractal faults. Phys Rev E 56:1346–1356

Hansen A, Schmittbuhl J, Batrouni GG, Oliveira FA (2000) Normal stress distribution of rough surfaces in contact. Geophys Res Lett 27:3639–3642

Harp EL, Jibson RW (1995) Inventory of landslides triggered by the 1994 Northridge, Califonia earthquake. Open File Report 95-213, US Geol. Survey, Washington D.C.

Harp EL, Jibson RW (1996) Landslides triggered by the 1994 Northridge, Califonia earthquake. Bull Seismol Soc Amer 86:319–322

Henley CL (1989) Self-organized percolation: a simpler model. Bull Am Phys Soc 34:838

Henley CL (1993) Statics of a "self-organized" percolation model. Phys Rev Lett 71:2741–2744

Hergarten S, Neugebauer HJ (1996) A physical statistical approach to erosion. Geol Rundsch 85:65–70

Hergarten S, Neugebauer HJ (1997) Homogenization of Manning's formula for modeling surface runoff. Geophys Res Lett 24:877–880

Hergarten S, Neugebauer HJ (1998) Self-organized criticality in a landslide model. Geophys Res Lett 25:801–804

Hergarten S, Neugebauer HJ (1999) Self-organized criticality in landsliding processes. In: Hergarten S, Neugebauer HJ (eds.), Process Modelling and Landform Evolution, vol. 78 of *Lecture Notes in Earth Sciences*, pp. 231–249. Springer, Berlin, Heidelberg, New York

Hergarten S, Neugebauer HJ (2000) Self-organized criticality in two-variable models. Phys Rev E 61:2382–2385

Hergarten S, Neugebauer HJ (2001) Self-organized critical drainage networks. Phys Rev Lett 86:2689–2692

Hergarten S, Paul G, Neugebauer HJ (2000) Modelling surface runoff. In: Schmidt J (ed.), Soil Erosion – Application of Physically Based Models, pp. 295–306. Springer, Berlin, Heidelberg, New York

Horton RE (1945) Erosional development of streams and their drainage basins; hydrophysical approach to quantitative morphology. Bull Geol Soc Am 56:275–370

Hovius N, Stark CP, Allen PA (1997) Sediment flux from a mountain belt derived by landslide mapping. Geology 25:231–234

Hovius N, Stark CP, Chu HT, Lin JC (2000) Supply and removal of sediment in a landslide-dominated mountain belt: Central Range, Taiwan. J Geol 108:73–89

Howard AD (1994) A detachment-limited model for drainage basin evolution. Water Resour Res 30:2261–2285

Huang J, Turcotte DL (1990) Are earthquakes an example of deterministic chaos? Geophys Res Lett 17:223–226

Huang J, Turcotte DL (1992) Chaotic seismic faulting with a mass-spring model and velocity-weakening friction. Pure Appl Geophys 138:569–589

Huber G (1991) Scheidegger's rivers, Takayasu's aggregates, and continued fractions. Physica A 209:463–470

Hurst HE (1951) Long-term storage capacity of reservoirs. Trans Am Soc Civil Engineers 116:770–808

Hurst HE (1957) A suggested statistical model of some time series which occur in nature. Nature 180:494

Hurst HE, Black RP, Simaika YM (1965) Long-Term Storage: An Experimental Study. Constable, London

Inaoka H, Takayasu H (1993) Water erosion as a fractal growth process. Phys Rev E 47:899–910

Ito K, Matsuzaki M (1990) Earthquakes as self-organized critical phenomena. J Geophys Res 95:6853–6860

Iverson RM (1997) The physics of debris flow. Rev Geophys 35:245–296

Jackson TJ, Vine DL (1996) Mapping surface soil moisture using an aircraft-based passive microwave instrument: algorithm and example. J Hydrol 184:85–99

Jan CD, Shen HW (1997) Review dynamic modelling of debris flows. In: Armanini A, Michiue M (eds.), Recent developments on debris flows, vol. 64 of *Lecture Notes in Earth Sciences*, pp. 93–116. Springer, Berlin, Heidelberg, New York

Jensen HJ (1991) 1/f noise from the linear diffusion equation. Physica Scripta 43:593–595

Jensen HJ (1998) Self-Organized Criticality – Emergent Complex Behaviour in Physical and Biological Systems, vol. 10 of *Lecture Notes in Physics*. Cambridge University Press, Cambridge, New York, Melbourne

Jensen HJ, Christensen K, Fogedby C (1989) 1/f noise, distribution of lifetimes, and a pile of sand. Phys Rev B 40:7425–7427

Jones LM, Molnar P (1979) Some characteristics of foreshocks and their possible relationship to earthquake prediction and premonitory slip on faults. J Geophys Res 84:3596–3608

Kagan YY, Knopoff L (1978) Statistical study of the occurrence of shallow earthquakes. Geophys J R Astron Soc 55:67–86

Kaminski E, Jaupart C (1998) The size distribution of pyroclasts and the fragmentation sequence in explosive volcanic eruptions. J Geophys Res Solid Earth 103:29,759–29,779

Kanamori H, Anderson DL (1975) Theoretical basis of some empirical relations in seismology. Bull Seismol Soc Am 65:1072–1096

Kent C, Wong J (1982) An index of littoral zone complexity and its measurement. Can J Fish Aquat Sci 39:847–863

Kiersch GA (1964) Vaiont reservoir disaster. Civil Engin 34:32–29

Kinouchi O, Prado CP (1999) Robustness of scale invariance in models with self-organized criticality. Phys Rev E 59:4964–4969

Knopoff L, Landoni JA, Abinante MS (1992) Dynamical model of an earthquake fault with localization. Phys Rev A 46:7445–7449

Kooi H, Beaumont C (1996) Large-scale geomorphology: classical concepts reconciled and integrated with contemporary ideas via a surface process model. J Geophys Res 101:3361–3386

Kosakowski G, Kasper H, Taniguchi T, Kolditz O, Zielke W (1997) Analysis of groundwater flow and transport in fractured rock – geometric complexity of numerical modelling. Z Angew Geol 43:81–84

Kramer S, Marder M (1992) Evolution of river networks. Phys Rev Lett 68:381–384

Krapivsky PL, Grosse I, Ben-Naim E (2000) Scale invariance and lack of self-averaging in fragmentation. Phys Rev E 61:R993–R996

Lay T, Wallace TC (1995) Modern Global Seismology, vol. 58 of *International Geophysics Series*. Academic Press, San Diego

Leopold LB, Langbein WB (1962) The concept of entropy in landscape evolution. No. 500-A in US Geol. Survey Prof. Papers. US Government Printing Office, Washington D.C.

Leopold LB, Maddock T (1953) The hydraulic geometry of stream channels and some physiographic implications. No. 252 in US Geol. Survey Prof. Papers. US Government Printing Office, Washington D.C.

Lin B, Taylor PL (1994) Model of spatiotemporal dynamics of stick-slip motion. Phys Rev E 49:3940–3947

Lorenz EN (1963) Deterministic nonperiodic flow. J Atmos Sci 20:130–141

Main IG, Burton PW (1986) Long-term earthquake recurrence constrained by tectonic seismic moment release rate. Bull Seismol Soc Am 76:297–304

Malamud BD, Guzzetti F, Turcotte DL, Reichenbach P (2001) Power-law correlations of Italian landslide areas. Geophys Res Abstracts 3

Malamud BD, Morein G, Turcotte DL (1998) Forest fires: an example of self-organized critical behavior. Science 281:1840–1842

Malamud BD, Turcotte DL (1999) Self-affine time series: I. generation and analyses. Adv Geophys 40:1–90

Malinverno A (1995) Fractal and ocean floor topography. A review and a model. In: Barton C, Pointe PRL (eds.), Fractals in the Earth Sciences, pp. 107–130. Plenum Press, New York

Mandelbrot BB (1967) How long is the coast of Britain? Statistical self-similarity and fractional dimension. Science 156:636–638

Mandelbrot BB (1982) The Fractal Geometry of Nature. W. H. Freeman & Co., San Francisco

Mandelbrot BB (1985) Self-affine fractals and fractal dimension. Physica Scripta 32:257–260

Maritan A, Colaiori F, Flammini A, Cieplak M, Banavar JR (1996a) Universality classes of optimal channel networks. Science 272:984–986

Maritan A, Rinaldo A, Rigon R, Giacometti A, Rodriguez-Iturbe I (1996b) Scaling laws for river networks. Phys Rev E 53:1510–1515

Matheron G (1963) Principles of geostatistics. Econom Geol 58:1246–1266

Matsuzaki M, Takayasu H (1991) Fractal features of the earthquake phenomenon and a simple mechanical model. J Geophys Res 96:19,925–19,931

McCloskey J, Bean CJ (1992) Time and magnitude predictions in shocks due to chaotic fault interactions. Geophys Res Lett 19:119–122

Meybeck M (1995) Global distribution of lakes. In: Lerman A, Imboden DM, Gat JR (eds.), Physics and Chemistry of Lakes, pp. 1–35. Springer, Berlin, Heidelberg, New York

Moßner W, Drossel B, Schwabl F (1992) Computer simulations of the forest-fire model. Physica A 190:205–217

Murray AB, Paola C (1994) A cellular model of braided rivers. Nature 371:54–57

Nakanishi H (1990) Cellular-automaton model of earthquakes with deterministic chaos. Phys Rev A 41:7086–7089

Nakanishi H (1991) Statistical properties of the cellular-automaton model for earthquakes. Phys Rev A 43:6613–6621

Nussbaum J, Ruina A (1987) A two degree-of-freedom earthquake model with static/dynamic friction. Pure Appl Geophys 124:629–656

Ohmori H, Sugai T (1995) Toward geomorphometric models for estimating landslide dynamics and forecast landslide occurrence in Japanese mountains. Z Geomorph N F Suppl. 101:149–164

Olami Z, Christensen K (1992) Temporal correlations, universality, and multifractality in a spring-block model of earthquakes. Phys Rev A 46:1720–1723

Olami Z, Feder HJS, Christensen K (1992) Self-organized criticality in a continuous, nonconservative cellular automation modeling earthquakes. Phys Rev Lett 68:1244–1247

Omori F (1894) On the aftershocks of earthquakes. J Coll Sci Imp Univ Tokyo 7:111–200

Papazachos BC (1975) Foreshocks and earthquake prediction. Tectonophysics 28:213–226

Pape H, Clauser C, Iffland J (1999) Permeability prediction for reservoir sandstones based on fractal pore space geometry. Geophysics 64:1447–1460

Pastor-Satorras R, Vespignani A (2000) Corrections to scaling in the forest-fire model. Phys Rev E 61:4854–4859

Paul G, Hergarten S, Neugebauer HJ (1999) Numerical modelling of surface runoff and infiltration of water. In: Hergarten S, Neugebauer HJ (eds.), Process Modelling and Landform Evolution, vol. 78 of *Lecture Notes in Earth Sciences*, pp. 95–107. Springer, Berlin, Heidelberg, New York

Pelletier JD (2000) Spring-block models of seismicity: review and analysis of a structurally heterogeneous model coupled to a viscous asthenosphere. In: Rundle JB, Turcotte DL, Klein W (eds.), GeoComplexity and the Physics of Earthquakes, vol. 120 of *Geophysical Monograph Series*, pp. 27–42. American Geophysical Union, Washington D.C.

Pelletier JD, Malamud BD, Blodgett T, Turcotte DL (1997) Scale-invariance of soil moisture variability and its implications for the frequency-size distribution of landslides. Engin Geol 49:255–268

Pelletier JD, Turcotte DL (1999) Self-affine time series: II. applications and models. Adv Geophys 40:91–166

Pikovsky A, Rosenblum M, Kurths J (2002) Synchronization: a Universal Concept in Nonlinear Science. Cambridge University Press, Cambridge, New York, Melbourne

Plouraboué F, Kurowski P, Hulin JP, Roux S, Schmittbuhl J (1995) Aperture of rough cracks. Phys Rev E 51:1675–1685

Rayleigh L (1916) On convection currents in a horizontal layer of fluid when the higher temperature is underside. Phil Mag 32:529–546

Rigon R, Rinaldo A, Rodriguez-Iturbe I, Bras RL, Ijjasz-Vasquez E (1993) Optimal channel networks: a framework for the study of river basin morphology. Water Resour Res 29:1635–1646

Rigon R, Rodriguez-Iturbe I, Maritan A, Giacometti A, Tarboton DG, Rinaldo A (1996) On Hack's law. Water Resour Res 32:3367–3374

Rinaldo A, Rodriguez-Iturbe I, Bras RL, Ijjasz-Vasquez E, Marani A (1992) Minimum energy and fractal structures of drainage networks. Water Resour Res 28:2181–2195

Rinaldo A, Rodriguez-Iturbe I, Rigon R (1998) Channel networks. Annu Rev Earth Planet Sci 26:289–327

Rinaldo A, Rodriguez-Iturbe I, Rigon R, Bras RL (1993) Self-organized fractal river networks. Phys Rev Lett 70:822–825

Rodriguez-Iturbe I, Ijjasz-Vasquez E, Bras RL, Tarboton DG (1992a) Power law distribution of mass and energy in river basins. Water Resour Res 28:1089–1093

Rodriguez-Iturbe I, Rinaldo A (1997) Fractal River Basins. Chance and Self-Organization. Cambridge University Press, Cambridge, New York, Melbourne

Rodriguez-Iturbe I, Rinaldo A, Rigon R, Bras RL, Marani A, Ijjasz-Vasquez E (1992b) Energy dissipation, runoff production, and the three-dimensional structure of river basins. Water Resour Res 28:1095–1103

Rodriguez-Iturbe I, Vogel GK, Rigon R, Entekhabi D, Castelli F, Rinaldo A (1995) On the spatial organization of soil moisture fields. Geophy Res Lett 22:2757–2760

Rosin P, Rammler E (1933) Laws governing the fineness of powdered coal. J Inst Fuel 7:29–36

Rundle JB, Jackson DD (1977) Numerical simulation of earthquake sequences. Bull Seismol Soc Am 67:1363–1377

Sammis CG, Nadeau RM, Johnson LR (1999) How strong is an asperity? J Geophys Res 104:10609–10619

Sapozhnikov VB, Foufoula-Georgiou E (1996a) Do the current landscape evolution models show self-organized criticality? Water Resour Res 32:1109–1112

Sapozhnikov VB, Foufoula-Georgiou E (1996b) Self-affinity in braided rivers. Water Resour Res 32:1429–1439

Sapozhnikov VB, Foufoula-Georgiou E (1997) Experimental evidence of dynamic scaling and indications of self-organized criticality in braided rivers. Water Resour Res 33:1983–1991

Scheidegger AE (1967) A stochastic model for drainage patterns into an intramontane trench. Bull Assoc Sci Hydrol 12:15–20

Scholz CH (1998) Earthquakes and friction laws. Nature 391:37–42

Scholz H (1990) The Mechanics of Earthquakes. Cambridge University Press, Cambridge, New York, Melbourne

Schoutens JE (1979) Empirical analysis of nuclear and high-explosive cratering and ejecta. In: Nuclear Geophysics Sourcebook, vol. 55 of *Rep. DNA OIH-4-2*. Def. Nuclear Agency, Bethesda, MD

Sinclair KC, Ball RC (1996) A mechanism for global optimization of river networks from local erosion rules. Phys Rev Lett 76:3359–3363

Sornette A, Sornette D (1989) Self-organized criticality and earthquakes. Europhys Lett 9:197–202

Sornette D (2000) Critical Phenomena in Natural Sciences – Chaos, Fractals, Selforganization and Disorder: Concepts and Tools. Springer, Berlin, Heidelberg, New York

Sornette D, Knopoff L, Kagan YY, Vanneste C (1996) Rank-ordering statistics of extreme events: application to the distribution of large earthquakes. J Geophys Res 101:13883–13893

Sornette D, Vanneste C, Sornette A (1991) Dispersion of b-values in the Gutenberg-Richter law as a consequence of a proposed fractal nature of continental faulting. Geophys Res Lett 18:897–900

Sparrow C (1982) The Lorenz Equations: Bifurcations, Chaos, and Strange Attractors. Springer, Berlin, Heidelberg, New York

Stark CP, Hovius N (2001) The characterization of landslide size distributions. Geophys Res Lett 28:1091–1094

Stetter HJ (1973) Analysis and Discretization Methods for Ordinary Differential Equations. Springer, Berlin, Heidelberg, New York

Strahler AN (1952) Hypsometric (area-altitude) analysis of erosional topography. Bull Geol Soc Am 63:1117–1142

Sugai T, Ohmori H, Hirano M (1994) Rock control on the magnitude-frequency distribution of landslides. Trans Japan Geomorphol Union 15:233–251

Sun T, Meakin P, Jøssang T (1994) The topography of optimal drainage basins. Water Resour Res 30:2599–2610

Sun T, Meakin P, Jøssang T (1995) Minimum energy dissipation river networks with fractal boundaries. Phys Rev E 51:5353–5359

Takayasu H, Inaoka H (1992) New type of self-organized criticality in a model of erosion. Phys Rev Lett 68:966–969

Tebbens S, Burroughs S (2000) Idenfying power laws in upper-truncated cumulative number-size distributions with applications to earthquakes and tsunamis. Geophys Res Abstracts 2

Tebbens S, Burroughs S, Barton CC, Naar DF (2001) Statistical self-similarity of hotspot seamount volumes modeled as self-similar criticality. Geophys Res Lett 28:2711–2714

Tokunaga E (1978) Consideration on the composition of drainage networks and their evolution. Geogr Rep Tokyo Metro Univ 13:1–27

Tokunaga E (1984) Ordering of divide segments and law of divide segment numbers. Jap Geomorph Un 5:71–77

Tucker GE, Bras RL (1998) Hillslope processes, drainage density, and landscape morphology. Water Resour Res 34:2751–2764

Turcotte DL (1986) Fractals and fragmentations. J Geophys Res 91:1921–1926

Turcotte DL (1997) Fractals and Chaos in Geology and Geophysics. Cambridge University Press, Cambridge, New York, Melbourne, 2nd edn.

Utsu T (1961) A statistical study on the occurrence of aftershocks. Geophys Mag 30:521–605

Voss RF (1985) Random fractal forgeries. In: Earnshaw RA (ed.), Fundamental Algorithms in Computer Graphics, pp. 805–835. Springer, Berlin, Heidelberg, New York

Whitehouse IE, Griffiths GA (1983) Frequencey and hazard of large rock avalanches in the central Southern Alps, New Zealand. Geology 11:331–334

Xu HJ, Knopoff L (1994) Periodicity and chaos in a one-dimensional dynamical model of earthquakes. Phys Rev E 50:3577–3581

Yokoi Y, Carr JR, Watters RJ (1995) Fractal character of landslides. Envir Engin Geoscience 1:75–81

Zapperi S, Lauritsen KB, Stanley HE (1995) Self-organized branching processes: mean-field theory for avalanches. Phys Rev Lett 75:4071–4074

Index