Berechnung von Information und Komplexität in Zeitreihen

—

Analyse des Wasserhaushaltes von bewaldeten Einzugsgebieten

Dissertation

zur Erlangung des Doktorgrades
der Fakultät für Biologie, Chemie und Geowissenschaften
der Universität Bayreuth

vorgelegt
von

Frank Wolf

geboren in Paderborn

Februar 1999

Diese Arbeit wurde am Lehrstuhl für Ökologische Modellbildung des Bayreuther Instituts für Terrestrische Ökosystemforschung (BITÖK) der Universität Bayreuth vom 1. März 1996 bis zum 15. Februar 1999 erstellt.

Inhaltsverzeichnis

Zusammenfassung

Bisher wurden und werden in der Ökosystemforschung, wie am BITÖK, vor allem die vielfältigen, innerhalb von Wäldern ablaufenden Prozesse untersucht. In dieser Arbeit werden Zeitreihen zum Wasserhaushalt (Niederschlag, Matrixpotential und Abfluss) bewaldeter Einzugsgebiete ausgewertet, ohne dabei ein solches Prozessverständnis vorauszusetzen. Ziel ist es, dabei alleine die in diesen Daten vorhandene Information für Aussagen über Systemeigenschaften und Vorhersagbarkeit zu nutzen. Die Beschränkung auf den Wasserhaushalt bei diesen Untersuchungen erfolgt aufgrund der Schlüsselfunktion des Wassers als Transportmedium der gelösten Stoffe und aufgrund der guten Datenlage.

Der vorgeschlagene Ansatz erfordert die Quantifizierung von Information und Komplexität in experimentellen Zeitreihen. Die dazu angewendeten Methoden sind neu in diesem Forschungsbereich. Sie werden ausführlich vorgestellt und bezüglich ihrer Anwendbarkeit auf experimentelle Daten untersucht. Dazu wurde die für eine bestimmte mittlere Genauigkeit erforderliche Datenmenge ermittelt. Diese Bedingung sowie weitere Betrachtungen führen zu einem weitgehend parameterfreien Verfahren. Die Methoden wurden für eine universelle Anwendung auf (lückenhafte) Zeitreihen programmiert. Das Programm SYMDYN steht neben einer ausführlichen Anleitung auf dem Internet allgemein zugänglich zur Verfügung.

Prinzipiell wurde eine Informationsabnahme des Wassers beim Durchlaufen der untersuchten Einzugsgebiete vom Niederschlag bis zum Abfluss festgestellt. Die Abfluss-Information nimmt mit zunehmender Regelmäßigkeit des Niederschlags ab und mit verminderter (naturnaher) Bewaldung zu. Sie nähert sich dabei der Information des Niederschlags an. Der Vergleich mit den Rohdaten und der Autokorrelation als klassischer Methode der Zeitreihenanalyse zeigt, dass die Abfluss-Information ein sensibles Maß für die Beurteilung der Bebauung oder forstlicher Eingriffe in die Bewaldung eines Gebietes ist.

Die Information des Matrixpotentials erlaubt im Vergleich mit der Abfluss-Information eine Beurteilung der für den Abfluss relevanten Eindringtiefe des Niederschlagssignals ohne explizite Kenntnis der hydraulischen Bodenbeschaffenheit oder der bevorzugten Fließwege. Ein Maximum an Komplexität kann als Indiz für eine effektive Zeitskala gewertet werden. Dies erlaubt die Bewertung von Messreihen bezüglich ihrer Redundanz und Zufälligkeit und die Beurteilung ihrer Modellierbarkeit und Vorhersagbarkeit. Damit lassen sich Kriterien für zukünftige Messkampagnen formulieren, die eine angemessene räumliche und zeitliche Auflösung der im Mittel relevanten Dynamik der untersuchten Größe garantieren und keine unnötigen Kosten durch eine zu hohe Messfrequenz verursachen. Es wurde festgestellt, dass eine gröber als stündliche Auflösung von Niederschlagsmengen die mittlere Dynamik nicht mehr auflösen kann, während Abflussmengen gegebenenfalls auch seltener als täglich gemessen werden können. Die ermittelten komplexitätsoptimalen Messauflösungen stimmen mit den heuristischen Erfahrungswerten der Monitoringpraxis gut überein.

Die Analyse der Daten stellt keine erschöpfende Betrachtung zur Information im Wasserhaushalt von bewaldeten Einzugsgebieten dar, sondern gibt Hinweise bezüglich der möglichen Aussagen, die eine solche äußere Betrachtungsweise der Systeme erlaubt. Mit dieser Arbeit wurden die methodischen Grundlagen und Voraussetzungen dafür geschaffen, die Vielzahl von hydrologischen Datenreihen aus dem Monitoring von Ökosystemen auf neue Indikatoren von Veränderungen systematisch zu überprüfen. Für dieses praxisrelevante Ziel der Umweltforschung wurde mit SYMDYN ein neues Werkzeug zur allgemeinen Verfügung gestellt.

Summary

The main object of ecosystem research as performed at the BITÖK are so far the various processes related to forests. Here, time series of water related parameters (precipitation, matrix potential, and runoff) of forested catchments are investigated without a priori knowledge of specific processes. The aim was to formulate statements about system properties and predictability purely on the basis of the information inherent in the data. The restriction to water budgets in this initial study was due to their central role as a transport medium of dissolved matter and the quality and quantity of available data.

The proposed concept requires the quantification of information and complexity in experimental time series. The applied methods are new in the field. They are presented in detail and investigated due to their applicability to experimental data. Therefore, the required amounts of data for a given mean accuracy were determined. This requirement and further considerations lead to an almost parameter-free procedure. The methods were programmed for a universal application to time series involving gaps. The program, SYMDYN, and operating instructions can be down-loaded from the internet free of charge by the public.

The percolation of the water through the investigated catchments was accopanied by a general decrease of the hydrologic information from precipitation to runoff. The runoff information decreases when the annual regularity of precipitation increases. It increases and thus contains more of the precipitation information for a declined (natural) forest stand. As was shown by a comparison with the raw data and the autocorrelation as a classical tool of time series analysis, the runoff information is a sensitive measure for judging the build up area or deforestation of a catchment.

The information of the soil matrix potential compared with the runoff information enables an estimation of the relevant percolation depth, of which the dynamics of precipitation is reduced to that of the runoff. This was possible without knowledge about soil structure and preferred flow paths. A maximum of complexity is interpreted as an index for an effective time scale. This allows the estimation of the redundancy and randomness and to judge the modellability and predictability of the measured data. Thus, future measurment campaigns should be planned for a resolution of the mean relevant dynamics of the observed parameter in the data records. Costs can be saved by avoiding non-reasonable high sampling rates. It was concluded that precipitation must be sampled hourly or more often to resolve its mean dynamic. Runoff can sufficiently be measured daily or even coarser. The estimated optimal sampling rates due to a maximum of complexity agree well with the heuristic experience of monitoring practice.

The data analysis of this study does not yield an exhaustive overview of the information inherent in the water budget of forested catchments, but illustrates what can be achieved by such an external viewpoint on the systems. This study provides the methodical requirements for a systematic analysis of the variety of hydrological monitoring data with respect to new indicators of changes. For this relevant aim in environmental research, SYMDYN was offered as a new tool for common usage.

1 Einleitung

Die Entwicklung der menschlichen Kultur ist an die Nutzung von natürlichen Ressourcen gebunden (GOUDIE, 1994). Im Laufe dieses Jahrhunderts wurden bei der Nutzung von Ökosystemen Umweltveränderungen als Folge einer industrialisierten Zivilisation zu einer neuen wichtigen Randbedingung. Die Verschmutzung von Böden und Gewässern durch ferntransportierte Luftverunreinigungen wurde zum Anlass umfangreicher Forschungsbemühungen. In den 80er Jahren wurden diese Forschungen ausgeweitet, als die Wirkungen dieser Umweltveränderungen auf Fischbestände oder in Wäldern zu einer öffentlichen Auseinandersetzung geführt hatten. Die untersuchten Ökosysteme erwiesen sich bald als ein besonders schwieriger (komplexer?) Forschungsgegenstand.

Ein häufig verfolgter Ansatz diese Komplexität zu bewältigen, liegt in der versuchten Entschlüsselung der inneren Vernetzungen zwischen Struktur und Funktion der Bestandteile und Kompartimente von Ökosystemen. Ziel ist es dabei, auf der Basis eines solchen Prozessverständnisses ihr Verhalten und ihre Reaktion auf Störungen vorhersagen zu können (KIMMINS, 1997). Mit dieser Begründung wurden Forschungsinstitute wie das BITÖK eingerichtet (PT BEO & PT UKF, 1997).

Aufgrund des engen Zusammenhangs zwischen den chemischen Stoffumsätzen und dem Wassertransport werden terrestrische Ökosysteme oft in einer biogeochemischen Sichtweise als belebte Wassereinzugsgebiete definiert (LIKENS & BORMANN, 1995). Wasser ist das wichtigste Transportmedium der umgesetzten Stoffe. Die Ökosystemforschung stützt sich daher häufig auf die Beobachtung und Analyse des Stoff- und Wasserhaushaltes von kleinen, häufig bewaldeten Einzugsgebieten. Mittlerweile liegen viele langjährige Messreihen von Niederschlag und Abfluss zu dieser Fragestellung vor (siehe z. B. Kapitel 4). Weitere Messungen aus dem Inneren dieser Einzugsgebiete, z.B. die Messungen des Bodenwasserzustandes als Matrixpotential, geben partiell Aufschluss über die Dynamik der Wasserbewegung in verschiedenen Tiefen. Die Verluste durch Evapotranspiration werden zumindest implizit über die Bilanzierung dichter Einzugsgebiete erfasst.

Der Anlass zu dieser Arbeit ist die Hypothese, dass es neben dem prozessorientierten Ansatz möglich ist, die vorhandenen Daten direkt und möglichst „modellfrei“ zu analysieren. Die Motivation dazu entspringt aus den vielfältigen Problemen, von Beobachtungsdaten zu einer eindeutigen, modellgestützten Interpretation im Hinblick auf interne Prozesse zu gelangen (BEVEN, 1996). Insbesondere lassen sich immer wieder sehr verschiedene Modellanpassungen durchführen, die sich in der Qualität der Datenreproduktion nicht unterscheiden. Ein Grund für diese vermutete Beliebigkeit der Modellierung von hydrologischen Daten auf der Einzugsgebietsebene (JAKEMAN & HORNBERGER, 1993) könnte bereits im Informationsgehalt oder der Redundanz der Beobachtungsdaten liegen.

In dieser Arbeit sollen daher Methoden untersucht werden, die es erlauben, den Informationsgehalt und die Komplexität von typischen hydrologischen Datensätzen möglichst parameterfrei zu quantifizieren. Damit kann eine Auswertung in zwei Schritte aufgeteilt werden: Erstens eine abstrakte Analyse des Datensatzes anhand von Kriterien wie Informationsgehalt oder Autokorrelation in Abhängigkeit von unterschiedlichen zeitlichen Messauflösungen oder Daten-Aggregationen und zweitens eine weitergehende Modellierung und Interpretation, die aber durch die Ergebnisse des ersten Schrittes limitiert sein kann (z.B. in der Zahl der möglichen Parameter eines zu kalibrierenden Modells). Mit anderen Worten: es soll festgestellt

werden, wieviel Information in den hydrologischen Zeitreihen enthalten ist und welche Aussagen sich alleine daraus über die Daten selbst, ihre Modellierbarkeit und Vorhersagbarkeit sowie über den Zustand des Einzugsgebietes ohne ein tieferes Prozessverständnis ableiten lassen. Konkret soll es möglich werden zu prüfen, ob die relevante Dynamik einer Messgröße in den Daten aufgelöst ist und diese demnach zu einer Modellierung geeignet sind. Aus einer hohen Information kann direkt eine schwierige Vorhersagbarkeit gefolgert werden.

Ein weitergehendes Ziel liegt in der Frage, ob es möglich ist, anhand der Beziehungen im Informationsgehalt verschiedener Beobachtungsvariablen direkt von der Systemebene auf den biologischen Zustand des Einzugsgebietes zu schließen. Es ist theoretisch klar, dass die Existenz belebter Ökosysteme auf jeder betrachteten Skala dissipative Stoff- oder Energieumsätze über den Rand des betreffenden Systems notwendig voraussetzt. Umgekehrt bleibt es aber offen, ob auch aus einer Analyse von Stoff- und Energieumsätzen auf die „Belebtheit" des Systems oder auf konkrete biologische Eigenschaften geschlossen werden darf.

Zur Umsetzung dieser Ziele müssen zuerst Methoden untersucht und bereitgestellt werden, die sich zu einer Quantifizierung der Information und Komplexität von Stoff- und Wasserflüssen in Ökosystemen eignen. Dabei muss die begrenzte Datenmenge sowie die Lückenhaftigkeit der tatsächlichen Messreihen berücksichtigt werden. In den folgenden Abschnitten der Einleitung wird erklärt, welche Annahmen für die Quantifizierung von Information und Komplexität möglich oder notwendig sind (Abschnitt 1.1). In Abschnitt 1.2 werden die bisherigen Anwendungen dieses Konzeptes vorgestellt. Anhand der Gliederung dieser Arbeit wird in Abschnitt 1.3 das Programm zur Umsetzung der genannten Ziele erläutert.

1.1 Zum Begriff von Information und Komplexität

Die Geschichte der quantifizierbaren Informationsbegriffe beginnt mit der Begründung der Informationstheorie durch SHANNON (1948) im Kontext von Nachrichtenübertragung. Dabei stand von Anfang an ein wahrscheinlichkeitstheoretischer Zugang für eine diskrete Menge von Symbolen (Buchstaben, Wörter) im Vordergrund. Unter plausiblen Annahmen über das gesuchte Funktional der einzelnen Wahrscheinlichkeiten, die u. a. von KHINCHIN (1957) präzisiert wurden, ergibt sich ein Ausdruck, der formal der thermodynamischen Entropie gleicht und daher auch als Shannon-Entropie bezeichnet wird. Der Informationsbegriff von Shannon bewertet nicht den semantischen Inhalt oder die Bedeutung einer Nachricht, sondern dessen Unvorhersagbarkeit (Unsicherheit, Zufälligkeit, Unkorreliertheit). Eine Nachricht ist umso informationsreicher, je unsicherer wir über ihren syntaktischen Inhalt sind. Shannons wegweisende Arbeiten wurden von BALATONI & RÉNYI (1956), RÉNYI (1960) u. a. mathematisch präzisiert und erweitert und haben die Entwicklung von ähnlichen Maßen initiiert.

Ein wichtiges Entropie-Maß ist der Informationsgewinn, der an einem Beispiel veranschaulicht werden soll: In einem deutschen Text wird die Buchstabenfolge „`Info`" beobachtet. Als nächsten Buchstaben können wir mit Sicherheit ein „`r`" vorhersagen, weil es sich wahrscheinlich um ein von „Information" abgeleitetes Wort handelt. Selbst ohne diese Sprachkenntnis würde eine sture Zählung der Buchstabenhäufigkeiten nach jedem Auftreten von „`Info`" in dem Text ergeben, dass „`r`" ein sicherer Tipp ist. Wir gewinnen durch die Beobachtung des Folgebuchstabens in diesem Fall also kaum Information dazu. Hätten wir die Zeichenfolge „`die_`" beobachtet, könnten wir uns über den Buchstaben nach dem Leerzeichen nicht sicher sein. Jeder Buchstabe wäre möglich. Durch die Kenntnis des neuen Buchstabens gewinnen wir viel Information dazu. EBELING (1997) berechnet solche lokalen Unsicherheiten an

Beispielen aus dem Text „Moby Dick“. Wenn wir für jede 4er-Kombination von Zeichen die Unsicherheit über das 5. Zeichen ermitteln, können wir eine mittlere Unsicherheit über das 5. Zeichen berechnen und erhalten so ein Maß für die Redundanz oder Information des Textes. Die Zahl 4 ist hier willkürlich gewählt. Je länger die Zeichenfolge ist, desto mehr Information ist bereits in ihr selbst enthalten und desto weniger Information können wir durch das nachfolgende Zeichen gewinnen.

KOLMOGOROV (1965) gefiel die Idee nicht, Information auf einer Wahrscheinlichkeitsverteilung zu berechnen und damit etwa einen Text als Realisation eines Zufallsprozesses zu betrachten. Er begründete den algorithmischen Informationsbegriff in der Länge des kürzesten Computerprogramms, das eine Nachricht exakt beschreibt. Theoretische Untersuchungen wurden dazu von CHAITIN (1987, 1990) von 1966 bis heute durchgeführt. Die praktische Realisation dieses für die Datenkomprimierung wichtigen Konzeptes von LEMPEL & ZIV (1976) und KASPAR & SCHUSTER (1987) liefert eine Bewertung von Information, die sich im Kontext der hier untersuchten Datensätze jedoch nicht qualitativ von der durch die Shannon-Entropie gelieferten unterscheidet.

Mit dem Begriff der Information hängt der Begriff der Komplexität zusammen. Dieser wurde in vielen früheren Arbeiten synonym für Information oder Zufälligkeit verwendet. Heute ist damit in intuitiver Weise ein Zustand hoher Struktur zwischen Ordnung und Chaos gemeint. Eine stark geordnete Nachricht enthält wenig Information, ist aber ähnlich wie eine sehr zufällige Nachricht, die viel Information enthält, strukturell einfach. GRASSBERGER (1986) erläutert diesen Zusammenhang anhand von ebenen Mustern. Der Zusammenhang zwischen Chaos und Komplexität erklärt die Entwicklung der Komplexitätstheorie, welche die Informationstheorie umfasst, aus der Chaosforschung (GELL-MANN, 1998). Als ein Zentrum der Komplexitätsforschung wurde Mitte der 80er Jahre das Santa Fe Institut (USA) gegründet.

CRUTCHFIELD (1994b) formuliert das (algorithmische) Idealmaß für Komplexität über die Zustände des kleinsten Automaten aus einer erschöpfenden Hierarchie sich ergänzender Rechenmaschinen, die eine Nachricht (Symbolfolge) stochastisch beschreiben. Die stochastischen Komplexitätsmaße beziehen sich auf die Anordnung oder Darstellung der Information in einer Nachricht. So bezeichnet GRASSBERGER (1986) eine hohe Gesamtmenge an Information, die gespeichert werden muss, um jederzeit eine optimale Vorhersage zu liefern, als komplex. BATES & SHEPARD (1993) bewerten hingegen eine hohe Schwankung in der Differenz von lokaler Informationszunahme und -abgabe als komplex.

Zur Illustration des in dieser Arbeit wichtigen Komplexitätsbegriffs von BATES & SHEPARD (1993) soll das obige Beispiel zum Informationsgewinn noch einmal aufgegriffen werden. Nach Beobachtung der 4 Buchstaben „`Info`“ in einem deutschen Text schließt sich voraussichtlich die Buchstabenfolge „`nfor`“ an. Neben dem Informationsgewinn durch die Beobachtung des „`r`“ wird durch das Vergessen des „`I`“ gleichzeitig Information abgegeben. Die neue Buchstabenfolge könnte auch Bestandteil der Wörter „`konform`“, „`hinfort`“ oder „`Anforderung`“ sein. Durch den Buchstaben „`I`“ geht ein wichtiger Teil der Information zur Spezifikation der Buchstabenfolge verloren. Der Informationsverlust durch das wegfallende „`I`“ ist hier höher als der Informationsgewinn durch das hinzukommende „`r`“. Diese Bilanz des Nettoinformationsgewinns kann für jede Zeichenfolge der gleichen Länge aufgestellt werden. Sie ist im Mittel über den ganzen Text ausgewogen. Nach BATES & SHEPARD (1993) ist ein Text dann komplex, wenn der Nettoinformationsgewinn zwischen den möglichen Zeichenfolgen eines Textes stark schwankt.

Unter dem Begriff „Komplexitätsmaße“ werden Maße für Information und Komplexität zusammengefasst. Gegenwärtig sind mindestens 31 oder sogar 45 unterschiedliche Komplexitätsmaße veröffentlicht (Seth Lloyd nach HORGAN, 1997, S. 316 u. 440f). Diese unterscheiden

sich außer in dem ihnen zugrunde liegenden Komplexitätsbegriff in ihrer Berechenbarkeit und Popularität. Einen aktuellen Überblick geben u. a. WACKERBAUER et al. (1994), BADII & POLITI (1997) und EBELING et al. (1998).

1.2 Praktische Anwendungen von Komplexitätsmaßen

Bevor eine Folge von Anwendungsbeispielen vorgestellt wird, die sich zunehmend von dem Untersuchungsobjekt von SHANNON (1948) entfernen, soll eine methodische, objektunabhängige Anwendung vorgestellt werden, die auch für diese Arbeit von Bedeutung ist: Das Maximum-Entropie-Prinzip (MEP) von JAYNES (1957). Nach dem MEP läßt sich eine unbekannte Wahrscheinlichkeitsverteilung optimal unter alleiniger Ausnutzung von gegebenen unvollständigen Randinformationen konstruieren (siehe auch HONERKAMP, 1994, S. 100ff). Die Verteilung ist demnach optimal, wenn alle Ereignisse möglichst gleichwahrscheinlich sind, d. h. wenn die Auswahl eines Ereignisses gerade nur so wenig eingeschränkt wird, wie dies die Randbedingungen erfordern. Es soll also genau und nur die vorgegebene Information ausgenutzt werden. Das wird erreicht, wenn die Verteilung die Shannon-Entropie maximiert. PRESS et al. (1992, S. 572ff u. 818ff) stellen die Anwendung des MEPs zur Spektren-Schätzung und zur Bildrestaurierung vor. Einen Überblick über Anwendungen des MEPs aus mathematischer Sicht gibt KAPUR (1994).

SHANNON (1948, 1976) beschäftigte sich mit der Messung der Informationsübertragung in Telefonleitungen. Er bestätigte mit seinem Entropie-Maß die Ergebnisse anderer Untersuchungen, dass die englische Sprache aufgrund ihrer Struktur (Grammatik, Semantik) zu etwa 50 % redundant ist. FUCKS (1955) führte eine umfangreichere vergleichende Untersuchung zu Sprachelementen, Sprachstil und Sprachen auf der Basis von Shannons Entropie-Information durch. Dabei spielt die Silbe als kleinstes Element des Sprechflusses eine besondere Rolle. Er stellte unter anderem fest, dass in der englischen Sprache im Vergleich zu insgesamt neun anderen Sprachen am wenigsten Information in der Verteilung der Silben pro Wort liegt, da einsilbige Wörter mit großem Abstand am häufigsten sind. Deutsch liegt dabei auf Platz 2. Russisch, Latein und Türkisch haben eine breitere Silbenverteilung und sind diesbezüglich am informationsreichsten. Möglicherweise liegt hierin eine objektive Begründung zur Wahl des Englischen als Weltsprache.

Die Untersuchungen von EBELING et al. (1996) beschränken sich nicht auf die Silben in einem Text, sondern auf Buchstabenfolgen einer festen Länge wie in den Beispielen aus 1.1. Sie bestätigen anhand von Stufen in der Information Zusammenhänge in Texten, die durch ihre hierarchische Organisation in Wörtern, Sätzen und Kapiteln bedingt sind. Dazu haben sie die Bibel, „Grimms Märchen“, „Moby Dick“ und „Die Brüder Karamazov“ untersucht.

Die Anwendung von Komplexitätsmaßen ist nicht auf Nachrichten oder Texte der menschlichen Sprache beschränkt. Auf die gleiche Weise wie Texte läßt sich auch der Quell- oder binäre Code einer Computersprache analysieren. Wegen seiner strengen Syntax zeichnet sich beispielsweise Fortran-Quellcode nach SCHMITT & HERZEL (1997) durch eine geringe Information aus. Auch die Notenfolge eines Musikstückes kann als Zeichenfolge kodiert und bezüglich ihres Informationsgehaltes untersucht werden. Dieser ist ungefähr mit dem von Texten vergleichbar (SCHMITT & HERZEL, 1997) und kann zu einer Klassifizierung etwa nach Komponisten verwendet werden (EBELING et al., 1995). Informationsanalysen von Sprachen, Texten, Autoren und Musik finden sich auch in dem Buch von EBELING et al. (1998).

DNA- und Protein-Sequenzen sind Beispiele für biologische Codes, die ebenfalls aus elementaren Bausteinen aufgebaut sind. Im Fall von DNA sind dies die vier Nukleotide Adenin, Cytosin, Guanin und Thymin, so dass DNA-Stränge als eine Zeichenfolge der Buchstaben A, C, G und T aufgeschrieben werden können. Protein-Sequenzen sind aus 20 Aminosäuren kombiniert. Diese Biosequenzen können ebenfalls auf ihren Informationsgehalt untersucht werden. Es besteht die Hoffnung, dass so kodierende von nicht-kodierenden Abschnitten auf einem DNA-Strang aufgespürt werden können. Übereinstimmend konnte bisher festgestellt werden, dass sich DNA durch einen sehr hohen Informationsgehalt, nahe dem von Zufallsfolgen, auszeichnet (SCHMITT et al., 1993; HERZEL et al., 1994: Hefe Chromosom III; EBELING, 1997: HIV-Virus; SCHMITT & HERZEL, 1997: Epstein-Barr-Virus). Trotzdem beobachteten u.a. HERZEL et al. (1994) langreichweitige Korrelationen.

Die Berechnung von Information und Komplexität in einer Zeitreihe mit quasi-kontinuierlichem Wertespektrum erfordert zuerst eine Transformation derselben in eine Zeichenfolge aus einem festen Alphabet von Symbolen. Eine solche Zeichenfolge ist allgemein die Voraussetzung zur Berechnung von fast allen Komplexitätsmaßen. Die Theorie von unendlich langen symbolischen Sequenzen, von sogenannten symbolischen dynamischen Systemen, ist Gegenstand der „Symbolischen Dynamik“, die von LIND & MARCUS (1995) beschrieben wird. Die Anwendung von Komplexitätsmaßen in der Signal- und Datenanalyse befindet sich erst in einer Anfangsphase. Es gibt noch viele offene Probleme bezüglich der theoretischen Grundlagen und der Behandlung experimenteller Daten (KURTHS et al., 1996). Nachfolgend werden einige Anwendungsbeispiele zur Zeitreihenanalyse mit diesen Methoden aufgezählt:

- WITT et al. (1994) konnten anhand von Informationsmaßen die dynamische Unzulänglichkeit eines Modells über die Magnetfeldumpolungen der Erde zeigen, welches mit statistischen Standardverfahren sowie mit Informationsmaßen strukturell mit den Messungen übereinstimmte (auch beschrieben bei KURTHS & WITT, 1994). Dieses Beispiel zeigt, dass Komplexitätsmaße ein wichtiges Werkzeug zur Validierung von Modellen sein können, das über bisherige Verfahren hinausgeht.
- Die Erkennung von Hochrisikopatienten für einen Herzinfarkt ist zur Zeit noch schwierig. Anhand von EKG-Messungen, die mit Informationsmaßen untersucht wurden, ließ sich in einer ersten Testphase die Unterscheidung von Gesunden und Risikopatienten deutlich gegenüber den linearen statistischen Standardverfahren verbessern (KURTHS et al., 1995 u. 1996). Nach mündlicher Auskunft von Jürgen Kurths wird derzeit die Brauchbarkeit dieser Methode in einer medizinischen Großstudie weiter untersucht.
- Weitere Anwendungen sind etwa die Untersuchung der Organisation von Aktivitäten auf der Sonnenoberfläche (SCHWARZ et al., 1993) oder die Untersuchung der menschlichen Motorik bei der Organisation und Synchronisation schneller Bewegungen (KURTHS et al., 1996).

1.3 Gliederung der Arbeit

Die Anwendung von Komplexitätsmaßen in der Ökosystemforschung ist neu. Erste Untersuchungen in dieser Richtung wurden im Rahmen einer Diplomarbeit von ROMAHN (1996) durchgeführt, deren Ergebnisse in LANGE et al. (1997) veröffentlicht sind. Neben methodischen Betrachtungen wurde dabei eine prinzipielle Informationsabnahme (Metrische Entropie) des Wassers mit zunehmender Bodentiefe für ein Einzugsgebiet (Lange Bramke) festgestellt.

In dieser Arbeit wird zunächst eine Auswahl von verschiedenen klassischen und neuen Methoden zur Quantifizierung von Information und Komplexität vorgestellt und bezüglich ihrer Bewertung von dynamischen Verhalten diskutiert (Kapitel 2). Es wurde ein Computerprogramm (SYMDYN, von Symbolischer Dynamik) geschrieben, das die Berechnung dieser Methoden in universeller Weise auch für lückenhafte Zeitreihen erlaubt. Das Programm ist ein wesentlicher Teil dieser Arbeit und wird in Abschnitt 2.7 kurz beschrieben. Es ist neben einer ausführlichen Anleitung und einem Anwendungsbeispiel über das Internet zu beziehen (siehe 2.7).

Grundsätzliche Fragen zur Anwendbarkeit der Methoden werden in Kapitel 3 behandelt. Dabei werden die Äquidistanz, Ergodizität, Stationarität und Messlücken der Daten und deren Einfluss auf die Berechnung von Komplexitätsmaßen sowie Schwankungen der Ergebnisse diskutiert. Aufgrund der oft nur für theoretisch unendlich lange Symbolfolgen definierten Methoden und des wesentlichen Einflusses der Datenmenge auf die Ergebnisse werden in Abschnitt 3.6 obere Grenzen für die erforderliche Datenmenge zur Einhaltung einer bestimmten Genauigkeit analytisch und numerisch hergeleitet. Einen weiteren Schwerpunkt bilden in Abschnitt 3.8 Untersuchungen zum Einfluss und zur Fixierung der Transformation von (hydrologischen) Zeitreihen in eine Symbolfolge, die einen wesentlichen Parameter der Methode darstellt und das Ergebnis entscheidend beeinflusst.

Das in dieser Arbeit verwendete Datenmaterial aus unterschiedlichen Einzugsgebieten wird in Kapitel 4 vorgestellt. Dabei werden Standorteigenschaften und -unterschiede beschrieben, die für die Bewaldung und den Wasserhaushalt von Bedeutung sind. Änderungen im Baumbestand durch Durchforstungsmaßnahmen werden ebenfalls genannt, insoweit die Daten auf einen entsprechenden Einfluss untersucht wurden.

In Kapitel 5 werden die hydrologischen Zeitreihen bezüglich der eingangs formulierten Ziele analysiert. Zunächst (5.1) wird dazu die Information des Wassers an den verschiedenen Orten im Einzugsgebiet allgemein quantifiziert und in Beziehung zu den verschiedenen Messstellen gesetzt. Anschließend (5.2) wird der Einfluss der Bewaldung auf die Abflussdynamik in einem gut instrumentierten Gebiet anhand von acht Teileinzugsgebieten untersucht. Dabei werden die Möglichkeiten der Komplexitätsmaße mit denen der Autokorrelation als einem Standardverfahren der Zeitreihenanalyse sowie mit denen der Wiederkehranalyse als einem alternativen Verfahren verglichen. Abschnitt 5.3 behandelt die Bestimmung einer effektiven Zeitauflösung von Messdaten anhand einer Maximierung der Komplexität. Abschließend (5.4) wird eine Klassifizierung aller hier vorliegenden hydrologischen Zeitreihen anhand ihrer Dynamik aus Sicht der Information und Komplexität bezüglich ihres Typs (Niederschlag und Abfluss) und ihrer Nutzung oder Bewaldung vorgenommen.

Im Kapitel 6 werden die Ergebnisse dieser Arbeit kurz reflektiert und diskutiert. Dabei wird ein Ausblick auf mögliche Anwendungen gegeben. Analytische Rechnungen und große Tabellen sind zur leichteren Lesbarkeit des Haupttextes in einen Anhang (7) ausgelagert. Häufig verwendete Symbole werden in einer Tabelle (8) mit einem Verweis auf ihre Definition aufgeführt.

Der Fortschritt dieser Arbeit während ihrer Förderphase durch das BMBF ist in den Forschungsberichten des BITÖK (WOLF et al., 1997 u. 1998) dokumentiert. Weitere Teilergebnisse sind als Tagungsbeitrag (LANGE et al., 1998) veröffentlicht.

Diese Arbeit orientiert sich an den Regeln der neuen Rechtschreibung nach DUDEN (1996). Die technische Umsetzung des Manuskriptes erfolgte mit MS Word 97 und der Anleitung von WISEMAN (1997). Die Grafiken wurden mit Word-Diagramm und Word-Grafik erstellt; die Diagramme mit Gnuplot im Postscript-Format. Alle Gleichungen im Text werden nummeriert

und mit dieser Nummer in runden Klammern () zitiert. Alle Internet-Adressen sind auf dem Stand von Januar 1999.

2 Methoden

In diesem Kapitel werden die Methoden beschrieben, mit denen die Daten — vor allem zum Wasserhaushalt von Ökosystemen — untersucht werden sollen. Zunächst werden die auf den Originaldaten aufbauenden Datenstrukturen beschrieben, auf denen die Komplexitätsmaße definiert sind. Anschließend werden Maße für Korrelation vorgestellt. Dann werden etablierte Referenzprozesse aufgeführt, die zur Einschätzung der Bewertung von Dynamik durch die Komplexitätsmaße dienen. Nach einer generellen Klassifizierung werden die hier betrachteten Maße für Information und Komplexität im Detail beschrieben. Neben der mathematischen Definition der jeweiligen Methode wird der zugehörige Komplexitätsbegriff erklärt. Die Bewertung von Zufälligkeit und Dynamik wird anhand der Referenzprozesse exemplarisch demonstriert. Ein genereller Überblick über jede Art von Dynamik ist unmöglich. Vor allem ist die den Daten zugrunde liegende Dynamik unbekannt. Ein Vergleich der Komplexitätsmaße untereinander bezüglich ihrer Bewertung von Dynamik ist nur sinnvoll, wenn diese bekannt ist. Dies ist bei den Referenzprozessen der Fall.

2.1 Hierarchie von Datenstrukturen

Standardmethoden zur Zeitreihenanalyse werden direkt mit den Messwerten berechnet. Die Methoden der Symbolischen Dynamik sind jedoch auf Symbolsequenzen definiert, die diskret in Raum und Zeit sind. Digitale Messdaten sind grundsätzlich räumlich und zeitlich diskret. Hier ist jedoch eine viel stärkere Vergröberung gemeint. Diese kann z. B. dadurch erreicht werden, dass der Wertebereich einer Messung in disjunkte Teilintervalle aufgeteilt wird, denen eine Indizierung zugeordnet wird. Die Messdatenreihe kann dann in einen Symbolsatz transformiert werden, indem jedem Messwert der Index, die Nummer oder das Symbol des Teilintervalls zugeordnet wird, in dem er sich befindet. Dies führt zu einem symbolisch dynamischen System und ist das Thema der Symbolischen Dynamik (LIND & MARCUS, 1995, S. xi). Die Transformation eines Datensatzes in eine Symbolfolge durch Partitionierung wird im nachfolgenden Abschnitt beschrieben.

Komplexitätsmaße werden auch auf anderen Datenstrukturen berechnet, die auf dem Symbolsatz aufbauen. Diese sind die Ebenen der folgenden Hierarchie und werden im Weiteren beschrieben:

Datensatz
↓Partitionierung↓
Symbolsatz
↓Wörter↓
Analysebaum
↓Morphs↓
Endlicher Automat

2.1.1 Partitionierung und Symbolsatz

Gegeben sei eine äquidistante Messreihe

$$X = \{x_0, x_1, x_2, \ldots, x_{N-1} \mid x_i \in [a,b], i = 0,1,\ldots N-1\} \quad (1)$$

von N Messwerten x_i innerhalb des Messbereichs $[a,b]$. Die Zeitpunkte der Messungen sind wegen der Äquidistanz und zur Vereinfachung lediglich durch einen Index dargestellt – angefangen mit 0 für den Zeitpunkt der ersten Messung. Für die nachfolgenden Betrachtungen ist der konkrete Zeitabstand zweier Messungen unerheblich. Falls zu jedem Messzeitpunkt mehrere Messgrößen erhoben werden, kann Gleichung (1) auch vektoriell gelesen werden. Hier soll jedoch nur eine Messgröße betrachtet werden, da sich auch alle nachfolgenden Betrachungen auf zunächst eine eindimensionale Datenreihe beziehen.

Mit einer geeigneten Vorschrift zur Partitionierung (s. u.) wird jedem Datenpunkt $x_i \in X$ eindeutig ein Symbol $s_i \in$ A aus einem gegebenen Alphabet

$$A = \{a_0, a_1, \ldots, a_{\lambda-1}\} \quad (2)$$

der Größe $\lambda \geq 2$ zugeordnet. Dadurch entsteht eine Symbolfolge

$$S = \{s_0, s_1, s_2, \ldots, s_{N-1} \mid s_i \in A, i = 0,1,\ldots N-1\}. \quad (3)$$

Der Symbolsatz ist die Repräsentation von Zeitreihen (Nicht-Symbolisch) Dynamischer Systeme in der Symbolischen Dynamik und Basis aller Methoden der Symbolischen Dynamik (vgl. Lind/Marcus, 1995, S. 201).

Übliche Alphabetgrößen sind $\lambda = 2$ (BADII & FINARDI, 1992; CRUTCHFIELD, 1992, 1994a, 1994b; CRUTCHFIELD & PACKARD, 1983; CRUTCHFIELD & YOUNG, 1989; EBELING, 1996; EBELING et al., 1996; FELDMAN & CRUTCHFIELD, 1998; GRASSBERGER, 1986; GRASSBERGER, 1988, KASPAR & SCHUSTER, 1987; KURTHS & WITT, 1994; KURTHS et al., 1996; LANGE et al., 1997; LANGE et al., 1998; LI, 1990; RATEITSCHAK et al., 1995; ROMAHN, 1996; WITT et al., 1994), $\lambda = 3$ (BATES & SHEPARD, 1993; RATEITSCHAK et al., 1995; ROMAHN, 1996), $\lambda = 4$ (EBELING, 1996; HERZEL et al., 1994; KURTHS & WITT, 1994; KURTHS et al., 1996; WITT et al., 1994) oder $\lambda = 32$ für lange Texte und Musikstücke (EBELING, 1996; EBELING & NEIMAN, 1995; EBELING et al., 1995; EBELING et al., 1996). Der Alphabetgröße sind – wie in Abschnitt 3.6 detailliert beschrieben – durch die Anzahl der Datenpunkte strenge Grenzen gesetzt. Dies ist ein Grund für die Popularität von binären Alphabeten. Für die logistische Abbildung (24) ist nach CRUTCHFIELD & PACKARD (1983) ein binäres Alphabet bereits ausreichend für eine generierende Partitionierung (siehe 2.1.1.1).

Falls die Messreihe lückenhaft ist oder Messwerte nicht vertrauenswürdig sind, wird dies in der Zeitreihe durch das Fehlen entsprechender Datenpunkte oder durch einen Wert außerhalb des Messbereichs markiert. Im Symbolsatz wird in solchen Fällen für jeden betreffenden Zeitpunkt ein besonderes Lückensymbol $a^* \notin A$ eingefügt, so dass der Symbolsatz äquidistant in der Zeit ist.

Mit einer Partitionierung ist prinzipiell eine Unterteilung des Zustandsraumes (Wertebereichs) eines Systems gemeint (LIND & MARCUS, 1995, S. 202; WACKERBAUER et al., 1994). Eine erste derartige Unterteilung erfolgt bereits durch die Auflösung des Messgerätes (CRUTCHFIELD, 1992). Neben dieser „statischen Partitionierung“ unterscheiden KURTHS et al. (1996) noch die praxisrelevante „dynamische Partitionierung“.

2.1.1.1 Statische Partitionierung

Bei der statischen Partitionierung wird der Wertebereich $[a,b]$ einer Zeitreihe in λ disjunkte Teilintervalle unterteilt. Dies erfordert $\lambda-1$ Parameter

$$\pi_j,\ j = 0, 1, ..., \lambda-2 \ \text{ mit } \ a < \pi_0 < \pi_1 < ... < \pi_{\lambda-2} < b, \tag{4}$$

welche die Grenzen der Teilintervalle bestimmen. Diese wurden hier konventionell – bis auf das rechte Teilintervall – nach oben halboffen gewählt, so dass gilt:

$$[a,b] = [a, \pi_0) \,(\, [\pi_0, \pi_1) \,(\, ... \,(\, [\pi_{\lambda-2}, b] \tag{5}$$

Den Teilintervallen (Zellen) werden die Symbole des Alphabets (2) zugeordnet. Üblicherweise werden die Zellen durchnummeriert, wobei die Nummern dann den Symbolen entsprechen. Jedem Datenpunkt der Zeitreihe (1) wird schließlich das Symbol der Zelle zugeordnet, die ihn enthält. Z. B. für eine binäre Partitionierung, $\lambda = 2$, $A = \{0,1\}$, wird jedem Datenpunkt x_i, $i = 0, 1, ..., n-1$ ein Symbol

$$s_i = \begin{cases} 0 & \text{falls } x_i < \pi_0 \\ 1 & \text{falls } x_i \geq \pi_0 \end{cases} \tag{6}$$

zugeordnet (siehe Abb. 2-1).

In der Theorie der Symbolischen Dynamik ist definiert, wie eine Partitionierung beschaffen sein muss, damit sie eine „symbolische Repräsentation“ eines Dynamischen Systems liefert (LIND & MARCUS, 1995, S. 203): Hierzu muss neben dem kompakten metrischen Zustandsraum des Dynamischen Systems auch die invertierbare Funktion bekannt sein, die einen Zustand in den nächsten überführt. Damit kann für jede Partitionierungszelle (Intervall), die einem Punkt x_i einer Trajektorie (x_0, x_1, x_2, ...) zugeordnet ist, die Urbildmenge, d. h. die Menge der möglichen Startwerte x_0, ermittelt werden. Wenn der abgeschlossene Durchschnitt

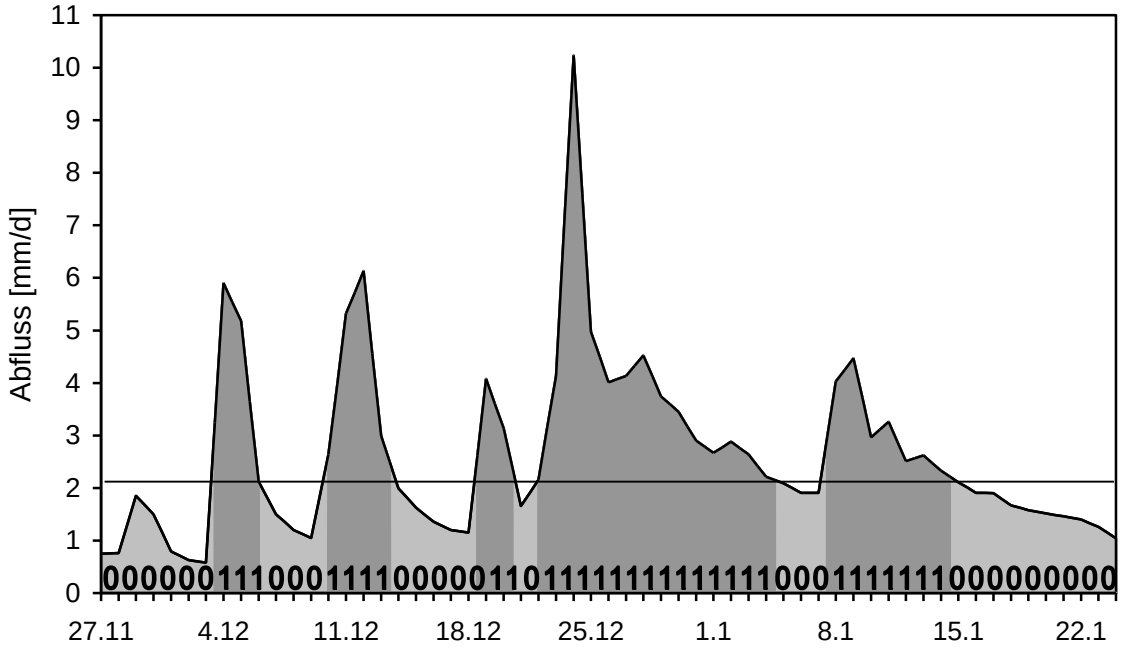

Abb. 2-1. Statische binäre Partitionierung. Transformation einer Datenreihe anhand eines Schwellenwertes in eine binäre Symbolfolge. Exemplarisch für den täglichen Gebietsabfluss des Lehstenbaches (Fichtelgebirge) vom 27.11.1988 – 24.01.1989.

der Urbilder aller Partitionierungszellen einer Trajektorie aus genau einem Punkt (Startwert) besteht und dies für jede unendlich lange Trajektorie des Systems gilt, dann liefert die Partitionierung eine symbolische Repräsentation des Dynamischen Systems. Eine solche Partitionierung wird im Fall beidseitig unendlicher Trajektorien „Markov Partitionierung“ genannt. Für Details sei auf die Quelle verwiesen. Eine weniger mathematische Darstellung zur symbolischen Repräsentation physikalischer Systeme findet sich bei BADII & POLITI (1997, S. 69ff) und EBELING et al. (1998, S. 124ff).

In der Praxis liegt oft nur eine einmalige kurze Zeitreihe als Realisation eines Prozesses mit unbekannten Bewegungsgleichungen vor. Daher läßt sich die Definition einer Symbolischen Repräsentation nicht direkt überprüfen. CRUTCHFIELD & PACKARD (1983) gehen von einer Arbeit Kolmogorovs aus und fordern eine „generierende“ Partitionierung als symbolische Repräsentation. Diese Bedingung ist erfüllt, wenn alle hinreichend langen Symbolfolgen individuelle Punkte bezeichnen, was an die Definition von LIND & MARCUS (1995) erinnert. Die Parameter einer solchen Partitionierung maximieren nach CRUTCHFIELD & PACKARD (1983) die Kolmogorov-Sinai-Entropie (siehe 2.5.3), die bei ihnen „Metrische Entropie“ genannt wird, was den Bezeichnungen in dieser und anderen Arbeiten widerspricht. Damit ist ein praktisch leicht überprüfbares Kriterium gegeben, das der Intuition gerecht wird, den Informationsverlust durch die starke Vergröberung bei der Transformation einer Original-Zeitreihe in eine Symbolfolge zu minimieren. Die Kolmogorov-Sinai-Entropie ist allerdings ein Grenzwert, der praktisch nur durch die Metrische Entropie oder den Informationsgewinn approximiert werden kann (siehe 2.5.3 und 2.5.4).

Nach CRUTCHFIELD & PACKARD (1983) genügt (für die logistische Abbildung) eine binäre Partitionierung, deren Parameter die Kolmogorov-Sinai-Entropie maximiert. Eine Verfeinerung der Partitionierung kann dann durch die Betrachtung von (kurzen) Teilfolgen einer bestimmten Länge — sogenannter Blöcke oder Wörter (siehe 2.1.2) — erreicht werden (siehe auch BADII & POLITI, 1997, S. 70ff; EBELING et al. 1998, S. 131). Die Verteilung dieser Wörter ist optimal, wenn jedes Wort möglichst gleichwahrscheinlich ist, d. h. wenn die Auswahl eines Wortes bei einer gegebenen Zeitreihe und bei gegebener Wortlänge gerade nur so wenig eingeschränkt wird, wie dies der Struktur der Daten entspricht. Es soll also genau und nur die in den Daten vorhandene Information ausgenutzt werden. Das wird erreicht, wenn die Verteilung (der Wörter) die Shannon-Entropie (siehe 2.5.1) maximiert und ist das Maximum-Entropie-Prinzip von JAYNES (1957). Dies bedeutet die gleiche Wahl des Partitionierungsparameters, wie von CRUTCHFIELD & PACKARD (1983) vorgeschlagen wurde, weil die Shannon-Entropie in ihrer normierten Form als Metrische Entropie im Shannon’schen Sinn die Kolmogorov-Sinai-Entropie — die Metrische Entropie von CRUTCHFIELD & PACKARD (1983) — approximiert (siehe 2.5.3). Die Arbeit von CRUTCHFIELD & PACKARD (1983) basiert nicht auf JAYNES (1957) — auch wenn sie den Autoren sicher bekannt war — und gelangt auf anderem Wege zum selben Resultat für diese Situation.

Auf der Ebene des Symbolsatzes — also bei Wortlänge 1 — wird die Shannon-Entropie bei Gleichhäufigkeit der Symbole des Alphabets maximal. Der Partitionierungsparameter muss dafür als Median der Werteverteilung der Original-Zeitreihe bei binärem Alphabet gewählt werden. Im allgemeinen Fall von (größeren) Alphabeten werden die Partitionierungsparameter als Grenzen der entsprechenden Quantile gewählt. Dies gilt für höhere Wortlängen nicht mehr (siehe 3.8). Dennoch werden in dieser Arbeit neben Entropie-maximalen auch äquiquantile Partitionierungen wegen ihrer leichteren Handhabbarkeit und schnelleren Berechenbarkeit verwendet. Der Wert eines Komplexitätsmaßes hängt entscheidend von der gewählten Partitionierung ab. Daher sollten für aussagekräftige und konsistente Werte verschiedenen Partitionierungen getestet werden.

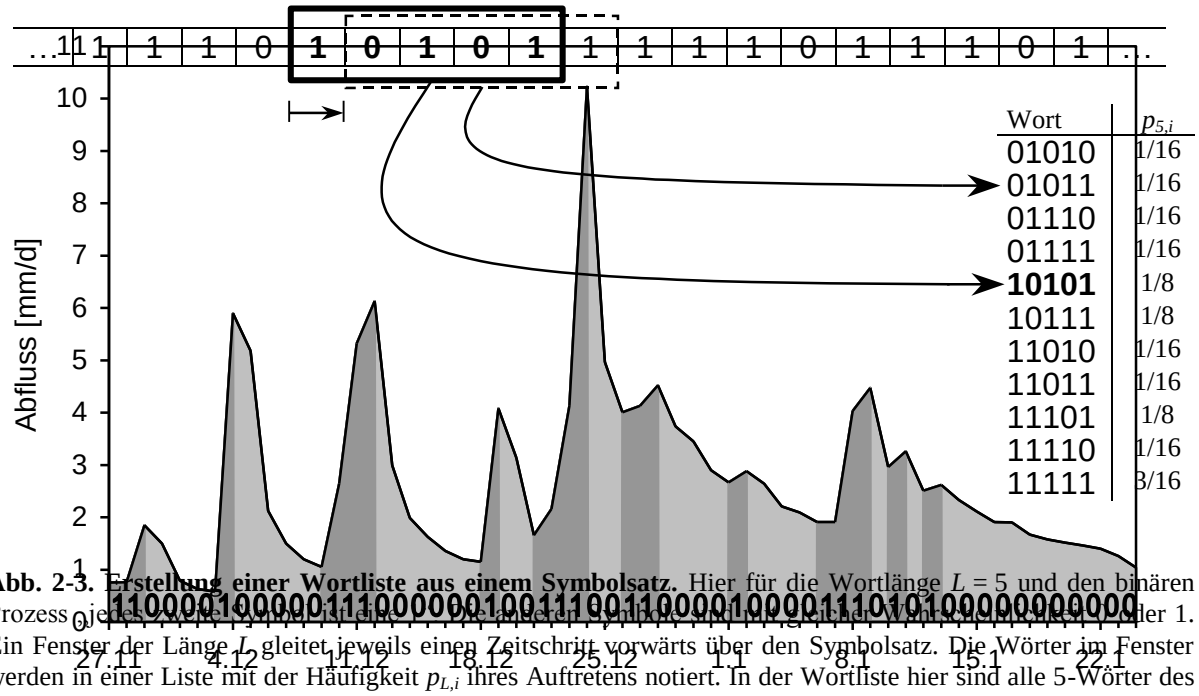

Abb. 2-3. Erstellung einer Wortliste aus einem Symbolsatz. Hier für die Wortlänge $L = 5$ und den binären Prozess jedes zweite Symbol ist eine 1. Die anderen Symbole sind zufällig Nullen oder 1. Ein Fenster der Länge L gleitet jeweils einen Zeitschritt vorwärts über den Symbolsatz. Die Wörter im Fenster werden in einer Liste mit der Häufigkeit $p_{L,i}$ ihres Auftretens notiert. In der Wortliste hier sind alle 5-Wörter des Beispielprozesses mit ihren theoretischen Wahrscheinlichkeiten eingetragen.

Abb. 2-2. Dynamische binäre Partitionierung. Transformation einer Datenreihe anhand der Steigung der Werte in eine binäre Symbolfolge. Exemplarisch für den täglichen Gebietsabfluss des Lehstenbaches (Fichtelgebirge) vom 27.11.1988 – 24.01.1989.

2.1.1.2 Dynamische Partitionierung

Die Konstruktion einer dynamischen Partitionierung kann auf die einer statischen Partitionierung zurückgeführt werden, wenn man diese nicht auf den Datensatz selbst, sondern auf die Folge der Differenzen $d_i \square x_{i+1} - x_i$ für i = 0, 1, ..., N–2 anwendet. Hierbei verringert sich die Anzahl der Datenpunkte im Symbolsatz um eins. Es handelt sich also eher um die (statische) Partitionierung der Differenzenfolge oder der ersten Ableitung als um einen wirklich neuen Partitionierungstyp.

Bei einer binären Partitionierung mit dem Alphabet $A = \{0,1\}$ kann beispielsweise zwischen Anstieg und Abstieg der Daten unterschieden werden. Dabei wird jedem Datenpunkt x_i, $i = 0$, 1, ..., N–2 ein Symbol

$$s_i = \begin{cases} 0 & \text{falls } x_i > x_{i+1} \\ 1 & \text{falls } x_i \le x_{i+1} \end{cases} \tag{7}$$

zugeordnet (siehe Abb. 2-2).

In einigen Fällen kann durch die Partitionierung der Differenzenfolge, die Struktur oder Dynamik einer Zeitreihe besser repräsentiert werden als durch eine statische Partitionierung. So konnten WITT et al. (1994) zeigen, dass ein Modell über das Auftreten von Magnetfeldumpolungen der Erde zwar die Struktur aber nicht die Dynamik der Beobachtungen beschreibt (siehe auch: KURTHS & WITT, 1994).

2.1.2 Wörter und Bäume

Zur Quantifizierung der Information oder Komplexität eines Symbolsatzes werden häufig Teilfolgen (Wörter, Blöcke) einer festen Länge L des Symbolsatzes betrachtet. Dabei interessieren die Verteilung dieser L-Wörter oder auch Übergänge zwischen den Wörtern. WACKERBAUER et al. (1994) klassifizieren Komplexitätsmaße, die auf den relativen Häufigkeiten $p_{L,i}$ der Wörter definiert sind, als „Strukturelle Maße“ im Gegensatz zu den „Dynamischen Maßen“, bei denen die Übergangswahrscheinlichkeiten zwischen Wörtern $p_{L,i\to j}$ oder Zellen einer Partitionierung (zusätzlich) betrachtet werden.

Bei bekannten Prozessgleichungen ist die Wahrscheinlichkeit des Auftretens eines Wortes im Prinzip bekannt. Bei realen Messreihen kann diese höchstens durch die relativen Häufigkeiten oder einen anderen Schätzer approximiert werden. Beide Größen werden mit dem Symbol p bezeichnet, da aus dem Zusammenhang klar ist, ob es sich um relative Häufigkeiten oder Wahrscheinlichkeiten handelt. Wenn die Wortlänge L bekannt ist, kann der Index L auch weggelassen werden: Also p_i statt $p_{L,i}$ und $p_{i\to j}$ statt $p_{L,i\to j}$. Für $L = 1$ entsprechen die Wörter den Symbolen des Alphabets A. Abb. 2-3 veranschaulicht die Ermittlung der L-Wörter aus einem Symbolsatz.

Wörter, die ein Lückensymbol enthalten, werden nicht in der Wortliste notiert. Bei Maßen und Datenstrukturen, die auf den Wörtern aufbauen, ist das Problem fehler- oder lückenhafter Daten — falls zuvor markiert — also auf elegante Weise gelöst. Falls der Datensatz viele (isolierte) Lücken enthält, kann bei großen Wortlängen die Anzahl verwertbarer Wörter klein werden. In solchen Fällen ist generell zu erwägen, ob die Auswertung derartiger Daten sinnvoll ist, und ob die fehlenden oder fehlerhaften Werte sinnvoll ersetzt werden können.

Die Wörter eines Symbolsatzes werden in einer Baumstruktur gespeichert und verwaltet. Bäume gehören zu den fundamentalen Datenstrukturen in der Informatik und werden in der Graphentheorie definiert. Unterschiedliche Zugänge dazu finden sich u. a. bei BRONŠTEJN et al. (1997, S. 317), HOROWITZ & SAHNI (1981, S. 68), und SEDGEWICK (1988, S. 36).

Bäume bestehen aus Knoten, die miteinander verbunden sind (siehe Abb. 2-4). Die Knoten sind in L+1 Schichten hierarchisch angeordnet. Die Wortlänge L wird auch als Baumtiefe bezeichnet. In der obersten Schicht gibt es nur einen (Wurzel-) Knoten. Jeder Knoten, ausgenommen die (Blätter-) Knoten in der untersten Schicht, besitzt maximal λ Unterknoten. Die Unterknoten werden auch Kinder genannt. Ihr Ursprungsknoten wird dann Elternknoten genannt. Die Verzweigungen zu den Knoten entsprechen den Symbolen der Wörter. Die

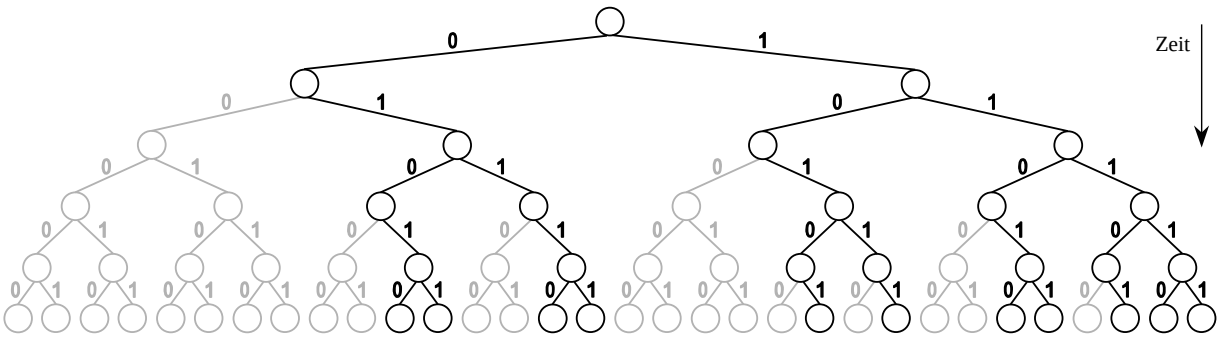

Abb. 2-4. Darstellung der Wortliste aus Abb. 2-3 in einer Baumstruktur. Von der Wurzel (oberster Knoten) verzweigt der Baum nach rechts, wenn das erste Symbol eines Wortes eine 1 ist, und nach links, wenn das erste Symbol eine 0 ist. Die nachfolgenden Symbole führen zu entsprechenden Verzweigungen. Da jedes weitere Zeichen einen weiteren Zeitschritt bedeutet, verläuft die Zeit von der Wurzel zu den Blättern (untere Knotenschicht). Die grauen Verzweigungen kommen in dem Beispielprozess nicht vor. Sie ergeben zusammen mit den anderen Verzweigungen einen voll besetzten Baum, wie z. B. bei weißem Rauschen.

Verzweigungen von der Wurzel zu den Knoten der ersten Schicht entsprechen dem ersten Symbol der Wörter. Analog entsprechen die Verbindungen der Knoten von der (*i*–1)-ten zur *i*-ten Schicht den *i*-ten Symbolen der *L*-Wörter für i = 1 ... *L*.

In den Knoten wird unter anderem notiert, welche von den λ möglichen Kindern existieren, und die Anzahl $n_{k,i}$ (k = 0 ... L Knotenschicht, i = 0 ... λ^k–1 Nummer des Knotens in der k-ten Schicht) der Wörter, die den Knoten durchlaufen. Die Wurzel wird von allen Wörtern durchlaufen: $n_{0,0} = N\text{-}L\text{+}1$. Anhand der $n_{k,i}$ werden die relativen Häufigkeiten der *L*-Wörter

$$p_{L,i} = \frac{n_{L,i}}{n_{0,0}} \tag{8}$$

und der Übergänge zwischen zwei *L*-Wörtern *i* und *j*

$$p_{L,i\to j} = \frac{n_{L+1,j}}{n_{L,i}} \tag{9}$$

bestimmt. Die letzte Berechnung setzt einen Baum der Tiefe *L*+1 voraus, in dem der Knoten *j* der Blattschicht *L*+1 ein Kind des Knotens *i* der *L*-ten Schicht ist. Sonst ist $p_{L,i\to j} = n_{L+1,j} = 0$, d. h. es gibt keinen Übergang von *L*-Wort *i* nach *j*. Für die bedingten Wahrscheinlichkeiten gilt, wie aus der Stochastik bekannt und von WACKERBAUER et al. (1994) verwendet:

$$p_{L,i\to j} = \frac{p_{L,ij}}{p_{L,i}} \tag{10}$$

Die relative Häufigkeit $p_{L,ij}$ für das aufeinanderfolgende Auftreten der Wörter *i* und *j* lässt sich nach Gleichung (10) durch Multiplikation von (8) und (9) berechnen

$$p_{L,ij} = \frac{n_{L+1,j}}{n_{0,0}} \tag{11}$$

und stimmt mit der Häufigkeit $p_{L+1,j}$ für das aus *i* und *j* gebildete (*L*+1)-Wort überein. Abb. 2-5 faßt die Berechnung der relativen Häufigkeiten über die Knotenbesuche noch einmal zusammen.

Die Anzahl n_K der Knoten eines voll besetzten Baumes berechnet sich als geometrische Reihe über die Anzahl λ^i der Knoten in den Schichten:

$$n_K = \sum_{i=0}^{L} \lambda^i = \frac{\lambda^{L+1} - 1}{\lambda - 1} \tag{12}$$

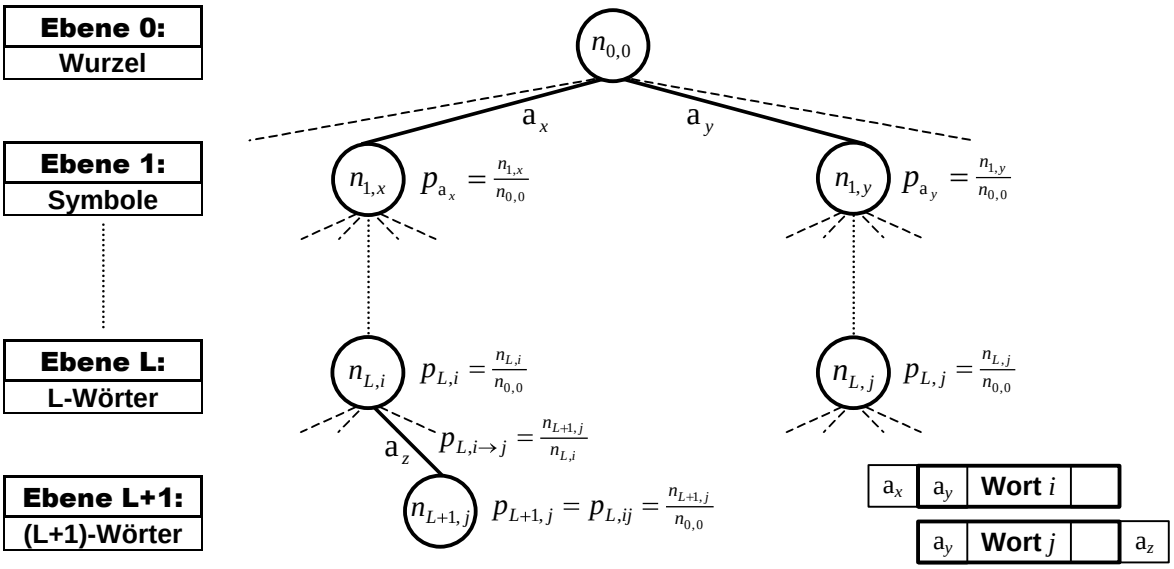

Abb. 2-5. Berechnung relativer Häufigkeiten $p_{...}$ über die Knotenbesuche $n_{...,...}$ der *L*-Wörter in einem Baum. Die *L*-Wörter *i* und *j* haben die Symbole in dem fett umrahmten Teil gemeinsam, also u. a. das Symbol a_y, an zweiter Stelle bei *i* und an erster Stelle bei *j*. Unterschiedlich sind a_x an erster Stelle von Wort *i* und a_z an letzter Stelle von *j*. Die relativen Häufigkeiten $p_{a...}$ der Symbole $a_{...}$ können anhand der Knotendurchläufe $n_{1,...}$ in der ersten Ebene ermittelt werden. Die relativen Häufigkeiten $p_{L,i}$ und $p_{L,j}$ der Wörter, ihres gemeinsamen Auftretens $p_{L,ij}$ und des Überganges $p_{L,i\to j}$ von Wort *i* nach *j* werden über die Anzahl der Knotenbesuche $n_{L,i}$, $n_{L,j}$ und $n_{L+1,i}$ berechnet.

Eine Baumstruktur wird auf natürliche Weise durch die Verzweigungen aufgebaut. Die Wörter werden kompakt gespeichert. Es gibt etablierte Verfahren für das Arbeiten auf Bäumen. In SYMDYN (siehe 2.7) werden die Bäume nach der Tiefe-Zuerst Suche (SEDGEWICK, 1988, S. 423) Postorder (BRONŠTEJN et al., 1997, S. 318) durchlaufen. Beginnend mit der Wurzel werden rekursiv zuerst die Kinder von links nach rechts und zuletzt der aktuelle Elternknoten durchlaufen. Desweiteren ist die Baumstruktur Ausgangspunkt für die nächste Datenebene:

2.1.3 Endliche Automaten

Die letzte Datenstruktur in dieser Hierarchie ist der endliche Automat. Über die Verteilung seiner Zustände wird die ε–Komplexität definiert. Diese Methoden basieren auf den Arbeiten von Crutchfield (CRUTCHFIELD & YOUNG, 1989, CRUTCHFIELD, 1991, 1992, 1994a, 1994b).

Zur Rekonstruktion eines Automaten betrachtet CRUTCHFIELD (1992) auf dem Analysebaum alle Unterbäume einer vorgegebenen Tiefe *D*, wobei $D \le L$. Eine übliche Wahl ist $D = L/2$ (CRUTCHFIELD, 1992). Abb. 2-6 zeigt die möglichen Unterbäume der Tiefe $D = 2$ für den von CRUTCHFIELD (1992) und hier bereits verwendeten Beispielprozess: „Jedes zweite Symbol ist eine 1“. Jeder Unterbaum repräsentiert ein charakteristisches Entwickungspotential für die *D* zukünftigen Zeitschritte. Daher werden sie von CRUTCHFIELD (1992, 1994a, 1994b) auch „future-morphs“ genannt. Sie definieren die Zustände *v* eines Automaten. Die auch von CRUTCHFIELD (1992) verwendete Bezeichnung *v* für „vertices“ stammt aus der Graphentheorie, da die Zustände eines Automaten als Knoten eines Graphen verstanden werden (HOROWITZ & SAHNI, 1981). Da die Menge der Zustände (in der Praxis) endlich ist, gehören die hier

konstruierten Automaten zur Klasse der endlichen (finiten) Automaten (HOPCROFT & ULLMAN, 1990).

Eine Verallgemeinerung der topologischen Äquivalenz von Unterbäumen ist die δ-Ähnlichkeit (CRUTCHFIELD & YOUNG, 1989). Demnach gehören zwei Unterbäume B_1 und B_2 zur selben Äquivalenzklasse (future-morph), wenn sie (i) topologisch äquivalent sind und (ii) sich die relativen Häufigkeiten aller Übergänge $p_{l,i\to j}$ zwischen den Knoten der Unterbäume um jeweils maximal $\delta > 0$ unterscheiden:

$$\mathrm{B}_1 \sim \mathrm{B}_2 \quad \text{genau dann, wenn } \left| p_{l,i\to j}^{(\mathrm{B}_1)} - p_{l,i\to j}^{(\mathrm{B}_2)} \right| < \delta, \ \forall l,i,j \tag{13}$$

Die erfolgreiche Rekonstruktion eines Automaten hängt entscheidend von den Zuständen ab, die durch die Äquivalenzrelation von Unterbäumen auf den Analysebaum gefunden werden. Bereits ein untypisches Wort kann die charakterischtische Topologie eines Baumes empfindlich stören. Daher kann es sinnvoll sein, die Zustände eines Automaten über eine δ-Ähnlichkeit der Unterbäume ohne topologische Gleichheit zu definieren. In SYMDYN (siehe 2.7) sind sowohl rein topologische, topologische δ–, als auch reine δ-Ähnlichkeit möglich und verwendet worden. Die δ-Ähnlichkeit wird auch von KURTHS et al. (1996) und WACKERBAUER et al. (1994) verwendet.

Die Verbindungen *e* („edges“, Kanten des Graphen) zwischen den Zuständen eines Automaten werden aus den relativen Häufigkeiten $p_{v \xrightarrow{s} v'}$ der Übergänge zwischen den Morphs *v* und *v'* für jedes Symbol *s* des Alphabets ermittelt. Sollte ein Teilbaum in der untersten Morph-Schicht eines Baumes (Ebene *L*–*D*) zum ersten Mal auftauchen, können von ihm keine Verbindungen zu anderen Morphs festgestellt werden. In diesem Fall versagt die Rekonstruktion eines Automaten. Derartige Zustände werden „Dangling states“ genannt. Zustände, von denen es nur Verbindungen zu anderen Zuständen hin gibt, werden transiente (vorübergehende) Zustände genannt.

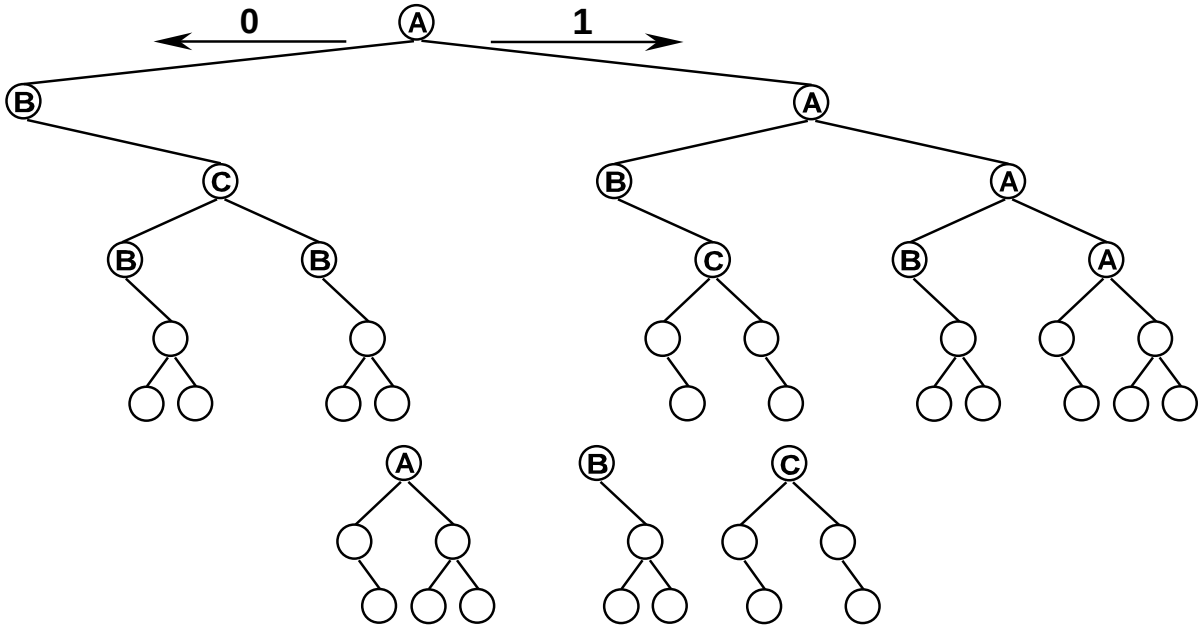

Abb. 2-6. Unterbäume der Tiefe *D* = 2 in dem Analysebaum der Tiefe *L* = 5 für den Prozess „Jedes zweite Symbol ist eine 1“. Es gibt genau drei topologisch verschiedene Unterbäume: **A**, **B** und **C**.

$$\mathbf{T}^{(0)} = \begin{pmatrix} 0 & \frac{1}{4} & 0 \\ 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \end{pmatrix}, \quad \mathbf{T}^{(1)} = \begin{pmatrix} \frac{3}{4} & 0 & 0 \\ 0 & 0 & 1 \\ 0 & \frac{1}{2} & 0 \end{pmatrix}, \quad \mathbf{T} = \begin{pmatrix} \frac{3}{4} & \frac{1}{4} & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

$$p_{\mathbf{A}\xrightarrow{0}\mathbf{B}} = \frac{1}{4}, \quad p_{\mathbf{C}\xrightarrow{0}\mathbf{B}} = \frac{1}{2}, \quad p_{\mathbf{A}\xrightarrow{1}\mathbf{A}} = \frac{3}{4}, \quad p_{\mathbf{B}\xrightarrow{1}\mathbf{C}} = 1, \quad p_{\mathbf{C}\xrightarrow{1}\mathbf{B}} = \frac{1}{2}$$

Abb. 2-7. Übergangsmatrizen für den Prozess „Jedes zweite Symbol ist eine 1". Zwischen den in Abb. 2-6 festgestellten Zuständen **A**, **B** und **C** gibt es nur fünf Verbindungen. Für diese sind die theoretischen Übergangswahrscheinlichkeiten $p_{\dots}$ für das jeweilige Symbol 0 oder 1 in den Matrizen $\mathbf{T}^{(0)}$ und $\mathbf{T}^{(1)}$ eingetragen. In der Übergangsmatrix **T** ist der Teil $\mathbf{T}^*$ der wiederkehrenden Zustände markiert. Eine ähnliche Darstellung findet sich bei Romahn (1996, S. 27).

Für jedes Symbol *s* werden die Übergangshäufigkeiten von jedem Zustand in jeden anderen in eine Transfermatrix $\mathbf{T}^{(s)}$ eingetragen. In den Zeilen werden die Zustände „von", und in den Spalten die Zustände „nach" notiert. Die Summe dieser Matrizen

$$\mathbf{T} = \sum_{s \in A} \mathbf{T}^{(s)} \tag{14}$$

liefert eine kompakte Beschreibung des Automaten. Da von jedem Zustand bei erfolgreicher Maschinen-Rekonstruktion zumindest Verbindungen weg führen, muß die Zeilensumme von **T** stets 1 ergeben. Abb. 2-7 stellt diese Matrizen für den Beispielprozess dar.

Der Zustand **A** in dem Beispiel ist ein transienter Zustand. Wenn er verlassen wird, verbleibt der Automat nur noch in den beiden wiederkehrenden Zuständen **B** und **C**. Abb. 2-8 zeigt den vollständigen Automaten für den Beispielprozess in der üblichen Darstellung (HOPCROFT & ULLMAN, 1990). Nach CRUTCHFIELD (1992) ist dies zugleich die minimale Beschreibung und damit die ε–Maschine für den Prozess.

Zur Beurteilung der Qualität eines Automaten wird sein stationärer Zustand **v** betrachtet, d. h. sein Verhalten nach vielen Zeitschritten, wenn alle vorübergehenden Zustände verlassen sind und sich die Aufenthaltswahrscheinlichkeiten nicht mehr ändern. Der Zustandsvektor **v** enthält die stationären Aufenthaltswahrscheinlichkeiten p_v aller wiederkehrenden Zustände des Automaten; diejenigen der transienten Zustände sind 0. Da Multiplikation eines beliebigen Zustandsvektors mit der Übergangsmatrix **T** den Zustandsvektor im nächsten Zeitschritt liefert, muss für die Multiplikation des stationären Wiederkehr-Zustandsvektors **v** mit dem Wiederkehrteil $\mathbf{T}^*$ von **T** gelten:

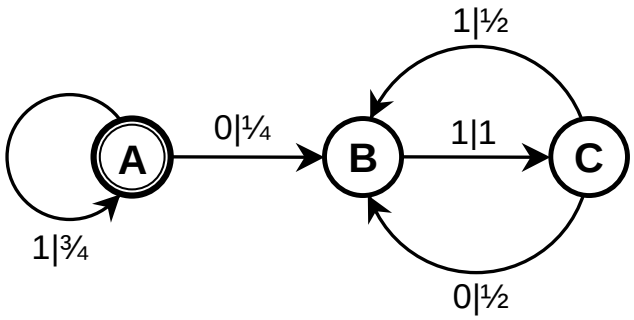

Abb. 2-8. Endlicher Automat und ε–Maschine für den Prozess „Jedes zweite Symbol ist eine 1". A ist der Anfangszustand. Die Pfeile beschreiben die Übergänge zwischen den Zuständen mit den Symbolen und ihren Wahrscheinlichkeiten. Vgl. auch die Darstellung von CRUTCHFIELD (1992).

$$\mathbf{v} = \mathbf{T}^* \cdot \mathbf{v} \tag{15}$$

v ist also ein Eigenvektor von $\mathbf{T}^*$ zum Eigenwert 1 und wird in SYMDYN (siehe 2.7) durch Lösung des zu (15) gehörenden homogenen Gleichungssystems durch Gauß-Elimination mit Pivot-Suche (STOER, 1994, S. 182ff) berechnet. Die Existenz des Eigenwertes 1 ergibt sich aus der Markov-Eigenschaft von $\mathbf{T}^*$ nach Definition. Da die stationären Wahrscheinlichkeiten aller (wiederkehrenden) Zustände zusammen 1 ergeben müssen, wird der Eigenvektor entsprechend normiert und ist damit eindeutig. Die wiederkehrenden Zustände stellen eine „strongly connected component" (SEDGEWICK, 1988, S. 481ff) in dem Automatengraphen dar, d. h. jeder Zustand ist von jedem anderen aus (über andere Zustände) erreichbar. Sie werden in SYMDYN mittels rekursiver Tiefe-Zuerst Suche (SEDGEWICK, 1988, S. 423) auf dem Baum aufgespürt.

2.1.4 Höhere Ebenen, ε-Maschinen

Bei CRUTCHFIELD (1994b) ist diese Hierarchie von Datenstrukturen der Anfang einer Hierarchie von Modellen. Der untersten Ebene entspricht der Symbolsatz, der bei CRUTCHFIELD (1994b) mit dem Datensatz übereinstimmt. Ein Kriterium für die Adäquatheit eines Modells ist seine beschränkte Größe bei zunehmender Genauigkeit. Demnach ist der Symbolsatz selbst das schlechteste Modell für den ihn generierenden Prozess, da für eine genauere Beschreibung eine stets zunehmende Anzahl von Symbolen erforderlich ist. Für periodische Prozesse liefert die nächsthöhere Ebene, die Baum-Modellklasse, ein adäquates Modell, wenn die Periodenlänge kleiner als die Baumtiefe ist. Die Baumgröße ändert sich auch bei theoretisch unendlicher Datenlänge nicht. Bei nicht-periodischen Prozessen wächst der Baum jedoch mit der Datenmenge. Wenn der Prozess nur kurz zeitlich korreliert ist, liefert die nächsthöhere Ebene der (stochastischen) finiten Automaten eine endliche Repräsentation. Aber es gibt auch Beispiele, bei denen die Beschreibungskapazität der endlichen Automaten nicht ausreicht (CRUTCHFIELD, 1994b).

Gibt es ein ausgezeichnetes Modell, das einen Prozess beschreibt? Falls ja, kann man einen Algorithmus angeben, der ein solches Modell liefert? Da alle Modelle in der Regel formal als Computerprogramm formuliert sind, liegt es nahe in der Automatentheorie und bei den formalen Sprachen, wie bei HOPCROFT & ULLMAN (1990) beschrieben, zu suchen. CRUTCHFIELD (1994b) schlägt die Rekonstruktion des

- *kleinsten* Modells
- auf der Ebene der *niedrigsten* Beschreibungskapazität,
- die eine *endliche* Repräsentation des Prozesses liefert,

vor. Dieses Modell nennt er „ε-Maschine". Den Weg dort hin beschreibt er als „hierarchische ε-Maschinen-Rekonstruktion" in fünf Schritten:

1. Auf der untersten Stufe steht der Symbolsatz (Datensatz), Modell $M_0 = S$.
2. Rekonstruktion des Modells M_l aus dem Modell M_{l-1} durch Zusammenfassen von Zuständen mit Regelmäßigkeiten von M_{l-1} als neue Zustände von M_l und den Verknüpfungen der Zustandsgruppen von M_{l-1} als Verknüpfungen der Zustände von M_l
3. Test der Beschreibungskapazität der l-ten Modellklasse durch sukzessiv genauere Modelle. ε bezeichnet den Grad der Approximation, mit $\varepsilon \to 0$ für beliebige Genauigkeit.
4. Wenn die Komplexität des Modells mit $\varepsilon \to 0$ divergiert, $\|M_l\| \to \infty$, gehe eine Modellklasse höher $l \rightarrow l+1$ und zurück zu 2.

5. Wenn $||M_l|| < \infty$ für $\varepsilon \to 0$, ist die erste und damit niedrigste Ebene erreicht, die genügend Beschreibungskapazität für eine endliche Darstellung des Prozesses hat. Eine ε–Maschine wurde rekonstruiert. Ende.

Die in Abschnitt 2.1.3 beschriebenen stochastischen finiten Automaten haben in der Automaten-Hierarchie nach CRUTCHFIELD (1994b) nur geringe Beschreibungskapazität und werden in dieser Hinsicht von vielen anderen Automatentypen übertroffen. Die höchste Beschreibungskapazität besitzt die Universelle Turing Maschine, die ein mathematisches Modell für die Arbeitsweise eines Computers darstellt (HOPCROFT & ULLMAN, 1990).

CRUTCHFIELD (1991, 1992, 1994a, 1994b) und CRUTCHFIELD & YOUNG (1989) geben keine konkrete Anweisung zur Konstruktion eines höheren Automaten (z. B. String Production) aus einem stochastischen endlichen Automaten an. In der Arbeit von 1991 werden Rekonstruktionsrelationen ausgewählter Automatentypen tabellarisch aufgezählt. KURTHS et al. (1996) und WACKERBAUER et al. (1994) verstehen unter einer ε–Maschine nach CRUTCHFIELD & YOUNG (1989) einen endlichen Automaten, wie in Abschnitt 2.1.3 beschrieben. Dabei bezeichnen sie den Diskriminanz-Parameter δ der Morph-Äquivalenz als ε. CRUTCHFIELD & YOUNG (1989) bezeichnen die Genauigkeit der Partitionierung eines Datensatzes in eine Symbolfolge mit ε. Es gibt also drei Bedeutungen der Variablen ε in diesem Zusammenhang.

Die Möglichkeit, höhere Automaten aus dem Finiten Automaten zu rekonstruieren bietet sich in der Praxis jedoch kaum, weil es in der Regel schwer genug ist, überhaupt einen finiten Automaten erfolgreich zu rekonstruieren. Für eine gelungene Automaten-Rekonstruktion der logistischen Funktion mit generierender Partitionierung fordern KURTHS et al. (1996) 50 000 bis 500 000 Datenpunkte und WACKERBAUER et al. (1994) sogar 10^8 Datenpunkte. Das ist für reale Messreihen nicht nur in der Ökosystemforschung praktisch unmöglich.

Crutchfields Idee, *das* (minimale) informationstheoretische Modell für einen Prozess zu konstruieren, ist der Traum eines kritischen Modellierers. Eine erfolgreiche hierarchische ε–Maschinen Rekonstruktion ist eine Alternative zur willkürlichen Auswahl einer Modellklasse mit einer beliebigen Anzahl von Parametern. Dabei wird genau und nur die in den Daten enthaltene Information und dessen Dynamik benutzt. Leider ist das Verfahren von Crutchfield (bisher) — von theoretischen Ausnahmen abgesehen — anscheinend praktisch nicht durchführbar, wie auch in dieser Arbeit festgestellt werden musste. Für theoretische Fragen ist es dennoch eine gute und wichtige Idee.

2.2 Maße für Korrelation

Von besonderem Interesse bei der Zeitreihenanalyse ist der zeitliche Zusammenhang (die Korrelation) von Datenpunkten bei einem bestimmten Zeitabstand (Lag). Wenn die Daten bei jedem Zeitabstand völlig unkorreliert sind, handelt es sich um reines Rauschen. Periodische Daten sind jeweils zu den ganzzahligen Vielfachen eines bestimmten Lags stark korreliert. Oftmals nimmt die Korrelation mit zunehmendem Zeitabstand ab. Der Zeitabstand, bei dem die Korrelation erstmals oder letztmals unter eine vorgegebene Signifikanzschwelle fällt, wird Korrelationslänge genannt.

Als Maße für Korrelation werden die Autokorrelation (Standardverfahren der Statistik) und die Transinformation (informationstheoretisches Pendant dazu) vorgestellt.

2.2.1 Autokorrelation

Für eine Zeitreihe $X = (x_0, x_1, ..., x_{N-1})$ mit dem Mittelwert (Erwartungswert)

$$\bar{x} = \frac{1}{N}\sum_{t=0}^{N-1} x_t \tag{16}$$

ist die empirische Autokovarianz zum Lag k (nach HARTUNG et al., 1998, S. 675; HIPEL & MCLEOD, 1994, S. 72; HONERKAMP, 1994, S. 383; und andere Bücher über Statistik):

$$c(k) = \frac{1}{N}\sum_{t=0}^{N-k-1}(x_t - \bar{x})(x_{t+k} - \bar{x}) \tag{17}$$

Der Faktor $1/N$ wird mitunter auch der Anzahl der Summanden entsprechend $1/(N–k)$ gewählt (HARTUNG et al., 1998, S. 675), was zu einer Aufwertung von $c(k)$ bei hohen Zeitabständen führt. Dies ist auch in SYMDYN möglich. Die Formel (17) liefert jedoch die beste Schätzung der Autokovarianz (HIPEL & MCLEOD, 1994, S. 72) und wird auch von HARTUNG et al. (1998, S. 675), HONERKAMP (1994, S. 382f), SCHLITTGEN & STREITBERG (1994, S. 7) u. a. bevorzugt. Es gilt $c(k) = c(–k)$. $c(0)$ ist die empirische Varianz mit dem Faktor $1/N$ anstatt $1/(N–1)$. Die empirische Autokorrelation zum Lag k ist:

$$r(k) = \frac{c(k)}{c(0)} = \frac{\sum_{t=0}^{N-k-1}(x_t - \bar{x})(x_{t+k} - \bar{x})}{\sum_{t=0}^{N-1}(x_t - \bar{x})^2} \tag{18}$$

In SYMDYN werden in den Formeln (16) bis (18) nur die lückenfreien Daten verwendet. Anstelle von N wird daher in den Faktoren die Anzahl der lückenfreien Daten eingesetzt.

$r(k)$ nimmt Werte von –1 bis +1 an. $r(0) = 1$. Ein positives Vorzeichen bedeutet eine positive Korrelation. Ein negatives Vorzeichen bedeutet, dass sich Werte mit dem Zeitabstand k entgegengesetzt verhalten: Wenn ein Wert hoch (gegenüber dem Erwartungswert) ist, ist er k Zeitschritte weiter — im Mittel — niedrig.

Für brauchbare Schätzungen der Autokorrelation empfiehlt HONERKAMP (1994, S. 383) eine Datenmenge von mindestens $N = 50$. Da die Werte von $r(k)$ wegen der abnehmenden Zahl von Summanden in (17) oder (18) mit zunehmendem Zeitabstand immer unsicherer werden, empfiehlt HONERKAMP (1994, S. 383) die Funktion nur bis $k = N/4$ auszuwerten.

Die Autokorrelationslänge gibt an, bis zu welchem Zeitabstand die Autokorrelation noch signifikant von 0 verschieden ist. Es wird also geprüft, wann $|r(k)|$ zuletzt oberhalb eines Schwellenwertes r_α liegt, der einem vorgegebenem Signifikanzniveau $\alpha \in [0,1]$ entspricht. α ist die Wahrscheinlichkeit für den Fehler 1. Art (siehe HARTUNG et al., 1998, S. 133ff), dass nicht signifikante Korrelationen fälschlicherweise als signifikant angenommen werden. Nach HONERKAMP (1994, S. 384, umgestellt) gilt

$$r_\alpha = \frac{\left|t_{N-2;1-\alpha/2}\right|}{\sqrt{N-2+t^2_{N-2;1-\alpha/2}}} \tag{19}$$

wobei $t_{N-2;1-\alpha/2}$ der Parameter aus der Student'schen t-Verteilung ist und in SYMDYN gemäß den von ABRAMOWITZ & STEGUN (1984, S. 409 u. 425) angegebenen Formeln berechnet wird. Die so berechneten t-Werte interpolieren die entsprechenden Tabellen in HONERKAMP (1994,

S. 511). Üblicherweise wird $\alpha = 0.05$ gewählt, d. h. 95 % aller nichtsignifikanten Korrelationen liegen unterhalb der Signifikanzschwelle.

In den Daten vorhandene Periodizitäten können elegant mittels Fourier-Analyse aufgespürt werden, wenn die Amplituden in einem Diagramm über die Frequenzen aufgetragen werden. Eine solche Darstellung wird Spektraldichte, Varianzspektrum oder Power-Spektrum genannt (HARTUNG et al., 1998, S. 702). Anhand des Amplitudenabfalls mit der Frequenz kann unterschiedliches chaotisches Verhalten oder farbiges Rauschen festgestellt werden (SCHROEDER, 1991). Die Spektren eignen sich so zur Klassifizierung von Daten. Auf diese Weise stellten PANDEY et al. (1998) den typischen Frequenz-Abfall von Abflusszeitreihen fest. In dieser Arbeit wird auf die Untersuchung von Spektren verzichtet, weil sie nicht (wesentlich) zur Klärung der hier gestellten Fragen beitragen. Ein Vergleich mit der Transinformation (siehe 2.2.2) — als informationstheoretische Alternative — ist außerdem nur mit der Autokorrelation direkt möglich. Entsprechende Berechnungen finden sich bei NEWIG (1998).

Die Berechnung der Autokorrelation setzt die Stationarität der Daten voraus, d. h. zeitlich konstante statistische Momente. Anderenfalls kann ein Trend oder eine Periodizität alle übrige Information überdecken (HARTUNG et al., 1998, S. 676). Gegebenenfalls sind vor einer Korrelationsanalyse Trends und Periodizitäten zu bereinigen. Dieses Problem wird in Abschnitt 3.3 ausführlich besprochen.

2.2.2 Transinformation

Die mittlere Informationsmenge, die in einer Vorhersage k Zeitschritte in die Zukunft enthalten ist, kann durch

$$I(k) = \sum_{i,j=0}^{\lambda-1} p_{ij}(k) \log_2 \frac{p_{ij}(k)}{p_i p_j} \tag{20}$$

berechnet werden (FARMER, 1982). $I(k)$ wird Transinformation zum Lag k oder wechselseitige Information (mutual information) genannt und ist auf Symbolfolgen (siehe 2.1.1) definiert. p_i und p_j sind die Wahrscheinlichkeiten für das Auftreten der Symbole s_i und s_j. $p_{ij}(k)$ ist die Wahrscheinlichkeit für das Auftreten von s_i und s_j k Zeitschritte später. Nach LI (1990) kann $I(k)$ auch über k Zeitschritte auseinander liegende Wörter definiert werden.

Die Transinformation ist ein Maß für die Abhängigkeit der Symbole. Nur bei statistischer Unabhängigkeit ist $p_{ij}(k) = p_i \cdot p_j$ und damit $I(k) = 0$. Dies gilt für die Autokorrelation $r(k)$ nur für $\lambda = 2$. HERZEL & GROßE (1995) zeigten allgemein, dass von den λ^2 Parametern der Korrelationsmatrix nur $(\lambda-1)^2$ unabhängig sind und dass Autokorrelationsfunktionen davon nur die $\lambda(\lambda-1)/2$ Parameter des symmetrischen Teils bestimmen können. Mit Kreuzkorrelationen können auch die antisymmetrischen Abhängigkeiten aufgespürt werden. Die Transinformation misst jede Art von Abhängigkeit der Symbole zum Lag k. Sie kann daher als Verallgemeinerung der Autokorrelation interpretiert werden (HERZEL & GROßE, 1995; KURTHS et al. 1996; LI, 1990). Andere Interpretationen sind die in einem Symbol über das andere gespeicherte Information oder der Grad der Vorhersagbarkeit des zweiten Symbols bei Kenntnis des ersten (LI, 1990).

Positive und negative Korrelationen werden im Gegensatz zur Autokorrelation nicht unterschieden. Jede Art von Abhängigkeit beim Lag k trägt zu einem positiven Wert von $I(k)$ bei. Bei Unkorreliertheit fällt $I(k)$ nach LI (1990) doppelt so schnell gegen 0 ab als $r(k)$. Die absoluten Werte von $I(k)$ sind oft klein gegenüber denen von $r(k)$ (HERZEL & GROßE, 1995).

Die endliche Datenlänge N führt bei den im Abschnitt 2.5 vorgestellten Entropie-Maßen bei höheren Wortlängen zu einer systematischen Unterschätzung des tatsächlichen Wertes. Der Wert der Transinformation wird nach HERZEL & GROßE (1995) im Mittel systematisch um

$$I_S = \frac{(\lambda - 1)^2}{2(N-k)\ln 2} \tag{21}$$

überschätzt. Eine signifikante Abhängigkeit $I(k)$ muss oberhalb dieser natürlichen Fluktuationen liegen.

Die statistische 95 % Signifikanz von Spitzen in der Transinformation ist nach KURTHS et al. (1996) erst oberhalb von

$$I_{95\%} = \frac{3.16}{N}\sqrt[4]{k} \tag{22}$$

gegeben. Die Signifikanzschwelle I_S oder $I_{95\%}$ liegt um so niedriger, je größer die Datenmenge N oder je kleiner der Lag k ist. Bei der Berechnung der Transinformation wird die Signifikanz der Werte nach (21) und (22) geprüft.

Die Transinformation kann auf dem Symbolsatz prinzipiell schon bei kleineren Datenmengen N oder mit größeren Alphabetumfängen λ berechnet werden, als die in Abschnitt 2.5 vorgestellten Wort-Entropien, da sie mit einer Entropie über 2-Wörter vergleichbar ist (HERZEL et al., 1994). Die Anzahl der Daten sollte deutlich (Faktor 10) größer als λ^2 sein (EBELING et al., 1995).

Mit Hilfe der Transinformation stellten EBELING et al. (1995) kurz- und mittelreichweitige Korrelationen in „Moby Dick" von Melville und „Grimms Märchen“ der Gebrüder Grimm fest, sowie einen prinzipiellen Unterschied im zeitlichen Zusammenhang von Mozarts Sonate KV 311 gegenüber einem Präludium in f-Moll von Bach und der Klaviersonate op. 31 Nr. 2 von Beethoven. In DNA-Sequenzen (Hefe Chromosom III) haben HERZEL et al. (1994) langreichweitige Korrelationen mittels Transinformation festgestellt.

2.3 Bewertung von Dynamik mit Referenzprozessen

Bevor die Maße für Information und Komplexität im Detail vorgestellt werden, sollen kurz drei wichtige theoretische Prozesse vorgestellt werden, an denen die unterschiedliche Bewertung von Dynamik durch die Maße getestet und klassifiziert wird. Das sind (i) streng periodische Prozesse, da in den Daten Tages- oder Jahresgänge häufig sind, (ii) Bernoulli-Prozesse, weil sich hier leicht über einen Parameter die Zufälligkeit steuern lässt und damit dessen Beurteilung durch die Maße abschätzen lässt, und (iii) die logistische Abbildung als gut untersuchtes Standardbeispiel und Testprozess für ein ganzes Spektrum an Dynamik, inklusive chaotischem Verhalten.

Es gibt weitere Modelle, die zum Testen von Methoden verwendet werden, die aber aufgrund ihrer Handhabbarkeit und Popularität hier nicht angewendet werden sollen. Dazu gehören insbesondere Markow-Modelle (siehe z. B. JETSCHKE, 1989, S. 207), die durch ihr Gedächtnis eine realistischere Modellierung von Zeitreihen erlauben, z. B. für DNA-Stränge: SCHMITT et al. (1993), HERZEL et al. (1994). Desweiteren gehören dazu die autoregressiven Prozesse der Zeitreihenanalyse (HARTUNG et al., 1998, S. 678ff) und die Hénon-Abbildung, z. B. in ECKMANN et al. (1987), KURTHS & SCHWARZ (1995). In den Abschnitten 2.1.2 und 2.1.3 wurde der Prozess „Jedes zweite Symbol ist eine 1“ bei der grafischen Veranschaulichung

verwendet und erklärt, der aber in der Literatur — außer bei CRUTCHFIELD (1992, 1994a, 1994b) — keine große Rolle spielt.

Die theoretische Untersuchung der Methoden liegt jedoch nur soweit im Interesse dieser Arbeit, wie sie für die Anwendung bedeutsam ist. Daher werden nur die im ersten Abschnitt genannten und nachfolgend beschriebenen Prozesse berücksichtigt. Sie sollen ein Gefühl für die Eigenschaften geben, die von den Komplexitätsmaßen erfasst werden, und dafür, wie diese bewertet werden. Dies ist nur bei bekannter Dynamik, also anhand der theoretischen Referenzprozesse, möglich. Es soll auch deutlich werden, welches Potential in der Methode steckt, trotz des großen Informationsverlustes bei binärer Partitionierung der Daten.

2.3.1 Periodische Prozesse

Die Daten x_t zu jedem Zeitpunkt t eines streng periodischen Prozesses mit Periode p wiederholen sich gemäß:

$$x_t = x_{t+p} \tag{23}$$

Periodizitäten gibt es in den untersuchten Daten als Tages- oder Jahresgänge. Diese sind aber in der Regel nie streng. (23) ist dafür nur ein idealisiertes Modell, das im Mittel zutrifft. Komplexitätsmaße können strenge Periodizitäten erkennen und nehmen meist einen Wert von 0 oder $\log_2 p$ an, wenn die Periodenlänge p nicht den Erfassungsbereich (z. B. die Wortlänge) der Methoden übertrifft.

2.3.2 Binärer Bernoulli-Prozess

Für Bernoulli-Prozesse (siehe z. B. HARTUNG et al., 1998, S. 199ff), wird hier der einfachste Fall von binären Symbolfolgen betrachtet, bei denen ein Symbol mit der Häufigkeit $p \in [0,1]$ und das andere Symbol mit der Häufigkeit $1 - p$ unabhängig von der Position im Symbolsatz auftritt. $p = 0$ oder $p = 1$ bedeutet, dass der Symbolsatz nur aus einem Symbol besteht. Mit Annäherung an $p = ½$ nimmt der Anteil des anderen Symbols zu. Bei $p = ½$ sind beide Symbole gleich häufig: Die Folge erinnert an einen Münzwurf. So lässt sich die Zufälligkeit direkt über den Parameter p steuern. In den Abbildungen in dieser Arbeit bezieht sich die Angabe „Zufälligkeit in Prozent“ auf $p \cdot 200$ % für $p \in [0,½]$ aus dem Bernoulli-Prozess.

Analog wurde die Bewertung von Zufälligkeit in den Daten durch die Komplexitätsmaße auch mit einem Markov-Modell erster Ordnung oder einem AR(1)-Prozess untersucht. Dabei bedeutet der Zufälligkeitsparameter nicht die Häufigkeit eines Symbols, sondern die Abhängigkeit (Korreliertheit) eines Symbols von seinem Vorgänger. Diese Modelle sind näher an den in dieser Arbeit untersuchten Zeitreihen als der Bernoulli-Prozess. Sie liefern jedoch eine sehr ähnliche Abhängigkeit der Komplexitätsmaße von der Zufälligkeit. Mit dem Bernoulli-Prozess lassen sich darüber hinaus leicht analytische Beziehungen formulieren.

In Anlehnung an den binären Bernoulli-Prozess wurde für einige Anwendungen, z. B. in Abb. 2-10, ein Prozess mit kontinuierlichem Wertespektrum definiert. Dazu wird jeweils eine Zufallszahl mit vorgegebener Wahrscheinlichkeit p wiederholt. Tritt das Ereignis „Zahl nicht wiederholen“ ein, so wird eine neue Zufallszahl ermittelt, die nach der Wahrscheinlichkeit p wiederholt wird.

2.3.3 Logistische Abbildung

Die logistische Abbildung diente ursprünglich u. a. der Modellierung von sich saisonal fortpflanzenden Populationen, in denen sich die Generationen nicht überlappen. Dazu gehören insbesondere Insekten der gemäßigten Zonen. Neben diesem Faktum stellt MAY (1976) die bis dahin bekannte Vielfalt des Attraktors dieser und verwandter Funktionen vor, die für ihre eigentliche Berühmtheit und Beachtung auch in anderen Disziplinen verantwortlich ist. Die Anzahl der Individuen x_{t+1} der Generation (t+1) berechnet sich aus derjenigen der vorangegangenen Generation t gemäß

$$x_{t+1} = r \cdot x_t \cdot (1 - x_t) \tag{24}$$

Dabei ist $r \in [0,4]$ der Wachstumsparameter des Modells, das die Populationsdynamik normiert beschreibt (siehe MAY, 1976), d. h. $x \in [0,1]$. Die Randwerte $x_t = 0$ und $x_t = 1$ bedeuten $x_\tau = 0$ für alle $\tau > t$, also das Aussterben der Population.

Von dem Wachstumsparameter hängt es ab, ob die Population ausstirbt ($0 \le r \le 1$), eine stabile Größe erreicht ($1 < r \le 3$), gegen einen stabilen Zyklus konvergiert oder sich chaotisch entwickelt. Abb. 2-9 zeigt den interessanten Bereich ($3.5 \le r \le 4$) des Attraktors[1]. Der Attraktor ist gekennzeichnet durch eine Kaskade von Periodenverdopplungen durch wiederholte, sogenannte „Heugabel"-Bifurkationen (nach MAY, 1976) mit wachsendem r, beginnend mit $r = 3$ und Periode 2. Diese endet in einem Akkumulationspunkt mit Periode ∞ bei $r \approx 3.56994567$ (nach RATEITSCHAK et al., 1995), dem sogenannten Feigenbaum-Punkt nach FEIGENBAUM (1978), der eine universelle Konvergenzrate der Periodenverdopplungen festgestellt hat. Danach treten mit wachsendem r chaotische Bänder auf, die sich verbreitern, übereinander laufen und verschmelzen. Der Bandverschmelzungspunkt, an dem zum ersten Mal ein gemeinsames Band erreicht wird, liegt bei $r \approx 3.6785$ (nach JETSCHKE, 1989, S. 122). Dieses Band verbreitert sich weiter, bis bei $r = 4$ (voll entwickeltes Chaos) der gesamte Wertebereich (0,1) abgedeckt wird und das Modell zur Erzeugung von Zufallszahlen in diesem Bereich verwendet werden kann. In den Bändern treten immer wieder periodische Fenster auf, die wiederum über eine Kaskade von Periodenverdopplungen ins Chaos übergehen. Der Attraktor zeigt also die für chaotische Prozesse typische Selbstähnlichkeit (vgl. auch JETSCHKE, 1989, S. 125). Die Fenster werden durch das Phänomen der Intermittenz eingeleitet, d. h. ein periodisches Grundmuster in der Zeitreihe wird zu scheinbar zufälligen Zeiten von irregulären Sprüngen unterbrochen (JETSCHKE, 1989, S. 125f). Diese Störungen nehmen vor einem Fenster mit wachsendem r ab und verschwinden zu Beginn des Fensters. Das Fenster der Periode 3 beginnt bei $r = 1 + \sqrt{8} \approx 3.82842712$ (nach WACKERBAUER et al., 1994). Der Akkumulationspunkt liegt dann bei $r \approx 3.849$. Bei $r \approx 3.857$ entsteht ein 1-Band-Attraktor ohne Bandverschmelzung, d. h. es tritt ein plötzlicher Wechsel in der chaotischen Dynamik auf. Dieser Punkt wird innere Krise genannt (WACKERBAUER et al., 1994).

[1] Die Darstellung in Abb. 2-9 entspricht den Werten von 1001 Zeitreihen der Länge 160. Bei der weiteren Verwendung von Abb. 2-9 als Hintergrundbild werden nur 90 Werte im Abstand von $\Delta r = 0.001$, also 501 „Zeitreihen", gezeigt.

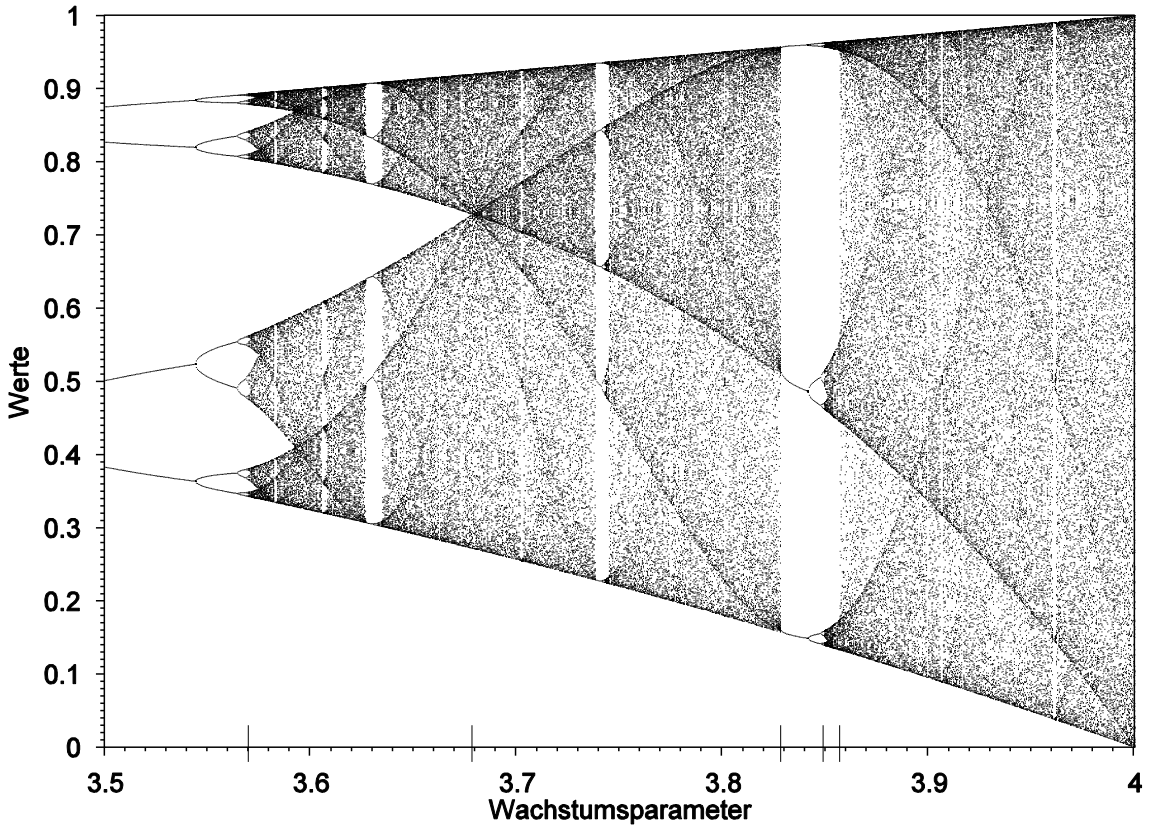

Abb. 2-9. Attraktor der logistischen Abbildung (24). Je 160 Datenpunkte (Werte) nach 10000 Voriterationen. Schrittweite für *r*: 0.0005. Damit sind 1001 „Zeitreihen" dargestellt. Die Markierungen auf der Ordinate entsprechen den im Text besprochenen Wachstumsparametern.

Alle im letzten Abschnitt genannten Punkte sind in Abb. 2-9 markiert. Aufgrund der Selbstähnlichkeit gibt es noch viele weitere Akkumulations-, Verschmelzungs-, Intermittenz- und Krisenpunkte im Attraktor der logistischen Abbildung. Für weitere Details und zur Erklärung der Dynamik sei auf die bereits erwähnte Literatur verwiesen. Einen Überblick über alle Phänomene mit Ausnahme der Krisen gibt JETSCHKE (1989, S. 117ff). Bereits MAY (1976) gibt zu Bedenken, dass trotz der deterministischen Struktur von (24) eine stochastische Beschreibung der Dynamik angemessen ist. Damit ist die logistische Abbildung ein Beispiel für ein deterministisches Modell eines biologischen Prozesses mit scheinbar stochastischem Verhalten. Dieses deterministische Chaos bedeutet eine schlechte Vorhersagbarkeit von Populationen mit Wachstumsparametern im chaotischen Bereich.

Die logistische Funktion ist eine der am häufigsten verwendeten Testfunktionen für Komplexitätsmaße und andere Methoden. Sie wurde unter anderem von CRUTCHFIELD & YOUNG (1989), CRUTCHFIELD (1994a, 1994b), GRASSBERGER (1986 u. 1988), KASPAR & SCHUSTER (1987), KURTHS et al. (1996), KURTHS & WITT (1994), TRULLA et al. (1996), WACKERBAUER et al. (1994) und WITT (1996) verwendet. Sie diente bei der Entwicklung von SYMDYN u. a. dazu, die korrekte Implementierung der Methoden anhand von veröffentlichten Resultaten zu prüfen. Die Arbeit von WACKERBAUER et al. (1994) eignete sich dazu wegen ihrer vergleichenden Darstellung der meisten auch hier verwendeten Methoden in besonderer Weise und wird für weitere Details empfohlen.

Zur Generierung von „Zeit"-reihen nach (24) wurde jeweils ein zufälliger Startwert $x_0 \in (0,1)$ vorgegeben. Um Einflüsse der Einschwingvorgänge zu vermeiden, wurden die ersten Iteratio-

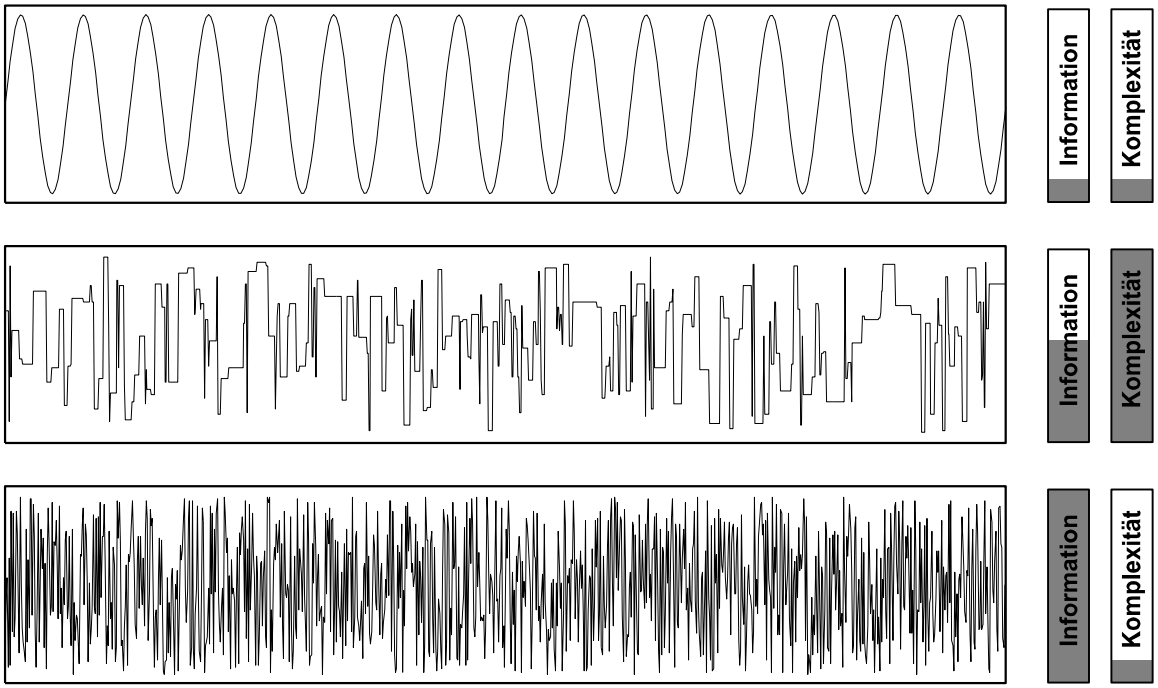

Abb. 2-10. Information und Komplexität von Zeitreihen. Qualitative Darstellung für die Sinus-Funktion (periodisch), einen Bernoulli-Prozess mit kontinuierlichen Zuständen und Wiederholungswahrscheinlichkeit 83 % (hoch strukturiert) und eine Zufallszahlenfolge (zufällig) nach PRESS et al. (1992, S. 282).

nen überlesen. Die Erzeugung einer Symbolfolge erfolgte, wie in der Literatur (s. o.) beschrieben, durch generierende binäre Partitionierung des Wertebereichs bei $x_p = ½$.

2.4 Typen von Komplexitätsmaßen

Grundsätzlich werden zwei verschiedene Typen von Komplexitätsmaßen unterschieden: Maße für *Information* (Zufälligkeit, Unvorhersagbarkeit, Unsicherheit, Unordnung, Maße erster Ordnung) und Maße für *Komplexität* (Maße für Struktur, Maße zweiter Ordnung). Die Zusammenfassung dieser Typen unter dem Oberbegriff „Komplexitäts"-Maß ist etwas verwirrend und hat zu verschiedenen Klarstellungen und zu der genannten Klassifizierung geführt (z. B. in: GRASSBERGER, 1986; CRUTCHFIELD, 1994a, 1994b; WACKERBAUER et al., 1994; KURTHS & WITT, 1994; KURTHS et al., 1996; ATMANSPACHER et al., 1997). Präzisierend werden in dieser Arbeit die Begriffe „Informationsmaß" oder „Komplexitätsmaß" verwendet.

Die prinzipielle Charakterisierung der Komplexitätsmaße erfolgt über deren Beurteilung der Zufälligkeit. Informationsmaße sind minimal bei konstanten Daten, nehmen mit der Zufälligkeit der Daten zu und erreichen bei völlig zufälligen Daten ihr Maximum (siehe später Abb. 2-11). Komplexitätsmaße verschwinden an beiden Extremen konstanter und völlig zufälliger Daten und nehmen dazwischen ein Maximum an (siehe später Abb. 2-20). FELDMAN & CRUTCHFIELD (1998) warnen allerdings davor, Komplexitätsmaße nur über diese Randbedingungen zu identifizieren, etwa indem man ein Maß zweiter Ordnung als umgekehrte Parabel-Funktion über ein Maß erster Ordnung definiert. Ein Komplexitätsmaß muß einen „strukturellen Gehalt" der Daten erfassen. Was genau als informativ oder komplex angesehen wird,

hängt von dem jeweiligen Maß ab und wird bei der Darstellung der einzelnen Maße in den nachfolgenden Abschnitten erklärt. In Anlehnung an ein gerne gezeigtes Bild von GRASSBERGER (1986) mit drei typischen räumlichen Mustern, veranschaulicht Abb. 2-10 die Beurteilung des Informations- und Komplexitätsgehaltes bei Zeitreihen[2].

WACKERBAUER et al. (1994) schlagen eine feinere Klassifizierung der Komplexitätsmaße vor. Sie unterscheiden *„strukturelle Maße"*, die keine explizite Information über die Dynamik eines Systems enthalten und beispielsweise auf (Wort-) Häufigkeiten (siehe 2.1.2) basieren, und *„dynamische Maße*", die explizit die Dynamik eines Systems berücksichtigen, z. B. durch Übergangswahrscheinlickeiten von Wörtern. Zusätzlich berücksichtigen sie bei der Klassifizierung die Art der Partitionierung, wobei sie — wie schon in 2.1.1.1 erwähnt — zwischen *homogen* und *generierend* unterscheiden. Die Kombination ergibt dann vier Typen von Komplexitätsmaßen.

2.5 Maße für Information

In diesem Abschnitt werden die in dieser Arbeit verwendeten Informationsmaße vorgestellt. Es handelt sich dabei um solche Maße, die in der Literatur häufiger beschrieben und verwendet werden. Neben der mathematischen Definition der Maße wird auch der Informationsbegriff, d. h. die Vorstellung von Information, erklärt, die zu dieser Definition geführt hat. Desweiteren werden fundamentale Eigenschaften der Maße sowie die Bewertung von Dynamik am Beispiel der in 2.3 genannten Referenzprozesse beschrieben.

Der Begriff „Entropie" stammt ursprünglich aus der Thermodynamik. Er wurde 1865 von Rudolf Clausius eingeführt und bezeichnet das Verhältnis von abgegebener oder aufgenommener Wärmemenge eines Systems zu seiner Temperatur (WOLKENSTEIN, 1990, S. 36f). Mit dieser fundamentalen Zustandsgröße kann der zweite Hauptsatz der Thermodynamik formuliert werden. Nach der altgriechischen Bedeutung des Wortes kann „Entropie" mit „Umwandlung" oder „Umwandelbarkeit" übersetzt werden (WOLKENSTEIN, 1990, S. 36). Mit Hilfe von Boltzmann's *H*-Theorem über die Verteilung der Mikrozustände eines Stoffes gelangt man (TOLMAN, 1967, S. 134 – 179) zur statistischen Entropie (25), S. 37. Wählt man für die Konstante in (25) die Boltzmann-Konstante $k = 1.38 \cdot 10^{-23}$ J/K, so entspricht unter Gleichgewichtsbedingungen die statistische Entropie genau der thermodynamischen Entropie (WILDE & SINGH, 1998, S. 42). Die Verknüpfung zwischen Mikrokosmos (z. B. Bewegung von Gasmolekülen) und Makrokosmos (z. B. Druck und Temperatur eines Gases) mittels stochastischer Methoden ist Gegenstand der statistischen Mechanik (DIU et al., 1994, S. 9). SHANNON (1948) hat aufgrund der Ähnlichkeit der von ihm gefundenen Formel für Information mit der statistischen Entropie als erster den Begriff Entropie in die Informationstheorie eingeführt (BALATONI & RÉNYI, 1956). Aber auch in der Thermodynamik spricht man von der Entropie als Maß der Unordnung (WOLKENSTEIN, 1990, S. 85), was vergleichbar mit der informationstheoretischen Bedeutung ist. Die Verbindungen zwischen beiden Entropiebegriffen sind Gegenstand einer umfangreichen Literatur (z. B.: ZUREK, 1990) und sollen hier nicht weiter vertieft werden. Bezüge zur Thermodynamik oder Statistischen Mechanik finden sich z. T. auch in neueren Arbeiten der Komplexitätstheorie, z. B. bei CRUTCHFIELD (1992).

Die Messung der Informationsmenge einer Nachricht muss unabhängig von der Form (z. B. Sprache) und dem Inhalt sein (vgl. RÉNYI, 1976, S. 435). Als einheitliches Bezugssystem zur Darstellung von Information wird das kleinstmögliche, das binäre System (0/1-System,

[2] Diese Abbildung wurde auch von NEWIG (1998, S. 24) übernommen.

Dualsystem, dyadische System), verwendet. Beispiel: Ein bestimmtes Ereignis von 16 möglichen Ereignissen, unter denen man frei wählen kann, kann durch vier ja/nein-Fragen spezifiziert werden. Es kann durch eine 4-stellige Dualzahl dargestellt werden und mit vier Zweipunkt-Relais (0/1-Schalter) gespeichert werden (vgl. WEAVER, 1976, S. 18f). Ein solches Ereignis hat einen Informationsgehalt (Shannon-Entropie oder topologische Entropie) von 4 Bit (von „binary digit“). Die fundamentale Operation zur Berechnung einer Information ist das Logarithmieren zur Basis 2: $4 = \log_2 16$. Dies ermöglicht die Additivität (siehe SHANNON, 1976, S. 60, WEAVER, 1976, S. 19, oder Abschnitt 2.5.1) der Information und liefert den Bezug zu einem binären System.

Die zuerst in diesem Abschnitt beschriebenen Informationsmaße, Shannon-, Rényi-, und Metrische Entropie, Informationsgewinn und wechselseitige Information, werden auch unter dem Begriff Entropie zusammengefasst, da sie alle von der Shannon-Entropie abgeleitet sind. Sie werden in dieser Arbeit auf einer Verteilung von *L*-Wörtern (siehe Abschnitt 2.1.2) berechnet. Die Shannon-Entropie selbst wird hier nicht auf Zeitreihen angewendet. Sie wird jedoch als Grundlage der auf ihr aufbauenden Maße beschrieben. Auch die Rényi-Entropie wird hier nicht direkt auf Zeitreihen angewendet. Sie veranschaulicht allerdings weitere Eigenschaften der Shannon-Entropie und dient als Grundlage der Rényi-Komplexität.

2.5.1 Shannon-Entropie

Die Shannon-Entropie ist das historisch erste berechenbare Maß für Information mit praktischer Relevanz. Es wurde 1948 von Claude E. Shannon bei den amerikanischen „Bell Telephone Laboratories“ im Rahmen einer mathematischen Theorie der Kommunikation (SHANNON, 1948, 1976) vorgestellt. Dabei ging es um die Messung der Information einer Nachrichtenquelle, um die Kapazität eines Übertragungskanals (Telefonleitung), um Störungen der Informationsübertragung, die Codierung von Nachrichten und kontinuierliche Nachrichten.

Der Ansatz von SHANNON (1976, S. 59f) zur Messung der Information einer Nachrichtenquelle war, ein Maß für die Wahlfreiheit oder Unsicherheit bei der Auswahl eines bestimmten Ereignisses aus einer Menge von *n* möglichen Ereignissen zu finden. Von einem solchen Maß *H* verlangt er die folgenden Eigenschaften, wenn die Wahrscheinlichkeiten $p_1, p_2, \ldots p_n$ der *n* möglichen Ereignisse bekannt sind:

1. *H* soll stetig in den p_i sein.
2. Wenn alle p_i gleich sind, $p_i = 1/n$, soll *H* monoton mit *n* wachsen.
3. Wenn eine Auswahl in zwei aufeinanderfolgende Wahlvorgänge aufgeteilt wird, soll *H* gleich der gewichteten Summe der individuellen *H*-Werte sein.

Diese Bedingungen können — wie von Shannon gezeigt — nur von

$$H = -k\sum_{i=1}^{n} p_i \log p_i \tag{25}$$

erfüllt werden. Wobei (25) bis auf eine positive Konstante *k* eindeutig ist, welche die Einheit festlegt. *H* wird als statistische Entropie bezeichnet, wenn für *k* die Boltzmann-Konstante gewählt wird (siehe Beginn von Abschnitt 2.5). Die Shannon-Entropie

$$H_S = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{26}$$

misst den Informationsgehalt einer Nachrichtenquelle in Bit, also bezogen auf einen binären Informationsspeicher. Die Konstante wird per Definition auf $k = 1/\log(2)$ gesetzt (BALATONI & RÉNYI, 1956). Wenn ein Ereignis mit Sicherheit eintritt und damit keine Wahl mehr bleibt ($p_j = 1$ und $p_i = 0$ für alle $i \neq j$), ist $H_S = 0$. Bei maximaler Wahlfreiheit und Unsicherheit sind alle $p_i = 1/n$ und damit $H_S = \log_2 n$. Wegen der Monotonie gilt dann:

$$0 \leq H_S \leq \log_2 n \tag{27}$$

Bei der Zeitreihenanalyse werden die Shannon-Entropie sowie die davon abgeleiteten Maße üblicherweise auf der Verteilung von Teilfolgen des Symbolsatzes (siehe 2.1.1) einer festen Länge L (Blöcke oder L-Wörter, siehe 2.1.2) berechnet. H_S wird dann auch Blockentropie oder Entropie per Block genannt (EBELING et al. 1998, S. 178). Bei einer Alphabetgröße von λ gibt es λ^L Wörter und damit gilt nach (27):

$$0 \leq H_S \leq L \log_2 \lambda \tag{28}$$

Die Shannon-Entropie nimmt (für Zufallsfolgen) also linear mit der Wortlänge und Steigung $\log_2 \lambda$ (= 1 bei binärem Alphabet) zu.

Shannon's Untersuchungen von 1949 waren mathematisch noch nicht ganz ausgereift (BALATONI & RÉNYI, 1956). Khinchin hat 1953 den Eindeutigkeitsbeweis der Entropie-Formel (25) mit nur zwei fundamentalen Annahmen geführt (KHINCHIN, 1957). BALATONI & RÉNYI (1956) haben die Beziehungen der Shannon-Entropie wahrscheinlichkeitstheoretisch präzisiert und ergänzt. Weitere Hinweise zur Shannon-Entropie — auch Shannon Information, Block Entropie, Entropie pro Teilwort genannt — finden sich bei EBELING (1997), EBELING et al. (1995), EBELING et al. (1996), FELDMAN & CRUTCHFIELD (1998), KURTHS et al. (1996), PRESS et al. (1992, S. 632), RATEITSCHAK et al. (1995), SHANNON (1976), WACKERBAUER et al.

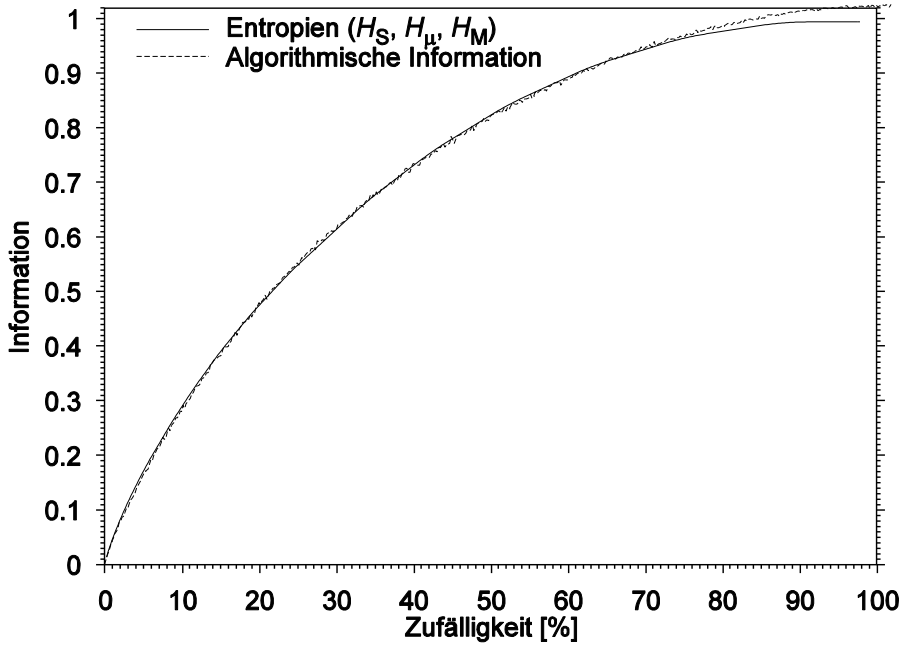

Abb. 2-11. Information und Zufälligkeit. Entropie und Algorithmische Information für den Bernoulli-Prozess. Zufälligkeit = $p \cdot 200$ % (siehe 2.3.2). Entropien: Shannon-Entropie nach (29) für $L = 1$, Metrische Entropie und Informationsgewinn nach (39), Wechselseitige Information nach (47) für $L = 2$. Algorithmische Information (siehe 2.5.6) empirisch für je 100000 Datenpunkte.

(1994), WITT et al. (1994).

Bewertung von Dynamik:

Wenn die Daten periodisch mit Periode $p \le L$ sind, gibt es nur p verschiedene L-Wörter mit gleicher Häufigkeit. Daraus folgt $H_S = \log_2 p$ (siehe WACKERBAUER et al., 1994). Falls $p > L$ kann die Periode von der Shannon-Entropie, sowie von allen anderen Methoden, die auf Wort-Häufigkeiten basieren, nicht entdeckt werden.

Für den binären Bernoulli-Prozess mit Zufälligkeitsparameter p (siehe 2.3.2) kann die Shannon-Entropie in Abhängigkeit von der Wortlänge L — wie in Anhang 7.1 gezeigt — direkt angegeben werden:

$$H_S(L,p) = -L\left[p\log_2 p + (1-p)\log_2(1-p)\right] \tag{29}$$

Diese Funktion wurde bereits von SHANNON (1976, S. 62) für $L = 1$ beschrieben. Sie ist in Abb. 2-11 und Abb. 2-13 ($\alpha = 1$) dargestellt. H_S nimmt demnach streng monoton mit der Zufälligkeit zu, ist also ein Maß für Zufälligkeit, wie man es von einem Informationsmaß erwarten kann.

Die Shannon-Entropie der logistischen Abbildung ist in Abb. 2-14 dargestellt und wird mit der Rényi-Entropie im nächsten Abschnitt diskutiert. Auch dabei wird die Messung von Zufälligkeit als die (wesentliche) Eigenschaft der Shannon-Entropie bestätigt.

2.5.2 Rényi-Entropie

RÉNYI (1960 u. 1961) schlägt eine Verallgemeinerung der Shannon-Entropie durch Gewichtung seltener oder häufiger Ereignisse (Wörter) vor. Für eine Zufallsvariable, dessen n Werte mit den Wahrscheinlichkeiten $p_1, p_2, \ldots p_n$ eintreffen, definiert er (verändert nach RÉNYI, 1976, S. 474) das Maß α-ter Ordnung der Information $H_R(\alpha)$ so:

$$\begin{aligned} H_R(\alpha) &= \frac{1}{1-\alpha}\log_2 \sum_{i=1}^{n} p_i^{\alpha} \quad \text{für} \quad \alpha \neq 1 \\ H_R(\alpha) &= -\sum_{i=1}^{n} p_i \log_2 p_i \qquad \text{für} \quad \alpha = 1 \end{aligned} \tag{30}$$

Durch die Shannon-Entropie $H_R(\alpha = 1)$ wird die Formel in der ersten Zeile von (30) stetig ergänzt. Insgesamt ist die Rényi-Entropie $H_R(\alpha)$ eine stetige, monoton fallende Funktion von α (siehe Abb. 2-12). Für $|\alpha| > 1$ erfahren die höheren Wahrscheinlichkeiten p_i in (30) eine Aufwertung gegenüber den niedrigeren, bis letztere bei extrem großem $|\alpha|$ gegenüber dem wahrscheinlichsten Ereignis vernachlässigt werden können. Für $|\alpha| < 1$ werden umgekehrt die geringeren Wahrscheinlichkeiten stärker gewichtet, bis bei $\alpha = 0$ alle Ereignisse gleich bewertet werden. Diese beiden Extremfälle werden nachfolgend genauer besprochen:

Neben der Shannon-Entropie für $\alpha = 1$ liefert die Rényi-Entropie für $\alpha = 0$ als weiteren Spezialfall die Hartley'sche Formel (RÉNYI, 1976, S. 437):

$$H_{top} = \log_2 n' \tag{31}$$

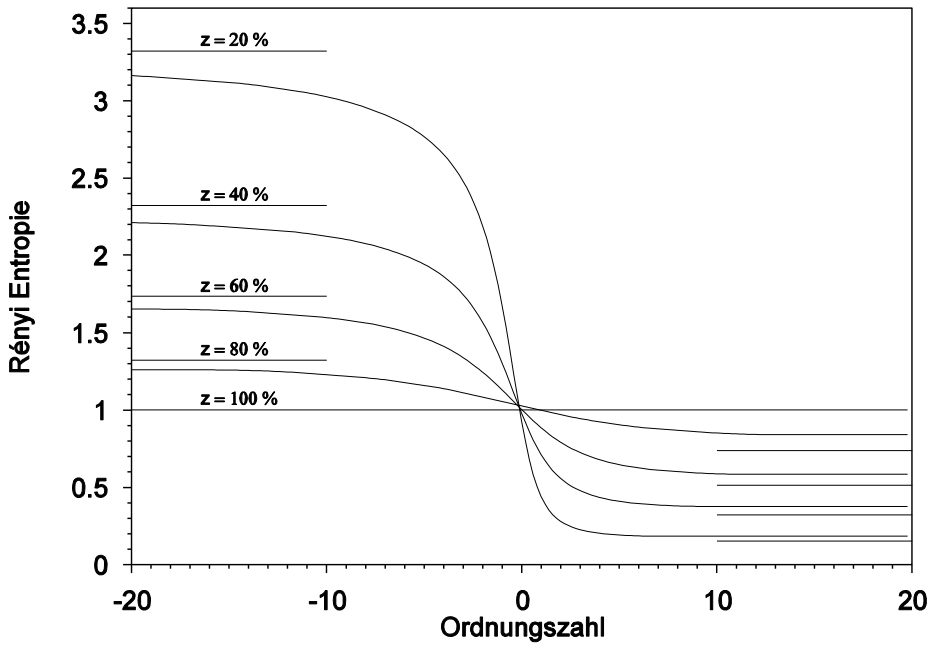

Abb. 2-12. Rényi-Entropie in Abhängigkeit von der Ordnungszahl α für verschiedene Zufälligkeiten *z*. Binärer Bernoulli-Prozess nach Gleichung (34) mit $z = p \cdot 200$ %. Die Asymptoten für $\alpha \to \pm\infty$ wurden nach Gleichung (33) berechnet. Da es für $p > 0$ stets zwei Zustände gibt, schneiden sich alle Kurven bei $\alpha = 0$ und einer topologischen Entropie von $H_{top} = 1$.

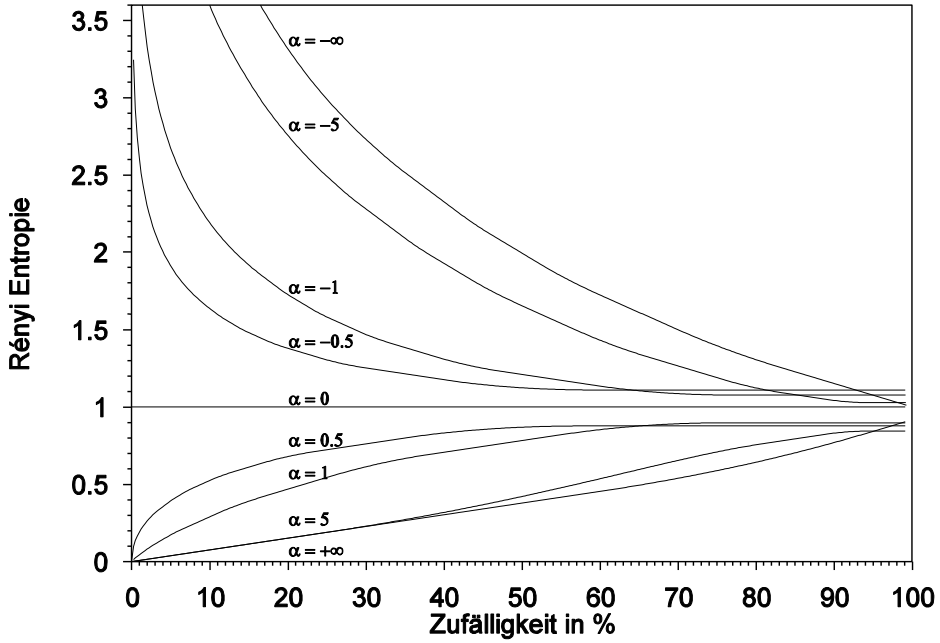

Abb. 2-13. Rényi-Entropie in Abhängigkeit von der Zufälligkeit für verschiedene Ordnungszahlen α. Binärer Bernoulli-Prozess nach Gleichung (34) mit Zufälligkeit = $p \cdot 200$ %.

Dabei ist n' die Anzahl der Werte der Zufallsvariablen mit positiver Eintrittswahrscheinlichkeit ($p_i > 0$). H_{top} gibt also die auf ein binäres System bezogene Informationsmenge an, die zur Charakterisierung eines möglicherweise eintreffenden Wertes der Zufallsgröße unabhängig von der Verteilung der Eintrittswahrscheinlichkeiten nötig ist. H_{top} bewegt sich in den selben Grenzen (27) oder (28) wie die Shannon-Entropie und wird *topologische Entropie* genannt,

weil nur die Anzahl der überhaupt möglichen Ereignisse berücksichtigt wird. CRUTCHFIELD (1994a) bezeichnet erst den Grenzwert

$$h_{\text{top}} = \lim_{L\to\infty} \frac{\log_2 n'}{L} \tag{32}$$

als topologische Entropie für einen Symbolsatz mit jeweils n' möglichen Teilsequenzen der Länge L. Dies ist das topologische Analogon der metrischen Entropie der Quelle (37).

Für große positive ($\alpha \to +\infty$) oder negative ($\alpha \to -\infty$) Werte von α nähert sich H_R asymptotisch jeweils einem Grenzwert. Es sei p_{max} die höchste und p_{min} die niedrigste Wahrscheinlichkeit in der Verteilung. Dann wird die Summe in Gleichung (30) bei großen Werten von α, $|\alpha| \to 1$, durch p_{max}^{α}, falls $\alpha \to 1$, und durch p_{min}^{α}, falls $\alpha \to -1$, dominiert. Entsprechend gilt unabhängig von der Anzahl der Ereignisse, die mit diesen extremen Wahrscheinlichkeiten auftreten:

$$\lim_{\alpha\to+\infty} H_R(\alpha) = -\log_2 p_{max} \quad \text{und} \quad \lim_{\alpha\to-\infty} H_R(\alpha) = -\log_2 p_{min} \tag{33}$$

Abb. 2-12 zeigt exemplarisch den Zusammenhang der Rényi-Entropie mit α und den Asymptoten.

Weitere Hinweise zur Rényi-Entropie — auch Rényi Information oder verallgemeinerte Information genannt — finden sich bei EBELING et al. (1995), KURTHS et al. (1996), KURTHS & WITT (1994), WACKERBAUER et al. (1994), WITT et al. (1994).

Bewertung von Dynamik:

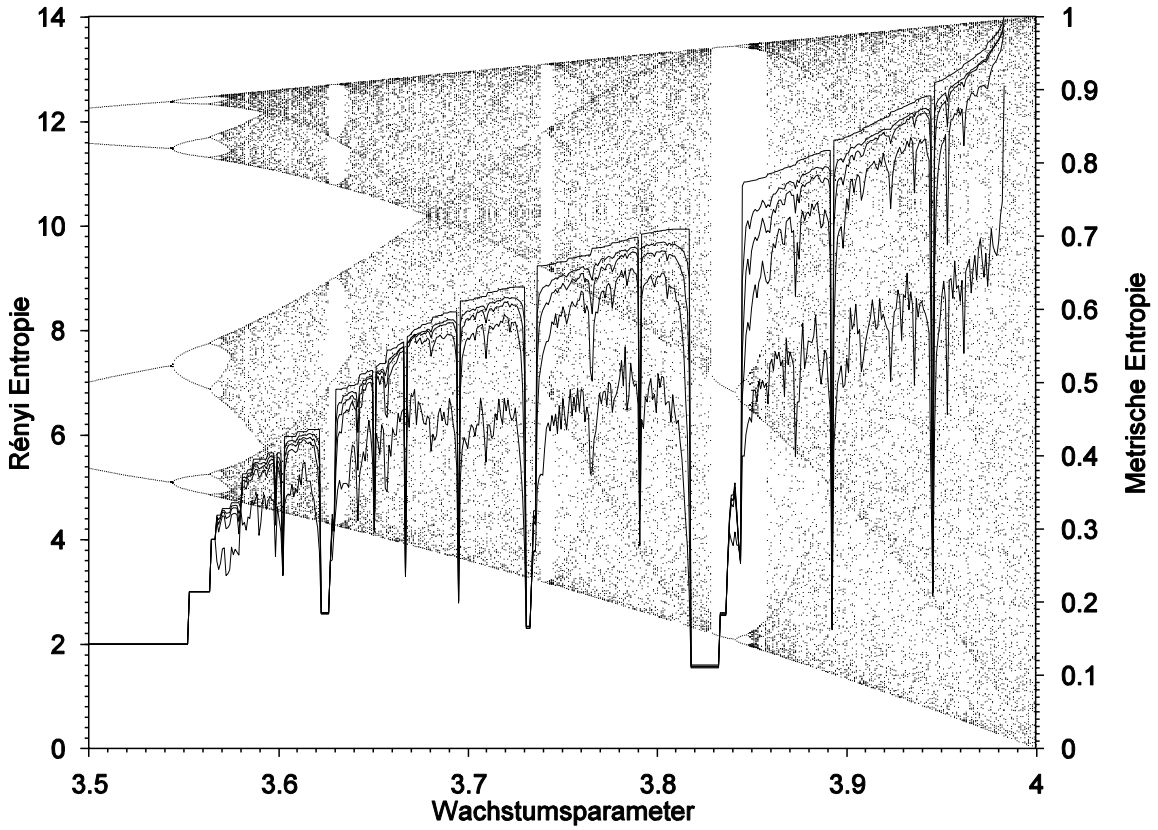

Abb. 2-14. Rényi-Entropie $H_R(\alpha)$ für die logistische Abbildung (24). $H_R(\alpha)$ ist für die Ordnungszahlen α = 0, 0.5, 1, 2, und 64 von oben nach unten abgebildet (linke Skala). $H_R(0)$ ist die **Topologische Entropie**; $H_R(1)$ die **Shannon-Entropie** und mit der rechten Werteskala die **Metrische Entropie**. Je N = 100000 Daten nach 10000 Voriterationen, statische binäre Partitionierung bei 0.5, Wortlänge L = 14. Hintergrund: Attraktor nach Abb. 2-9.

Für den binären Bernoulli-Prozess mit Zufälligkeitsparameter p (siehe 2.3.2) ist die Rényi-Entropie der L-Wörter in Abhängigkeit von p und α nach Anhang 7.2:

$$H_R(\alpha, p) = \frac{L}{1-\alpha} \log_2 \left(p^\alpha + (1-p)^\alpha \right) \tag{34}$$

Abb. 2-13 zeigt, dass die Rényi-Entropie für $\alpha \in (0,\infty)$ streng monoton mit der Zufälligkeit zunimmt, also ein Maß für Zufälligkeit und Information ist. Für $\alpha < 0$ ist die Rényi-Entropie ein Maß für Ordnung und Redundanz. Die praktisch interessante Gewichtung seltenerer Wörter durch Wurzelziehen ($0 < \alpha < 1$) und häufiger Wörter durch Potenzieren ($\alpha > 1$) kennzeichnet das Informationsmaß. Der Fall $\alpha < 0$ wird daher nicht weiter diskutiert.

Periodizitäten der Länge p werden unabhängig von α mit $\log_2 p$ bewertet (vgl. 2.5.1 und WACKERBAUER et al., 1994), falls $p \leq L$. Die periodischen Bereiche bei der logistischen Abbildung (siehe Abb. 2-14) fallen durch konstante tiefe Stufen in der Funktion H_R über den Wachstumsparameter auf. Die Akkumulationspunkte der Periodenverdopplung werden wegen $p > L$ nicht entdeckt. Der wesentliche Verlauf von H_R in Abb. 2-14 wird durch die Anzahl der Wörter bestimmt, die mit dem Wachstumsparameter tendenziell zunimmt und ein Indiz für die ebenfalls zunehmende Zufälligkeit (Breite der Chaosbänder) ist. Dies zeigt die Topologische Entropie $H_R(0)$, die nur von der Anzahl der Wörter abhängt. Ihr stufiger Verlauf ist durch die diskreten Werte der sich nur allmählich ändernden Wortmenge bedingt. Mit zunehmender Ordnung α macht sich die Ungleichverteilung der Wörter bemerkbar. Dadurch werden insbesondere die noch bei $H_R(0)$ vorhandenen Ecken vor und nach den periodischen Fenstern abgerundet. Ursache dafür ist die zunehmende Seltenheit der Nicht-Periodischen Wörter vor Beginn der Fenster (Intermittenz) und der umgekehrte Vorgang nach Ende der Fenster. Insgesamt bestätigen diese Beobachtungen die Rényi-Entropie als Maß für Zufälligkeit und Information.

Ab $\alpha = 64$ ändert sich $H_R(\alpha)$ mit weiterer Zunahme von α kaum noch, so dass 64 als Grenzwert für $\alpha \to \infty$ in Abb. 2-14 gewählt wurde. Die tendenzielle Zunahme von $H_R(64)$ mit dem Wachstumsparameter ist dann nur noch wenig oder kaum ausgeprägt, da H_R nur noch durch die größte Worthäufigkeit bestimmt wird. Die Schwankung der Werte nimmt mit wachsendem α zu. Die Shannon-Entropie $H_R(1)$ ist daher ein Informationsmaß, das sowohl die topologische Wortstruktur und Zufälligkeit als auch die spezielle Wortverteilung moderat berücksichtigt. Sie entdeckt auch zwischen den Fenstern durch signifikante Einbrüche dynamische Besonderheiten, die der Attraktor in Abb. 2-9 alleine nicht erkennen lässt.

2.5.3 Metrische Entropie

Gemäß der zweiten Forderung zur Definition der Shannon-Entropie (siehe 2.5.1) nimmt diese monoton mit der Anzahl der möglichen Ereignisse zu. Da bei der Zeitreihenanalyse die Shannon-Entropie auf der Verteilung der L-Wörter (siehe 2.1.2) berechnet wird, bewegt sie sich in den in Ungleichung (28) formulierten Grenzen. Um die Entropien unabhängig von der Wortlänge vergleichen zu können, bietet sich eine Normierung durch den Quotienten H_S/L an:

$$H_\mu = -\frac{1}{L} \sum_{i=1}^{n} p_i \log_2 p_i \tag{35}$$

Hierbei bezeichnen p_i die Häufigkeiten der maximal $n = \lambda^L$ L-Wörter. Die Metrische Entropie H_μ ist die mittlere Entropie oder Unsicherheit pro Symbol eines L-Wortes (EBELING et al. 1998, S. 98, 178) und liegt in den Grenzen:

$$0 \leq H_\mu \leq \log_2 \lambda \tag{36}$$

Der Wertebereich der metrischen Entropie hängt also nur noch von der Alphabetgröße λ der Partitionierung (siehe 2.1.1) ab. Bei binärer Partitionierung, $\lambda = 2$, ist $H_\mu \in [0,1]$. Eine Normierung bezüglich der Alphabetgröße wird nicht vorgenommen, um — wie bei Informationsmaßen üblich — den Bezug zu einem binären Informationsspeicher und der Einheit Bit zu bewahren.

Eine echte Unabhängigkeit von der Wortlänge L wird durch die Normierung nicht erreicht. H_μ ist eine monoton fallende Funktion von L (SHANNON, 1976, S. 66). In der Praxis, d. h. bei endlicher Datenmenge, konvergiert die Metrische Entropie mit der Wortlänge gegen 0, weil ab einer bestimmten Wortlänge nur noch wenige einmalige Wörter vorkommen und die Shannon-Entropie dann nicht mehr mit der Wortlänge wächst. Nach SHANNON (1976, S. 66) konvergiert H_μ für $L \to \infty$ gegen die Entropie der Nachrichtenquelle, d. h. hier also gegen die Entropie des die Zeitreihe erzeugenden Prozesses. Die Entropie der Quelle ist die Größe, an der man eigentlich interessiert ist. Die Entropie pro Symbol eines L-Wortes ist nur eine mehr oder weniger gute Approximation daran. Ein solcher Grenzübergang ist, wie bereits angedeutet, jedoch in der praktischen Zeitreihenanalyse nicht operationalisierbar, da alle Wort-Entropien und -Komplexitäten nur für, verglichen mit der Datenlänge, sehr kleine Wortlängen vertrauenswürdige Werte liefern (siehe 3.6). Der Grenzwert

$$h = \lim_{L \to \infty} H_\mu \tag{37}$$

wird als (Shannon) Entropie der Quelle, Shannon-Entropie (!), Entropierate oder Metrische Entropie (!) bezeichnet (EBELING et al. 1998, S. 178, FELDMAN & CRUTCHFIELD, 1998, WITT, 1996, S. 10), wenn die Systemgröße gegen unendlich geht. Wird gleichzeitig das Supremum über alle Partitionierungen Π (mit zunehmender Genauigkeit) gebildet, gelangt man zur Kolmogorov-Sinai Entropie (siehe: BADII & POLITI, 1997, S. 113, EBELING et al., 1998, S. 139 u. 178, GRASSBERGER, 1986, S. 920, KURTHS et al., 1996, RATEITSCHAK et al. (1995), WACKERBAUER et al., 1994, S. 145):

$$H_{\mathrm{KS}} = \sup_{\Pi} \lim_{L \to \infty} H_\mu \tag{38}$$

H_{KS} kann auch über die bedingte Information (siehe 2.5.4) definiert werden, die für $L \to \infty$ denselben Grenzwert wie mit H_μ hat (SHANNON, 1976, S. 66). Wie bereits erwähnt können h und H_{KS} für reale Messreihen nicht berechnet werden. In der Statistischen Mechanik etwa, in der die Systeme (z. B. ein Gas) typischerweise aus 10^{23} Atomen oder Molekülen bestehen (WILDE & SINGH, 1998, S. 3), sind derartige Definitionen durchaus sinnvoll. Auch in der praktischen Zeitreihenanalyse weist die Kolmogorov-Sinai Entropie den Weg zur Entropie des generierenden Prozesses unabhängig von einer bestimmten Wortlänge oder Partitionierung: Die Wortlänge sollte maximal gewählt werden, so dass möglichst viel Struktur von der Verteilung der Wörter erfasst wird. Und als Partitionierung ist diejenige zu wählen, die maximale Entropiewerte liefert, d. h. der Symbolsatz sollte möglichst viel Information von der Zeitreihe erhalten.

Bewertung von Dynamik:

Für einen binären Bernoulli-Prozess mit dem Zufälligkeitsparameter p gilt für die Metrische Entropie gemäß (29) und (35) unabhängig von der Wortlänge:

$$H_\mu(p) = -p \log_2 p - (1-p) \log_2 (1-p) \tag{39}$$

Bei dieser Funktion handelt es sich gleichzeitig um die Kolmogorov-Sinai Entropie, die Entropie des Bernoulli-Prozesses. Sie ist in Abb. 2-11 grafisch dargestellt.

Da die Metrische Entropie die Shannon-Entropie lediglich auf das Intervall $[0, \log_2 \lambda]$ skaliert, sei für die weitere Diskussion zur Bewertung von Dynamik auf die Shannon-Entropie verwiesen, die in diesem Zusammenhang mit der Rényi-Entropie in Abschnitt 2.5.2 besprochen wird. Abb. 2-14 (mittlere Linie, rechte Skala) stellt die Metrische Entropie für die logistische Abbildung dar.

2.5.4 Mittlerer Informationsgewinn

Ein besonders häufig verwendetes Maß für Information ist der mittlere Informationszuwachs oder -gewinn. Er misst die Information, die erforderlich ist, um einen Zustand *j* auszuwählen, wenn der vorangegangene Zustand *i* bekannt ist (WACKERBAUER et al., 1994). Bei einer Verteilung von *L*-Wörtern wird die bedingte Entropie als mittlere Ungewissheit des Symbols, das auf ein *L*-Wort folgt, berechnet (EBELING et al., 1998). Der Informationsgewinn durch eine zusätzliche Messung ist um so geringer, je besser diese aus den vorangegangenen Messungen vorhergesehen werden kann. Diese Eigenschaft wird durch den mittleren Informationsgewinn quantifiziert (siehe auch das Beispiel in der Einleitung 1.1).

Die Approximation durch relative Häufigkeiten der Wahrscheinlichkeit $p_{L,i\to j}$ für das L–1 Zeitschritte überlappende Auftreten eines *L*-Wortes *j* nach einem *L*-Wort *i* wurde bereits in Abschnitt 2.1.2 beschrieben. Neu an dem Wort *j* ist nur das letzte Symbol, das auf das Wort *i* folgt. Für diese Situation ist nach WACKERBAUER et al. (1994) der Informationsgewinn

$$G_{ij} = -\log_2 p_{L,i\to j} \tag{40}$$

Der mit den Häufigkeiten $p_{L,ij}$ für das Auftreten der Wörter *i* und *j* gewichtete Durchschnitt ist der mittlere Informationsgewinn:

$$H_{\mathrm{G}} = -\sum_{i,j=1}^{\lambda^L} p_{L,ij} \log_2 p_{L,i\to j} \tag{41}$$

Mit Gleichung (10), S. 23, ergibt sich, wie in Anhang 7.3 gezeigt, die zu (41) äquivalente Formel für den mittleren Informationsgewinn:

$$H_{\mathrm{G}}(L) = H_{\mathrm{S}}(L+1) - H_{\mathrm{S}}(L) \tag{42}$$

mit der Shannon-Entropie $H_{\mathrm{S}}(L)$ nach Gleichung (26) über die Verteilung der *L*-Wörter. Die Berechnung von H_{G} nach den beiden Formeln ist jedoch unterschiedlich: Für die Differenzen-Formel (42) werden zwei Bäume der Tiefe *L* und *L*+1 benötigt, während für die kompakte Formel (41) nur ein Baum der Tiefe *L*+1 erforderlich ist (siehe 2.1.2). Da die Berechnungszeit der „Shannon"-Maße im wesentlichen vom Aufbau der Bäume abhängt, liegt hierin ein numerischer Vorteil der Ein-Baum-Formel (41) gegenüber der Zwei-Baum-Formel (42). Dieser Vorteil hängt stark von der speziellen Anwendung ab. Bei der Berechnung von H_{G} für sukzessive höhere Wortlängen werden die Ergebnisse von H_{S} in (42) im nächsten Schritt wieder verwendet, so dass es kaum einen Zeitunterschied zwischen den beiden Formeln gibt. In anderen Fällen konnte ein Zeitvorteil von bis zu 66 % bei den verwendeten hydrologischen Zeitreihen festgestellt werden. Ein weiterer Vorteil von (41) ist eine höhere numerische Stabilität, da bei etwa gleich großen Summanden in der Differenz (42) Auslöschung eintreten kann (siehe: STOER, 1994, S. 8f), d. h. ein Ergebnis $H_{\mathrm{G}} \approx 0$ kann stark von Rundungsfehlern beeinflusst werden. Dieser Effekt dürfte jedoch kaum bedeutsam sein, da Entropien nahe 0 in

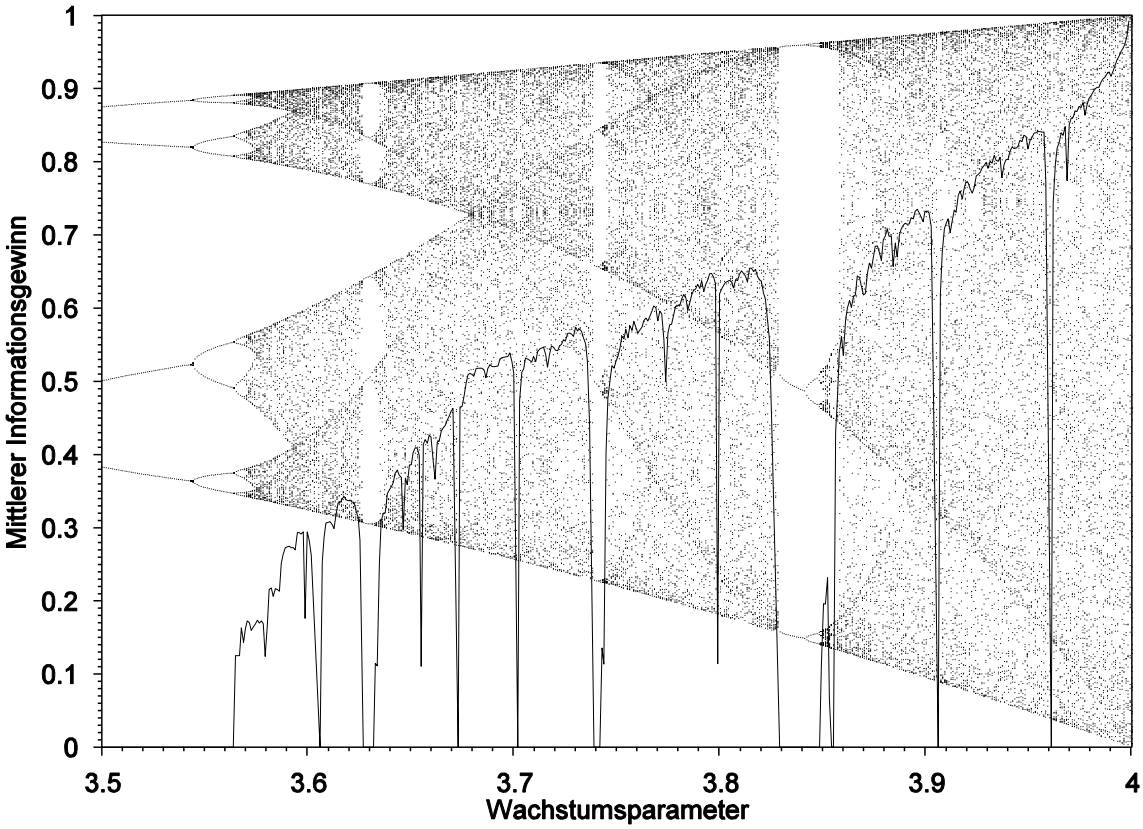

Abb. 2-15. Mittlerer Informationsgewinn für die logistische Abbildung (24). Je $N = 100000$ Daten nach 10000 Voriterationen, statische binäre Partitionierung bei 0.5, Wortlänge $L = 10$. Hintergrund: Attraktor nach Abb. 2-9.

der Praxis selten sind und üblicherweise (z. B. in SYMDYN) mit einer Genauigkeit von 10^{-16} gerechnet wird. Bei Wortlängen oberhalb des durch die begrenzte Datenmenge gerechtfertigten Bereichs (siehe 3.6) werden mit Formel (42) allerdings negative Werte für H_G ermittelt, weil die Shannon-Entropie hier mit zunehmender Wortlänge abnimmt. Formel (41) liefert dann $H_G = 0$. Die Übereinstimmung der beiden Formeln gilt, wenn man die endliche Datenlänge berücksichtigt, nur für kleine Wortlängen, solange die Statistik noch gesättigt ist. Bei einigen Autoren (z. B.: EBELING, 1996, EBELING et al., 1995, EBELING et al., 1996, RATEITSCHAK et al. 1995), die den mittleren Informationszuwachs auf einer Verteilung von Wörtern oder Blöcken berechnen — dies ist z. B. bei PRESS et al. (1992, S. 633) nicht der Fall — wird nur die Formel (42) als Unsicherheit oder bedingte Entropie angegeben, die elegant mit der Entropie pro Block (Shannon-Entropie H_S) definiert werden kann. Daher ist in SYMDYN die Berechnung von H_G nach beiden Formeln möglich.

Der mittlere Informationsgewinn ist wie die Metrische Entropie eine monoton fallende Funktion von der Wortlänge. Intuitiv ist klar, dass die Unsicherheit über ein zusätzliches Symbol (eine zusätzliche Messung) nicht zunehmen kann, wenn immer mehr Vorgänger bekannt sind (GRASSBERGER, 1986, S. 919). Da mit zunehmender Wortlänge bei endlicher Datenmenge das Verhältnis vorhandener zu möglichen Wörtern immer ungünstiger wird (siehe 3.6), bedeutet dies ab einer kritischen Wortlänge eine Annäherung an 0. Dieser Effekt ist bei H_G ausgeprägter als bei H_μ. Grundsätzlich sind die Werte von H_G stabiler im Vertrauensbereich und dann schneller bei 0 als dies bei H_μ der Fall ist. H_G konvergiert mit $L \to \infty$ ebenso wie H_μ theoretisch gegen die Entropie der Quelle (SHANNON, 1976, S. 66, siehe Abschnitt 0 unten), allerdings schneller als H_μ (CRUTCHFIELD, 1994a). Dies erklärt die bevor-

zugte Verwendung des Informationsgewinns H_G gegenüber den anderen Shannon-Informationsmaßen in der Literatur.

Wie gut der Wert von H_G bei der höchsten für die Datenlänge noch vertrauenswürdigen Wortlänge die Entropie der Quelle, d. h. des die Zeitreihe erzeugenden Prozesses, approximiert, ist nicht bekannt. Damit dies überhaupt möglich ist, muss die Zeitreihe lang genug sein, um das ganze potentielle Spektrum an Struktur, welches der Prozess erzeugen kann, erlebt und aufgezeichnet haben zu können (siehe 3.2).

Die Beziehung der metrischen H_μ zur bedingten Entropie H_G ist mit Gleichung (42) und $H_S(L{=}0) = 0$ leicht einzusehen und wurde bereits von SHANNON (1976, S. 66) festgestellt:

$$H_\mu(L) = \frac{1}{L}\sum_{l=0}^{L-1} H_G(l) \tag{43}$$

Bewertung von Dynamik:

Für einen binären Bernoulli-Prozess berechnet sich der mittlere Informationsgewinn nach (42) mit (29) zu derselben Funktion (39) wie die Metrische Entropie. Diese ist unabhängig von der Wortlänge und entspricht der Entropie (37) des Bernoulli-Prozesses (siehe Abb. 2-11). Der mittlere Informationsgewinn entspricht bei allen Markov Prozessen, dessen Ordnung unterhalb der Wortlänge liegt, der Entropie der Quelle (GRASSBERGER, 1986, S. 921, WACKERBAUER et al., 1994, S. 146).

Die Bewertung der logistischen Dynamik zeigt Abb. 2-15. Auch hier ist wieder prinzipiell dieselbe Zunahme des Maßes mit dem Wachstumsparameter (der Zufälligkeit) zu beobachten, wie bei Shannon-, Rényi- und Metrischer Entropie. Ein wichtiger Unterschied liegt in der Bewertung von Periodizität (Länge p): Die Metrische Entropie nimmt hier den Wert $(\log_2 p)/L$ an, der erst für sehr große Wortlängen und utopische Datenmengen nahe 0 ist. Der Informationsgewinn ist jedoch exakt 0, falls $p \le L$ ist. Also auch hier zeigt sich eine höhere Stabilität des Informationsgewinns gegenüber der Metrischen Entropie. Positive Werte von H_G links von den Feigenbaum-Punkten sind Artefakte der endlichen Wortlänge, die eine Auflösung längerer Perioden verhindert.

2.5.5 Mittlere wechselseitige Information

Die mittlere wechselseitige Information (engl.: mean mutual information) ist ein Sonderfall der Transinformation (siehe 2.2.2), die auf die Verteilung der L-Wörter und für einen Lag von einen Zeitschritt angewendet wird. Sie gibt an, wieviel Information im Mittel in der Abhängigkeit (Korreliertheit) von zwei aufeinanderfolgenden Wörtern enthalten ist, und wird nach WACKERBAUER et al. (1994) wie folgt definiert:

$$H_M = \sum_{i,j=1}^{\lambda^L} p_{L,ij} \log_2 \frac{p_{L,ij}}{p_{L,i}\, p_{L,j}} \tag{44}$$

Wie bereits in der Quelle angegeben und in Anhang 7.4 gezeigt ist (44) äquivalent zu einer Differenz von Shannon-Entropien:

$$H_M(L) = 2H_S(L) - H_S(L+1) \tag{45}$$

Zwei aufeinanderfolgende L-Wörter haben $L-1$ Symbole gemeinsam (siehe 2.1.2, insbes. Abb. 2-5). Die Korreliertheit von L-Wörtern mit $L > 1$ wird also vor allem durch diese

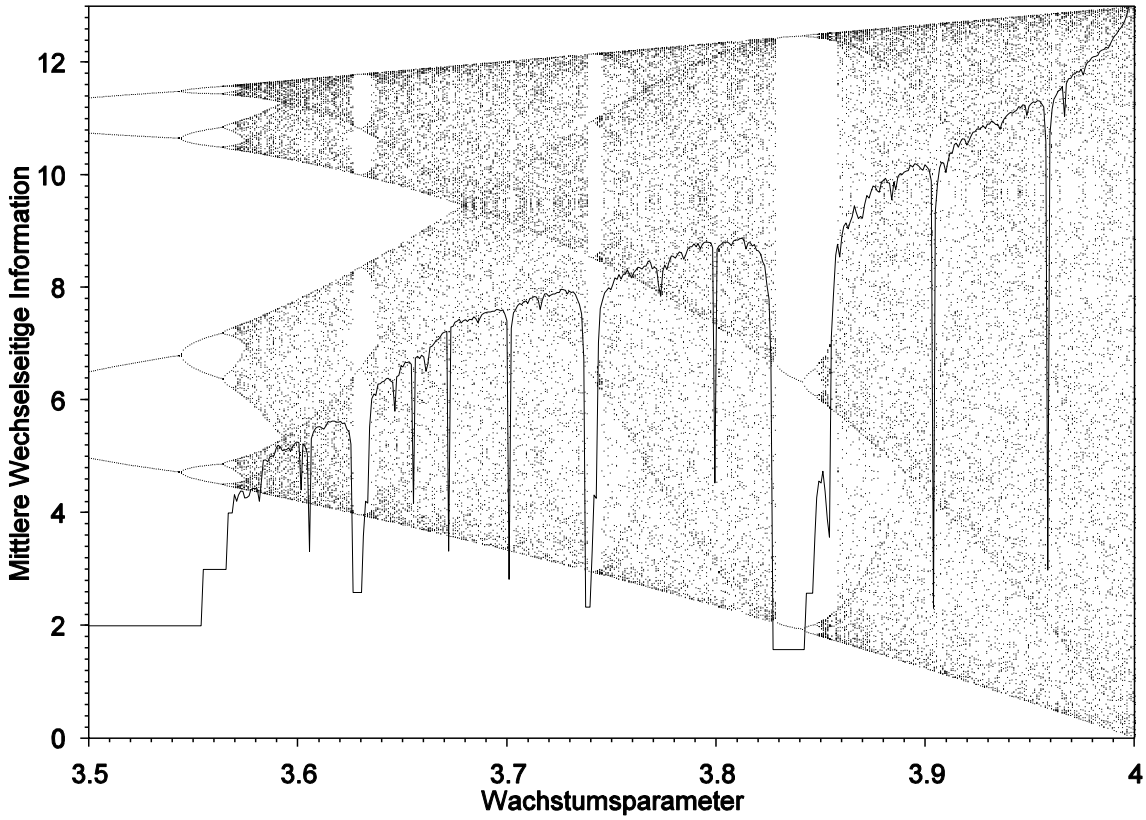

Abb. 2-16. Mittlere Wechselseitige Information für die logistische Abbildung (24). Je $N = 100000$ Daten nach 10000 Voriterationen, statische binäre Partitionierung bei 0.5, Wortlänge $L = 14$. Hintergrund Attraktor nach Abb. 2-9.

gemeinsamen $L - 1$ Symbole bestimmt. Besonders deutlich wird dies, wenn Gleichung (42) in (45) eingesetzt wird:

$$H_{\mathrm{M}}(L) = H_{\mathrm{S}}(L) - H_{\mathrm{G}}(L) \tag{46}$$

Die mittlere Information der Abhängigkeit zweier aufeinanderfolgender L-Wörter entspricht demnach der mittleren Information eines L-Wortes (Shannon-Entropie, H_{S}) vermindert um die mittlere Information eines zusätzlichen Symbols (mittlerer Informationsgewinn, H_{G}).

Die beiden äquivalenten Formeln (44) und (45) führen zu den gleichen numerischen Unterschieden in der Berechnung der mittleren wechselseitigen Information wie bei der Berechnung des mittleren Informationsgewinns nach (41) und (42) (siehe 2.5.4). Die kompakte Ein-Baum-Formel (44) war bei Testrechnungen mit hydrologischen Zeitreihen um maximal 75 % schneller als die Zwei-Baum-Differenzen-Formel (45).

Bewertung von Dynamik:

Periodizitäten der Länge p werden nach (46) und wegen $H_{\mathrm{G}} = 0$ (siehe 2.5.4) wie bei der Shannon-Entropie (siehe 2.5.1) durch $H_{\mathrm{M}} = \log_2 p$ bewertet (siehe WACKERBAUER et al., 1994).

Für einen binären Bernoulli-Prozess in Abhängigkeit von dem Zufälligkeitsparameter p berechnet sich die mittlere wechselseitige Information durch Einsetzen von (29) in (45) zu

$$H_{\mathrm{M}}(L, p) = -(L-1)\left(p \log_2 p + (1-p)\log_2(1-p)\right) = H_{\mathrm{S}}(L-1, p)\,, \tag{47}$$

siehe Abb. 2-11. H_M ist in diesem Fall also genau durch die Information der gemeinsamen $L - 1$ Symbole aufeinanderfolgender Wörter definiert (s. o.) oder kurz gesagt: durch die Shannon-Entropie der $(L - 1)$-Wörter. Auf der Symbol-Ebene ($L = 1$) äußert sich die Unabhängigkeit der Symbole des Bernoulli-Prozesses — wie zu erwarten war — durch $H_M = 0$. Es verbirgt sich also keine Information in der Korreliertheit der Symbole, weil sie beim Bernoulli-Prozess nicht korreliert sind.

Mit der Mittleren Wechselseitigen Information werden im Vergleich zu den bisher diskutierten Maßen keine neuen Muster in den Iterationen der logistischen Abbildung (siehe Abb. 2-16) entdeckt. Die Bewertung der Dynamik erfolgt auch auf die bereits von der Metrischen Entropie und vom Informationsgewinn bekannte Weise. Dies war wegen der Beziehung (46) auch nicht anders zu erwarten.

2.5.6 Algorithmische Information

Fast alle hier vorgestellten Maße für Information und Komplexität basieren auf einer Wahrscheinlichkeitsverteilung von L-Wörtern. Die Betrachung individueller Szenen eines Textes als Zufallsfolge empfand KOLMOGOROV (1965) als unnatürlich und begründete so den algorithmischen Ansatz zur Messung von Information. Als Maß dafür schlug er die Länge des kürzesten Computerprogramms zur Erzeugung eines Datensatzes vor. Er selbst bemerkte jedoch die Schwierigkeit — im allgemeinen Unmöglichkeit — diese Größe zu bestimmen. Der Grund dafür ist das Fehlen eines konkreten Bezugsystems (Basis) und die Unmöglichkeit der Einsicht in den erzeugenden Prozess (Endoperspektive). Durch die Beschränkung auf die Operationen „Einfügen“ und „Kopieren“, welche hintereinander ausgeführt werden, wird die algorithmische Information berechenbar. Dieser Vorschlag stammt von LEMPEL & ZIV (1976), die die Existenz eines absoluten Maßes für Information anzweifeln.

Durch diese Festlegung wird z. B. der algorithmische Informationsgehalt der ersten 1 000 000 Nachkommastellen der irrationalen Kreiszahl π maximal, da die Ziffern wie Zufallszahlen angeordnet sind und stets neue Teilfolgen „eingefügt“ werden müssen. Da aus diesem Grund auch jede kurze Teilfolge (L-Wort) gleichwahrscheinlich ist, ist auch die Entropie-Information maximal. Beide Informationskonzepte geben wie alle anderen Methoden in dieser Arbeit und in der beschreibenden Statistik nur eine äußere Beschreibung (Exoperspektive) des beobachteten Datensatzes. Im Kolmogorov’schen Sinne ist die Formulierung „die ersten 1 000 000 Nachkommastellen von π“ oder ein kurzes Programm, dass diese erzeugt, die kürzeste Beschreibung der Daten und damit ist der Kolmogorov-algorithmische „Informations“-gehalt klein. Dies verlangt jedoch ein viel größeres Bezugsystem, dass u. a. π erkennen kann, und ist eine Beschreibung aus der Innenperspektive des erzeugenden Prozesses. Abgesehen von der bereits erwähnten Unmöglichkeit einer solchen Definition, würde hierdurch keine Information, sondern eine Komplexität definiert.

Zur Definition der Algorithmischen Information muss zunächst die Anzahl der Komponenten eines Symbolsatzes bestimmt werden, mit denen dieser durch Einfügen neuer Symbole oder Kopieren bereits bekannter Sequenzen erzeugt werden kann. Diese Zahl wird Lempel-Ziv Komplexität C_{LZ} genannt (ZIV & LEMPEL, 1978). LEMPEL & ZIV (1976) setzten hier Komplexität mit Zufälligkeit gleich. Die Komponenten eines Symbolsatzes werden wie folgt ermittelt:

1. Das erste Symbol s_0 muss immer eingefügt werden und ist die erste Komponente.
2. Es sei s_{l-1} das letzte eingefügte Symbol, d. h. der erste Teil des Symbolsatzes (s_0, s_1, ..., s_{l-1}) ist bereits bekannt und in Komponenten zerlegt. Nun soll ein möglichst langer Teil (s_l,

s_{l+1}, ..., s_{l+k-1}) des restlichen Symbolsatzes (s_l, s_{l+1}, ..., s_{N-1}) aus dem bekannten kopiert werden. Dazu wird die längste Folge (s_i, s_{i+1}, ..., s_{i+k-1}) mit $i < l$, beginnend mit $i = 0$, ermittelt, die mit (s_l, s_{l+1}, ..., s_{l+k-1}) übereinstimmt. $i+k$ kann größer als l sein. Entweder ist dann $l+k-1 = N-1$, d. h. die letzte Komponente (s_l, s_{l+1}, ..., s_{N-1}) wurde ermittelt: Fertig. Oder das Symbol s_{l+k} kann nicht kopiert werden. Es muss eingefügt werden. Eine neue Komponente (s_l, s_{l+1}, ..., s_{l+k}) wurde gefunden. Damit ist s_{l+k} das letzte eingefügte Symbol. Falls $l+k < N-1$, geht es bei 2. weiter.

Die Bestimmung der Komponenten eines Symbolsatzes nach dieser Vorschrift ist relativ einfach und soll an einigen Beispielen illustriert werden:

Symbolsatz	Komponenten	N	C_{LZ}	I_A
0000000000000000	0·000000000000000	16	2	0.5
1010101010101010	1·0·10101010101010	16	3	0.75
0001101001000101	0·001·10·100·1000·101	16	6	1.5
1101001111010010	1·10·100·111·1010010·	16	5	1.25
1011001110000101	1·0·11·00·111·000·010·1	16	8	2
0·000001·11·0001111·00000110·11111·1111111110·0011111·11000000·00		56	10	1.04
1·10·0001·00000·111·000000·1001·11001·10000100001·1101·0100000·00000		57	12	1.23

Tabelle 2-1. Beispiele von Symbolsätzen der Länge N, ihrer Lempel-Ziv Komplexität C_{LZ} und algorithmischen Information I_A. 1. konstanter Symbolsatz, 2. periodischer Symbolsatz, 3. Bsp. aus LEMPEL & ZIV (1976), 4. Bsp. aus WACKERBAUER et al. (1994), 5. „zufälliger" Symbolsatz, 6. statisch partitionierter Datensatz von Abb. 2-1, 7. dynamisch partitionierter Datensatz von Abb. 2-2.

KASPAR & SCHUSTER (1987) geben zur Beschreibung des Lempel-Ziv Algorithmus ein Flussdiagramm an. Die Implementierung in SYMDYN zeigt der nachfolgende Pseudo-C-Quellcode. Darin werden Lückensymbole $a*$ als kopierbar interpretiert, weil sie die Lempel-Ziv Komplexität C_{LZ} nicht (künstlich) erhöhen sollen[3].

```
C_LZ ← 1
l ← 1
k ← 0
WHILE ( l < N )
{
        k_max ← 0
        i ← 0
        WHILE ( i < l  AND  l+k < N )
        {
                k ← 0
                WHILE ( l+k < N  AND  ( s_{i+k} = s_{l+k}  OR  s_{i+k} = a*  OR  s_{l+k} = a* ) )
                        k ← k+1
                IF ( k > k_max ) THEN k_max ← k
                i ← i+1
        }
        C_LZ ← C_LZ+1
        l ← l+k_max+1
}
```

[3] Bei einer großen Anzahl von Lücken, die sich relevant auf C_{LZ} auswirkt, wäre eine Abschätzung der Zunahme von C_{LZ} bei der entsprechenden Anzahl von Daten und eine anschließende Korrektur vorstellbar. Dies ist jedoch fehlerträchtig, so dass von einer Auswertung der Daten eher abzuraten ist.

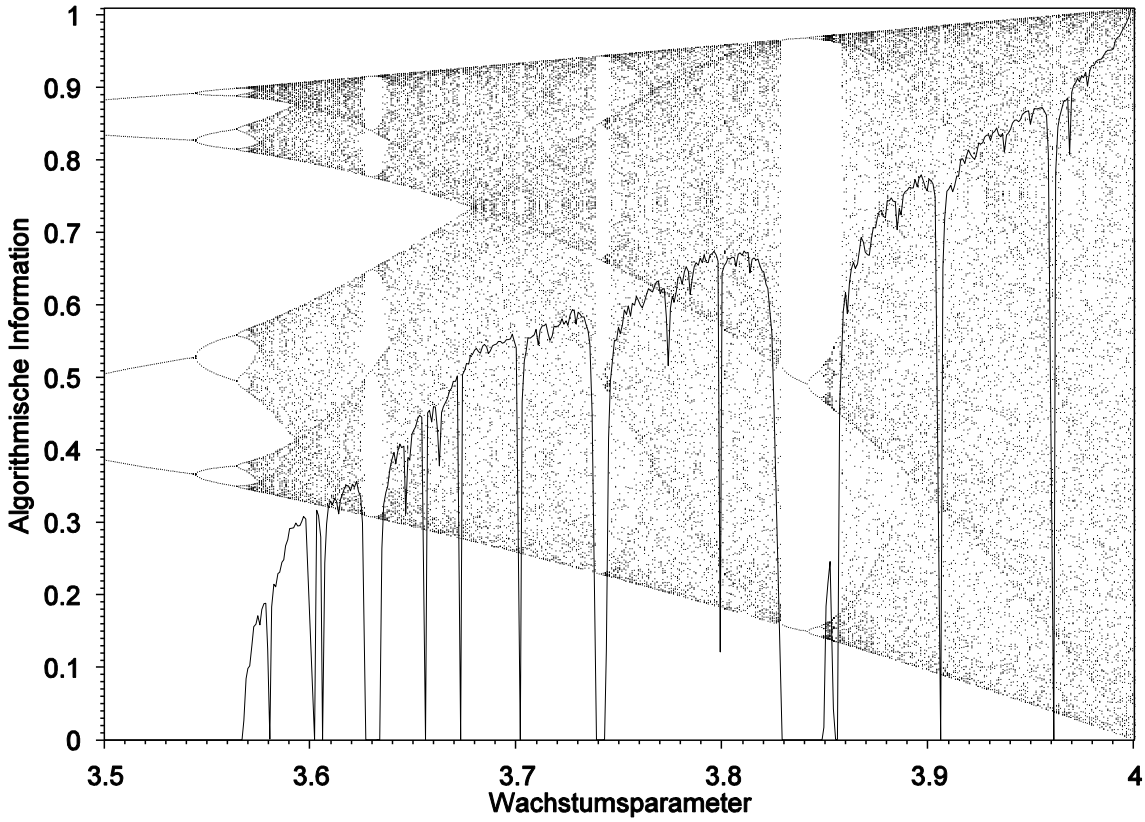

Abb. 2-17. Algorithmische Information für die logistische Abbildung (24). Je N = 100000 Daten nach 10000 Voriterationen, statische binäre Partitionierung bei 0.5. Hintergrund: Attraktor nach Abb. 2-9.

Die Lempel-Ziv Komplexität C_{LZ} ist eine monoton mit der Anzahl N der Daten steigende Größe, falls diese nicht periodisch sind. C_{LZ} ist bei zufälliger Anordnung der Symbole maximal und nähert sich mit zunehmendem N dem Grenzwert (LEMPEL & ZIV, 1976):

$$\frac{N}{\log_\lambda N} \tag{48}$$

wobei λ die Alphabetgröße des Symbolsatzes ist. Die algorithmische Information I_A wird dann definiert als (KASPAR & SCHUSTER, 1987):

$$I_A = \frac{C_{LZ} \log_\lambda N}{N} \tag{49}$$

Bei hinreichend großer Datenmenge N ist $I_A \in [0,1]$. In den Beispielen aus Tabelle 2-1 ist N zu klein, so dass bei nicht-periodischen Daten stets Werte größer als 1 erreicht werden. Für eine Abweichung vom theoretischen Grenzwert um weniger als 5 % haben KASPAR & SCHUSTER (1987) eine minimale Datenlänge von N = 1000 ermittelt.

Für hinreichend lange Datensätze einer ergodischen Quelle (siehe 3.2) mit Entropie h (37) stellen LEMPEL & ZIV (1976) den Zusammenhang

$$I_A \leq h \tag{50}$$

mit ihrem Informationsmaß fest. Die enge Beziehung zwischen dem (wahrscheinlichkeits-) maßtheoretischen und algorithmischen Informationsmaß ist nach GRASSBERGER (1986, S. 937) nur bei instationären Daten (siehe 3.3) gefährdet: Das Konzept der Kolmogorov Komplexität liefert kein Wahrscheinlichkeitsmaß, aber es induziert ein solches Maß (CHAITIN nach GRASSBERGER, 1986, S. 937).

Mit der Lempel-Ziv Komplexität hängt, wie von ZIV & LEMPEL (1978) gezeigt, die Kompressibilität eines Datensatzes zusammen, die mit der Entropie der Quelle gleich gesetzt werden kann. PÖSCHEL (1996) schlägt die Messung der Kompressibilität eines Datensatzes als Abschätzung für die Entropie mittels kommerzieller Kompressionsprogramme, z. B. „gzip“, vor. WOLFRAM (1985) definiert den effektiven Informationsgehalt einer Folge als Länge ihrer kürzesten Spezifizierung durch alle (in polynomialer Zeit) möglichen Berechnungen.

Weitere Hinweise zur Algorithmischen Information — auch Algorithmische Komplexität genannt — finden sich bei KURTHS et al. (1996), KURTHS & WITT (1994), WITT et al. (1994).

Bewertung von Dynamik:

Bei streng periodischen Daten erreicht die Lempel-Ziv-Komplexität einen festen endlichen Wert der durch den Aufbau einer Saison bestimmt ist. Eine Obergrenze dafür kann nach (48) mit $N \gg p$ (Periodenlänge) bestimmt werden. Wegen der Normierung für I_A in (49) bedeutet dies, dass für $N \gg p$ die Algorithmische Information in etwa verschwindet. Es gilt $I_A = 0$ für $N \to \infty$ (WACKERBAUER et al., 1994). Ein großer Vorteil der Algorithmischen Information gegenüber den Wort-Entropien ist, dass die Entdeckung von Periodizitäten — und mittelreichweitigen Korrelationen im Allgemeinen — nicht durch eine kurze Wortlänge beschränkt ist. Leider kann eine Periodizität durch $I_A \approx 0$ aber erst bei hinreichend langen Datensätzen erkannt werden (s. o. und erste Beispiele in Tabelle 2-1).

Eine Formel für die Algorithmische Information des Bernoulli-Prozesses konnte nicht hergeleitet werden. Die empirischen Berechnungen in Abb. 2-11 zeigen jedoch eine sehr gute Übereinstimmung mit den Entropie-Formeln für den Bernoulli-Prozess. Für diesen Fall wird also der theoretisch bereits von LEMPEL & ZIV (1976) erkannte Zusammenhang zu den Entropien (s. o.) bestätigt.

Auch die Algorithmische Information liefert keine neue Information über die logistische Abbildung wie Abb. 2-17 zeigt. Der Verlauf von I_A stimmt gut mit dem der zuvor betrachteten Informationsmaße überein. Dies bestätigt den bereits theoretisch festgestellten Zusammenhang mit der Entropie der Quelle (s. o.), die offenbar schon gut durch die Metrische Entropie (Abb. 2-14) und besonders den Informationsgewinn (Abb. 2-15) approximiert wird.

Das algorithmische Konzept von Information unterscheidet sich in seiner Definition grundlegend von dem der Entropie. Trotzdem führen beide Konzepte zu einem einheitlichen Maß für Information. Diese Feststellung wurde auch mit experimentellen Zeitreihen aus Kapitel 4 bestätigt.

2.6 Maße für Komplexität

2.6.1 Effektive Maßkomplexität

Wie bereits in Abschnitt 2.5.4 erwähnt, nimmt die Unsicherheit über eine zusätzliche Messung höchstens ab, je mehr vorangegangene Messungen bekannt sind. Anders ausgedrückt: Der mittlere Informationsgewinn, H_G nach Gleichung (41) oder (42), nimmt mit zunehmender Wortlänge ab. Der Betrag dieser Abnahme ist im Mittel gerade:

$$\delta h_L = H_G(L-1) - H_G(L) \tag{51}$$

Diese Informationsmenge muss wenigstens gespeichert werden, um die zusätzliche Messung, d. h. das zusätzliche Symbol, optimal vorherzusagen (GRASSBERGER, 1986, S. 918). Sie wird nach der Beobachtung des Symbols nicht mehr benötigt. Für jeden Zeitschritt und jede Wortlänge L muss zumindest eine Informationsmenge von δh_L über L Zeitschritte gespeichert werden. Die minimale Gesamtmenge an Information, die gespeichert werden muss, um zu jeder Zeit eine optimale Vorhersage zu liefern ist also

$$C_{\mathrm{EM}} = \sum_{L=1}^{\infty} L\,\delta h_L = \sum_{L=1}^{\infty} L\left(H_{\mathrm{G}}(L-1) - H_{\mathrm{G}}(L)\right) \tag{52}$$

Diese Größe wurde von GRASSBERGER (1986, S. 920) als Effektive Maßkomplexität (EMC) eingeführt. Hierzu wird — wie bereits zu Gleichung (43) bemerkt — aus Plausibilität $H_{\mathrm{S}}(L{=}0) = 0$ definiert. Gleichung (52) kann nach GRASSBERGER (1986, S. 920) auch mittels der Quellenentropie, h (37), die er als Grenzwert von H_{G} definiert, formuliert werden:

$$C_{\mathrm{EM}} = \sum_{L=0}^{\infty} \left(H_{\mathrm{G}}(L) - h\right) \tag{53}$$

Einsetzten von (42) in (53) ergibt eine Teleskopsumme und (s. WACKERBAUER et al., 1994):

$$C_{\mathrm{EM}} = \lim_{L\to\infty} \left(H_{\mathrm{S}}(L) - Lh\right) \tag{54}$$

Die Effektive Maßkomplexität ist nach GRASSBERGER (1986) eine untere Abschätzung für die tatsächliche Maßkomplexität (TMC), welche die Menge der tatsächlich zur Vorhersage benötigten Information angibt. Für eine optimale Auswahl muss die Information sauber codiert werden, was eine zusätzliche Speicherung von Information erfordert. TMC ist eine Obergrenze für die Mengen Komplexität (SC), die als Shannon-Entropie der Knotenbesuche eines minimalen Automaten definiert ist (GRASSBERGER, 1986), und stimmt sogar in einfachen Fällen mit dieser überein. SC wird von der Algorithmischen Komplexität (AC) nach oben bregrenzt, die — analog zur topologischen Entropie (31) — dem Logarithmus der Anzahl der Knoten des Automatengraphen (vgl. Abb. 2-8) entspricht. Dies ergibt eine Verbindung zu den Automaten-Komplexitäten, wozu insbesondere die in Abschnitt 2.6.4 beschriebene ε-Komplexität nach CRUTCHFIELD & YOUNG (1989), CRUTCHFIELD (1992, 1994a, 1994b) gehört.

Die Effektive Maßkomplexität erschien GRASSBERGER (1986, S. 936) als das einzige beobachtbare (berechenbare) Maß unter den im letzten Abschnitt genannten Methoden, wenn die Grammatik des Prozesses unbekannt ist. Davon ist bei Messdaten in der Praxis — insbesondere in der Ökosystemforschung — auszugehen. Damit ist auch Crutchfields Idee einer ε-Komplexität (siehe 2.6.4) prinzipiell zum Scheitern in der Praxis verurteilt, wenn diese nur auf einer bestimmten Automatenklasse, z. B. den in Abschnitt 2.1.3 beschriebenen endlichen stochastischen Automaten, basiert. Abhilfe könnte hierbei eine hierarchische Maschinen-Rekonstruktion über alle Automatenklassen schaffen. Dies scheiterte in dieser Arbeit jedoch bereits an der gerade genannten, niedrigen Automatenklasse.

Während die Shannon-Entropie (als H_μ oder H_{G}) die Informationsmenge pro Symbol zur Spezifikation eines Strings misst, gibt die Effektive Maßkomplexität die Information pro Symbol an, die benötigt wird, um zu garantieren, dass ein String zu der Gesamtheit dazugehört, ohne diesen näher zu spezifizieren (GRASSBERGER, 1986, S. 936).

Nach WACKERBAUER et al. (1994) beschreibt die Effektive Maßkomplexität das Konvergenzverhalten der lokalen Steigungen, d. h. des Informationsgewinns (42), gegen die Entropie h des Prozesses: Kleine Werte von C_{EM} bedeuten eine schnelle Konvergenz; große Werte eine langsame Konvergenz.

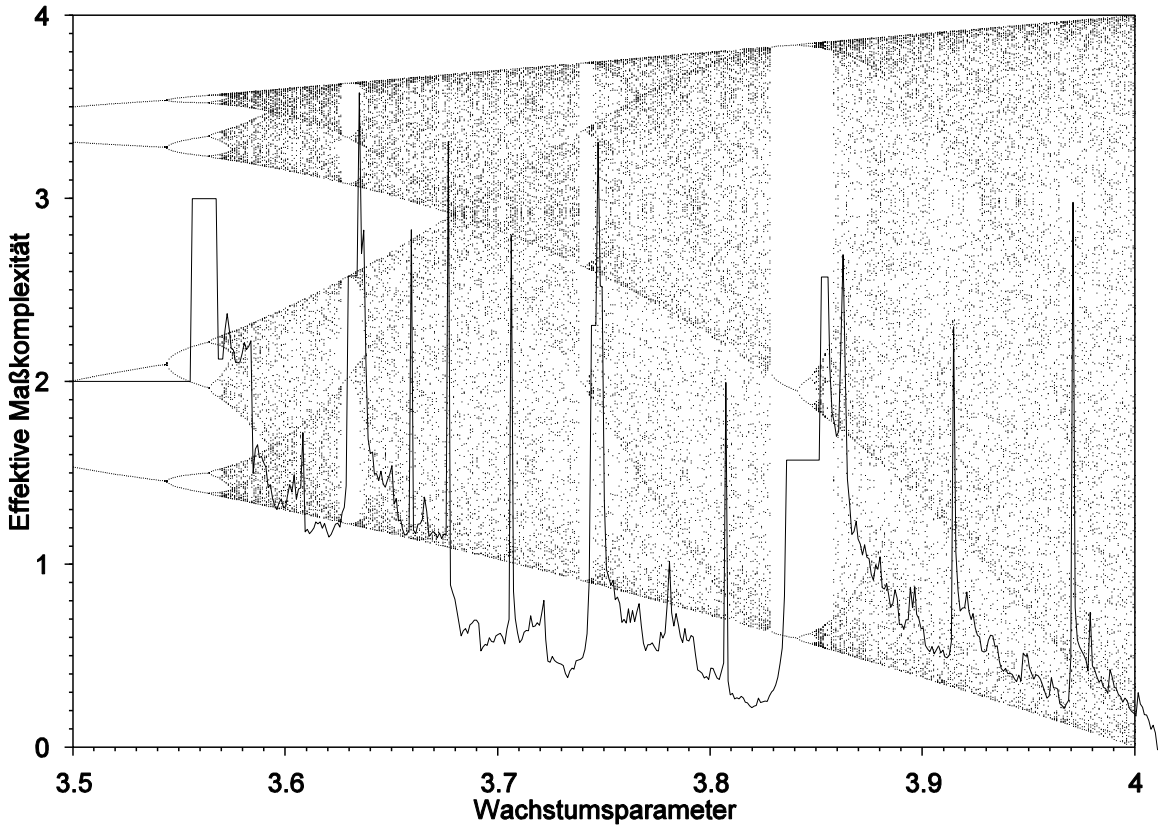

Abb. 2-18. Effektive Maßkomplexität für die logistische Abbildung (24). Je N = 100000 Daten nach 10000 Voriterationen, statische binäre Partitionierung bei 0.5, Wortlänge L = 7. Hintergrund: Attraktor, Abb. 2-9.

Bei der praktischen Berechnung der Effektiven Maßkomplexität können keine echten Grenzwerte, wie in (52), (53) oder (54) vorhanden, ermittelt werden. Statt dessen wird die Summe in (52) nur bis zu einer der Datenmenge entsprechend maximalen Wortlänge L ausgewertet (siehe 3.6). Dies liefert nach Anhang 7.7

$$C_{\mathrm{EM}} \approx (L+1)H_{\mathrm{S}}(L) - LH_{\mathrm{S}}(L+1) \tag{55}$$

mit der Shannon-Entropie H_{S} nach (26). Diese Formel ist, wie ebenfalls in Anhang 7.7 gezeigt, äquivalent zu:

$$C_{\mathrm{EM}} \approx \sum_{i,j=1}^{\lambda^L} p_{L,ij} \log_2 \frac{p_{L,i\to j}^{L}}{p_{L,i}} \tag{56}$$

mit den L-Wort-Häufigkeiten $p_{L,\ldots}$ (siehe 2.1.2). In SYMDYN kann die Effektive Maßkomplexität nach beiden Formeln berechnet werden. Die Zwei-Baum-Differenzen-Formel (55) steht der Definition (52) näher. Die kompakte Ein-Baum-Formel (56) ist numerisch stabiler und wird schneller berechnet. Dieser Unterschied wurde bereits in Abschnitt 2.5.4 am Beispiel von H_{G} ausführlich diskutiert. Die Übereinstimmung der Formeln (55) und (56) konnte auch empirisch für unterkritische Wortlängen gezeigt werden. Mit (56) konnte ein Zeitvorteil von bis zu 66 % gegenüber (55) für hydrologische Daten beobachtet werden.

Anwendungen der Effektiven Maßkomplexität auf die logistische Funktion finden sich bei GRASSBERGER (1986) (auch auf zelluläre Automaten), RATEITSCHAK et al. (1995) und WACKERBAUER et al. (1994).

Bewertung von Dynamik:

Für streng periodische Daten mit Periode p ist $C_{EM} = \log_2 p$ (WACKERBAUER et al., 1994), wie mit (55) und H_S (siehe 2.5.1) einzusehen ist, falls $L \leq p$ gilt. Das Erkennen von Periodizitäten hängt also auch hier — wie bei den Wort-Entropien und allen anderen Wort-Maßen — davon ab, ob die Periode unterhalb der Wortlänge liegt.

Für einen Bernoulli-Prozess ist der Informationsgewinn $H_G = H_\mu$ nach Gleichung (39) unabhängig von der Wortlänge L. Damit ist $\delta h_L = 0$ nach Gleichung (51) und $C_{EM} = 0$ nach Gleichung (52): Die Effektive Maßkomplexität verschwindet.

Die Effektive Maßkomplexität der logistischen Abbildung ist geprägt von den stufigen periodischen Bereichen, die in ein Maximum am Akkumulationspunkt übergehen, siehe Abb. 2-18. In den chaotischen Bereichen werden deutlich kleinere Werte erreicht, die mit Verbreiterung der Chaosbänder tendenziell abnehmen. Diese Eigenschaft ist bei den Informationsmaßen genau gegenläufig. Zu Beginn der periodischen Fenster steigt C_{EM} bereits leicht an. Damit ist die Effektive Maßkomplexität ein Maß für Ordnung (speziell: Periodizität), das für unendliche Perioden- und Wortlänge theoretisch singulär ist. Dies wird jedoch aufgrund der endlichen Daten- und Wortlänge nie beobachtet werden. C_{EM} stellt nach den Tabellen in Abschnitt 7.9 die strengsten Anforderungen an die Wortlänge.

2.6.2 Fluktuationskomplexität

In Abschnitt 2.5.4 wurde der Informationsgewinn G_{ij} (40) für den Übergang eines L-Wortes i zu einem L-Wort j durch Anhängen eines zusätzlichen Symbols an Wort i und Streichen seines ersten Symbols erklärt. Analog kann der Informationsverlust L_{ij} beim Übergang von Wort j nach i durch Voranstellen eines Symbols und Streichen des letzten Symbols erklärt werden:

$$L_{ij} = -\log_2 p_{L,i\leftarrow j} \tag{57}$$

Der Netto-Informationsgewinn bezüglich der Übergänge zwischen den Wörtern i und j ist dann mit Gleichung (10), die analog für $p_{L,i\leftarrow j}$ gilt:

$$\Gamma_{ij} = G_{ij} - L_{ij} = \log_2 \frac{p_{L,i\leftarrow j}}{p_{L,i\rightarrow j}} = \log_2 \frac{p_{L,i}}{p_{L,j}} \tag{58}$$

Der Mittelwert des Netto-Informationsgewinns über alle Paare i und j ist:

$$\langle \Gamma \rangle = \sum_{i,j=1}^{\lambda^L} p_{L,ij} \Gamma_{ij} \tag{59}$$

Wie leicht einzusehen ist und in Anhang 7.5 gezeigt wird, gilt $\langle \Gamma \rangle \equiv 0$. Informationsgewinn und -verlust sind also auf die gesamte Verteilung der L-Wörter gemittelt gleich. Dies gilt jedoch nicht generell für die Varianz:

$$C_\Gamma = \sigma_\Gamma^2 = \sum_{i,j=1}^{\lambda^L} p_{L,ij} \left(\log_2 \frac{p_{L,i}}{p_{L,j}} \right)^2 \tag{60}$$

Diese Größe wurde von BATES & SHEPARD (1993) zur Beurteilung der Dynamik anhand der Informations-Fluktuation von nicht-deterministischen finiten Automaten am Beispiel von zellulären Automaten eingeführt. Sie ist als Fluktuationskomplexität jedoch auch auf beliebi-

ge andere Dynamische Systeme anwendbar, wie von WACKERBAUER et al. (1994) angeführt. Nach BATES & SHEPARD (1993) erfasst C_Γ die zeitliche Fluktuation in der relativen Dominanz von Chaos oder Ordnung des Systems. Positive Werte kennzeichnen eine Netto-Speicherung von externer Information. Bei Systemen mit maximaler Entropie (Zufallsprozesse) gilt stets $G_{ij} = L_{ij}$ und deshalb $C_\Gamma = 0$. Bei Systemen mit verschwindender Entropie H_G (periodische Prozesse, dessen Periodenlänge unterhalb der Wortlänge liegt) ist wegen $G_{ij} = L_{ij} = 0$ stets $C_\Gamma = 0$. Damit sind die Randbedingungen an ein Maß zweiter Ordnung (Komplexitätsmaß), wie in Abschnitt 2.4 formuliert, erfüllt. Komplexe Systeme zeichnen sich im Sinne dieses Maßes durch eine hohe zeitliche Schwankung der Netto-Information oder eine hohe Netto-Informationsspeicherungskapazität aus. Diese Interpretation bezieht sich hier auf die sukzessiven Übergänge von Teilfolgen der Länge L des partitionierten Datensatzes ineinander, d. h. der Verteilung der L-Wörter. Da die Fluktuationskomplexität über Zustandswahrscheinlichkeiten und Übergangswahrscheinlichkeiten berechnet wird und nicht wie etwa beim mittleren Informationsgewinn (siehe 2.5.4) auf Zustandswahrscheinlichkeiten alleine reduziert werden kann, wird C_Γ von WACKERBAUER et al. (1994) als dynamisches Maß klassifiziert.

Da die Partitionierung häufig nach dem Prinzip der Informationsmaximierung gewählt wird, sind die Symbole oft gleichhäufig. In diesem Fall verschwindet C_Γ bei Wortlänge $L = 1$ für alle Zeitreihen, wie mit (60) leicht einzusehen ist. C_Γ sollte daher nur für $L \geq 2$ ausgewertet werden. Charakteristische Werte werden bereits bei $L = 2$ erreicht, wie empirische Untersuchungen an Messreihen zeigen.

Bewertung von Dynamik:

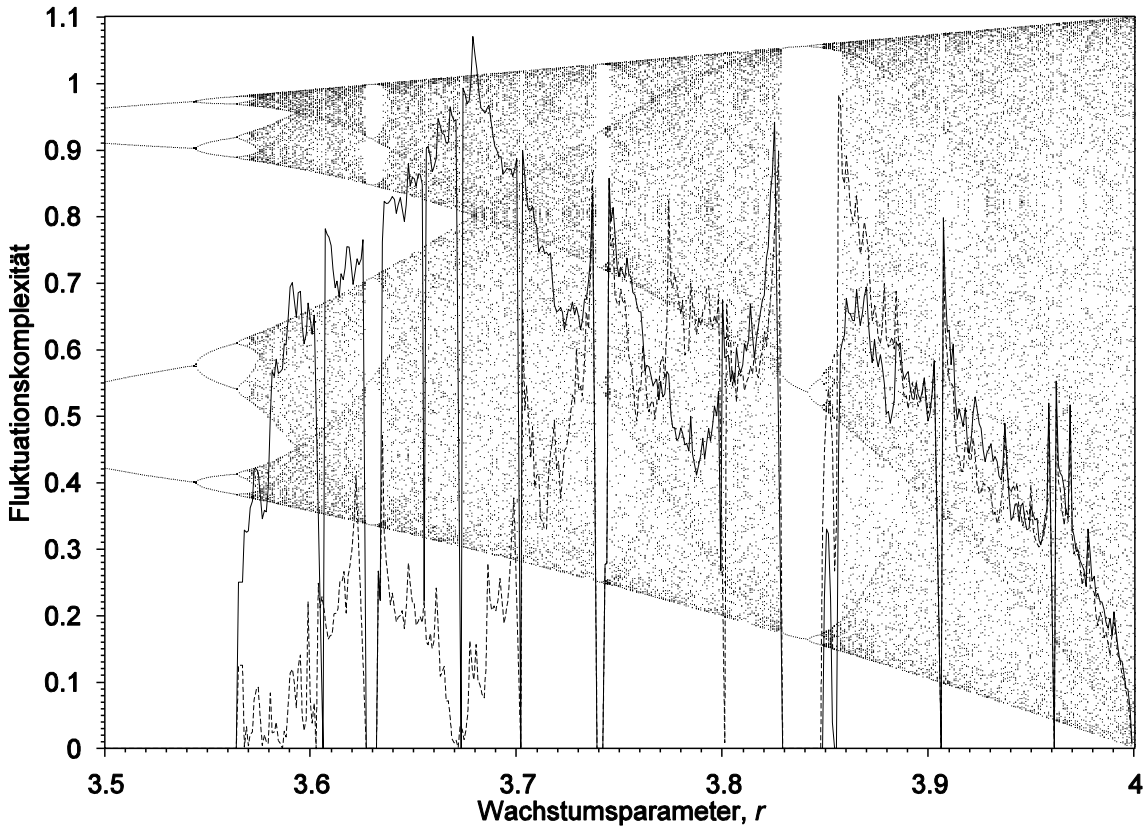

Abb. 2-19. Fluktuationskomplexität für die logistische Abbildung (24). Je $N = 100000$ Daten nach 10000 Voriterationen, statische binäre Partitionierung bei 0.5, Wortlängen $L = 8$ (gestrichelte Linie) und $L = 9$ (durchgezogene Linie). Hintergrund: Attraktor, Abb. 2-9.

Die Fluktuationskomplexität eines binären Bernoulli-Prozesses berechnet sich, wie in Anhang 7.6 gezeigt, zu

$$C_\Gamma(p) = 2p(1-p)\left(\log_2 \frac{p}{1-p}\right)^2 \tag{61}$$

in Abhängigkeit vom Zufallsparameter p unabhängig von der Wortlänge. Sie erreicht ihr Maximum bei einer Zufälligkeit von 16.6 % ($p \approx 0.0832217202$, siehe Anhang 7.6) und einem Wert von 1.8283945658. Diese Funktion ist in Abb. 2-20 dargestellt und veranschaulicht die Bewertung von Zufälligkeit durch Komplexität.

Bei der Betrachtung der Fluktuationskomplexität für die logistische Abbildung in Abb. 2-19 fällt sofort der alternierende Kurvenverlauf für gerade und ungerade Wortlängen auf. Dies gilt in gleicher Weise auch für die nicht in Abb. 2-19 dargestellten Wortlängen. Insbesondere ist C_Γ für $r < 3.75$ deutlich größer bei ungeraden Wortlängen in Vergleich zu geraden Wortlängen. Dies ist ein Indiz für Bandverschmelzung und wird von WACKERBAUER et al. (1994) erklärt. Besonders deutlich wird dies bei der Betrachtung der Differenzfunktion, etwa $C_\Gamma(L=9) - C_\Gamma(L=8)$, die am Bandverschmelzungspunkt $r = 3.6785$ ein globales Maximum aufweist.

Bei Periodizität verschwindet die Fluktuationskomplexität, falls die Periode im Bereich der Wortlänge liegt. Die periodischen Fenster im Attraktor der logistischen Abbildung werden von Spitzenwerten der Fluktuationskomplexität umrahmt, die eine Sensitivität auf Intermittenz und innere Krisen zeigen (siehe 2.3.3, Abb. 2-19 und WACKERBAUER et al., 1994). Die bereits beim Bernoulli-Prozess beobachtete Abnahme von C_Γ bei zunehmender hoher Zufälligkeit ist auch bei der logistischen Abbildung zu erkennen. Die Fluktuationskomplexität nimmt also insgesamt hohe Werte bei komplexer (Änderung der) Dynamik unterschiedlicher Art an.

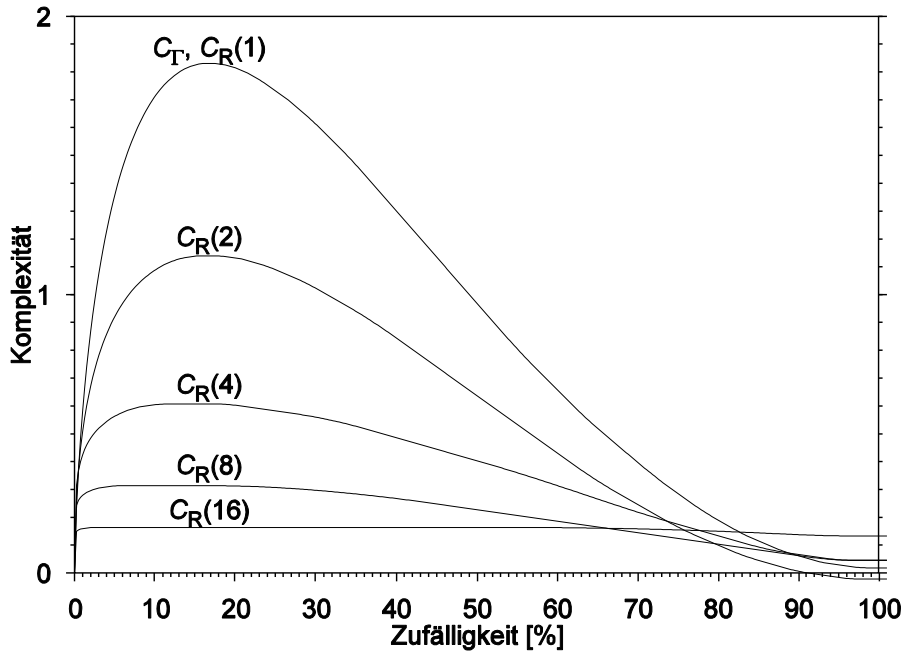

Abb. 2-20. Komplexität und Zufälligkeit. Fluktuationskomplexität C_Γ nach (61) und Rényi-Komplexität $C_R(\alpha)$ nach (62) mit (34) für den Bernoulli-Prozess. Zufälligkeit = p·200 % (siehe 2.3.2).

2.6.3 Rényi-Komplexität

Nach einer mündlichen Empfehlung von Jürgen Kurths können die Differenzen von Rényi-Entropien (siehe 2.5.2) konjugierter Ordnungen, also $H_R(1/\alpha) - H_R(\alpha)$ für $\alpha > 1$, als Maß für Komplexität verwendet werden. Dabei sollte für α ein Wert von 3 oder 4 gewählt werden. Anhand von Abb. 2-12 (und Abb. 2-13) ist leicht vorstellbar, dass eine solche Definition die Randbedingungen an ein Komplexitätsmaß erfüllt. Komplexität bedeutet hier eine hohe Vielfalt in den Häufigkeiten der Wörter, eine in etwa gleiche Anzahl von seltenen wie häufigen Wörtern. Denn: Die Ordnung $1/\alpha$ in H_R (α-te Wurzel) gewichtet kleinere Wahrscheinlichkeiten stärker als größere, falls $\alpha > 1$. Bei der Ordnung α ist es umgekehrt.

Die Rényi-Komplexität $C_R(\alpha)$ der Ordnung $\alpha > 1$ einer Verteilung von L-Wörtern sei:

$$C_R(\alpha) = \frac{2}{(\alpha-1)L\ln 2}\left(H_R(\tfrac{1}{\alpha}) - H_R(\alpha)\right) \tag{62}$$

Die Rényi-Komplexität C_R sei dann:

$$C_R = \lim_{\alpha\to 1,\,\alpha>1} C_R(\alpha) \tag{63}$$

Diese Definitionen beruhen auf folgenden Feststellungen:

- Für den binären Bernoulli-Prozess, also mit $H_R = H_R(p)$ nach Gleichung (34) und $C_\Gamma = C_\Gamma(p)$ nach (61), ist $C_R \equiv C_\Gamma$. Der Beweis erfolgte durch Nachrechnen mit der Funktion „Limit“ in Mathematica 3.0“ (WOLFRAM, 1996, S. 834). Die Rényi-Komplexität der Ordnung α nach (62) konvergiert also für Bernoulli-Prozesse gegen die Fluktuations-

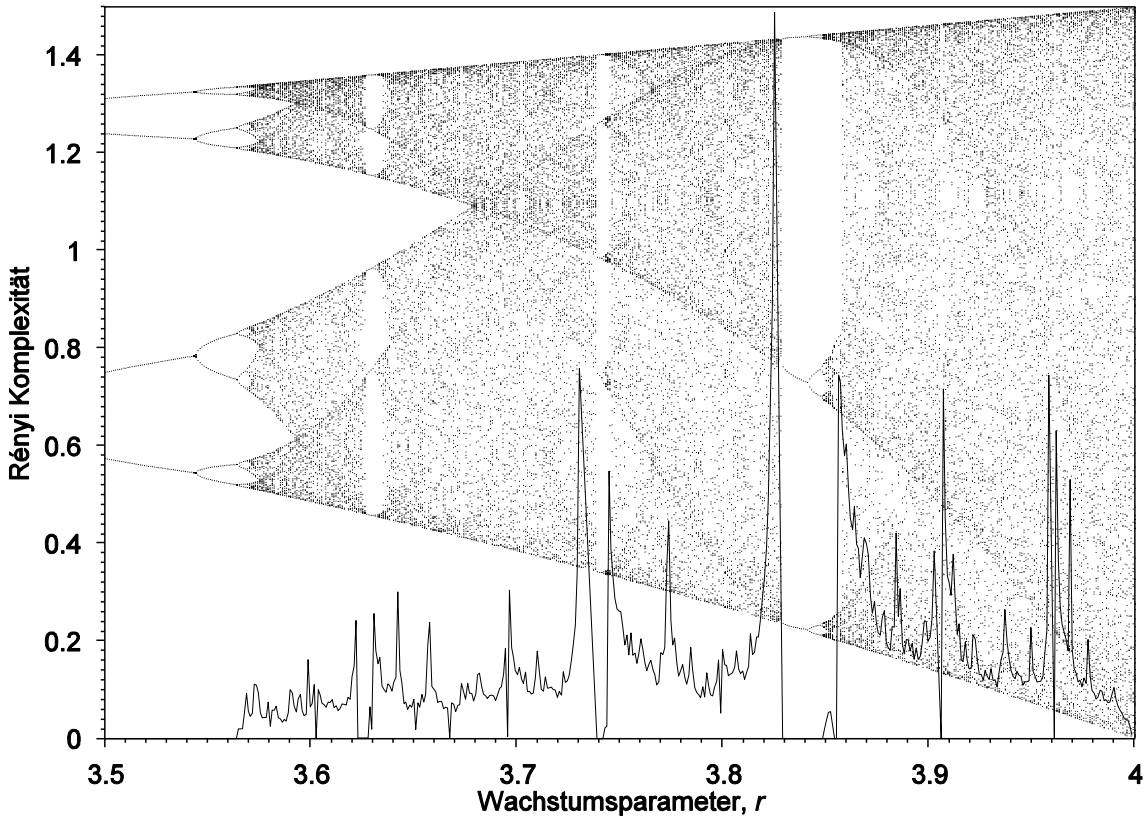

Abb. 2-21. Rényi-Komplexität für die logistische Abbildung (24). Je N = 100000 Daten nach 10000 Voriterationen, statische binäre Partitionierung bei 0.5, Wortlänge L = 11. Hintergrund: Attraktor, Abb. 2-9.

komplexität mit $\alpha \to 1$ (siehe auch Abb. 2-20). Die Rényi-Komplexität ist für diesen Fall eine Verallgemeinerung der Fluktuationskomplexität.

- $C_R(\alpha)$ mit $\alpha > 1$ erreicht für Bernoulli-Prozesse ab $p = 0$ sehr schnell ein bestimmtes Komplexitätsniveau (siehe Abb. 2-20). Mit zunehmendem α wird der Abstand von diesem Niveau zum Komplexitätsmaximum geringer und die Position des Maximums wird schon bei hohem Determinismus erreicht. Deterministische Prozesse werden in ihrer Komplexität gegenüber zufälligeren Prozessen damit deutlich überbewertet. Im Extremfall $\alpha \to \infty$ wird das Maximum sofort erreicht: Von einem Komplexitätsmaß kann dann keine Rede mehr sein. C_Γ ist also nur für $\alpha \approx 1$ spezifisch im Sinne eines Maßes für Komplexität.
- Ohne den Faktor $1/(\alpha-1)$ konvergiert $C_R(\alpha)$ gegen konstant 0 mit $\alpha \to 1$. Der Faktor $1/L$ normiert gegen die Wortlänge L.
- Im Gegensatz zum Bernoulli-Prozess unterscheidet sich $C_R(\alpha)$ für alle α quantitativ von C_Γ bei der logistischen Abbildung (vgl. Abb. 2-21). Die Definition der Rényi-Komplexität führt also zu einer neuen Bewertung von Dynamik, die aber qualitativ ähnlich ist wie bei C_Γ. Bei Periodizität ist $C_R(\alpha) = 0$, falls die Periode innerhalb der Wortlänge liegt. $C_R(\alpha)$ ist besonders sensitiv auf Intermittenz und innere Krisen (siehe 2.3.3), welche die periodischen Fenster im Attraktor der logistischen Abbildung umrahmen, wie Abb. 2-21 zeigt. Die Spitzen von $C_R(\alpha)$ an diesen Punkten — insbesondere vor den Fenstern — überragen die anderen Werte um ein Vielfaches. Diese Eigenschaft zeichnet $C_R(\alpha)$ aus. Sie ist um so ausgeprägter, je näher α bei 1 liegt.

Aus numerischen Gründen kann C_R nicht nach (63) berechnet werden. Diese Definition gibt aber die Richtung für eine Fixierung des Parameters α an. Eine stabile Approximation von C_R durch $C_R(\alpha)$ kann noch für $\alpha = 1.0001$ gewährleistet werden. Dieser Wert wurde empirisch für die logistische Abbildung ermittelt und auch für Gleichverteilung bestätigt (siehe 3.6). Wenn α noch näher an 1 gewählt wird, kann es zu Ungenauigkeiten in den Werten kommen bis zu einer völligen Dominanz der numerischen Artefakte. Damit ist nicht die vermutete Divergenz der Formel bei verschwindender Intermittenz gemeint. Im Folgenden wird die Rényi-Komplexität als $C_R(1.0001)$ berechnet.

2.6.4 ε-Komplexität

Die Idee ist nicht neu, die Komplexität eines Prozesses über die Konfiguration eines geeigneten minimalen Automaten zu definieren. So sah KOLMOGOROV (1965) die Komplexität eines Datensatzes in der Länge des kürzesten Computerprogramms, das denselben exakt beschreibt. Chaitin berechnete bereits 1966 die Größenordnung dieser Zahl für eine *N*-Zustands *M*-Band Turing Maschine (CHAITIN, 1990, S. 219ff) und entwickelte diesen Komplexitätsbegriff weiter (CHAITIN, 1987). Mit Komplexität ist dabei Zufälligkeit gemeint (LEMPEL & ZIV, 1976). Nur eine deterministische Beschreibung der Daten garantiert die geforderte vollständige Wiederherstellbarkeit der Originaldaten aus dem komprimierten Bild (ZIV & LEMPEL, 1978). WOLFRAM (1984) definierte eine „Reguläre-Sprachen-Komplexität“ über die Anzahl der Zustände eines minimalen deterministischen finiten Automaten zur Beschreibung (nicht nur) von durch zellulären Automaten erzeugte Sequenzen[4]. Auch GRASSBERGER (1986) bezieht sich auf diesen Automatentyp (siehe 2.6.1).

[4] Eine formale Sprache hängt unmittelbar mit der sie beschreibenden Automatenklasse zusammen (HOPCROFT & ULLMAN, 1990).

Der Komplexitätsbegriff nach Crutchfield (CRUTCHFIELD & YOUNG, 1989, CRUTCHFIELD, 1992, 1994a, 1994b) basiert auf minimalen nicht-deterministischen finiten Automaten. Die Konstruktion dieser Automaten wurde bereits in Abschnitt 2.1.3 beschrieben. Die minimale Maschine wird sogar auf sukzessive höheren Automatenebenen gesucht, falls die Beschreibungskapazität der aktuellen Ebene nicht ausreicht (siehe 2.1.4). CRUTCHFIELD (1994b) gibt für diese hierarchische ε-Maschinen Rekonstruktion eine Vorschrift an.

In Analogie zur deterministischen Universellen Turing Maschine (Computer), auf dem die Kolmogorov Komplexität definiert ist, benennt CRUTCHFIELD (1994b) die nicht-deterministische „Bernoulli-Turing Maschine“, auf der sein Maß für Komplexität definiert ist. Die deterministische Definition der Kolmogorov Komplexität lässt diese monoton mit der Zufälligkeit wachsen, was sie als Informationsmaß charakterisiert (siehe 2.5 oder Abb. 2-11). Die stochastischen Automaten werden bei höherer Zufälligkeit wieder einfacher, d. h. sie haben u. a. weniger Zustände (Knoten). Die darauf definierte ε-Komplexität nach Crutchfield nimmt ein Maximum bei hohem Determinismus aber noch nicht zu hoher Zufälligkeit an, also an der Grenze von Determinismus und Zufälligkeit, an der Grenze von Ordnung und Chaos.

Die ε-Komplexität wird nach CRUTCHFIELD (1992, 1994a, 1994b) als Shannon-Entropie (siehe 2.5.1) auf den Besuchshäufigkeiten $p(v)$ der stationären Zustände $v \in \mathbf{v}$ des finiten Automaten, wie in Abschnitt 2.1.3 beschrieben, definiert[5]:

$$C_\varepsilon = -\sum_{v \in \mathbf{v}} p(v) \log_2 p(v) \tag{64}$$

Diese Größe wird von CRUTCHFIELD (1994a, 1994b) selbst als statistische Komplexität C_μ bezeichnet. μ bezeichnet dabei das Wahrscheinlichkeitsmaß. Da auch andere Komplexitäten, z. B. C_Γ und C_{EM}, statistische Größen sind, wird hier die Bezeichnung C_ε bevorzugt. Wenn die zugrunde liegende Maschine minimal ist, misst C_ε die Gedächtnislänge des generierenden Prozesses (CRUTCHFIELD, 1992), d. h. die Speichergröße des Prozesses in Bits (CRUTCHFIELD, 1994a) oder genauer: die Speichergröße in Bits, die zur Vorhersage der Umgebung bei gegebener Genauigkeit ε erforderlich ist (CRUTCHFIELD, 1994b).

Bei CRUTCHFIELD & YOUNG (1989) wird C_ε allgemeiner als Renyi Entropie (siehe 2.5.2) der Ordnung α über die Zustände des Automaten definiert. Im Fall von $\alpha = 1$, also Gleichung (64), wird C_ε dann Graph Komplexität genannt. Der Fall $\alpha = 0$ liefert die stochastische Algorithmische Komplexität oder topologische Komplexität, die nur von der Anzahl $\|\mathbf{v}\|$ der Zustände des Automaten abhängt:

$$C_0 = \log_2 \|\mathbf{v}\| \tag{65}$$

Durch die stochastische Transfermatrix **T** (siehe 2.1.3) der Maschine wird ihr zugehöriger Markov Prozess beschrieben (CRUTCHFIELD, 1992). Dabei geht die detaillierte (deterministische) Struktur der ursprünglichen Symbolfolge verloren. Es bleibt lediglich die Übertragungsstruktur als Markov Kette erhalten. Die Entropierate der Markov Kette kann über die Übergangswahrscheinlichkeiten $p(v \to v')$ von Zustand v nach v' und die stationären Zustandswahrscheinlichkeiten $p(v)$, $v, v' \in \mathbf{v}$ berechnet werden (CRUTCHFIELD, 1992):

$$h_\mu(\mathbf{T}) = -\sum_{v \in \mathbf{v}} p(v) \sum_{v' \in \mathbf{v}} p(v \to v') \log_2 p(v \to v') \tag{66}$$

[5] Die stationären Aufenthaltswahrscheinlichkeiten der transienten Zustände sind 0 und würden keinen Beitrag zu C_ε liefern.

Sie misst die Informationsproduktionsrate in Bits pro Zeitschritt und entspricht nach CRUTCHFIELD (1992) der Entropierate *h*, gemäß Gleichung (37), der stochastischen Maschine.

Zur Beurteilung der Güte des Automaten schlagen CRUTCHFIELD & YOUNG (1989) die Berechnung des Graph-Indeterminismus vor:

$$I_G = \sum_{v \in \mathbf{v}} p(v) \sum_{s \in A} p(s \mid v) \sum_{v' \in \mathbf{v}} p(v \to v'; s) \log_2 p(v \to v'; s) \tag{67}$$

Darin bedeuten $p(v)$ die stationären Zustandswahrscheinlichkeiten, $p(v \to v'; s)$ die Übergangswahrscheinlichkeiten von Zustand v nach v' mit dem Symbol s aus dem Alphabet A und $p(s|v)$ die Wahrscheinlichkeit, dass der Zustand v mit dem Symbol s verlassen wird. Der Indeterminismus misst den Grad der Mehrdeutigkeit in den Übergängen zwischen den Zuständen des Automaten (Knoten des Graphen). Eine ε-Maschine ist rekonstruierbar, wenn I_G für einen Parametersatz verschwindet. Die Parameter der ε-Maschine sind die Feinheit der Partitionierung ε, die Wortlänge *L*, die Morph-Tiefe *D* und die Diskriminanz δ. Sie beschreiben jeweils ein bestimmtes Approximationsniveau, welches normalerweise einen positiven Indeterminismus bedeutet, der auf einen Rest an äußerem Rauschen hinweist.

Weitere Hinweise zur ε-Komplexität finden sich bei KURTHS et al. (1996), KURTHS & WITT (1994), WACKERBAUER et al. (1994).

Bemühen und Scheitern der ε-Maschinen Rekonstruktion:

Anwendungen der ε-Komplexität auf Messdaten gibt es in der verwendeten Literatur nicht. In dieser Arbeit wurde viel Zeit investiert, um die ε-Komplexität auf ökosystemare Zeitreihen anwenden zu können. Schließlich wäre sie das Komplexitätsmaß der Wahl, da sie auf dem minimalen Modell der „natürlichen" Modellklassenhierarchie zur Beschreibung der vom generierenden Prozess verwendeten Sprache beruht.

Die Automaten von ROMAHN (1996) basierten auf rein topologischer Morph-Äquivalenz (siehe 2.1.3). Ihre Qualität wurde alleine am Indeterminismus gemessen. In dieser Arbeit wurde an einer flexibleren Zustandsdefinition (δ-Ähnlichkeit mit oder ohne Topologie, siehe 2.1.3) zugelassen. Außerdem erwies sich der Indeterminismus als Qualitätskriterium für ungeeignet. Es wurden daraufhin alternativ der Quotient aus ε-Komplexität und Indeterminismus untersucht sowie die quadratische Abweichung oder die Modell Effizienz *R* (JANSSEN & HEUBERGER, 1995) für einen von der rekonstruierten Maschine erzeugten künstlichen Symbolsatz im Vergleich zum ursprünglichen Symbolsatz. Es wurde sogar eine Verteilung von künstlichen Symbolsätzen erzeugt. Die Ergebnisse waren aber diesbezüglich stabil genug, so dass nur ein solcher Symbolsatz ausreichte. Die quadratische Abweichung von ursprünglichen und generierten Symbolsatz erwies sich darunter als das geeignetste Qualitätskriterium. Dieses wurde auch von Annette Witt nach persönlicher Auskunft an Holger Lange verwendet.

Schließlich wurden bei bis zu 1 000 000 Iterationen der logistischen Abbildung verschiedenste Parametervariationen von Wortlänge, Morph-Tiefe und Diskriminanz untersucht, um die Bewertung des Attraktors der logistischen Abbildung durch die ε-Komplexität nach WACKERBAUER et al. (1994) zu reproduzieren. Dies ist nur ansatzweise gelungen. Immer wieder gab es mehrere Wachstumsparameter, für welche die Maschinen-Rekonstruktion versagte. Darüber hinaus waren die lokalen Schwankungen des Maßes relativ hoch, was auch in der Arbeit von WACKERBAUER et al. (1994) auffällt.

Insgesamt musste also in dieser Arbeit, wie auch in der Dissertation von WITT (1996, S. 45), festgestellt werden, dass die ε-Komplexität zur Analyse von realen Messdaten (aufgrund des dort vorhandenen geringen Datenumfanges) nicht geeignet war. Die Anforderungen an die

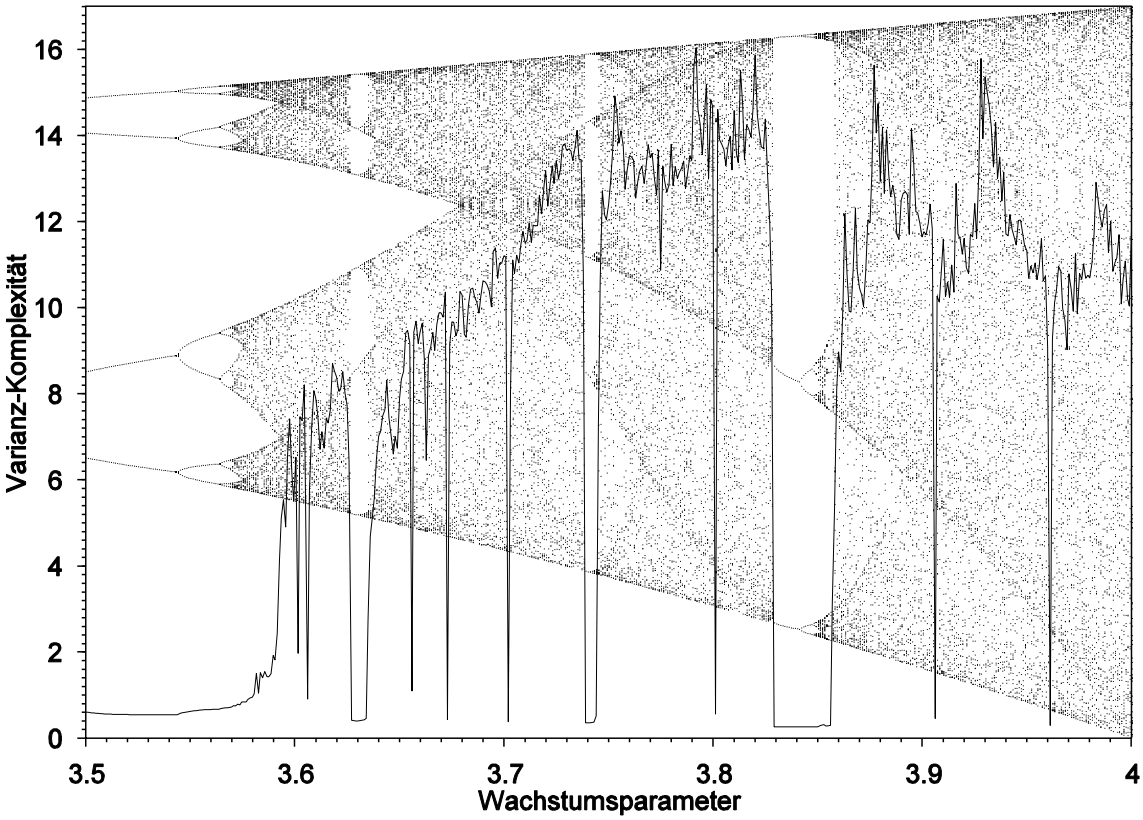

Abb. 2-22. Varianz-Komplexität für die logistische Abbildung (24). Je $N = 10000$ Daten nach 10000 Voriterationen und Normierung. Hintergrund: Attraktor, Abb. 2-9.

Rekonstruktion eines Automaten aus den Daten sind erwartungsgemäß höher als die Anforderungen an eine statistische Beschreibung.

2.6.5 Metastatistische Komplexität

Der Begriff „Metastatistisch“ wird hier im Sinne von ATMANSPACHER et al. (1997) gebraucht. Damit ist gemeint, dass über mehrere Teile eines Datensatzes eine statistische Größe erster Ordnung (z. B. Information) berechnet wird und von diesen Werten anschließend wieder die gleiche Größe berechnet wird, was zu einem Maß zweiter Ordnung (Komplexität) führen soll.

2.6.5.1 Varianz-Komplexität

Eine Alternative zu den bereits vorgestellten Komplexitätsmaßen ist die Varianz-Komplexität von ATMANSPACHER et al. (1997). Dieses Maß ist parameterfrei und wird direkt auf den Daten berechnet. Die Schwierigkeiten, eine optimale Parameterkombination (Partitionierung, Wortlänge, u. a.) zu finden, entfallen also. Die Varianz-Komplexität basiert auf der Erkenntnis, dass Maße für Komplexität metastatistisch formuliert sind, d. h. es wird eine Statistik über eine Statistik betrieben. ATMANSPACHER et al. (1997) definieren dieses strukturelle Maß (siehe 2.4) zur Analyse räumlicher Muster (Pixelgrafiken). Nach einer Empfehlung von Harald Atmanspacher wird hier die Konstruktionsvorschrift auf Zeitreihen angewendet.

Der Datensatz wird mit einem Fenster fester Länge abgefahren (vgl. Abb. 2-3). Für jedes Fenster *i* der Länge *n* (also: jede Teilfolge *i* der Länge *n*) wird die Standardabweichung $\sigma_{i,n}$ der Daten berechnet. Von den $\sigma_{i,n}$ mit $n/2 < i < N–3n/2+1$ wird wiederum die Standardabweichung σ_n berechnet. Für ein Feld mit Strukturen auf allen Längen — also für ein komplexes räumliches Muster — stellen ATMANSPACHER et al. (1997) im Gegensatz zu periodischen oder zufälligen Mustern fest, dass sich σ_n kaum mit *n* ändert. Daher definieren sie die Varianz-Komplexität über den Kehrwert der gesamten Krümmung der Funktion σ_n:

$$C_{\mathrm{V}} = \frac{1}{\sum_n \left|\sigma_{n+1} - 2\sigma_n + \sigma_{n-1}\right|} \tag{68}$$

In ihrem Beispiel erfüllt diese Definition die Anforderungen an ein Maß für Komplexität.

Bei der Untersuchung eindimensionaler Daten (der logistischen Abbildung, einem Bernoulli-Prozess mit beliebigen Zuständen und Zeitreihen) in dieser Arbeit konnte weder ein deutliches Erste-Ordnung-Verhalten (Anstieg mit Zufälligkeit) noch ein Zweite-Ordnung-Verhalten (Minimum bei konstanten/periodischen und zufälligen Daten, dazwischen Maximum) für die Varianz-Komplexität festgestellt werden. Insbesondere fiel auf, dass C_{V} bei konstanten Daten (z. B. logistische Funktion mit einem Fixpunkt) gegen ∞ divergiert, was nach Gleichung (68) zu erwarten war, und dass C_{V} bei zufälligen Daten nicht verschwindet, sondern dort sogar höhere Werte annimmt als in den periodischen Fenstern der logistischen Funktion. Insgesamt bewertet C_{V} die logistische Abbildung in einer Weise, die zwischen den bisher beobachteten Charakteristika für Informations- und Komplexitätsmaße liegt (siehe Abb. 2-22). Außerdem hängt C_{V} von der Skalierung der Daten ab. Um dies zu vermeiden, wurden die Daten vor der Analyse normiert, d. h. von jedem Datenpunkt wurde der Mittelwert des Datensatzes abgezogen und anschließend wurde durch die Standardabweichung dividiert. Damit hatten alle Datensätze den Mittelwert 0 und die Standardabweichung 1. An der prinzipiellen Beurteilung der Dynamik änderte aber auch die Normierung nichts.

2.6.5.2 Bandverschmelzungskomplexität

Die Varianz-Komplexität nach Gleichung (68) verhält sich nur wenig spezifisch. Daher wurde versucht, eine andere Eigenschaft von σ_n für eine parameterfreie Definition von Komplexität zu nutzen:

Um stets alle Fenster zu berücksichtigen, wurde σ_n jetzt mit allen Werten $\sigma_{i,n}$ berechnet. Im Gegensatz zur Beobachtung von ATMANSPACHER et al. (1997) zeigte die Funktion σ_n bei allen untersuchten Datensätzen eine prinzipiell hyperbolische Abhängigkeit von *n* (siehe Abb. 2-23 oben). Außerdem zeigten die Werte von σ_n bereits ab kleinen Fenstergrößen *n* deutliche Schwankungen bei verschiedenen Realisationen von Zufallsprozessen (siehe Abb. 2-23 unten). Bei größeren Datenmengen *N* lagen schon ab kleinen Fenstergrößen *n* im Mittel auch die Werte von σ_n deutlich höher (siehe Abb. 2-23 unten). Als einigermaßen stabil erwies sich lediglich die Lage und Höhe des ersten und globalen Maximums von σ_n (siehe Abb. 2-23 oben rechts). Dieses wurde bei der logistischen Funktion stets für $n = 2$ erreicht. Bei periodischen Daten mit Periodenlänge *l* liegt es zwischen $2 \le n \le l/2$. Die Höhe dieses Maximums ist charakteristisch für die Komplexität der Daten. Leider verschwindet sie nicht bei zufälligen Daten. Ansonsten ist sie 0 bei konstanten Daten, klein bei periodischen Daten und nimmt ein scheinbar stetiges Maximum bei Bandverschmelzung der logistischen Funktion an. Eine Unabhängigkeit von der Skalierung der Daten wird am besten erreicht, indem das mit den

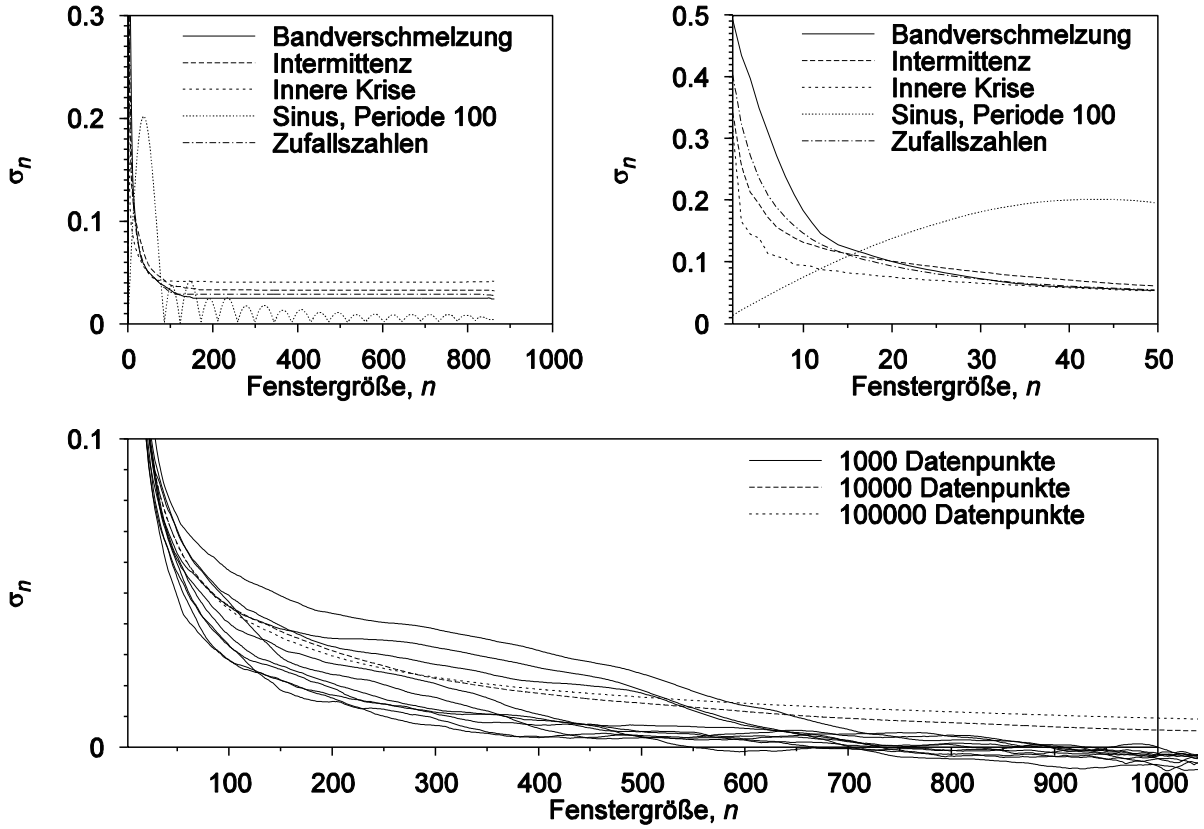

Abb. 2-23. Standardabweichung der Standardabweichungen σ_n in Abhängigkeit von der Fenstergröße *n*. *Oben:* Je 1000 standardisierte Daten von fünf Prozessen, u. a. der logistischen Abbildung (24) für $r = 3.678$, 3.828 und 3.857. σ_n ist links für $n \in [2,1000]$ und rechts für $n \in [2,50]$ dargestellt. *Unten:* Die logistische Abbildung (24) am Bandverschmelzungspunkt $r = 3.678$. 10×1000 aufeinander folgende Iterationen sowie 10000 und 100000 separate Iterationen.

Rohdaten ermittelte σ_n-Maximum mit der Standardabweichung der Daten normiert wird. Diese Größe wird Bandverschmelzungskomplexität genannt.

Trotz der Schwankungen von σ_n und der Abhängigkeit von der Datenmenge schon ab kleinen Fenstergrößen *n* gab es Hinweise, dass sich komplexe Daten durch hohe Werte von σ_n gerade bei hohem *n* auszeichnen. Dies konnte mit entsprechenden „Komplexitäts“-Funktionen, z. B.:

$$\sum_n \sigma_n^2, \quad \sum_n \frac{n-1}{n}\sigma_n \quad \text{oder} \quad \sum_n \left(\sigma_n - f_n\right)^2, \text{ falls } \sigma_n > f_n := \frac{2.76}{n+6.91} \tag{69}$$

bestätigt werden. Sie wiesen auch ein stetiges Maximum beim Bandverschmelzungspunkt der logistischen Abbildung auf. Allerdings waren diese Komplexitäten — aufgrund der Schwankungen von σ_n — sehr stark verrauscht und zeigten überhaupt keine Abnahme mit zunehmender Zufälligkeit.

Die Bandverschmelzungskomplexität erinnert in ihrer Beurteilung der Dynamik der logistischen Abbildung an diejenige der Fluktuationskomplexität für ungerade Wortlängen (vgl. WACKERBAUER et al., 1994).

2.6.5.3 Entropie-Verteilungskomplexität

Bei den beiden letzten Versuchen einer metastatistischen Definition von Komplexität geschah dies über die Verteilung der Standardabweichungen von Teilfolgen des Datensatzes. Es liegt nahe, auf ähnliche Weise eine Komplexität der Informationsverteilung zu definieren. Hierzu stehen die (etablierten) Informationsmaße aus Abschnitt 2.5 zur Verfügung. Das einfachste Maß für Information oder Ungleichverteilung, welches auch bei nur wenigen Datenpunkten noch berechenbar ist, ist die Shannon-Entropie auf den Symbolen. Damit wurde wie folgt eine Entropie-Verteilungskomplexität berechnet:

Der Datensatz wird zunächst durch Partitionierung (siehe 2.1.1) in einen Symbolsatz transformiert. Dann wird für jedes Fenster i von Teilsymbolfolgen der Länge n die Shannon-Entropie $H_{S,i,n}$ nach (26) berechnet (Wortlänge $L = 1$). Als Basis für den Logarithmus wird die Alphabetgröße λ genommen, damit der Wertebereich einheitlich [0,1] ist. Über die Häufigkeiten der $H_{S,i,n}$ in [0,0.5] und (0.5,1] für alle i bei festem n wird die Shannon-Entropie $H_{S,n}$ berechnet. (Dies ist der einfachste Fall. Als Verfeinerung könnte $H_{S,n}$ über kleinere Teilintervalle von [0,1] berechnet werden.) Die Funktion $H_{S,n}$ in Abhängigkeit von n ist analog zu σ_n aus Abschnitt 2.6.5.1 und 2.6.5.2 zu verstehen. Auch $H_{S,n}$ erreicht ein ausgezeichnetes Maximum bei geringer Fenstergröße, falls die Funktion nicht konstant 0 ist, wie z. B. bei der logistischen Iteration (siehe 2.3.3) für $r < 3.678$. Allerdings erreicht $H_{S,n}$ dann schnell ein Niveau von konstant 0.

Leider ist es nicht gelungen anhand der Funktion $H_{S,n}$ ein Maß für Information, Komplexität oder eine andere Eigenschaft (z. B. Bandverschmelzung) zu kreieren. Auch das Maximum von $H_{S,n}$ weist nicht die charakteristischen Eigenschaften des Maximums von σ_n auf. Wenn unter Komplexität die Ungleichmäßigkeit der Verteilung der Information einer Nachricht (Zeitreihe) verstanden wird, wie z. B. in der Definition der Fluktuationskomplexität, so lässt sich dieser Komplexitätsbegriff nicht auf die hier vorgestellte Weise umsetzen.

Eine Alternative zu dem hier vorgestellten Ansatz stellt die Jensen-Shannon-Abweichung nach LIN (1991) dar. Mit ihr könnte die Abweichung der Shannon-Entropien in den Fenstern von der Shannon-Entropie des gesamten Datensatzes berechnet werden, wobei noch eine Gewichtung möglich ist. Dieses Konzept wurde hier jedoch nicht weiter verfolgt.

2.6.5.4 Fazit zu den metastatistischen Methoden

Insgesamt konnte mit keiner der hier vorgestellten metastatistischen Methoden ein sicheres Maß für Information, Komplexität oder eine andere Eigenschaft konstruiert werden. Die metastatistischen Methoden werden in dieser Arbeit daher nicht weiter angewendet.

2.7 Das Programm SYMDYN

Zur Berechnung der in diesem Kapitel vorgestellten Methoden sowie zur Durchführung der darauf basierenden Analysen wurde das Programm SYMDYN (von Symbolische Dynamik) geschrieben, das ein wesentlicher Teil dieser Arbeit ist. Es baut ursprünglich auf dem Programm von ROMAHN (1996) zur Berechnung der Metrischen Entropie und topologischer ε–Maschinen auf. Die Methoden in SYMDYN können in universeller Weise auf Zeitreihen

angewendet werden. Messlücken sowie die begrenzte Datenmenge werden dabei gemäß Abschnitt 3.6 und der Erläuterungen in diesem Kapitel berücksichtigt.

Mit SYMDYN können sowohl (Mess-) Datenreihen, die in einer Zeitreihen-Datenbank näher beschrieben sind, als auch vom Programm generierte künstliche Datenreihen wie Zufallsfolgen, Bernoulli-Prozesse, logistische Funktion und andere analysiert werden. Die Parameter der künstlichen Daten können Schleifen durchlaufen. Es sind einige Methoden zur Vorverarbeitung der Daten implementiert, wie z. B. Aggregation, Glättung, Vertauschen, Entsaisonalisierung u. a. Für jeden Funktionsparameter und/oder Vorverarbeitungsschritt wird dann das ausgewählte Komplexitätsmaß berechnet. Dabei können die maßspezifischen Parameter ebenfalls variieren. Die Maße der „Shannon-Gruppe“ (H_S, H_μ, H_G, H_M, C_{EM}, C_Γ, C_R) mit gleichen Parametern können sogar in einem Programmschritt gleichzeitig berechnet werden. Wenn Funktionsparameter oder Parameter einer oder mehrerer Vorverarbeitungsroutinen variieren, wird eine Statistik über das ausgewählte Komplexitätsmaß ausgegeben; alternativ kann auch nur der Wert des jeweiligen Optimums bei Variation der maßspezifischen Parameter für jeden äußeren Parameter ausgegeben werden.

SYMDYN arbeitet in der MS-DOS Eingabeaufforderung von Windows 95/NT. Der Programmablauf von SYMDYN wird durch eine externe Steuerungsdatei bestimmt. Diese wird vor Programmstart mit einem Texteditor editiert und anschließend als Argument beim Programmaufruf übergeben[6]. Außer dem ausführbaren Programm SYMDYN.EXE werden also noch die Zeitreihen-Datenbank SYMDYN.TS (Textdatei) und die Steuerdatei SYMDYN.KEY benötigt. Darüber hinaus wird zur Bestimmung der maximalen Wortlänge die Resource-Datei SYMDYN.RES verwendet, die die Tabellen aus Abschnitt 3.6 enthält. Die Sprache der Programmausgabe und Kommentierung im Quellcode ist wegen der Universalität englisch. Entsprechend wurden Variablennamen und Abkürzungen gewählt. Das Programm selbst ist in C++ nach KERNIGHAN & RITCHIE (1990) und STROUSTRUP (1995) geschrieben. Mit den objektorientierten Datenstrukturen dieser modernen Programmiersprache läßt sich eine nach außen gleiche Behandlung verschiedener Analyse-Methoden elegant realisieren. Das Programm wurde zuerst mit „Borland-C++ 4.5“ compiliert. Die mittlerweile 59 Quelldateien in drei Ordnern wurden zuletzt mit „Microsoft Visual-C++ 5.0“ compiliert. Das ausführbare Programm ist etwa 340 kB groß. Die Programmierung von SYMDYN stellt (zeitlich) einen wesentlichen Teil dieser Arbeit dar. Für Details sei auf die Anleitung (s. u.) oder die Kommentare im Quellcode verwiesen.

Zur Bedienung des Programms liegt eine Anleitung (28 Seiten) vor. Diese ist mit dem Programm, der Resource-Datei, einer Standard-Steuerdatei, einer Beispiel-Datenbank und Beispiel-Daten auf dem FTP-Server des BITÖK unter der Adresse ftp://ftp.bitoek.uni-bayreuth.de/pub/mod/symdyn/ zu beziehen. Zugang dazu ist auch über die BITÖK-Projekt-Seite http://www.bitoek.uni-bayreuth.de/Forschung/Projekte/000180/DE.html, meine BITÖK-Homepage http://www.bitoek.uni-bayreuth.de/Mitarbeiter/000139/DE.html und meine private Homepage http://www.bitoek.uni-bayreuth.de/~Frank.Wolf/ möglich.

[6] Änderungen in der Steuerdatei können auch für einen Batch-Job automatisiert werden, z. B. mit Visual Basic oder Perl.

3 Anforderungen an die Daten und Methode

In diesem Kapitel sollen Anforderungen an die Datenqualität und Datenmenge formuliert werden, die zur Berechnung von Komplexitätsmaßen erforderlich sind. Messdaten sind oftmals nicht stationär, enthalten Lücken und liegen nur in begrenzter Anzahl vor. Außerdem sind sie einmalig, d. h. die Erhebung kann unter den gleichen Bedingungen nicht wiederholt werden. Daher können stabile Analysewerte nicht etwa durch eine Monte-Carlo-Simulation und anschließender Verwendung des Mittelwertes erreicht werden. Die natürliche Schwankung der Analysewerte des untersuchten Prozesses kann also nicht exakt bestimmt werden.

Ziel dieses Abschnittes ist es unter anderem auch Anforderungen an die Parameter (Wortlänge und Partitionierung) der Verfahren zu formulieren. Idealerweise können Parameter unter den gegebenen Umständen fixiert werden, was die Objektivität der Methoden erhöht.

3.1 Äquidistanz

Grundsätzlich werden bei einer Zeitreihenanalyse äquidistante Daten untersucht, d. h. der Zeitabstand benachbarter Messpunkte muss gleich sein (HARTUNG et al., 1998). Ein nichtäquidistanter Zeitabstand gibt die Dynamik des Prozesses in der Zeitreihe verzerrt (und verfälscht) wieder. Hinzu kommt, dass bei den meisten Methoden — so auch bei den Komplexitätsmaßen — explizit keine Zeitskala der Daten berücksichtigt wird. D. h. diese Methoden setzten Äquidistanz voraus.

Das Zeitintervall sollte so klein sein, dass auch alle wesentlichen Phänomene erfasst werden (HARTUNG et al., 1998). Zur Größe des relevanten Zeitintervalls wurden in dieser Arbeit eigene Untersuchungen angestellt (siehe 5.3). In dieser Arbeit wurden ausschließlich äquidistant gemessene oder auf die kleinste gemeinsame äquidistante Auflösung ausgelesene Daten (Saugspannungen, Steinkreuz, siehe 4.2) verwendet.

3.2 Ergodizität

Eine der Grundvoraussetzungen zur Anwendung der Methoden der Statistischen Mechanik ist die Ergodizität eines Systems. Für die exakte Formulierung der starken und schwachen *ergodischen Hypothese* in der Statistischen Mechanik sei auf TOLMAN (1967, S. 65ff) oder WILDE & SINGH (1998, S. 17f) verwiesen. Bei der Analyse von Zeitreihen oder Symbolfolgen wird mit Ergodizität gefordert, dass eine Zeitreihe oder Symbolfolge eine typische Realisierung des generierenden Prozesses oder der zugrunde liegenden Nachrichtenquelle ist (EBELING et al., 1998).

Beispielsweise sollte bei der Untersuchung des Abflusses eines Wassereinzugsgebietes gewährleistet sein, dass die Zeitreihen sowohl Trockenperioden, die Schneeschmelze, Starkregenereignisse sowie alle Jahreszeiten u. a. erfassen. Anderenfalls enthält eine Abflussreihe vermutlich nicht alle oder die typischen Charakteristika des Einzugsgebietes.

Die Gewährleistung der Ergodizität kann nicht statistisch überprüft werden. Die Einhaltung dieser Forderung muß bei der Auswahl der Daten und des Messzeitraumes beachtet werden. Eine Analyse, ein Modell oder eine Vorhersage kann nur die Struktur liefern, die in den zugrunde liegenden Daten (zur Validierung) enthalten ist.

SHANNON (1976, S. 134) und KHINCHIN (1957) setzten die Ergodizität der Nachrichtenquellen zum Beweis ihrer Theoreme über die Informationsentropie voraus. Nur dann konvergieren die Entropieraten gegen die Entropie der Quelle (siehe 2.5.3).

3.3 Stationarität

Eine ähnlich fundamentale Eigenschaft wie die Ergodizität ist die Stationarität. Sie wird von vielen statistischen Verfahren, wie auch von den stochastischen Komplexitätsmaßen, über die Daten vorausgesetzt. Eine Zeitreihe ist stationär, wenn ihre wesentlichen statistischen Eigenschaften nicht von der Zeit abhängen (WITT et al., 1998). Die starke Stationarität verlangt die Zeitunabhängigkeit aller statistischen Momente, was für endliche Zeitreihen praktisch nicht überprüft werden kann. Die schwache Stationarität verlangt nur die Invarianz der Momente bis zu einer bestimmten (i. d. R. der zweiten) Ordnung. Zur Definition der Stationaritätsbegriffe siehe z. B. HIPEL & MCLEOD (1994, S. 69). Für die schwache Stationarität der Ordnung 2 dürfen Mittelwert, Varianz und Autokorrelation (von Teilfolgen der Zeitreihe) nur von der Zeitdifferenz aber nicht von der Position in der Zeitreihe abhängen. Diese Annahme liegt den Untersuchungen hydrologischer stationärer Prozesse bei HIPEL & MCLEOD (1994, S. 69) zugrunde. WITT et al. (1998) schlagen einen neuen Test auf Stationarität vor, der auf „nur“ einige 1000 Datenpunkte ausgelegt ist, und die zeitliche Unabhängigkeit der Häufigkeitsverteilung und des Frequenz-Spektrums der Autokorrelation in benachbarten Fenstern untersucht.

Wenn Stationarität gefordert wird, müssen die relevanten Zeitskalen (z. B. eine Periode) des Prozesses klein bezüglich des Beobachtungszeitraumes sein (WITT et al., 1998). Die relevanten Zeitskalen des Prozesses hängen mit der Korrelationslänge (siehe 2.2) zusammen. WITT (1996, S. 31) schliesst nicht aus, dass eine instationäre Zeitreihe durch eine spezielle Transformation (Partitionierung, siehe 2.1.1) in eine stationäre Symbolfolge umgewandelt werden kann.

Üblicherweise werden Instationaritäten (Trends und Saisonalitäten) zunächst durch sogenannte Filter bereinigt, damit die Ergebnisse der Analyse dadurch nicht verfälscht oder dominiert werden (HIPEL & MCLEOD, 1994; HARTUNG et al., 1998; u. a.). Diese Vorgehensweise kann jedoch die Analyse auch ungünstig beeinflussen, wenn z. B. ein klimatisch bedingter Jahresgang mit einer angenommenen Periode von 365 Tagen eliminiert werden soll. Da eine Klimaperiode nie exakt 365 Tage dauert und sich das Klima nicht harmonisch, sondern mit Verzögerungen und Schüben entwickelt, kann den Daten hiermit eine zusätzliche Frequenz aufgeprägt werden, die in der Autokorrelation sichtbar ist (s. u.).

Trends in den hier verwendeten hydrologischen Daten (siehe Kapitel 4) sind unwahrscheinlich, wenn kurzreichweitige anthropogene Eingriffe ausgeschlossen werden können. Langfristige Veränderungen des Klimas, etwa eine Temperaturerhöhung sowie die Evolution der Flora, Fauna und des Bodens und Gesteins, wirken sich auf Zeitskalen jenseits der maximal 47 Jahre für die längste hier verwendete Zeitreihe aus.

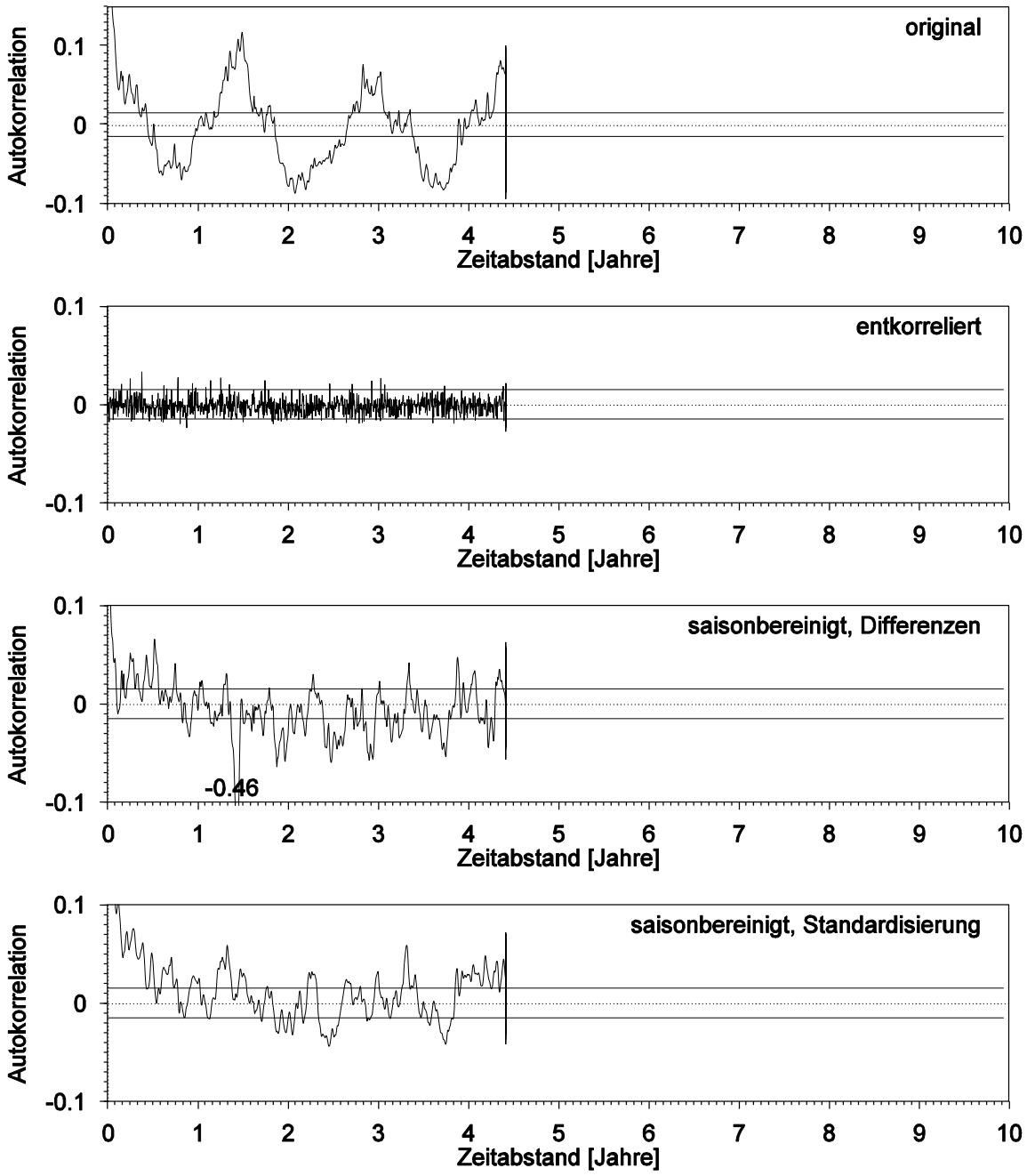

Abb. 3-1. Autokorrelation für den täglichen Abfluss der Langen Bramke 1948 – 1995. Von oben nach unten: Für die Originalzeitreihe, nach zufälliger Neuanordnung der Messwerte, Bereinigung des Jahresganges durch Differenzen nach (70) mit $p = 365$, Bereinigung des Jahresganges durch Standardisierung nach (71) mit $p = 365$. Die Waagerechten geben den 5 % Signifikanzbereich an.

Saisonabhängigkeit von Komplexitätsmaßen:

Die in dieser Arbeit verwendeten Abflussdaten haben fast alle einen ausgeprägten Jahresgang (siehe Kapitel 4 oder 5.4.1). Im Folgenden wird der Einfluss einer Saisonbereinigung nach zwei verschiedenen Verfahren auf die Berechnung von Komplexitätsmaßen für die längste hier vorliegende Zeitreihe, den täglichen Abfluss der Langen Bramke (siehe 4.3), untersucht.

Bei der *Saisonbereinigung durch Differenzenbildung* wird eine unterstellte Periode p (hier: Jahresgang $p = 365$ Tage) durch Subtraktion von jeweils $t = p$ Zeitschritte auseinander liegen-

den Datenpunkten *x* eliminiert (HARTUNG et al., 1998, S. 668f). Die neue Zeitreihe *Y* verkürzt sich dabei um *p* Werte:

$$y_t = x_{t+p} - x_t \tag{70}$$

Bei der *Saisonbereinigung durch Standardisierung* wird die angenommene Periode *p* dadurch eliminiert, dass von jedem Datenpunkt x_t der zugehörige saisonale Mittelwert μ_s abgezogen wird und anschließend durch die saisonale Standardabweichung σ_s dividiert wird (HIPEL & MCLEOD, 1994, S. 465), $s = 0, 1, ... p–1$:

$$z_t = \frac{x_t - \mu_s}{\sigma_s} \tag{71}$$

Abb. 3-1 zeigt die Autokorrelation der Original- und saisonbereinigten Daten. Der Jahresgang in der Originalzeitreihe ist deutlich zu erkennen und signifikant. Bei zufälliger Neuanordnung der Messwerte ist die Autokorrelation praktisch nirgends signifikant: Die Zeitreihe ist völlig entkorreliert. Bei den beiden Methoden zur Saisonbereinigung verschwindet der Jahresgang, aber die Autokorrelation ist noch über den gesamten Zeitraum signifikant, d. h. außer dem Jahresgang wurden keine Korrelationen zerstört. Die negative Spitze in der Autokorrelation nach genau einem Jahr bei der Differenzenmethode ist durch das Verfahren bedingt. Dies kann durch Einsetzen von (70) in (18) eingesehen werden. Nach einer Rechnung von Holger Lange muss die Autokorrelation bei der angenommenen Periode einen Wert von etwa –0.5 annehmen. Dies ist bei der Standardisierung nicht der Fall.

Sowohl bei den Originaldaten wie auch bei der Jahresgangbereinigung durch Standardisierung wird das 5 %-Signifikanzniveau zum ersten Mal nach etwa drei Monaten erreicht. Dies kann auf die Dauer langfristiger klimatischer Ereignisse zurückgeführt werden, wie die Trägheit des Grundwasserspeichers, die Dauer der Schneedecke im Winter und die Vegetationsperiode im Sommer. Bei der Jahresgangbereinigung durch Differenzen wird die Signifikanzschwelle bereits nach 28 Tagen zum ersten Mal unterschritten. Dies deutet auf die Dauer mittelfristiger Ereignisse, wie das Trockenfallen des Baches nach längerem Niederschlag oder die Schneeschmelze, hin. Bis zu diesem Zeitraum ist auch noch eine Informationserhöhung mit der Aggregation der Daten möglich (siehe 5.3.1).

Zur Quantifizierung der Auswirkung einer Saisonbereinigung auf die Berechnung der Komplexitätsmaße wurden diese sowohl für die originalen täglichen Abflussmessungen der Langen Bramke 1948 – 1995 berechnet als auch für die Daten nach Bereinigung des Jahresganges nach den beiden oben vorgestellten Methoden. Um die Signifikanz der Abweichungen beurteilen zu können, ist ein Maß für die Schwankung der Komplexitätsmaße erforderlich. Dazu wurden die Komplexitätsmaße jeweils für die ersten 11 Intervalle von vier Jahren (1. Intervall: 1.11.1948 – 31.10.1952, 2. Intervall: 1.11.1952 – 31.10.1956, ... , 11. Intervall: 1.11.1988 – 31.10.1992) berechnet. Davon konnten Mittelwert und Standardabweichung berechnet werden, welche ein Maß für die Schwankung ist. Die Differenz von Werten eines Komplexitätsmaßes von Original- und bereinigten Daten kann als signifikant angenommen werden, wenn diese größer ist als die Summe der jeweiligen Standardabweichungen. Dies gilt deswegen, weil die Werte annähernd normalverteilt sind, worauf die kleinen Momente 3. und 4. Ordnung (Schiefe und Wölbung, vgl. HARTUNG et al., 1998, S. 118) hindeuten. Um die Signifikanz leichter einschätzen zu können wurde ein Signifikanz-Index i_S aus dem Komplexitätsmaß-Mittel μ_O der Originaldaten, dessen empirischer Standardabweichung σ_O und den entsprechenden Größen μ_S und σ_S nach Saisonbereinigung definiert:

Tabelle 3-1. Einfluss der Bereinigung des Jahresganges auf verschiedene Komplexitätsmaße für den täglichen Abfluss der Langen Bramke 1948 – 1992. Binäre statische Median-Partitionierung. Die Maße wurden jeweils in benachbarten Intervallen von vier Jahren (1461 Tage) mit Wortlänge 4 berechnet. Davon sind Mittelwert und Standardabweichung für die *Original*daten sowie nach Bereinigung des Jahresganges durch *Differenzen* (70) oder *Standard*isierung (71) notiert. Die relative Abweichung (RA) in Prozent sowie der Index i_S nach (72) bezieht sich auf die Methode in der Spalte davor bezüglich der Originaldaten. Die Mittel beziehen sich auf absoluteWerte.

Maß	Original	Differenzen	RA [%]	i_S	Standard	RA [%]	i_S
I_A	0.260 ± 0.029	0.319 ± 0.047	+22.7	0.776	0.325 ± 0.047	+25.0	0.855
H_μ	0.444 ± 0.018	0.486 ± 0.036	+9.5	0.778	0.492 ± 0.041	+10.8	0.814
H_G	0.245 ± 0.023	0.302 ± 0.046	+23.3	0.826	0.311 ± 0.051	+26.9	0.892
H_M	1.533 ± 0.050	1.645 ± 0.100	+7.3	0.747	1.658 ± 0.110	+8.2	0.781
C_{EM}	0.798 ± 0.029	0.739 ± 0.045	–7.4	0.797	0.725 ± 0.045	–9.1	0.986
C_Γ	1.481 ± 0.079	1.588 ± 0.089	+7.2	0.637	1.647 ± 0.093	+11.2	0.965
C_R	1.411 ± 0.095	1.497 ± 0.114	+6.1	0.411	1.493 ± 0.122	+5.8	0.378
Mittel			11.9	0.710		13.9	0.810

$$i_S = \frac{|\mu_S - \mu_O|}{\sigma_S + \sigma_O} \tag{72}$$

Für $i_S > 1$ kann von einer signifikanten Abweichung ausgegangen werden.

Tabelle 3-1 stellt die Mittelwerte und Standardabweichungen der in dieser Arbeit wichtigen Komplexitätsmaße mit und ohne Saisonbereinigung sowie die relative prozentualen Abweichungen gegenüber den Originaldaten und die Indizes i_S dar. Es wurde eine binäre statische Median-Partitionierung verwendet. Die Wortlänge wurde bei allen Maßen gleich 4 gewählt, um die Abweichungen verschiedener Maße vergleichen zu können. Wortlänge 4 garantiert bei 1461 Datenpunkten (Tagen) nach Tabelle 7-4 eine mittlere Genauigkeit von 5 %. Die gleichen Berechnungen wurden außerdem für statische binäre H_μ- und H_G-maximale Partitionierungen sowie für dynamische 0- und H_G-maximale Partitionierungen durchgeführt.

Da die statischen Partitionierungen (siehe 2.1.1.1) per Konstruktion grundsätzlich von anderer Qualität sind — z. B. das Messrauschen weniger berücksichtigen — als die dynamischen Partitionierungen (siehe 2.1.1.2) waren die Ergebnisse entsprechend unterschiedlich: Dynamische Partitionierungen bewirken einen hohen Informationsgehalt (I_A, H_μ, H_G, H_M) und eine niedrige Komplexität (C_{EM}, C_Γ, C_R), was ein Indiz für hohe Zufälligkeit (Messrauschen) ist. Dies soll hier nicht das Objekt der Betrachtung sein. Außerdem zeigte sich bei dynamischer Partitionierung eine besonders hohe Abweichung der Informationen und Komplexitäten der Jahresgang-bereinigten Daten nach der Standardisierungsmethode von denen der Originaldaten (RA = 30 %, $i_S = 2.2$ im Mittel bei H_G-Maximierung, RA = 118 %, $i_S = 7.1$ bei 0-Partitionierung). Die Abweichungen der Werte nach der Differenzen-Methode bei dynamischer Partitionierung waren gering und innerhalb der Standardabweichungen (RA = 6 %, $i_S = 0.4$ im Mittel bei H_G-Maximierung, RA = 11 %, $i_S = 0.8$ bei 0-Partitionierung).

Die Ergebnisse der Berechnungen nach den drei statischen Partitionierungen waren sehr ähnlich. Die besten Ergebnisse wurden bei H_G-Maximierung erreicht (RA = 9.9 %, $i_S = 0.6$ bei Differenzen-Methode, RA = 12.7 %, $i_S = 0.8$ bei Standardisierung). Am wenigsten schwankten die Indizes i_S zwischen den Maßen bei der Median-Partitionierung. Deshalb wurden diese Werte für Tabelle 3-1 ausgewählt. Erwartungsgemäß liegen die Werte der Informationsmaße (I_A, H_μ, H_G, H_M) bei informationsmaximierenden Partitionierungen höher als bei Median-Partitionierung (siehe auch 3.8.1).

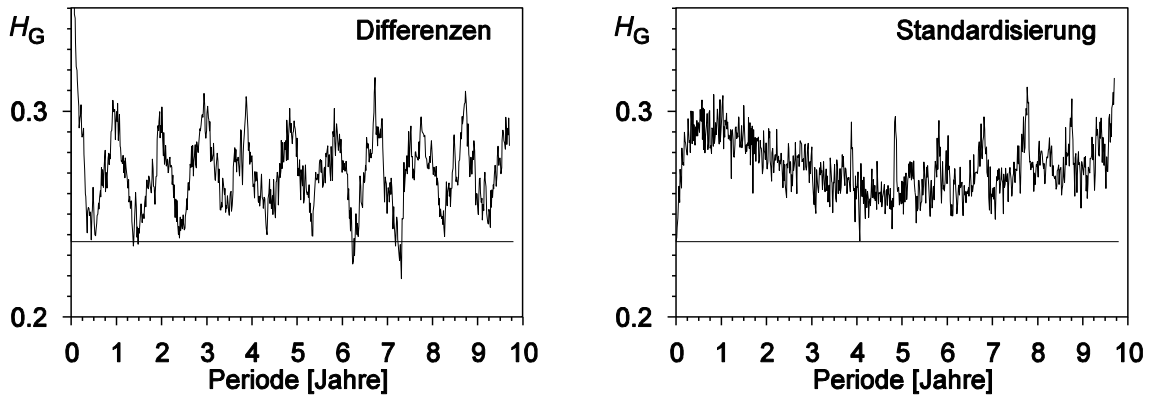

Abb. 3-2. Auswirkung einer Saison-Bereinigung mit verschiedener Periode auf den Informationsgewinn H_G für den täglichen Abfluss der Langen Bramke 1948 – 1995. *Links:* Differenzen-Methode nach (70). *Rechts:* Standardisierung nach (71). Statische binäre Median-Partitionierung, Wortlänge 8.

Insgesamt liegt eine systematische Abweichung (höhere Informationen, geringere Komplexitäten, außer C_{EM}) der Jahresgang-bereinigten Werte gegenüber denen der Originaldaten vor. Da der Index i_S im Mittel kleiner als 1 ist, kann über die Signifikanz dieser Abweichung gestritten werden oder mit einem statistischen Test entschieden werden. Abb. 3-2 zeigt eine systematische Abweichung des Informationsgewinns bei fast jeder unterstellten Periode zur Entsaisonalisierung. Die größte Abweichung wird bei den Jahresvielfachen erreicht. Bei der Differenzenmethode ergibt sich ein periodischer Verlauf, bei dem der H_G-Wert der Originaldaten bei den Halbjahres-Vielfachen erreicht wird. Diese Beobachtung wurde auch bei den anderen Komplexitätsmaßen gemacht. Die Werte nach dem Standardisierungsverfahren lassen vermuten, dass die unterstellte Periode nicht sehr wesentlich ist. Die Differenzen-Entsaisonalisierung in Abb. 3-2 bestätigt jedoch den Jahresgang als die überragende Periode in den Daten.

Die Untersuchungen in diesem Abschnitt bestätigen insgesamt die dem Hydrologen und Wasserwirtschaftler bekannte Tatsache, dass der Abfluss einen deutlichen Jahresgang enthält. Die Zeitreihe ist also nicht stationär. Aber muss deswegen generell eine Entsaisonalisierung vor der Analyse durchgeführt werden? Der Jahresgang im Abfluss ist durch den Jahresgang der Temperatur bedingt, der eine höhere Verdunstung und insbesondere die Transpiration der Vegetation im Sommer zur Folge hat. Aber genau die soll in dieser Arbeit u. a. untersucht werden: Eine veränderte Transpiration, die sich im Abfluss eines Einzugsgebietes zeigt, ist ein Bestandes-Hinweis auf eine Veränderung der biologischen Aktivität in dem Gebiet. Dies wird in Abschnitt 5.2 untersucht. In Abschnitt 5.1 wird gezeigt, dass die typische Struktur des Abflusses, zu der auch der Jahresgang gehört, erst beim Durchlaufen des Wassers durch das System entsteht. Daher ist eine generelle Jahresgang-Bereinigung der Daten vor der Analyse nicht im Sinne der hier untersuchten Fragestellungen.

Ein weiterer Grund, der gegen eine generelle Entsaisonalisierung spricht, sind Artefakte der Verfahren, welche die Untersuchung stören können. Die beiden hier verwendeten Methoden prägen den Daten eine ungewollte Struktur auf, die bei der Differenzen-Methode als Spitze in der Autokorrelation sichtbar wird und bei der Standardisierung durch eine gewisse Gleichgültigkeit / Insensibilität gegenüber der unterstellten Periode in den Daten. Es gibt noch andere Verfahren zur Entsaisonalisierung — etwa Spektral-Methoden. Es liegt allerdings nicht im Fokus dieser Arbeit eine entsprechende störungsfreie Methode zu finden, da auf eine Entsaisonalisierung generell verzichtet werden soll (s. o.).

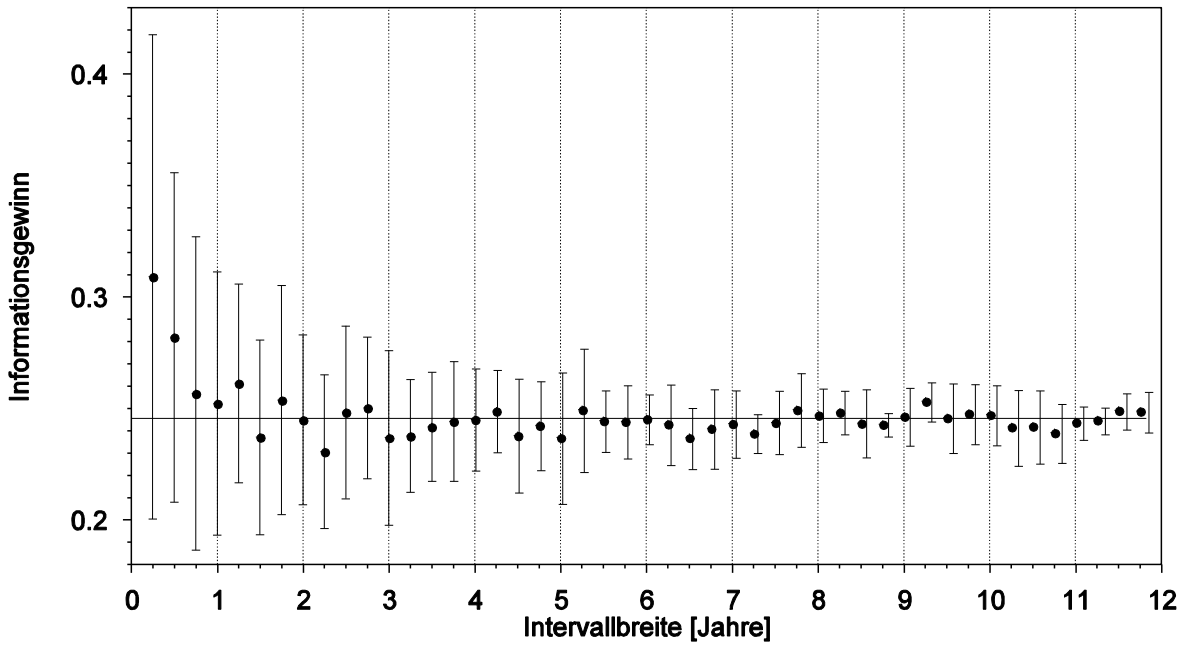

Abb. 3-3. Mittelwerte und Standardabweichungen des Informationsgewinns von benachbarten Intervallen verschiedener Breite für den täglichen Abfluss der Langen Bramke 1948 – 1995. Die Waagerechte gibt den Wert für den vollen Zeitraum an. Mit zunehmender Intervallbreite nimmt die Anzahl der Intervalle ab. Statische binäre Median-Partitionierung, Wortlänge 4.

Dennoch muss die Instationarität der Daten durch den Jahresgang beachtet werden. So sollten die Komplexitätsmaße stets über Zeitabschnitte, die einem Vielfachen der Periode entsprechen (vgl. Tabelle 3-1), oder die deutlich länger als die Periode sind, berechnet werden. Abb. 3-3 zeigt, dass die Mittelwerte derart berechneter Komplexitätsmaße kaum von dem über die gesamte Zeitreihe gewonnenem Wert abweicht. Die Schwankung solcher Werte ist natürlich um so höher, je kürzer die betrachteten Zeitintervalle sind. Ab einer Periode von vier Jahren sind die Informationswerte für den Lange-Bramke-Datensatz relativ stabil.

Bei den kurzen Intervallen (z. B. 91 Tage) in Abb. 3-3 ist eine mittlere Genauigkeit von 5 % für die verwendete Wortlänge 4 nach Tabelle 7-4 nicht mehr gewährleistet. Eine zu hohe Wortlänge würde sich in einem geringeren Wert für den Informationsgewinn äußern. Tatsächlich ist dieser aber höher. Die geringe Datenmenge wirkt sich hier also noch nicht limitierend aus, weil die Information insgesamt nicht so hoch ist. Die Tabellen 7-1 bis 7-6 wurden für den „schlimmsten“ Fall (maximale Information), der am meisten Daten erfordert, erstellt.

3.4 Messlücken

Messlücken in den Daten müssen durch einen besonderen Wert außerhalb des Messbereichs gekennzeichnet sein. Im Symbolsatz weist SYMDYN diesem Wert ein spezielles Lückensymbol außerhalb des Alphabets (2) zu. In der Verteilung der Wörter werden nur die lückenfreien Wörter berücksichtigt (siehe 2.1.2). Die meisten in Kapitel 2 vorgestellten Methoden basieren auf dieser Wort-Verteilung. Wenn es nicht zu viele verstreute Lücken in den Daten gibt und die Wortlänge nicht zu groß ist — eine Lücke infiziert L Wörter der Länge L —, bleiben somit noch genügend Wörter zur Berechnung der Maße übrig.

Die einzigen in Kapitel 2 vorgestellten relevanten Methoden, die nicht auf einer Wort-Verteilung basieren, sind die Korrelationsfunktionen und die Algorithmische Information. Die Korrelationsfunktionen aus Abschnitt 2.2 werden nur anhand der lückenfreien Daten- oder Symbol-Paare berechnet. Die Datenmenge in den Formeln reduziert sich entsprechend. Bei der Algorithmischen Information werden Lücken als kopierbare Symbole interpretiert, da dies auf den Großteil der Symbole zutrifft. Damit wird das Maß eventuell leicht unterschätzt. Die Alternativen wären, die Lücken auf Kosten der Äquidistanz zu überspringen oder sie als neue Komponente zu verstehen, was das Maß mit jeder Lücke mehr überschätzt.

Insgesamt bestätigt die vorgenommene Lückenbehandlung die Erwartung, dass nur wenige Messausfälle im Vergleich zur Datenmenge die Berechnung der Maße kaum beeinflusst. Dabei ist allerdings die Verteilung der Lücken wichtig. Wenn mehrere Lücken isoliert über den Messzeitraum verstreut auftreten, ist deren Einfluss auf die Algorithmische Information und Maße, die auf langen Wörtern basieren, größer als wenn dieselbe Anzahl von Lücken in einem Block hintereinander auftritt, falls in dem ausgefallenen Zeitraum kein bisher unbeobachtetes Verhalten auftritt (siehe 3.2). Es ist leider unmöglich für alle Prozesse und Verteilungen von Messlücken dessen Auswirkung auf die Analyse-Methoden zu quantifizieren.

Unter der Annahme, dass der Effekt bei informationsreichen Prozessen (Bernoulli, gleichverteilt) am stärksten ist und die Lücken-Verteilung im wesentlichen 2-parametrig ist, d. h. es gibt eine Wahrscheinlichkeit für einen unerwarteten Messausfall und eine Wahrscheinlichkeit, dass die nächste Messung dann auch ausfällt, könnten Obergrenzen für die Anzahl der Lücken bei einer bestimmten Genauigkeit (für den relativen Fehler) in einer 2-dimensionalen Matrix aufgetragen werden.

3.5 Schwankungen der Werte

Obwohl in der Praxis oft nur eine historisch einmalige Zeitreihe vorliegt, wird sie als Realisation eines stochastischen Prozesses betrachtet (HIPEL & MCLEOD, 1994), dessen Mittelwert, Standardabweichung und andere Momente unbekannt sind. Um einen signifikanten Unterschied in der Information oder Komplexität zweier Zeitreihen feststellen zu können, sollte zumindest die Standardabweichung dieser Werte bekannt sein.

WITT (1996, S. 35) hat die zufälligen Fehler der Lempel-Ziv-Komplexität (siehe 2.5.6) für 3 theoretische Prozesse, dessen statistische Eigenschaften bekannt sind, untersucht. Die Fehler hingen nicht-linear sowohl von dem Wert der Lempel-Ziv-Komplexität als auch von dem Prozess ab. Dieses Beispiel zeigt, dass es unmöglich ist, ohne Kenntnis des zugrunde liegenden Prozesses auf die Abweichungen der Werte eines Komplexitätsmaßes zu schließen.

Die Schwankungen der Werte können also nur geschätzt werden, indem die zumeist ohnehin schon kurze Messreihe in benachbarte Teilintervalle gleicher Länge zerlegt wird, in denen das Maß jeweils berechnet wird. Von diesen Werten können dann Mittelwert und Standardabweichung berechnet werden. Die Länge der Teilintervalle sollte oberhalb der relevanten Zeiskala (Korrelationslänge) liegen. Für die hier untersuchten Daten mit Saisonalität (Tagesgang, Jahresgang) wird ein kleines Vielfaches der Periode empfohlen. Da die Berechnung von Komplexitätsmaßen grundsätzlich viele Datenpunkte erfordert, kann dieses Vorgehen (empfindliche) Konsequenzen für die Berechenbarkeit der Maße haben (siehe 3.6).

3.6 Erforderliche Datenmenge und maximale Wortlänge

KASPAR & SCHUSTER (1987) stellten fest, dass der theoretische Wert der Algorithmischen Information für Zufallsfolgen erst ab etwa 1000 Datenpunkten mit einer Genauigkeit von 5 % erreicht wird. Derartige Angaben konnten für die anderen hier verwendeten Methoden in der verwendeten Literatur nicht gefunden werden. Sie sind jedoch für die Anwendbarkeit von Komplexitätsmaßen auf Messdaten, bei denen die Datenmenge oft gering ist, von fundamentalem Interesse. In diesem Abschnitt soll diese Lücke geschlossen werden.

Alle hier verwendeten Komplexitätsmaße, mit Ausnahme der Algorithmischen Information,

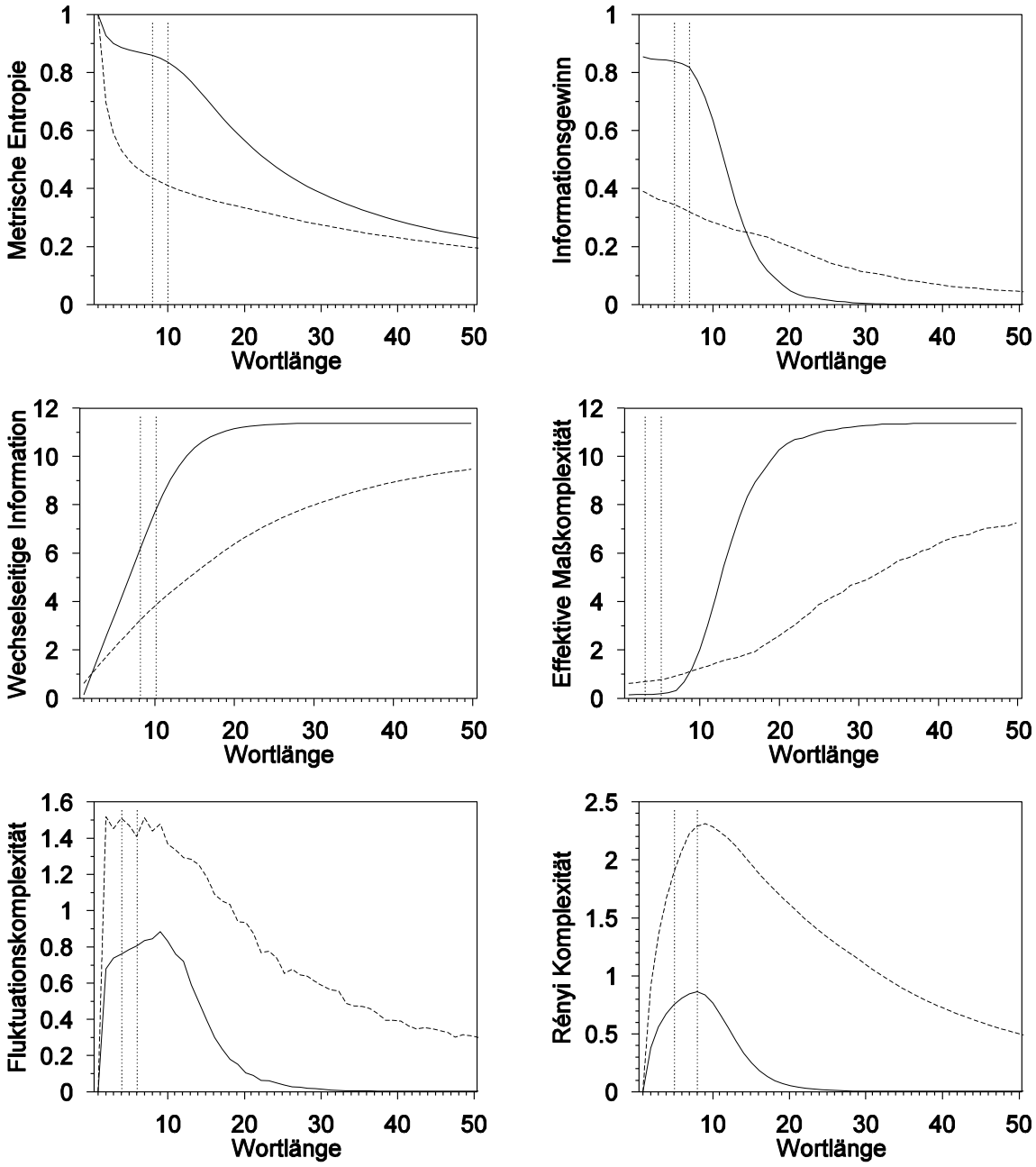

Abb. 3-4. Abhängigkeit von Komplexitätsmaßen experimenteller Zeitreihen von der Wortlänge bei endlicher Datenmenge. Niederschlag (durchgezogen) und Abfluss (gestrichelt) des Lehstenbaches 1987 – 1995 (2921 Datenpunkte). Maximale Wortlänge für 1 % (linke Senkrechte) und 5 % (rechte Senkrechte) mittlere Genauigkeit nach Tabellen 7-1 und 7-4.

basieren auf einer Verteilung von *L*-Wörtern (siehe 2.1.2). Diese wird aus den relativen Häufigkeiten der Wörter geschätzt. Bei *N* Datenpunkten gibt es *N*–*L*+1 Wörter. Theoretisch sind maximal λ^L verschiedene Wörter möglich. Daher können bereits bei kleinen Wortlängen nicht mehr alle theoretisch möglichen Wörter (genügend oft) beobachtet werden. Zur Berechnung eines Komplexitätsmaßes mit einer bestimmten Genauigkeit sind um so mehr Daten erforderlich, je umfangreicher das Alphabet λ (siehe 2.1.1) ist und je länger die Wörter *L* sind. Umgekehrt gibt es bei gegebener Datenmenge eine maximale Wortlänge $L_{\max}$, die nicht überschritten werden darf, damit die Werte der Maße eine bestimmte Genauigkeit einhalten. Um möglichst viel von der typischen Struktur der Zeitreihe zu erfassen, ist man an möglichst hohen Wortlängen zur Auswertung der Maße interessiert. Die Wortlänge als Parameter der Komplexitätsmaße ist damit auf $L_{\max}$ festgelegt. Einen Eindruck von der Auswirkung der endlichen Datenmenge auf die Werte von Komplexitätsmaßen bei unterschiedlicher Wortlänge vermittelt Abb. 3-4.

Nach KHINCHIN (1957, S. 57) gibt es bei großen Wortlängen und theoretisch unendlichen Datenmengen ungefähr

$$\lambda^{hL} \tag{73}$$

verschiedene Wörter. Dabei bezeichnet *h* die Entropie der Quelle (37). Die Approximation von *h* durch die Metrische Entropie $H_\mu(L) = H_S(L)/L$ nach (35) mit der Shannon-Entropie H_S nach (26) wird von EBELING et al. (1995) als Modifikation von (73) berücksichtigt:

$$\lambda^{H_S(L)} \tag{74}$$

Beide Formeln sagen aus, dass die Anzahl vorhandener Wörter schon bei um so kleineren Wortlängen erreicht wird, je höher die Entropie und damit die Unordnung der Daten ist. Dies entspricht der intuitiven Erwartung. Die Entropie ist aber gerade eine der zunächst unbekannten Größen, die (mit $L_{\max}$) erst bestimmt werden sollen. (Methoden zur Extrapolation der Entropie werden in Abschnitt 3.7 beschrieben.) Außerdem sollte bei einem Vergleich von Komplexitäten verschiedener Datensätze eine einheitliche Wortlänge verwendet werden, da z. B. wegen der monotonen Konvergenz der relativen Entropien (siehe 2.5.4 und LANGE et al., 1998) niedrige Werte näher am Grenzwert *h* liegen, wenn höhere Wortlängen verwendet werden. Diese beiden Argumente sprechen dafür, die erforderliche Datenmenge und maximale Wortlänge für den Fall maximaler Entropie, $h = 1$, $H_S(L) = L$, also wenn der Wörtervorrat nach (73) und (74) am schnellsten erschöpft ist, zu ermitteln. Dies ist der gleichverteilte Bernoulliprozess (Zufallsprozess).

Bei einem gleichverteilten Zufallsprozess sind alle *L*-Wörter gleich häufig mit der Wahrscheinlichkeit $p_L = 1/\lambda^L$. Es sei $q_L \square 1–p_L$. Die Wahrscheinlichkeit unter *N*–*L*+1 Wörtern ein bestimmtes *L*-Wort w_i *k*-mal zu beobachten ist gemäß einer Binomial-Verteilung:

$$p_{L,i}(k) = \binom{N-L+1}{k} p_L^k q_L^{N-L+1-k} \tag{75}$$

(Dies ist auch der Ausgangspunkt für die Approximationsformeln von SCHMITT et al., 1993; HERZEL et al., 1994; SCHMITT & HERZEL, 1997.) Der Erwartungswert einer Funktion *f*, die mit der beobachteten Häufigkeit $k/(N–L+1)$ dieses Wortes anstelle von dessen Wahrscheinlichkeit p_L ausgewertet wird, ist dann:

$$f_{\exp} = \sum_{k=0}^{N-L+1} p_{L,i}(k)\, f\left(\tfrac{k}{N-L+1}\right) \tag{76}$$

Für einen integralen Wert F von f als Summe über alle Wörter muss (76) noch mit der theoretischen Anzahl verschiedener Wörter multipliziert werden:

$$F_{\exp} = \lambda^L f_{\exp} \tag{77}$$

Auf diese Weise können die Erwartungswerte der Wort-Komplexitätsmaße in Abhängigkeit von λ, L und N ermittelt werden. Für die Shannon-Entropie H_S entspricht f den Summanden aus Formel (26). Es gilt dann gemäß (77) wie in Anhang 7.8 gezeigt:

$$H_{S,\exp} = \log_2(N-L+1) - \sum_{k=0}^{N-L} \binom{N-L}{k} p_L^k q_L^{N-L-k} \log_2(k+1) \tag{78}$$

Diese Formel wird auch für die Berechnung der Erwartungswerte der Metrischen Entropie H_μ nach (35) sowie des Informationsgewinns H_G, der Wechselseitigen Information H_M und der Effektiven Maßkomplexität C_{EM} nach den Differenzenformeln (42), (45) und (55) verwendet.

Für die Rényi-Entropie H_R nach (30), $\alpha \neq 1$, ergibt sich mit (77):

$$H_{R,\exp}(\alpha) = \frac{1}{1-\alpha} \log_2 \left[\lambda^L \sum_{k=1}^{N-L+1} \binom{N-L+1}{k} p_L^k q_L^{N-L+1-k} \left(\frac{k}{N-L+1} \right)^\alpha \right] \tag{79}$$

Diese Gleichung lässt sich mit der Funktionalgleichung des Logarithmus etwas umschreiben und wird zur Berechnung des Erwartungswertes der Rényi-Komplexität C_R nach (63) verwendet. Auch hier stellte sich heraus, dass numerisch stabile Werte von C_R nur durch $C_R(\alpha)$ mit $\alpha = 1.0001$, nicht kleiner, garantiert sind (siehe 2.6.3).

In den kompakten Formeln (41), (44) und (56) für H_G, H_M und C_{EM} sowie bei der Fluktuationskomplexität C_Γ werden neben den Wahrscheinlichkeiten $p_{L,i}$ der L-Wörter auch die Wahrscheinlichkeiten $p_{L,j}$ der darauf folgenden Wörter oder der bedingten oder verknüpften Wahrscheinlichkeiten $p_{L,i\to j}$ und $p_{L,ij}$ verwendet. Diese werden in SYMDYN durch relative Häufigkeiten in nur einem Baum der Tiefe (L+1) ermittelt, wie in Abb. 2-5 dargestellt wurde. Die Anzahl der (L+1)-Wörter ist N–L und bestimmt auch die der L-Wörter im (L+1)-Baum. Die Wahrscheinlichkeit unter den N–L L-Wörtern mit den theoretischen Wahrscheinlichkeiten $1/\lambda^L$ ein Wort k-mal zu beobachten ist analog zu (75):

$$\binom{N-L}{k} p_L^k q_L^{N-L-k} \tag{80}$$

Die Wahrscheinlichkeit, danach ein bestimmtes Symbol l-mal zu beobachten, $l \le k$, ist wegen der Unabhängigkeit der Ereignisse des Bernoulli-Prozesses:

$$\binom{k}{l} p_1^l q_1^{k-l} \tag{81}$$

In diesem Fall ist $p_{L,i} \approx k/(N-L)$, $p_{L,ij} \approx l/(N-L)$ und $p_{L,i\to j} = l/k$. Damit lässt sich der Erwartungswert des Informationsgewinns $H_{G,\exp,c}$ nach (41) analog zu (77) berechnen:

$$H_{G,\exp c} = \lambda^{L+1} \sum_{k=1}^{N-L} \binom{N-L}{k} p_L^k q_L^{N-L-k} \sum_{l=1}^{k-1} \binom{k}{l} p_1^l q_1^{k-l} \frac{l}{N-L} \log_2 \frac{k}{l} \tag{82}$$

Die Indizes $l = 0$ und $l = k$ wurden nicht aufgeführt, da sie keinen Beitrag zu der Summe liefern. Der Index „c“ in $H_{G,\exp,c}$ weist auf die kompakte Formel (41) für H_G hin.

Auf die gleiche Weise lässt sich der Erwartungswert der Approximation der Effektiven Maßkomplexität $C_{\mathrm{EM,exp,c}}$ nach (56) berechnen:

$$C_{\mathrm{EM,exp,c}} = \lambda^{L+1} \sum_{k=1}^{N-L} \binom{N-L}{k} p_L^k q_L^{N-L-k} \sum_{l=1}^{k-1} \binom{k}{l} p_1^l q_1^{k-l} \frac{l}{N-L} \log_2 \frac{(N-L)l^L}{k^{L+1}} \tag{83}$$

Zur Berechnung der Erwartungswerte der kompakten Form der Wechselseitigen Information sowie der Fluktuationskomplexität wird zusätzlich $p_{L,j}$ benötigt. Dazu wird zuerst die Häufigkeit der ersten Symbols von Wort w_i ermittelt. Dieses wird mit der Wahrscheinlichkeit

$$\binom{N-L}{k} p_1^k q_1^{N-L-k} \tag{84}$$

k-mal beobachtet. Anschließend werden die gemeinsamen (L–1) Symbole der Wörter w_i und w_j betrachtet. Dieses (L–1)-Wort wird in Wort w_i mit der Wahrscheinlichkeit

$$\binom{k}{l} p_{L-1}^l q_{L-1}^{k-l} \tag{85}$$

l-mal beobachtet, wobei $l \le k$ ist. Das letzte Symbol kann dann nur noch m-mal beobachtet werden, mit $m \le l$. Die Wahrscheinlichkeit dafür ist:

$$\binom{l}{m} p_1^m q_1^{l-m} \tag{86}$$

Damit ist $p_{L,i} = l/(N–L)$, $p_{L,ij} = m/(N–L)$, $p_{L,i\to j} = m/l$ und $p_{L,j} = p_{L,ij}/p_{1,i} = m/k$. Bei der letzten Gleichung wurde die Unabhängigkeit der Bernoulli-Ereignisse ausgenutzt. Damit lässt sich der Erwartungswert der kompakten Formel (44) der Wechselseitigen Information $H_{\mathrm{M,exp,c}}$ wie gewohnt berechnen:

$$H_{\mathrm{M,exp,c}} = \lambda^{L+1} \sum_{k=1}^{N-L} \binom{N-L}{k} p_1^k q_1^{N-L-k} \sum_{l=1}^{k} \binom{k}{l} p_{L-1}^l q_{L-1}^{k-l} \sum_{m=1}^{l} \binom{l}{m} p_1^m q_1^{l-m} \frac{m}{N-L} \log_2 \frac{k}{l} \tag{87}$$

Entsprechend gilt für den Erwartungswert der Fluktuationskomplexität $C_{\Gamma,\mathrm{exp}}$ nach (60):

$$C_{\Gamma,\mathrm{exp}} = \lambda^{L+1} \sum_{k=1}^{N-L} \binom{N-L}{k} p_1^k q_1^{N-L-k} \sum_{l=1}^{k} \binom{k}{l} p_{L-1}^l q_{L-1}^{k-l} \sum_{m=1}^{l} \binom{l}{m} p_1^m q_1^{l-m} \frac{m}{N-L} \left(\log_2 \frac{lk}{(N-L)m} \right)^2 \tag{88}$$

Formeln (87) und (88) sind nur für $L > 1$ gültig. Für $L = 1$ verschwindet die mittlere Summe und das Argument im Logarithmus bei $H_{\mathrm{M,exp,c}}$ ist 1 und damit gilt $H_{\mathrm{M,exp,c}} \equiv 0$. Die Fluktuationskomplexität liefert für $L = 1$ keine charakteristischen Werte (siehe 2.6.2) und wird nur der Vollständigkeit halber für $L = 1$ erwähnt:

$$C_{\Gamma,\mathrm{exp}} = \lambda^{2} \sum_{k=1}^{N-2} \binom{N-2}{k} p_1^k q_1^{N-2-k} \sum_{l=1}^{k} \binom{k}{l} p_1^l q_1^{k-l} \frac{l}{N-2} \left(\log_2 \frac{k^2}{(N-2)l} \right)^2 \tag{89}$$

Die Gültigkeit der hier dargestellten Formeln über die Erwartungswerte von Komplexitätsmaßen wurde empirisch überprüft. Die Funktionen (Maß über Wortlänge) vermitteln zwischen den theoretischen Grenzkurven bei kleinen Wortlängen, wenn die relativen Häufigkeiten der L-Wörter gut den Wahrscheinlichkeiten $1/\lambda^L$ entsprechen, und den Sättigungsgrenzen, wenn jedes beobachtete Wort einmalig mit der Häufigkeit $1/(N–L+1)$ ist.

Die theoretischen Werte der betrachteten Maße für den gleichverteilten Bernoulli-Prozess sind:

$$\begin{aligned} H_{\mathrm{S,theo}} &= L\log_2\lambda \\ H_{\mathrm{R,theo}} &= L\log_2\lambda \\ H_{\mu,\mathrm{theo}} &= \log_2\lambda \\ H_{\mathrm{G,theo}} &= \log_2\lambda \\ H_{\mathrm{M,theo}} &= (L-1)\log_2\lambda \\ C_{\mathrm{EM,theo}} &= 0 \\ C_{\Gamma,\mathrm{theo}} &= 0 \\ C_{\mathrm{R,theo}} &= 0 \end{aligned} \tag{90}$$

Das Kriterium zur Berechnung der erforderlichen Datenmenge soll die Genauigkeit ε für die relative Abweichung des zu erwartenden Wertes C_{exp} vom theoretischen Wert C_{theo} sein:

$$\frac{\left|C_{\mathrm{theo}} - C_{\mathrm{exp}}\right|}{\max\left\{1, C_{\mathrm{theo}}\right\}} \leq \varepsilon \tag{91}$$

Die Formulierung (91) ist speziell für die hier vorliegende Situation und drückt aus, dass es sich bei den Informationsmaßen um relative Genauigkeiten handelt, während für die Komplexitätsmaße wegen (90) nur eine absolute Genauigkeit gefordert werden kann. Bei vorgegebener Alphabetgröße λ und Genauigkeit ε läßt sich anhand der in diesem Abschnitt hergeleiteten Formeln für jede Wortlänge L die mindestens erforderliche Datenmenge N_{min} bestimmen, für die das Kriterium (91) eingehalten wird. Bei $L = 1$ bedeutet dies die Datenmenge, die erforderlich ist, um ein Maß mit Sicherheit (auch bei maximaler Information) anwenden zu können. Wenn N_{min} für verschiedene Wortlängen bekannt ist, lässt sich daraus für eine bestimmte Datenmenge N auch die maximale Wortlänge L_{max} ableiten.

Leider kann (91) in keinem Fall analytisch nach N oder L aufgelöst werden. Es gibt daher keine exakte Formel für N_{min} oder L_{max}. Diese Größen müssen numerisch bestimmt werden. Auch dies ist mit der hier dargestellten Form der Formeln nicht möglich: Bei großen Werten von N verursachen die Binomialkoeffizienten einen Speicherüberlauf[7] und die Potenzen der p_n das Gegenteil. Daher wurden diese Größen im Logarithmus berechnet und beim Einsetzen zurück transformiert (vgl. PRESS et al., 1992, S. 215). Die Binomialkoeffizienten wurden iterativ nach (1.35e) aus BRONŠTEJN et al. (1997, S. 10) berechnet, um Zeit zu sparen. Trotzdem war die Berechnungszeit von den Mehrfachsummenformeln nicht akzeptabel. Durch Schätzung der Erwartungswerte der Summationsindizes und Beschränkung auf die relevanten Summanden konnten auch die Dreifachsummen auf einem Pentium 130 MHz oder 200 MHz bei N bis > 1000000 in wenigen Tagen berechnet werden.

N_{min} wurde für die hier relevanten Alphabetgrößen $\lambda = 2$, 3 und 4 sowie für die üblichen Genauigkeiten 1 % und 5 % ($\varepsilon = 0.01$, 0.05) und die Wortlängen $L = 1$, 2, 3,... ermittelt, bis eine Datenmenge von $N_{\mathrm{min}} > 1000000$ erreicht war. Damit liegen die für die Praxis relevanten Werte von N_{min} für die hier verwendeten Komplexitätsmaße vor. Die erforderlichen Datenmengen N_{min} sind in den Tabellen 7-1 bis 7-6 im Anhang aufgeführt und liegen in SYMDYN zur Berechnung der maximalen Wortlänge für die jeweilige Datenmenge vor. Abb. 3-5 stellt die Werte für binäre Alphabete und 5 % Genauigkeit dar.

[7] Die größte Maschinenzahl lag bei 10^{306} und die kleinste bei 10^{-306}.

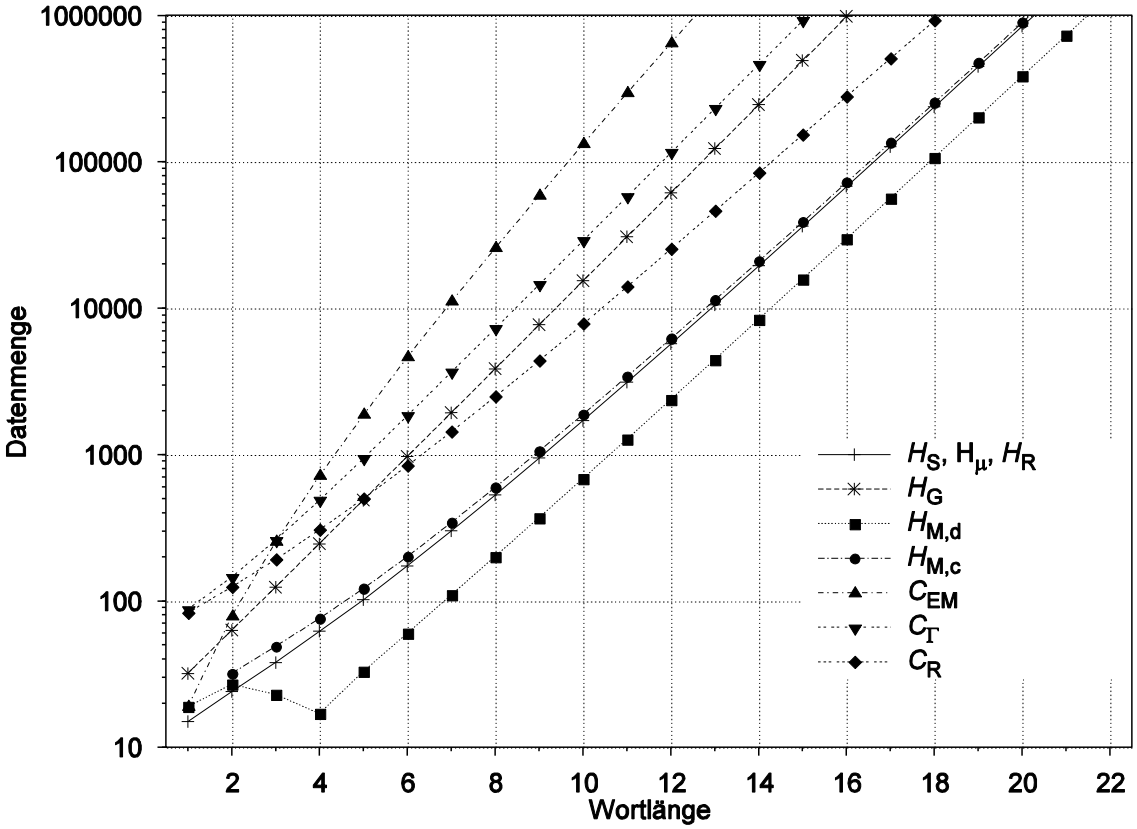

Abb. 3-5. Erforderliche Datenmengen zur Berechnung von Komplexitätsmaßen in Abhängigkeit von der Wortlänge für binäre Alphabete und 5 % mittlere Genauigkeit nach Tabelle 7-4.

Die in den Tabellen dargestellten Ergebnisse sind für die Maße mit nicht-verschwindendem theoretischem Wert (90) unabhängig von der Basis des Logarithmus, weil sich diese in (91) herauskürzt. Ebenso kürzt sich die Wortlänge in metrischer Entropie und Shannon-Entropie weg, so dass beide Maße die gleiche Datenmenge erfordern. Diese Datenmenge fordert auch die Rényi-Entropie in der Nähe der Ordnung $\alpha = 1$ wegen ihrer Konvergenz gegen die Shannon-Entropie. Bei einer Verdopplung von $\alpha \geq 1$ verdoppelt sich auch die geforderte Datenmenge $N_{\min}$. Bei Halbierung von $0 < \alpha \leq 1$ halbiert sich auch $N_{\min}$. Dies ist ein Artefakt der angenommenen Gleichverteilung, da kleine α (Wurzeln) den Charakter der Gleichverteilung begünstigen, während große α (Potenzen) die Verteilung zum anderen Extrem verzerren. Eine allgemein gültige Aussage über die erforderliche Datenmenge von $H_R(\alpha)$ für beliebige α und beliebige Verteilungen ist daher nicht möglich. $N_{\min}$ dürfte aber an den Werten für die Shannon-Entropie orientiert sein, wenn α in der Nähe von 1 liegt.

Eine interessante Feststellung ist, dass die Erwartungswerte des Informationsgewinns und der Effektiven Maßkomplexität, die mit den kompakten Formeln (41) und (56) gewonnen wurden, im unterkritischen Wortlängenbereich mit denen der Differenzenformeln (42) und (55) übereinstimmen, darüber hinaus jedoch abweichen. Dies wurde auch bei der Anwendung der Maße auf die hydrologischen Zeitreihen festgestellt. Daher sind die erforderlichen Datenmengen in beiden Fällen jeweils gleich. Bei der Wechselseitigen Information ist das anders: Die Erwartungswerte nach der kompakten Formel (44) nähern sich bei jeder Wortlänge streng monoton von unten mit zunehmender Datenmenge N an den theoretischen Wert $H_{M,theo}$ an. Die Erwartungswerte $H_{M,exp}$ nach der Differenzenformel (45) nähern sich für $L = 1$ und $L = 2$ streng monoton von oben den theoretischen Werten. Bei $L > 2$ erfolgt zunächst eine Approximation von unten, dann wird $H_{M,theo}$ überschritten, bis $H_{M,exp}$ ein Maximum erreicht und

sich bei weiter zunehmendem N von oben $H_{\mathrm{M,theo}}$ nähert. Dieses Maximum kann bei kleinem ε und L über dem Toleranzbereich um $H_{\mathrm{M,theo}}$ liegen. Bei höheren Wortlängen verlässt es den Toleranzbereich nicht mehr. Dieses besondere Konvergenzverhalten erklärt den maximal zweimaligen Rückgang der erforderlichen Datenmenge bei zunehmender Wortlänge. Dies ist eine einmalige Kuriosität der Wechselseitigen Information nach (45), welche die prinzipielle exponentielle Zunahme der erforderlichen Datenmenge mit der Wortlänge verzerrt darstellt. Daher sollte dieses Maß gemäß seiner ursprünglichen Definition (44) berechnet werden.

Prinzipiell nimmt die erforderliche Datenmenge bei allen Maßen exponentiell mit der Wortlänge gemäß der Khinchin‘ schen Formel (73) zu. Ein exakt exponentieller Zusammenhang würde in der logarithmischen Darstellung in Abb. 3-5 einen linearen Kurvenverlauf bedeuten. Tatsächlich sind die „Geraden“ in Abb. 3-5 aber mehr oder weniger gekrümmt. Eine Approximation der Tabellenwerte erfordert also eine Erweiterung des Funktionstyps. Ein geeigneter Ansatz dafür ist:

$$N = \alpha_1 \lambda^{\alpha_2 L + \alpha_3 L^2} \tag{92}$$

wobei Verbesserungen durch höhere Polynomgrade im Exponenten erreicht werden. Die Parameter (α_1, α_2, α_3) müssen für jedes Maß an die Werte angepasst werden. Die Genauigkeit ε wirkt sich im Wesentlichen nur auf α_1 (reziprok) aus. Wenn die Parameter nach der Methode der kleinsten Quadrate angepasst werden, sind sie universell für alle Alphabetgrößen λ. Kleine N werden dann aber systematisch unterschätzt. Eine Normierung der quadratischen Fehler durch das Quadrat der Tabellenwerte führt zu einer ausgeglichenen Approximation auf Kosten der Universalität für λ.

Die größten Ansprüche an die Datenmengen schon bei kleinen Wortlängen haben die Komplexitätsmaße, allen voran die Fluktuationskomplexität. Sie erfordert mindestens 146 Datenpunkte für eine Genauigkeit von 5 % sowie minimales Alphabet und Wortlänge 2 (siehe 2.6.2). Ihre Ansprüche werden jedoch bald von denen der Effektiven Maßkomplexität übertroffen. Unter den Informationsmaßen erfordert der Informationsgewinn die meisten Daten: mindestens 32 für 5 % Genauigkeit. Shannon-, Metrische und Rényi-Entropie haben insgesamt die geringsten Ansprüche an die Datenmenge (mindestens 15 für $\varepsilon = 0.05$).

Wenn ein hoher Informationsgehalt (hohe Zufälligkeit) in den Daten sicher ausgeschlossen werden kann, sollten die Datenmengen für 5 % Genauigkeit ausreichen, da sich bei gleicher Datenmenge und Wortlänge mit abnehmender Zufälligkeit gemäß (73) und (74) die Genauigkeit durch eine abnehmende Anzahl verschiedener Wörter verbessert. Für die informationsreichen DNA-Sequenzen müssen die Werte in den Tabellen mit $\lambda = 4$ und $\varepsilon = 0.01$ beachtet werden. EBELING et al. (1995) geben für Texte mit einem 32er-Alphabet und einigen Millionen Zeichen eine Obergrenze von $L = 30$ für die Wortlänge an, welche nur aufgrund der Regelhaftigkeit (Grammatik, Semantik) möglich ist. Wenn man jedoch Zeitreihen von Niederschlag und Abfluss mit gleicher Wortlänge vergleichen will, sollte man sich an den hier ermittelten Wortlängen orientieren.

3.7 Extrapolationen und Korrekturformeln

Es wurden verschiedene Methoden zur Verbesserung der Schätzung von (Shannon-) Entropien und verwandten Maßen, die über eine Verteilung von Wörtern berechnet werden, vorgeschlagen. Sie basieren jeweils auf einer Approximation der, nach Größe sortierten (Zipf-, Rank-geordneten), Verteilung der beobachteten Worthäufigkeiten. Das fragliche Entropie-

Maß wird dann anhand der Approximation anstelle der beobachteten Wort-Verteilung ausgewertet.

Im einfachsten Fall wird angenommen, dass eine bestimmte Anzahl von Wörtern gleichhäufig ist und die anderen gar nicht vorkommen. Diese Annahme beruht auf dem Theorem der asymptotischen Gleichverteilung bezüglich hoher Wortlängen von McMillan. Sie wurde von SCHMITT et al. (1993) verwendet, um die Berechnung von Shannon-Entropien für DNA-Sequenzen zu verbessern. Wegen der nahen Gleichverteilung der Wörter bei DNA-Sequenzen konnten sie damit die Entropie bei größeren Wortlängen auswerten. Sie bemerkten jedoch selbst, dass diese Methode an nahe Gleichverteilung gebunden ist. SCHMITT & HERZEL (1997) haben am Beispiel des Textes von „Alice im Wunderland“ gezeigt, dass diese Methode für andere Verteilungen kaum eine oder gar keine Verbesserung bringen muss.

HERZEL et al. (1994) schlagen die Approximation der Wortverteilung durch eine Exponential- oder Potenz-Funktion vor, wenn der Funktionstyp gut zur beobachteten Verteilung paßt. In diesem Fall können sie einen analytischen Korrekturterm angeben, mit dem die Berechnung der Shannon-Entropie — in ihrem Fall für Texte — verbessert werden kann. Auch PÖSCHEL et al. (1995) schlagen die Approximation der Verteilung mit einer geeigneten Funktion vor, z. B. stückweise lineare oder Potenz-Funktionen. Sie geben einen Algorithmus zur optimalen Parameterschätzung der Funktion an, die gut zu den Daten passen muss. Die Shannon-Entropie für den Text von „Moby Dick“ ließ sich so noch für deutlich höhere Wortlängen berechnen.

Am Beispiel des Textes von „Alice im Wunderland“ konnten SCHMITT & HERZEL (1997) zeigen, dass auch mit der Methode von PÖSCHEL et al. (1995) die Shannon-Entropie mit nur wenig höherer Wortlänge berechnet werden kann. Sie verzichten daher auf die Annahme einer bestimmten Funktion für die ganze Verteilung. Da hohe Worthäufigkeiten in der Regel eine gute Schätzung für die Wahrscheinlichkeiten liefern, schlagen sie vor, die größten i_0 Häufigkeiten direkt zu übernehmen und anstelle der niedrigeren Häufigkeiten n konstante Wahrscheinlichkeiten anzunehmen. Für jeden Wert i_0 wird n dann so bestimmt, dass die Entropie dieser Verteilung der beobachteten entspricht; gleichzeitig wird eine Erwartungsentropie anhand der künstlichen Verteilung berechnet. Gemäß dem Maximum-Entropie-Prinzip ist die maximale Erwartungsentropie der verbesserte Wert für die beobachtete Entropie. Oberhalb der kritischen Wortlänge ist aber auch diese Korrektur als Schätzung mit Vorsicht zu genießen, da z. B. die Monotonie der Werte mit der Wortlänge nicht mehr gegeben ist. Am Beispiel des Textes von „Alice im Wunderland“ konnten SCHMITT & HERZEL (1997) die Überlegenheit dieser Methode gegenüber der beobachteten Entropie, sowie den Verfahren von SCHMITT et al. (1993) und PÖSCHEL et al. (1995) demonstrieren.

GROßE (1996) und HOLSTE et al. (1998) schlagen eine bessere Approximation der Wortwahrscheinlichkeiten durch Bayes Schätzer anstelle der relativen Worthäufigkeiten vor. Damit lässt sich die Varianz der Shannon-Entropie-Schätzungen verringern, wie GROßE (1996) für das *Caenorhabditis elegans* Chromosom III und für allgemeine Verteilungen zeigt. HOLSTE et al. (1998) geben Bayes Schätzer für die Tsallis- und die Rényi-Entropie an, für die sie ebenfalls eine Abnahme der Varianz nachweisen.

Die genannten Korrekturverfahren sind, wie die Autoren z. T. selbst bemerken, mit Unsicherheiten verbunden. Die allgemein mögliche Erhöhung der maximalen Wortlänge durch die genannten Verfahren wird nicht genannt. Die Wahl einer möglichst großen Wortlänge im Vergleich zu den in 3.6 berechneten Grenzen ist für die hier verwendeten Zeitreihen zudem nicht wesentlich, weil die Komplexitätsmaße im Vertrauensbereich niedriger Wortlängen vergleichsweise stabil sind oder systematisch von der Wortlänge abhängen (Abb. 3-4 und LANGE et al. 1998). Für eine vergleichende Berechnung von Komplexitätsmaßen im Rahmen

dieser Arbeit ist daher eher eine gemeinsame als eine über das garantierte Limit hinausgehende Wortlänge entscheidend. Daher wird in dieser Arbeit von Korrekturmaßnahmen abgesehen. Die Komplexitätsmaße werden nach den in Kapitel 2 genannten Formeln anhand von relativen (Wort-) Häufigkeiten berechnet. Dabei werden die Wortlängen gemäß der Kriterien in Abschnitt 3.6 maximal gewählt.

3.8 Zur Wahl der Partitionierung

Nach den Bemerkungen in Abschnitt 2.1.1.1 reicht zur Partitionierung des Wertebereiches der Daten ein binäres Alphabet aus. Der Partitionierungsparameter sollte Entropie-maximal gewählt werden. Eine Verfeinerung der Partitionierung kann dann durch Wörter der Länge $L > 1$ erreicht werden. Die Wortlänge kann — wie in Abschnitt 3.6 gezeigt — durch einen Maximalwert fixiert werden. Damit ist insgesamt auch die Partitionierung ein wohlbestimmter Parameter der Komplexitätsmaße.

In diesem Abschnitt sollen einige Abhängigkeiten der Maße von der Partitionierung anhand von relevanten Beispielen gezeigt werden, da es von Vorteil sein kann, verschiedene Partitionierungen zu verwenden. Es macht durchaus Sinn auch nicht-Entropie-maximale Partitionierungen zu verwenden, z. B. wird man zur Partitionierung einer Steigung den Parameter 0 wählen, um Anstieg und Abstieg zu unterscheiden, auch wenn das Entropie-Maximum einen anderen Wert vorschlägt. Auch eine Median- oder Quantil-Partitionierung ist sinnvoll, da sie zu einer Gleichverteilung der Symbole, aber in der Regel nicht zu einem Entropie-Maximum führt. Die geeignete Partitionierung hängt letztlich von der jeweiligen Fragestellung ab. Ein universelles Kriterium dafür gibt es nicht. Das Ergebnis einer Untersuchung hängt wesentlich von der Partitionierung ab. Daher sollte diese wohl durchdacht sein.

3.8.1 Binäre Alphabete

Zuerst wird die Abhängigkeit der Komplexitätsmaße von einem Partitionierungsparameter bei binärem Alphabet anhand von Abflusszeitreihen gezeigt. Dazu wurden die täglichen Abflussmessungen der fünf Einzugsgebiete: Lehstenbach, Lange Bramke, Birkenes, Hubbard Brook W 1 und Andrews W 3 untersucht (siehe Kapitel 4). Stellvertretend für die unterschiedlichen Beobachtungen werden die Berechnungen für das Einzugsgebiet des Lehstenbaches (Abb. 3-6) und Hubbard Brook, Watershed 1, (Abb. 3-7) gezeigt.

Die Verteilung der Abflusspegel in Hubbard Brook ist sehr schief: 0 mm/d bis 0.232 mm/d sind mit 23 % am häufigsten. Der Lehstenbach hingegen läuft nie trocken und hat eine etwas breitere Verteilung der Abflusswerte. Dies spiegelt sich auch bei den Komplexitätsmaßen wieder, die bei Hubbard Brook durch eine empfindlichere Abhängigkeit von π_0 als beim Lehstenbach gekennzeichnet sind. Die Verteilung der Abflusshöhen und der Verlauf der Komplexitäten von Lehstenbach und Lange Bramke sind in etwa vergleichbar. Dies gilt auch für die Abfluss-Verteilung und den Komplexitätsverlauf von Hubbard Brook (Watershed 1), Andrews und Birkenes untereinander.

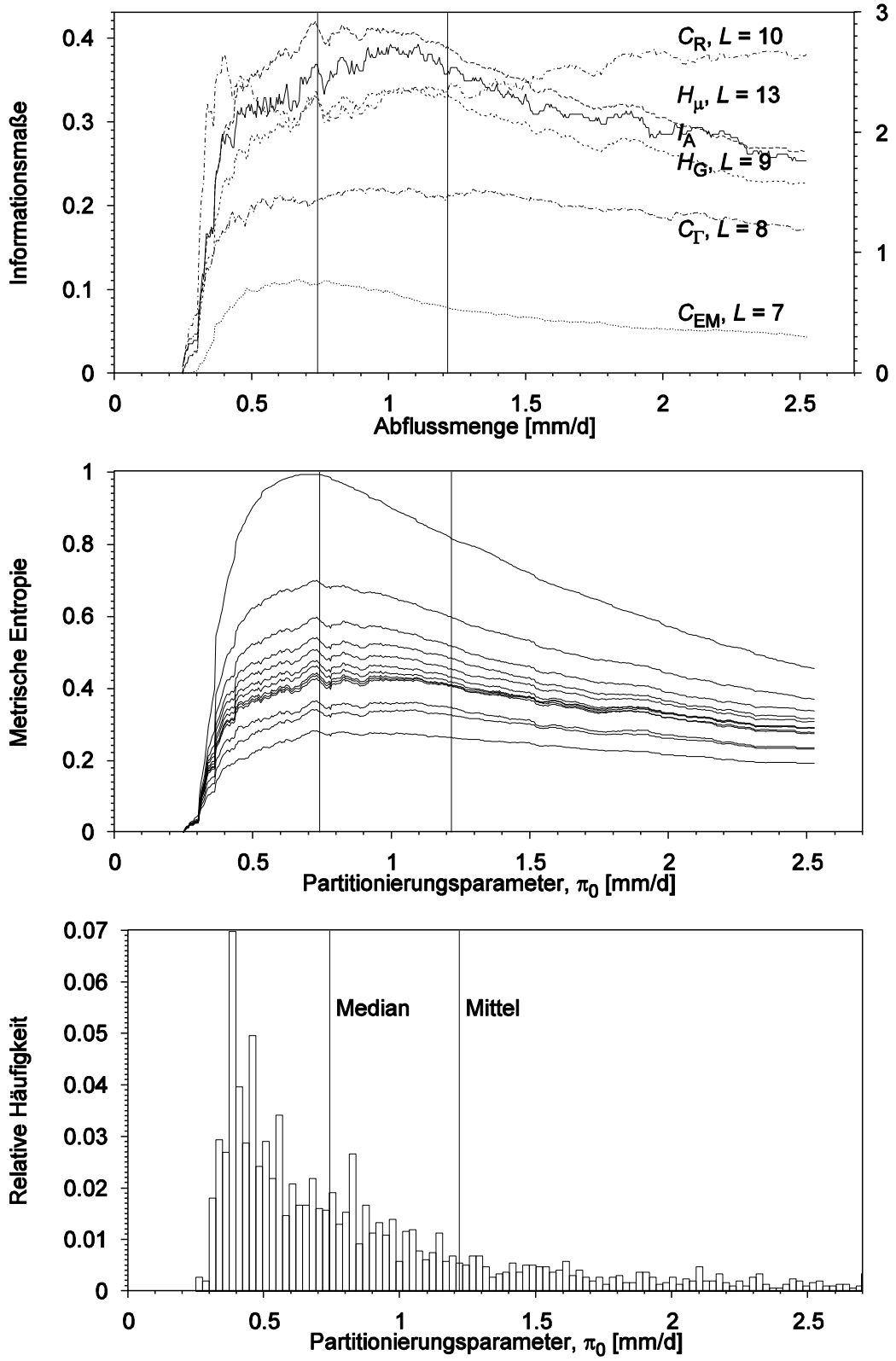

Abb. 3-6. Komplexitätsmaße in Abhängigkeit vom Partitionierungsparameter π_0 für den Abfluss des Lehstenbaches, 1987 – 1995. *Unten:* Erste 20 % des Wertebereichs der Abflusswerte, Min.: 0.2472 mm/d, Max.: 12.5342 mm/d. *Mitte:* Metrische Entropie für Wortlänge 1 bis 10, 15, 20 und 30 (von oben nach unten). *Oben:* Verschiedene Komplexitätsmaße bei jeweils maximaler Wortlänge für 5 % Genauigkeit nach Tabelle 7-4. Berechnet wurden jeweils 500 äquidistante Werte für π_0.

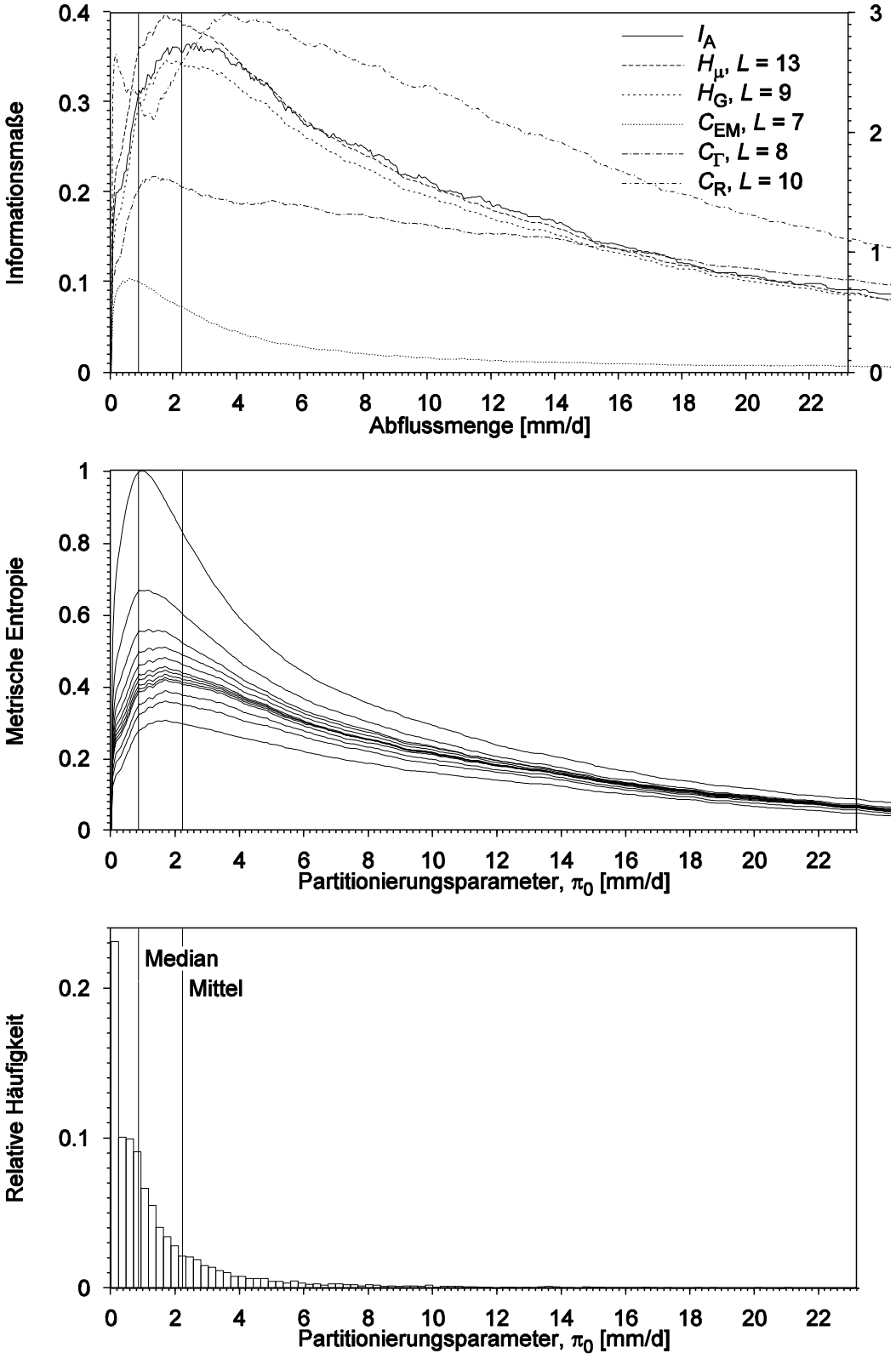

Abb. 3-7. Komplexitätsmaße in Abhängigkeit vom Partitionierungsparameter π_0 für Hubbard Brook, Watershed 1, Abfluss 1956 – 1993. *Unten:* Erste 20 % des Wertebereichs der Abflusswerte, Min.: 0 mm/d, Max.: 116 mm/d. *Mitte:* Metrische Entropie für Wortlänge 1 bis 10, 15, 20 und 30 (von oben nach unten). *Oben:* Verschiedene Komplexitätsmaße bei jeweils maximaler Wortlänge für 5 % Genauigkeit nach Tabelle 7-4. Berechnet wurden jeweils 500 äquidistante Werte für π_0.

Die Metrische Entropie H_μ nimmt auf der Symbolebene (Wortlänge 1) erwartungsgemäß beim Median ein Maximum an. Dieses verschiebt sich mit zunehmender Wortlänge nach rechts in den Bereich, in dem auch andere (Informations-) Maße maximal werden. Für den Informationsgewinn H_G wurde im Unterschied dazu eine etwa deckungsgleiche Übereinstimmung der Kurvenverläufe bei unterschiedlicher Wortlänge im zulässigen Wortlängenbereich beobachtet. Hier wird also bestätigt, dass H_G ein stabileres Maß als H_μ ist (vgl. 2.5.4). Darüber hinaus wird festgestellt, dass H_G bereits bei kleinen Wortlängen eine zuverlässige Schätzung des Entropie-Maximums liefert. In dem Bereich um Median und Mittelwert nehmen alle Komplexitätsmaße ein (individuelles) Maximum an.

Das Entropie-Maximum wird grundsätzlich nicht beim Median oder bei den Grenzen gleicher Quantile bei äquiquantiler Partitionierung für höhere Alphabete angenommen. Die äquiquantile Partitionierung ist nur für den Symbolsatz aus Sicht der Metrischen Entropie optimal. Dies entspricht bei stationären Daten der Erwartung einen optimalen Symbolsatz zu erhalten, wenn die Zahl der Symbolwechsel maximal ist oder wenn alle Symbole gleichhäufig sind. Allerdings werden die Maße in der Regel nicht auf dem Symbolsatz ausgewertet, sondern auf der Verteilung der L-Wörter, $L > 1$. Durch diese Wortverteilung wird eine Verfeinerung der Partitionierung erreicht, für die eine maximale Anzahl von Symbolwechseln oder eine Gleichverteilung von Symbolen für ein Informationsmaximum nicht unbedingt optimal ist. Dies haben die Berechnungen in diesem Abschnitt gezeigt.

Da die Grenzen der Äquiquantile einfach separat bestimmt werden können (z. B. in SYMDYN) und zumindest für den Symbolsatz optimal sind, werden auch in Zukunft äquiquantile (Median) Partitionierungen untersucht. Allerdings ist die Entropie-maximale Partitionierung auf der Wort-Ebene, auf der die Maße ausgewertet werden, die Partitionierung der Wahl.

3.8.2 Höhere Alphabete

Nun soll die Abhängigkeit der Komplexitätsmaße von der Alphabetgröße festgestellt werden. Dazu soll eine statische äquiquantile Partitionierung genügen, da diese einfach zu bestimmen ist und zumindest aus Sicht der Metrischen Entropie, wie im letzten Abschnitt gesehen, optimal für den Symbolsatz ($L = 1$) ist. Hohe Alphabetgrößen lassen zudem nur minimale Wortlängen L zu. Die Bestimmung eines Entropie-Maximums bei großen Alphabeten, also vielen Partitionierungsparametern, ist sehr rechenzeitaufwendig.

Da große Alphabete nur bei einer großen Datenmenge möglich sind (siehe 3.6 oder Tabellen 7-1 bis 7-6), wurden hierfür nur die täglichen Abflusszeitreihen der Langen Bramke (17226 Tage) und von Hubbard Brook, Watershed 1 (13880 Tage), betrachtet. Bei Berechnung der Informations- und Komplexitätsmaße für unterschiedliche Alphabetgrößen ändert sich der binäre Wertebereich der Maße. Um diese noch miteinander vergleichen zu können, wurden die Wertebereiche normiert, indem die Logarithmen in den Formeln zur Basis λ, anstatt zur Basis 2 unabhängig von λ, gewählt wurden. Die Einheiten der Maße sind damit nicht mehr binär, falls $\lambda \neq 2$, sondern ternär ($\lambda = 3$), quartär ($\lambda = 4$), pentär ($\lambda = 5$) usw. Dadurch verringern sich auch die Anforderungen an die erforderliche Datenmenge für die Komplexitätsmaße C_{EM}, C_Γ und C_R, weil nach Bedingung (91) für diese Maße nur eine absolute Genauigkeit gefordert werden kann, die von den kleineren Komplexitäten in angepassten Einheiten eher erreicht wird. Dies bedingt eine höhere maximale Wortlänge. Z. B. konnten für C_Γ und die dafür bei äquiquantiler Partitionierung minimal erforderliche Wortlänge 2 für die Lange Bramke 5 % Genauigkeit bei binären Einheiten nur bis $\lambda = 8$ und bei angepassten Einheiten

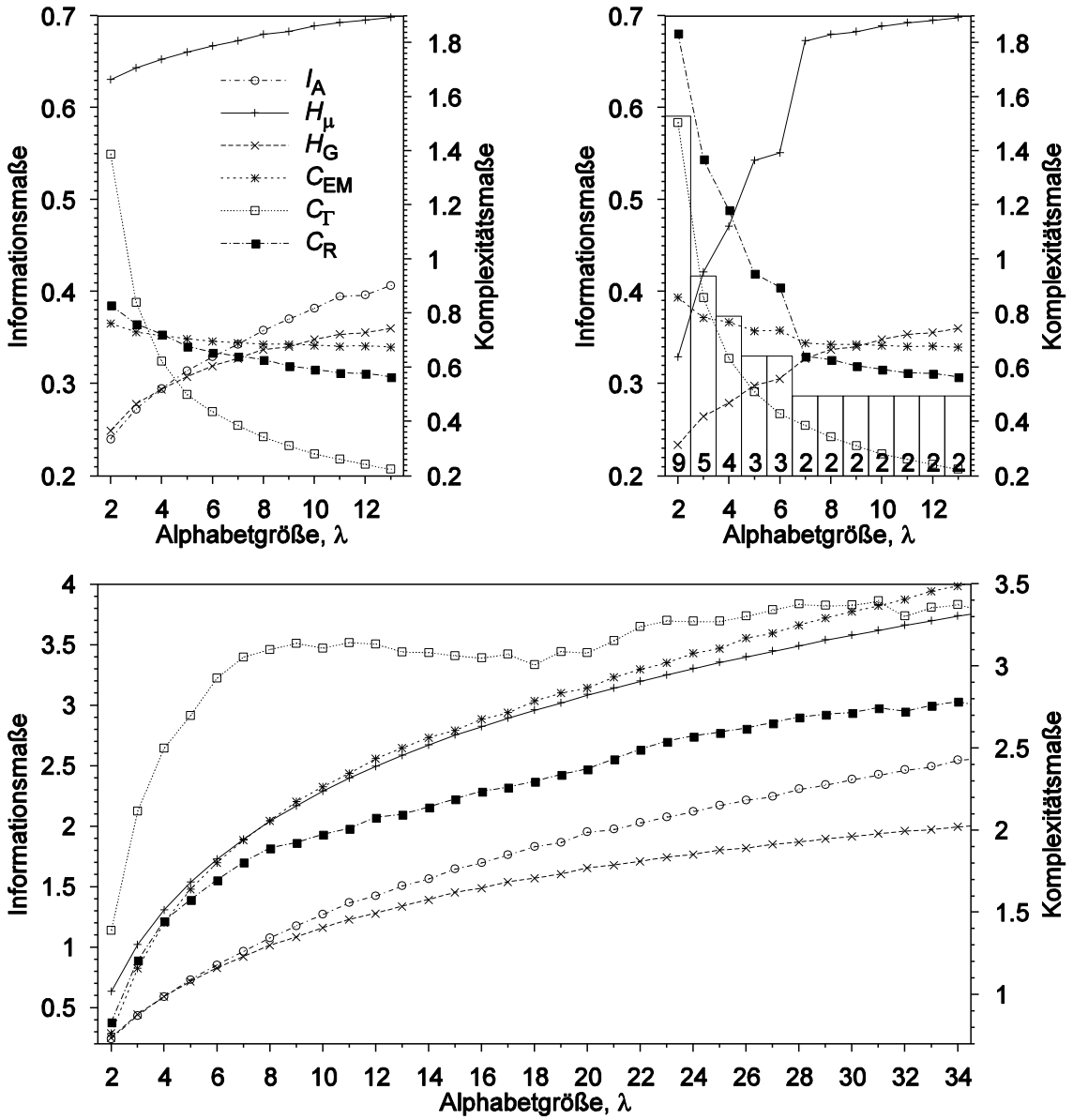

Abb. 3-8. Information und Komplexität des Abflusses der Langen Bramke 1948 – 1995 bei verschiedener Alphabetgröße λ. *Oben:* Angepasste Einheiten, d. h. Logarithmen zur Basis λ. *Links:* Konstante Wortlänge $L = 2$. *Rechts:* Maximale Wortlänge als Säulen-Diagramm mit zugehörigem L-Wert, nach Gleichung (88) und (91) für mindestens 5 % Genauigkeit bei angepassten Einheiten. *Unten:* Binäre Einheiten, Wortlänge 2.

bis λ = 13 gewährleistet werden. Für die oberen Bilder in Abb. 3-8 wurden angepasste Einheiten verwendet. Das untere Bild ist in binären Einheiten erstellt.

Die Ergebnisse der Berechnungen fielen für Hubbard Brook und die Lange Bramke sehr ähnlich aus. Sie sind deshalb nur für die Lange Bramke in Abb. 3-8 dargestellt. Die linke obere Grafik zeigt die Alphabetabhängigkeit der wichtigsten Informations- und Komplexitätsmaße bei konstanter Wortlänge 2. Dies ist die kleinste Wortlänge, bei der diese Maße bei äquiquantiler Partitionierung noch sinnvolle Werte ergeben. Alle Maße zeigen eine mehr oder weniger starke systematische Abhängigkeit von der Alphabetgröße λ: Die Informationsmaße nehmen mit λ zu, was der intuitiven Erwartung entspricht, dass die Information bei höherer Auflösung zunimmt. Die Komplexitätsmaße nehmen mit λ ab. Berechnungen mit unterschiedlicher Alphabetgröße können also nicht ohne weiteres quantitativ verglichen werden. Die Abbildung zeigt, mit welcher grundsätzlichen Änderung eines Wertes zu rechnen ist, wenn λ

erhöht oder verringert wird. Bei binären Einheiten nehmen alle Maße prinzipiell mit der Alphabetgröße zu (Abb. 3-8 unten).

Die Werte der Rényi-Komplexität C_R und Effektiven Maßkomplexität C_{EM} zeigen mit einer jeweils leichten Abnahme die geringste Abhängigkeit von λ. Die Metrische Entropie H_μ steigt leicht mit λ an. Die Algorithmische Information I_A und der Informationsgewinn H_G nehmen etwas stärker mit λ zu. Beide Maße liegen noch bis etwa $\lambda = 5$ nahe beieinander und laufen bei höherem λ auseinander. Die Nähe von I_A und H_G zur Entropie der Quelle und damit auch untereinander wurde bereits in den Abschnitten 2.5.4 und 2.5.6 besprochen. Am stärksten ist die Abnahme der Fluktuationskomplexität C_Γ mit wachsendem λ. Sie nähert sich von 1.4 bei $\lambda = 2$ nahezu hyperbolisch einem Wert von nahe 0 für $\lambda > 13$. Die Ursache dafür wird in der Darstellung mit binären Einheiten (Abb. 3-8 unten) sichtbar: Die Fluktuationskomplexität erreicht ab $\lambda = 7$ einen konstanten Wert von etwa 3.1. Die Interpretation dieses Verhaltens ist schwierig. Berechnungen bis zu einer Alphabetgröße von $\lambda = 230$ zeigen, dass auch andere Maße in binären Einheiten ein Plateau erreichen. Die Konsequenzen einer möglicherweise ungesättigten Statistik bei dieser gigantischen Alphabetgröße konnten nicht beurteilt (erkannt) werden. Eine Vergleichsrechnung mit Zufallszahlen bestätigte für diesen Fall jedoch die theoretische Grenze von $\lambda = 13$ bei gleicher Datenmenge. Das frühe Erreichen eines konstanten Wertes könnte an der Definition der Fluktuationskomplexität als Varianz (quadratischer Wert) liegen oder an einer tatsächlichen Sättigung der lokalen Informationsschwankungen. Die drastische Abnahme der Informationsfluktuationen mit zunehmendem λ in angepassten Einheiten könnte bedeuten, dass solche Fluktuationen bei binärem Alphabet am sichersten erfasst werden. Dies würde die von CRUTCHFIELD & PACKARD (1983) vorgeschlagene Verwendung binärer Alphabete (und Verfeinerung der Partitionierung durch höhere Wortlängen) 10 Jahre später durch das Maß von BATES & SHEPARD (1993) bestätigen. Die in diesem Abschnitt beschriebenen Beobachtungen sind typisch für den untersuchten Datentyp. Inwieweit sie universell gelten, muss weiter untersucht werden.

Was ist nun die geeignete Alphabetgröße? In Abschnitt 3.6 wurde u. a. festgestellt, wie groß die Wortlänge bei einer gegebenen Anzahl von Datenpunkten gewählt werden kann, um eine bestimmte mittlere Genauigkeit der Werte der Komplexitätsmaße zu gewährleisten. Bei kleinen Alphabetgrößen sind demzufolge größere Wortlängen möglich als bei großen Alphabeten. Bei höherer Wortlänge wird mehr Struktur von der Zeitreihe durch die Berücksichtigung von lokalen zeitlichen Zusammenhängen erfasst. Bei höheren Alphabeten wird mehr Struktur durch eine höhere Auflösung der Werte erfasst. Welches ist der größere Vorteil? Abb. 3-8, oben rechts, im Vergleich zur linken Teilabbildung zeigt, dass es sowohl sehr Wortlängen-sensitive Maße gibt wie auch solche, die kaum qualitativ davon abhängen. Zu den stark Wortlängen-sensitiven Maßen zählt insbesondere die Metrische Entropie H_μ und die Rényi-Komplexität C_R. Die Fluktuationskomplexität C_Γ und der Informationsgewinn H_G reagieren kaum sichtbar auf die sich ändernde Wortlänge. Die Werte der Komplexitäten, C_Γ und C_R, sowie die der Informationen, H_μ und H_G, liegen bei größerer Wortlänge und kleinerer Alphabetgröße jeweils deutlich näher beieinander als bei großem Alphabet und kleiner Wortlänge. Dies spricht für eine Berechnung der Maße bei binärem Alphabet mit maximaler Wortlänge, was im Einklang mit dem Hinweis aus dem letzten Abschnitt steht, dass bei C_Γ Informationsfluktuationen möglicherweise besser bei binärem Alphabet erfasst werden. Nach diesen Kriterien ist also die Berücksichtigung der lokalen zeitlichen Zusammenhänge von größerem Vorteil als eine höhere Auflösung der Messwerte. In dieser Arbeit soll daher binären Alphabeten mit maximaler Wortlänge gemäß 3.6 der Vorzug gegeben werden. Sie schließt sich somit der Vorgehensweise von CRUTCHFIELD & PACKARD (1983) an. Höhere Alphabete bleiben jedoch interessant zur Gewinnung konsistenter Ergebnisse, z. B. bei der

Bestimmung verrauschter Maxima (siehe 5.3.1) oder zur Feststellung qualitativer Unterschiede, z. B. mit der Transinformation (siehe 5.2.1).

Diese Betrachtungen sind unabhängig von den Situationen zu verstehen, bei denen sich höhere Alphabete automatisch oder aus Plausibilitätsgründen anbieten. Damit sind beispielsweise Biosequenzen, Texte, Musikstücke und Zeitreihen gemeint, bei denen sich bestimmte Schwellenwerte aufgrund ihrer Bedeutung als Partitionierungsgrenzen anbieten.

3.8.3 Tolerante Partitionierungen

Das Rauschen in den Messdaten führt bei scharfen Partitionierungsgrenzen zu Symbolwechseln, die nicht der tendenziellen Dynamik entsprechen. Dies gilt insbesondere für dynamische Partitionierungen, bei denen vor allem das Messrauschen sichtbar wird (siehe z. B. 5.3.1.1.3). Daher liegt es nahe eine tolerante Partitionierung zuzulassen, bei der ein Symbolwechsel erst registriert wird, wenn eine Partitionierungsgrenze um einen bestimmten Toleranzwert überschritten wird. Solche Partitionierungen wurden bereits von ROMAHN (1996) untersucht (siehe auch LANGE et al., 1997) und sind in SYMDYN mit relativen oder absoluten Toleranzbreiten möglich. In der Arbeit von ROMAHN (1996), in der die Untersuchungen ausführlich dargestellt sind, wie auch in dieser Arbeit konnte lediglich eine systematische Informationsabnahme mit zunehmender Toleranzbreite festgestellt werden. Es gab kein Indiz für eine ausgezeichnete Toleranzbreite. Daher muss von einer unscharfen Partitionierung abgesehen werden. Zur Rauschreduzierung gibt es etablierte Filter (siehe z. B. in PRESS et al., 1992). Von der Anwendung solcher Methoden wird jedoch mit dem gleichen Argument der Informationserhaltung abgesehen, das auch gegen eine Saisonbereinigung spricht (siehe 3.3).

4 Daten

In diesem Kapitel werden die in dieser Arbeit untersuchten Daten zum Wasserhaushalt von (bewaldeten) Wassereinzugsgebieten vorgestellt. Dazu wird zunächst das jeweilige Gebiet beschrieben, soweit dies von hydrologischem Interesse und für einen Vergleich der Gebiete anhand der hydrologischen Dynamik relevant ist. Es werden die geografische Lage, der geologische Untergrund und Boden, der Bewuchs (dominierende Baumarten) und das Klima beschrieben. Mit Niederschlag ist stets der Freilandniederschlag des Einzugsgebietes gemeint. Dieser wird mit mindestens einem Niederschlagssammler im Gebiet gemessen. Der oberirdische Abfluss wird an einem Abflusswehr in dem im Gebiet entstehenden Bach gemessen.

4.1 Lehstenbach (Fichtelgebirge)

Das Wassereinzugsgebiet des Lehstenbaches liegt am Großen Waldstein im Fichtelgebirge (50° 9‘ nördliche Breite, 12° 52‘ östliche Länge). Es ist das erste der beiden Hauptuntersuchungsgebiete des BITÖK und wird bei MANDERSCHEID & GÖTTLEIN (1995) beschrieben. Das 419 ha große Gebiet liegt zwischen 690 und 880 m über NN. Es öffnet sich nach Südost (HEINDL et al., 1995). Nahe der Wasserscheide liegen die Intensiv-Messflächen Coulissenhieb

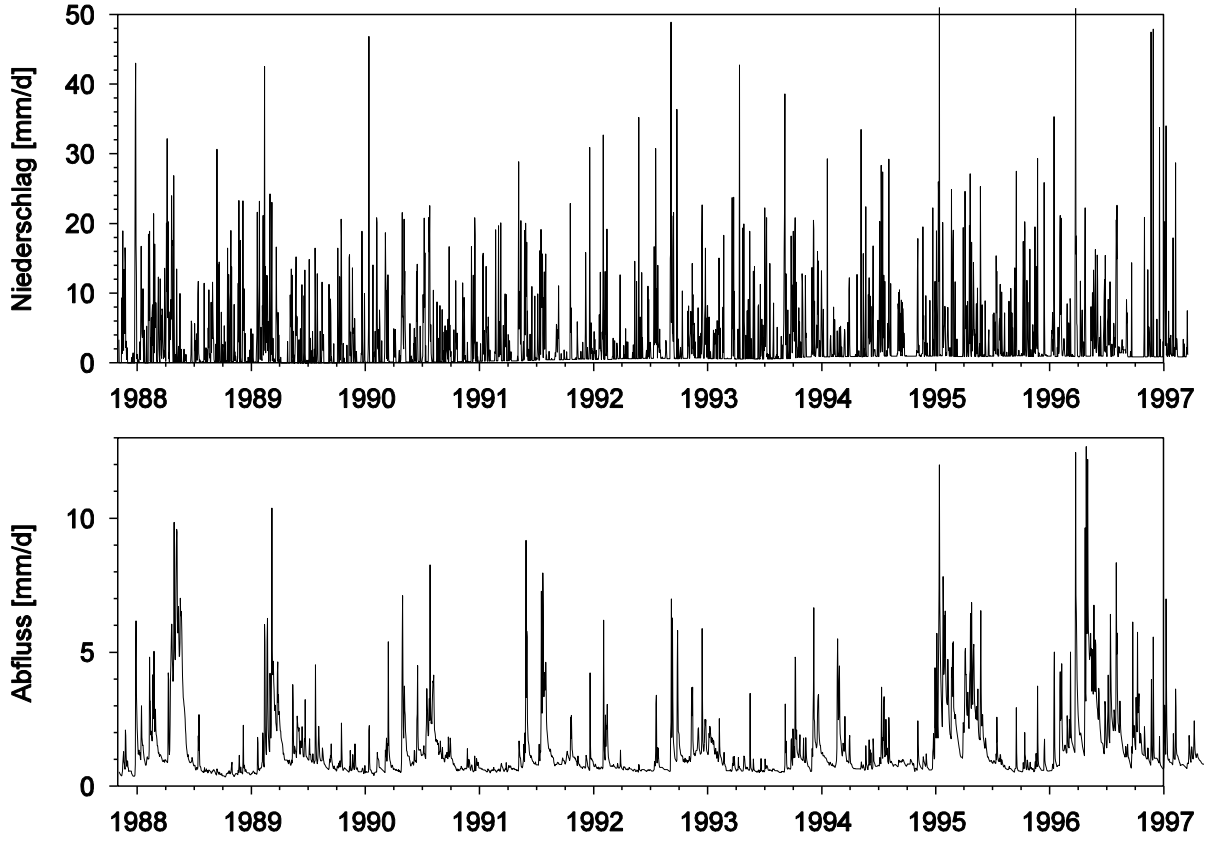

Abb. 4-1. Tägliche Mengen von Niederschlag und Abfluss im Einzugsgebiet des Lehstenbaches vom 2.11.1987 bis 31.10.1996. Niederschlag nur bis 31.10.1995.

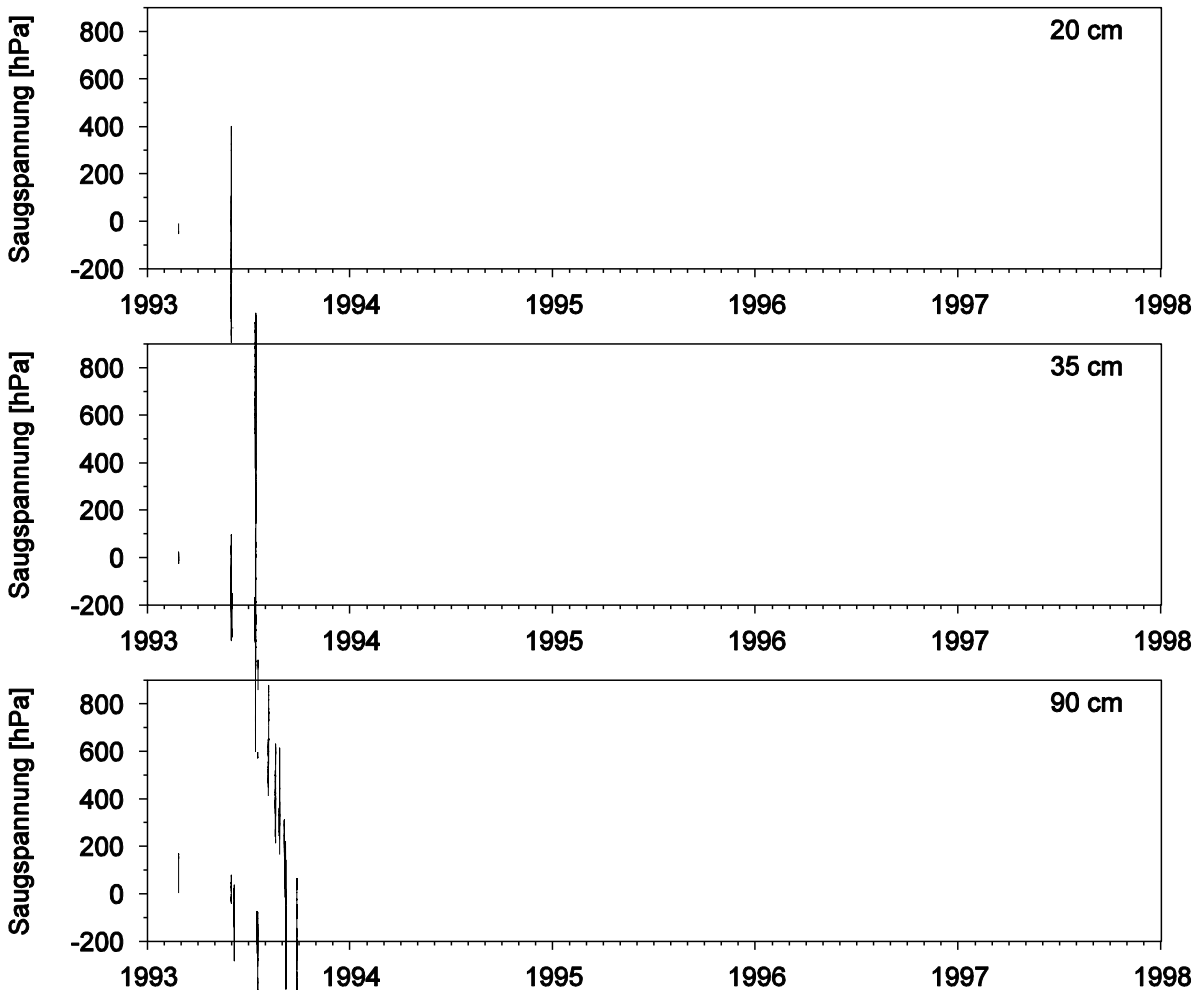

Abb. 4-2. Saugspannungen in 20 cm, 35 cm und 90 cm Tiefe am Tensiometer-Standort 12 auf der Fläche „Coulissenhieb“ von 1993 bis 1997.

(außerhalb) und Weidenbrunnen (innerhalb). Eine meteorologische Messstation des BITÖK befindet sich seit Januar 1994 im benachbarten Pflanzgarten des Forstamtes Weissenstadt.

Der geologische Untergrund des Lehstenbach-Einzugsgebietes besteht größtenteils aus dem etwa 291 Mio. Jahre altem Randgranit, der während der variskischen Gebirgsbildung fast senkrecht aufgestellt wurde. Daneben befindet sich der etwa 288 Mio. Jahre alte grobkörnige Kerngranit, der auch im Einzugsgebiet im Dachbereich zu finden ist. Die Granite sind von Störungen (Klüften) durchzogen, die z. T. Quarzgänge enthalten. Die Verwitterungszone der Granite läßt sich z. T. bis zu 40 m tief nachweisen. Dieser Bereich enthält sowohl mehrere Meter große Findlinge, wie auch Granitgrus und feinkörnigen Granitzersatz. Im Pleistozän entstanden die dichten bis zu 3 m mächtigen Solifluktionsböden. Etwa ein Drittel der Fläche des Einzugsgebietes ist heute vermoort. Im unverwitterten Granit zirkuliert das Wasser fast ausschließlich in den Klüften, während der vergruste Granit als Wasserspeicher dient. Die sauren Böden, sowie saure Niederschläge verursachen im Bachwasser einen pH-Wert von 5.0 – 6.5 bei niedrigem Abfluss und von weniger als 4 bei hohen Abflüssen. Diese und weitere Informationen zur Gewässerchemie finden sich bei BITTERSOHL & LISCHEID (1995).

Das Einzugsgebiet des Lehstenbaches ist zu etwa 90 % mit Fichte (*Picea abies* (L.) KARST.) bewachsen. Der Anteil der Laubbäume ist sehr gering. Das mittlere Bestandesalter liegt bei 40 bis 50 Jahren und ist auf einer Fläche von 21 % vertreten. Der Anteil an Beständen über 130 Jahre beträgt etwa 6 % (HEINDL et al., 1995).

Das Klima im Fichtelgebirge ist humid kontinental mit kurzen kühlen Sommern und langen kalten Wintern. Im Sommer wird das Gebiet durch milde atlantische Luftmassen mit Westwinden beeinflusst. Im Winter herrschen häufig kontinentale Bedingungen mit Ostwinden vor. Die Temperatur beträgt 5 – 6.5 °C im Jahresmittel. Von April 1992 bis September 1994 wurde am Coulissenhieb die höchste Temperatur im Juli mit 15.5 °C erreicht und die niedrigste Temperatur im Dezember mit –1.6 °C (monatliche Mittel). Es gibt etwa 100 bis 200 Nebeltage im Jahr (PETERS & GERCHAU, 1995). Von den 950 – 1050 mm Jahresniederschlag gelangen nur 550 – 650 mm bis zum Abfluss (BITTERSOHL & LISCHEID, 1995).

Messungen zur Hydrologie des Lehstenbach-Einzugsgebietes werden vom Bayerischen Landesamt für Wasserwirtschaft (LfW) seit 1987 durchgeführt. Die Abteilung Hydrogeologie des BITÖK führt erst seit ihrer Einrichtung 1993 eigene Messungen in dem Gebiet durch (BITTERSOHL & LISCHEID, 1995). Niederschläge werden in hoher Auflösung erst ab 1994 im Pflanzgarten von der Abteilung Klimatologie des BITÖK gemessen. Aus der zentralen Datenbank des BITÖK und vom LfW (Abfluss) wurden die folgenden Daten verwendet:

- Niederschlag in täglicher Auflösung, vom 02.11.1987 bis 31.10.1995 (siehe Abb. 4-1),
- Niederschlag von der Fläche „Coulissenhieb“ in stündlicher Auflösung, vom 04.04.1992 bis 30.08.1994,
- Niederschlag von der Station „Pflanzgarten“ in 10-minütlicher Auflösung, vom 28.01.1994 bis 07.12.1998,
- Gebietsabfluss in täglicher Auflösung vom 02.11.1987 bis 31.10.1996 (siehe Abb. 4-1),
- Abfluss in stündlicher Auflösung vom 02.11.1993 bis 31.10.1996

Auf der Fläche „Coulissenhieb“ wird von der Abteilung Bodenkunde die Saugspannung des

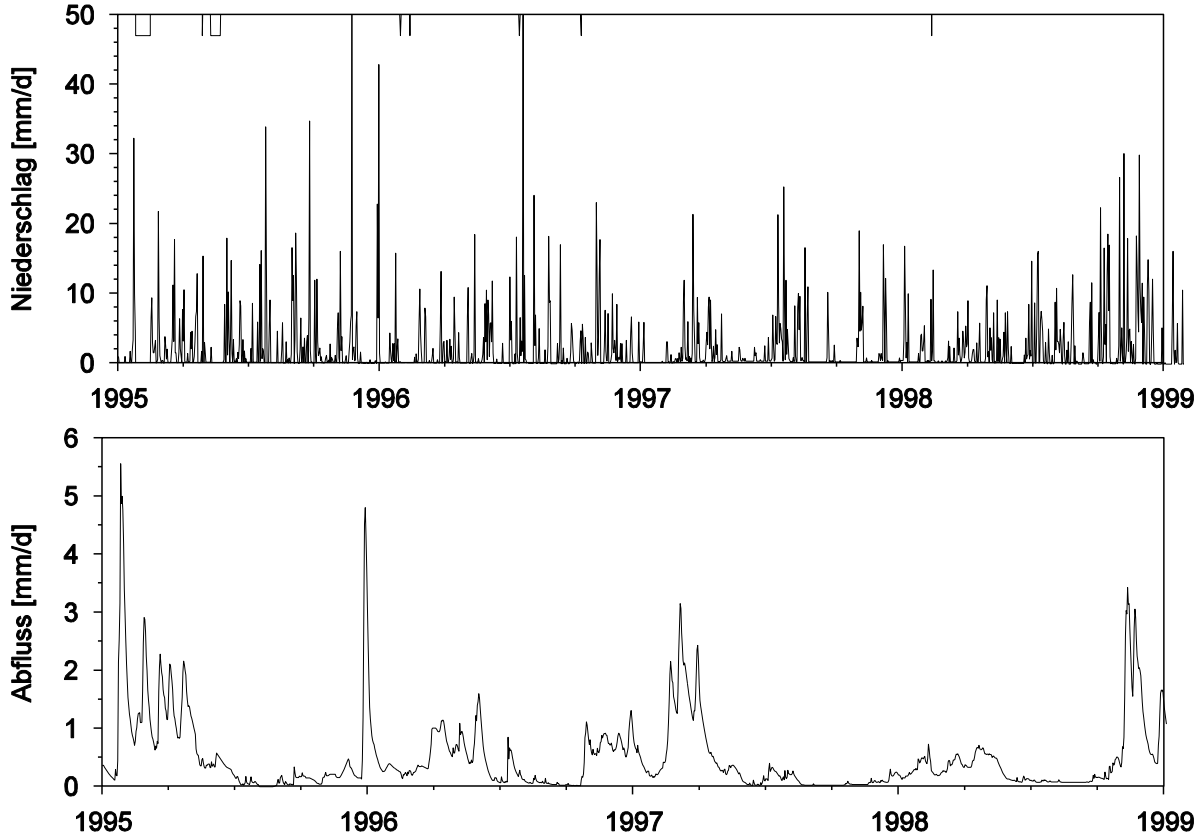

Abb. 4-3. Tägliche Mengen von Niederschlag und Abfluss im Einzugsgebiet „Steinkreuz“ vom 01.01.1995 bis 29.12.1998. Die Tageswerte wurden durch Aggregation der im Text beschriebenen Daten gewonnen. Ausfälle bei der Messung der Niederschlagsmenge sind am oberen Rand markiert.

Bodens in einem Feld mit 20 Tensiometern in drei Tiefen (20, 35 und 90 cm) gemessen. Davon wurden die Aufzeichnungen von fünf Tensiometern (jeweils alle drei Tiefen) in stündlicher Auflösung vom 26.02.1993 bis zum 27.06.1997 verwendet (Manderscheid, persönliche Mitteilung). Ausfälle der Tensiometer oder Daten-Logger haben zu Lücken in den Daten geführt. Abb. 4-2 zeigt die Saugspannungen der Tensiometer am Messort Nr. 12.

4.2 Steinkreuz (Steigerwald)

Das Wassereinzugsgebiet „Steinkreuz“ liegt im Steigerwald (49° 52‘ nördliche Breite, 10° 27‘ östliche Länge). Es ist die zweite Hauptuntersuchungsfläche des BITÖK und wird bei LISCHEID & GERSTBERGER (1997) beschrieben. Das 55 ha große Teileinzugsgebiet der Regnitz liegt zwischen 400 und 460 m über NN. Es öffnet sich von Südwesten.

Der geologische Untergrund besteht aus Sedimenten des Mittleren Keupers (Oberer Trias) und wird aus wechselnden tonigen Schichten und Sandstein aufgebaut. Aufgrund von nur wenigen tektonischen Störungszonen sowie einer 30 m mächtigen Tonschicht der Lehrbergstufe kann das Gebiet als dicht gelten. Dem Ausgangsmaterial entsprechend dominieren sandige Braunerden. Die nur schwach sauren Böden sind tiefgründig mit einer Durchwurzelungstiefe von

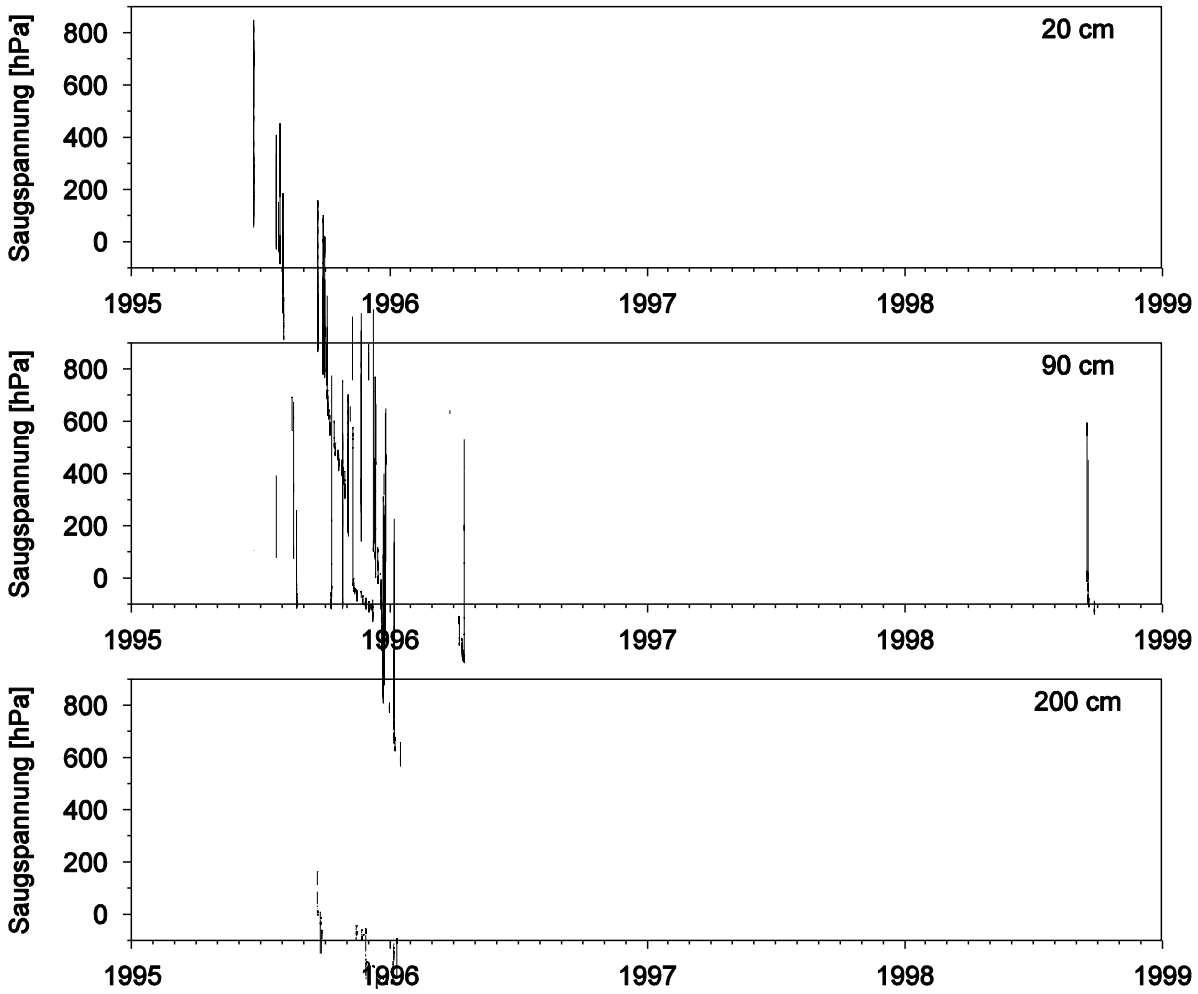

Abb. 4-4. Saugspannungen in 20 cm, 90 cm und 200 cm Tiefe am Tensiometer-Standort 01 im Einzugsgebiet Steinkreuz von 1995 bis 1998.

geschätzten mindestens 2 m (Lischeid, persönlich). Das Gebiet ist vornehmlich von Buche (*Fagus sylvatica* L.) und Traubeneiche (*Quercus petraea* (MATT.) LIEBL.) sowie anderen Laub-Baumarten bewachsen.

Der Steigerwald liegt in der Übergangszone zwischen subatlantischem (im Westen) und subkontinentalem Klima (im Osten). Die bevorzugte Windrichtung ist West bis Süd-West. Die mittlere Temperatur beträgt 7.5 °C im Jahr, –1.5 °C im Januar und 16.5 °C im Juli. Im Jahr fällt im Mittel 650 – 800 mm Niederschlag. Hohe Niederschlagsraten gibt es von Juni bis August und (weniger stark) von Dezember bis Januar.

LISCHEID & GERSTBERGER (1997) stellen eine Vielzahl von Parametern vor, welche von der Abteilung Hydrogeologie des BITÖK im Steinkreuz-Einzugsgebiet seit August 1994 gemessen werden. Davon wurden die folgenden Daten verwendet (Lischeid, persönliche Mitteilung):

- Niederschlag in 10-minütlicher Auflösung vom 01.01.1995 – 29.12.1998 (siehe Abb. 4-3),
- Gebietsabfluss, stündlich, vom 01.01.1995 – 25.12.1998 (siehe Abb. 4-3) und
- Saugspannungen im Boden in 20 cm, 90 cm und 200 cm Tiefe von jeweils fünf Standorten in der 1.29 ha großen Intensiv-Messfläche im Zentrum des Steinkreuz-Einzugsgebietes. Die weitgehend äquidistanten Aufzeichnungen wurden auf einen Zeitabstand von einer Stunde, der gröbsten Probenahmerate, ausgelesen. Die Daten enthalten teilweise viele Lücken von wenigen Stunden, Tagen oder sogar Monaten, die mehr oder weniger gleichmäßig über den Messzeitraum von März, Juni oder September 1995 bis Ende 1998 verteilt sind. Abb. 4-4 zeigt die Saugspannungen von Tensiometergruppe 01, die einen der vollständigsten Datensätze liefert.

4.3 Lange Bramke (Harz)

Das Wassereinzugsgebiet der Langen Bramke liegt im Oberharz (51° 52‘ nördliche Breite, 10° 26‘ östliche Länge). Es wird in der Dissertation von SCHMIDT (1997) beschrieben, der auch die Informationen über das Gebiet entnommen wurden. Das 76 ha große Gebiet liegt zwischen 543 und 700 m über NN. Es öffnet sich nach Osten mit einem mittleren Gefälle des Haupttales von 11 %.

Das Einzugsgebiet liegt im Oberharzer Devonsattel (Unterdevon). Die anstehenden quarzitischen Sandsteine sind teilweise von Kalksandsteinen und Tonschiefern durchsetzt. Darüber befindet sich eine Hangschuttdecke von 2 – 5 m Mächtigkeit. Der Boden wird aus einer überwiegend schluffig-lehmigen Fließerde gebildet. Es herrschen Podsole und Braunerden mit einer Humusauflage von 9 bis 12 cm vor.

Das Gebiet wurde 1947 vollständig kahlgeschlagen. 34 ha wurden ab 1949 mit 3- bis 4-jährigen Fichten (*Picea abies* (L.) KARST.) wieder aufgeforstet. Es folgten weitere Pflanzungen und Saaten, so dass 1970 wieder die volle Deckung der Fläche erreicht war. Daran schlossen sich diverse Durchforstungsmaßnahmen bis heute an. Details sind bei SCHMIDT (1997, S. 18f) genannt. 1996 wies der Fichtenbestand ein Alter von 38 bis 46 Jahren auf und bestockte die Untersuchungsfläche homogen zu 90 % mit Ausnahme der Schneisen.

Der Harz liegt im Übergangsbereich zwischen ozeanischem und kontinentalem Klima. Die mittlere Jahrestemperatur liegt bei 6 °C. Aufgrund der vorherrschenden west- bis südwestlichen Windrichtung liegt das Einzugsgebiet im Lee der Schalke (762 m über NN). Ein subat-

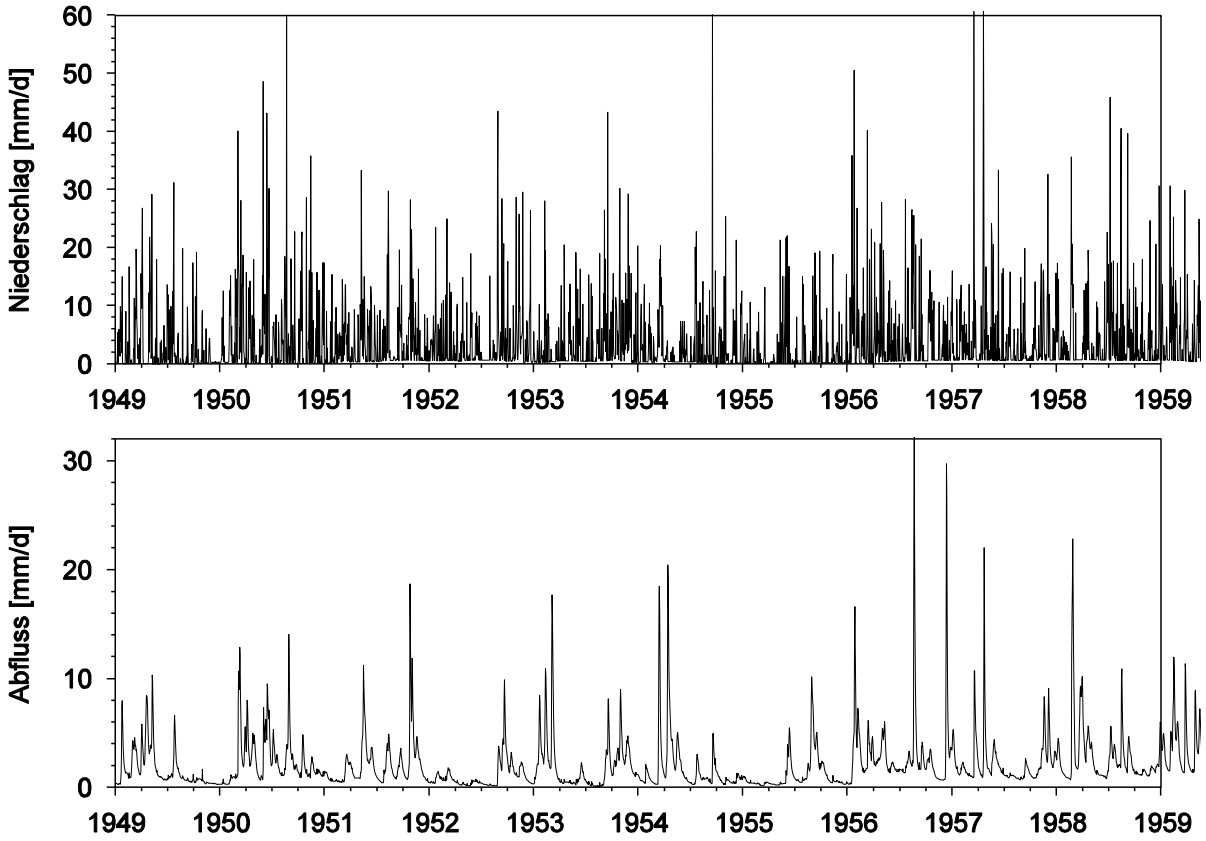

Abb. 4-5. Tägliche Mengen von Niederschlag und Abfluss im Einzugsgebiet der Langen Bramke vom 01.01.1949 bis 31.12.1958.

lantisches Klima verursacht eine über das Jahr gleichmäßige Niederschlagsverteilung. Die Staulage bewirkt ganzjährig hohe Niederschläge mit einem Jahresmittel von 1230 mm.

Die Messungen von täglicher Niederschlags- und Abflussmenge im Lange-Bramke-Gebiet begannen bereits im November 1948, da nach der Abholzung Auswirkungen auf Hochwasserspitzen und Erosion befürchtet wurden, die untersucht werden sollten. Seit 1981 obliegt das Einzugsgebiet, sowie die benachbarten Untersuchungsgebiete „Steile Bramke“ und „Dicke Bramke“, der Arbeitsgemeinschaft Oberharzer Untersuchungsgebiete, an der neben dem Institut für Bodenkunde und Waldernährung (Universität Göttingen) die Bundesanstalt für Gewässerkunde (Koblenz), die Harzwasserwerke des Landes Niedersachsen sowie das Institut für Geographie, Abteilung für Physische Geographie und Hydrologie der TU Braunschweig (IfG), beteiligt sind. Vom Einzugsgebiet der Langen Bramke wurden die folgenden Daten verwendet (Schmidt, persönliche Mitteilung):

- Niederschlag in täglicher Auflösung vom 01.11.1948 – 31.12.1988,
- Abfluss in täglicher Auflösung vom 01.11.1948 – 31.12.1995,
- Niederschlag in stündlicher Auflösung vom 1983 – 1992 und
- Abfluss in stündlicher Auflösung vom 1986 – 1995

Die über 47-jährige Abflussmessung der Langen Bramke ist die längste Abflusszeitreihe in dieser Arbeit. Sie wird in Abb. 4-5 zur Zeit der Wiederaufforstung mit dem Niederschlag dargestellt.

4.4 Birkenes (Norwegen)

Das Wassereinzugsgebiet „Birkenes“ liegt 35 km nordöstlich von Kristiansand im südlichsten Norwegen (58° 23‘ nördliche Breite, 8° 15‘ östliche Länge, aus: LÜKEWILLE et al., 1998, Tabelle B.1). Das Untersuchungsgebiet des Norwegischen Institutes für Wasserforschung (NIVA) und des Norwegischen Institutes für Luftforschung (NILU) wird unter anderem bei MULDER et al. (1990), MULDER et al. (1991) und MÜLLER et al. (1993) beschrieben. Das 41 ha große Gebiet liegt zwischen 190 und 270 m über NN und öffnet sich nach Nordosten.

Auf dem granitischen Grundgestein haben sich flache, zumeist weniger als 1 m mächtige Mineralböden (Podsole bis saure Braunerden) aus eiszeitlichem Geschiebelehm entwickelt. Die Bodentiefe nimmt mit zunehmender Höhe ab. Entlang des Hauptbaches und in Vertiefungen an den Hängen haben sich Torfschichten von wenigen cm bis wenigen m Mächtigkeit gebildet. Die tiefgründigeren Bodenpartien sind überwiegend von 60 – 80 Jahre alter Fichte (*Picea abies* (L.) KARST.) mit einem Unterwuchs von Blaubeeren, Farnen und Moosen bewachsen.

Das Gebiet liegt im Übergangsbereich zwischen gemäßigtem kontinentalen und gemäßigtem ozeanischen Klima (NEWIG, 1998, S. 40), ist aber aufgrund der Küstennähe überwiegend ozeanisch beeinflusst. Im Jahr fallen im Mittel 1400 mm Niederschlag. Davon finden sich 1100 mm im Abfluss wieder.

In Birkenes werden seit 1972 tägliche Niederschlags- und Abflussmengen gemessen. Diese Daten stellte Dr. Rolf Vogt vom NIVA dem BITÖK zur Verfügung. Die Niederschlagsmengen liegen vom 01.07.1972 bis 31.07.1993 vor; die Abflussmengen vom 12.07.1972 bis 30.12.1992. Letztere fehlen für die zwei Jahre vom 01.01.1987 bis 31.12.1988. Abb. 4-6 zeigt

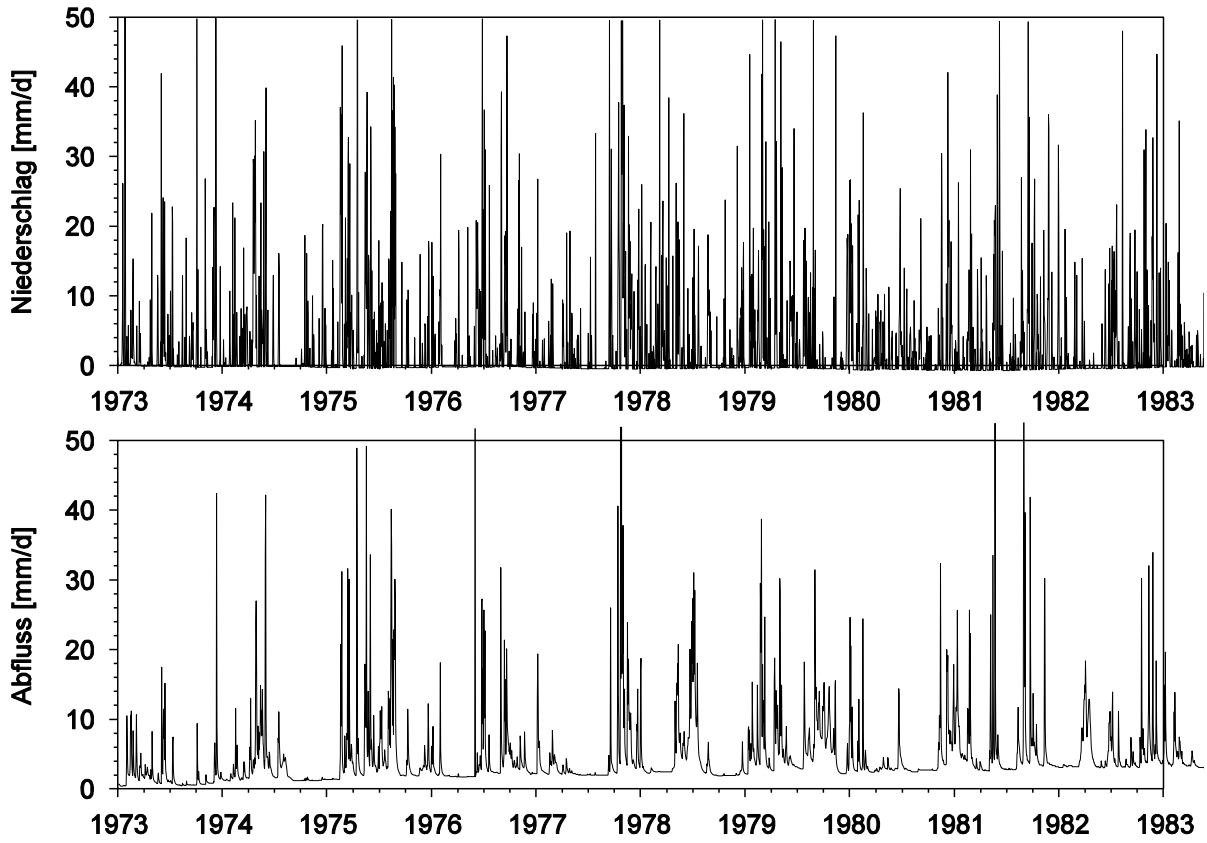

Abb. 4-6. Tägliche Mengen von Niederschlag und Abfluss in Birkenes vom 01.01.1973 bis 31.12.1982.

exemplarisch die Daten der ersten 10 Jahre.

4.5 Hubbard Brook Experimental Forest (USA)

Das Untersuchungsgebiet „Hubbard Brook Experimental Forest“ liegt im White Mountain National Forest in New Hampshire (43° 56‘ nördliche Breite, 71° 45‘ westliche Länge) im Nordosten der USA. Es wird von BORMANN & LIKENS (1979), LIKENS & BORMANN (1995) und in der Broschüre USDA (1996) beschrieben. Das hügelige 3037 ha große Einzugsgebiet des Hubbard Brook liegt zwischen 222 und 1015 m über NN. Es ist in mehrere Teileinzugsgebiete (Watersheds, W) gegliedert, von denen acht seit mehr als 20 Jahren hydrologisch untersucht werden. Das längliche Gebiet erstreckt sich von West nach Ost, was auch der Ausrichtung des Haupttales und der Fließrichtung des Hubbard Brook entspricht. Die Watersheds 1 – 6 liegen im Nordosten des Gebietes (Südhanglage); die Watersheds 7 und 8 liegen im Süden (Nordhanglage). Es treten Hangneigungen von 70 % und mehr auf; die mittleren Steigungen sind aber gering (21 % bei W 3). Die acht Teilgebiete bieten aufgrund unterschiedlicher Behandlungsmethoden eine einzigartige Möglichkeit zur vergleichenden Untersuchung der Auswirkung forstlicher Eingriffe (Rodungsmaßnahmen) und der damit verbundenen veränderten Transpirationsleistungen der Bäume auf den Wasserhaushalt von Einzugsgebieten.

Der Untergrund des östlichen Gebiets, das die Watersheds 1 – 6 umfasst, besteht aus Quarz-Glimmerschiefern und Quarziten des Silur. Das metamorphe Gestein wurde im Devon von verschiedenen Gesteinen intrudiert (u. a. Granite). Der westliche Teil des Gebietes mit Teilen der Watersheds 7 und 8 ist von einem gefalteten Granit aus dem Devon geprägt. Das Gebiet ist von eiszeitlichen Ablagerungen aus dem Pleistozän überdeckt, die 0 – 50 m dick sind und alle Korngrößen von Ton bis zu Felsen von 10 m Durchmesser enthalten.

Tabelle 4-1. Größe, Messzeitraum und Behandlung der Watersheds (W) 1 – 8 im Hubbard Brook Experimental Forest. Die täglichen Niederschlagsdaten liegen jeweils bis zum 31.12.1994 und die Abflussdaten bis zum 31.12.1993 vor. Die Tabelle wurde erstellt nach USDA (1996, Tabelle 2) und der Datenlage in http://www.hbrook.sr.unh.edu/data/data.htm.

W	Größe [ha]	Beginn Niederschl.	Beginn Abfluss	Behandlung
1	11.8	01.01.1956	01.01.1956	keine
2	15.6	01.01.1957	01.10.1957	Kahlschlag ohne Räumung im Winter 1965 – 1966; Herbizid-Behandlung im Sommer 1966, 1967 und 1968; neues Wachstum seit 1969
3	42.4	01.01.1958	01.10.1957	keine
4	36.1	01.01.1960	01.07.1960	Kahlschlag bis zu 2 cm Mindest-Durchmesser in Schneisen in 3 Phasen: 1970, 1972 und 1974; je 1/3 der Fläche. Geschlagenes Holz wurde entfernt.
5	21.9	01.01.1964	01.01.1962	Kahlschlag bis 5 cm Durchmesser 1983 – 1984. Geschlagenes Holz wurde entfernt.
6	13.2	01.01.1964	01.01.1963	keine
7	76.4	01.01.1965	01.01.1965	keine
8	59.4	01.01.1969	01.05.1968	keine

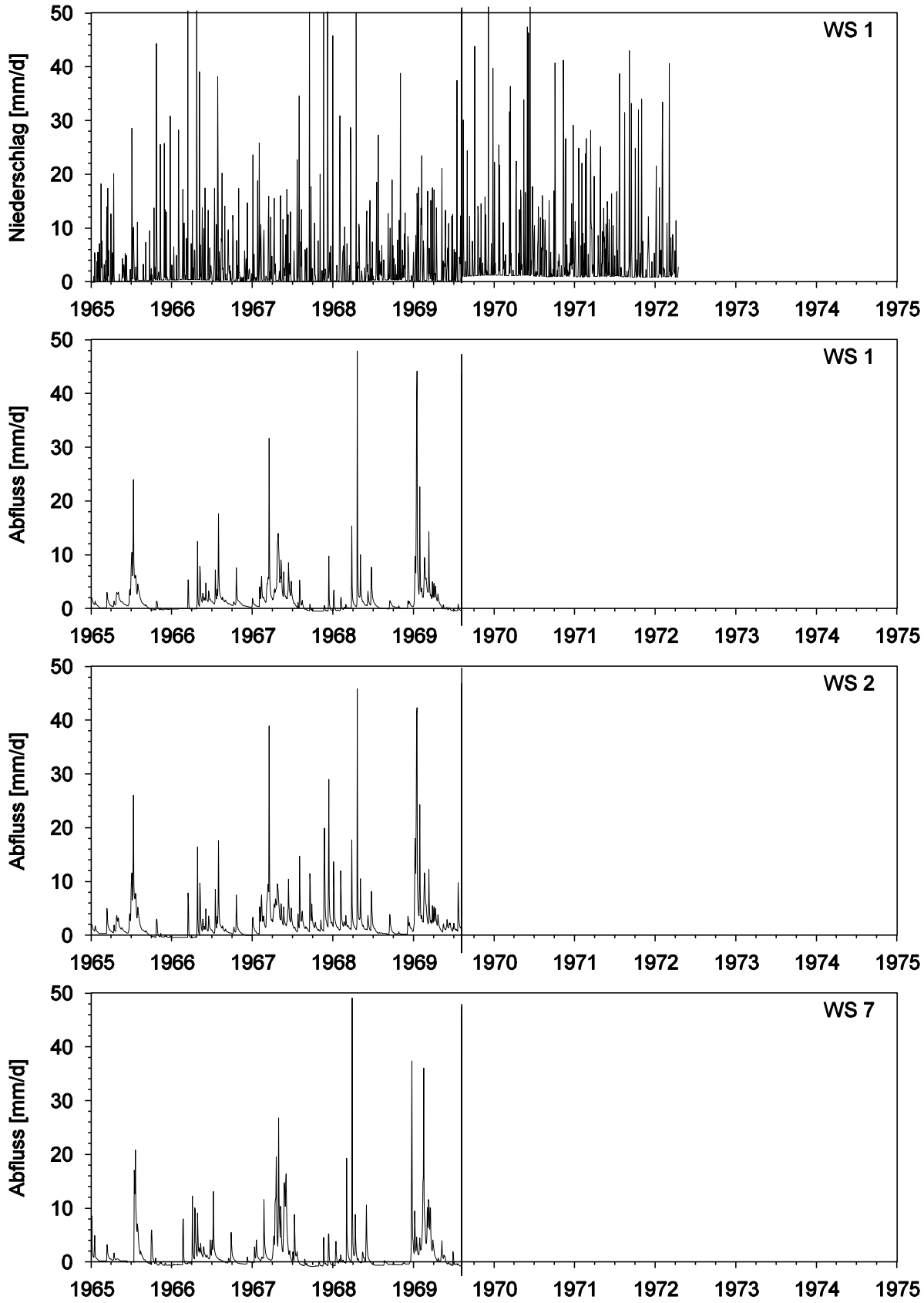

Abb. 4-7. Tägliche Mengen von Niederschlag und Abfluss in den Watersheds (W) 1, 2 (Kahlschlag) und 7 (Nordhang) von Hubbard Brook vom 01.01.1965 bis 31.12.1974. Da sich der Niederschlag kaum zwischen den Watersheds unterscheidet, wurde W 1 als Referenz gewählt.

Bei den sauren (pH-Wert 4.5 oder weniger) und relativ unfruchtbaren Böden handelt es sich hauptsächlich um gut durchlässige Podsole. Der 7 cm tiefen Streuauflage folgt ein etwa 50 cm tiefer Mineralboden, dem sich der eiszeitliche Gefügelehm anschließt, so dass nach etwa 2 m das anstehende Gestein erreicht wird. Damit gilt das Gebiet als dicht, so dass das Niederschlagswasser nur über Verdunstung (Evapotranspiration) und den Abfluss des Hubbard Brook entweichen kann.

Das Gebiet war vor den Durchforstungsmaßnahmen (siehe Tabelle 4-1) vollständig bewaldet. Dabei dominieren zu 80 – 90 % sommergrüne nordische Harthölzer: Vor allem Zucker-Ahorn (*Acer saccharum* MARSH.), Buche (*Fagus grandifolia*), Gelb-Birke (*Betula alleghaniensis* BRIT.) und einige Weiß-Eschen (*Fraxinus americana* L.) auf den unteren und mittleren Hängen (botan. Bez. nach MITCHELL, 1979). Nadelbäume sind nur zu 10 – 20 % am Baumbewuchs beteiligt.

Trotz des nur 116 km in östlicher Richtung entfernten Ozeans ist das Klima aufgrund der vorherrschenden Windrichtung kontinental dominiert und sehr variabel auf allen Zeitskalen. Pro Jahr treten etwa 111 größere Niederschlagsereignisse auf. Es gibt lange kalte Winter und kurze kalte (für südliche Lage) Sommer. Die mittlere Temperatur beträgt im Januar –9 °C und im Juli 18 °C. Der Niederschlag ist über die Monate gleichmäßig verteilt. Im Jahr fallen durchschnittlich 1400 mm Niederschlag, davon ein Drittel bis ein Viertel als Schnee. Die gute Durchlässigkeit des Oberbodens verhindert in der Regel ein Gefrieren des Bodens, bevor sich Mitte Dezember eine schützende Schneedecke darüber legt. Diese schmilzt im April und bewirkt das Jahresmaximum im Wasserstand des Hubbard Brook.

Der „Hubbard Brook Experimental Forest“ wird von der Northeastern Forest Experiment Station, U. S. Department of Agriculture, Randor, Pennsylvania betrieben. Er gehört neben 21 anderen Untersuchungsgebieten zum „Long Term Ecological Research (LTER)“ Netzwerk der USA (http://lternet.edu/network/sites), das von der National Science Foundation gefördert wird. Die 8 oben genannten Teilgebiete im Hubbard Brook-Gebiet sind im Vergleich zu anderen Untersuchungsgebieten (des LTER) sehr gut bezüglich der Menge und Verteilung von Regensammlern und der Qualität der Abflussmessung durch V-Wehre instrumentiert (POST et al., 1998). Messdaten sind auf dem Internet unter der URL http://www.hbrook.sr.unh.edu/data/data.htm verfügbar. Von dort stammen die in dieser Arbeit verwendeten Zeitreihen von täglichem Freilandniederschlag und Abfluss der acht Watersheds. Tabelle 4-1 gibt einen Überblick über die Größe der Gebiete, deren Behandlung und die zur Verfügung stehenden Messzeiträume. In Abb. 4-7 werden in einem 10-jährigem Zeitraum exemplarisch Südhang- (W 1, 2), Nordhang- (W 7) und Kahlschlag- (W 2) Gebiete bezüglich des Abflusses und Niederschlags verglichen.

4.6 H. J. Andrews Experimental Forest (USA)

Das Untersuchungsgebiet „H. J. Andrews Experimental Forest“ liegt im westlichen Bereich der Kaskaden in Oregon (44° 12‘ nördliche Breite, 122° 12‘ westliche Länge) im Nordwesten der USA. Es wird auf der Internet-Seite http://www.fsl.orst.edu/lter/research/compplfr.htm beschrieben, der die Informationen in diesem Abschnitt entnommen sind. Zugang zu weiteren Informationen über die Forschungsaktivitäten und den Daten erfolgt unter der URL: http://www.fsl.orst.edu/lterhome.html. Das 6400 ha große Einzugsgebiet des Lookout Creek liegt zwischen 410 und 1630 m über NN.

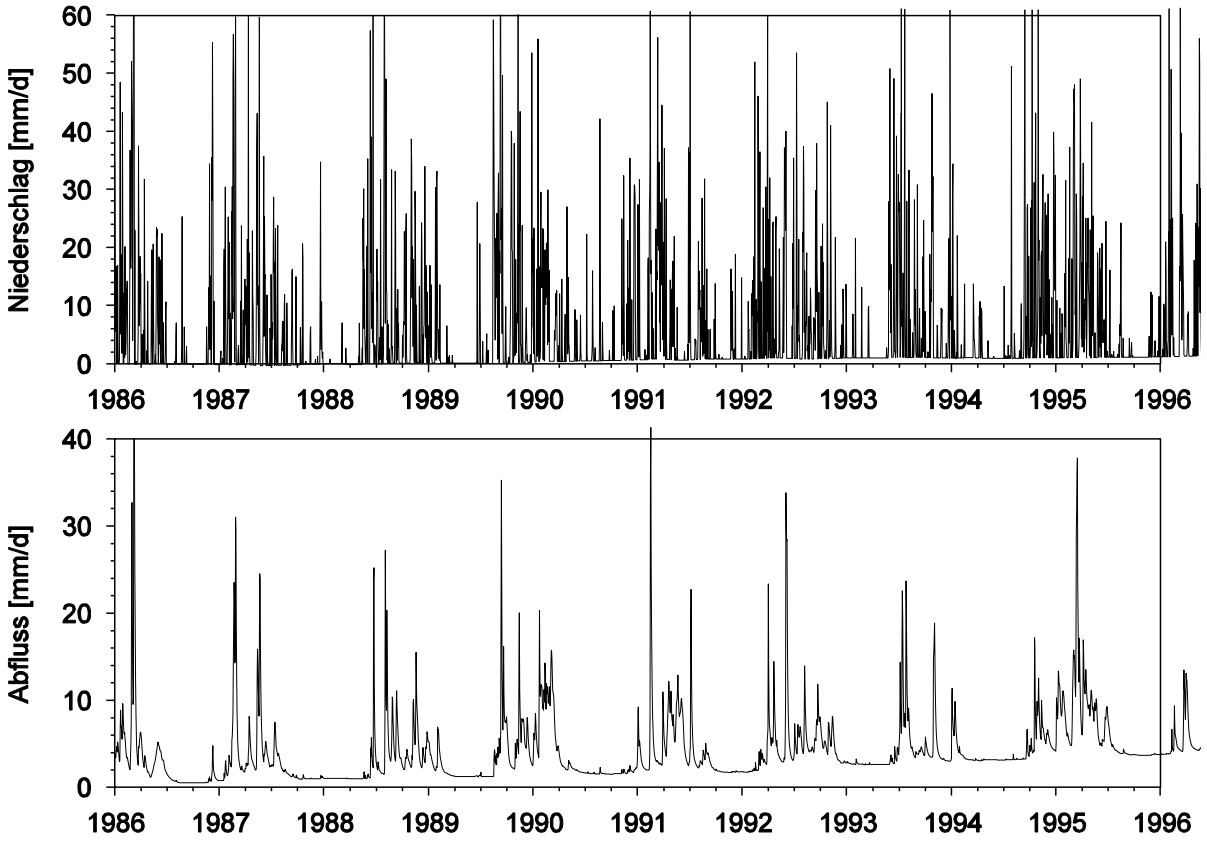

Abb. 4-8. Tägliche Mengen von Niederschlag (W 2) und Abfluss (W 8) in Andrews Forest vom 01.01.1986 bis 31.12.1995.

Das anstehende Gestein wurde im Miozän gebildet und besteht aus vulkanischen Ablagerungen in den tieferen Lagen und andesitischer Lava sowie jüngeren Gesteinen der Hochkaskaden in den höheren Lagen des Andrews Forest. Die Eiszeiten haben auf diesem Untergrund eine stark zerklüftete, lokal steile Landschaft geformt. Aus dem Ausgangsmaterial haben sich hauptsächlich Inceptisol- (Roh-) Böden und lokal Alfisols (Braunerden) und Spodosols (beginnende Podsolierung) entwickelt.

Noch bis 1948 war das Andrews-Gebiet von unberührtem Wald bedeckt: zu 65 % mit 400 – 500 Jahre alten Bäumen und zu 35 % mit 50 – 150 Jahre alten Beständen aufgrund von Bränden. Der Holzeinschlag begann 1950 und betraf 30 % der Fläche, auf der heute ein junger heterogener Wald wächst. Alte Waldbestände mit über 400 Jahre alten Bäumen stehen noch auf 40 % der Fläche und 100- bis 140-jährige Bestände auf 20 % der Fläche. Die Coniferen-Wälder bestehen in den unteren Lagen vor allem aus Douglasien (*Pseudotsuga*), Westlicher Hemlockstanne (*Tsuga heterophylla* (RAF.) SARG.) und „Western Red"-Zeder und in den oberen Lagen aus Edel-Tanne (*Abies procera* REHD.), „Pacific Silver"-Tanne, Douglasien und Westlicher Hemlockstanne (dt. und bot. Namen nach MITCHELL, 1979). Die Wälder der unteren und mittleren Lagen zählen zu den größten und produktivsten Wäldern der Welt mit einer mittleren Höhe von 75 m.

Nasse, milde Winter und trockene, kühle Sommer prägen das maritime Klima. Die Temperatur in 430 m Höhe liegt bei +1 °C im Januar und 18 °C im Juli. Der mittlere Jahresniederschlag variiert von 2300 mm in tieferen Lagen (hauptsächlich als Regen) bis 3550 mm in höheren Lagen (hauptsächlich als Schnee). Er fällt vor allem von November bis März. Abflussspitzen treten im Allgemeinen im November bis Februar auf, wenn warmer Regen auf Schnee fällt.

Andrews gehört wie Hubbard Brook (siehe 4.5) zum LTER-Netzwerk der USA. Seit 1953 werden die Abfüsse von neun kleineren Einzugsgebieten, Watersheds (W 1, 2, 3, 6, 7, 8, 9, 10 und Mack Creek), des Lookout Creek kontinuierlich aufgezeichnet. Diese haben eine Größe von 9 – 600 ha und eine Höhe von 460 – 960 m. In den Teilgebieten wurden Durchforstungsmaßnahmen durchgeführt, deren Auswirkungen auf die Hydrologie untersucht werden. In die bewaldeten Kontrollgebiete 2, 8 und 9 wurde nicht eingegriffen. Weitere Details sind von D. Henshaw 1994 unter der URL http://www.fsl.orst.edu/lter/navigafr.htm beschrieben.

Daten zum „H. J. Andrews Experimental Forest“ werden von der Forest Science Data Bank, einer Partnerschaft des Department of Forest Science, Oregon State University, und der U. S. Forest Service Pacific Northwest Research Station, Corvallis, Oregon, auf den oben angegebenen Internet-Seiten zu Verfügung gestellt. Von dort liegen die täglichen Abflussmessungen der Gebiete

- W 3 vom 01.10.1952 – 13.02.1996,
- W 6 vom 01.10.1963 – 30.09.1996,
- W 7 vom 01.10.1963 – 30.09.1987 und 01.10.1994 – 30.09.1996,
- W 8 vom 01.10.1963 - 30.09.1996,
- W 9 vom 01.10.1968 – 30.09.1996 und
- W 10 vom 01.10.1968 – 30.09.1996

vor. Die Abflussmessung im H. J. Andrews Experimental Forest erfolgt durch Kanäle und wird von POST et al. (1998) als ungenau bezeichnet. Zudem gibt es in dem gesamten Gebiet nur zwei Niederschlagssammler, die den Zeitraum der Abflussmessungen abdecken. Davon wurde in dieser Arbeit nur die längste Aufzeichnung vom 01.10.1957 – 31.12.1996 in Watershed 2 verwendet. Abb. 4-8 zeigt einen Ausschnitt dieser Zeitreihe im Vergleich mit dem Abfluss des Kontrollgebietes 8. Die anderen Aufzeichnungen von Niederschlägen erfolgten nicht in den Watersheds und beginnen 1963, 1979, 1987, 1994 und 1995.

Wegen der ungenauen Abflussmessungen und unvollständigen Aufzeichnungen der Niederschläge sollen die vergleichenden Untersuchungen zum Einfluss unterschiedlicher Vegetationsdecken durch Rodungen auf den Wasserhaushalt von Einzugsgebieten nur mit den Watersheds von Hubbard Brook, nicht aber mit denen von Andrews, durchgeführt werden. Die Abflussmessungen von Andrews sollen lediglich für einen Vergleich der verschiedenen in diesem Kapitel beschriebenen Einzugsgebiete untereinander herangezogen werden.

4.7 Gwynns Falls (USA)

Zur Untersuchung der Frage inwieweit der Bewuchs die Dynamik des Abflusses eines Einzugsgebietes (aus Sicht der Komplexitätsmaße) beeinflusst, sollten auch nahezu unbewachsene oder urbane Einzugsgebiete betrachtet werden. Als extremes Beispiel dafür wurde zunächst an die minütlichen Messungen von Straßenabflüssen einzelner Niederschlagsereignisse in Bayreuth von STRIEBEL (1994) gedacht. Diese episodenhaften Aufzeichungen von kleinen Gulli-Einzugsgebieten eignen sich jedoch nicht für einen Vergleich mit den langjährigen Aufzeichnungen der ansonsten untersuchten Einzugsgebiete im Hektar-Maßstab. Als Alternative bieten sich langjährige Abflussmessungen urbaner Einzugsgebiete an, z. B.:

Das Einzugssgebiet „Gwynns Falls“ liegt in Baltimore City und Baltimore County im Bundesstaat Maryland (39° 15‘ nördliche Breite, 76° 30‘ westliche Länge) im Osten der USA. Es wird auf der Internet-Seite http://baltimore.umbc.edu/lter/description/working/description.

Tabelle 4-2. Tägliche Abflussmessungen der USGS in Gwynns Falls. Neben Nummer und Name der Messstation ist die Größe der Entwässerungsfläche, sowie Beginn und Ende der lückenlosen Aufzeichnungen genannt.

Station Nr.	Name	Größe [ha]	Beginn	Ende
01589200	Gwynns Falls Nr Owings Mills	1269	01.07.1958	30.09.1975
01589300	Gwynns Falls At Villa Nova	8418	01.02.1957	30.09.1988
01589330	Dead Rn At Franklintown	1430	01.10.1959	30.09.1987

htm beschrieben, der die Informationen zu diesem Abschnitt entnommen sind. Das 17150 ha große urbane Gebiet ist von Nordwest nach Südost geneigt und entwässert in die Chesapeake Bay. Von drei Teil-Einzugsgebieten liegen langjährige Messreihen des Abflusses in täglicher Auflösung vor. Diese sind in Tabelle 4-2 kurz charakterisiert. Die Daten gehören zum Messprogramm der U.S. Geological Survey (USGS), das von WAHL et al. (1995) beschrieben wird. Sie sind im Internet auf den Seiten der USGS unter http://waterdata.usgs.gov/nwis-w/MD/index.cgi?statnum= verfügbar, wobei hinter dem Gleichheitszeichen die Stationsnummer (aus Tabelle 4-2) einzufügen ist.

Die Topographie von Gwynns Falls ist von fluvialer Erosion mit dem gemäßigten und feuchten Klima der mittelatlantischen Küsten geprägt. Auf diese Weise ist eine Landschaft mit sanften Neigungen sowie Bergen mit steilen Hängen und herausragenden Felsen entstanden. Die jährlichen 1090 mm Niederschlag sind insgesamt gleichmäßig über das Jahr verteilt. Die Evapotranspiration ist im Juli maximal und mit für die nur 380 mm Abfluss verantwortlich. Im Sommer fallen etwa 10 % mehr Niederschlag als zu den anderen Jahreszeiten. Lokale, kurze und heftige Niederschläge von Stürmen und Hurricanes sind ebenfalls im Sommer typisch. Das große Wasservorkommen und der Einfluss südlicher Winde tragen zu einer relativ hohen Luftfeuchtigkeit über das ganze Jahr bei.

Im Gwynns Falls Wassereinzugsgebiet leben zwischen 405955 (1970) und 356165 (1990) Menschen. Die meisten davon sind im unteren Bereich (z. B. Franklintown, siehe Abb. 4-9) an der Chesapeake Bay angesiedelt. Dort ist bis zu 90.4 % der Fläche mit Siedlungen, Industrie und Gewerbe verbaut. Die oberen Teil-Einzugsgebiete (z. B. Owings Mills, siehe Abb. 4-9) bestehen zu 90.8 % aus bewaldeten, landwirtschaftlichen und offenen Flächen. 1970 setzte sich die Fläche aus 10.5 % Landwirtschaft, 24.8 % Wald, 64.6 % Bebauung und 0.1 % Wasser zusammen. 1990 waren es 6.7 % Landwirtschaft, 18.9 % Wald, 74.3 % Bebauung und 0.1 % Wasser. Die bebaute Fläche hat also trotz sinkender Einwohnerzahlen zugenommen. Mit seiner Flächennutzung unterscheidet sich Gwynns Falls deutlich von den zuvor beschriebenen nahezu vollständig bewaldeten Einzugsgebieten. Es soll daher mit diesen Gebieten bezüglich der Abflussdynamik aus informationstheoretischer Sicht verglichen werden.

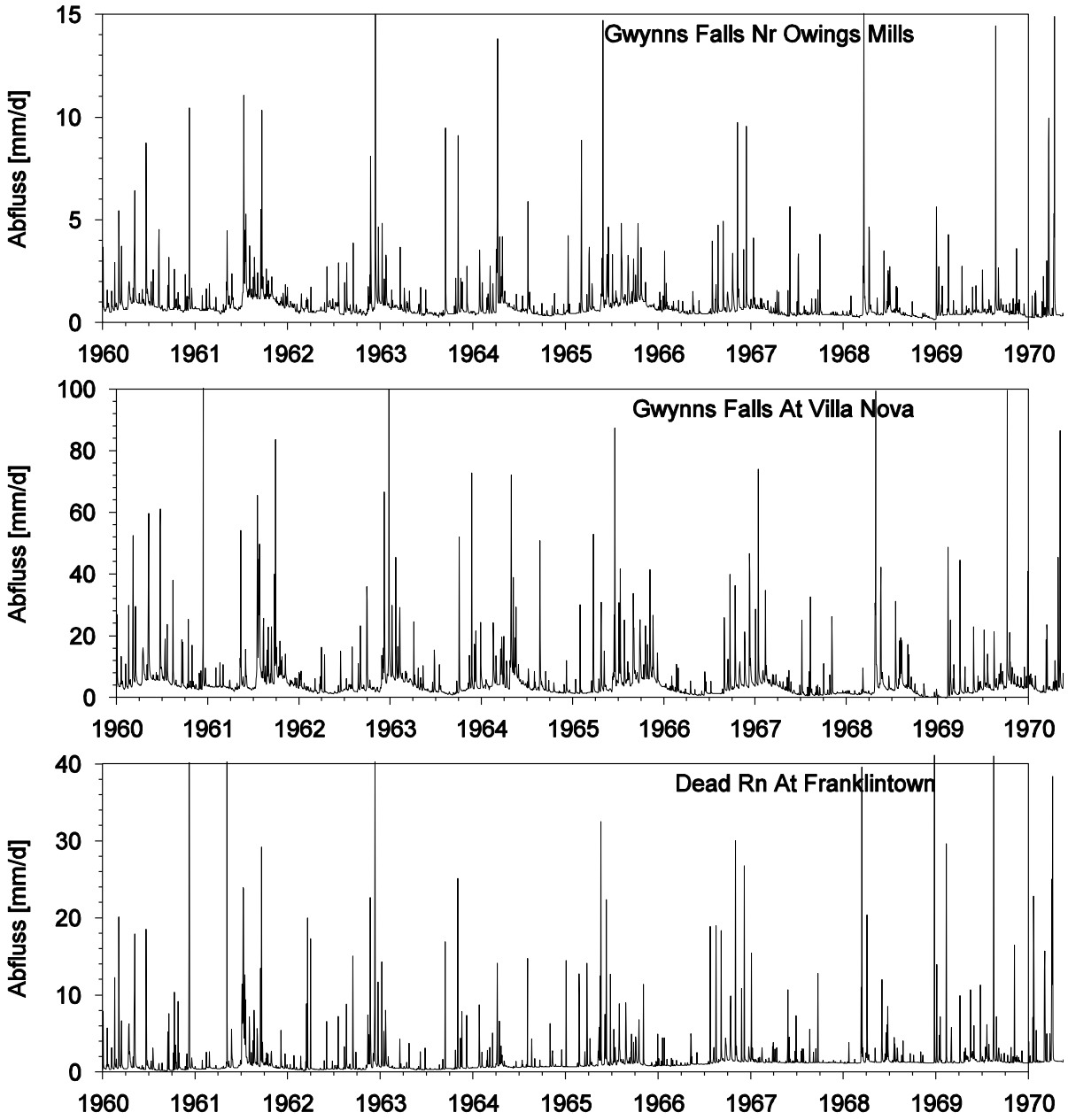

Abb. 4-9. Tägliche Mengen von Abfluss in Gwynns Falls vom 01.01.1960 bis 31.12.1969. *Oben und Mitte:* ein bewirtschaftetes ländliches Teilgebiet. *Unten:* ein urbanes Teilgebiet.

5 Analysen und Ergebnisse

5.1 Information entlang von Fließwegen

Der Vergleich der Rohdaten in den Gebieten „Lehstenbach“ und „Steinkreuz“ (siehe 4.1 und 4.2) läßt darauf schließen, dass die Dynamik im Sinne einer höheren Variabilität der Saugspannungen mit zunehmender Bodentiefe abnimmt. Die Dynamik der Niederschlagsmengen ist besonders hoch und hat zumindest bei der 8-jährigen Zeitreihe für das Lehstenbach-Gebiet den Charakter eines Rauschprozesses. Der niedrige Sommer-Abfluss aufgrund der Transpiration der Vegetation bedingt einen Jahresgang in den Abflusszeitreihen, der aufgrund der längeren Aufzeichnungen beim Lehstenbach besonders deutlich zu erkennen ist. Insgesamt wird entlang des Transportweges des Wassers durch das Einzugsgebiet eine Informationsabnahme vermutet (HAUHS & LANGE, 1996a, 1996b; ROMAHN, 1996; LANGE et al., 1997).

In diesem Abschnitt wird die Dynamik des Wassers in verschiedenen Bereichen der Einzugsgebiete „Lehstenbach“ und „Steinkreuz“ durch die Information und Komplexität der Zeitreihen quantifiziert. Damit soll festgestellt werden, wie sich die Dynamik des Wassers beim Durchlaufen des Gebietes, vom Niederschlag durch den Boden zum Bach, ändert. Es werden die Zeitreihen von Niederschlag und Abfluss sowie die Saugspannungen in verschiedenen Bodentiefen verglichen.

Für das Einzugsgebiet der Langen Bramke wurde bereits eine prinzipielle Informationsabnahme (Maxima der Metrischen Entropie) der Saugspannungen mit zunehmender Bodentiefe festgestellt (ROMAHN, 1996; LANGE et al., 1997). Dies galt für unterschiedliche Auflösungen (15 Minuten bis 12 Stunden). Die Entropie-Maxima der Saugspannungen waren bei hoher Auflösung sehr klein und nahmen systematisch (doppellogarithmisch-linear) mit gröberer Auflösung zu. Nur die Entropie-Maxima der Saugspannung am tiefsten Tensiometer (300 cm) lagen je nach Auflösung zwischen den Werten in 15 cm und 150 cm Tiefe. Die Ursache dafür ist noch unbekannt.

Bei der Berechnung der Informations- und Komplexitätsmaße von Saugspannungen für das Lehstenbach- und Steinkreuz-Gebiet soll von einer täglichen Auflösung ausgegangen werden. Die Informationen und Komplexitäten sollen mit den entsprechenden Werten von Niederschlag und Abfluss verglichen werden. Höhere Auflösungen weisen eine hohe Redundanz auf, wie die Untersuchungen von ROMAHN (1996) und in Abschnitt 5.3 zeigen. Außerdem sind wegen des von ROMAHN (1996) festgestellten systematischen Zusammenhangs der Entropie-Maxima von der Auflösung keine weiteren Erkenntnisse durch die Berücksichtigung anderer Auflösungen zu erwarten. Die tägliche Auflösung ermöglicht darüber hinaus einen Vergleich mit den in dieser Auflösung vorliegenden Niederschlags- und Abflussmessungen anderer Gebiete (siehe Abschnitt 5.4).

Tägliche Niederschlags- und Abflussmengen wurden durch Aggregation aus höher aufgelösten Messungen gewonnen. Bei den Saugspannungen handelt es sich um eine Intensitätsvariable, die mit den zugehörigen Kapazitätsvariablen (Wassergehalt, Wasserfluss) durch die hydraulischen Bodencharakteristiken stark nicht-linear verknüpft ist. Für die einzelnen Messpunkte sind diese Charakteristiken (pF-Kurve, Leitfähigkeiten) außerdem unbekannt. Es

ist daher nicht möglich für das Matrixpotential eine Aggregation oder Mittelung der Werte durchzuführen. Unterschiedliche Messauflösungen werden daher durch Dezimierung (Ausdünnung) erzielt, d. h. für ein tägliches Messintervall wird nur jeder 24. Stundenwert berücksichtigt.

Die Information und Komplexität einer Zeitreihe hängt wesentlich von ihrer zeitlichen Auflösung ab. Daher ist für den Vergleich verschiedener Zeitreihen eine einheitliche und möglichst optimale (komplexitätsmaximale) Auflösung unbedingt erforderlich. Dieses Thema wird in Abschnitt 5.3 behandelt.

5.1.1 Lehstenbach

Für das Einzugsgebiet des Lehstenbaches liegen neben Messungen von Niederschlag und Abfluss auch Messungen der Saugspannung des Bodens an je fünf Standorten und in je drei Tiefen (20, 35 und 90 cm) in stündlicher Auflösung vor (siehe 4.1). Diese Daten wurden, wie oben beschrieben, auf tägliche Auflösung dezimiert. Anschließend wurden die Informations- und Komplexitätsmaße H_μ, H_G, H_M, C_{EM}, C_Γ und C_R für jeweils drei binäre statische Partitionierungen (Median, H_μ-maximal und H_G-maximal) berechnet. Dazu waren 51 Programmläufe von SYMDYN erforderlich, da die genannten Maße in jeweils einem Lauf zusammen berechnet werden konnten. Von den Informationen und Komplexitäten der Saugspannungen wurden für jede Tiefe Mittelwert und Standardabweichung berechnet. Damit kann die Signifikanz von Unterschieden beurteilt werden.

Bei allen Partitionierungen und Entropien (H_μ, H_G, H_M) wurde eine prinzipielle Informationsabnahme der hydrologischen Dynamik vom Niederschlag mit zunehmender Bodentiefe festgestellt (siehe Abb. 5-1 für H_G). Dies bestätigt zumindest hier die Gültigkeit der Filterhypothese nach HAUHS & LANGE (1996a, 1996b). Der Niederschlag war in allen neun Fällen mit Abstand das informationsreichste Signal. Sein Entropiewert unterscheidet sich nur gering von weissem Rauschen (H_μ, H_G um 0.9). Die Niederschlags-Entropien lagen jedoch um mehrere Standardabweichungen über denen der Saugspannung in 20 cm Tiefe. Letztere lagen wiederum über den Entropien der Saugspannungen in 35 cm. Dieser Unterschied war aber nur wenig signifikant und bei H_G und H_G-maximaler Partitionierung am größten. Die Information

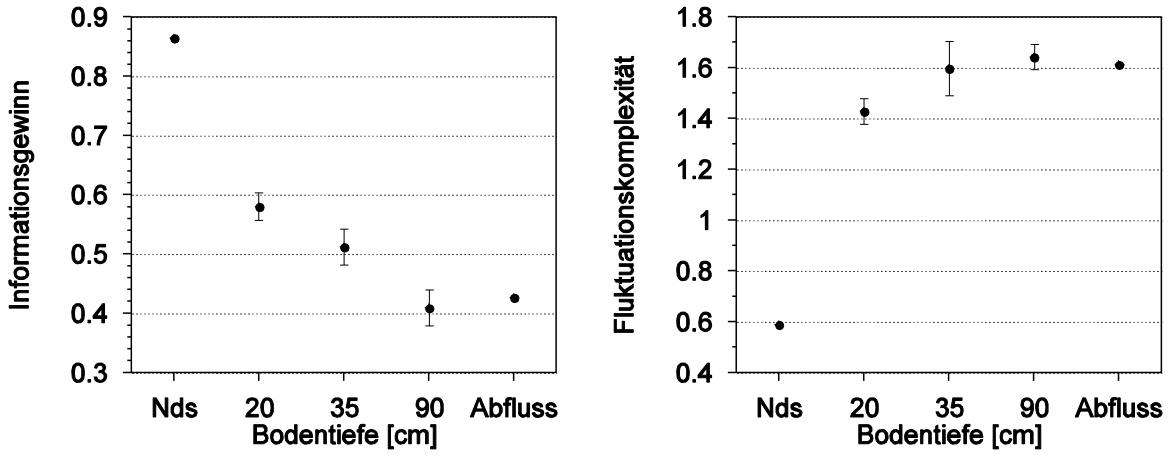

Abb. 5-1. Information und Komplexität hydrologischer Zeitreihen im Einzugsgebiet des Lehstenbaches. Tägliche Auflösung. Für die Saugspannungen sind Mittelwerte und Standardabweichungen der Maße von den jeweils fünf Standorten in jeder Tiefe eingetragen. Binäre statische H_G-maximale Partitionierung. Wortlänge 4. Maximal überlappender 4-jähriger Vergleichszeitraum (siehe 4.1).

des Abflusses lag in allen Fällen zwischen der der Saugspannungen in 35 cm und 90 cm Tiefe. Diese unterschieden sich jeweils signifikant. In 90 cm Tiefe wurde ein Informationsgewinn von 0.3 nicht unterschritten. Die Standardabweichungen waren bei Entropie-maximaler Partitionierung geringer als bei Median-Partitionierung.

Aus dem Informationsverlauf kann eine effektive mittlere Eindringtiefe des Niederschlagssignals von 35 cm bis 90 cm gefolgert werden. Dieses Ergebnis ist plausibel vor dem Hintergrund der teilweise oberflächennah anstehenden Granite in diesem Einzugsgebiet. Aufgrund der wegen der heterogenen hydraulischen Eigenschaften unbekannten Fließwege kann dieses Aussage nicht im Detail überprüft werden. Es liegt aber ein Vorteil der hier angewendeten Auswertungsverfahren darin, dass Aussagen über die angemessenen Einbautiefen für Tensiometer und Zeitauflösungen (siehe 5.3.2) für Matrixpotentiale möglich sind, ohne dass dazu bereits Kenntnisse über die Geometrie und Dynamik der Fliesswege notwendig sind.

Allenfalls kann man nachträglich eine Interpretation der Ergebnisse im Hinblick auf möglich Fliesswege versuchen, wobei aber das Problem der Nicht-Eindeutigkeit im Vergleich zu Ergebnissen prozessorientierter Modelle unverändert bestehen bleibt. In diesem Sinne wäre eine mögliche Erklärung für die höhere Information des Abflusses gegenüber der Saugspannung in 90 cm Tiefe ein oberflächennahes Abfließen des Wassers bei starken Regenfällen. Der vertikal versickernde Niederschlag wird auf dem Weg zum anstehenden Tiefenwasser in seiner Dynamik gedämpft. Dieses Tiefenwasser trägt aber erst deutlich später (Tage) und länger als Reaktion auf den Niederschlag zum Bachabfluss bei (DUNNE & BLACK, 1970; FREEZE, 1972). Die Abflussdynamik hängt also mit der Dynamik des Tiefenwassers und der Saugspannung des Bodens in 90 cm Tiefe zusammen. Durch den oberflächennahen Grundwasserabfluss, der das Bachwasser eher erreicht als das Tiefenwasser, erhält die Dynamik des Abflusses ein zusätzliches Signal.

Die Fluktuations- und Rényi-Komplexitäten der Saugspannungen und des Abflusses unterschieden sich kaum voneinander (siehe Abb. 5-1 für C_Γ). Sie lagen mit einem Wert um 1.6 auf einem einheitlich für C_Γ hohem Niveau. Die Maximalität der Komplexität hängt mit der Optimalität der täglichen Auflösung für die Saugspannungen und den Abfluss zusammen und wird in Abschnitt 5.3 erklärt. Die Komplexitäten (C_Γ, C_R) des Niederschlags lagen mit einem Wert um 0.6 jeweils deutlich und signifikant niedriger. Mit der hohen Information der Niederschläge weist dies auf den Zufallscharakter der täglichen Niederschlagsmengen hin.

Die Effektive Maßkomplexität zeigte bei Entropie-maximaler Partitionierung den schon früher (siehe 2.6.1) beobachteten spiegelbildlichen Verlauf zu den Informationsmaßen, d. h. ihr Wert nahm vom Niederschlag mit der Bodentiefe zu und der Abfluss-C_{EM}-Wert liegt zwischen den C_{EM}-Werten der Saugspannungen in 35 cm und 90 cm Tiefe.

5.1.2 Steinkreuz

Für das Einzugsgebiet „Steinkreuz“ ist die Datenlage ähnlich wie im Lehstenbach-Gebiet. Neben Niederschlag und Abfluss liegen Messungen der Saugspannung des Bodens an je fünf Standorten und in je drei Tiefen (20, 90 und 200 cm) in stündlicher Auflösung vor (siehe 4.2). Die Daten wurden ebenfalls auf tägliche Auflösung gebracht und der gleichen Analyse unterzogen wie beim Einzugsgebiet des Lehstenbaches.

Auch hier war die Information des Niederschlags bei allen Partitionierungen mit Abstand am höchsten und lag ebenfalls (wie im Lehstenbach-Gebiet) bei einem Wert (H_μ, H_G) von 0.9. Die Information nahm mit zunehmender Bodentiefe ab, wobei nur der Unterschied zwischen

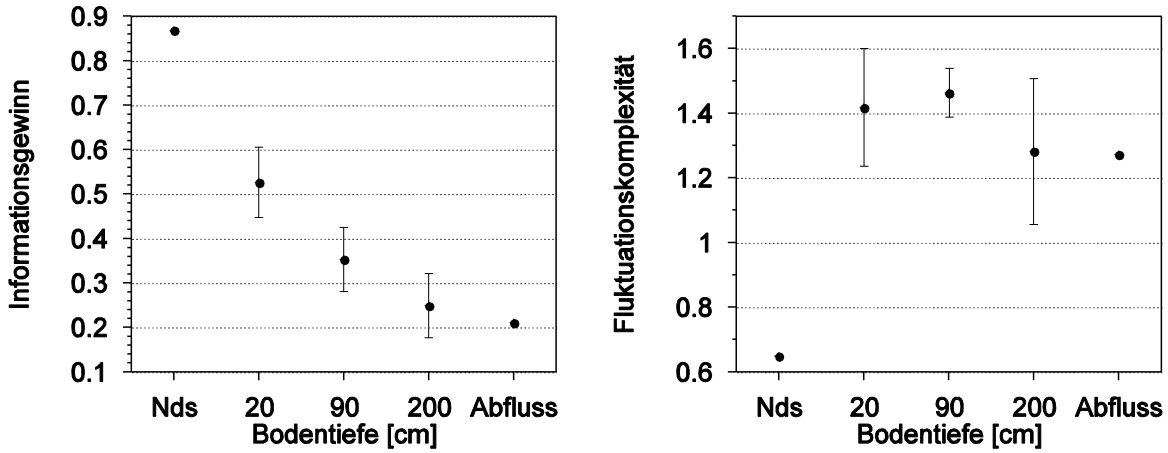

Abb. 5-2. Information und Komplexität hydrologischer Zeitreihen im Einzugsgebiet „Steinkreuz". Für die Saugspannungen sind Mittelwerte und Standardabweichungen der Maße von den jeweils fünf Standorten in jeder Tiefe eingetragen. Binäre statische H_G-maximale Partitionierung. Wortlänge 4.

20 cm und 90 cm Tiefe als signifikant gewertet werden kann, nicht aber der Unterschied zwischen 90 cm und 200 cm Tiefe (siehe Abb. 5-2 für H_G). Die Informationen der Saugspannung in 200 cm Tiefe und des Abflusses sind nicht signifikant verschieden. Bei H_G und H_G-maximaler Partitionierung wurde das mittlere Informationsminimum beim Abfluss angenommen; bei H_μ war die mittlere Saugspannungsinformation etwa gleich groß wie die Abfluss-Information. Die Informationen der Saugspannung in 20 cm und 90 cm Tiefe sind in den Gebieten Lehstenbach und Steinkreuz ähnlich, wenn auch im Mittel etwas geringer beim Steinkreuz. Mit einem H_G-Wert von 0.2 liegt das Informationsminimum im Steinkreuz deutlich unterhalb des Minimums beim Lehstenbach. Es findet also eine stärkere Glättung des Niederschlagssignals zum Abfluss hin statt.

Aus dem Informationsverlauf kann auch hier eine effektive vertikale Eindringtiefe des Niederschlagssignals definiert werden. Diese liegt etwa bei 200 cm und damit deutlich tiefer als beim Lehstenbach, was auf die tiefengründig verwitterten und vor allem durchlässigeren Böden im Steinkreuz-Gebiet zurückgeführt werden kann. Dort liegen tiefgründige sandige Braunerden über Keupersandsteinen vor (siehe 4.2), während im Lehstenbach-Gebiet tiefgründige aber dichte Solifluktionsböden über Granit vorherrschen (siehe 4.1). Der hohe Vermoorungsgrad im Lehstenbach-Gebiet ist ein weiterer Hinweis für die geringere Durchlässigkeit des Bodens, bzw. für das oberflächennah anstehende Ausgangsgestein.

Die Komplexität des Niederschlags ist auch im Steinkreuzgebiet in allen Fällen mit Abstand am geringsten (siehe Abb. 5-2). Sie erreicht wie die Information die gleiche Größenordnung wie im Lehstenbachgebiet, das nur 106 km Luftlinie entfernt ist. Die Komplexitäten (C_Γ, C_R) der Saugspannungen und des Abflusses sind vergleichsweise hoch und nicht signifikant voneinander verschieden. Bei Entropie-maximaler Partitionierung nehmen die Komplexitäten systematisch (aber nicht signifikant) mit der Bodentiefe bis zum Abfluss ab. Die Effektive Maßkomplexität zeigt wieder einen prinzipiell spiegelbildlichen Verlauf zu den Informationen bei diesem Partitionierungstyp.

5.2 Abhängigkeit der Abfluss-Information von der Vegetation

Der Einfluss von forstlichen Eingriffen auf den Wasser- und Stoffhaushalt von bewaldeten Wassereinzugsgebieten ist bzw. war ein Forschungsschwerpunkt in den Gebieten „Lange Bramke“, „Hubbard Brook“ und „Andrews“ (siehe 4.3, 4.5 und 4.6). Durch die gute Datenlage (siehe POST et al., 1998) ist „Hubbard Brook“ in besonderer Weise für eine solche Untersuchung geeignet. Von diesem Gebiet liegen die langjährigen täglichen Aufzeichnungen des Niederschlags und Abflusses für acht Teileinzugsgebiete vor. In dreien dieser Teilgebiete wurden dokumentierte Kahlschläge vorgenommen. Von den fünf Referenzgebieten befinden sich zwei in Nordhanglage.

Anhand dieses Datenmaterials und mit Hilfe von Komplexitätsmaßen soll der Einfluss der Bewaldung auf die Dynamik des Wasseraustrages im Abfluss untersucht werden. Die Möglichkeiten der Untersuchungsmethode werden mit anderen Methoden — insbesondere der Autokorrelation als einer klassischen Methode zur Zeitreihenanalyse — verglichen.

5.2.1 Die Entwaldung des Watersheds 2 von 1966 bis 1968

Der stärkste experimentelle Eingriff erfolgte im Teilgebiet 2 (W 2) des Hubbard Brook Experimental Forest (siehe Tabelle 4.1 in 4.5). Das Gebiet wurde im Winter 1965/1966 — also in der Vegetationspause — völlig kahl geschlagen. Die gefällten Bäume wurden nicht abtransportiert. In den Sommern 1966, 1967 und 1968 wurden unspezifisch wirkende Herbizide eingesetzt, die eine neue Vegetationsentwicklung verhinderten (USDA, 1996). Erst ab 1969 konnte sich wieder eine neue Vegetation entwickeln.

In der Abflussganglinie ist das Fehlen der Vegetation durch einen konstanten Basisabfluss und eine vermehrte Anzahl von Abflussspitzen im Vergleich zu W 1 und W 7 (Abb. 4-7) zu erkennen. Die Schneeschmelze im April bleibt im Abfluss sichtbar. Im Sommer 1970 ist wieder ein niedrigerer Wasserstand zu erkennen, der auf die Transpiration der jungen Vegetation zurückzuführen ist. Im Sommer 1971 wird wieder Niedrigwasser ähnlich wie in W 1 und W 7 erreicht.

Ein Vergleich der Autokorrelation über den gesamten vorliegenden Messzeitraum von Watershed 2 (1957 – 1993) mit denen der Watersheds 1, 3, 4, 5 und 6 ließ kaum einen Unterschied erkennen. Die Autokorrelationen über den vollständigen Beobachtungszeitraum sind nicht charakteristisch für die Teilgebiete, sondern für das gesamte Einzugsgebiet des Hubbard Brook und werden in Abschnitt 5.4.1 mit denen der anderen Gebiete verglichen. Hier wurden die Autokorrelationen in dem relevanten 3-jährigen Zeitraum (1966 – 1968) und in den vier Jahren vorher, also 1962 – 1965, verglichen. Für den Zeitraum vor der Maßnahme waren nur Daten aus den Gebieten 1 bis 5 und für den Zeitraum während der Maßnahme aus den Gebieten 1 bis 6 vorhanden. Im Zeitraum von 1962 – 1965 waren die Autokorrelationsfunktionen praktisch identisch. Dies gilt auch für die Watersheds 1, 3, 4, 5 und 6 im Zeitraum von 1966 – 1968. Repräsentativ wurde Watershed 3 ausgewählt und in Abb. 5-3 mit Watershed 2 verglichen.

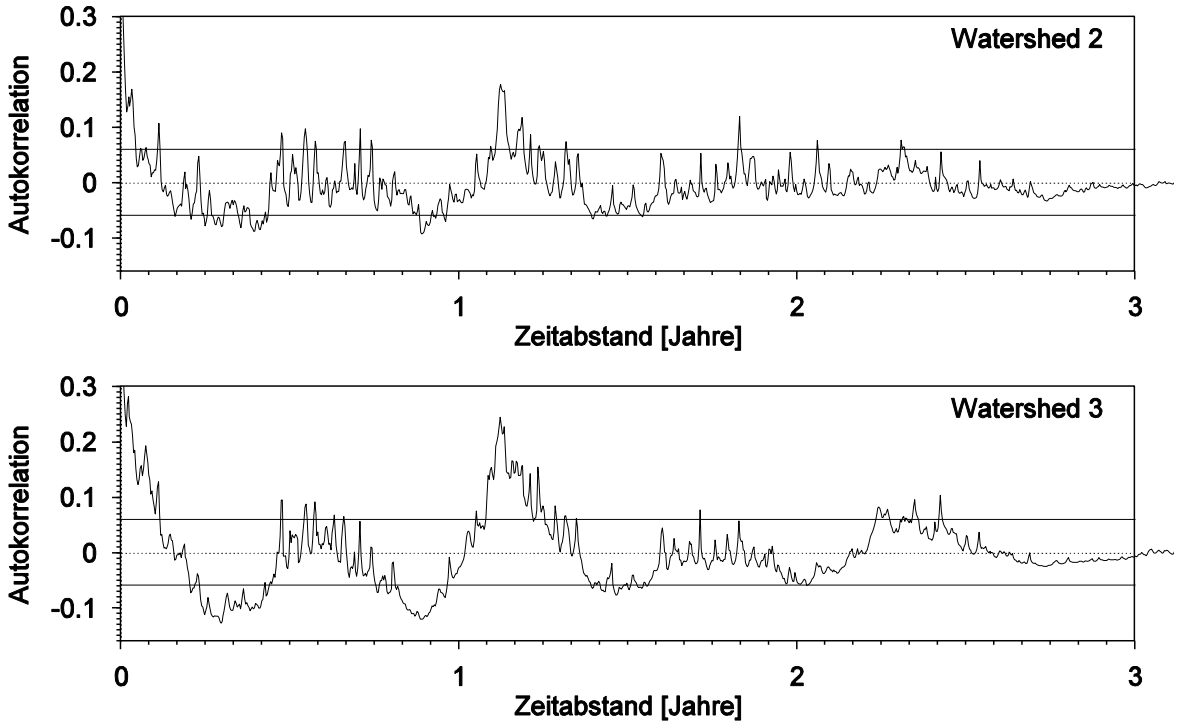

Abb. 5-3. Autokorrelation im Abfluss von Hubbard Brook, W 2 und W 3 von 01.01.1966 – 31.12.1968 (Vegetationspause in W 2).

Die Datenmenge von 1096 Tagen (drei Jahre) ist für eine Jahresganganalyse nach der Empfehlung von HONERKAMP (1994, S. 383) nicht ausreichend. Danach sollte die Autokorrelation nur bis zu ¼ der Datenmenge ausgewertet werden. Die Konsistenz der Analysen für die Referenzgebiete begründet jedoch auch eine Interpretation größerer Zeitabstände. Folgende Unterschiede fallen auf: Der Jahres- und Halbjahresgang bei Watershed 2 (W 2) ist gegenüber dem von W 3 nur gedämpft (weniger signifikant) und nur bis zu einem Lag von einem Jahr erkennbar. Während die Autokorrelation bei W 3 erstmals nach 37 Tagen unter die 5 %-Signifikanzlinie fällt, ist dies bei W 2 schon nach 14 Tagen der Fall. Insgesamt nehmen also die kurz- und mittelreichweitigen Korrelationen bei Fehlen einer aktiven Vegetation ab.

Da die Transinformation zumindest auf der Symbolebene nach HERZEL & GROßE (1995) ein sensibleres Maß für statistische Abhängigkeit ist als die Autokorrelation (siehe 2.2.2), wurde

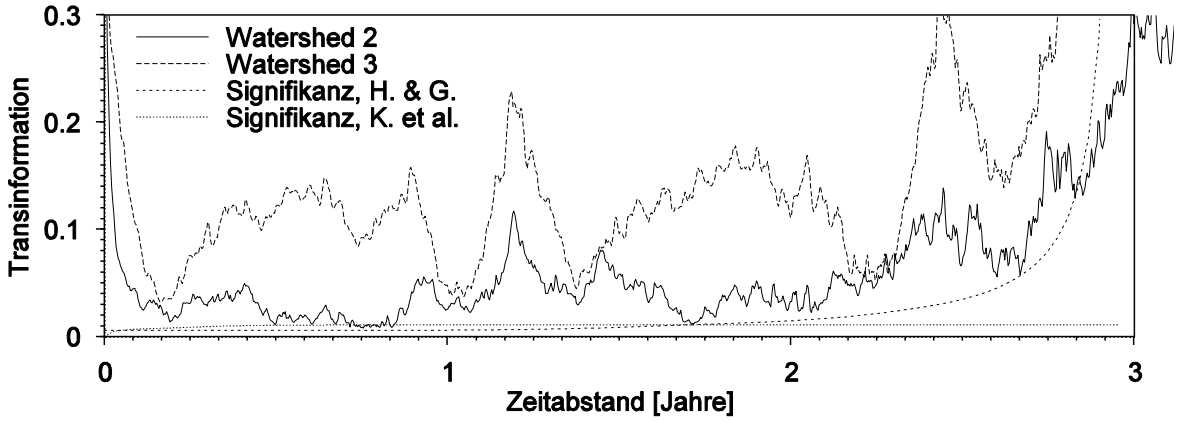

Abb. 5-4. Transinformation für den täglichen Abfluss in Hubbard Brook, W 2 und W 3 von 1966 – 1968 (Vegetationsruhe in W 2). Statische quartäre äquiquantile Partitionierung. Signifikanzschwellen nach HERZEL & GROßE (1995) und KURTHS et al. (1996).

der gerade beschriebene Vergleich auch für die Transinformation durchgeführt. Dazu wurden binäre bis quartäre statische und dynamische äquiquantile Partitionierungen verwendet, d. h. im binären dynamischen Fall wurde bei 0 partitioniert. Die dynamischen Partitionierungen verursachten jeweils ein starkes Rauschen der Transinformation. Trotzdem waren bei W 3 im Gegensatz zu W 2 insbesondere bei ternärem Alphabet deutliche Vierteljahresspitzen der Transinformation erkennbar. Die Interpretation der mit statischer Partitionierung gewonnenen Werte war aufgrund des geringeren Rauschens einfacher: Die Transinformationen von W 3 waren fast überall deutlich größer als die von W 2. Bei binärem Alphabet wurden im interpretierbaren Zeitabstand von bis zu zwei Jahren für W 3 vierteljährliche Spitzenwerte erreicht, die nicht in allen Fällen ein Pendant bei W 2 hatten. Abb. 5-4 zeigt die Transinformation für ein quartäres Alphabet, bei dem die Unterschiede zwischen W 2 und W 3 am deutlichsten waren. Aber nicht nur in diesem Fall, sondern allgemein muss festgestellt werden, dass die Unterschiede zwischen W 2 und W 3 im relevanten Zeitraum wesentlich deutlicher mit der Transinformation erkennbar sind als mit der Autokorrelation.

5.2.2 Vergleich aller acht Hubbard Brook Teileinzugsgebiete

5.2.2.1 Grundsätzliche Unterschiede

Bei der Betrachtung der Rohdaten sind bis auf kurzzeitige vorübergehende Perioden (siehe 5.2.1.1) kaum Unterschiede zwischen den acht Teileinzugsgebieten erkennbar. Auffällig sind kurze Hochwasserperioden im Frühjahr (Schneeschmelze) sowie im Herbst/Winter, während im Sommer Niedrigwasser herrscht. Die Auswirkungen von Starkregenereignissen auf die Abflüsse sind erkennbar. Ein unmittelbarer Zusammenhang von Niederschlag und Abfluss ist nicht immer zu beobachten. Der typische Verlauf der Hochwasserperioden mit einem steilen

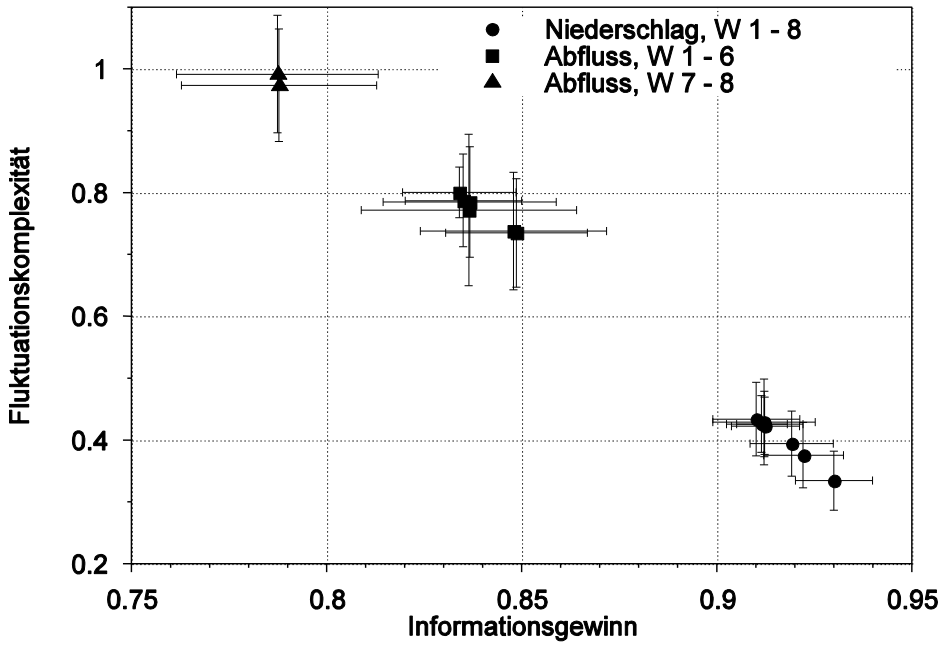

Abb. 5-5. Klassifikation von Standorten sowie Eingangs- und Ausgangssignalen in Hubbard Brook. Mittelwerte und Standardabweichungen von allen benachbarten 4-Jahresintervallen. Binäre, dynamische 0-Partitionierung. Wortlänge 4.

Anstieg und einem allmählichen Rückgang der Abflusspegel läßt sich überall wiederfinden.

Die Autokorrelation wurde sowohl über den jeweiligen maximalen Gesamtzeitraum wie auch über den Zeitraum 1969 – 1993, für den bei allen Teilgebieten Messungen vorliegen (siehe 4.5), berechnet. Diese Funktionen waren ebenfalls wenig spezifisch für die Teilgebiete. Das Gleiche gilt für die entsprechenden Power-Spektren. Die Autokorrelationen charakterisieren das Gebiet insgesamt durch einen dominanten Jahresgang und schwach signifikanten Halbjahresgang (siehe 5.4.1). Lediglich die beiden Gebiete am Nordhang, W 7 und W 8, fallen gegenüber den anderen Gebieten durch einen etwas stärkeren Halbjahresgang auf.

Mit Hilfe von Komplexitätsmaßen ließen sich bei statischen Partitionierungen und Betrachtung der vollständigen Beobachtungszeiträume keine Unterschiede zwischen den Teilgebieten erkennen. Bei dynamischen Partitionierungen fielen die Abflüsse der Nordhanggebiete (W 7 und W 8) durch signifikant höhere Komplexitäten und Informationen auf, wie Abb. 5-5 zeigt. Der Standortunterschied wird auf diese Weise — im Gegensatz zur Autokorrelation — klar herausgearbeitet.

5.2.2.2 Korrelationsanalyse

Um zu einer feineren Unterscheidung von Veränderungen in den Teilgebieten zu gelangen, wurden kurze Teilintervalle (wenige Jahre) für jedes der acht Gebiete analysiert. Die Intervalle wurden, um jeweils einen Monat verschoben, für den ganzen Untersuchungszeitraum berechnet. Um an die Beobachtungen aus Abschnitt 5.2.1 anzuknüpfen, wurde zunächst das

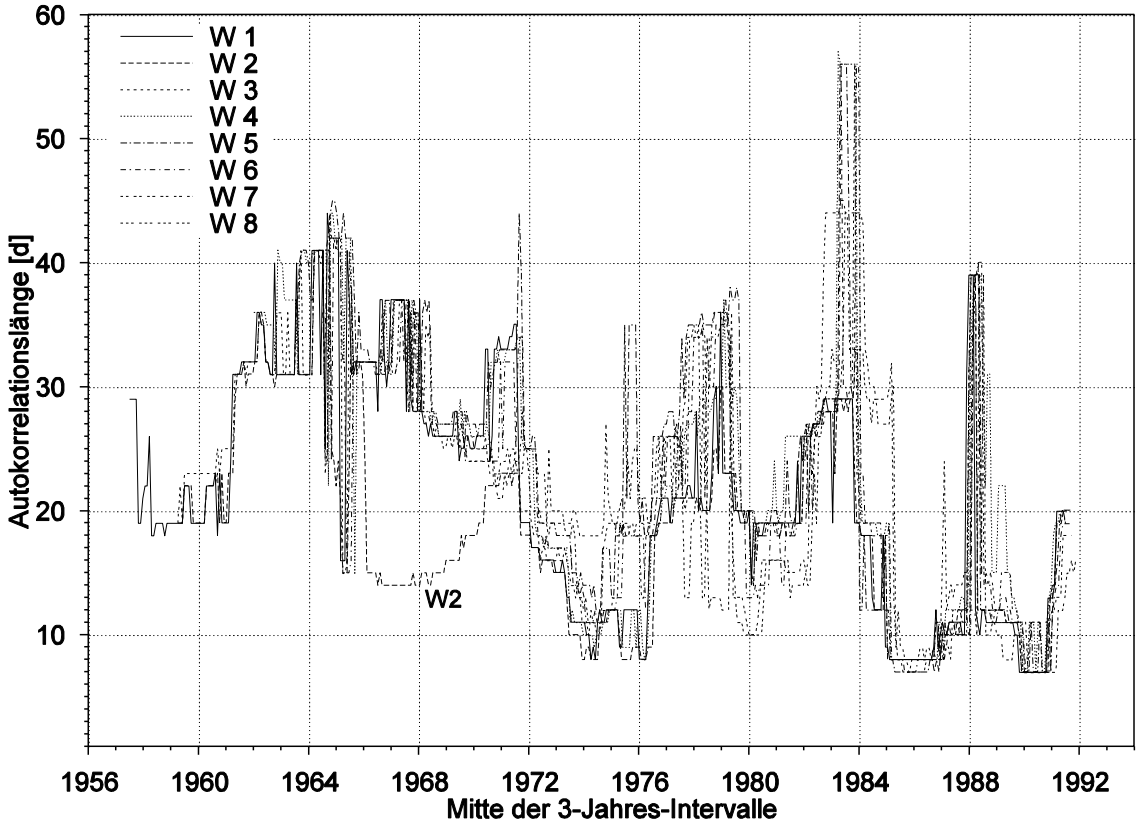

Abb. 5-6. Autokorrelationslängen in 3-Jahresintervallen für die acht Watersheds (W) von Hubbard Brook. Als Autokorrelationslänge gilt hier der Zeitabstand für das erste Absinken der Autokorrelationsfunktion unter das 5 %-Signifikanzniveau. Die 3-Jahresintervalle wurden jeweils um einen Monat überlappend verschoben.

erstmalige Absinken der Autokorrelationsfunktion unter das 5 % Signifikanzniveau für jedes 3-Jahresintervall berechnet. Diese Korrelationslängen sind in Abb. 5-6 dargestellt. Dabei fällt ein — wahrscheinlich klimatisch bedingter — tendenzieller Verlauf der Korrelationslängen auf, der bei allen Teilgebieten ähnlich ist. Auffallend ist aber auch die davon abweichende sprungartige Abnahme der Korrelationslänge im Jahr 1966 bei W 2. Diese bleibt bis 1969 nahezu konstant niedrig und steigt danach wieder an. 1971 / 1972 erreicht die Autokorrelationslänge des Abflusses von W 2 wieder den Wert der anderen Gebiete. Diese Beobachtung stimmt mit dem Zeitpunkt der Rodung von W 2 im Winter 1965 / 1966 und der anschließenden Behandlung mit Herbiziden in den Sommern 1966, 1967 und 1968 überein. Die Vegetation braucht dann etwa drei Jahre, um die Dynamik des Abflusses aus Sicht der Autokorrelation in der gleichen Weise wie vor dem Kahlschlag zu beeinflussen.

BORMANN & LIKENS (1979, S. 143ff) beschreiben die natürliche Wiederbegrünung nach der Behandlung von W 2 und dessen Auswirkungen auf den Wasser- und Nährstoffhaushalt des Gebietes. Auf den Fotos ist bereits 1971 wieder ein erheblicher Bedeckungsgrad mit Sekundärvegetation zu erkennen. Im Mai 1972 ist auf dem Foto kein Boden mehr zu erkennen: Es fallen insbesondere mannshohe Büsche von „Pin cherry“ (*Prunus pensylvanica*) auf. Drei bis vier Jahre nach der Behandlung werden wieder die mit Regression von W 3 vorhergesagten Abflussmengen wie vor der Entfernung jeglicher Vegetation erreicht. Diese Veränderung im Abflussregime geht höchstwahrscheinlich auf den Einfluss der Evapotranspiration zurück.

Im Unterschied zu der relativ geringen Reaktion der Abflussdynamik auf die Entwaldung hatte der Kahlschlag grossen Einfluss auf die im Abfluss gelösten Inhaltsstoffe. drei Jahre nach dem Ende der Behandlung erreicht auch die zuvor deutlich erhöhte Auswaschung von Nährstoffen (Calcium, Kalium, Nitrat u. a.) mit dem Abfluss wieder das Niveau von vor dem

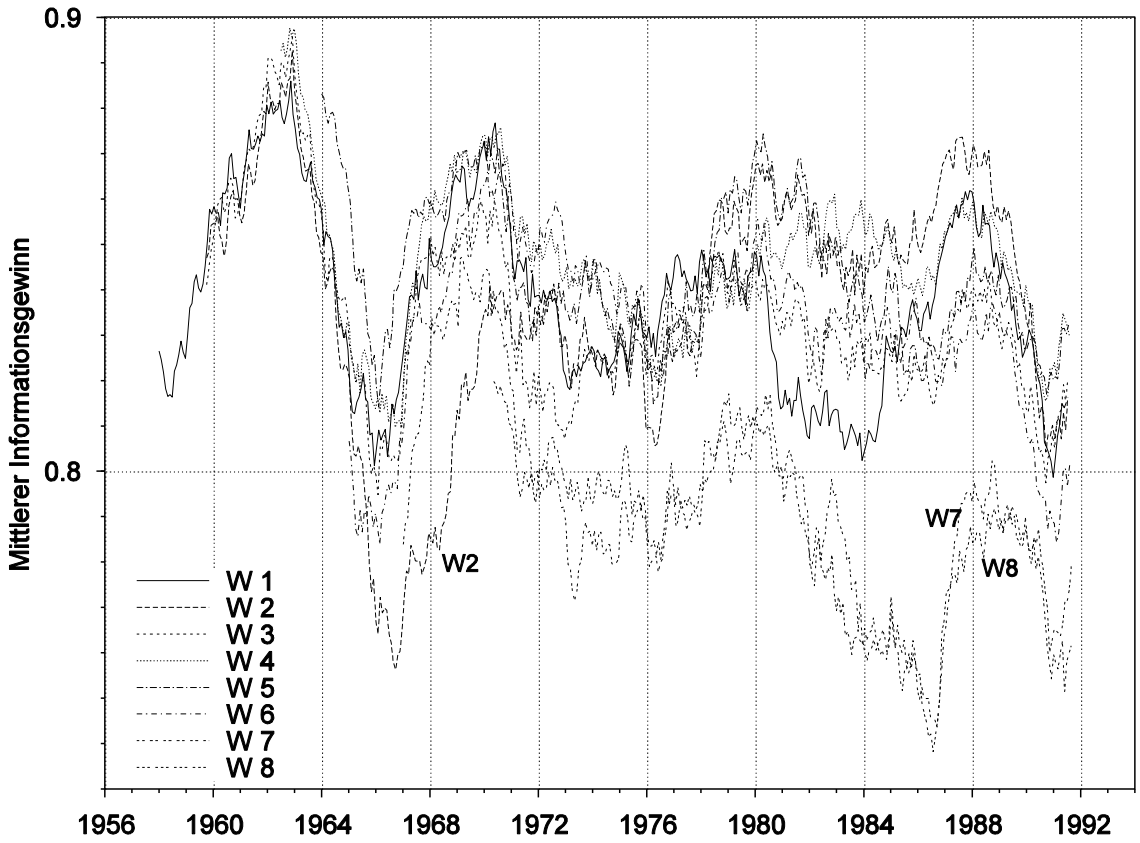

Abb. 5-7. Informationsgewinn des Abflusses von 4-Jahresintervallen für die acht Watersheds (W) von Hubbard Brook bei dynamischer binärer 0-Partitionierung. Wortlänge 4. Die Intervalle wurden jeweils um einen Monat überlappend verschoben.

Kahlschlag.

Die Autokorrelation der Abflüsse ist also eng mit der Transpirationsleistung der Vegetation und den davon beeinflussten Abflussmengen korreliert. Größere Bäume sind zur Erzeugung der hydrologischen Dynamik bezüglich der Korrelationslänge offensichtlich nicht erforderlich. Das erklärt, warum die etappenweise Reduzierung des Waldes auf einen Stammdurchmesser von 2 cm in den Jahren 1970 – 1974 in W 4 sowie auf 5 cm in den Jahren 1983 – 1984 in W 5 hier unentdeckt bleibt.

5.2.2.3 Informationsanalyse

Zur weiteren Analyse wurden nun Informations- (H_μ, H_G) und Komplexitätsmaße (C_{EM}, C_Γ, C_R) in Intervallen von ein, zwei und vier Jahren für die acht Teilgebiete berechnet. Die Betrachtung von nur einem oder zwei Jahren erschwerte die Interpretation der Werte durch vermehrte Schwankungen (vgl. auch Abb. 3-3). Daher beschränkt sich die Darstellung der Ergebnisse auf 4-Jahresintervalle. Der Partitionierungstyp (dynamisch oder statisch) erwies sich als wesentlich für die Unterscheidung von Merkmalen (s. u.). Die verschiedenen Maße lieferten unterschiedliche Bewertungen der Dynamik. Bei Fluktuations- und Rényi-Komplexität und statischer Partitionierung waren die acht Gebiete kaum zu unterscheiden. Von den hier untersuchten Komplexitätsmaßen erwies sich der Informationsgewinn als das sensitivste Maß zur Feststellung von Unterschieden. Daher beschränkt sich die weitere Diskussion und Beschreibung auf H_G.

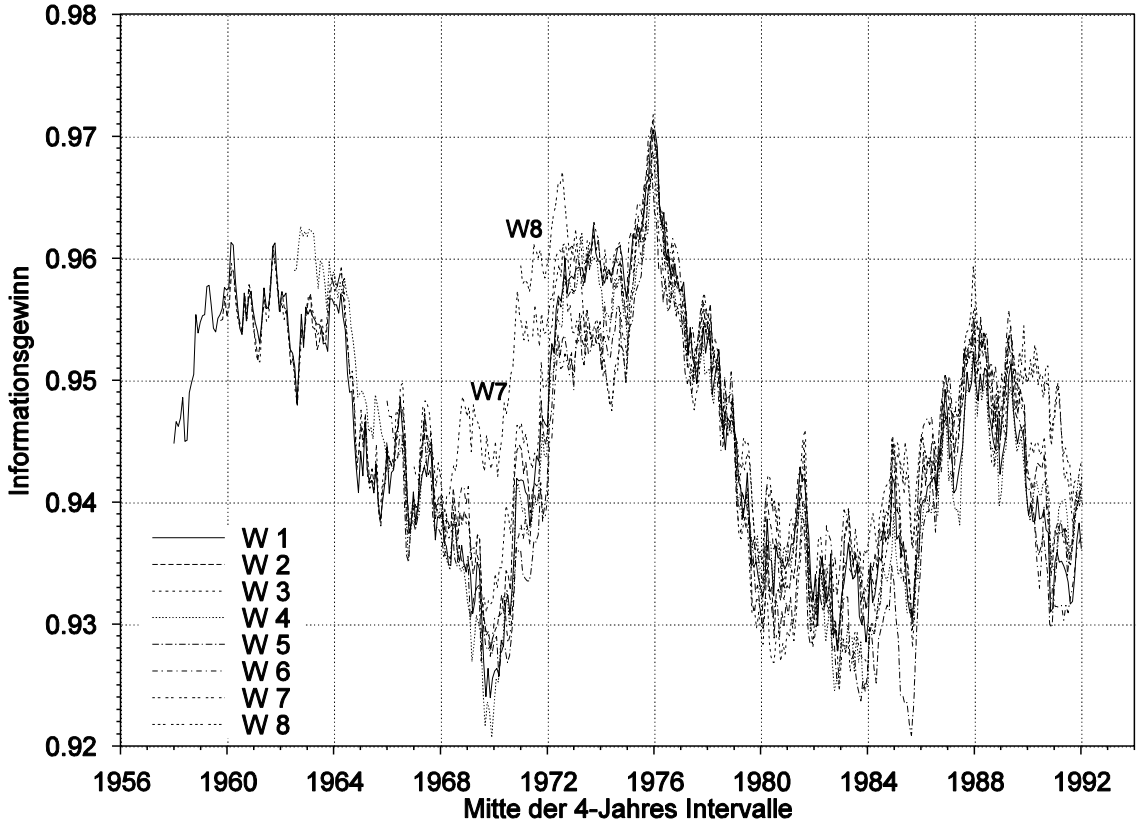

Abb. 5-8. Informationsgewinn des Niederschlags von 4-Jahresintervallen für die acht Watersheds (W) von Hubbard Brook bei statischer H_G-maximaler Partitionierung. Wortlänge 4. Die Intervalle wurden jeweils um einen Monat überlappend verschoben.

Wie bei der Autokorrelation kann auch hier ein gleicher tendenzieller Verlauf der Information bei allen Teilgebieten festgestellt werden. Für den größten Teil des Messzeitraumes sind die Teilgebiete kaum zu unterscheiden. Dieser Informationsverlauf deckt sich nicht mit den Informationsverläufen für die Niederschläge (vgl. Abb. 5-8 mit Abb. 5-7 und Abb. 5-9). Das nach LIKENS & BORMANN (1995) im Beobachtungszeitraum trockenste Wasser-Jahr vom 01.06.1964 bis 31.05.1965 könnte jedoch für den deutlichen Rückgang der Information der Abfluss-Änderungen verantwortlich sein, der in Abb. 5-7 sichtbar ist. Das feuchteste Wasser-Jahr vom 01.06.1973 bis 31.05.1974 könnte mit der hohen Information der Abflussmengen zusammenhängen, die in Abb. 5-9 erkennbar sind. Temperaturmessungen wurden nicht untersucht. Falls die Informations-Verläufe der Abflüsse stark mit den entsprechenden Verläufen einer Klima-Variablen korreliert sind, könnte damit der tendenzielle Kurvenverlauf heraus normiert werden, so dass nur noch Abweichungen von der klimatischen Dynamik auffielen.

Bei dynamischer 0-Partitionierung (Abb. 5-7) wird der Eingriff in W 2 durch eine geringere Information der Abfluss-Änderungen im gleichen Umfang sichtbar wie bei der Autokorrelationslänge. Neu ist hier die deutliche Unterscheidung der beiden Nordhanggebiete (W 7 und W 8), die bereits in der Gesamtbetrachtung in Abb. 5-5 auffielen, die aber in ihrer Korrelationslänge unauffällig waren. Die Maßnahmen in W 4 und W 5 bleiben aber auch hier unentdeckt. Die Niederschläge blieben bis auf einige unerklärte Abweichungen von W 6 und W 8 weitgehend ununterscheidbar bezüglich der Information ihrer Änderungen. Somit fielen bei Betrachtung der Informationsdifferenzen zwischen den Niederschlags- und Abflussänderungen die gleichen Unterschiede (W 2 und Hanglage) im gleichen Umfang auf, wie bei der alleinigen Betrachtung der Abflussänderungen. Eine dynamische H_G-maximale Partitionie-

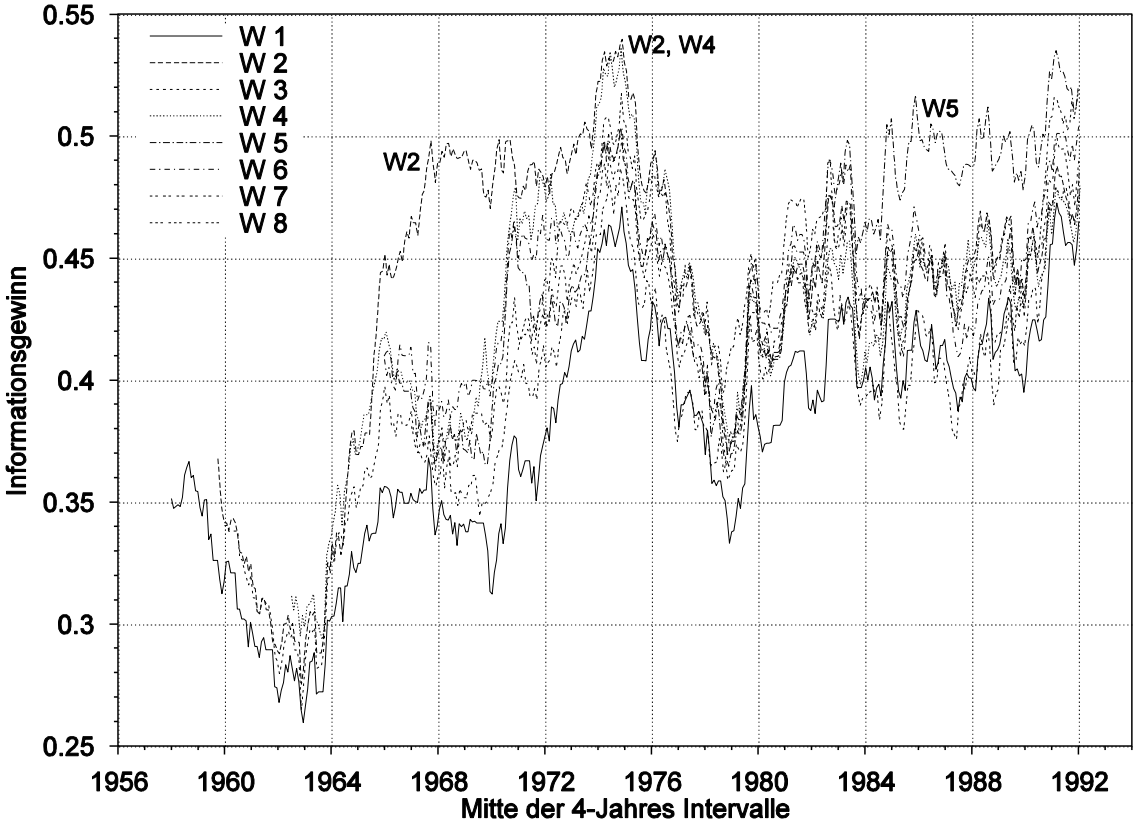

Abb. 5-9. Informationsgewinn des Abflusses von 4-Jahresintervallen für die acht Watersheds (W) von Hubbard Brook bei statischer H_G-maximaler Partitionierung. Wortlänge 4. Die Intervalle wurden jeweils um einen Monat überlappend verschoben.

rung erwies sich als wenig spezifisch und daher ungeeignet zur Feststellung von Unterschieden.

Bei statischer Partitionierung (Abb. 5-9) bleiben die Nordhanggebiete im Informationsgewinn unentdeckt. Bei H_G-maximaler Partitionierung liegen die acht Teilgebiete deutlich näher beieinander als bei Median-Partitionierung. Dies gilt insbesondere für den Niederschlag. Allerdings werden bei H_G-Maximierung die nachfolgend beschriebenen Unterschiede deutlicher sichtbar. Die Informationsdifferenz zwischen Niederschlag und Abfluss ist in gleicher Weise geeignet diese Unterschiede zu beschreiben wie die Information der Abflüsse selbst. Die folgenden Beobachtungen beziehen sich daher auf den Informationsgewinn des Abflusses bei binärer statischer H_G-maximaler Partitionierung:

Die Behandlung in W 2 äußert sich hier in einer deutlich erhöhten Information von 1965/66 bis 1970/71. Ab 1971 ist die Information gegenüber den Abflüssen der unbehandelten Gebiete noch bis mindestens 1975 sichtbar aber weniger stark als im Behandlungszeitraum und den drei Folgejahren erhöht. Dies stimmt mit den noch vorhandenen Abweichungen vom vorhergesagten Abfluss bei BORMANN & LIKENS (1979, Figure 5-6) überein. Auch die Behandlung von W 4 — zumindest die Rodung der letzten Schneisen des verbleibenden Drittels der Fläche 1974 — ist in der Abfluss-Information zu erkennen. Die Rodung von W 5 im Jahr 1984 bis auf einen Stammdurchmesser von 5 cm kann für die erhöhte Abfluss-Information des Gebietes von 1984 bis 1990 verantwortlich gemacht werden.

Nach BORMANN & LIKENS (1979, Figure 5-5) erfolgt die Wiederbesiedlung mit Pflanzen nach der 3-jährigen Herbizid-Behandlung in W 2 um etwa ein bis zwei Jahre verzögert gegenüber einer natürlichen Wiederbegrünung unmittelbar nach einem Kahlschlag. Daher ist bei W 2

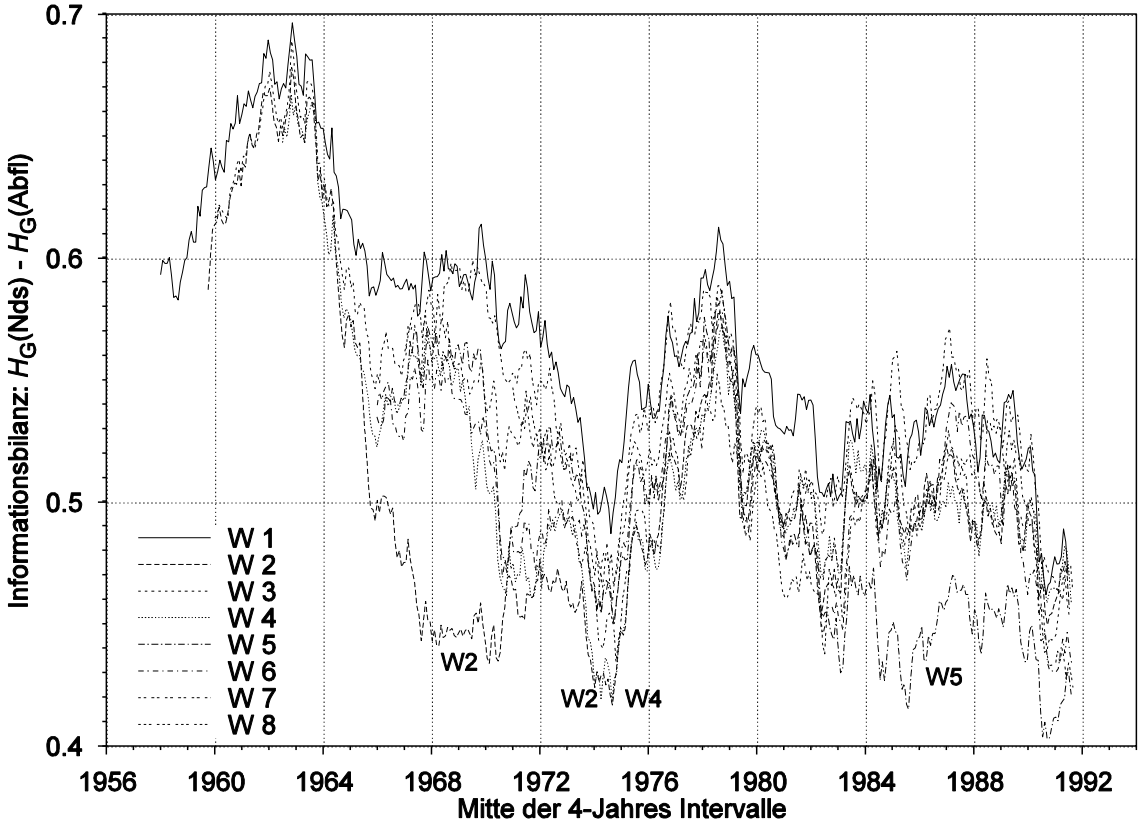

Abb. 5-10. Informationsdifferenz (H_G) zwischen Niederschlag (Nds) und Abfluss (Abfl) von 4-Jahresintervallen für die acht Watersheds (W) von Hubbard Brook bei statischer H_G-maximaler Partitionierung. Wortlänge 4. Die Intervalle wurden jeweils um einen Monat überlappend verschoben.

eine längere Störung der Abflussdynamik zu erwarten als bei W 5. In W 5 wurden sogar Bäume bis zu einem Stammdurchmesser von 5 cm nicht gefällt. Diese dürften allerdings durch die Rodung und den Abtransport der großen Bäume geschädigt worden sein. Die Maßnahme in W 5 äußert sich jedoch im gleichen zeitlichen Umfang wie die Maßnahme bei W 2.

Die Abfluss-Information erlaubt also eine Unterscheidung der Hanglage sowie ein genaues Erkennen der Durchforstungsmaßnahmen. Das Fehlen einer (ausgewachsenen) Bewaldung bewirkt eine höhere Information der Abflussmengen durch das Ausbleiben der sommerlichen Transpiration und der damit verbundenen größeren Nähe der Abflussdynamik zu der Dynamik des Niederschlags (vgl. Abb. 4.7). Interessanterweise fällt — wie schon in Abschnitt 5.2.2.1 festgestellt — die Hanglage, die bei den nördlich exponierten Gebieten eine geringere Transpiration vermuten lässt als bei Südhanglage, erst bei einer Untersuchung der Änderungen der Abflussmengen auf.

5.2.2.4 Wiederkehranalyse

In diesem Abschnitt soll kurz auf die Möglichkeiten der Wiederkehranalyse als ein weiteres Nicht-Standard-Verfahren für die saisonale Unterscheidung der Abflüsses der acht Teilgebiete in Hubbard Brook eingegangen werden. Diese Methode basiert auf der Arbeit von ECKMANN et al. (1987). Für eine skalare Zeitreihe wird eine Korrelationsmatrix erstellt, dessen Einträge $w(i, j)$ die Stärke des Zusammenhangs (Differenz) eines kurzen Teilstücks der Zeitreihe ab dem i-ten Zeitschritt mit einem Teilstück ab dem j-ten Zeitschritt messen. Die Länge dieser Teilstücke wird (Einbettungs-) Dimension genannt. In der Regel wird $w(i, j)$ nur binär gemessen: Wenn $w(i, j)$ einen Schwellenwert (Radius) überschreitet, wird der Punkt (i, j) Wiederkehrpunkt genannt, sonst nicht. Die grafische Darstellung dieser Matrix wird Wiederkehr-Diagramm (recurrence plot) genannt und ermöglicht ein sensibles Aufspüren vorübergehender Ereignisse in der Zeitreihe. Auf diese Weise waren z. B. klimatische Schwankungen und der Kahlschlag in W 2 zu erkennen (siehe LANGE, 1999, Abb. 46). Es verwundert nicht, dass die erhebliche Datenmenge, die eine solche 2-dimensionale Darstellung enthält, sensible Analysen erlaubt. Diese große Informationsmenge muss letztlich aber wieder auf einfache Aussagen reduziert werden. Eine solche Reduzierung stellt die Wiederkehr-Quantifizierungs-Analyse (RQA) dar, die bei TRULLA et al. (1996) beschrieben wird und eine Weiterentwicklung der Wiederkehr-Diagramm-Analyse ist. Sie extrahiert aus der Korrelationsmatrix einige Maße (Zahlen), welche die Matrix charakterisieren. Eines von diesen Maßen ist der Anteil des Determinismus, der den prozentualen Anteil der Wiederkehrpunkte angibt, die eine diagonale Anordnung von einer bestimmten Mindestlänge (Line) haben. Eine solche Anordnung deutet auf eine fortlaufende Abhängigkeit von Datenpunkten mit konstantem Zeitabstand hin. Zur Erstellung der Wiederkehr-Diagramme sowie der RQA-Maße wurde das RQA-Programmpaket, Version 4.1 von C. L. Webber, Jr, verwendet.

Für die acht Abflusszeitreihen von Hubbard Brook wurden von jeweils 3- oder 4-jährigen Abschnitten mit überlappender monatlicher Verschiebung die (1095×1095)- oder (1461×1461)-Korrelationsmatrizen erstellt und die RQA-Maße berechnet. Auf diese Weise wurden methodisch ähnliche Abbildungen erzeugt, wie bei der Korrelations- und Komplexitätsanalyse (s. o). Der Anteil der Wiederkehrpunkte war bei den Nordhanggebieten und bei verschiedenen Parametern der Methode grundsätzlich höher als bei den Südhanggebieten. Der Anteil des Determinismus erwies sich als das wichtigste RQA-Maß. Es war stabil gegenüber Parameter-Variationen der Methode, sensitiv auf die Hanglage und auf die Durchforstungen in W 2 und W 5. Abb. 5-11 zeigt, dass mit der Wiederkehranalyse keine neueren Erkenntnisse

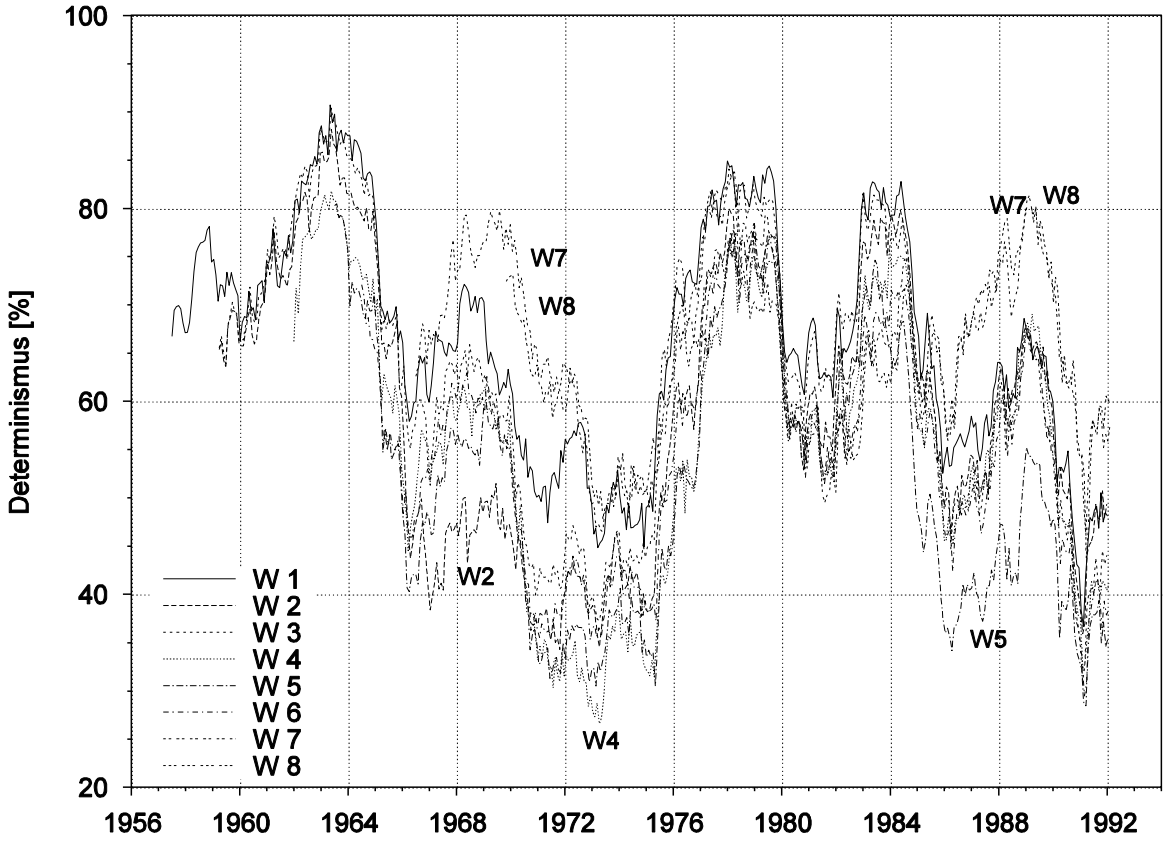

Abb. 5-11. Anteil des Determinismus in Wiederkehr-Diagrammen über jeweils drei Jahre für die acht Watersheds (W) von Hubbard Brook. Die Intervalle wurden jeweils über einen Monat überlappend verschoben. Dimension 5, Radius 10, Line 10 (Erläuterung im Text).

zur Unterscheidung der Teilgebiete und deren Behandlung gewonnen werden konnten als mit der Quantifizierung der Information in Abschnitt 5.2.2.3. Die Unterschiede wurden bei der Informationsanalyse sogar deutlicher herausgearbeitet als hier.

5.3 Optimale Messauflösung und effektive Zeitskala

Bei langfristigen Monitoringprogrammen werden Größen oft täglich, stündlich oder in noch höherer zeitlicher Auflösung gemessen. Grundsätzlich gilt: Je höher aufgelöst, desto besser aber auch teurer. Nur so lassen sich auch kurzfristige extreme Ereignisse erfassen. Doch was ist die im Mittel relevante Zeitskala eines Prozesses? Wenn die Daten zu oft gemessen werden, sind die Werte hoch redundant, vorhersagbar und der zusätzliche Informationsgewinn steht in keinem Verhältnis zum erhöhten Messaufwand. Solche Daten benötigen oft unnötig viel Speicherplatz. Sind die Daten zu selten gemessen, wird ein wesentlicher Teil der Dynamik übersehen und die Messwerte gleichen möglicherweise einer Zufallsfolge. Eine Modellierung und Vorhersage der Werte ist in diesem Fall besonders schwierig.

Die Messauflösung bestimmt den Informationsgehalt und allgemein die statistische Bewertung der Daten. ROMAHN (1996) stellte eine systematische Abhängigkeit der Maxima der Metrischen Entropie von der Messauflösung für Saugspannungen des Bodens im Einzugsgebiet der Langen Bramke fest (siehe auch LANGE et al., 1997). Bei einem Vergleich von Zeitreihen ist auf eine einheitliche Auflösung zu achten. Diese kann auch künstlich durch Aggregierung, Mittelung oder Dezimierung hochaufgelöster Daten erreicht werden. Dabei

sollte die gemeinsame vergröberte Auflösung einem ganzzahligen Vielfachen der jeweils ursprünglichen Auflösung entsprechen. In Abschnitt 5.1 wurden beispielsweise 10-minütliche Daten auf tägliche Werte aggregiert und stündliche Daten entsprechend ausgedünnt.

In diesem Abschnitt soll anhand von hoch aufgelösten Messdaten festgestellt werden, was die relevante Zeitskala von hydrologischen Zeitreihen in Wassereinzugsgebieten ist. Dazu wird die Komplexität der Daten auf verschiedenen Aggregationsniveaus (Vergröberungsstufen) betrachtet. Das Kriterium für die relevante Zeitskala der Beobachtungsgröße ist ein Maximum an Komplexität, da eine geringere Komplexität bei höherer Information ein Indiz für zunehmende Zufälligkeit ist und geringere Komplexität bei niedrigerer Information ein Indiz für zunehmende Redundanz ist.

Besonders anschaulich wird dieses Kriterium am Beispiel der Fluktuationskomplexität (siehe 2.6.2): Bei sehr redundanten Daten sind sowohl der lokale Informationsgewinn wie auch der Informationsverlust im Mittel gering. Bei relativ zufälligen Daten sind lokaler Informationsgewinn und -verlust im Mittel hoch. In beiden Fällen sind die Differenzen aus lokalem Informationsgewinn und -verlust sowie deren Schwankungen und damit die Fluktuationskomplexität niedrig. Erst bei einer im Mittel maximal unausgewogenen lokalen Informationsbilanz ist die Fluktuationskomplexität maximal. Dann können die Daten als interessant angesehen werden.

5.3.1 Vorgehensweise am Beispiel der Langen Bramke

5.3.1.1 Stündlicher Gebietsabfluss

5.3.1.1.1 Statische äquiquantile Partitionierung

Zunächst wurden nur die stündlichen Abflussmessungen in der Langen Bramke von 1986 – 1995 ($N = 87645$ Werte) untersucht, um die geeignete Vorgehensweise festzustellen. Dazu wurden bei sukkzessive höheren Aggregationsniveaus von einer Stunde, 2 Stunden, 3, 4, 5, usw. verschiedene Maße mit verschiedenen Parametern berechnet. Bei der Aggregation einer bestimmten Stufe s werden jeweils s aufeinanderfolgende Datenpunkte aufsummiert und zu einem neuen Datenpunkt zusammengefasst. Dadurch verkleinert sich die ursprüngliche Datenmenge N auf nur noch N/s Werte. Dies macht die Analyse bei höherem s durch hohe Schwankungen der statistischen Maße immer schwieriger und schließlich unmöglich. Die Grenzen der Analysen ergeben sich aus der bei der jeweils verfügbaren Datenmenge maximal möglichen Wortlänge, die für die Fluktuationskomplexität, Rényi-Komplexität und Metrische Entropie (bei äquiquantiler Partitionierung) mindestens $L = 2$ betragen muss (siehe 3.6). Für die ersten Untersuchungen wurden statische äquiquantile Partitionierungen (siehe 2.1.1.1) verwendet.

Zunächst war zu klären, ob bei jeder Aggregationsstufe die jeweils maximal mögliche Wortlänge (5 % mittlere Genauigkeit) verwendet werden sollte oder ob einheitlich für alle Aggregationsniveaus das Minimum $L = 2$ ausreichend ist. Dabei stellte sich heraus, dass die Wechsel der Wortlängen, also dessen Reduzierung ab jeweils einem bestimmten s, zu Sprüngen in den Maßen führen, welche die Identifizierung der Komplexitätsmaxima behindern. Die Sprünge waren bei der Metrischen Entropie besonders hoch und nahmen mit der Alphabet-

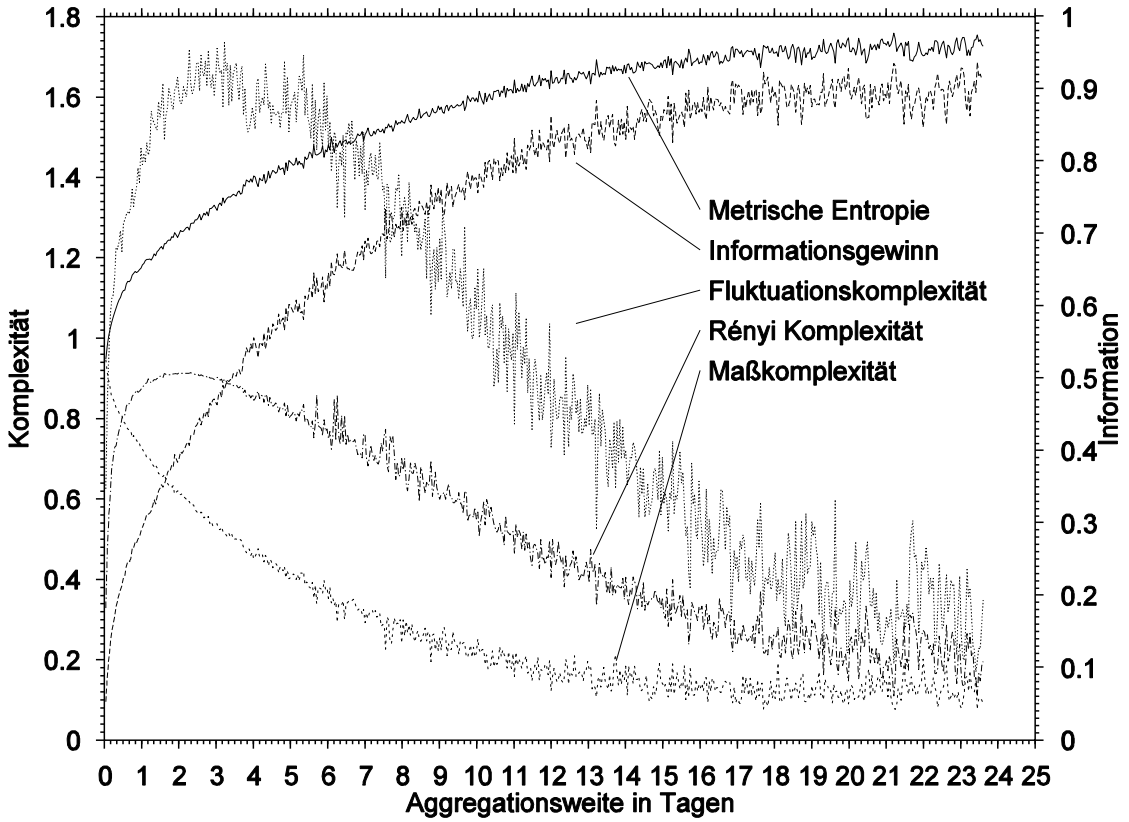

Abb. 5-12. Aggregation der stündlichen Abflussmessungen der Langen Bramke 1986 – 1995 und Berechnung von Komplexitätsmaßen bei binärer statischer H_μ-maximaler Partitionierung. Wortlänge 2.

größe zu. Im Folgenden wurde daher nur eine konstante minimale Wortlänge von $L = 2$ verwendet.

Es wurden Alphabete der Größen $\lambda = 2$, 3 und 4 verwendet. Im Fall von $\lambda = 4$ konnte die Fluktuationskomplexität C_Γ jedoch nur bis zur Aggregationsstufe $s = 59$ h berechnet werden, um für $L = 2$ noch 5 % Genauigkeit im Mittel zu garantieren. Da das Maximum von C_Γ bei $\lambda = 2$ und $\lambda = 3$ im Bereich von zwei bis drei Tagen lag (s. u.), war ein Vergleich mit $\lambda = 4$ nicht möglich.

Der Vergleich der Methoden untereinander zeigt, dass alle Informationsmaße im Mittel, d. h. bis auf ein Rauschen, mit dem Aggregationsniveau ansteigen und ab etwa 20 Tagen Aggregation nur noch um einen konstanten hohen Wert schwanken (siehe Abb. 5-12). Das Rauschen ist bei der Metrischen Entropie am geringsten und bei der Algorithmischen Information am stärksten. Vor allem bei niedrigen Aggregationsniveaus, also bei noch hinreichender Datenmenge, ist der Verlauf des Informationsgewinns (im Gegensatz zur Metrischen Entropie) dem der Algorithmischen Information sehr ähnlich. Wegen der Nähe dieser Maße zur Entropie der Quelle (siehe 2.5.4 und 2.5.6) kann der Informationsgewinn als Repräsentant der Informationsmaße mit moderatem Rauschniveau gelten.

Die Effektive Maßkomplexität zeigt einen in etwa an einer Horizontalen gespiegelten Verlauf zu den Informationsmaßen, der auch schon in den Abschnitten 2.6.1 und 5.1 beobachtet wurde. Dies bedeutet keinen Erkenntnisgewinn, weshalb auf das Maß verzichtet werden kann. Der Verlauf von Fluktuationskomplexität C_Γ und Rényi-Komplexität C_R ist qualitativ ähnlich. C_Γ ist allerdings deutlich stärker verrauscht als C_R. Beide erreichen ein globales Maximum

zwischen zwei bis drei Tagen Aggregation: $C_{\Gamma,\max}(\lambda=2) = 58$ h, $C_{\Gamma,\max}(\lambda=3) = 60$ h[8], $C_{R,\max}(\lambda=2) = 55$ h, $C_{R,\max}(\lambda=3) = 60$ h. Die Bestimmung der Maxima wird durch das Rauschen erschwert. Die Konsistenz der Ergebnisse spricht für eine optimale Messauflösung des Abflusses der Langen Bramke von etwa 55 bis 60 Stunden — also 2½ Tagen.

5.3.1.1.2 Statische Entropie-maximale Partitionierung

Zur Überprüfung des bisherigen Ergebnisses wurden die Rechnungen für binäre und ternäre Entropie-maximale Partitionierungen mit fester Wortlänge $L = 2$ wiederholt. Die Maße wurden bei den Partitionierungsparametern π_0 (und π_1) ausgewertet, bei denen die Metrische Entropie, H_μ, oder der Informationsgewinn, H_G, maximal war. Die Rechenzeit-intensive Suche nach den optimalen π_i wurde aufgrund der bisherigen Beobachtungen (siehe 3.8.1) auf ein Teilintervall von jeweils 10 % der Länge des Wertebereichs um den Median eingeschränkt. Darin wurden 200 äquidistante Werte für π_i angenommen.

An dem Verlauf der Informationsmaße und der Effektiven Maßkomplexität änderte sich bei der Entropie-maximalen im Vergleich zur äquiquantilen Partitionierung qualitativ kaum etwas: Die Informationswerte lagen erwartungsgemäß etwas höher und das Rauschniveau des Informationsgewinns bei H_G-Maximierung nahm auf die Größenordnung der Metrischen Entropie ab, welche sich kaum änderte. Das Rauschniveau von Fluktuations- und Rényi-Komplexität änderte sich nicht. Das Maximum der Fluktuationskomplexität wurde schmaler, was seine Bestimmung erleichterte.

Bei H_G-Maximierung springt die Rényi-Komplexität bei $\lambda = 2$ und $s \approx 5$ Tagen auf ein höheres Niveau, welches auch zunächst das vorherige Maximum übertrifft. Bei $\lambda = 3$ ist dieses Verhalten bei C_Γ und C_R für $s \approx 6$ Tage zu beobachten. Das bisherige Maximum wird dann aber nicht übertroffen. Diese Sprünge (bei $\lambda = 2$ und $\lambda = 3$) liegen also jenseits der „eigentlichen“ globalen Maxima und beeinflussen deren Interpretation nicht. Die globalen Maxima sind: $C_{\Gamma,\max}(\lambda=2) = 82$ h, $C_{\Gamma,\max}(\lambda=3) = 51$ h, $C_{R,\max}(\lambda=2) = 59$ h, $C_{R,\max}(\lambda=3) = 26$ h. Bei $\lambda = 3$ ist ein Sprung nach oben für C_Γ und nach unten für C_R beobachtbar, der bedeuten würde, dass eher $C_{R,\max}(\lambda=3) = 64$ h anzunehmen ist. Dies würde eine bessere Übereinstimmung mit den bisherigen Werten bedeuten.

Bei H_μ-Maximierung wurden ebenfalls $C_{\Gamma,\max}(\lambda=2) = 82$ h und $C_{R,\max}(\lambda=2) = 59$ h als globale Maxima festgestellt. Einen Sprung in der Rényi-Komplexität, wie bei der H_G-Maximierung, gab es nicht. Auf H_μ-maximale Partitionierungen für $\lambda = 3$ wurde verzichtet. Eine Verdopplung der Anzahl der Partitionierungsparameter π_0 in dem 10-%-Intervall zur Bestimmung des Informationsmaximums für $\lambda = 2$ ändert an den beschriebenen Beobachtungen nichts.

5.3.1.1.3 Dynamische Partitionierungen

Zusätzlich zu den statischen Partitionierungen, d. h. Partitionierungen des Wertebereiches der Daten, wurden auch dynamische Partitionierungen, d. h. Partitionierungen der Steigung (1. Differenz, siehe 2.1.1.2), untersucht. Dazu wurden bei jeweils fester Wortlänge $L = 2$ eine binäre 0-Partitionierung (Trennung von Zu- und Abnahme der Werte), eine binäre Metrische-Entropie-maximale-Partitionierung und eine ternäre äquiquantile Partitionierung verwendet. Der Partitionierungsparameter beim Entropie-Maximum, wie auch der Median, lag jeweils

[8] Das tatsächliche globale Maximum liegt hier bei einer Auflösung von 104 Stunden. Der mittlere Kurvenverlauf identifiziert dies jedoch als extreme Spitze im Rauschen auf einer im Mittel bereits abfallenden Kurve.

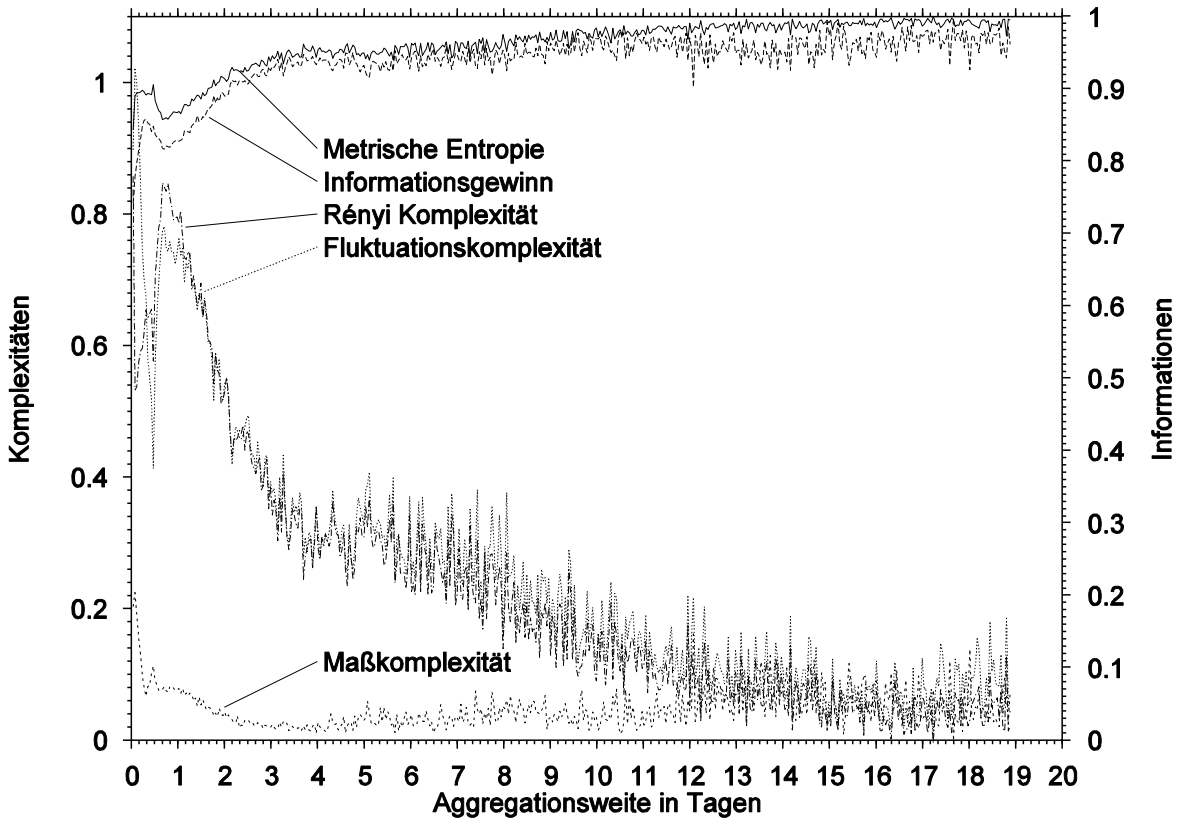

Abb. 5-13. Aggregation der stündlichen Abflussmessungen der Langen Bramke 1986 – 1995 und Berechnung von Komplexitätsmaßen bei binärer dynamischer 0-Partitionierung. Wortlänge 2.

etwas unterhalb von 0 mm/d, was durch den schnellen Anstieg und langsamen Rückgang des Abflusses über mehrere Tage nach Niederschlagsereignissen zu erklären ist.

Bei allen drei Partitionierungstypen ergab sich ein konsitentes Bild (siehe Abb. 5-13): Die Information lag bereits bei den Stundendaten auf hohem Niveau, stieg bis zu einer Aggregation von drei Tagen nur noch wenig an und blieb dann — bis auf ein Rauschen — konstant maximal. Auch hier war die Effektive Maßkomplexität nur ein Spiegel der Informationsmaße. Die Komplexitäten gingen von einem globalen Maximum aus, d. h. bei einer Stunde (C_R) oder zwei Stunden (C_Γ), und erreichten ein auffälliges lokales Maximum bei 17 bis 19 Stunden (C_R) oder 18 bis 24 Stunden (C_Γ). Bei noch höheren Aggregationen nehmen die Komplexitäten ab. Das lokale Minimum zwischen den genannten Maxima liegt bei 12 Stunden. Dieser Wert wird von den drei Partitionierungstypen sowohl von der Fluktuations- als auch der Rényi-Komplexität bestätigt. Lediglich bei der binären 0-Partitionierung gibt es einen Ausreißer mit zwei Stunden für C_R und bei der binären Entropie-maximalen Partitionierung mit 10 Stunden bei C_Γ.

Die Interpretation dieser Ergebnisse ist schwierig. Die Differenzendaten haben bereits einen hohen Informationsgehalt, so dass von Redundanz keine Rede sein kann. Daher kann eine Aggregation dieser Daten nicht mehr zu einer effektiven Zeitskala führen. Durch die Differenzenbildung wird das hochfrequente Messrauschen stärker erfasst als bei statischer Partitionierung. Eine geringe Aggregierung der Differenzen könnte dieses Rauschen kompensieren und so den Einbruch in den Komplexitätskurven verursachen. Falls dies der Fall ist, kann das lokale Minimum der Komplexität als Indiz für eine minimale Messrate angesehen werden, bei der das Messrauschen der stündlichen Messung kompensiert wird. Diese minimale Messrate läge dann für den Abfluss der Langen Bramke bei 12 Stunden.

5.3.1.1.4 Die effektive Zeitskala für den Abfluss der Langen Bramke

Bei unterschiedlichem Partitionierungstyp (statisch äquiquantil und Entropie-maximal), verschiedener Alphabetgröße ($\lambda = 2, 3$) und verschiedenen Komplexitätsmaßen (C_Γ und C_R) kann einheitlich ein Komplexitätsmaximum bei Aggregierung der Daten auf zwei bis drei Tage festgestellt werden. Ein Komplexitätsmaximum bedeutet, dass die Daten weder zu redundant (informationsarm) noch zu unkorreliert (zufällig, informationsreich) sind. Abb. 5-14 verdeutlicht, wie die Daten mit zunehmender Aggregation informationsreicher werden, wie sich die Information zunächst zunehmend komplexer darstellt und schließlich durch eine höhere Zufälligkeit an strukturellem Gehalt verliert und wieder einfacher organisiert ist. Daraus kann eine effektive Zeitskala von zwei Tagen für den Abfluss der Langen Bramke gefolgert werden.

Ein Zeitabstand von zwei bis drei Tagen liegt in der Größenordnung des Abstandes von der durch das oberflächennahe Abfließen verursachten Vorwelle im Bachabfluss mit der vom Tiefenwasser gespeisten Hauptwelle (DUNNE & BLACK, 1970). Bei einer Zeitauflösung der Daten in diesem Bereich ist nur noch die Hauptwelle als das charakteristische Signal des Abflusses zu erkennen.

Bei der direkten Interpretation der Abflussganglinien typischer Hochwasserereignisse aus der Bramke wird zwischen einer Vorwelle, die in unmittelbarem Zusammenhang mit dem Niederschlag auftritt, und einer verzögerten Hauptwelle unterschieden. Die Hauptwelle tritt erst bei Ereignissen von mehr als 16-25 mm Niederschlag auf und erreicht ein Maximum nach 30-50 Stunden (HAUHS, 1985). In den bisherigen Anwendungen prozess-orientierter Modelle wurde

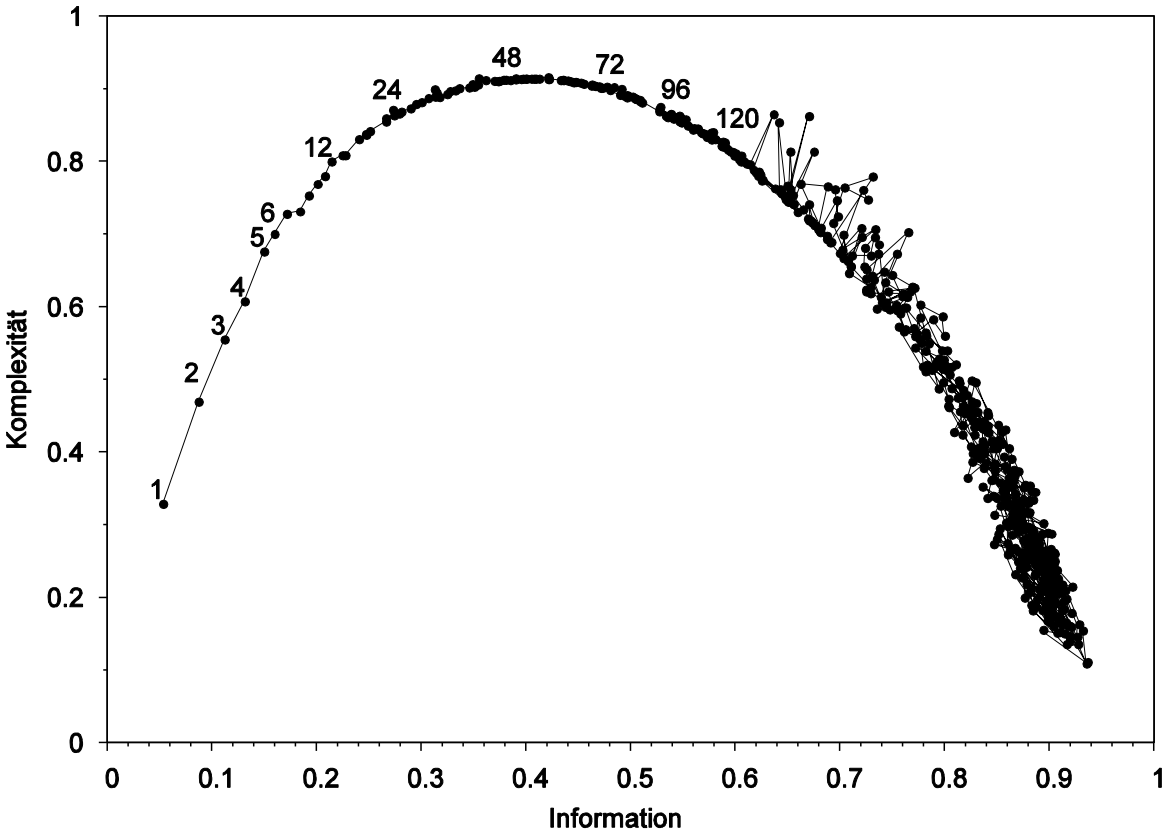

Abb. 5-14. Information (H_μ) und Komplexität (C_R) bei Vergröberung der Zeitauflösung für den stündlichen Abfluss der Langen Bramke 1986 – 1995. Verwendung der Metrischen Entropie und Rényi-Komplexität aus Abb. 5-12. Die Zahlen entsprechen den Aggregationsstufen in Stunden.

ausschliesslich der Verlauf dieser Hauptwellen rekonstruiert (SCHMIDT, 1997). Die aus den Komplexitätsmaßen abgeleitete optimale Messauflösung liegt auffällig nah auf der Zeitskala der Reaktionszeit bis zur Hauptwelle. Wie bei der Analyse der Messauflösung von Niederschlagsereignissen (siehe 5.3.1.2) wird hier also eine Auflösung als optimal identifiziert, die auch im Rahmen der praktischen Hydrologie heuristisch ermittelt wurde.

Wie hängt dieser Wert mit der Korrelationslänge zusammen? Wegen der Saisonalität (Jahresgang, siehe 3.3) ist die Korrelationslänge praktisch unendlich. Für die Originaldaten sinkt die Autokorrelationsfunktion nach 2511 Stunden — also etwa 3½ Monaten — zum ersten Mal unter das 5 %-Signifikanzniveau. Für die Jahresgang-bereinigten Daten nach der Differenzenmethode (siehe 3.3) ist dies nach 516 Stunden (21½ Tagen) der Fall. Eine saisonunabhängige Autokorrelationslänge von 21 Tagen erscheint plausibel, weil dann in der Autokorrelationsfunktion der Originaldaten (siehe Abb. 5-17) die Spitzenhöhe erreicht ist, die für die Jahresvielfachen charakteristisch ist. Der Anstieg der Information durch zunehmende Unkorreliertheit der aggregierten Daten erfolgt nur bis zu dieser Autokorrelationslänge. Bei Bereinigung des Jahresganges erreichen die Informationsmaße nach dieser Zeit sogar ihr theoretisches Maximum und die Komplexitätsmaxima ändern sich nicht, wie eine Vergleichsrechnung zeigt. Eine Zeitauflösung von zwei bis drei Tagen ist insgesamt unabhängig von der Autokorrelationslänge zu sehen, die mit 21 Tagen deutlich länger ist.

Die Konsistenz der Ergebnisse legt eine abgekürzte Vorgehensweise zur Bestimmung der effektiven Zeitskala nahe. Eine Beschränkung auf binäre Partitionierungen bei fester Wortlänge 2 scheint ausreichend zu sein. Die Bestimmung Entropie-maximaler Partitionierungen bei höheren Alphabeten ist besonders rechenaufwendig. Die gelegentlichen Sprünge und das Rauschen in den Werten legen jedoch die vergleichende Betrachtung von Entropie-maximalen Partitionierungen (H_μ, H_G) und Median-Partitionierungen nahe.

Die Komplexitätsmaxima sollten anhand der Fluktuations- und Rényi-Komplexität bestimmt werden, die in einem Programmlauf von SYMDYN zusammen bestimmt werden können. Die automatisierte Maximum-Bestimmung war oft unbrauchbar. Alle Maxima wurden daher grafisch visuell überprüft und bestimmt. Ein Algorithmus, der (i) lokale Fluktuationen nach unten und oben ausgleicht, (ii) einseitige Spitzen abschneidet und (iii) die Relevanz von Sprüngen im globalen Kurvenverlauf beurteilt stand nicht zur Verfügung und wurde auch nicht entwickelt.

5.3.1.2 Stündlicher Gebietsniederschlag

Die stündlichen Gebietsniederschlags-Messungen der Langen Bramke von 1983 – 1992 wurden ebenfalls künstlich aggregiert und einer Komplexitätsanalyse unterzogen. Dabei wurden binäre und ternäre, äquiquantile und Entropie-maximale (H_G u. H_μ) Partitionierungen bei fester Wortlänge $L = 2$ verwendet. Abb. 5-15 zeigt die Komplexitätsmaße für einen der sechs Partitionierungstypen mit besonders niedrigem Rauschniveau im interessanten Aggregations-bereich. Die Informationsmaße zeigten jeweils einen schnellen Anstieg auf einen fast maximalen Wert schon bei drei Tagen Aggregation. Ab etwa sieben Tagen ist kein Anstieg der Informationsmaße mehr erkennbar. Die Effektive Maßkomplexität brachte — wie beim Abfluss — keinen Erkenntnisgewinn. Die Fluktuations- und Rényi-Komplexität waren durch ein globales Maximum bereits bei zwei Stunden charakterisiert, auf das ein Abfall über mehrere Tage folgte. Bei den ternären Entropie-maximalen Partitionierungen lag das Maximum bei drei (C_R) und fünf (C_Γ) Stunden. Die Kurven zur ternären äquiquantilen Partitionierung waren nicht interpretierbar, weil aufgrund der überwiegenden Anzahl niederschlagsfreier

Stunden zunächst zwei Partitionierungszellen unbesetzt sind und sich diese Situation erst ab einer bestimmten Aggregationsstufe ändert, was einen Sprung in den Werten verursacht, der die Bestimmung des Maximums verhindert.

Die Rechnungen wurden für dynamische Partitionierungen (binär 0 und Entropie-maximal bzgl. H_G u. H_μ sowie ternär äquiquantil) wiederholt. Die Kurvenverläufe hierbei waren von denen bei statischer Partitionierung kaum zu unterscheiden. Dies ist ein deutlicher Unterschied zu den Abflussmessungen (siehe 5.3.1.1.3) und durch das hohe Rauschniveau der Niederschlagsdaten zu erklären, das sowohl die Niederschlagsmengen selbst wie auch deren Änderungen auszeichnet.

Die Autokorrelation des Niederschlags sinkt nach 151 Stunden (6.3 Tagen) zum ersten Mal unter das 5 %-Signifikanzniveau. Ein Anstieg der Information (s. o.) erfolgt also auch beim Niederschlag nur etwa bis zur Autokorrelationslänge (vgl. 5.3.1.1.4). Durch die Aggregation der Daten lassen sich also auch mit den nur auf kurze Korrelationen sensitiven Komplexitätsmaßen längerreichweitige Zusammenhänge erfassen. Der Korrelationsradius liegt für Wortlänge 2 bei 2 (H_μ, C_R) oder 3 (H_G, H_M, C_{EM}, C_Γ) (aggregierten) Zeiteinheiten.

Eine effektive Zeitskala der Niederschlagsmenge für die Lange Bramke soll aus diesen Untersuchungen nicht gefolgert werden. Die konsistente Lage des Komplexitätsmaximums bei zwei Stunden ist zu unsicher für eine solche Aussage, weil dies die erste Aggregationsstufe ist. Um eine effektive Zeitskala der Niederschlagsmenge sicher zu ermitteln, sind höher aufgelöste Messungen erforderlich (s. u.). Ein Zeitintervall von zwei Stunden kann allerdings als obere Grenze der effektiven Zeitskala der Niederschlagsmenge angenommen werden.

Die kleinste raum-zeitliche Skala, auf der die Regenintensität definiert werden kann, ist nach

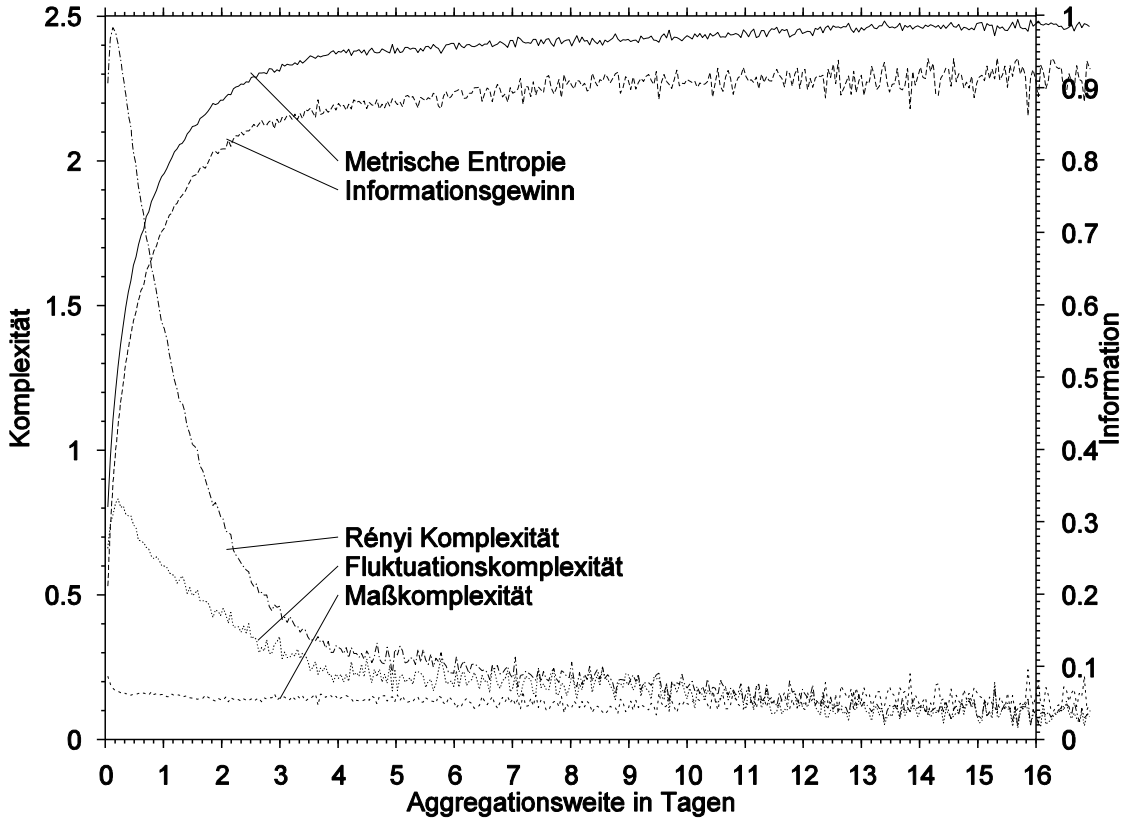

Abb. 5-15. Aggregation der stündlichen Niederschlagsmenge der Langen Bramke 1983 – 1992 und Berechnung von Komplexitätsmaßen. Ternäre statische H_μ-maximale Partitionierung. Wortlänge 2.

RODRIGUEZ-ITURBE et al. (1984) diejenige, bei der die Regenintensität als stetige Funktion in Raum und Zeit betrachtet werden kann. Sie ist nach ihrer Meinung nicht größer als eine Minute. Die zeitliche Korrelation in Regenfeldern entspricht nach Taylor's Hypothese nur bis etwa 40 Minuten der räumlichen Korrelation (WAYMIRE et al., 1984). WAYMIRE et al. (1984) stellen mit ihrem Modell eine Übereinstimmung der beiden Korrelationsfunktionen bis 14 Minuten für tropische Niederschlagsereignisse fest. Diese Zeiten können als Einflussdauer von Regenzellen auf einen festen Ort interpretiert werden und stehen somit mit der effektiven Zeitskala von Niederschlägen in Zusammenhang. Sie liegen unterhalb der stündlichen Messauflösung beim Niederschlag im Einzugsgebiet der Langen Bramke und können dafür nicht überprüft werden. Für die Gebiete „Lehstenbach" und „Steinkreuz" liegen 10-minütliche Niederschlagsmengen vor. Sie werden bezüglich der Effektiven Zeitskala in den nächsten Abschnitten untersucht.

Alleine die raum-zeitliche Struktur der Niederschläge ist bereits für die Schwierigkeiten bei der Modellierung von Zeitreihen der Regenintensität verantwortlich (siehe RODRIGUEZ-ITURBE et al., 1984; VALDES et al., 1985). Stark aggregierte — spätestens wöchentliche — Niederschlags-Zeitreihen sind nach RODRIGUEZ-ITURBE et al. (1989) statistisch nur noch als Zufallsfolgen interpretierbar. Die Informationsmaxima für den Niederschlag im Lange-Bramke-Gebiet wurden ebenfalls nach spätestens wöchentlicher Aggregation erreicht und beibehalten und lagen auf dem Niveau von Zufallsfolgen (s. o.). Nur bei hoch aufgelösten Regenintensitäten — etwa 15 s bei einem mehrstündigen Niederschlagsereignis bei RODRIGUEZ-ITURBE et al. (1989) — ist ein anderes statistisches Verhalten möglich. Der Nachweis einer chaotischen Dynamik im Unterschied zu farbigem Rauschen ist aber insgesamt schwierig, wie die Kritik von GHILARDI & ROSSO (1990) an der Arbeit von RODRIGUEZ-ITURBE et al. (1989) und die Gegendarstellung von RODRIGUEZ-ITURBE et al. (1990) zeigen.

5.3.2 Lehstenbach und Steinkreuz

5.3.2.1 Hypothese und Programm

Für den Abfluss im Lange-Bramke-Gebiet wurde eine gröbere effektive Zeitskala festgestellt als für den Niederschlag. In Abschnitt 5.1 wurde eine prinzipielle Informationszunahme des Wassers beim Durchlaufen der Einzugsgebiete „Lehstenbach" und „Steinkreuz" bei fester täglichen Auflösung beobachtet. Daher wird vermutet, dass die Aggregationsstufen der Komplexitätsmaxima, denen hier die Bedeutung einer optimalen Messauflösung oder effektiven Zeitskala unterstellt wird, mit zunehmender Bodentiefe ebenfalls zunehmen, d. h. dass die optimale Auflösung gröber wird.

Diese Vermutung soll nun anhand der Messungen des Niederschlags, der Saugspannungen und des Abflusses in den Gebieten „Lehstenbach" und „Steinkreuz" überprüft werden. Die Niederschlagsmengen liegen in beiden Gebieten sogar in 10-minütlicher Auflösung vor. Damit sind die Bedingungen für eine Bestimmung der Komplexitätsmaxima hier günstiger als bei der Langen Bramke. Die anderen Größen wurden jeweils stündlich gemessen.

5.3.2.2 Vorgehensweise

Gemäß der Folgerungen in 5.3.1.1.4 wurden die Fluktuations- und Rényi-Komplexitäten aller Zeitreihen für verschiedene Vergröberungsstufen[9] mit binären statischen äquiquantilen, H_μ- und H_G-maximalen Partitionierungen berechnet. Für die drei Partitionierungstypen, drei Tensiometer-Tiefen, fünf Tensiometer-Standorte, sowie Niederschlag und Abfluss waren 51 Programmläufe von SYMDYN für jedes der beiden Einzugsgebiete erforderlich. Dies dauerte jeweils etwa einen Tag auf einem 133 MHz Pentium mit 64 MB RAM. Die jeweils 102 Komplexitätsmaxima (C_Γ und C_R) wurden mit einem zusätzlichen Programm bestimmt und in jedem Fall grafisch visuell überprüft und gegebenenfalls korrigiert, wenn einzelne Ausreißer und Artefakte das Ergebnis verfälschten. Derartige Korrekturen waren bei der Mehrheit der Saugspannungsmessungen im Steinkreuz-Gebiet erforderlich, da die Vielzahl der verstreuten Lücken dieser Daten und die insgesamt geringere Datenmenge verglichen mit den Aufzeichnungen im Coulissenhieb (siehe 4.1 und 4.2) gerade bei den hohen Vergröberungsstufen nur noch wenige vergröberte Datenpunkte erlaubten. Von der Lage und Höhe der Komplexitätsmaxima für die jeweils fünf Messorte in einer Tiefe wurden die Mittelwerte und Standardabweichungen berechnet, um stabile Werte für diese Größen und ein Maß für ihre Schwankung zu erhalten.

5.3.2.3 Beobachtungen und Ergebnisse

Als Kriterium für die Auswahl eines Komplexitätsmaßes und Partitionierungstyps zur Bestimmung der Komplexitätsmaxima dient neben der Konsistenz der Ergebnisse über verschiedene Maße und Partitionierungen auch die Stabilität der Werte. Die Schwankungen der Komplexitätsmaxima im Steinkreuz-Gebiet waren aufgrund der geringeren Datenmenge erwartungsgemäß höher als im „Coulissenhieb“. Grundsätzlich waren die Schwankungen der Aggregationsweiten der Maxima bei äquiquantiler Partitionierung höher als bei Entropiemaximaler Partitionierung. Im Gegensatz dazu schwankte die Höhe der Maxima bei äquiquantiler Partitionierung nur wenig, besonders bei C_R. Die Mittelwerte der Rényi-Komplexitäten der Saugspannungen und des Abflusses stimmten hier und bei H_μ-maximaler Partitionierung sehr gut überein. Die Verteilung der Höhe der Maxima sowie deren Schwankungen ist jedoch nicht mit ihrer Lage und dessen Schwankung korreliert. Letztere ist hier aber die interessante Größe. Daher kann auf eine Analyse der Werte der Maxima selbst verzichtet werden.

Prinzipiell nimmt die effektive Zeitauflösung beim Durchlaufen des Wassers durch das Gebiet zu. Diese Beobachtung ist vergleichbar zu der Informationsabnahme, die in Abschnitt 5.1 beschrieben wurde. Doch nun zu den Gebieten im Einzelnen:

5.3.2.3.1 Lehstenbach

Für das Lehstenbach-Gebiet konnte einheitlich eine Verschiebung der Komplexitätsmaxima zu höheren Vergröberungsstufen beim Durchlaufen des Wassers durch das Gebiet festgestellt werden (siehe Abb. 5-16 für C_Γ bei H_G-Maximierung). Diese ist mit der in Abschnitt 5.1.1 beschriebenen Informationsabnahme vergleichbar. Das bedeutet, dass die effektive Zeitskala des Abflusses derjenigen der Saugspannung in 90 cm entspricht oder nur wenig signifikant

[9] D. h.: Aggregation der Niederschlags- und Abflussmessungen und Ausdünnung der Saugspannungsmessungen

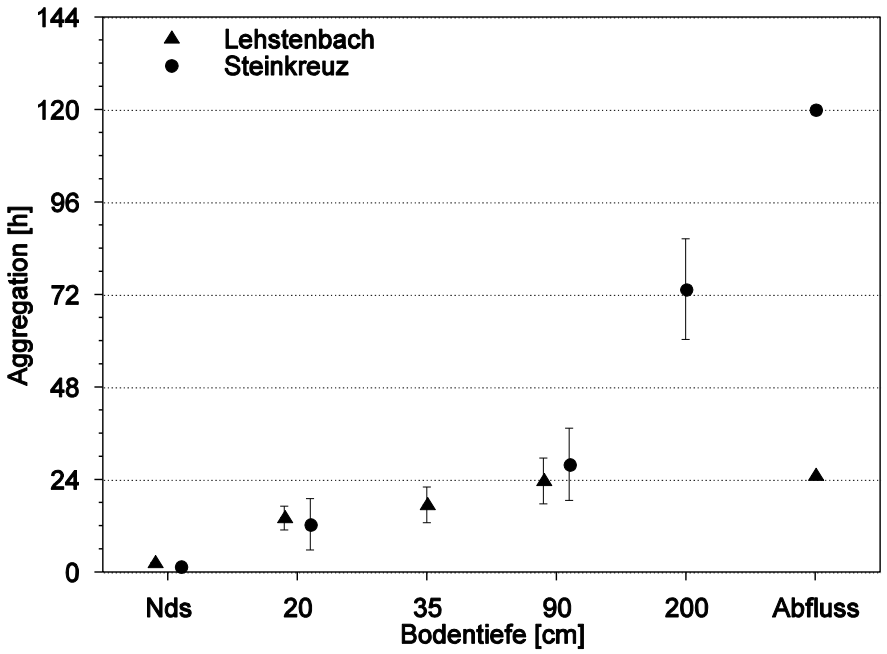

Abb. 5-16. Vergröberungsstufen der zeitlichen Auflösung beim Maximum der Fluktuationskomplexität für Niederschlag, Saugspannungen und Abfluss in den Gebieten „Lehstenbach“ und „Steinkreuz“. Für die Saugspannungen sind Mittelwerte und Standardabweichungen der Dezimierungsoptima von den jeweils fünf Standorten in jeder Tiefe eingetragen. Binäre statische H_G-maximale Partitionierung. Wortlänge 2.

darunter liegt. Die Unterschiede in den optimalen Zeitauflösungen der Saugspannungen in unterschiedlicher Tiefe sind nur wenig signifikant.

Die Aggregation des 10-minütigen Niederschlags führte bei allen Partitionierungen zu einem Maximum der Fluktuationskomplexität bei 140 Minuten und der Rényi-Komplexität bei 20 Minuten. Bei den anderen Zeitreihen lagen die Vergröberungsstufen der Maxima bei Median-Partitionierung insgesamt höher als bei Entropie-maximaler Partitionierung. Die konsistentesten Ergebnisse wurden bei H_μ-Maximierung erreicht, die deswegen zur Bestimmung der optimalen Messauflösungen herangezogen wurden. Die Median-Partitionierungen schieden wegen der hohen Standardabweichungen (s. o.) aus.

Die optimalen Auflösungen für die Saugspannungen in 20 cm Tiefe lagen bei 11 bis 14 Stunden, in 35 cm Tiefe bei 13 bis 17 Stunden und in 90 cm Tiefe bei 21 bis 24 Stunden (jeweils C_R und C_Γ). Für den Abfluss wurden bei einer Aggregation von 20 bis 25 Stunden maximale Komplexitäten erreicht. Demnach wäre für die Messung des Abflusses und der Saugspannung in 90 cm Tiefe eine tägliche Auflösung ausreichend zur Erfassung der mittleren Dynamik. Für die Messung der Saugspannungen in 20 cm und 35 cm Tiefe wäre noch eine halbtägliche Auflösung ausreichend. Der Niederschlag muss dagegen hochaufgelöst gemessen werden, mindestens 2-stündlich.

5.3.2.3.2 Steinkreuz

Auch im Steinkreuz-Gebiet nehmen die Vergröberungsstufen der Komplexitätsmaxima beim Durchlaufen des Wassers durch das Gebiet zu (siehe Abb. 5-16 für C_Γ bei H_G-Maximierung). Diese Zunahme unterscheidet sich jedoch von dem Verlauf der in Abschnitt 5.1.2 beschriebenen Informationsabnahme dadurch, dass beim Abfluss mit signifikantem Abstand die höchste optimale Auflösung erreicht wird. Die Unterschiede in den effektiven Zeitauflösungen waren hier jeweils signifikant.

Die Aggregation des in 10-minütlicher Auflösung gemessenen Niederschlags führte bei allen Partitionierungen zu einem Maximum der Fluktuationskomplexität bei 90 Minuten und der Rényi-Komplexität bei 70 Minuten. Bei Median-Partitionierung wurden wie beim Lehstenbach-Gebiet für die anderen Messreihen höhere optimale Auflösungen ermittelt als bei Entropie-maximaler Partitionierung. Wegen der höheren Standardabweichungen werden auch hier die Ergebnisse der Median-Partitionierungen nicht ausgewertet. Auch hier werden die Ergebnisse mit H_G-maximaler Partitionierung wegen der Konsistenz zur Bestimmung der effektiven Zeitskala verwendet.

Die effektive Zeitskala für die Saugspannungen in 20 cm Tiefe lagen bei 12 bis 14 Stunden, in 90 cm Tiefe bei 28 bis 30 Stunden und in 200 cm Tiefe bei 62 bis 73 Stunden. Für den Abfluss wurde eine optimale Messauflösung von 82 bis 120 Stunden ermittelt. Demnach wäre eine nur 3-tägige Messung des Abflusses im Steinkreuz-Gebiet ausreichend zur Erfassung der mittleren Dynamik. Die Messungen den Saugspannungen in 200 cm Tiefe könnten mindestens 2-täglich, in 90 cm Tiefe täglich und in 20 cm Tiefe halbtäglich erfolgen. Der Niederschlag wäre bei stündlicher Messung noch ausreichend aufgelöst. Diese Größenordnung ist mit den von WAYMIRE et al. (1984) genannten 14 bis 40 Minuten für die raum-zeitliche Korrelation in Regenfeldern vergleichbar (siehe 5.3.1.2), wenn man bedenkt, dass sich ihre Untersuchung auf tropische Niederschläge bezieht, denen eine höhere Dynamik unterstellt werden kann als den Niederschlägen im gemäßigten Klima des Steigerwaldes.

Beim Vergleich (Abb. 5-16) zwischen den optimalen Zeitauflösungen von Lehstenbach und Steinkreuz fallen vor allem die deutlich höheren Auflösungen auf, die beim Steinkreuz in 200 cm Tiefe und beim Abfluss erreicht werden. Die Stärke dieses Unterschiedes war nach dem in Abschnitt 5.1 festgestellten Informationsverlauf nicht zu vermuten. Beim Niederschlag und bei den Bodentiefen bis 90 cm unterscheiden sich die Werte nicht signifikant. Über die Ursachen der Unterschiede kann in ähnlicher Weise spekuliert werden, wie in Abschnitt 5.1.

5.4 Klassifikation von Einzugsgebieten

Bereits bei der Beschreibung der Gebiete und der Betrachtung der Originaldaten von Niederschlag und Abfluss in Kapitel 4 fallen grundsätzliche Unterschiede auf. In Abschnitt 5.1 wurde für zwei Gebiete der prinzipielle Informationsreichtum des Niederschlags gegenüber dem Abfluss festgestellt. Ein Ergebnis von Abschnitt 5.2 war, dass bei geringerer Bewaldung die Information im Abfluss eines Wassereinzugsgebietes höher, also näher an der Information des Niederschlags, ist. Diese Prinzipien werden in diesem Abschnitt anhand aller in dieser Arbeit vorliegenden Daten von Niederschlag und Abfluss überprüft. Auf diese Weise kommt

es zu einer Klassifizierung der verschiedenen Einzugsgebiete bezüglich der Information und Komplexität ihrer Niederschläge und Abflüsse.

5.4.1 Saisonalität

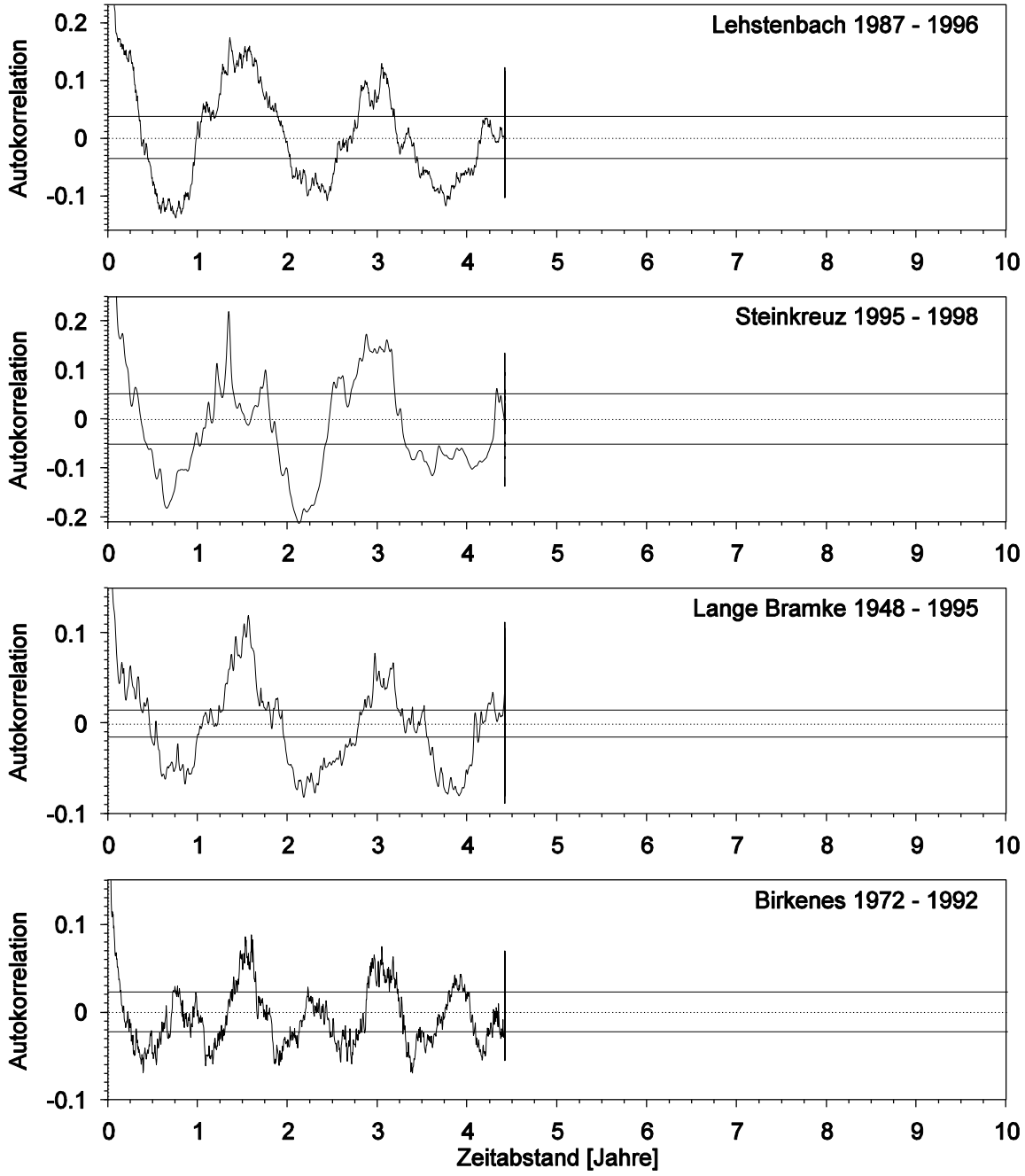

Abb. 5-17. Autokorrelationen täglicher Abflüsse verschiedener europäischer Einzugsgebiete mit 5 % Signifikanzbereich.

Die Saisonalität der Gebietsabflüsse ist ein wesentliches Merkmal von Einzugsgebieten (siehe 3.3). Die Hauptursache dafür sind die höheren Abflüsse durch die Schneeschmelze im Frühjahr und das Niedrigwasser durch die Transpiration der Pflanzen im Sommer. Diese Ereignisse können direkt an den Zeitreihen der Abflüsse in Kapitel 4 identifiziert werden. Dabei fallen Unterschiede auf, die charakteristisch für das jeweilige Einzugsgebiet sind.

In diesem Abschnitt werden zunächst die Abflüsse der verschiedenen Gebiete bezüglich ihrer Saisonalität anhand der Autokorrelationsfunktionen untersucht. Die Autokorrelation ist ein etabliertes Standardverfahren der Zeitreihenanalyse (siehe 2.2.1). Ihr informationstheoretisches Pendant ist die Transinformation (siehe 2.2.2). Zur Hervorhebung von kurzzeitigen bewuchsbedingten Unterschieden in der Abflussdynamik erwies sich die Transinformation als

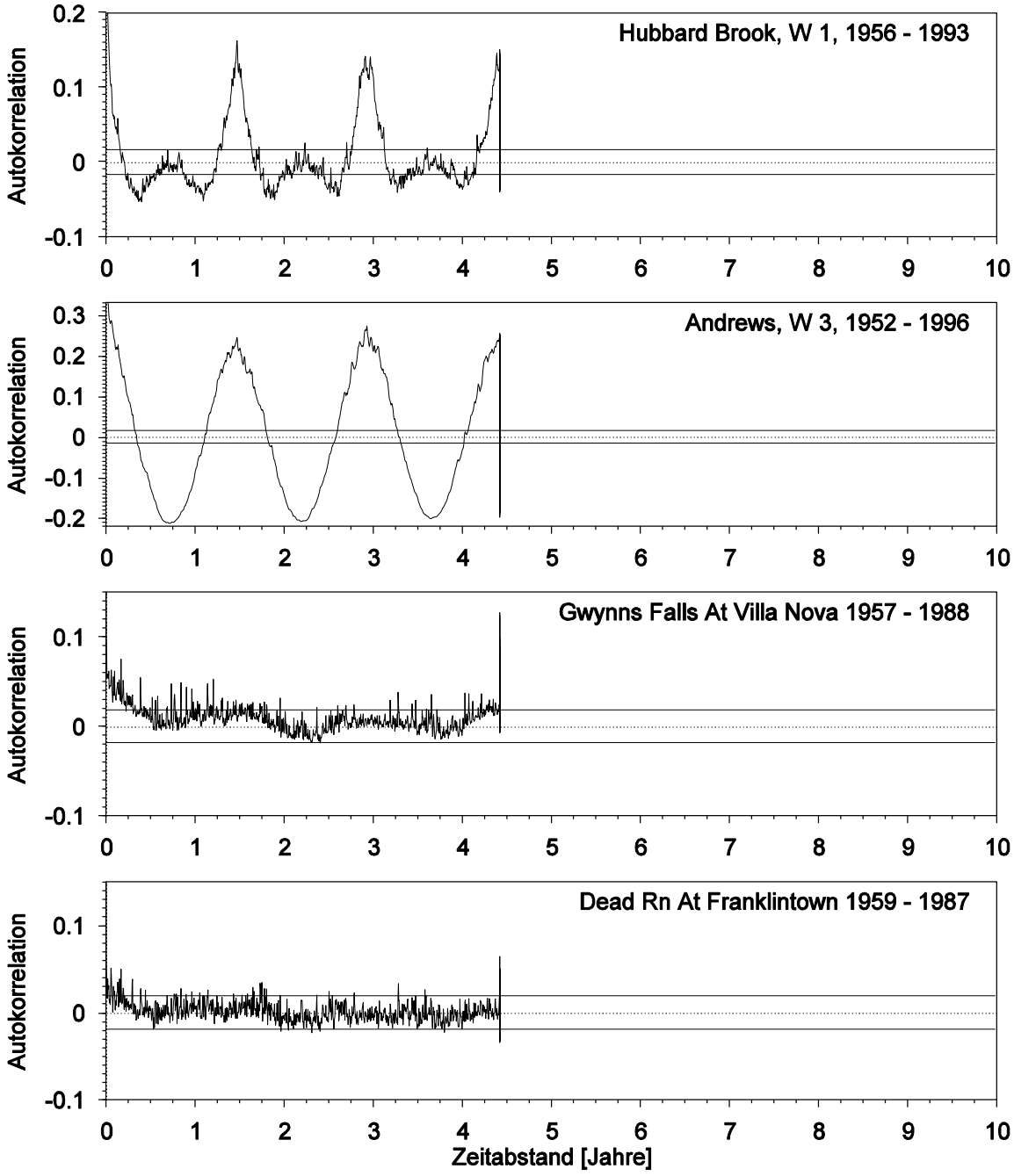

Abb. 5-18. Autokorrelationen täglicher Abflüsse verschiedener amerikanischer Einzugsgebiete mit 5 % Signifikanzbereich.

das sensiblere Maß (siehe 5.2.1). Dies könnte auf den von HERZEL & GROßE (1995) festgestellten analytischen Vorteil der Transinformation bei der Entdeckung von Korrelationen in Symbolsequenzen zurückzuführen sein. Bei dem von NEWIG (1998) und LANGE et al. (1998) im Zusammenhang mit dieser Arbeit durchgeführten Vergleich von Autokorrelation und Transinformation von Abflüssen über den vollständigen langjährigen Messzeitraum wurde kein prinzipieller Vorteil von einer der beiden Methoden festgestellt. Wegen der Differenzierung zwischen positiven und negativen Korrelationen, welche die Lesbarkeit der Korrelationskurven erhöht, wird die Korrelations- und Saisonanalyse in diesem Abschnitt mit der Autokorrelation durchgeführt.

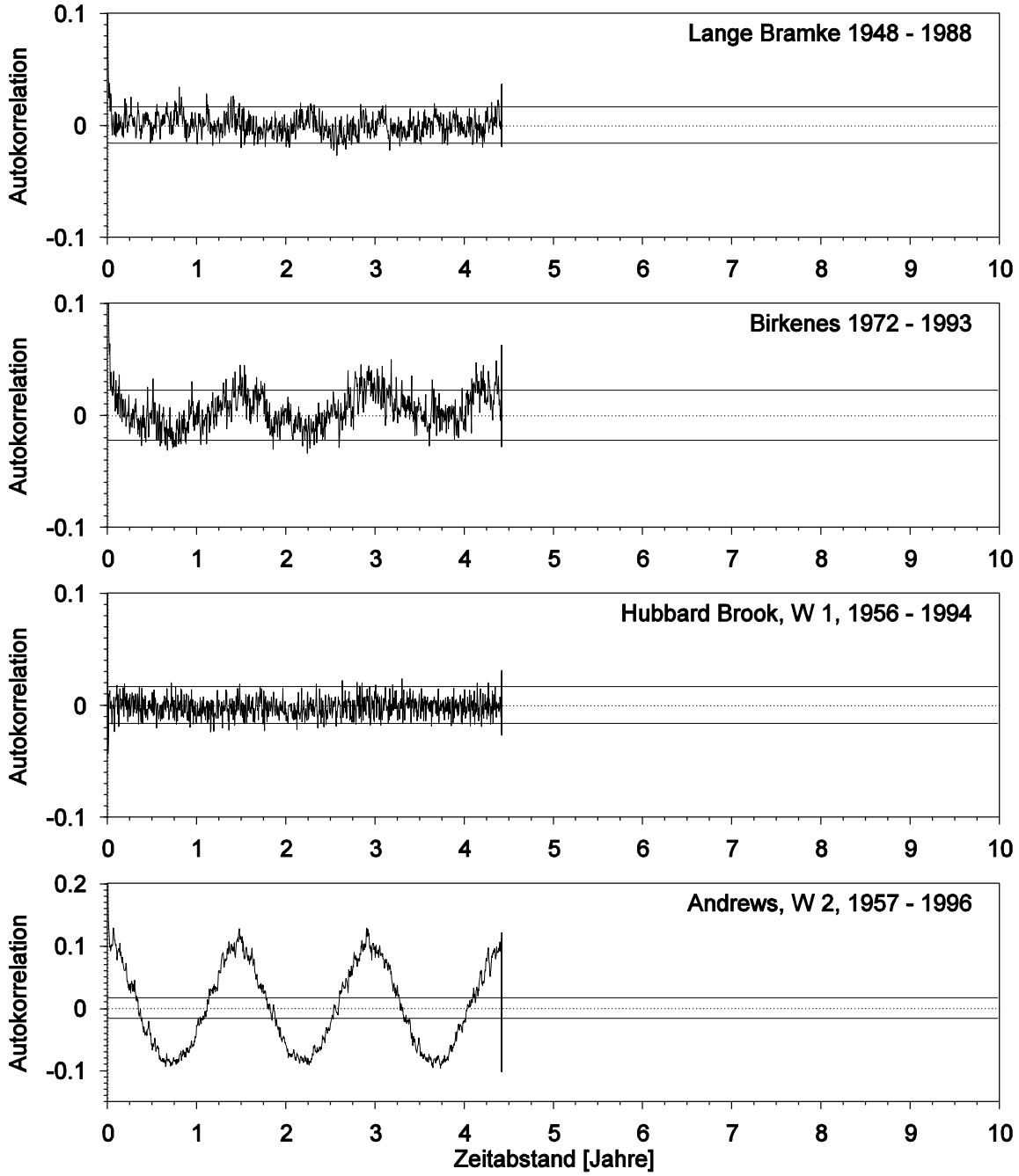

Abb. 5-19. Autokorrelationen täglicher Niederschlagsmengen mit 5 % Signifikanzbereich. Die Autokorrelationen der Niederschläge von „Lehstenbach“ und „Steinkreuz“ sind denen von „Lange Bramke“ sehr ähnlich.

In Abb. 5-17 und Abb. 5-18 ist ein signifikanter Jahresgang im Abfluss von allen naturnahen bewaldeten Wassereinzugsgebieten erkennbar. Die Autokorrelationen der drei deutschen Gebiete, Lehstenbach, Steinkreuz und Lange Bramke, lassen außer einem Rauschen darüber hinaus kein anderes Muster erkennen. Sie sind abgesehen von den durch die Datenmenge bedingten Unterschieden aus Sicht der Autokorrelation nicht zu unterscheiden. Dies gilt auch für die Niederschläge (siehe Abb. 5-19), in denen allerdings kein Jahresgang zu erkennen ist.

Auch der Abfluss in Andrews (siehe Abb. 5-18) zeichnet sich durch einen signifikanten Jahresgang aus[10]. Allerdings fällt hier der nahezu Sinus-förmige glatte Kurvenverlauf mit hoher, kaum variierender Amplitude auf. Ein solcher, nicht ganz so glatter, Sinus-förmiger Verlauf ist auch beim Niederschlag im Andrews-Gebiet zu beobachten (siehe Abb. 5-19). Damit unterscheidet sich der Niederschlag in diesem Gebiet grundlegend von Niederschlägen in den anderen untersuchten Gebieten. Der besonders strenge Jahresrhythmus der Niederschläge und damit auch der Abflüsse im Andrews-Gebiet ist in den Niederschlags- und Abflussmengen in Abb. 4-8 direkt zu erkennen. Die Sommer sind dort beinahe niederschlagsfrei und von Niedrigwasser geprägt.

Die Autokorrelationen der Abflüsse in Birkenes und Hubbard Brook unterscheiden sich von den bisher beschriebenen durch einen zusätzlichen Halbjahresgang. Dieser erreicht bei Birkenes fast die Amplitudenhöhe des Jahresganges, während er bei Hubbard Brook nicht signifikant ist[10]. Dies kann bei Birkenes durch zwei Niedrigwasser-Perioden erklärt werden, die in Abb. 4-6 sichtbar sind: Im Sommer hält Niederschlagsarmut (siehe Abb. 4-6) und die Transpiration der Vegetation den Wasserstand niedrig. Im Winter verhindert die Schneedecke einen Nachschub für das Abflusswasser, weil die flachgründigen Böden nur wenig Wasser speichern können und das vor Bildung der Schneedecke enthaltene Wasser schnell abgeben. Die sommerliche Niederschlagsarmut wird von der Autokorrelation in Abb. 5-19 als schwach signifikanter Jahresgang identifiziert.

Ähnlich verhält es sich im Hubbard Brook-Gebiet. In Abb. 4-7 ist Niedrigwasser im Sommer und — etwas höher — im Winter zu erkennen. Das Hochwasser zwischen der Vegetationsperiode und der Schneedecke ist niedriger und von kürzerer Dauer als das Hochwasser bei Schneeschmelze. Durch die tiefgründigeren Böden wird offensichtlich ein bestimmter Mindestwasserstand im Winter, der — im Unterschied zu Birkenes — stets erkennbar höher ist als zur Vegetationszeit, nicht unterschritten. Ein weiterer Unterschied zu Birkenes ist das Fehlen eines Jahresganges im Niederschlag. Der Niederschlag ist in Hubbard Brook gleichmäßig über das Jahr verteilt (siehe 4.5) und in der Autokorrelation in Abb. 5-19 nur als Rauschprozess identifizierbar. Dies führt insgesamt zu einem — im Vergleich zum Jahresgang durch die Schneeschmelze — schwachen Halbjahresrhythmus in der Abflussdynamik der Einzugsgebiete von Hubbard Brook.

Die Autokorrelationsfunktionen der Abflüsse in den urbanen Einzugsgebieten (Abb. 5-18) unterscheiden sich grundlegend von denen der naturnahen Einzugsgebiete. Bis auf eine höchstens für drei Monate signifikante Anfangskorrelation sind diese der Autokorrelation eines Rauschprozesses sehr ähnlich. Nur bei „Villa Nova“, dem größten der drei hier untersuchten urbanen Gebiete, fällt ein kaum signifikanter Jahresgang auf. Eine Vergleichsrechnung mit zufällig angeordneten Abflussmengen (entkorrelierte Daten) führt zu nur wenig veringerten Amplituden, aber einem Verschwinden der Anfangskorrelationen und einem stabilen lokalen Mittelwert der Autokorrelationen von Null. Schon bei einem direkten Vergleich der Abflussmengen in Abb. 4-9 werden Unterschiede der urbanen Gebiete zu den naturnahen Gebieten deutlich: Der Wasserstand der Bäche in den urbanen Gebieten ist im

[10] Dies und das Folgende gilt auch für die anderen Teilgebiete, die nicht in Abb. 5-18 dargestellt sind.

Mittel gleichmäßiger. Bei den Gebieten „Owings Mills“ und „Villa Nova“ ist dieser Verlauf nahezu Sinus-förmig mit etwa dem 3-fachem Abfluss im Winter wie im Sommer. Dieser Verlauf wird von einigen zufällig angeordneten kurzzeitigen Hochwasserspitzen, die erheblich höher als der Basisabfluss sind, unterbrochen. Die Spitzen treten mit einer so kurzen Dauer und relativen Höhe bei den bewaldeten, unbebauten Gebieten nicht auf. Der Abfluss des mit etwa 90 % am stärksten bebauten Einzugsgebietes „Franklintown“ ist fast ausschließlich durch diese Hochwasserspitzen geprägt und führt sonst nur sehr wenig Wasser. Seine Autokorrelation ist der eines Rauschprozesses am nächsten.

Die Saisonalität erlaubt eine Klassifikation der Abflüsse bezüglich der Kategorien

1. keine oder kaum ausgeprägte Saisonalität (urbane Gebiete),
2. Jahresgang (deutsche Gebiete und Andrews) und
3. Jahres- und Halbjahresgang (Birkenes und Hubbard Brook).

Eine feinere Unterteilung ist bei der Bewertung einer ganzen Funktion (Autokorrelation) schwierig. Auch der Skalenexponent für den Amplitudenabfall im Amplituden-Spektrum der Autokorrelationsfunktion ist wenig gebietsspezifisch, sondern eher charakteristisch für Abflusszeitreihen im Allgemeinen, wie PANDEY et al. (1998) für 19 sehr unterschiedliche Einzugsgebiete in den USA gezeigt haben. Im folgenden Abschnitt sollen die Möglichkeiten von Maßzahlen für Information und Komplexität in dieser Hinsicht getestet werden.

5.4.2 Information und Komplexität

In Abschnitt 5.4.1 wurden die Abflüsse der Einzugsgebiete bezüglich ihrer Saisonalität mit Hilfe der Autokorrelation und der Abflussmengen selbst untersucht. Nun werden die Niederschläge und Abflüsse aller (Teil-) Einzugsgebiete — soweit die Daten vorliegen — bezüglich ihrer Information und Komplexität verglichen. Insgesamt waren dies 34 Zeitreihen. Auf dieser Basis wird eine Klassifikation der Daten und Gebiete vorgenommen.

5.4.2.1 Vorgehensweise

Da die meisten Daten von Niederschlag und Abfluss in täglicher Auflösung vorlagen, wurden zunächst die höher aufgelösten Daten von Steinkreuz auf tägliche Werte aggregiert. Eine tägliche Auflösung ist für Abflussmessungen nahezu optimal hinsichtlich eines Maximums an Komplexität, wie für die drei in stündlicher Auflösung vorliegenden Zeitreihen von Lehstenbach, Steinkreuz und Lange Bramke in Abschnitt 5.3 festgestellt wurde.

Die vorliegenden Messzeiträume wurden in benachbarte Intervalle von jeweils vier Jahren unterteilt. Für jedes Intervall wurden der Informationsgewinn und die Fluktuationskomplexität bei binärer statischer Partitionierung und Wortlänge 4 berechnet. Beide Maße erwiesen sich bei den bisherigen Untersuchungen als zuverlässig zur Bestimmung von Information und Komplexität. Die Verwendung von dynamischen Partitionierungen scheidet nach den Erfahrungen in Abschnitt 5.3.1.1.3 aus. Höhere Alphabete lohnen sich im Allgemeinen nicht, wie die bisherigen Analysen und die Untersuchungen in Abschnitt 3.8.2 gezeigt haben. Von den für jedes 4-Jahresintervall berechneten Werten wurden die Mittelwerte und Standardabweichungen berechnet, um ein Maß für die Schwankungen der Werte zu erhalten, mit dem die Signifikanz von Unterschieden beurteilt werden kann.

Da der Messzeitraum für Steinkreuz nur genau vier Jahre beträgt, konnten hierfür keine Mittelwerte und Standardabweichungen berechnet werden. Dies wurde jedoch nicht zum Anlass genommen, die Intervalle auf zwei Jahre zu halbieren, da die Schwankung der Maße bei einem solchen Zeitraum noch groß ist, wie die Untersuchungen zur Stationarität in Abschnitt 3.3 (Abb. 3-3) gezeigt haben. Für den Niederschlag und Abfluss von Lehstenbach waren genau zwei 4-Jahresintervalle möglich; für den Abfluss von Owings Mills konnten vier und für Birkenes fünf Intervalle eingerichtet werden. Für die anderen Gebiete waren sechs bis elf (Lange Bramke) Intervalle möglich.

Zur Überprüfung der Konsistenz der Ergebnisse wurden die Rechnungen für äquiquantile (Median-) und für Entropie-maximale (H_G) Partitionierungen durchgeführt. In beiden Fällen wird der Informationsverlust durch die extreme Vergröberung (binäre Digitalisierung) des Wertebereiches der Daten minimiert. Es wird also eine maximale Informationserhaltung angestrebt. Bei der Median-Partitionierung ist die Information des Symbolsatzes aus Sicht der Metrischen Entropie (H_μ) maximal. Bei der H_G-maximalen Partitionierung ist die Information der Verteilung der Wörter, auf der die Maße (H_G und C_Γ) berechnet werden, aus Sicht des Informationsgewinns maximal. Diese Prinzipien basieren auf den Arbeiten von JAYNES (1957) und CRUTCHFIELD & PACKARD (1983) und wurden in Abschnitt 2.1.1.1 ausführlich erklärt. Die Erfahrungen in diesem Kapitel sprechen eher für die H_G-Maximierung als der stabileren Methode.

5.4.2.2 Ergebnisse

5.4.2.2.1 Anordnung der Zeitreihen im Informations-Komplexitäts-Diagramm (IKD)

Die Ergebnisse der Berechnungen sind in den Informations-Komplexitäts-Diagrammen (IKD) (Abb. 5-20 und Abb. 5-21) grafisch dargestellt. Diese Darstellungsart wird von LANGE (1999) Zufälligkeits-Komplexitäts-Diagramm (Z.-K. Diagramm) genannt. Bei beiden Partitionierungen ergibt sich das folgende einheitliche Bild: Die einzelnen Gebiete sind im IKD entlang einer schiefen Parabel-ähnlichen Funktion angeordnet. Diese Anordnung fiel auch schon bei der Aggregation stündlicher Messwerte in Abb. 5-14 auf. Trotz unabhängiger Definitionen gibt es also scheinbar einen funktionalen Zusammenhang von Information und Komplexität.

Dieser Zusammenhang kann durch eine Parameterkurve mit dem Informationsgewinn $H_G(p)$ für die Informationskoordinate und der Rényi-Komplexität $C_R(\alpha, p)$ der Ordnung α für die Komplexitätskoordinate approximiert werden. Dabei ist p der Parameter des Bernoulli-Prozesses (siehe 2.3.2). $H_G(p)$ entspricht dem Ausdruck in Gleichung (39) und $C_R(\alpha, p)$ wird mit dem Ausdruck (34) in Gleichung (62) verwendet. $C_R(\alpha, p)$ ist durch die Ordnungszahl α besser zur Approximation geeignet als die Fluktuationskomplexität C_Γ, die im Bernoulli-Fall von $C_R(\alpha, p)$ mit $\alpha \to 1$ approximiert wird (siehe 2.6.3).

Die Approximationskurven sind in Abb. 5-20 und Abb. 5-21 ebenfalls dargestellt. Die Informations-Komplexitäts-Punkte (IK-Punkte) der Zeitreihen liegen stets unterhalb der $(H_G(p), C_\Gamma(p)) = (H_G(p), C_R(1, p))$ -Kurve, wobei die Abweichung mit zunehmender Komplexität ebenfalls wächst. Der Bernoulli-Prozess ist ein Grenzprozess von völlig unabhängigen Ereignissen. Daher ist es wenig überraschend, wenn Zeitreihen, die ebenfalls sehr unabhängig, unkorreliert oder zufällig organisiert sind, eine ähnliche Information und Komplexität haben. Komplexere Zeitreihen weisen Korrelationen auf, die im Bernoulli-Prozess nicht vorhanden sind und zu einem größerem Abstand von der analytischen Kurve im IKD führen.

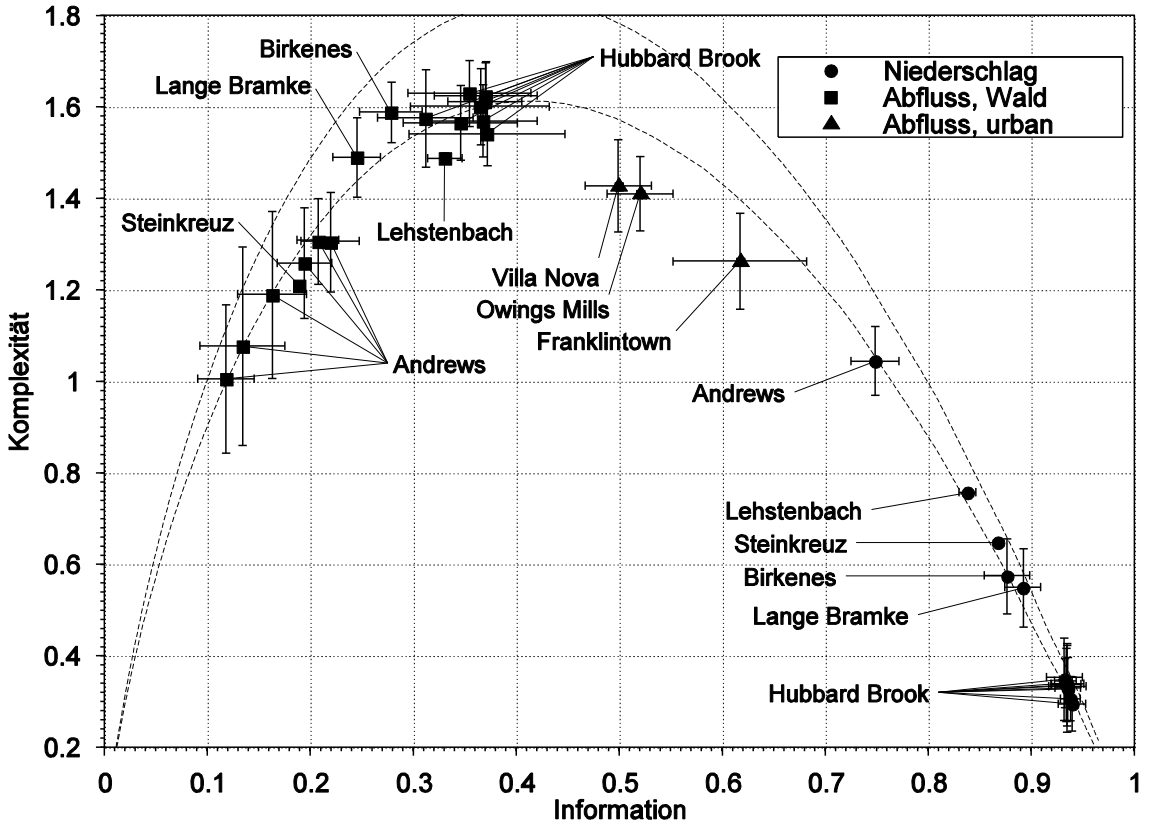

Abb. 5-20. Fluktuationskomplexität und Informationsgewinn von täglichem Niederschlag und Abfluss. Mittelwerte und Standardabweichungen aus benachbarten Zeiträumen von vier Jahren. Binäre statische **Median-Partitionierung**. Wortlänge 4. Gestrichelt sind die analytischen Kurven für C_Γ (oben) und $C_R(1.25)$ (Approximation) eingezeichnet.

Die $(H_G(p), C_R(\alpha, p))$-Kurve mit $\alpha = 1.25$ für die Median- und $\alpha = 1.3$ für die H_G-maximale Partitionierung approximiert die IK-Punkte — besonders im letzteren Fall — gut. Die H_G-Maximierung führt erwartungsgemäß zu einer höheren Bewertung der Information und damit insgesamt zu einem Vorwärtsrücken auf der Approximationskurve durch die IK-Punkte, aber auch zu einer Verzerrung. Die IK-Punkte entsprechen bei H_G-maximaler Partitionierung besser dem theoretischen Kurvenverlauf als bei der Median-Partitionierung. Dies kann neben den anderen Hinweisen in diesem Kapitel als Indiz für die bessere Eignung der H_G-maximalen Partitionierung gegenüber einer Median-Partitionierung gewertet werden, wenn der Funktion $(H_G(p), C_R(\alpha, p))$ die entsprechende Bedeutung zugesprochen werden kann. Ein weiteres Indiz dafür sind die höheren Standardabweichungen von H_G und C_Γ bei Median-Partitionierung. Diese waren auch schon bei der Bestimmung der Komplexitätsmaxima in Abschnitt 5.3.2.3 aufgefallen. Bei H_G-maximaler Partitionierung werden also insgesamt stabilere Werte der Maße erreicht.

Interessant ist hier, dass die Diversität der verschiedenen Zeitreihen beinahe jeden Punkt der analytischen Kurve liefert, der ein einfaches Nachzeichnen der Kurve erlaubt. Sehr einfache, d. h. wenig informative und komplexe Prozesse liegen aber nicht vor, so dass der linke Ast der Kurve nur ansatzweise besetzt ist. Nach der Filterhypothese ist zu erwarten, dass Kandidaten für diesen Ast in stark verwitterten Landschaften wie dem tropischen Tieflandregenwald zu suchen sind. Derartige Datensätze standen im Rahmen dieser Arbeit nicht zur Verfügung.

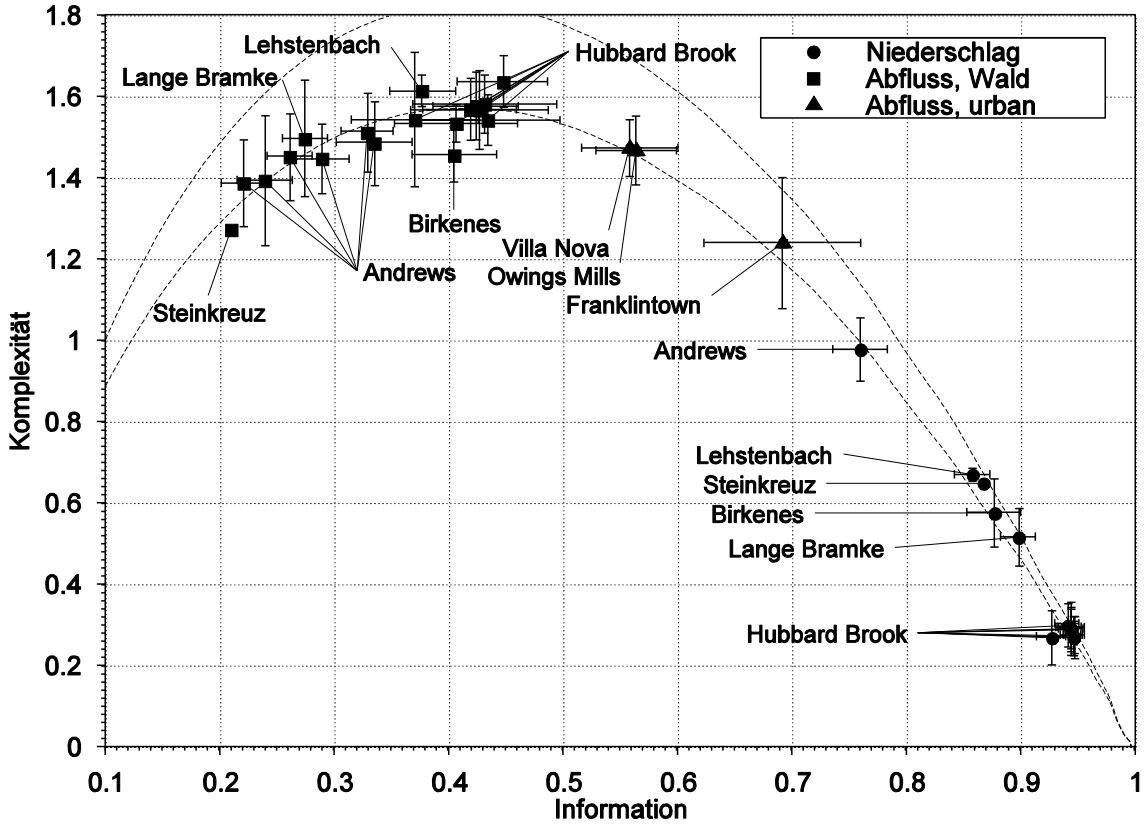

Abb. 5-21. Fluktuationskomplexität und Informationsgewinn von täglichem Niederschlag und Abfluss. Mittelwerte und Standardabweichungen aus benachbarten Zeiträumen von vier Jahren. Binäre statische $\boldsymbol{H_G}$**-maximale Partitionierung**. Wortlänge 4. Gestrichelt sind die analytischen Kurven für C_Γ (oben) und $C_R(1.3)$ (Approximation) eingezeichnet.

5.4.2.2.2 Unterscheidung von Gebieten und Prozessen

Der rechte Kurven-Ast ist durch die hoch-informativen Zeitreihen der Niederschläge nahezu vollständig besetzt. Hierbei zeichnet sich bereits eine klare Klassifizierung der hydrologischen Zeitreihen ab: Alle Niederschläge zeichnen sich durch eine hohe Information und geringe bis mäßige Komplexität aus, sind also auf dem rechten Kurven-Ast zu finden. Alle Abflüsse der bewaldeten, naturnahen Gebiete ordnen sich dagegen auf der linken Seite bei geringer bis mittlerer Information und (mittlerer bis) hoher Komplexität ein. Dies überrascht nach den bisherigen Analysen nicht und entspricht der intuitiven Erwartung.

Die Abflüsse der urbanen Gebiete liegen im IKD zwischen den Niederschlägen und Abflüssen. Dabei liegt das mit 90 % am stärksten bebaute Gebiet signifikant näher bei den Niederschlägen. Die höhere Unregelmäßigkeit / Dynamik der Abflüsse dieses Gebietes fällt schon bei der Betrachtung der Daten in Abb. 4-9 auf und wird hier klar quantifiziert. Bei der Autokorrelation (Abb. 5-18) konnte kaum ein Unterschied außerhalb des Signifikanzbereiches festgestellt werden.

Beim Vergleich der Niederschläge hebt sich Andrews durch eine geringere Information und höhere Komplexität deutlich von den anderen Gebieten ab. Die Ursache dafür ist der extrem ausgeprägte Jahresgang im Niederschlag des Andrews-Gebietes, der in Abb. 4-8 und in der Autokorrelation Abb. 5-19 zu erkennen ist und in Abschnitt 5.4.1 beschrieben wurde. Auf-

grund des ebenfalls in den Rohdaten (Abb. 4-6) und in der Autokorrelation (Abb. 5-19) erkennbaren Jahresganges im Niederschlag von Birkenes wäre hier die höchste Komplexität und niedrigste Information der Niederschlagsmessungen nach Andrews zu erwarten. Interessanterweise liegen die Gebiete Lehstenbach und Steinkreuz im IKD zwischen Andrews und Birkenes. Die drei deutschen Gebiete Lehstenbach, Steinkreuz und Lange Bramke weisen eine ähnliche Autokorrelation ihrer Niederschläge auf und liegen auch im IKD nahe beieinander. Deren Nähe zum Niederschlag in Birkenes kann zunächst nicht erklärt werden.

Die Niederschläge der acht Teileinzugsgebiete des Hubbard Brook sind die informationsreichsten und einfachsten Zeitreihen in diesem Vergleich. Dies entspricht der Erwartung an die gleichmäßige Verteilung dieser Niederschläge (siehe 4.5). In ihrer Autokorrelation sind sie fast nicht von einem Rauschprozess zu unterscheiden, wenn man von der signifikanten Korrelation von bis zu drei Tagen absieht, auf die in 5.4.1 nicht eingegangen wurde. Auch im IKD erreichen sie nicht das Informationsmaximum, dass sie als Zufallsprozess identifizieren würde. Davon sind sie noch etwa drei Standardabweichungen entfernt. Bei Median-Partitionierung weisen sich die Niederschläge auf den beiden Nordhang-Gebieten durch eine besonders geringe Komplexität aus. Insgesamt sind die IKP der acht Teilgebiete aber bei keiner Partitionierung signifikant verschieden.

Die Abflüsse der naturnahen, bewaldeten Wassereinzugsgebiete befinden sich auf dem linken Ast der Kurve im IKD bei insgesamt hoher Komplexität. Die beiden Extrema im Vergleich der Abflüsse sind die Teilgebiete von Hubbard Brook mit maximaler Komplexität und mitterer Information und die Teilgebiete von Andrews mit geringerer Komplexität und minimaler Information. Als Ursache dafür ist wieder die besonders strenge Saisonalität im Wasserhaushalt von Andrews zu sehen. Für die komplexe und vergleichsweise informationsreiche Dynamik bei Hubbard Brook gibt es keine konkrete Erklärung. Möglicherweise ist der in 5.4.1 festgestellte, schwach signifikante Halbjahresgang zusammen mit dem dominanten Jahresgang eine Ursache dafür. Die IK-Punkte von Hubbard Brook sind nicht signifikant verschieden, während die IK-Punkte von Andrews weiter aufgefächert sind und zwischen einigen Gebieten signifikante Unterschiede erkennen lassen.

Die IK-Punkte der anderen Abflüsse sind vor allem bei Median-Partitionierung zwischen denen von Andrews und Hubbard Brook angeordnet. Ihre spezielle Lage kann aber zunächst nicht erklärt werden. Auffällig ist das Gebiet Steinkreuz, das seine Position zwischen den beiden Partitionierungstypen kaum ändert und zwischen den Andrews-Gebieten im IKD angeordnet ist. Bei H_G-Maximierung liegt der Steinkreuz-Abfluss zwar noch im Variationsbereich der Andrews-Gebiete, hat aber die geringste Information und Komplexität.

Zumindest bezüglich der im Abfluss und Niederschlag extremen Gebiete Hubbard Brook und Andrews kann festgestellt werden, dass eher die Lage im IKD charakterisierend für ein Einzugsgebiet ist als die Informationsdifferenz. Letztere bleibt annähernd konstant. Sie könnte als Maß für die Gestörtheit eines Systems dienen, da sie für die urbanen Gebiete deutlich kleiner ist als für die naturnahen Gebiete, wenn der Niederschlag der urbanen Gebiete in IKD zwischen Hubbard Brook und Andrews vermutet werden kann. In diesem Sinne wird die von HAUHS & LANGE (1996b) vermutete Bedeutung der Informationsdifferenz zwischen Eingangs- und Ausgangsgröße eines Einzugsgebietes zugunsten der Bedeutung der IKD-Lage dieser Größen relativiert.

5.4.2.2.3 Klassifizierung

Anhand ihrer Lage im IKD läßt sich aufgrund der bisherigen Beobachtungen die folgende Grobeinteilung einer Zeitreihe des oberirdischen Wasserein- oder -austrages in einem Wassereinzugsgebiet vornehmen:

1. Niederschlag (hohe Information, geringe Komplexität)
2. Abfluss eines urbanen Gebietes (mittlere Information und Komplexität)
3. Abfluss eines naturnahen, bewaldeten Gebietes (geringe Information, hohe Komplexität)

Diese Feststellung ist nicht erstaunlich und zeigt eher die Zulässigkeit der Methode als ihre Möglichkeiten. Darüber hinaus konnten Hinweise für die Fähigkeiten zu einer feineren Unterteilung gewonnen werden. So läßt sich gegebenenfalls der Bebauungsgrad einer Fläche in seinem Wasserabfluss feststellen oder die Stärke eines Jahresrhythmus im Abfluss oder Niederschlag.

Um diese Hinweise prüfen und konkretisieren zu können, sind langjährige Messreihen von mehreren verschiedenen Gebieten erforderlich. Hier wurden nur sieben verschiedene nicht direkt benachbarte Einzugsgebiete untersucht. Die Beschaffung von entsprechendem Datenmaterial aus Deutschland oder Nachbarländern ist z. T. problematischer als das freie Herunterladen von Daten aus dem Internet vom LTER-Netzwerk der USA.

Die Informationen und Komplexitäten von unterschiedlichen Einzugsgebieten müssten dann auf statistisch signifikante Korrelationen mit einem Vektor von Gebietseigenschaften getestet werden. In diesem Vektor sollten unter anderem enthalten sein: geografische Koordinaten, Höhe über NN, Bodentiefe, Bewaldungsgrad, Bebauungsgrad, Klimazone, eventuell die Stärke und Frequenz einer Saisonalität.

Wenn die Korrelationen bekannt sind, kann die Methode in umgekehrter Weise genutzt werden. Von Veränderungen im Abfluss könnte auf Veränderungen in einem Einzugsgebiet geschlossen werden. So ließen sich beispielsweise die Auswirkungen neuer Flächennutzungen überprüfen und beurteilen. Eine hohe Abfluss-Information erschwert die Vorhersage des Abflusses, z. B. von Hoch- oder Niedrigwassern und kann in dieser Hinsicht gewertet werden.

6 Schlussbemerkungen

6.1 Berechnung von Information und Komplexität in Zeitreihen

Insgesamt wurden in dieser Arbeit sechs verschiedene Maße für Information und sieben Maße für Komplexität vorgestellt. Diese systematisch untersuchten Maße wurden für eine flexible Anwendung auf experimentelle, d. h. auf lückenhafte und endliche Zeitreihen programmiert. Das Programm (SYMDYN) ist ein Ergebnis und Bestandteil dieser Arbeit und steht über den FTP-Server des BITÖK mit einer ausführlichen Anleitung zur Verfügung.

Bei den untersuchten Informationsmaßen handelt es sich ausschließlich um bereits etablierte Methoden, die vergleichsweise problemlos anwendbar sind. Sie führen alle zu einer übereinstimmenden Bewertung des dynamischen Verhaltens: Eine Zeitreihe ist um so informationsreicher, je zufälliger ihre Werte aus einer äußeren Perspektive zeitlich angeordnet zu sein scheinen. Aufgrund der Stabilität seiner Werte bei fast allen Berechnungen erwies sich der Informationsgewinn, den eine zusätzliche Messung im Mittel liefert, als das wichtigste Maß für Information. Es wurde bereits in der Einleitung (1.1) an einem Beispiel veranschaulicht.

Die Komplexitätsmaße sind neueren Datums. Das erste Maß dieser Art ist die Effektive Maßkomplexität von GRASSBERGER (1986). Die „Komplexität“ nach KOLMOGOROV (1965) und LEMPEL & ZIV (1976) ist mangels einer Einsicht in die zugrunde liegenden Prozesse ein Maß für Information. Unter den Komplexitätsmaßen aus der Literatur erwies sich alleine die Fluktuationskomplexität nach BATES & SHEPARD (1993) als operationalisierbar und nützlich für experimentelle Zeitreihen im Sinne einer Bewertung von dynamischem Verhalten, die sich von den Informationsmaßen unterscheidet. Die Effektive Maßkomplexität nach GRASSBERGER (1986) führte bezüglich der Informationsmaße zu einer spiegelbildlichen Bewertung von Dynamik.

Eine erfolgreiche Rekonstruktion von ε-Maschinen nach CRUTCHFIELD & YOUNG (1989) ist trotz einer Erweiterung des Automaten-Zustandsbegriffs und der Zielfunktion für realistische Datenmengen nicht gelungen. Eine ε-Komplexität konnte somit auch hier (vgl. ROMAHN, 1996) nicht berechnet werden. Die Varianz-Komplexität nach ATMANSPACHER et al. (1997) konnte zwar berechnet werden, führte aber zu keiner mit dem Konzept von Information oder Komplexität konformen Bewertung von Dynamik. Daraufhin wurden eigene Versuche einer metastatistischen Definition von Komplexität unternommen, da diese Ansätze die Aussicht auf ein parameterfreies Verfahren boten. Mangels einer eindeutig interpretierbaren Bewertung der Dynamik von bekannten Testprozessen, wie der logistischen Abbildung, wurden auch diese Versuche als nicht erfolgreich aufgegeben.

Die Definition der Rényi-Komplexität in dieser Arbeit hat zu einem neuen Maß für Komplexität geführt. Es liefert die gleiche qualitative Bewertung der Dynamik wie die Fluktuationskomplexität und stimmt für Bernoulli-Prozesse exakt mit ihr überein. Letztere bewertet hohe Schwankungen in der Bilanz zwischen dem lokalen Informationsgewinn durch eine zusätzliche Messung und dem Informationsverlust durch das gleichzeitige Vergessen einer älteren

Messung als komplex. Die Rényi-Komplexität orientiert sich an der Verteilung von Information in einem Datensatz. Sie basiert auf der Rényi-Entropie nach RÉNYI (1960), die eine Gewichtung von selteneren oder häufigeren Ereignissen erlaubt und damit die Shannon-Entropie nach SHANNON (1948) verallgemeinert. Nach der Rényi-Komplexität wird ein hohes Ungleichgewicht in der Informationsbilanz von seltenen und häufigen Ereignissen als komplex bewertet. Mit einem Parameter kann die Stärke dieser Bewertung beeinflusst werden.

In einer vergleichenden Analyse wurden Zeitreihen vom Niederschlag und Abfluss verschiedener Einzugsgebiete bezüglich ihrer Information (Informationsgewinn) und Komplexität (Fluktuationskomplexität) in einem Informations-Komplexitäts-Diagramm gegenübergestellt. Dabei stellte sich heraus, dass die Punkte der verschiedenen Gebiete und Zeitreihen entlang einer Kurve angeordnet sind. Diese Kurve kann erstaunlich gut durch eine Parameterkurve aus Informationsgewinn und Rényi-Komplexität für einen Bernoulli-Prozess approximiert werden. Als entscheidender Vorteil der Rényi-Komplexität gegenüber der Fluktuationskomplexität erwies sich dabei die Möglichkeit der gewichteten (parametrisierten) Bewertung der Informationsverteilung. Dieses Gewicht könnte möglicherweise charakteristisch für den untersuchten Datentyp sein. Diese Behauptung muss jedoch anhand umfangreicher Vergleichsrechnungen mit anderen Datentypen überprüft werden.

Die Rényi-Komplexität erwies sich im Vergleich zur Fluktuationskomplexität außerdem als die stabilere Alternative zur Berechnung von Komplexität: Schwankungen der Rényi-Komplexität waren bei einer Vergröberung der Messauflösung bei noch hinreichender Datenmenge oft kaum festzustellen, was die Bestimmung von Komplexitätsmaxima deutlich erleichterte. Diese Bestimmung war bei der Fluktuationskomplexität durch die lokalen Schwankungen deutlich erschwert. Dementsprechend streuten die Rényi-Komplexitäten deutlich weniger um einen Mittelwert als die Fluktuationskomplexitäten.

Die Forderung einer bestimmten mittleren Genauigkeit für die Berechnung eines Komplexitätsmaßes führte zur Fixierung eines Parameters (der Wortlänge) der Methoden. Dazu wurden zuerst analytische Ausdrücke für die maximal erforderliche Datenmenge hergeleitet. Die numerische Lösung dieser Gleichungen führte zu den Tabellen im Anhang 7.9, die auch in SYMDYN zur automatischen Berechnung der (maximalen) Wortlänge enthalten sind. Der Informationsgewinn kann bereits ab 63, die Fluktuationskomplexität ab 146 und die Rényi-Komplexität ab 125 Datenpunkten für eine minimale Wortlänge von 2 und mit einer mittleren Genauigkeit von 5 % bei beliebig zufälligen Daten berechnet werden. Diese Grenze wird für nicht-informationsmaximale Daten als ausreichend erachtet. Für die 1 % mittlere Genauigkeit sind bei maximaler Information bereits 294, 723 und 618 Datenpunkte erforderlich, also fast zwei Jahre täglicher Messwerte.

Desweiteren konnte die Partitionierung als wichtigster Parameter der Verfahren aus prinzipiellen Motiven und nach Stabilitätskriterien weitgehend fixiert werden. Entropie-maximale Partitionierungen lieferten in fast allen Fällen die stabilsten Ergebnisse im Vergleich zu Median-Partitionierungen. Auch hierfür kann der Informationsgewinn als Favorit gelten, obwohl eine Metrische-Entropie-Maximierung im Sinne von JAYNES (1957) und nach den Erfahrungen aus Abschnitt 5.3 dieser Arbeit eine lohnende Alternative sein kann.

Der Jahresrhythmus in den Zeitreihen zum Gebietsabfluss oder noch längere Trends bedeuten Instationaritäten, welche die Grundannahmen der Methoden prinzipiell verletzten. Es konnte jedoch gezeigt werden, dass die Bereinigung des Jahresganges nur wenig signifikante Abweichungen in den Komplexitätsmaßen bewirkt und die Zeitreihen mit einem künstlichen Signal versieht, da die Periode nie exakt 365 Tage lang ist. Der Jahresgang verhindert die Berechnung der Maße also nicht. Er darf nicht eliminiert werden, da er ein Kennzeichen des Klimas

und der Transpiration der Bäume ist, dessen Einfluss auf den Wasserhaushalt hier untersucht werden soll.

Mit SYMDYN und den Untersuchungen zur Anwendbarkeit der Methode wurde in dieser Arbeit eine Grundlage für eine äußere und praktisch parameterfreie Bewertung von Dynamik im Sinne von Information und Komplexität in experimentellen Zeitreihen geschaffen. Die untersuchten Methoden gehen über das hinaus, was verbreitete statitsische Verfahren liefern können und stehen offenbar in Beziehung zu Eigenschaften der betreffenden Ökosysteme (siehe folgenden Abschnitt).

6.2 Analyse des Wasserhaushaltes von bewaldeten Einzugsgebieten

Prinzipiell konnte eine Informationsabnahme des Wassers beim Durchlaufen der untersuchten Einzugsgebiete vom Niederschlag bis zum Abfluss festgestellt werden. Die Abfluss-Information nimmt mit zunehmender Regelmäßigkeit des Niederschlags ab. Sie nimmt aber mit verminderter (naturnaher) Bewaldung zu und nähert sich damit der Information des Niederschlags an. Die generelle Informationsabnahme sowie die erhöhte Informationsdifferenz zwischen Niederschlag und Abfluss bei vollständiger Bewaldung bestärkt die von HAUHS & LANGE (1996a, 1996b) postulierte Hypothese einer Informationsfilterung und dessen Beziehung zum „biologischen Organisationsgrad“ eines Ökosystems.

Eine geringe (naturnahe) Bewaldung sowie die Auswirkung von Kahlschlägen kann auch in den Abflussmengen direkt erkannt werden. Die Autokorrelationslängen als Beispiel für eine klassische Methode der Zeitreihenanalyse sind bei deutlich reduziertem Bewuchs ebenfalls verringert. Die Berechnung der Abfluss-Information erlaubt im Vergleich dazu jedoch eine feinere und einfachere Beurteilung der Transpirationsleistung durch die Vegetation anhand der Dynamik des Abflusses. Die mit der Information gemessene Kurzzeitdynamik von Abflussmengen erlaubt somit eine Beurteilung der Belebtheit des Systems oder anderer biologischer Eigenschaften.

Möglicherweise eignet sich die Abfluss-Information als Indikator für eine Erkennung entsprechender Veränderungen im System. Die Bebauung von Einzugsgebieten bewirkt eine höhere Abfluss-Information und damit eine schlechtere Vorhersagbarkeit (von Hochwassern). Somit könnte für die Bebauung von sensiblen Gebieten die Einhaltung eines Grenzwertes der Abfluss-Information gefordert werden, um eine gewisse Vorwarnzeit zu gewährleisten. Vor einem diesbezüglich praxisreifen Konzept müssen die initialen Erkenntnisse in dieser Arbeit durch weitere Untersuchungen bestärkt und erweitert werden.

Die Quantifizierung der Information des Matrixpotentials in verschiedenen Bodentiefen und dessen Vergleich mit der Abfluss-Information erlaubt eine aggregierte Beurteilung der für den Abfluss relevanten „effektiven“ Eindringtiefe des Niederschlags ohne explizite Kenntnis der hydraulischen Bodeneigenschaften. Die Unterschiede dieser Eindringtiefen für die zwei diesbezüglich untersuchten Gebiete konnten in Beziehung zu den unterschiedlichen Bodeneigenschaften gesetzt werden. Dieses Beispiel zeigt, dass es möglich ist durch ein objektives Verfahren alleine anhand der hydrologischen Dynamik Hinweise auf diese Gebietseigenschaften zu erhalten.

Während für die bisherigen Ergebnisse alleine die Betrachtung der Information ausreichend war, erwies sich die Komplexität als ein nützliches Konzept zur Bestimmung einer effektiven

Zeitskala der Messung. Ein Maximum an Komplexität kann als ideale Darstellung von Information interpretiert werden, da eine geringere Komplexität bei geringerer Information ein Indiz für Redundanz ist und eine geringere Komplexität bei höherer Information ein Indiz für Zufälligkeit ist. Eine Modellierung und Vorhersage von zu informationsreichen Daten ist schwierig oder gar unmöglich. Eine Extrapolation von redundanten Daten ist trivial. Die Erhebung redundanter Daten ist teuer und verbraucht unnötig Speicherplatz, wenn außer der mittleren Dynamik kein Interesse an vorübergehenden Extremereignissen besteht. Die Erhebung von zu zufälligen Daten ist nahezu sinnlos, da die relevante Dynamik unsichtbar bleibt.

Entsprechend der allgemeinen Informationsabnahme des Wassers beim Durchlaufen eines Einzugsgebietes bei fester täglicher Zeitauflösung wurde anhand von höher aufgelösten Daten eine Zunahme der effektiven Zeitauflösung festgestellt. Diese liegt beim Niederschlag im Bereich von 20 bis 140 Minuten, beim Abfluss zwischen einem und fünf Tagen und beim Matrixpotential dazwischen. Die Werte für den Niederschlag rechtfertigen die Messung von Niederschlag in 10-minütlicher und eventuell stündlicher Auflösung. Tägliche Niederschlagsaufzeichnungen werden der Dynamik des Niederschlags jedoch nicht mehr gerecht. Die effektiven Auflösungen des Abflusses waren stärker gebietsabhängig als die des Niederschlags. Eine tägliche Abflussmessung war in allen Fällen zur Beschreibung der mittleren Dynamik ausreichend. Bei der Langen Bramke wären auch 2-tägliche Messungen ausreichend. Im Steinkreuz-Gebiet sollte der Abfluss mindestens alle fünf Tage gemessen werden. Die hier ermittelten Messintervalle entsprechen den üblichen Beobachtungen der operativen Hydrologie, die aber aus einer völlig anderen heuristischen Begründung heraus so gewählt wurden.

Ein Anwendungsfeld der Methode, das in dieser Arbeit nicht verfolgt wurde, liegt in der Beurteilung von Modellen. Erste Untersuchungen in dieser Hinsicht auf Ökosystemebene wurden von THIES (1998) im Rahmen einer Diplomarbeit durchgeführt.

Die Untersuchungen zum Wasserhaushalt bewaldeter Einzugsgebiete stellen in dieser Arbeit keine erschöpfende Analyse dar. Sie sollen vielmehr die Möglichkeiten aufzeigen, die eine Quantifizierung von Information und Komplexität aus einer äußeren Betrachtung als eine neue Methode in der Ökosystemforschung bietet. Mit dieser Arbeit wurden die methodischen Grundlagen und Voraussetzungen dafür geschaffen, die Vielzahl von Datenreihen aus dem Monitoring von Ökosystemen auf ihre Eignung für eine Modellierung und Vorhersage und auf neue Indikatoren von Veränderungen systematisch zu überprüfen.

7 Anhang

7.1 Die Shannon-Entropie des Bernoulli-Prozesses

Bei einem binären Bernoulli-Prozess treten die beiden Ereignisse (Symbole) 0 und 1 mit den Häufigkeiten $p \in [0,1]$ und $q = 1 - p$ unabhängig voneinander auf. Eine Folge von L Symbolen, ein L-Wort $w_{L,i}$, hat dann die Häufigkeit $p(w_{L,i}) = p^k q^{L-k}$, wenn in $w_{L,i}$ das Symbol 0 k-mal vorkommt. Es gibt genau

$$\binom{L}{k} = \frac{L!}{(L-k)!k!} \tag{93}$$

verschiedene L-Wörter, in denen das Symbol 0 k-mal vorkommt. Insgesamt gibt es 2^L verschiedene Wörter. Für die Shannon-Entropie (26) der L-Wörter gilt dann:

$$\begin{aligned}
H_S(L) &= -\sum_{i=1}^{2^L} p(w_{L,i}) \log_2 p(w_{L,i}) \\
&= -\sum_{k=0}^{L} \binom{L}{k} p^k q^{L-k} \log_2 p^k q^{L-k} \\
&= -p^L \log_2 p^L - \sum_{k=0}^{L-1} k \binom{L}{k} p^k q^{L-k} \log_2 p - \sum_{k=0}^{L-1} (L-k) \binom{L}{k} p^k q^{L-k} \log_2 q \\
&= -Lp^L \log_2 p - L \sum_{k=1}^{L-1} \frac{(L-1)!}{(L-k)!(k-1)!} p^k q^{L-k} \log_2 p - L \sum_{k=0}^{L-1} \frac{(L-1)!}{(L-1-k)!k!} p^k q^{L-k} \log_2 q \qquad \text{mit (93)} \\
&= -Lp^L \log_2 p - L \sum_{k=0}^{L-2} \binom{L-1}{k} p^{k+1} q^{L-1-k} \log_2 p - L \sum_{k=0}^{L-1} \binom{L-1}{k} p^k q^{L-k} \log_2 q \\
&= -Lp(\log_2 p) \sum_{k=0}^{L-1} \binom{L-1}{k} p^k q^{L-1-k} - Lq(\log_2 q) \sum_{k=0}^{L-1} \binom{L-1}{k} p^k q^{L-1-k} \\
&= -L(p \log_2 p + q \log_2 q)(p+q)^{L-1} \qquad \text{binomischer Satz} \\
&= -L(p \log_2 p + (1-p) \log_2 (1-p)) \qquad \text{mit } q = 1 - p
\end{aligned}$$

Damit ist Formel (29) von Seite 39 bewiesen.

7.2 Rényi-Entropie des Bernoulli-Prozesses

Mit den Bezeichnungen aus dem letzten Abschnitt 7.1 gilt für die Rényi-Entropie (30) der L-Wörter eines Bernoulli-Prozesses:

$$
\begin{aligned}
H_{\mathrm{R}}(\alpha,p,L) &= \frac{1}{1-\alpha}\log_2\sum_{i=1}^{2^L}\left(p(w_{L,i})\right)^{\alpha} \\
&= \frac{1}{1-\alpha}\log_2\sum_{k=0}^{L}\binom{L}{k}\left(p^k q^{L-k}\right)^{\alpha} \\
&= \frac{1}{1-\alpha}\log_2\sum_{k=0}^{L}\binom{L}{k}\left(p^{\alpha}\right)^{k}\left(q^{\alpha}\right)^{L-k} \\
&= \frac{1}{1-\alpha}\log_2\left(p^{\alpha}+q^{\alpha}\right)^{L} \\
&= \frac{L}{1-\alpha}\log_2\left(p^{\alpha}+(1-p)^{\alpha}\right)
\end{aligned}
$$

7.3 Formeln für den mittleren Informationsgewinn

Die Äquivalenz der Formeln (41) und (42) für den mittleren Informationsgewinn H_{G} wurde bereits von SHANNON (1976, S. 63 u.66) und WACKERBAUER et al. (1994) beschrieben. Sie wird hier für eine Verteilung von L-Wörtern (siehe 2.1.2) ausgehend von Gleichung (41) gezeigt:

$$
\begin{aligned}
H_{\mathrm{G}}(L) &= -\sum_{i,j=1}^{\lambda^L} p_{L,ij}\log_2 p_{L,i\to j} \\
&= -\sum_{i,j=1}^{\lambda^L} p_{L,ij}\log_2\frac{p_{L,ij}}{p_{L,i}} && \text{nach Gleichung (10), S. 23} \\
&= -\sum_{i,j=1}^{\lambda^L} p_{L,ij}\log_2 p_{L,ij} + \sum_{i,j=1}^{\lambda^L} p_{L,ij}\log_2 p_{L,i} \\
&= -\sum_{i=1}^{\lambda^{L+1}} p_{L+1,i}\log_2 p_{L+1,i} + \sum_{i=1}^{\lambda^L}\left(\sum_{j=1}^{\lambda^L} p_{L,ij}\right)\log_2 p_{L,i} && \text{siehe Bemerkung} \\
&= -\sum_{i=1}^{\lambda^{L+1}} p_{L+1,i}\log_2 p_{L+1,i} + \sum_{i=1}^{\lambda^L} p_{L,i}\log_2 p_{L,i} \\
&= H_{\mathrm{S}}(L+1) - H_{\mathrm{S}}(L) && \text{mit Gleichung (26)}
\end{aligned}
$$

Damit ist die Äquivalenz der Formeln (41) und (42) gezeigt.

Bemerkung: Wegen der Überlappung zweier aufeinanderfolgender L-Wörter um $(L-1)$ Symbole können diese als ein $(L+1)$-Wort interpretiert werden, also auch: $p_{L,ij} = p_{L+1,i}$ (siehe Abb. 2-5).

7.4 Formeln für die mittlere wechselseitige Information

Die Äquivalenz der beiden Formeln (44) und (45) für die mittlere wechselseitige Information H_{M} über eine Verteilung von L-Wörtern (siehe 2.1.2) ergibt sich ausgehend von Gleichung (44) wie folgt:

$$
\begin{aligned}
H_{\mathrm{M}}(L) &= \sum_{i,j=1}^{\lambda^L} p_{L,ij} \log_2 \frac{p_{L,ij}}{p_{L,i}\, p_{L,j}} \\
&= \sum_{i,j=1}^{\lambda^L} p_{L,ij} \log_2 p_{L,ij} - \sum_{i,j=1}^{\lambda^L} p_{L,ij} \log_2 p_{L,i} - \sum_{i,j=1}^{\lambda^L} p_{L,ij} \log_2 p_{L,j} \\
&= \sum_{i=1}^{\lambda^{L+1}} p_{L+1,i} \log_2 p_{L+1,i} - \sum_{i=1}^{\lambda^L} \left(\sum_{j=1}^{\lambda^L} p_{L,ij} \right) \log_2 p_{L,i} - \sum_{j=1}^{\lambda^L} \left(\sum_{i=1}^{\lambda^L} p_{L,ij} \right) \log_2 p_{L,j} \qquad \text{Bem. von 7.3} \\
&= \sum_{i=1}^{\lambda^{L+1}} p_{L+1,i} \log_2 p_{L+1,i} - \sum_{i=1}^{\lambda^L} p_{L,i} \log_2 p_{L,i} - \sum_{j=1}^{\lambda^L} p_{L,j} \log_2 p_{L,j} \\
&= 2H_{\mathrm{S}}(L) - H_{\mathrm{S}}(L+1) \qquad \text{nach Def. (26) für } H_{\mathrm{S}}
\end{aligned}
$$

7.5 Mittelwert des Netto-Informationsgewinns

Für den Mittelwert des Netto-Informationsgewinns gilt ausgehend von Gleichung (59):

$$
\begin{aligned}
\langle \Gamma \rangle &= \sum_{i,j=1}^{\lambda^L} p_{L,ij} \Gamma_{ij} \\
&= \sum_{i,j=1}^{\lambda^L} p_{L,ij} \log_2 p_{L,i} - \sum_{i,j=1}^{\lambda^L} p_{L,ij} \log_2 p_{L,j} \qquad \text{nach Gleichung (58)} \\
&= \sum_{i=1}^{\lambda^L} \left(\sum_{j=1}^{\lambda^L} p_{L,ij} \right) \log_2 p_{L,i} - \sum_{j=1}^{\lambda^L} \left(\sum_{i=1}^{\lambda^L} p_{L,ij} \right) \log_2 p_{L,j} \\
&= \sum_{i=1}^{\lambda^L} p_{L,i} \log_2 p_{L,i} - \sum_{j=1}^{\lambda^L} p_{L,j} \log_2 p_{L,j} \\
&= -H_{\mathrm{S}}(L) + H_{\mathrm{S}}(L) \qquad \text{mit Shannon-Entropie (26)} \\
&= 0
\end{aligned}
$$

7.6 Fluktuationskomplexität für einen Bernoulli-Prozess

Die Fluktuationskomplexität (60) wird mit den Übergangshäufigkeiten von je zwei L-Wörtern $\mathrm{w}_{L,i}$ und $\mathrm{w}_{L,j}$ berechnet. Diese überlappen sich um die letzten $L-1$ Symbole von $\mathrm{w}_{L,i}$ bzw. die ersten $L-1$ Symbole von $\mathrm{w}_{L,j}$ (siehe 2.1.2, insbes. Abb. 2-5). Bei einem binären Bernoulli-Prozess treten die Symbole 0 und 1 mit den Häufigkeiten $p \in [0,1]$ und $q = 1 - p$ unabhängig voneinander auf. Der Überlappungsteil von $\mathrm{w}_{L,i}$ und $\mathrm{w}_{L,j}$ ist ein (L-1)-Wort mit der Häufigkeit $p^k q^{L-1-k}$, wenn das Symbol 0 darin k-mal vorkommt (vgl. Beginn von Anhang 7.1). Die Anzahl solcher (L-1)-Wörter berechnet sich nach dem Binomialkoeffizienten (93) mit L-1 anstatt L. Die Wörter $\mathrm{w}_{L,i}$ und $\mathrm{w}_{L,j}$ können zusätzlich jeweils das Symbol 0 oder 1 enthalten. Ihre Häufigkeiten $p_{L,i}$, $p_{L,j}$ sind entsprechend $p^{k+1}q^{L-1-k}$ oder $p^k q^{L-k}$. Bei der Berechnung der Fluktuationskomplexität nach Gleichung (60) müssen alle vier Kombinationen für das erste Symbol von $\mathrm{w}_{L,i}$ und das letzte Symbol von $\mathrm{w}_{L,j}$ beachtet werden:

$$
\begin{aligned}
C_\Gamma &= \sum_{i,j=1}^{\lambda^L} p_{L,ij}\left(\log_2\frac{p_{L,i}}{p_{L,j}}\right)^2 \\
&= \sum_{k=0}^{L-1}\binom{L-1}{k}p^k q^{L-1-k}\left[p^2\left(\log_2\frac{p}{p}\right)^2 + pq\left(\log_2\frac{p}{q}\right)^2 + qp\left(\log_2\frac{q}{p}\right)^2 + q^2\left(\log_2\frac{q}{q}\right)^2\right] \\
&= 2pq\left(\log_2\frac{p}{q}\right)^2\sum_{k=0}^{L-1}\binom{L-1}{k}p^k q^{L-1-k} \\
&= 2pq\left(\log_2\frac{p}{q}\right)^2(p+q)^{L-1} && \text{binomischer Satz} \\
&= 2p(1-p)\left(\log_2\frac{p}{1-p}\right)^2 && \text{mit } q = 1-p
\end{aligned}
$$

Dies ist eine in dem interessierenden Bereich $p \in [0,1]$ stetige und positive Funktion, die symmetrisch zu $p = ½$ ist. Für die monosymbolischen Sequenzen gilt erwartungsgemäß $C_\Gamma(p{=}0) = C_\Gamma(p{=}1) = 0$. Im Falle maximaler Zufälligkeit gilt ebenfalls $C_\Gamma(p{=}½) = 0$. Zwischen diesen Eckpunkten muss es wegen der Stetigkeit und Positivität ein Maximum geben. Dieses kann über die Nullstellen der ersten Ableitung

$$
\begin{aligned}
\frac{dC_\Gamma}{dp} &= \frac{2(1-2p)}{\ln^2 2}\left(\ln\frac{p}{1-p}\right)^2 + \frac{4p(1-p)}{\ln^2 2}\left(\ln\frac{p}{1-p}\right)\frac{1-p}{p}\frac{1}{(1-p)^2} \\
&= \frac{2}{\ln^2 2}\left[(1-2p)\left(\ln\frac{p}{1-p}\right)+2\right]\ln\frac{p}{1-p}
\end{aligned}
$$

ermittelt werden. Die Nullstellen liegen bei $p = ½$ (Minimum) und den Lösungen von

$$\left(\frac{p}{1-p}\right)^{1-2p} = e^{-2}$$

Diese wurden mit einem Newton-Verfahren (Routine „FindRoot“ von Mathematica 3.0, WOLFRAM, 1996, S. 884 u. 1088; zum Verfahren: z. B. STOER, 1994, S. 288ff) numerisch ermittelt:

$$p_1 \approx 0.0832217202 \quad \text{und} \quad p_2 \approx 0.9167782798$$

Die Höhe des Maximums ist $C_\Gamma(p_1) = C_\Gamma(p_2) \approx 1.8283945658$.

Wenn man den Bereich $p \in [0,½]$ als Zufälligkeitsskala von 0 % bis 100 % interpretiert, erreicht die Fluktuationskomplexität des binären Bernoulli-Prozesses für $p_1 \cdot 200\ \%$ = 16.644 % mit 1.828 ihr Maximum.

7.7 Zur Effektiven Maßkomplexität

Für die Effektive Maßkomplexität C_{EM} über eine Verteilung von L-Wörtern (siehe 2.1.2) gilt nach Definition (52):

$$C_{\mathrm{EM}} = \lim_{l\to\infty} \sum_{n=1}^{L} n \left(H_{\mathrm{G}}(n-1) - H_{\mathrm{G}}(n) \right)$$

Die Teilsummen S_L approximieren C_{EM}. Einsetzen des Informationsgewinns H_{G} nach (42) liefert:

$$C_{\mathrm{EM}} \approx S_L = \sum_{n=1}^{L} n \left(-H_{\mathrm{S}}(n+1) + 2H_{\mathrm{S}}(n) - H_{\mathrm{S}}(n-1) \right)$$

Für jedes n mit $2 \le n \le \mathrm{L}-1$ enthält die Summe S_L je drei Summanden der n-Wort Shannon-Entropie $H_{\mathrm{S}}(n)$. Mit $H_{\mathrm{S}}(0) = 0$ gilt dann:

$$\begin{aligned} S_L &= -LH_{\mathrm{S}}(L+1) + 2LH_{\mathrm{S}}(L) - (L-1)H_{\mathrm{S}}(L) + \sum_{n=2}^{L-1} \left(-(n+1) + 2n - (n-1) \right) H_{\mathrm{S}}(n) \\ &\quad - 2H_{\mathrm{S}}(1) + 2H_{\mathrm{S}}(1) - H_{\mathrm{S}}(0) \\ &= (L+1)H_{\mathrm{S}}(L) - LH_{\mathrm{S}}(L+1) \end{aligned}$$

Dies liefert die Approximation in Gleichung (55) und ist Grundlage zur Berechnung von C_{EM}. Eine Kompaktifizierung kann ähnlich wie für H_{G} (siehe 7.3) und H_{M} (siehe 7.4) mit der Definition der Shannon-Entropie (26) über die Häufigkeiten $p_{L,i}$ der L-Wörter und die Häufigkeiten $p_{L+1,i} = p_{L,ij}$ der $(L+1)$-Wörter als zwei aufeinanderfolgende L-Wörter erreicht werden:

$$\begin{aligned} S_L &= -(L+1) \sum_{i=1}^{\lambda^L} p_{L,i} \log_2 p_{L,i} + L \sum_{i=1}^{\lambda^{L+1}} p_{L+1,i} \log_2 p_{L+1,i} \\ &= -(L+1) \sum_{i=1}^{\lambda^L} \left(\sum_{j=1}^{\lambda^L} p_{L,ij} \right) \log_2 p_{L,i} + L \sum_{i,j=1}^{\lambda^L} p_{L,ij} \log_2 p_{L,ij} \\ &= \sum_{i,j=1}^{\lambda^L} p_{L,ij} \left(L \log_2 p_{L,ij} - (L+1) \log_2 p_{L,i} \right) \\ &= \sum_{i,j=1}^{\lambda^L} p_{L,ij} \log_2 \frac{p_{L,ij}^{L}}{p_{L,i}^{L+1}} \\ &= \sum_{i,j=1}^{\lambda^L} p_{L,ij} \log_2 \frac{p_{L,i\to j}^{L}}{p_{L,i}} \end{aligned}$$

Damit ist die zweite Formel zur praktischen Berechnung von C_{EM} gezeigt.

7.8 Erwartungswert der Shannon-Entropie

Mit den Bezeichnungen in Abschnitt 3.6 gilt für den Erwartungswert der Shannon-Entropie H_{S} nach (26) und (77):

$$
\begin{aligned}
H_{\mathrm{S,exp}} &= -\lambda^L \sum_{k=1}^{N-L+1} \binom{N-L+1}{k} p_L^k q_L^{N-L+1-k} \frac{k}{N-L+1} \log_2 \frac{k}{N-L+1} \\
&= \frac{\lambda^L \log_2 (N-L+1)}{N-L+1} \sum_{k=1}^{N-L+1} \binom{N-L+1}{k} p_L^k q_L^{N-L+1-k} k \\
&\quad -\frac{\lambda^L}{N-L+1} \sum_{k=1}^{N-L+1} \binom{N-L+1}{k} p_L^k q_L^{N-L+1-k} k \log_2 k
\end{aligned}
$$

Der Index $k = 0$ wurde in der ersten Zeile ignoriert, weil er keinen Beitrag zur Summe liefert. Der Faktor k wird nun in die Binomialkoeffizienten gezogen und durch Umindizieren erhält man (für $p_L = \lambda^{-L}$):

$$
\begin{aligned}
H_{\mathrm{S,exp}} &= \log_2 (N-L+1) \sum_{k=0}^{N-L} \binom{N-L}{k} p_L^k q_L^{N-L-k} - \sum_{k=0}^{N-L} \binom{N-L}{k} p_L^k q_L^{N-L-k} \log_2 (k+1) \\
&= \log_2 (N-L+1) - \sum_{k=0}^{N-L} \binom{N-L}{k} p_L^k q_L^{N-L-k} \log_2 (k+1)
\end{aligned}
$$

Die letzte Zeile erhält man durch Anwendung des binomischen Satzes auf den ersten Summanden.

7.9 Tabellen zur erforderlichen Datenmenge

Hier sind die Tabellen aufgeführt, die mit den Formeln in Abschnitt 3.6 berechnet wurden.

Tabelle 7-1: Erforderliche Datenmengen bei verschiedenen Wortlängen *L* für 1 % Genauigkeit und Alphabetgröße 2. Maße: Shannon-Entropie H_S, Metrische Entropie H_μ, Rényi-Entropie $H_R(\approx 1)$, Informationsgewinn H_G (beide Formeln), Wechselseitige Information H_M ($H_{M,c}$: kompakte Formel (44); $H_{M,d}$: Differenzenformel (45)) Effektive Maßkomplexität C_{EM} (beide Formeln), Fluktuationskomplexität C_Γ und Rényi-Komplexität $C_R(\alpha = 1.0001)$.

L	H_S, H_μ, H_R	H_G	$H_{M,d}$	$H_{M,c}$	C_{EM}	C_Γ	C_R
1	73	147	77		77	434	420
2	111	294	85	148	368	723	618
3	172	586	69	222	1237	1301	967
4	277	1168	89	344	3550	2454	1559
5	458	2331	122	552	9331	4759	2571
6	775	4657	95	913	23199	9368	4346
7	1339	9307	186	1546	55552	18586	7490
8	2355	18607	360	2673	129488	37020	13147
9	4202	37206	697	4704	295825	73888	23376
10	7589	74403	1350	8398	665429	147622	42089
11	13845	148796	2620	15171	1478500	295090	76438
12	25471	297580	5091	27682		590025	139951
13	47195	595148	9906	50933		1179893	257828
14	87982	1190283	19300	94380			478456
15	164880		37639	175952			892155
16	310398		73473	329747			1669920
17	586677		143536	620782			
18	1112759		280611	1173340			
19			548948				
20			1074520				

Tabelle 7-2: Erforderliche Datenmengen bei verschiedenen Wortlängen *L* für 1 % Genauigkeit und Alphabetgröße 3. Maße: wie Tabelle 7-1.

L	H_S, H_μ, H_R	H_G	$H_{M,d}$	$H_{M,c}$	C_{EM}	C_Γ	C_R
1	92	277	294		294	1155	821
2	185	828	469	278	2030	2885	1640
3	402	2480	677	555	9835	8070	3581
4	929	7434	1370	1203	41038	23625	8224
5	2253	22292	3180	2781	158013	70286	20144
6	5669	66864	7906	6752	579049	210269	51738
7	14659	200579	20428	16998	2052489	630214	124781
8	38703	601723	54061	43967		1890048	296581
9	103838	1805152	145346	116097			1196908
10	282148		395111	311499			
11	774526		1082614	846429			
12	2143981			2323555			

Tabelle 7-3: Erforderliche Datenmengen bei verschiedenen Wortlängen *L* für 1 % Genauigkeit und Alphabetgröße 4. Maße: wie Tabelle 7-1.

L	H_S, H_μ, H_R	H_G	$H_{M,d}$	$H_{M,c}$	C_{EM}	C_Γ	C_R
1	110	438	656		656	2164	1258
2	275	1748	1213	2001	5861	7350	3090
3	772	6984	2414	1097	37068	28089	8696
4	2351	27925	6538	3081	203438	111043	26447
5	7584	111687	20089	9395	1035107	442854	84699
6	25465	446730	65998	30327		1770096	282259
7	87975	1786900	225662	101847			966094
8	310390		791811	351882			3376335
9	1112750		2828243	1241538			

Tabelle 7-4. Erforderliche Datenmengen bei verschiedenen Wortlängen *L* für 5 % Genauigkeit und Alphabetgröße 2. Maße: wie Tabelle 7-1.

L	H_S, H_μ, H_R	H_G	$H_{M,d}$	$H_{M,c}$	C_{EM}	C_Γ	C_R
1	15	32	19	x	19	88	83
2	24	63	27	32	79	146	125
3	38	124	23	49	256	262	193
4	62	246	17	76	723	491	308
5	102	487	33	122	1887	949	504
6	174	969	60	202	4676	1863	844
7	301	1931	110	344	11176	3692	1444
8	531	3855	201	598	26020	7347	2510
9	948	7702	370	1056	59400	14656	4423
10	1713	15396	684	1890	133543	29274	7877
11	3120	30781	1272	3418	296598	58509	14145
12	5723	61550	2377	6231	652247	116978	25561
13	10559	123088	4456	11437	1422626	233914	46405
14	19570	246163	8380	21107		467785	84508
15	36413	492311	15796	39129		935527	154151
16	67973	984607	29834	72813		1871010	281226
17	127248	1969197	56443	135933			512227
18	238799		106942	254482			929386
19	449108		202886	477582			1674156
20	846252		385351	898198			
21	1597316		732682	1692488			
22			1394396				

Tabelle 7-5. Erforderliche Datenmengen bei verschiedenen Wortlängen *L* für 5 % Genauigkeit und Alphabetgröße 3. Maße: wie Tabelle 7-1.

L	H_S, H_μ, H_R	$H_{G,d}$	$H_{M,d}$	$H_{M,c}$	C_{EM}	C_Γ	C_R
1	19	59	63	x	63	232	166
2	40	174	108	60	415	576	330
3	88	518	174	120	1988	1606	710
4	207	1546	344	263	8261	4692	1625
5	505	4630	677	615	31751	13951	3902
6	1277	13880	279	1509	116239	41723	9711
7	3300	41628	764	3821	411759	125038	24715
8	8669	124868	2105	9887	1424501	374982	63990
9	23052	374588	5823	25993		1124811	171372
10	61866	1123745	16162	69141			456994
11	167232		44960	185580			1229603
12	454690		125311	501677			
13	1242205		349818	1364048			
14			977906				
15			2737071				

Tabelle 7-6. Erforderliche Datenmengen bei verschiedenen Wortlängen *L* für 5 % Genauigkeit und Alphabetgröße 4. Maße: wie Tabelle 7-1.

L	H_S, H_μ, H_R	H_G	$H_{M,d}$	$H_{M,c}$	C_{EM}	C_Γ	C_R
1	23	92	137	x	137	433	249
2	60	366	267	93	1188	1463	618
3	171	1457	578	237	7465	5579	1717
4	527	5818	1584	678	40879	22041	5174
5	1708	23258	4657	2099	207763	87885	16478
6	5717	93016	13844	6819	1008223	351258	54531
7	19563	372045	38782	22855		1404748	185438
8	67965	1488156	12617	78235			643267
9	238790		46237	271840			2264877
10	846242		169765	955135			
11	3019579		624284	3384944			
12			2298790				

8 Symbolverzeichnis

Symbol	Bedeutung	Seite
A	Alphabet von Symbolen	18
a_j	Buchstabe des Alphabets	18
$a*$	Lückensymbol	18
$[a,b]$	Wertebereich der Daten, Minimum a, Maximum b	18
C_ε	ε-Komplexität	59
C_{EM}	Effektive Maßkomplexität	52
C_Γ	Fluktuationskomplexität	54
C_{R}	Rényi-Komplexität	57
C_{V}	Varianz-Komplexität	62
D	Morph-Tiefe, Unterbaumtiefe zur Definition von Automatenzuständen	24
δ	Diskriminanz, maximaler Abstand der Übergangswahrscheinlichkeiten in Unterbäumen zur Definition äquivalenter Zustände eines Automaten	25
e	Kante, Verbindung zwischen Zuständen eines Automaten	25
h	Entropie der Quelle / des generierenden Prozesses	43
H_{G}	mittlerer Informationsgewinn, bedingte Entropie	44
H_{KS}	Kolmogorov-Sinai Entropie	43
H_{M}	mittlere wechselseitige Information	46
H_μ	metrische Entropie	42
H_{R}	Rényi-Entropie	39
H_{S}	Shannon-Entropie	37
H_{top}	topologische Entropie	39
$I(k)$	Transinformation zum Lag k	30
I_{A}	Algorithmische Information	50
λ	Alphabetgröße, $\lvert A\rvert$	18
L	Wortlänge	21
N	Anzahl der Daten / Zeitschritte	18
$n_{k,i}$	Anzahl der Wörter durch Knoten i aus Schicht k in Baumstruktur	23
n_K	Anzahl der Knoten eines Baumes	23
π_{i}	i-ter Partitionierungsparameter	19
p_i	relative Häufigkeit des i-ten Wortes bekannter Länge	21
$p_{L,i}$	relative Häufigkeit des i-ten Wortes der Länge L	21
$p_{i\to j}$	relative Übergangshäufigkeit von Wort i nach Wort j bei bekannter Länge	21
$p_{L,i\to j}$	relative Übergangshäufigkeit von L-Wort i nach L-Wort j	21
$p_{L,ij}$	rel. Häufigkeit für das aufeinanderfolgende Auftreten von L-Wort i und j	21
$r(k)$	Autokorrelation zum Lag k	29
s_i	Symbol aus dem Symbolsatz zum Zeitschritt i	18
S	Symbolsatz	18
x_i	Datenpunkt / Messwert zum Zeitpunkt i	18
X	Datensatz, Messreihe, Zeitreihe	18

9 Abbildungsverzeichnis

10 Literaturverzeichnis

Abramowitz, M; Stegun, IA (Hrsg.) (1984): *Pocketbook of mathematical functions.* Material selected by Michael Danos and Johann Rafelski. Abridged edition of Handbook of mathematical functions. Thun; Frankfurt am Main: Harri Deutsch.

Atmanspacher, H; Räth, C; Wiedenmann, G (1997): *Statistics and Meta-Statistics in the Concept of Complexity.* Physica A, 234: 819–829.

Badii, R; Finardi, M (1992): *Hierarchical resolution of power spectra.* Physica D, 58: 304–324.

Badii, R; Politi, A (1997): *Complexity: hierarchical structures and scaling in physics.* Cambridge: University Press.

Balatoni, J; Rényi, A (1956): *On the notion of entropy.* In: Pál Turán (Hrsg.): Selected Papers of Alfréd Rényi. Volume 1. Budapest: Akadémiai Kiadó, 1976: 558–586.

Bates, JE; Shepard, HK (1993): *Measuring complexity using information fluctuation.* Physics Letters A, 172: 416–425.

Beven, K (1996): *The limits of splitting: Hydrology*. The Science of the Total Environment, 183: 89–97.

Bittersohl, J; Lischeid, G (1995): *Hydrogeologie und Abflußverhalten im Einzugsgebiet Lehstenbach*. In: Manderscheid, B; Göttlein, A (Hrsg.): Wassereinzugsgebiet ‚Lehstenbach' – das BITÖK-Untersuchungsgebiet am Waldstein (Fichtelgebirge, NO-Bayern). Bayreuther Forum Ökologie. Band 18: 40–48.

Bormann, FH; Likens, GE (1979): *Pattern and Process in a Forested Ecosystem*. New York, Heidelberg, Berlin: Springer.

Bronštejn, IN; Semendjajew, KA; Musiol, G; Mühlig, H (1997): *Taschenbuch der Mathematik.* 3. überarb. und erw. Aufl. der Neubearb. Thun, Frankfurt am Main: Deutsch.

Chaitin, GJ (1987): *Algorithmic Information Theory*. Cambridge, New York, New Rochelle, Melbourne, Sydney: Cambridge University Press.

Chaitin, GJ (1990): *Information, Randomness and Incompleteness — Papers on Algorithmic Information Theory*. 2nd ed. Singapore, New Jersey, London, Hong Kong: World Scientific.

Crutchfield, JP (1991): *Reconstructing Language Hierarchies.* In: Atmanspacher, H; Scheingraber, H: Information dynamics. New York: Plenum Press.

Crutchfield, JP (1992): *Discovering Coherent Structures in Nonlinear Spatial Systems.* In: Brandt, A; Ramberg, S; Shlesinger, M: Nonlinear Ocean Waves. Singapore: World Scientific: 190–216.

Crutchfield, JP (1994a): *Observing Complexity and the Complexity of Observation.* In: Atmanspacher, H; Dalenoort, GJ (Hrsg.): Inside versus Outside. Series in Synergetics. Berlin: Springer: 235–272.

Crutchfield, JP (1994b): *The Calculi of Emergence: Computation, Dynamics, and Induction.* Physica D, 75: 11–54.

Crutchfield, JP; Young, K (1989): *Inferring Statistical Complexity*. Physical Review Letters, 63 (2): 105–108.

Crutchfield, JP; Packard, NH (1983): *Symbolic Dynamics of noisy chaos*. Physica D, 7: 201–223.

Diu, B; Guthmann, C; Lederer, D; Roulet, B (1994): *Grundlagen der statistischen Physik. Ein Lehrbuch mit Übungen.* Übers. aus dem Französischen von F. Marschner. Berlin, New York: Walter de Gruyter.

Duden (Wissenschaftlicher Rat der Dudenredaktion, Hrsg.) (1996): *Duden: Rechtschreibung der deutschen Sprache.* Band 1. 21. völlig neu bearb. und erw. Auflage. Mannheim, Leipzig, Wien, Zürich: Dudenverlag.

Dunne, T; Black, RG (1970): *Partial area contributions to storm runoff in a small New England watershed*. Water Resources Research, 6 (5): 1296–1311.

Ebeling, W (1997): *Prediction and entropy of nonlinear dynamical systems and symbolic sequences with LRO.* Physica D, 109: 42–52.

Ebeling, W; Neiman, A (1995): *Long-range correlations between letters and sentences in texts.* Physica A, 215: 233–241.

Ebeling, W; Pöschel, T; Albrecht, KF (1995): *Entropy, Transinformation and Word Distributions of Information-Carrying Sequences.* International Journal of Bifurcation and Chaos, 5 (1): 51–61.

Ebeling,W; Pöschel, T; Neiman, A (1996): *Entropy and compressibility of symbol sequences.* In: Toffoli, T; Biafore, M; Leao, J (eds.): Physics of Computation, Cambridge, MA: New England Complex Systems Institute.

Ebeling, W; Freund, J; Schweitzer, F (1998): *Komplexe Strukturen: Entropie und Information.* Stuttgart, Leipzig: Teubner.

Eckmann, JP; Kamphorst, SO; Ruelle, D (1987): Recurrence Plots of Dynamical Systems. Europhysics Letters, 4 (9): 973–977.

Farmer, JD (1982): *Information Dimension and the Probabilistic Structure of Chaos*. Zeitschrift für Naturforschung, 37a: 1304–1325.

Feigenbaum, MJ (1978): Quantitative Universality for a Class of Nonlinear Transformations. Journal of Statistical Physics, 19 (1): 25–52.

Feldman, DP; Crutchfield, JP (1998): *Measures of Statistical Complexity: Why?* Physics Letters A, 238 (4-5): 244–252.

Freeze, RA (1972): *Role of Subsurface Flow in Generating Surface Runoff: 2. Upstream Source Areas*. Water Resources Research, 8 (5): 1272–1283.

Fucks, W (1955): *Mathematische Analyse von Sprachelementen, Sprachstil und Sprachen*. In: Brandt, L (Hrsg.): Arbeitsgemeinschaft für Forschung des Landes Nordrhein-Westfalen. Heft 34a. Westdeutscher Verlag Köln und Opladen.

Gell-Mann, M (1998): *Das Quark und der Jaguar*. Vom Einfachen zum Komplexen. Die Suche nach einer neuen Erklärung der Welt. 2. Auflage. München: Piper.

Ghilardi, P; Rosso, R (1990): *Comment on „Chaos in Rainfall" by I. Rodriguez-Iturbe et al.*. Water Resources Research, 26 (8): 1837–1839.

Goudie, A (1994): *Mensch und Umwelt. Eine Einführung*. aus dem Engl. übers. u. bearb. von C. Niemitz. Originaltitel: The Human Impact on the Natural Environment. Fourth Edition. Heidelberg, Berlin, Oxford: Spektrum Akademischer Verlag.

Grassberger, P (1986): *Toward a Quantitative Theory of Self-Generated Complexity.* International Journal of Theoretical Physics, 25 (9): 907–938.

Grassberger, P (1988): *Complexity and Forecasting in Dynamical Systems.* In: Peliti, L; Vulpiani, A (Hrsg.): Measures of Complexity. Proceedings of the Conference, held in Rome, September 30 – October 2, 1987. In: Araki, H; Ehlers, J; Hepp, K; Kippenhahn, R; Weidenmüller, HA; Wess, J; Zittartz: Lecture Notes in Physics 314. Berlin, Heidelberg, New York, London, Paris, Tokyo: Springer: 1–21.

Große, I (1996): *Estimating Entropies from Finite Samples*. In: Freund, JA (Hrsg.) Dynamik, Evolution, Strukturen: Nichtlineare Dynamik und Statistik komplexer Strukturen. 1. Aufl. Berlin: Köster: 181–190.

Hartung, J; Elpelt, B; Klösener, K-H (1998): *Statistik: Lehr- und Handbuch der angewandten Statistik; mit zahlreichen, vollständig durchgerechneten Beispielen.* 11. durchgesehene Auflage. München, Wien: Oldenbourg.

Hauhs, M (1985): *Wasser- und Stoffhaushalt im Einzugsgebiet der Langen Bramke (Harz).* Dissertation. Universität Göttigen. Berichte des Forschungszentrums Waldökosysteme/Waldsterben, Band 17.

Hauhs, M; Lange, H (1996a): *Das Problem der Prozeßidentifikation in Waldökosystemen am Beispiel Wassertransport*. Internationales Hochschulinstitut Zittau. IHI-Schriften 2.

Hauhs, M; Lange, H (1996b): *Ecosystem dynamics viewed from an endoperspective*. The Science of the Total Environment, 183: 125–136.

Heindl, B; Ostendorf, B; Köstner, B (1995): *Lage und forstliche Charakterisierung des Einzugsgebietes Lehstenbach*. In: Manderscheid, B; Göttlein, A (Hrsg.): Wassereinzugsgebiet ‚Lehstenbach' – das BITÖK-Untersuchungsgebiet am Waldstein (Fichtelgebirge, NO-Bayern). Bayreuther Forum Ökologie. Band 18: 7–14.

Herzel, H; Große, I (1995): *Measuring correlations in symbol sequences*. Physica A, 216: 518–542.

Herzel, H; Schmitt, AO; Ebeling, W (1994): *Finite Sample Effects in Sequence Analysis.* Chaos, Solitons & Fractals; 4 (1): 97–113.

Hipel, KW; McLeod, AI (1994): *Time series modelling of water resources and envirnonmental systems.* Amsterdam, London, New York, Tokyo: Elsevier.

Holste, D; Große, I; Herzel, H (1998): *Bayes' Estimators of Generalized Entropies*. J. Phys. A: Math. Gen., 31: 2551–2566.

Honerkamp, J (1994): *Stochastic Dynamic Systems. Concepts, Numerical Methods, Data Analysis.* Translated by Katja Lindenberg. New York, Weinheim, Cambridge: VCH.

Hopcroft, JE; Ullman, JD (1990): *Einführung in die Automatentheorie, formale Sprachen, und Komplexitätstheorie.* 2. Aufl. Titel der engl. Originalausgabe: "Introduction to automata theory, languages and computation". Bonn; München; Reading, Mass.; Menlo Park, Calif.; New York; Don Mills, Ontario; Wokingham, England; Amsterdam; Sydney; Singapore; Tokyo; Madrid; San Juan: Addison-Wesley.

Horgan, J (1997): *An den Grenzen des Wissens. Siegeszug und Dilemma der Naturwissenschaften*. Dt. Übersetzung von: "The End of Science. Facing the Limits of Knowledge in the Twilight of the Scientific Age." von T. Schmidt. München: Luchterhand.

Horowitz, E; Sahni, S (1981): *Algorithmen: Entwurf und Analyse.* Dt. Übersetzung d. am. Ausgabe: "Fundamentals of computer algorithms" von M. Czerwinski. Berlin, Heidelberg, New York: Springer.

Jakeman, AJ; Hornberger, GM (1993): *How much Complexity Is Warranted in an Rainfall-Runoff Modell?* Water Resources Research, 29 (8): 2637–2649.

Janssen, PHM; Heuberger, PSC (1995): *Calibration of process-oriented models*. Ecological Modelling, 83: 55–66.

Jaynes, ET (1957): *Information Theory and Statistical Mechanics*. Physical Review, 106 (4): 620–630.

Jetschke, G (1989): *Mathematik der Selbstorganisation: Qualitative Theorie nichtlinearer dynamischer Systeme und gleichgewichtsferner Strukturen in Physik, Chemie und Biologie*. Braunschweig, Wiesbaden: Vieweg.

Kapur, JN (1994): *Measures of Information and their Application*. New York, Chichester, Brisbane, Toronto, Singapore: Wiley.

Kaspar, F; Schuster, HG (1987): *Easily calculable measure for the complexity of spatiotemporal patterns.* Physical Review A, 36 (2): 842–848.

Kernighan, BW; Ritchie, DM (1990): *Programmieren in C: mit dem C-reference-Manual in deutscher Sprache.* Dt. Ausg. von A. T. Schreiner u. E. Janich. 2. Ausg., ANSI C. München, Wien: Hanser; London: Prentice-Hall International.

Khinchin, AI (1957): *Mathematical Foundations of Information Theory*. Translated by R. A. Silverman and M. D. Friedman. New York: Dover Publications.

Kimmins, JP (1997): *Forest Ecology. A Foundation for Sustainable Management*. 2. Ed. New Jersey: Prentice-Hall.

Kolmogorov, AN (1965): *Three Approaches to the quantitative definition of information.* Problems of information transmission, 1 (1): 1–7.

Kurths, J; Witt, A (1994): *On Complexity Measures.* World Futures, 42: 177–192.

Kurths, J; Schwarz, U (1995): *On Nonlinear Signal Processing*. ESA SP–371: 15–21.

Kurths, J; Voss, A; Witt, A; Saparin, P; Kleiner, HJ; Wessel, N (1995): *Quantitative analysis of heart rate variability.* Chaos, 5 (1): 88–94.

Kurths, J; Schwarz, U; Witt, A; Krampe, RT; Abel, M (1996): *Measures of Complexity in Signal Analysis.* In: Kratz, RA (Hrsg.): Chaotic, Fractal, and Nonlinear Signal Processing. AIP Conference Proceedings 375: 33–54.

Lange, H (1999): *Charakterisierung ökosystemarer Zeitreihen mit nichtlinearen Methoden*. Habilitationsschrift. Universität Bayreuth. Fakultät für Biologie, Chemie und Geowissenschften.

Lange, H; Hauhs, M; Romahn, C (1997): *Classification of terrestrial ecosystems with complexity measures.* In: Schweitzer, F. (Hrsg.): Self-Organization of Complex Structures: From Individual to Collective Dynamics. London: Gordon and Breach: 293–306.

Lange, H; Newig, J; Wolf, F (1998): *Comparison of complexity measures for time series from ecosystem research.* In: Kastner-Maresch,A; Kurth, W; Sonntag, M; Breckling, B (Hrsg.): Individual-based structural and functional models in ecology. Contributions of a workshop helt at the University of Bayreuth. Bayreuth: Bayreuther Forum Ökologie. Band 52: 99–116.

Lempel, A; Ziv, J (1976): *On the Complexity of Finite Sequences*. IEEE Transactions on information theory, IT-22 (1): 75–81.

Li, W (1990): *Mutual Information Functions Versus Correlation Functions.* Journal of Statistical Physics, 60: 823–837.

Likens, GE; Bormann, FH (1995): *Biogeochemistry of a Forested Ecosystem*. 2nd Edition. New York, Berlin, Heidelberg, London, Paris, Tokyo, Hong Kong, Barcelona, Budapest: Springer.

Lin, J (1991): *Divergence Measures Based on the Shannon Entropy*. IEEE Transactions on Information Theory, 37 (1): 145–151.

Lind, M; Marcus, B (1995): *An introduction to symbolic dynamics and coding.* Cambridge University Press.

Lischeid, G; Gerstberger, P (1997): *The Steinkreuz catchment as a BITÖK main investigation site in the Steigerwald region: experimental setup and first results*. In: BITÖK (Hrsg.): Forschungsbericht 1996. Bayreuther Forum Ökologie. Band 41: 73–81.

Lükewille, A; Manø, S; Tørseth, K (1998): *Overvåking av langtransportert forurenset luft og nedbør. Atmosfærisk tilførsel, 1997*. Norsk institutt for luftforskning (NILU). OR 33/98.

Manderscheid, B; Göttlein, A (Hrsg.) (1995): *Wassereinzugsgebiet ‚Lehstenbach' – das BITÖK-Untersuchungsgebiet am Waldstein (Fichtelgebirge, NO-Bayern).* Bayreuther Forum Ökologie. Band 18.

May, RM (1976): *Simple mathematical models with very complicated dynamics*. Nature 261: 459–467.

Mitchell, A (1979): *Die Wald- und Parkbäume Europas*. Ein Bestimmungsbuch für Dendrologen und Naturfreunde. Übers. u. bearb. von G. Krüssmann. 2. Aufl. Hamburg, Berlin: Parey.

Müller, D-I; Wohlfeil, IC; Christophersen, N; Hauhs, M; Seip, HM (1993): *Chemical reactiveness of soil water pathways investigated by point source injections of chloride in a peat bog at Birkenes*. Journal of Hydrology, 144: 101–125.

Mulder, J; Christophersen, N; Hauhs, M; Vogt, RD; Andersen, S; Andersen, DO (1990): *Water Flow Paths and Hydrochemical Controls in the Birkenes Catchment as Inferred From a Rainstorm High in Seasalts*. Water Resources Research, 26 (4): 611–622.

Mulder, J; Pijpers, M; Christophersen, N (1991): *Water Flow Paths and the Spatial Distribution of Soils and Exchangeable Cations in an Acid Rain-Impacted and a Pristine Catchment in Norway*. Water Resources Research, 27 (11): 2919–2928.

Newig, J (1998): *Charakterisierung von Wassereinzugsgebieten durch Komplexitätstheorie und nichtlineare Zeitreihenanalyse.* Diplomarbeit. Universität Bayreuth, Fakultät für Biologie, Chemie und Geowissenschaften.

Pandey, G; Lovejoy, S; Schertzer, D (1998): *Multifractal analysis of daily river flows including extremes for basins of five to two million square kilometres, one day to 75 years*. Journal of Hydrology, 208: 62–81.

Peters, K; Gerchau, J (1995): *Klima und luftchemische Situation des Fichtelgebirges unter besonderer Berücksichtigung des Einzugsgebietes Lehstenbach*. In: Manderscheid, B; Göttlein, A (Hrsg.): Wassereinzugsgebiet ‚Lehstenbach' – das BITÖK-Untersuchungsgebiet am Waldstein (Fichtelgebirge, NO-Bayern). Bayreuther Forum Ökologie. Band 18: 15–39.

Pöschel, T (1996): *Kann die Entropie von Sequenzen vermittelst der Kompressibilität gemessen werden?*. In: Freund, JA (Hrsg.) Dynamik, Evolution, Strukturen: Nichtlineare Dynamik und Statistik komplexer Strukturen. 1. Aufl. Berlin: Köster: 191–201.

Pöschel, T; Ebeling, W; Rose, H (1995): *Guessing Probability Distributions from Small Samples*. Journal for Statistical Physics, 80: 1443–1452.

Post, DA; Grant, DE; Jones, JA (1998): *New Developments in Ecological Hydrology Expand Research Opportunities*. EOS 79/43: 517+526.

Press, WH; Teukolsky, SA; Vetterling, WT; Flannery, BP (1992): *Numerical recipes in C: the art of scientific computing.* 2. Aufl. Cambridge University Press.

Projektträger Biologie, Energie, Umwelt des Bundesministeriums für Bildung, Wissenschaft, Forschung und Technologie Forschungszentrum Jülich GmbH (PT BEO); Projektträger Umwelt- und Klimaforschung des Bundesministeriums für Bildung, Wissenschaft, Forschung und Technologie GSF Forschungszentrum für Umwelt und Gesundheit München (PT UKF) (1997): *Jahresbericht 1997. Umweltforschung – Ökologische Forschung*. Eggenstein-Leopoldshafen: Fachinformationszentrum Karlsruhe.

Rateitschak, K; Freund, J; Ebeling, W (1995): *Entropy of Sequences Generated by nonlinear Processes: The logistic map.* In: Shiner, JS (Hrsg.): Entropy and Entropy Generation. Fundamentals and Applications. Dordrecht, Boston, London: Kluwer Academic Publishers: 11–26.

Rényi, A (1960): *Some Fundamental Questions of Information Theory.* In: Pál Turán (Hrsg.): Selected Papers of Alfréd Rényi. Vol. 2: 1956 – 1961. Budapest: Akadémiai Kiadó, 1976: 526–552.

Rényi, A (1961): *On measures of entropy and information.* In: Pál Turán (Hrsg.): Selected Papers of Alfréd Rényi. Vol. 2: 1956 – 1961. Budapest: Akadémiai Kiadó, 1976: 565–580.

Rényi, A (1976): *Wahrscheinlichkeitsrechnung. Mit einem Anhang über Informationstheorie.* 5. Auflage. Berlin: VEB Deutscher Verlag der Wissenschaften.

Rodriguez-Iturbe, I; Gupta, VK; Waymire, E (1984): *Scale Considerations in the Modeling of Temporal Rainfall*. Water Resources Research, 20 (11): 1611–1619.

Rodriguez-Iturbe, I; De Power, BF; Sharifi, MB; Georgakakos, KP (1989): *Chaos in Rainfall*. Water Resources Research, 25 (7): 1667–1675.

Rodriguez-Iturbe, I; De Power, BF; Sharifi, MB; Georgakakos, KP (1990): *Reply*. Water Resources Research, 26 (8): 1841–1842.

Romahn, C (1996): *Untersuchungen zur Komplexität von Zeitreihen mit informationstheoretischen und thermodynamischen Methoden an Beispielen aus der Ökosystemforschung*. Diplomarbeit. Universität Bayreuth. Fakultät für Mathematik und Physik.

Schlittgen, R; Streitberg, BHJ (1994): *Zeitreihenanalyse*. 5. völlig überarb. und erw. Aufl. München, Wien: Oldenbourg.

Schmidt, S (1997): *Zusammenhang von Wasser- und Stoffhaushalt in der Langen Bramke — Vergleich unterschiedlicher zeitlicher und räumlicher Maßstäbe.* Dissertation. Universität Göttingen. Berichte des Forschungszentrums Waldökosysteme, Reihe A, Band 146.

Schmitt, AO; Herzel, H (1997): *Estimating the Entropy of DNA Sequences*. Journal for Theoretical Biology, 1888: 369–377.

Schmitt, AO; Herzel, H; Ebeling, W (1993): *A New Method to Calculate Higher-Order Entropies from Finite Samples*. Europhysics Letters, 23 (5): 303–309.

Schroeder, M (1991): *Fractals, Chaos, Power Laws. Minutes from an infinite Paradise*. New York: Freeman and Company.

Schwarz, U; Benz, AO; Kurths, J; Witt, A (1993): *Analysis of solar spike events by means of symbolic dynamics methods.* Astronomy and Astrophysics, 277: 215–224.

Sedgewick, R (1988): *Algorithms.* 2. Aufl. Reading, Massachusetts; Menlo Park, California; New York, Don Mills, Ontario; Wokingham, England; Amsterdam; Bonn; Sydney; Singapore; Tokyo; Madrid; San Juan: Addison-Wesley.

Shannon, CE (1948): *A Mathematical Theory of Communication*. Bell System Technical Journal, 27: 379–423.

Shannon, CE (1976): *Die mathematische Theorie der Kommunikation.* In: Shannon, CE; Weaver, W: Mathematische Grundlagen der Informationstheorie. München: Oldenbourg: 41–143.

Stoer, J (1994): *Numerische Mathematik 1: Eine Einführung — unter Berücksichtigung von Vorlesungen von F. L. Bauer*. 7. neubearb. und erw. Aufl. Berlin, Heidelberg, New York, London, Paris, Tokyo, Hong Kong, Barcelona, Budapest: Springer.

Striebel, T (1994): *Konzentrationen und physikochemisches Verhalten von Schwermetallen und Hauptionen in Regenabflüssen städtischer Straßen*. Dissertation, Fakultät für Biologie, Chemie und Geowissenschaften der Universität Bayreuth. Aachen: Shaker.

Stroustrup, B (1995): *Die C++ Programmiersprache*. Dt. Übersetzung von „The C++ programming language". 2. überarb. Auflage. 7 unveränd. Nachdruck. Bonn, München, Paris [u. a.]: Addison-Wesley.

Tolman, RC (1967): *The principles of statistical mechanics.* Oxford: University Press.

Trulla, LL; Giuliani, A; Zbilut, JP; Webber Jr., CL (1996): *Recurrence quantification analysis of the logistic equation with transients.* Physics Letters A, 223: 255–260.

United States Departement of Agriculture (USDA), Forest Service (Hrsg.) (1996): *Hubbard Brook Ecosystem Study. Site Description and Research Activities.* Northeastern Forest Experiment Station NE-INF-96-96R. Second Edition.

Valdes, JB; Rodriguez-Iturbe, I; Gupta, VK (1985): *Approximations of Temporal Rainfall From a Multidimensional Model*. Water Resources Research, 21 (8): 1259–1270.

Wackerbauer, R; Witt, A; Atmanspacher, H; Kurths, J; Scheingraber, H (1994): *A Comparative Classification of Complexity Measures.* Chaos, Solitons & Fractals, 4: 133–173.

Wahl, KL; Thomas, WO; Hirsch, RM (1995): *Stream-Gaging Program of the U.S. Geological Survey*. U.S. Geological Survey Circular 1123. Reston, Virginia. URL: http://h2o.usgs.gov/public/pubs/circ1123/index.html

Waymire, E; Gupta, VK; Rodriguez-Iturbe, I (1984): *A Spectral Theory of Rainfall Intensity at the Meso-β Scale*. Water Resources Research, 20 (10): 1453–1465.

Weaver, W (1976): *Ein aktueller Beitrag zur mathematischen Theorie der Kommunikation.* In: Shannon, CE; Weaver, W: Mathematische Grundlagen der Informationstheorie. München: Oldenbourg: 11–39.

Wei, WWS (1990): *Time Series Analysis. Univariate and Multivariate Methods.* Redwood City, Menlo Park, Reading, New York, Amsterdam, Don Mills, Sydney, Bonn, Madrid, Singapore, Tokyo, San Juan: Addison-Wesley.

Wiseman R (1997): *Das Word 97 Buch.* 2. Auflage. Düsseldorf, San Francisco, Paris, Soest (NL): Sybex.

Wilde, RE; Singh, S (1998): *Statistical Mechanics. Fundamentals and Modern Applications.* New York, Chichester, Weinheim, Brisbane, Singapore, Toronto: John Wiley & Sons, Inc.

Witt, A (1996): *Komplexitätsmaße und ihre Anwendungen*. Mathematisch-Naturwissenschaftliche Fakultät der Universität Potsdam. Dissertation.

Witt, A; Kurths, J; Krause, F; Fischer, K (1994): *On the validity of a model for the reversals of the Earth's magnetic field.* Geophysics Astrophysics Fluid Dynamics, 77: 79–91.

Witt, A; Kurths, J; Pikovsky, A (1998): *Testing Stationarity in Time Series*. Physical Review E, 58: 1800–1810.

Wolf, F; Lange, H; Hauhs, M (1997): *Ecosystem analysis by means of complexity theory*. In: BITÖK (Hrsg.): BITÖK Forschungsbericht 1996. Bayreuther Forum Ökologie, Band 41: 184–187.

Wolf, F; Lange, H; Hauhs, M (1998): *Ökosystem-Analysen mit komplexitätstheoretischen Ansätzen*. In: Matzner, E (Hrsg.): BITÖK Forschungsbericht 1995-97. Bayreuther Forum Ökologie, Band 56: 212–213.

Wolfram, S (1984): *Computation Theory of Cellular Automata.* Communications in Mathematical Physics, 95: 15–57.

Wolfram, S (1985): *Origins of Randomness in Physical Systems.* Physical Review Letters, 55 (5): 449–452.

Wolfram, S (1996): *The Mathematica book.* 3. Aufl. Cambridge: University Press.

Wolkenstein, M (1990): *Entropie und Information.* Übers. aus d. Russ.: H. Müller. Bearb. d. deutschsprachigen Ausg.: W. Ebeling. Deutsch Taschenbücher, Band 67. Thun, Frankfurt am Main: Verlag Harri Deutsch.

Ziv, J; Lempel, A (1978): *Compression of Individual Sequences via Variable-Rate Coding.* IEEE Transactions on information theory, IT-24 (5): 530–536.

Zurek, WH (Hrsg.) (1990): *Complexity, Entropy and the Physics of Information*. The Proceedings of the 1988 Workshop on Complexity, Entropy, and the Physics of Information held May – June, 1989 in Santa Fe, New Mexico. Redwood City, Menlo Park, California; Reading, Massachusetts; New York; Don Mills, Ontario; Wokingham, United Kingdom; Amsterdam; Bonn; Sydney; Singapore; Tokyo; Madrid; San Juan: Addison-Wesley.

11 Danksagung

Bei Prof. Michael Hauhs bedanke ich mich für die Möglichkeit als Mathematiker an seinem Lehrstuhl für Ökologische Modellbildung promovieren zu können. Die vielen von ihm verfolgten alternativen Ansätze haben mir einen nachhaltigen Eindruck von den Möglichkeiten und Grenzen der Modellierung vermittelt.

Dr. Holger Lange hat diese Arbeit im Rahmen seiner Habilitation: „Charakterisierung ökosystemarer Zeitreihen mit nichtlinearen Methoden“ betreut. Er war jederzeit bereit auf Fragen und Diskussionen einzugehen.

Dr. Gunnar Lischeid hat mich in hydrogeologischen Fragen beraten. Als Zimmergenosse danke ich ihm außerdem für die angenehme Arbeitsatmosphäre.

Die Untersuchungen zur Stationarität (Abschnitt 3.3) wurden durch die lebhafte Diskussion von Dr. Hans Piehler angeregt.

Für seine Hilfe bei Problemen während der Programmierung von SYMDYN bedanke ich mich bei Walter Dörwald.

Ich möchte mich ausdrücklich bei allen in Kapitel 4 genannten Personen, Instituten und deren Mitarbeitern für die Erhebung und Bereitstellung des verwendeten Datenmaterials bedanken. Ohne ihren Einsatz wäre diese Arbeit nicht möglich gewesen.

An der Durchsicht des Manuskriptes war Thorsten Schmid beteiligt. Für weitere Unterstützung danke ich meiner Frau, Claudia.

Dieses Projekt wurde teilweise durch das Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF) unter der Nummer PT BEO 51 – 0339476B, Projekt S11, gefördert.

Erklärung

Hiermit erkläre ich, dass ich diese Dissertation selbständig verfasst und keine anderen als die von mir angegebenen Quellen und Hilfsmittel benutzt habe.

Ich erkläre weiterhin, dass ich nicht diese oder eine gleichartige Doktorprüfung an einer anderen Hochschule endgültig nicht bestanden habe.

Bayreuth, den