FLOOD MITIGATION, MODEL UNCERTAINTY AND PROCESS DIAGNOSTICS

BRIDGING THE GAP BETWEEN OPERATIONAL PRACTICE AND RESEARCH

SIMON P. SEIBERT

Institut für Wasser und Gewässerentwicklung, Bereich Hydrologie
Fakultät für Bauingenieur-, Geo- und Umweltwissenschaften
Karlsruher Institut für Technologie (KIT)

Karlsruhe, 2016

FLOOD MITIGATION, MODEL UNCERTAINTY AND PROCESS DIAGNOSTICS

BRIDGING THE GAP BETWEEN OPERATIONAL PRACTICE AND RESEARCH

Zur Erlangung des akademischen Grades eines

DOKTOR-INGENIEURS
(Dr.-Ing.)

von der Fakultät für
Bauingenieur-, Geo- und Umweltwissenschaften
des Karlsruher Instituts für Technologie (KIT)
genehmigte

DISSERTATION

von
M.Sc. Ing.-Ök. Simon P. Seibert
aus Nürnberg

Tag der mündlichen Prüfung: 28. Juli 2016

REFERENT: Prof. Dr.-Ing. Erwin Zehe
KORREFERENT: Dr.-Ing. Uwe Ehret
KORREFERENT: Dr. habil. Laurent Pfister

Karlsruhe, 2016

To all those who aspire to develop but aren't able to achieve,
to those who are held back by limitations that are beyond them,
to those who are subject to conditions that cramp their lives.

CONTENTS

I INTRODUCTION 1

1 INTRODUCTION 3
- 1.1 Gaining new knowledge 5
- 1.2 At the interface of science and practice 7
- 1.3 Closeness on the ordinate and on the abscissa 8
- 1.4 Understanding the nature of runoff production 9

II THE POTENTIAL OF COORDINATED RESERVOIR OPERATION FOR FLOOD MITIGATION IN LARGE BASINS 13

2 PART 2: FLOOD MITIGATION IN LARGE BASINS 15
- 2.1 Introduction 16
- 2.2 Materials and Methods 18
  - 2.2.1 Study region 18
  - 2.2.2 Models 19
  - 2.2.3 Data 22
  - 2.2.4 Evaluation of model performance 23
  - 2.2.5 Reservoir impact estimation 27
- 2.3 Results and discussion 30
  - 2.3.1 Hydrological simulations 30
  - 2.3.2 Coupled hydrological and hydrodynamic simulations 32
  - 2.3.3 Grade-based evaluation of model performance 34
  - 2.3.4 Reservoir impact assessment 37
- 2.4 Summary and conclusions 42
  - 2.4.1 Evaluation of model performance 42
  - 2.4.2 The potential of coupled hydrological and hydrodynamic simulations 43
  - 2.4.3 The potential of coordinated reservoir operation for regional flood mitigation 43

III DISENTANGLING TIMING AND AMPLITUDE ERRORS IN STREAMFLOW SIMULATIONS 45

3 PART 3: TIMING AND AMPLITUDE UNCERTAINTIES 47
- 3.1 Introduction 48
  - 3.1.1 Single and multiple criteria for hydrograph evaluation 48
  - 3.1.2 Uncertainty assessment and model diagnostics – learning from model deficiencies 49
- 3.2 Series Distance – concept and modifications 51
  - 3.2.1 Hydrograph preprocessing 52
  - 3.2.2 Identification and pairing of events 53
  - 3.2.3 Pattern matching: identification, matching, and coarse-graining of segments 54
  - 3.2.4 Modifications in the SD error model 59

3.3 Error dressing: A heuristic for the construction of uncertainty ranges 60
3.3.1 The one-dimensional case 61
3.3.2 The two-dimensional case 62
3.4 Case study 64
3.4.1 Data and site properties 64
3.4.2 Conceptual setup 64
3.4.3 Evaluation of deterministic uncertainty ranges 65
3.5 Results and discussion 66
3.5.1 Potential and limitations of the core SD concept 66
3.5.2 Potential and limitations of the error dressing method 69
3.5.3 Case study results 70
3.6 Conclusions 76

IV EXPLORING THE INTERPLAY BETWEEN STATE, STRUCTURE AND RUNOFF BEHAVIOUR OF LOWER MESOSCALE CATCHMENTS 79

4 PART 4: INTERPLAY AMONG STRUCTURE, STORAGE AND RUNOFF 81
4.1 Introduction 82
4.1.1 Hydrological similarity as weak form of causality 82
4.1.2 Challenges in defining structural similarity at the lower mesoscale 82
4.1.3 Storage estimators and state estimators - how to normalize and how to achieve coherence? 83
4.1.4 Dimensionless response measures for and beyond capacity controlled runoff formation 84
4.1.5 Objectives and research questions 85
4.2 Conceptual framework and candidate diagnostics 86
4.2.1 Requirements of functional diagnostics 86
4.2.2 Candidate storage, response and intensity estimators for baseflow, runoff events and the seasonal water balance 88
4.3 Study area and dataset 92
4.3.1 Data quality and selection of headwater catchments 92
4.3.2 Landscape setting and perceptual models of runoff generation 94
4.4 Results 97
4.4.1 Storage and structure control on baseflow generation 97
4.4.2 Storage control on rainfall-runoff response 98
4.4.3 Seasonal interplay of storage and release 102
4.4.4 Intensity controlled runoff formation 107
4.5 Discussion and conclusions 111
4.5.1 Normalized double mass curves discriminating seasonal runoff behaviour 111

4.5.2 Pre-event discharge as best predictor for capacity controlled runoff production 112
4.5.3 Heterogeneous performance of storage-baseflow relations 113
4.5.4 "Edge filtering" of low passed data to detect high frequent runoff processes 114
4.5.5 Conclusion and Outlook 114

V SYNTHESIS AND DISCUSSION 117

5 PART 5: SYNTHESIS AND DISCUSSION 119
5.1 Summary 119
5.2 Discussion and outlook 124
5.2.1 New avenues in assessing closeness 124
5.2.2 Improving the understanding of runoff production at different time scales 126
5.3 Synthesis 128

VI APPENDIX 131

A APPENDIX 133
A.1 Appendix of Part II 133
A.1.1 Hydrological processes and man-made water regulation issues in the Bavarian part of the Danube basin 133
A.1.2 Structure of the LARSIM model 136
A.2 Appendix of Part IV 137
A.2.1 Physiographic site properties 137
A.2.2 Re-scaling and consistency of integrative storage measures 140
A.2.3 Automated delineation of rainfall driven events 140
A.2.4 Linkage between site identifiers and gauge names 142
A.2.5 Normalized double mass curves of remaining sites 143

VII BIBLIOGRAPHY AND BACK MATTER 147

BIBLIOGRAPHY 149
OWN PUBLICATIONS 173
ACKNOWLEDGMENTS 175
DECLARATION 177
COLOPHON 179

LIST OF FIGURES

Figure 1.1 Learning cycle in hydrological modelling 6
Figure 2.1 The Danube catchment in southern Germany 18
Figure 2.2 Return periods of recent historic floods 23
Figure 2.3 The Peak-Level-Certainty criterion 25
Figure 2.4 Simulation errors along the Danube 31
Figure 2.5 Hydrological modeling results 32
Figure 2.6 Comparison of hydrological and hydrodynamic results 33
Figure 2.7 Impact of boundary conditions on hydrodynamic simulations 34
Figure 2.8 Histograms of different performance statistics 35
Figure 2.9 Spatial patterns of model performance 37
Figure 2.10 Operation of the Forggensee reservoir 38
Figure 2.11 Local and regional impact of the Forggensee reservoir 39
Figure 2.12 Maximum possible water level reductions 41
Figure 2.13 Specific reservoir retention volumes 41
Figure 3.1 The Series Distance concept 52
Figure 3.2 Time-ordered matching of hydrograph segments 57
Figure 3.3 The coarse-graining scheme 59
Figure 3.4 The error dressing concept 62
Figure 3.5 Coarse-graining using equal weights 68
Figure 3.6 Case study results: Error distributions 71
Figure 3.7 Comparison of uncertainty envelopes 73
Figure 3.8 Vertical and horizontal error bars 75
Figure 4.1 Headwater catchments in the Bavarian part of the Danube catchment 93
Figure 4.2 Regime curves of precipitation, discharge and evapotranspiration 96
Figure 4.3 Empirical storage-baseflow relationships 98
Figure 4.4 Impact of storage on event-runoff coefficients 101
Figure 4.5 Normalized double mass curves 103
Figure 4.6 Normalized triple mass curves 106
Figure 4.7 Evidence of intensity controlled runoff formation mechanisms 107
Figure 4.8 Diagnostics for the detection of intensity controlled runoff mechanisms 108
Figure 4.9 Explorative modelling of intensity controlled conditions 110
Figure A.1 LARSIM structure 136
Figure A.2 Physiographic catchment properties 138
Figure A.3 Soil properties of the headwater catchments 139
Figure A.4 Coherent normalisation of dynamic storage 140
Figure A.5 Automated detection of rainfall-runoff events 141
Figure A.6 Double mass curve appendix (1/3) 143

Figure A.7 Double mass curve appendix (2/3) 144
Figure A.8 Double mass curve appendix (3/3) 145

LIST OF TABLES

Table 2.1 Reservoirs in the Danube basin 19
Table 2.2 Properties of the hydrological models 20
Table 2.3 Conversion table: NASH into grade levels 26
Table 2.4 Conversion table: PTDF and PLCY into grade levels 27
Table 2.5 Event specific modelling results 30
Table 3.1 The weighting factors' impact on the coarse-graining procedure 67
Table 3.2 Statistics of the derived error distributions 72
Table 3.3 Statistics of the uncertainty envelope 74
Table 4.1 Impact of storage on runoff response during radiation-driven conditions 99
Table 4.2 Impact of storage on runoff response during rainfall-driven conditions 100
Table 4.3 Seasonal runoff coefficients and inter-annual variations 104
Table A.1 Linkage between site identifiers (IDs) and gauge names 142

ABSTRACT

My dissertation is about *flood mitigation*, *model uncertainty* and *process diagnostics*. I develop and apply methods relevant for *reservoir operation* and *flood forecasting*, I introduce *pattern matching* procedures for streamflow time series and I analyze a comprehensive environmental dataset with regard to catchment *runoff production*.

Part I deals with flood mitigation in large-scale river basins by means of *coupled hydrological-hydrodynamic* modelling and *reservoir operation*. Here, I assess the hypothesis that flood protection reservoirs can, in addition to *local* points of interest, be operated at distant locations to improve *regional* flood mitigation. As a case study I use three major floods in the Bavarian part of the Danube basin (45.000 km^2) and nine larger reservoirs situated there (total retention volume $\geqslant 127 \cdot 10^6$ m^3). After identifying reservoirs which do have a *regional* impact I assess whether *regional* reservoir operation strategies are conformable with *local* reservoir operation strategies. The latter protect the direct downstream vicinity against flooding and their protection must not change for the worse. Furthermore, I evaluate model precision and relate it to reservoir impact. For the study site I find that model accuracy is satisfactory on average, although timing issues make the precise operation of reservoirs on *regional* locations partly challenging. I also find that only one out of nine reservoirs has a significant potential to impact the water levels at *regional* locations. While these findings are specific to the selected study site, the generally valid finding is that reservoir operation strategies optimized for *local* and *regional* flood mitigation are not necessarily mutually exclusive. Sometimes they can, due to temporal offsets, be pursued simultaneously. I conclude that *regional* operation of reservoirs in large-scale river basins can have a significant potential to improve flood mitigation. The individual potential of a reservoir is however case specific and requires accurate models and knowledge of the associated (timing) uncertainties.

Part II of my dissertation is on the assessment of model uncertainty. Here, the focus is on the simultaneous evaluation of timing and magnitude errors in streamflow simulations. The study was motivated by the fact that timing (*horizontal*) errors are rarely considered in hydrological streamflow simulations, though they are highly relevant e.g. for the operation of flood protection reservoirs. In this part I introduce improvements of the *Series Distance* (SD) approach which emulates *visual hydrograph comparison*. The latter is a powerful though complex evaluation method which rests on the *meaningful* comparison of observed and simulated streamflow time series. Fur this purpose SD distinguishes different flow conditions (periods of low-flow and periods of rise and recession in rainfall-runoff events) and determines the distance of two hydrographs not between points of equal time, but between points that are considered *hydrologically similar*. This is

achieved by means of a *pattern matching* procedure. The improvements comprise an automated procedure to emulate visual *coarse-graining*, i.e. the determination of an optimal level of generalization when comparing two hydrographs, a *scaled error model*, and *error dressing*, a concept to construct two-dimensional uncertainty ranges around deterministic simulations or forecasts. Applying the revised SD approach to a case study suggests that the proposed method closely resembles the way a hydrologist would visually evaluate the agreement of observation and model output. The results also show significant differences in the time-magnitude error statistics for different flow conditions, which standard methods are not able to reveal. I hence suppose that the importance of timing uncertainties in streamflow simulations is commonly underestimated. I conclude that the improved version of SD offers novel and elaborate techniques for both practical applications and model diagnostics and evaluation.

In *part III* I introduce a set of *diagnostics* for runoff production on the headwater scale. Its development was motivated by spatial patterns of model performance which I observed in the flood mitigation study. The proposed diagnostics, i.e. *data-driven signatures*, characterize the generation of baseflow, event-runoff and the seasonal water balance with respect to the corresponding physiographic controls by relating different components of the *input-state-output* triple. Key issues in this context are to derive meaningful surrogates for *state*, i.e. estimates for deep and near surface moisture content and to develop proper *normalization* schemes. The latter are required to consider the impact of physiographic catchment properties such as *bulk structural conductivity* or that of *biotic controls* of runoff production. Normalization is also required to be able to compare different sites. Applying the proposed *signatures* to a small inter-comparison study of catchments from southern Germany ($n = 22$, $17 \ldots 160\,km^2$) I find evidence for *functional similarity* among different sites and processes. This applies particularly for the seasonal water balance. Here, *normalized double mass curves* reveal significant and invariant regime shifts between winter and summer runoff regimes which coincide with the onset of vegetation. Temperature sums furthermore explain > 70 % of the variance in the seasonal summer runoff coefficients, suggesting a strong control of biotic controls across scales and across a considerable gradient of pyhsiographic conditions. I conclude that the proposed diagnostics can improve the understanding of runoff production on the headwater scale. They furthermore stimulate novel *kinds* of data analysis e.g. by evaluating *temporal derivatives* of rainfall and streamflow to identify the activation of rapid flow processes and/or by specifically evaluating *variable groups* instead of treating each property as single explanatory variable.

Based upon modelling large spatial domains I hence present both elaborate evaluation techniques and a set of process diagnostics which point towards new and/or alternative process hypotheses. The three topics hence close a learning cycle which is directed towards the improvement of hydrological models.

ZUSAMMENFASSUNG

In meiner Dissertation befasse ich mich mit drei Themenbereichen. Im ersten untersuche ich praktische Ansätze zur *Minderung von Hochwasserschäden durch flussgebietsweite Speichersteuerung*. Im zweiten Teil steht die Beurteilung hydrologischer Abflusssimulationen mit einem Fokus auf der simultanen *Erfassung von Zeit- und Wertefehlern* im Vordergrund. Der dritte Teil trägt zur Grundlagenforschung über *Abflussbildungsprozesse in mesoskaligen Kopfeinzugsgebieten* bei. Der erste Arbeitsschwerpunkt war durch ein Projekt des Bayerischen Landesamtes für Umwelt (BLfU) inhaltlich weitgehend vorgegeben. Die anderen beiden entwickelten sich aus der laufenden Bearbeitung heraus.

MINDERUNG VON HOCHWASSERSCHÄDEN

Im ersten Themenfeld, der *Speicherstudie*, stand die Frage im Vordergrund, ob durch eine (flussgebietsweite) Steuerung von Rückhaltebecken und Speichern Schäden durch Hochwasser potentiell verringert werden können. Die Frage wurde anhand von Daten aus dem Bayerischen Donaueinzugsgebiet untersucht, das in den letzten zwei Jahrzehnten mehrfach von schweren Hochwasserereignissen (HW) betroffen war. Innerhalb des rund 45.000 km^2 großen Gebietes sind insgesamt neun größere Rückhaltebecken und Speicher (im folgenden als *Speicher* zusammengefasst) mit einem Gesamtrückhaltevolumen von $\geqslant 127 \cdot 10^6$ m^3 vorhanden. Zur Beantwortung der Fragestellung wurden acht operationelle Hochwasservorhersagemodelle der Bayerischen Wasserwirtschaftsverwaltung gekoppelt und drei historische Hochwasserereignisse nachgerechnet. Zudem wurden hydrodynamische Simulationen durchgeführt. Zunächst erfolgte eine Beurteilung der Modellgüte. Darauf aufbauend wurde die *physikalisch maximal mögliche* (Fern)Wirkung der einzelnen Speicher anhand der historischen HW-Ereignisse untersucht. Hierzu wurden die Speicher nicht nur auf (*lokale*) Ziele im direkten Unterlauf des jeweiligen Speichers, sondern auch auf entfernte, in der Donau liegende (*regionale*) Pegel gesteuert. Über Referenzszenarien wurde die (Fern)wirkung der Speicher quantifiziert. Speicher, die den Wasserstand am *regionalen* Pegel signifikant ($\geqslant 10$ cm) beeinflussen können, wurden als *regional* wirksam klassifiziert. Sie kommen für eine *regional* abgestimmte Speichersteuerung in Betracht.

Im Untersuchungsgebiet zeigte nur einer von neun Speichern, der Forggensee, eine *regionale* Wirkung. An zwei von drei HW-Ereignissen ließ sich der Wasserstand durch ihn am Pegel Ingolstadt, 200 km vom Speicherauslass entfernt, noch um > 50 cm gegenüber dem Referenzszenario reduzieren. Weiter zeigte sich, dass die *lokalen* und *regionalen* Steuerungsstrategien zeitlich entkoppelt und damit vereinbar sind. Eine *regionale* Steuerung des Forggensees scheint damit möglich, auch ohne den Schutz der *lokalen* Bevölkerung zu gefährden. Zusätzlich verdeutlichen die Ergebnisse, dass zeitliche Aspekte für die Speicher-

steuerung eine große Rolle spielen und genaue Kenntnisse über die Simulationsunsicherheit zwingend erforderlich sind. Besonders deutlich wurde dies an der mittleren Donau. Hier fällt das größte Potential einer *regionalen* Speichersteuerung mit dem Auftreten großer Zeitfehler zusammen. Allgemeingültige Schlussfolgerungen der Studie sind:

- *Regionale* Speichersteuerungsstrategien können Hochwasserschäden verringern. Die Umsetzung einer solchen Steuerung ist jedoch komplex und stark vom Einzelfall abhängig.
- Der Einfluss eines Speichers ist allgemein umso größer, je größer sein Volumen im Verhältnis zum Volumen der zu beeinflussenden Hochwasserwelle ist.
- Das *Zeitfenster* für eine effektive (*regionale*) Steuerung von Speichern kann sehr klein sein. Genaue Kenntnisse über Wellenlaufzeiten und Zeitfehler in der Simulationskette sind daher von hoher Bedeutung.

MODELLUNSICHERHEIT

Zur besseren Quantifizierung von *Zeit- und Wertefehlern* habe ich im zweiten Teil meiner Dissertation das Series Distance (SD) Verfahren (Ehret und Zehe, 2011) um grundlegende Aspekte weiter entwickelt. SD erfasst simultan Zeit- und Wertefehler in Abflusssimulationen und -vorhersagen. Im Gegensatz zu klassischen Gütekriterien wie der Nash-Sutcliffe-Effizienz ist SD keine einzelne Gleichung, sondern vielmehr ein Verfahren, das die visuelle Ganglinieninterpretation nachempfindet. Im Kern wird versucht *hydrologisch ähnliches* miteinander zu vergleichen. Konkret bedeutet dies eine Differenzierung der Fehler nach Abflusssituation. Die Simulationsgüte in Niedrigwasserperioden wird also unabhängig von der Genauigkeit der Simulation in steigenden und/oder auch fallenden Zeitreihenabschnitten innerhalb von Niederschlag-Abflussereignissen (NA-Ereignissen) beurteilt. Der Vergleich der Zeitreihen basiert dazu auf *hydrologisch ähnlichen* Punkten und nicht auf Punkten mit identischem Abszissenwert. Dazu werden steigende Segmente der simulierten Zeitreihe mit den *zugehörigen* steigenden Segmenten der Messung verglichen (Analoges gilt für fallende Abschnitte). Eine solch differenzierte Betrachtung erfordert i) die Trennung von Niedrigwasserperioden und NA-Ereignissen, ii) die Identifizierung von steigenden und fallenden Segmenten innerhalb der NA-Ereignisse und iii) die Festlegung einer eindeutigen und chronologischen Abfolge von Segmenten, die miteinander verglichen werden.

In der ersten Version des SD Verfahrens mussten diese Arbeitsschritte z.T. von Hand gelöst und mit Hilfe eines stark vereinfachenden Schwellenwertverfahrens nachempfunden werden. Im Ergebnis war die automatisierte Anwendnung von SD auf längere Zeitreihen fehleranfällig und mit hohem Arbeitsaufwand verbunden. Um diese Hindernisse zu überwinden, wurde nun ein automatisiertes *coarse-graining* Verfahren entwickelt. Dieses empfindet die menschliche Fähigkeit nach, dominante Muster in Zeitreihen zu erkennen, diese in

Beziehung zu setzen und auf sinnvolle Art und Weise zu vergleichen. Technisch wurde das coarse-graining Verfahren iterativ über die Optimierung einer (parametrisierbaren) Zielfunktion gelöst.

Die Auswertungen zeigen, dass das *coarse-graining* die visuelle Interpretation von Ganglinien sehr gut nachempfindet. Zur Verdeutlichung des praktischen Nutzens wurde zudem eine Fallstudie gerechnet, bei der, basiert auf den SD Ergebnissen, ein zwei-dimensionaler Unsicherheitsbereich um eine Abflusssimulation konstruiert wurde. Die Fallstudie verdeutlicht zweierlei: i) auch bei vermeintlich genauen Abflusssimulationen können große Zeitfehler auftreten und ii) Zeit- und Wertefehlercharakteristika variieren zwischen unterschiedlichen Abflusssituationen (steigende vs. fallende Segemente) mitunter stark. Es liegen die Schlüsse nahe, dass die Bedeutung von Zeitfehlern weitgehend unterschätzt wird und dass eine differenzierte Fehlerbetrachtung von hohem Wert sein kann. Dies gilt nicht nur für praktische Anwendungszwecke wie die Konstruktion von Vertrauensbereichen, sondern auch für die Modelldiagnose.

PROZESSFORSCHUNG

Der dritte Teil meiner Dissertation widmet sich der Untersuchung von Abflussbildungsmechanismen auf Kopfeinzugsgebietsskala. Motiviert wurde er durch die räumlich heterogene Qualität der Modellierungsergebnisse in der Speicherstudie und dem Umstand, dass für mesoskalige Gebiete kaum Methoden zur Untersuchung der räumlichen Variabilität der Abflussbildungsmechanismen verfügbar sind. Ein Vergleich unterschiedlicher Einzugsgebiete hinsichtlich ihres *Verhaltens* in der Abflussbildung ist gegenwärtig nur schwer möglich. Erklärtes Ziel des dritten Arbeitsschwerpunkt war es, Bildung und Kontrollen von Basis- und Ereignisabfluss wie auch der saisonalen Wasserbilanz besser zu verstehen. Kenntnisse über diese Mechanismen wären nicht nur für die Regionalisierung, sondern auch für die Verbesserung verfügbarer Modelle eine große Hilfe.

Zu diesem Zweck habe ich dimensionslose und datengetriebene *Signaturen* entwickelt, die einzelne Komponenten des *Input-State-Output* Triples von Einzugsgebieten in Relation setzen. Zur Beschreibung des Gebietszustands (state) wurden unterschiedliche Schätzgrößen für den tiefen und oberflächennahen Gebietsspeicher abgeleitet und als erklärende Variable zur Prognose des *Basisabflusses* bzw. der *Ereignisabflussbeiwerte* herangezogen (output). Im Gegensatz zu diesen *speicher- bzw. kapazitätskontrollierten* Prozessen, deren Systemantwort (output) monoton mit der im System gespeicherten Wassermenge (state) steigt, wurde auf *Ereignisskala* zusätzlich versucht, die Bedeutung *intensitäts-kontrollierter* Abflussbildungsmechanismen wie Infiltrationsüberschuss oder präferentieller Fluss zu erfassen. Letztere sind weitgehend unabhängig vom *kapazitiven* Gebietszustand und werden vor allem (wenn auch nicht nur) durch die Intensität des Niederschlags, i.d.R. starke, konvektive Ereignisse ausgelöst. Zur Detektion dieser Mechanismen wurden unter anderem die zeitlichen Ableitungen von Ereignisniederschlag (input) und -abfluss (output) ausgewertet. Auf *saisonaler* Skala kamen *double-mass-curves* (DMCs) zum

Einsatz. Diese setzen den kumulierten Gebietsniederschlag (input) in Relation zur kumulierten Abflussspende (output). Um die Einflüsse biotischer und abiotischer Einflüsse zu trennen, wurden die DMCs separat für das Winterhalbjahr bzw. die Vegetationsperiode ausgewertet und mit 24 physiographischen Gebietseigenschaften korreliert.

Zum besseren Verständnis der Kontrollen der jeweiligen Abflussbildungsmechanismen und um unterschiedliche Gebiete vergleichen zu können, wurden die Signaturen *dimensionslos* formuliert. Dazu wurden alle Variablen, die Eingang in die Analyse fanden, mit strukturellen Gebietseigenschaften und/oder prozesslimitierenden Größen normiert. Zur Normierung des Basisabflusses wurden beispielsweise Schätzer für die strukturelle Leitfähigkeit des Untergrunds herangezogen. Die Schätzgrößen für die unterschiedlichen Speicherkompartimente wurden mit mittleren Porenvolumina des Bodens normiert und auf saisonaler Skala wurden die Achsen der DMCs mit Hilfe des kumulierten Jahresniederschlags in den Wertebereich zwischen Null und Eins übersetzt.

Angewendet auf 22 Kopfeinzugsgebiete ($17 \ldots 160$ km^2) im Bayerischen Donaueinzugsgebiet, identifizierten die entwickelten Signaturen *funktional* ähnliche Abflussbildungsmechanismen auf allen berücksichtigten Prozessskalen. Vielversprechende Ergebnisse wurden vor allem auf Ereignis- und Saisonalerskala gefunden. Im ersten Fall erklärte der normierte mittlere Abfluss vor einem NA-Ereignis in einigen Gebieten bis zu 70 % der Varianz der Abflussbeiwerte. In wenigstens zwei alpinen Gebieten wurden, trotz der nicht unerheblichen Einzugsgebietsgröße, Hinweise für intensitätskontrollierte Abflussbildungsprozesse gefunden. Auf saisonaler Skala erwiesen sich die normierten DMCs als besonders geeignet, um die Aufteilung von Niederschlag in Abfluss bzw. Evapotranspiration zu untersuchen. Das Verfahren detektierte eine starke und skaleninvariante Kontrolle der Evapotransipration auf die saisonale Abflussbildung. Über die Temperatur konnte nicht nur der Zeitpunkt des Regimewechsels zwischen Winter- und Sommerperiode deutlich genauer vorhergesagt werden als über gregorianische oder meteorologische Definitionen, sondern auch über 70 % der Varianz der saisonalen Sommerabflussbeiwerte erklärt werden. Ein wichtiges Ergebnis in der Auswertung der saisonalen Winterabflussbeiwerte war, dass der Median des topographischen Einzugsgebietsgradienten, multipliziert mit der gesättigten hydraulischen Leitfähigkeit des Bodens, 22 % der Varianz erklärte, wohingegen beide Variablen allein jeweils nicht signifikant korreliert waren und weniger als 5 % der Varianz erklärten. Ähnliche Ergebnisse wurden auch für die Basisabflussbildung ermittelt. Die vorgestellten Auswertungen eröffnen viele neue Forschungsperspektiven im Hinblick auf die Analyse von Daten (Verwendung von Variablengruppen anstelle der individuellen Auswertung) und/oder die Formulierung von *Prozesssignaturen*. Im Hinblick auf letztere ist zu erwarten, dass sich die Aussagekraft der vorgeschlagenen Methoden durch *ausgereifte* Normierungsansätze und die Integration anderer Daten noch weiter schärfen lässt.

Zusammenfassend betrachtet, schließt die Arbeit durch die drei behandelten Themen einen *Lernzyklus*. Ausgehend von *großskaligen Modellierungen* zur Minderung von Hochwasserschäden (IST-Zustand), folgen Analysen zur *Modellunsicherheit*. Das hierzu (weiter)entwickelte Series Distance Verfahren erlaubt eine differenzierte Bewertung der Simulationsgüte im Hinblick auf Zeit- und Wertefehler und zeigt somit Möglichkeiten zur *Modellverbesserung* auf. Letzteres erfordert jedoch auch fundierte Kenntnisse über die Abflussbildungsmechanismen in der Fläche. Hierzu entwickelte ich Methoden, die diese Prozesse räumlich differenziert erfassen.

Part I

INTRODUCTION

1 INTRODUCTION

Technical development (e.g. in computing power) and socio-economic needs (e.g. flood protection, hydro-electric energy supply or the mitigation of climate change impacts) provoke that hydrological models are nowadays applied to increasingly large domains and with increasingly high spatial resolution (see e.g. Biancamaria et al., 2009; Bravo et al., 2012; Markstrom, Hay, and Clark, 2016; Nester et al., 2011). Alongside, the requirements of model evaluation techniques increase as the importance of spatial patterns and the range of different perspectives on model performance need to be considered likewise. Melsen et al., (2016) even highlight that the calibration and validation time intervals do not keep pace with the increase in spatial resolution as they do not resolve the processes that are relevant at the applied spatial resolution. Though the usage of signatures (Carrillo et al., 2011; Kollat, Reed, and Wagener, 2012; Spence, 2007) and multi-criteria approaches (Efstratiadis and Koutsoyiannis, 2010; Gupta, Sorooshian, and Yapo, 1998; Kollat, Reed, and Wagener, 2012; Vrugt et al., 2003) for both calibration and validation are a clear step forward in the evaluation of environmental models, the hydrological community is still facing knowledge gaps e.g. in terms of the evaluation of timing errors, in the development of criteria which are understandable for laymen and/or in the emulation of visual hydrograph comparison techniques.

High resolution modelling of large spatial domains.

Improved methods for model evaluation are required.

The widespread application of hydrological models is however not only a blessing as hydrological theory lags behind technological progress in the way that processes causing e.g. flash floods (Merz, 2003), preferential flow (Beven and Germann, 2013), convective rainfall events (Ruiz-Villanueva et al., 2012), solute transport (Klaus et al., 2014), intensity triggered rainfall-runoff processes (Struthers and Sivapalan, 2007) and others are still not well understood and can thus not be modelled adequately, causing severe uncertainty. The reason for this is that many basic *hydrological functions* (and their spatio-temporal variation) remain unobservable underneath the surface and therefore remain unknown. In consequence, the selection of a hydrological model, i.e. the selection of a set of equations to represent hydrological processes is usually an under-determined problem.

Limitations in hydrological process understanding.

Guidance in terms of diagnostic approaches which shed light on the nature of the underlying processes are required - not only for the evaluation of hydrological models but also to improve our understanding of the *functional behaviour* of our watersheds and hence, to improve the available models. This is basically what (hydrological) similarity theory searches for and upon which a wealth of approaches originated. Past key studies to define and to detect hydrological similarity include the very popular topographic index (Kirby, 1975) which describes similarity of points within a catchment with respect to event

Diagnostics as learning tools for evaluation and improved understanding

Key studies on hydrological similarity and classification

scale runoff formation (Beven and Kirkby, 1979). The underlying key assumptions, i.e. that the topographic gradient is the most important control factor for runoff generation and that saturated hydraulic conductivity decreases exponentially with depth, are very appropriate concepts for humid basins with moderate to steep slopes and shallow, permeable soils overlying an impermeable bedrock. The concept appears inappropriate, however, for those environments which do not fit to the underlying assumptions, particularly if runoff formation is dominated by other factors as for instance the connectivity of saturated *patches* (Grayson et al., 1997; Zehe and Sivapalan, 2014), which is in turn controlled by soil hydraulic properties, geomorphological properties i.e. riparian zones, colluvial filled hollows, wetlands and others.

Also the concept of hydrological response units (HRU) has since Leavesley, (1973) inspired many scientists to detect functional entities and use them as building blocks for hydrological models. Flügel, (1996) and Flügel, (1995) later defined "Hydrological Response Units as distributed, heterogeneously structured entities having a common climate, land use and underlying pedo-topo-geological associations controlling their hydrological transport dynamics". Up to now, a large set of HRU separation methods has been suggested. Among them are for instance topographic indicators to support geomorphology-based predictive mapping of soil thickness (Pelletier and Rasmussen, 2009), explanations of the variability of base flow response based on climatic, soil and land use characteristics (Santhi et al., 2008) or decision trees to predict the locally dominating runoff processes based on soil, topography, landuse and small-scale experiments (Peschke et al., 1999; Scherrer and Naef, 2003; Schmocker-Fackel, Naef, and Scherrer, 2007). Also the REW concept (Reggiani, Sivapalan, and Hassanizadeh, 2000) can be seen as a mathematically rigorous and thermodynamically consistent interpretation of the HRU idea. Recently, Zehe et al., (2014) proposed a hierarchy of more specific functional units, defined on the basis of similarity of terrestrial and atmospheric controls on driving gradients and resistance terms controlling either the land surface energy balance or rainfall-runoff production, as refinement on the HRU idea. Other important approaches are the hydrology of soil types (HOST) classification for the United Kingdom (Boorman, Hollis, and Lilly, 1995) which is similar to the HRU concept in that sense that it is based on a number of perceptions describing dominant pathways of water movement through the soil and, where appropriate in the substrate. This also applies for the concept of hydrological landscapes proposed by Winter, (2001). The latter assumes that common patterns of surface runoff, ground water flow, and interchange of surface water and ground water with one another and with atmospheric water can be associated with fundamental hydrologic landscape units (FHLU).

Signatures as functional index or diagnostic approach.

Many recent studies propose the use of *signatures* for similarity assessment and model evaluation (Casper et al., 2012; Hrachowitz et al., 2013; Pfannerstill, Guse, and Fohrer, 2014). In most of these studies signatures are defined as specific characteristics of the hydrograph

such as autocorrelation, slope of/ or bias in the flow duration curve (or different segments thereof), rising limb density, peak distribution (Euser et al., 2013) and/or as flow statistics such as mean, variance, skewness or the coefficient of variation (Ley et al., 2011). Complementary to these classification schemes others proposes the use of *diagnostic signatures* or *functional indices* to study catchment response data and to improve the understanding of hydrological processes (Li, Sivapalan, and Tian, 2012; McMillan et al., 2011a; McMillan et al., 2014; Sawicz et al., 2011; Tian, Li, and Sivapalan, 2012). In these studies signatures are defined in a more comprehensive sense and next to properties of the hydrograph, characteristics of the water balance, recession characteristics and hydrological thresholds are also considered.

It is obvious that these approaches to define and to detect hydrological and/or functional similarity differ considerably with respect to the underlying assumptions, methods and proposed similarity measures (He, Bárdossy, and Zehe, 2011b; Hundecha and Bárdossy, 2004; Merz and Blöschl, 2004). However, so far there is no convergence of approaches and there is still a lack of robust data-driven diagnostics since many methods resting on similarity theory fall short of their expectations when it comes to practical applications. Ali et al., (2012) highlight this in their catchment inter-comparison study where "catchment groupings obtained using physical properties only did not match those obtained using flow indices, mean transit times or storage estimates".

No convergence of approaches.

1.1 GAINING NEW KNOWLEDGE

Modelling, evaluation and process diagnostics are the major topics I address in this thesis. They are all aspects of hydrological modelling which can be regarded as an iterative and hypothesis driven *learning cycle* pointing towards the generation of new knowledge (Fig. 1.1). Therein reality is represented through quantitative and qualitative observations (Gupta, Wagener, and Liu, 2008). Upon these we gain a mental understanding of the environmental system under consideration and *on the way things work* as Gupta, Wagener, and Liu, (2008) put it. The authors further coined the term *perceptual model* for this process. The latter is context-specific and subjective due to previous personal experiences and education. In the model building process the perceptual model is conceptualized and translated (through a range of different steps) into a numerical model which ultimately allows us to derive quantitative simulations. The learning cycle is closed by an *evaluation* of the derived model regarding its behaviour, form and function. Agreement, i.e. similarity, between the model and the observations of the environmental process under study suggests the acceptance of the derived model as a simplified but suitable representation of the system. Differences suggest rejection of the model, which always remains a hypothesis and to iterate on the learning cycle by

Modelling as an iterative and hypothesis driven learning cycle.

including new/other observations and/or by refining the perceptual and thus, the numerical model.

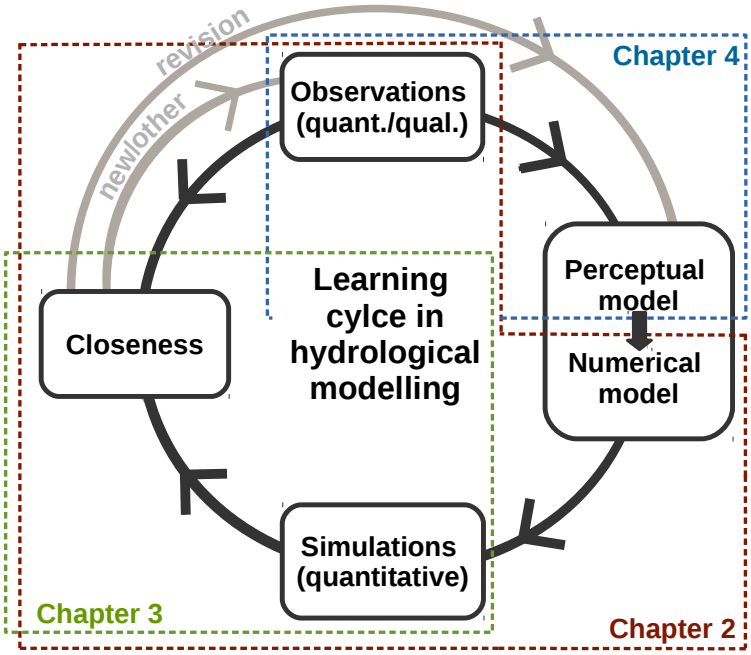


Figure 1.1: Structure of the thesis illustrated upon the learning cycle in hydrological modelling (modified after Gupta, Wagener, and Liu, (2008)).

Model evaluation is often unsatisfactory as standard statistics neglect timing errors and do not point towards possible model improvements.

In practice, model evaluation is often limited to the comparison of (historic) observations and simulations by measuring their degree of closeness using some statistical performance measures such as the *Nash efficiency* (Nash and Sutcliffe, 1970) or the root-mean-squared-error. Such a (fairly) weak evaluation based upon *standard* statistics bears, next to the well-known deficiencies of the NASH or more generally those of all *mean-square-error* based distance metrics (Gupta et al., 2009; Schaefli and Gupta, 2007; Seibert, 2001), two additional shortcomings: i) Nearly all standard statistics measure closeness in a vertical sense in that they somehow determine the differences in the ordinate between observed and simulated points with an identical abscissa. Such a vertical definition of closeness neglects the horizontal component of the error and is thus incomplete. ii) standard statistics are poor in a diagnostic sense in the way that they do not illustrate the nature of the problem under consideration and thus, do not point towards an alternative perceptual (and numerical) model hypothesis.

Flood mitigation and model application (Part II).

In my thesis I do not fully iterate through the learning cycle, but propose methods and techniques which focus on different aspects thereof and generate new knowledge in various respects. With reference to Fig. 1.1 the thesis is structured as follows: I first introduce and apply a large-scale coupling of (numerical) hydrological and hydrodynamic models which are currently the best available representations of the hydrological processes in the catchment of interest in chapter 2. Using these tools, I assess different aspects of flood mitigation in large

basins including *regional* reservoir operation and spatial performance evaluation by means of different performance statistics. The findings of this application-orientated study motivated the (further) development of a distance metric which captures both, the vertical and the horizontal component of the simulation error (chapter 3) as timing errors proved to be important. Though the main focus of this chapter is on closeness I relate the observed deviations back to the model output and present a method to construct uncertainty envelopes around streamflow simulations and forecasts. The last part of the thesis (chapter 4) closes the cycle in that it presents diagnostic signatures for process identification in lower mesoscale catchments which relates to perceptual model building. Therein I explore the interplay between state, structure and runoff behaviour from the similarity perspective. This is for instance done by disentangling the components of dynamical laws, i.e. parameters describing *gradients* and *resistances* (Zehe et al., 2014) and by treating them as parameter groups and not as single explanatory variable and/or by explicitly exploring the instationary role of ecological controls on runoff generation. In this way I present novel techniques which are not yet described by other approaches on similarity assessment. The work presented in this chapter was motivated by the distinct spatial performance patterns that I observed in the modelling exercise (chapter 2).

Model uncertainty (Part III).

Process diagnostics (Part IV).

By covering the range of these topics the thesis bridges the gap between application-orientated questions and basic research. The three different topics are self-contained and published separately in peer-reviewed international journals. In the following I introduce the different chapters in more detail. In chapter 5 the topics are discussed jointly.

1.2 AT THE INTERFACE OF SCIENCE AND PRACTICE

In chapter 2 I introduce a large-scale coupling of hydrological and hydrodynamic models. The purpose of this modelling exercise is to analyze the potential of *coordinated regional* reservoir operation for improving flood mitigation in large catchments. Contrary to regular operational practice of flood protection reservoirs, where a single reservoir is operated such that a certain area in the near *(local)* downstream vicinity of the reservoir is protected against flooding, I analyze here whether the joint operation of multiple reservoirs bears the potential to improve flood mitigation at rivers of higher order and thus, at multiple and distant *(regional)* spots. The *reservoir study* has thus a large practical focus and relevance and was carried out in the Bavarian part of the Danube basin, which was affected by several severe floods throughout the last two decades causing damages in the multi-digit billion area (Rimböck et al., 2014).

Coupled hydroogical-hydrodynamic modelling in a large spatial domain.

Optimizing reservoir operation at local and regional locations.

Formally, the study was commissioned by the Bavarian Environmental Agency (BLfU) and carried out in cooperation with the Chair of Hydraulic Engineering and Water Resources Management of the TU München (TUM). There was also a close collaboration with the

Bavarian Flood Forecasting Agency which provided hourly historic rainfall-runoff data and different operational models. These covered, in the hydrological case, the Bavarian part of the Danube basin ($\approx$ 45.000 km^2) with a resolution of one square kilometer. Two-dimensional hydrodynamic models for the Danube and its major tributaries Wertach, Lech, Isar, Naab and Regen (covering a total river length of $\approx$ 1300 km) were also provided. These tools and different methodological approaches were used to analyze the *local* and *regional* impact of nine different flood protection reservoirs which are distributed along the major northern and southern tributaries of the Danube based upon data from historic flood events. Specifically I address the following research questions:

Research questions related to the modelling of large spatial domains and (regional) reservoir operation.

- *Q 1.1*: Which reservoirs do have a regional impact on flood mitigation and might be suitable for a coordinated reservoir operation?
- *Q 1.2*: Are regional operation strategies conformable with local flood mitigation?
- *Q 1.3*: Is large-domain hydrological modelling accurate enough to allow for a regional operation of reservoirs?
- *Q 1.4*: To what extent does the coupling of hydrological and 2d-hydrodynamic models improve the simulation accuracy?

1.3 CLOSENESS ON THE ORDINATE AND ON THE ABSCISSA

Emulating visual hydrograph inspection using a pattern matching procedure.

In chapter 3 I focus on the evaluation of streamflow simulations and forecasts which are still the most important outcomes of hydrological models. Following up on the analysis of spatial model performance and the merging of different performance statistics into a single criterion as described in the previous chapter, I concentrate on the improvement of a novel distance metric which emulates visual hydrograph inspection. The latter is a powerful though subjective evaluation technique which is widely used in hydrology. In the comparison of time series it allows for the simultaneous consideration of various aspects like the occurrence of hydrological (rainfall-runoff) events, the timing of peaks and troughs, the agreement in shape and the comparison of individual rising or falling limbs within an event. The main strength of visual hydrograph comparison results from the human ability to identify and compare matching, i.e. hydrologically similar elements in hydrographs and to differentiate between magnitude *(vertical)* and timing *(horizontal)* agreement of hydrographs. Visual hydrograph inspection thus rests on two fundamental steps: In the first, rising and falling *segments* of the two time series are identified and intuitively and meaningfully matched. This requires a harmonization of the temporal resolution of the two hydrographs and the identification and mapping of dominant patterns. We call this process *coarse-graining*. The second step involves the actual comparison of the two hydrographs and refers to a joint but individual consideration of timing and magnitude errors.

Up to now only very few methods allow emulating what human reasoning solves in such an intuitive way. With few exceptions, all statistics that are traditionally used for model evaluation in hydrology *vertically* compare points with identical abscissa. These methods hence neither ensure that *apples are compared with apples* nor do they account for the horizontal error component.

Standard statistics may compare apples with oranges.

To close this gap in knowledge or at least, to narrow it, I introduce an adaptation of the *Series Distance* (SD) approach (Ehret and Zehe, 2011). The latter is a deterministic approach for the simultaneous but separate quantification of timing and magnitude errors in streamflow simulations and forecasts which was substantially revised in recent years. This includes in particular the development of an automated *coarse-graining* scheme which allows for a robust and continuous application of the formally event-based method. However, the core of the procedure was subject to several modifications as well.

Further development of the Series Distance approach.

I hence provide a novel evaluation technique for assessing closeness which is fundamental to the learning cycle (Fig. 1.1). The SD method is particularly relevant whenever knowledge of timing uncertainties is of importance such as in the operation of reservoirs, hydropower plants or for the planning of dike defense measures during floods.

The SD chapter is accompanied by a case study in which I apply the method to data from a small alpine catchment in order to assess the role of timing uncertainties in streamflow simulations. Based upon the SD results I construct 2-dimensional uncertainty envelopes around a historic streamflow simulation and compare it to a benchmark error model. Here I address the following research questions:

Research questions regarding the assessment of closeness.

- *Q 2.1*: How to emulate (human reasoning in) visual hydrograph inspection?
- *Q 2.2*: What is the role of timing uncertainties in hydrological streamflow simulations and forecasts?
- *Q 2.3*: How to consider horizontal error components in the construction of uncertainty envelopes and what is their impact on the region of confidence?

1.4 UNDERSTANDING THE NATURE OF RUNOFF PRODUCTION

The development of a perceptual model for an unknown catchment, i.e. obtaining a mental understanding of how a watershed actually *functions* in terms of storage and release of water is highly relevant but a difficult and challenging task. As pointed out by Sawicz et al., (2011), knowledge of both catchment functioning and their causes would allow us to (hierarchically) classify our catchments, to transfer (regionalize) information and permit generalization. It would further allow us to build more realistic (or minimally adequate) models which would in turn promote that we get more *right answers for the right reasons* (Dooge, 1986; Gottschalk, 1985; Grigg, 1965; Kirchner, 2006). The latter is of particular importance for the modelling

Knowledge of catchment functioning and the corresponding causes and controls is highly relevant.

of climate change impacts as changing boundary conditions, i.e. non-stationarity, "undermine a basic assumption that historically has facilitated management of water supplies, demands, and risks" as emphasized by Milly et al., (2008) (see also Wagener et al., (2010)).

Assessing runoff production at different time scales.

In order to move a step forward in the functional classification of catchments and motivated by distinct spatial patterns of model performance which I observed in the reservoir study, I explore the nature of different runoff formation processes at the catchment scale. Specifically, I assess catchment runoff production at both the event and seasonal time scale. On the seasonal scale the focus is on the water balance, i.e. the partitioning of rainfall into evapotranspiration (green water) and discharge (blue water) which I assess with respect to timing, amount and the corresponding controls. At the event scale the focus is on the importance of different sub-surface storage compartments and on the detection of *intensity* controlled mechanisms. The latter refers to large intensive runoff responses e.g. due to the activation of rapid flow paths/processes such as surface runoff and/or preferential flow which are triggered by an intensive, convective rainfall forcing. Contrary, the rates of subsurface matrix flow or base flow production depend primarily on the amount of water that is stored in the respective control volume and are thus, independent from the intensity of the forcing. Following Struthers and Sivapalan, (2007) we call the latter *capacity* controlled runoff formation mechanisms.

Assessment of (functional) similarity using dimensionless quantities.

Motivated by the success of similarity assessment based upon dimensionless quantities for scaling throughout a range of different disciplines including hydraulics (e.g. Reynolds, Froude, Peclet number), acoustics (e.g. Helmholtz number) or chemistry (e.g. Damköhler numbers) I approach the *functional classification* of catchment runoff formation using the dimensionless quantities as well. The latter proved useful as demonstrated in many studies (Bahram, Pierre, and Odgen, 1995; Berne, Uijlenhoet, and Troch, 2005; Budyko, 1956; Reggiani, Sivapalan, and Hassanizadeh, 2000; Struthers, Hinz, and Sivapalan, 2007a; Woods, 2003) among others.

Dimensionless representations require proper normalization.

The concept I propose rests on dimensionless *state-response* and *forcing-response* plots. Therein, I relate different forcing and storage descriptors to selected response measures, again separately for different runoff production timescales. The storage(forcing)-response plots are formulated in dimensionless form. This is vital as it i) allows to compare different catchments and ii) to explore the interplay of state and structure on catchment runoff formation. Dimensionless quantities are obtained using proper normalization in that I normalize the variable of interest by its limiting terrestrial or forcing characteristic. The resulting storage(forcing)-response plots can be interpreted as 2-dimensionless signatures (or *fingerprints*) for catchment scale runoff production as they may reveal (in)variance across scales. Specifically, I address the following research questions:

Research questions relevant for the formulation of functional diagnostics.

- *Q 3.1*: Are dimensionless state(forcing)-response diagrams feasible to characterize and to detect differences in event scale

runoff production, baseflow generation and the seasonal water balance?

- *Q 3.2*: Is it possible to detect evidence for intensity controlled runoff formation, which is essentially a high frequency process, based upon (hourly aggregated) operational data, which are poorly resolved in this context?
- *Q 3.3*: Which structural, climatic and ecological catchment characteristics explain the differences between different catchments and among different years and do any of them operate in groups?

I assess these questions using a small catchment inter-comparison study. For this I use the same data source as in the reservoir study (chapter 2) but instead of focusing on large streams I sample the smallest available (gauged) headwater catchments. Compared to the first chapter which is fairly application-oriented this part contains basic research. It provides data-driven diagnostics (signatures) which seek to improve our understanding of the *functional* runoff formation mechanisms at different time scales. Thereupon, it seeks to inspire a novel way of thinking in terms of analyzing variables as *parameter groups* and not as single explanatory variables.

Part II

THE POTENTIAL OF COORDINATED RESERVOIR OPERATION FOR FLOOD MITIGATION IN LARGE BASINS

In this part I assess the hypothesis that flood protection reservoirs can be operated on distant, i.e. *regional* locations and river reaches to improve flood mitigation in large catchments. The *reservoir study* was commissioned by the Bavarian Environmental Agency and carried out in cooperation with the Chair of Hydraulic Engineering and Water Resources Management from the Technical University of Munich (TUM). The study involved the large-scale coupling of hydrological and hydrodynamic models and different techniques on reservoir impact assessment. Whereas the focus of my former PhD colleague Daniel Skublics from the TUM was on the propagation of flood waves and their representation through hydrodynamic models, I here present the results of the hydrological part of the project. These include the findings obtained for model evaluation and *regional* reservoir operation.

The study is published in the Journal of Hydrology. The remainder of part II is a reprint of:

Seibert SP, Skublic D and Ehret U (2014): The Potential of coordinated reservoir operation for flood mitigation in large basins - a case study at the Bavarian Danube using coupled hydrological-hydrodynamic models. Journal of Hydrology 517, 1128-1144.

2

PART 2: FLOOD MITIGATION IN LARGE BASINS

ABSTRACT

The coordinated operation of reservoirs in large-scale river basins has great potential to improve flood mitigation. However, this requires large scale hydrological models to translate the effect of reservoir operation to downstream points of interest, in a quality sufficient for the iterative development of optimized operation strategies. And, of course, it requires reservoirs large enough to make a noticeable impact. In this paper, we present and discuss several methods dealing with these prerequisites for reservoir operation using the example of three major floods in the Bavarian Danube basin (45,000 km^2) and nine reservoirs therein: We start by presenting an approach for multi-criteria evaluation of model performance during floods, including aspects of local sensitivity to simulation quality. Then we investigate the potential of joint hydrologic-2d-hydrodynamic modeling to improve model performance. Based on this, we evaluate upper limits of reservoir impact under idealized conditions (perfect knowledge of future rainfall) with two methods: Detailed simulations and statistical analysis of the reservoirs' specific retention volume. Finally, we investigate to what degree reservoir operation strategies optimized for local (downstream vicinity to the reservoir) and regional (at the Danube) points of interest are compatible. With respect to model evaluation, we found that the consideration of local sensitivities to simulation quality added valuable information not included in the other evaluation criteria (Nash-Sutcliffe-Efficiency and Peak timing). With respect to the second question, adding hydrodynamic models to the model chain did, contrary to our expectations, not improve simulations, despite the fact that under idealized conditions (using observed instead of simulated lateral inflow) the hydrodynamic models clearly outperformed the routing schemes of the hydrological models. Apparently, the advantages of hydrodynamic models could not be fully exploited when fed by output from hydrological models afflicted with systematic errors in volume and timing. This effect could potentially be reduced by joint calibration of the hydrological-hydrodynamic model chain.

Finally, based on the combination of the simulation-based and statistical impact assessment, we identified one reservoir potentially useful for coordinated, regional flood mitigation for the Danube. While this finding is specific to our test basin, the more interesting and generally valid finding is that operation strategies optimized for local and regional flood mitigation are not necessarily mutually exclusive, sometimes they are identical, sometimes they can, due to temporal offsets, be pursued simultaneously.

2.1 INTRODUCTION

Reservoirs for flood protection are typically designed and operated with the goal to protect the areas in the downstream vicinity, i.e. along the river where the reservoir is situated (this is what we will refer to in the text as "local"). However, especially during floods at subsequent, larger rivers of higher order, the question arises whether existing reservoirs can also be used for flood mitigation at more distant points of interest along these large rivers (we will refer to this as "regional"), without compromising local flood mitigation. For example, such requests for regional operation have come from communities along the Bavarian Danube for the Forggensee, a reservoir which is located more than 200 km upstream on the Lech River, an alpine tributary to the Danube (Fig. 2.1). Historically, the Forggensee has been operated during floods with a focus on protecting the local communities along the Lech river.

If floods at regional points of interest can be influenced by several reservoirs, the additional question arises about how to optimally coordinate the operation of these reservoirs. So far, the operation of multiple reservoirs on a single gauge or the joint operation of a single reservoir to multiple local and regional gauges (in the text, we will refer to this as coordinated (multi)-reservoir operation) during floods is not widespread. One reason for this is probably that modeling tools sufficiently fast and precise for large-scale and coordinated reservoir management have not been available in the past. Recently, several large-scale hydrological models have been set up, as mentioned by Nester et al., (2011), Collischonn et al., (2007) and Mauser and Bach, (2009). The latter have used the PROMET model at 1 km^2 and 1 h resolution in a 77,000 km^2 river basin. Also, the LISFLOOD model (Knijff, Younis, and De Roo, 2010) has been used to set up the European Flood Alert System "EFAS" on a 25 km^2 grid Thielen et al., 2009. However, all of these examples lack an option to consider and to represent reservoir operation. In the models, flood propagation in larger rivers is typically represented by simplified hydrological routing schemes such as the Williams method (Williams, 1969), the Muskingum method (Cunge, 1969; McCarthy, 1938; Todini, 2007) or multilinear approaches (Szolgay, 2004). Hence, the ability of such models to deal with the complex flow processes that occur during floodplain inundation along large, flat rivers is limited. These processes can significantly change the shape and timing of a flood wave and, with it, also have an impact on finding the optimal reservoir operation strategy for regional flood mitigation. Hydrodynamic modeling approaches based on the Navier Stokes equations and suitable simplifications (e.g. the 2d-shallow water equations) are much better capable of resolving these processes and can, coupled with hydrological models, be used to overcome the limitations of the latter in this respect. This has already frequently been done, however, most of these studies have either been done at very high resolution and small spatio-temporal extent (e.g. Bradley et al., 1996; Kim et al., 2012; Nicholas and Mitchell, 2003; Ogden et al., 2000) or at large extent but

small spatio-temporal resolution (e.g. Biancamaria et al., 2009; Bravo et al., 2012; Paiva, Collischonn, and Buarque, 2013). Large-scale and high resolution couplings are still rare (e.g. Bao and Zhao, 2012), but are exactly what is needed for reservoir operation on the regional scale. In this paper, we present and discuss the application and results of such an extensive hydrologic-hydrodynamic model coupling. Another issue that arises in the context of regional reservoir operation is the need for evaluation of the underlying models in high resolution and on large scales. This is required for two reasons: Firstly to identify (and possibly remove) weak points in the modeling chain, and secondly to set the basis for quantification of the overall uncertainty associated with reservoir operation and the contributions of the various model chain components (e.g. uncertainties from weather forecasts and observations, uncertainties from representation of the rainfall-runoff process and routing, etc.). There is a wealth of literature on the topic of uncertainties related to hydrological modeling (Beven and Binley, 1992; Georgakakos et al., 2004; Matott, Babendreier, and Purucker, 2009; McMillan et al., 2011b), to hydrodynamic modeling (Cloke and Pappenberger, 2009; Marka et al., 2004; Pappenberger et al., 2005; Pappenberger et al., 2006; Zehetmair et al., 2008) and to reservoir operation (Raje and Mujumdar, 2010; Soares, Covas, and Reis, 2011; Yazdi and Salehi Neyshabouri, 2012), however it is beyond the scope of this study to establish a full treatment of uncertainties. Rather, we present a multi-criteria (see also Moriasi et al., (2007)) approach which includes a new metric termed "Peak Level Certainty" (PLCY) and which relates the forecast uncertainty at a given point to the locally required precision to make useful decisions.

It is also beyond the scope of this study to develop a complete system for coordinated multi-reservoir operation system (Labadie, 2004; Opan, 2011; Tilmant, Goor, and Kelman, 2011), instead we investigate methods to identify candidate reservoirs for regional flood mitigation by analyzing their maximum potential impact under idealized conditions. We apply both elaborate simulation-based and simplified statistically-based approaches and investigate to which degree the latter can be used as an easy-to-obtain approximation of the former.

The overall goal of this paper is therefore to present and to discuss methods which can be used for coordinated reservoir operation on the regional scale, rather than establishing such a system. We do so by using the example of three major flood events in the Bavarian Danube basin (45,000 km^2). The basin is completely covered by hydrological models and by 1300 km of 2-d hydrodynamic models and contains nine larger reservoirs.

The remainder of the paper is structured as follows: An overview of the basin, the data and all applied methods and models are provided in Section 2.2. Results from the modeling exercises are presented and are discussed in Section 2.3, being followed by conclusions in Section 2.4.

2.2 MATERIALS AND METHODS

2.2.1 *Study region*

The study area comprises the Bavarian Danube catchment, from its entry at the Baden-Württemberg border (river kilometer 2586) to gauge PILZ (Passau Ilzstadt, river kilometer 2225) which is located next to the Austrian border (Fig. 2.1). The basin encompasses 45,000 km^2 and exhibits strong hydro-metrological gradients in both north-south and east-west directions, with mean annual precipitation ranging from 600 mm in the northern sub-catchments to more than 2000 mm in the southern, alpine areas: the basin average is around 1000 mm (BMU, 2002). The rainfall regime is characterized by distinct seasonal cycles and in the southern alpine areas, 50 % and more of the precipitation falls as snow.

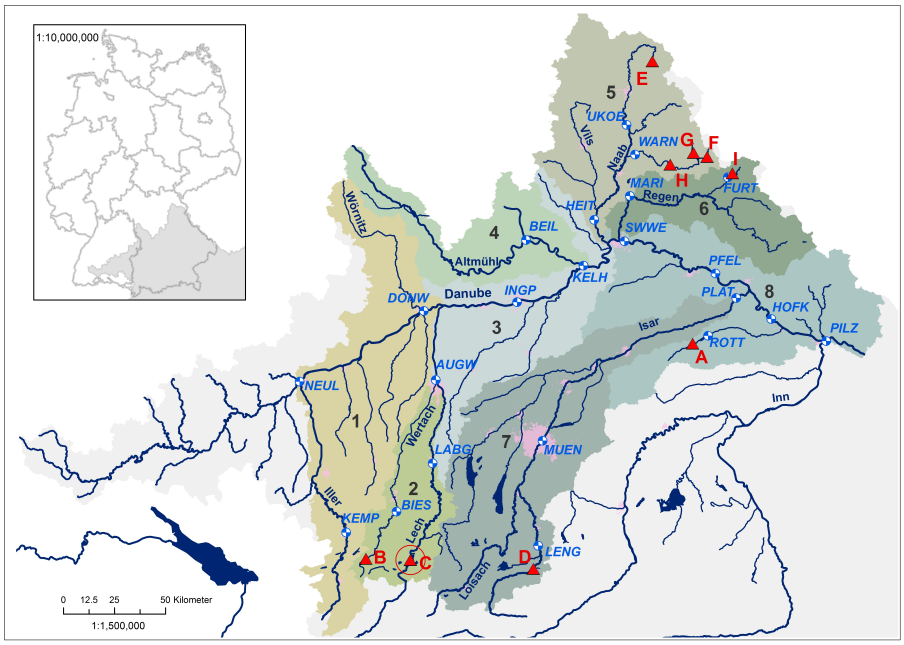


Figure 2.1: *Inset*: Germany with federal state boundaries (gray lines) and the Danube catchment (gray area). *Large map*: Bavarian Danube basin. The colored areas represent different sub-basins. The corresponding numbers refer to the hydrological sub-model identifier (MID, see Table 2.2). Important reservoirs (red triangles) are indicated by capital letters A – I (see reservoir identifier RID, Table 2.1). The encircled triangle highlights the Forggensee reservoir. Selected gauges are indicated by blue-white circles, the blue italic gauge identifiers (GID) link to Tables 2.1 and 2.2. The blue lines and polygons represent the major river network and important lakes.

The basin is composed of distinct physiographic regions: The southern areas are located in the northern Limestone Alps, the eastern parts lie in the Bavarian and Bohemian Forest. The majority of the basin (about 56 %) between the Alps in the South and the Danube in the center is underlain by faulted Molasse sediments. The area north

of the Danube is more heterogeneous, but limestone and karst formations prevail.

The most important tributaries of the Bavarian Danube from the Alps in the South are the rivers Iller, Lech and Isar, while from the North it are the Altmühl, Naab and Regen. Almost all larger rivers in the Bavarian Danube catchment, in particular the river Lech, have been intensively regulated during the last two centuries, and their potential for hydro-power production is nearly fully exploited (Pall and Janauer, 2003; Rösl, Rosemann, and Stockenhuber, 1984).

Nine flood protection reservoirs within the Bavarian Danube basin are large enough to be relevant for this study. The largest ones are located on the southern tributaries, in particular the reservoirs Forggensee (Lech) and Sylvensteinspeicher (Isar) (triangles C and D in Fig. 2.1). The reservoirs differ considerably with respect to size, technical aspects and catchment properties (Table 2.1).

Table 2.1: Characteristics of selected reservoirs within the Bavarian Danube catchment. For each reservoir, the name of the main water course and two gauge identifiers (GID) are provided. The local GID refers to the local, primary target of reservoir operation; the regional one represents a distant gauge in the Danube. V_R is the regular retention volume, $V_{R,max}$ the (theoretical) maximum possible retention volume. Area refers to the reservoirs' catchment size, HQ100 to the 1-in-100 year reservoir inflow. MAP is the mean long term annual precipitation in the reservoir catchment. The reservoir identifier (RID) and the GIDs link to Fig. 2.1. Order is according to $V_{R,max}$. Estimates are labeled by asterisks (*).

Reservoir	Water course	GID Local	GID Regional	V_R ($10^6 m^3$)	$V_{R,max}$ ($10^6 m^3$)	Area (km^2)	HQ100 ($m^3\ s^{-1}$)	MAP mm	RID
Sylvenstein-speicher	Isar	MUEN	HOFK	59.4	88.6	1138	950	1928	D
Forggensee	Lech	AUGW	INGP	22.1	83.6	1582	985	1597	C
Eixendorfer See	Naab	WARN	SWWE	12.1	16.9	399	130	853	H
Grüntensee	Wertach	BIES	INGP	11.1	15.6	85	161	1857	B
Vilstalsee	Vils	ROTT	PILZ	8.1	9.6	623	*315	863	A
Liebenstein-speicher	Naab	UKOE	SWWE	3	4.5	66	*35	893	E
Silbersee	Naab	WARN	SWWE	3.6	4.4	58	31	903	G
Drachensee	Regen	FURT	SWWE	3.9	3.9	178	*100	823	I
Perlsee	Naab	WARN	SWWE	3.2	3.7	61	36	943	F

2.2.2 *Models*

2.2.2.1 *Hydrological rainfall-runoff models*

The Bavarian Danube basin is covered by eight conceptual hydrological (HY) models (Table 2.2) of type LARSIM ('*L*arge *A*rea *R*unoff *S*imulation *M*odel'; Ludwig and Bremicker, (2006)). They are in operational use at the Bavarian flood forecasting agency and are used in this study in a coupled way to simulate distributed runoff formation, runoff concentration and flood routing. We use the models in

an event-based (i.e. non-continuous) mode (Ludwig, 1982). The models are operated on hourly time steps and except for the Lech and Isar, all models are grid-based (resolution 1 km^2) (see Fig A.1). The Lech and Isar models consist of variable-sized sub-catchments, typically 2-10 km^2 large. The channel routing is represented by simplified routing schemes (usually after Williams, (1969)). Including the routing parameters for channel roughness and geometry, altogether 13 model parameters (times the number of elements) are available for calibration.

Table 2.2: Overview of the hydrological sub-models. Important gauges are provided by the gauge identifiers (GID). Both GID and the model identifier (MID) refer to Fig. 2.1

MID	Model name	Water course	Size (km^2)	GID	Reservoirs	next MID
1	Iller and upper Danube	Iller, Danube	11410	KEMP, NEUL	-	3
2	Lech	Lech	3800	BIES, AUGW	Grüntensee, Forggensee	3
3	Central Danube	Danube	7228	INGP, SWWE	-	8
4	Altmühl	Altmühl	3812	BEIL	-	3
5	Naab	Naab	6362	HEIT	Eixendorfer See, Liebensteinspeicher, Perlsee Silbersee	3
6	Regen	Regen	3091	MARI	Drachensee	3
7	Isar	Isar	8435	MUEN, PLAT	Sylvensteinspeicher	8
8	Lower Danube	Danube	7931	HOFK, PILZ	Vilstalsee	-

Due to their operational use for flood forecasting, all models are calibrated with a focus on large floods, typically a one-in-one-hundred-year event. The application of the models to low or average flow conditions is of course possible but usually prone to higher uncertainty since the models are applied "out of scope". In every such case an event-specific adaptation of the default model configuration is required. This is done by an event-specific scaling factor (SF) which is applied to the default runoff-coefficients (RC) obtained from calibration. The SF is the most important and most sensitive model parameter. Typically it differs between 0.1 (for conditions with extremely small RC values such as extreme low flow) and 1 (for conditions with RC values similar to a one-in-one-hundred-years flood). In operational forecast mode, the SFs are determined for each new forecast run by expert's choice, mainly based on present discharge observations and predicted rainfall. As our study involved many model runs, we had to establish an automated method to mimic experts' reasoning. To achieve this, we used the single best predictor for RC, i.e. observed discharge which was converted into three classes: small (return period < 2 yrs), medium (return period $\geqslant 2$ and $\leqslant 10$ yrs) and high (return period > 10 yrs). Accordingly, we applied the SFs 0.4, 0.7 and 1.0. These values were obtained in a prior study, where we identified best event-specific SF values across many events and sub catchments by optimization and then averaged the results. Season-dependency of SF was not a critical issue in our study as all floods oc-

curred during the same season (May-Aug). The so determined SF values (or RC values, respectively) were estimated individually for each event, but kept constant throughout the periods of simulation. Using this simplified approach we emulated operational conditions, but in an automated way. As this study deals with hindcasts, it would have easily been possible to identify a perfect SF value for each event by optimizing it based on the available rainfall and discharge data. However, this would have meant to neglect the forecast uncertainty that under operational conditions was introduced by expert's choice of the SF value. This would of course have made simulation results better, but would have missed the purpose of the study. For the remaining parameters, we used (default) operational model configurations. To account for historical reservoir operation, discharge observations at the outlets of the reservoirs Grüntensee, Forggensee, Sylvensteinspeicher, Liebensteinspeicher, Perlsee and Silbersee (Table 2.1 and Fig. 2.1) were fed into the models, unless they were used for reservoir optimization (see section 2.2.5.1). The same applies for the Danube discharge as it enters Bavaria and for the hydrograph of the river Inn. The latter was fed into our models at the gauge Passau Ingling which is the located about 2.5 km upstream of the Inn river mouth. The simulation periods for the three events were 20 days (10.05.-01.06.1999), 13 days (05.08.-18.08.2002) and 11 days (18.08.-29.08.2005).

2.2.2.2 *Hydrodynamic models*

To test our hypothesis that coupled hydrologic-hydrodynamic modeling improves simulations in larger catchments, we used altogether 45 two-dimensional hydrodynamic (HD) models covering a total river length of approximately 1300 km. Most of them were either developed for floodplain mapping or for the analysis of flood protection measures. Due to the large number of models a large variety of bathymetry and floodplain data were used. Most HD models are based on precise airborne laser scanning data at 1 x 1 m resolution and river cross section surveys at 200 m intervals. The resolution of the computational grid ranges from 10 x 25 m in the channel to 50 x 50 m on the floodplains. For model calibration the flood event May 1999 was used for most HD models. With few exceptions, all models were provided by the Bavarian water management authorities and set up in Hydro AS-2d (Nujic, 2003). The latter is a commercial 2d-hydrodynamic software which solves the shallow water equations assuming a hydrostatic pressure distribution with a finite volume approach.

Detailed information on the hydrodynamic models can be obtained from the PhD thesis of Skublics, (2014).

The HD models were combined and homogenized. This included a simplification of the topology of most models by iteratively reducing the number of elements (from a total of eight million to four million), but preserving the precise elevation of hydraulic decisive terrain. This way, much shorter computation times were achieved, while sufficient accuracy of the results was maintained. Boundary conditions like material and land use definitions remained unchanged. Despite the homogenizations, the simulation quality of the models remained heterogeneous since their underlying data were of different quality.

The combined models (nine altogether) allowed flood wave simulations along more than 1300 river kilometers along the entire Bavarian Danube and along its tributaries Wertach, Lech, Isar, Naab and Regen. The (one-way) coupling between the HY and the HD model was established by directly feeding the HY model outputs into the HD models. Runoff generated in a catchment between two tributaries was summarized and was combined with the closest tributary. This way, the total number of interface nodes between HY and HD was reduced from 900 to 200.

2.2.3 *Data*

2.2.3.1 *Precipitation and streamflow data*

The hydrological models were forced with observed rain gauge data. Time series from 144 (May 1999), 141 (August 2002) and 280 (August 2005) meteorological stations in hourly resolution were available. Additionally, daily totals of precipitation time series of 701 (May 1999) and 1015 (August 2002) meteorological stations were used for densification, so that a mean areal coverage of 59, 43 and 179 km^2 per station was achieved for the respective events. The daily rainfall sums were disaggregated to hourly sums according to the temporal rainfall distribution of the nearest station with hourly data. Precipitation data were spatially interpolated using the "Rasterpunktverfahren" a method which is implemented within the LARSIM model (Ludwig, 1978). It is essentially an inversedistance weighted combination of the four closest stations, one in each quadrant, surrounding the point of interest.

For model evaluation, observed hourly stream flow data from about 90 gauges throughout the basin were considered. Apart from data quality criteria, the gauges were selected in such a way that all subcatchments were represented as adequately as possible.

2.2.3.2 *Characteristics of the selected flood events in the Danube Basin*

During the last fifteen years, four larger flood events with return periods exceeding ten years occurred in the Bavarian Danube catchment. They originated in the southern, alpine areas of Bavaria (May 1999, August 2005), in the (south-)eastern Bavarian Forest and alpine areas (August 2002), and in the central lowland and northern hills (January 2011). The May 1999 flood was triggered by heavy rainfall in the alpine headwaters falling on soils saturated from a previous period of intense snowmelt (most affected basins: Iller and Lech) (BLfU, 2003). The flood events of 2002 and 2005 were exclusively caused by long and intense rainfall (most affected basins 2002: Inn and Regen; most affected basins 2005: Iller, Lech and Isar) (BLfU, 2002, 2007). The January 2011 flood was mainly induced by intense snowmelt and moderate rainfall (most affected basins: Naab, Altmühl and Regen) (BLfU, 2011). Even though the highest return periods (Fig. 2.2), and also most of the damages occurred at the tributaries, all the events

also caused severe damages along the Danube itself. We selected the flood events of May 1999, August 2002 and 2005 to investigate model performance and reservoir impact since, for them, the largest contributions to runoff volume in the Danube came from the rivers with the largest reservoirs: Iller, Lech (Forggensee) and Isar (Sylvensteinspeicher) contributed by 81 % (1999), 65 % (2002) and 85 % (2005) to the runoff volume in the Danube, as recorded at gauge HOFK (Hofkirchen). BLfU, (2002, 2003, 2011) provide more detailed information such as spatial distributions of rainfall or hydrographs of the major gauges. Böhm and Wetzel, (2006) and Mikhailova, Mikhailov, and Morozov, (2012) compiled general information on hydrological extremes in the Danube basin.

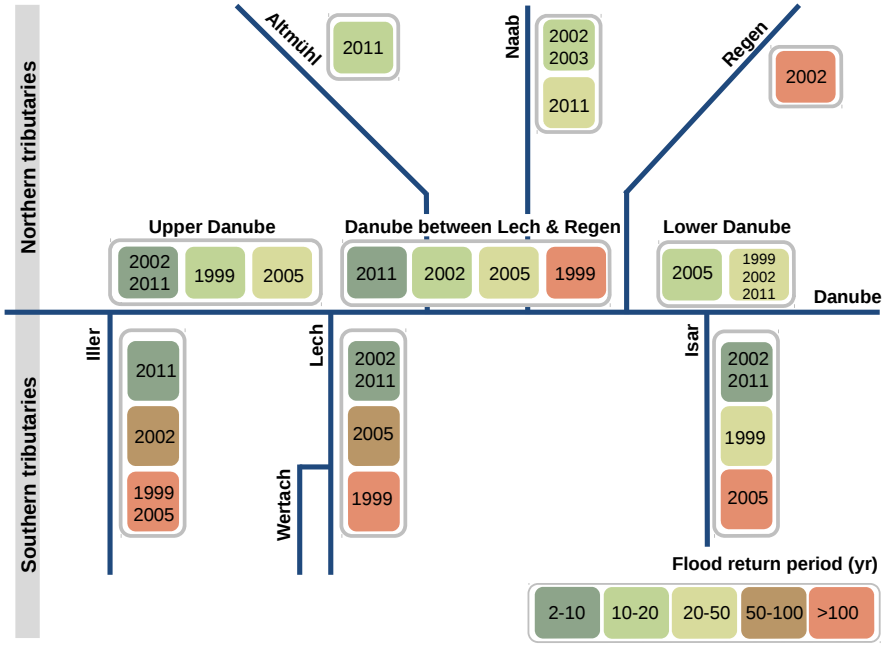


Figure 2.2: Sketch of the major river network (blue lines) of the Bavarian Danube catchment (excluding the Inn basin). The boxes represent the occurrence of floods and their respective return periods (color coded) in the period 1999-2011. Only floods with return periods larger than 10 years are shown.

2.2.4 *Evaluation of model performance*

Coordinated regional reservoir operation requires sufficiently precise simulations, which in turn requires methods to evaluate model performance. However, evaluating simulations in large basins represents a multi-dimensional problem. Numerous sites need to be judged numerous times with respect to numerous criteria (on numerous individual scales) and potentially even by numerous groups of users. To arrive at an evaluation that (i) includes all criteria considered important, (ii) allows a quick but comprehensive overview over all sites and events and, importantly, (iii) comes in a form interpretable also by non-experts, we have developed a method which (i) combines differ-

ent complementary evaluation criteria (Section 2.2.4.1), (ii) transforms and aggregates them into/on the same relative scale and (iii) allows for intuitive interpretation (Section 2.2.4.2).

2.2.4.1 *Single evaluation criteria*

The answer of how to adequately validate and quantify model performance is ambiguous, and many authors have dealt with it (Dawson, Abrahart, and See, 2007; Dawson, Abrahart, and See, 2010; Ehret and Zehe, 2011; Reusser et al., 2009; Smith, Georgakakos, and Liang, 2004; Vrugt et al., 2003; Weglarczyk, 1998; Willmott, 1981). An adequate representation of peak flow is of great interest in high flow simulations (Ramirez, 2000). For this reason we have included the amplitude-based Nash-Sutcliffe-Efficiency (NASH) (Nash and Sutcliffe, 1970). Despite its limitations (Gupta et al., 2009; Schaefli and Gupta, 2007) it is widely used as it has the advantages of non-dimensionality and high correlation to other metrics evaluating peak flow. The second criterion we have added to our set of evaluation criteria is the Peak-Time-Difference (PTDF). It accounts for timing errors and it is defined as the time difference between the simulated and the observed peak flow. Negative values indicate that the simulated peak flow has occurred prior to the observed one. Correct timing is an important aspect of model performance, as it strongly influences the superposition of flood waves at river confluences. Moreover, the PTDF is of high importance for the planning of dyke defense measures and reservoir operation.

The two criteria selected so far reflect how a hydrologist would evaluate model simulations, without additional knowledge of local conditions at points of interest in the catchment. However, the perspective of local residents or flood managers in a particular community on simulation quality may be substantially different: From their point of view, a model is a good model if it correctly discriminates between cases where no action is required and/or no damage occurs and such cases where large action is required and/or large damage occurs. In Bavaria, these cases are distinguished by the so-called "Meldestufen" (Flood Impact Levels (FIL)), available for all major gauges. FIL 3 denotes that built-up areas are flooded to a smaller degree, FIL 4 indicates large-scale flooding of built-up areas and that substantial efforts for dyke defense are required. Depending on local topography, land use and local flood protection, the water level difference between FIL 3 and FIL 4 may be smaller or larger. The point we want to make here is that local decision making is sensitive to local water level differences. Hence, it is advisable to also evaluate model performance based on these local sensitivities, For this purpose, we have developed a new evaluation criterion termed Peak-Level-Certainty (PLCY) and have added it to our set of criteria. It comprises the following steps and assumptions: We only consider peak water levels, as this is the decisive quantity for the extent of local damages and flood protection measures. We express the quality of a model by its ability to correctly distinguish cases where FIL 3 or FIL 4 is exceeded. This could

be done by counting "hits" and "missed events" in a classical contingency table. However, due to the limited number of floods available for analysis, this table would be very sparsely populated. Instead, we determine the mean absolute error between all available simulated and observed flood peak water levels (n) (this expresses model error) and divide it by the water level difference of FIL 4 and FIL 3 (this expresses local sensitivities). We have termed this ratio "relative error", re.

$$re = \frac{\frac{1}{n}\sum_{i=1}^{n}(|max(w_{sim} - w_{obs})|)}{\text{FIL 4} - \text{FIL 3}} \tag{2.1}$$

Under the simplifying assumption that the model error follows a uniform distribution and that it is valid for the entire range of flood water levels beyond FIL 3, the probability (PLCY) with which a peak flow predicted to be between FIL 3 and FIL 4 will actually be there in reality (and not below FIL 3 or above FIL 4), can be expressed as a function of re (Fig. 2.2). PLCY is high if re is low, which can be the case if either model errors are small, and/or if local conditions are not sensitive (i.e. the distance between FIL 3 and 4 is large).

A comparable approach has recently been suggested by Zappa, Fundel, and Jaun, (2013), who express local sensitivities by the difference between the 10-year and 5-year flood.

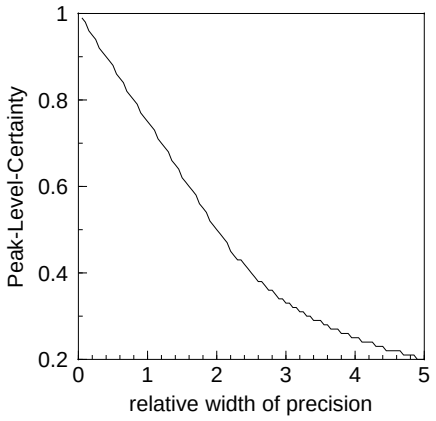


Figure 2.3: Relation between Peak-Level-Certainty and relative error (Eq. 2.1). The latter is the ratio of the mean absolute difference between simulated and observed peak water levels and the absolute difference between the local flood impact levels 3 and 4.

2.2.4.2 *Model performance expressed by multi-criteria grades*

To simplify the elaborate analysis, evaluation, aggregation and communication of the different criteria, we have developed a method which allows transforming, aggregating and combining them on a consistent relative scale. This way the procedure provides a meaningful overview of the spatial patterns of model performance in large basins. The relative evaluation scale fosters intuitive interpretation, which is relevant e.g. for non-experts and decision-makers. It should be noted, though, that the grading procedure is not intended to substitute the analysis of the individual performance statistics. Rather it complements it by providing a simple yet comprehensive overview which facilitates communication to non-experts and helps to focus

on relevant areas. Our approach combines and translates the three previously described evaluation criteria NASH, PTDF and PLCY into a single grade. From conversation with the forecasters, we have assumed the three criteria to represent all relevant aspects of flood-related model performance (magnitude, timing, and local sensitivities), to be non-redundant (see also Section 2.3.3) and to have equal importance (weights). A conversion matrix (Tables 2.3 and 2.4) translates the evaluation criteria into a five-category grade system, as it is used for example in the European Credit Transfer System and related grading tables (ECTS) or in many psychological questionnaires (Bühner, 2011). The categories are intuitively classified as "very good", "good", "satisfactory, "sufficient" and "insufficient".

Table 2.3: Nash-Sutcliffe-Efficiency grades from the literature.

Grade	Moriasi et al., (2007)	Haag et al. 2005, Hohenrainer et al., (2009)	Crochemore, (2011)	Applied ranges
Very good	> 0.75	> 0.9	> 0.77	> 0.85
Good	$]0.65 - 0.75]$	$]0.8 - 0.9]$	$]0.68 - 0.77]$	$]0.75 - 0.85]$
Satisfactory	$]0.50 - 0.65]$	$]0.7 - 0.8]$	$]0.63 - 0.68]$	$]0.66 - 0.75]$
Sufficient	-	$]0.6 - 0.7]$	$]0.55 - 0.63]$	$]0.55 - 0.66]$
Insufficient	< 0.50	< 0.6	< 0.55	< 0.55

The conversions we have selected (Tables 2.3) are adapted from existing literature (Crochemore, 2011; Crochemore et al., 2014; Moriasi et al., 2007) and expert knowledge from practitioners in the Bavarian Water Authority and engineering companies (Haag, Vollmer, and Heß, 2005; Hohenrainer et al., 2009). Even though the precise threshold values which are used to distinguish between two grades may be somewhat arbitrary, the general magnitude of the individual thresholds is not. This applies for all three criteria. For the sake of simplicity we have employed precise threshold values instead of fuzzy representations.

For timing errors, there was less literature available than for NASH (Cunge, 1969; Ehret and Zehe, 2011). Zappa, Fundel, and Jaun, (2013) provides some guidance concerning the evaluation of timing errors. In their study they defined tolerable timing errors as one fourth of the time between the initialization of a forecast and the median of the predicted peak timing. Since we did not work with forecasts, we defined the ranges of the PTDF values based on the idea that flood precautions required a certain lead time for preparation, irrespective of catchment size.

For this reason experiences gained by the Bavarian flood forecasting agency concerning the time required to e.g. initiate a warning, to conduct dyke defense and safety measures or to initiate a reservoir drawdown were used as criteria to define the value ranges for the PTDF statistic (Table 2.4) uniformly across all catchment sizes.

For PLCY, there were no literature values available. Hence we made reasonable assumptions based on our own judgment and comparable statistics such as the percent bias (Moriasi et al., 2007). Moreover, we tried to acknowledge measurement uncertainty. For streamflow this

is typically in the range of 10-15 % Harmel et al., 2006, peak level uncertainties in the same order of magnitude should still be evaluated as "very good" or "good" (Table 2.4).

Table 2.4: Ranges of values and corresponding grades for the Peak-Time-Difference (PTDF) and Peak-Level-Certainty (PLCY) criterion.

Grade	Peak-Time-Difference (h)	Peak-Level-Certainty (-)
Very good	< 3	> 90
Good	$[3 - 6[$	$]80 - 90]$
Satisfactory	$[6 - 9[$	$]65 - 80]$
Sufficient	$[9 - 12[$	$]50 - 65]$
Insufficient	$\geqslant 12$	$\leqslant 50$

The conversion of NASH (Tables 2.3), PTDF and PLCY (Table 2.4) to grades supports comparison and combination of different statistics. As one example for a simple multi-criteria evaluation we calculated the mean of the equally weighted three grades as additional criterion. It would of course also be possible to apply individual criterion weights according to specific user preferences. This procedure is in line with the suggestions given by Moriasi et al., (2007), who propose considering several and complementary criteria for single-event simulations. The weak correlations we have found between our performance statistics (the maximum r^2 was < 0.25 among NASH and PTDF) indicate the independence of our criteria. Nevertheless, it is important to state that every aggregation comes along with a loss in detail. Hence, any multi-criteria grades need to be interpreted with care since e.g. two positive grades may mask a negative one. A further limitation of our concept is that the data required for the calculation of the PLCY criterion (namely the FILs) may not be available and/or that FILs cannot be defined in a meaningful way. The latter would e.g. apply in remote areas where larger flooding does not cause any material damage.

2.2.5 *Reservoir impact estimation*

One goal of this study is to evaluate the potential of reservoirs for regional flood mitigation, and to evaluate to what degree this is possible without compromising reservoir operation optimized for local points of interest. We approached this question by investigating the maximum potential impact of each reservoir under "laboratory conditions". Here we defined the reservoir "range of impact" as the maximum river length downstream of the reservoir, where a noticeable (i.e. larger than the water level measurement uncertainty during high flow conditions, which is about 10 cm) peak water level reduction of $\Delta W_{crit} \geqslant 10$ cm was still possible. Within this range we assumed a potential for flood mitigation due to reservoir operation. If the range of impact extends into the Danube, the reservoir is potentially useful

for coordinated regional flood mitigation. As previously mentioned, the term "coordinated reservoir operation" can either refer to the operation of multiple reservoirs on a single regional gauge (e.g. reservoirs Grüntensee (RID = B) and Forggensee (RID = C) on gauge INGP (Ingolstadt), Fig. 2.1 and Table 2.1) or the joint operation of a single reservoir to multiple local and regional gauges (e.g. reservoir Forggensee (RID = C) to the gauges AUGW (Augsburg u.d. Wertach) and INGP). We applied both elaborate simulation-based and simplified statistically-based approaches and investigated to what degree the latter could be used as an easy-to-obtain approximation of the former.

2.2.5.1 *Reservoir impact estimation based on simulations*

The simulation based approach was applied to the three reservoirs we considered, due to their size and alpine setting, as potentially the most effective for our test floods: Sylvensteinspeicher (RID = D), Forggensee (RID = C) and Grüntensee (RID = B). To quantify their maximum possible impact, we defined the following idealized setting: Only physical reservoir characteristics like maximum possible water withdrawal (as function of water level and reservoir outlet characteristics), maximum tolerable lake water level lowering rates or critical velocities required for opening or closing of sluice gates, etc. were considered. Existing management plans and water law regulations were ignored since these would have restricted the evaluation to local sites (regional reservoir operation is at the moment not targeted in the Danube basin). By using observed rainfall, we further assumed full knowledge of the spatio-temporal distribution of rainfall during the flood event, i.e. we omitted the uncertainty associated with rainfall forecasts. Additionally, we maximized the usable retention volume of each reservoir prior to each flood event. For this purpose we allowed reservoir drawdowns starting 48 h prior to the peak inflow at the maximum possible rate to gain as much free retention volume as possible ($V_{R,max}$ in Table 2.1). Whereas the water level drawdown rate was pre-determined by the reservoir characteristics, the starting time 48 h prior to a flood event was not. It was determined by the operational experience gained at the Bavarian flood forecasting agency: 48 h prior to a flood, forecasts typically indicate with a high certainty that a flood will come. This justifies reservoir waterlevel drawdown even though it is often still uncertain how large the flood will be. During impounding, no spillway activation was allowed at any time. The same applied during the optimization of the reservoir release during the flood events.

Owing to the large number of boundary conditions and the well-known strengths of manual methods Boyle, Gupta, and Sorooshian, 2000, the optimization of the reservoir release was done manually and separately for each event and reservoir. Each reservoir was optimized for two target points, one local (the traditional) and one regional (the new) at the Danube (Table 2.1 and Fig. 2.1), applying the hydrological models (Table 2.2). The objective for optimization was to reduce the

water level at the selected target gauge below FIL 3, or, if impossible, to minimize the duration of its exceedance. To avoid local flooding as an unwanted side-effect of regional reservoir optimization, we defined an additional constraint: Any increase in water level above FIL 3 due to reservoir operation was not permitted at any point between the reservoir outlet and the target gauge.

Since it was not possible to analyze the impact of a reservoir independent from the impact of the other reservoirs and independent from the contribution of the different sub-basins, we considered them as neutral as possible. This was done by using observed discharge time series as model input at the outlets of all other reservoirs and at the last gauge in the Danube prior to the regional target gauge of the reservoir of interest.

The impact of the reservoirs was evaluated not only at the local and regional targets, but for each gauge downstream of the reservoir outlet by comparing the optimized water level to a reference water level. The spatial range of the effective reservoir impact was then given by the last gauge where the resulting impact was $\geqslant \Delta W_{crit}$. Reference water levels were generated using the same data and models but with a different reservoir release ($Q_{ref,release}$). The latter was defined as $Q_{ref,release} = min[max.\ physically\ possible\ release, inflow]$. Thus, in the reference case, reservoir outflow was equal to reservoir inflow except for very few cases, thereby mimicking a non-existent reservoir. Calculated discharge was converted to water levels using stage-discharge relationships following the Eta-approach (Pegelvorschrift, 1991).

2.2.5.2 *Reservoir impact estimation based on specific retention volume*

Estimating the potential impact of a reservoir by iterative manual optimization as described in the previous section is a laborious task. Therefore, we also sought for simpler alternatives based on easy-to-obtain reservoir indices to establish a connection between the simplified and advanced approach. The aim was then to use the former as a surrogate for the latter.

We used a simple but meaningful index, the specific retention volume (V^*), see Fischer, (2008). It relates the maximum possible retention volume ($V_{R,max}$) of a reservoir to the size of its catchment (A) and to the relative mean annual precipitation within its catchment:

$$V^* = \frac{V_{R,max}}{A} \cdot \frac{MAP}{MAP_D} \tag{2.2}$$

We expressed the relative mean annual precipitation by the ratio of the mean annual precipitation (MAP) in a given catchment and the mean annual precipitation within the entire Danube catchment (MAP_D). Consequently, V^* can be computed for each reservoir, but also for each point (gauge) downstream of the reservoir outlet by simply modifying A and MAP accordingly. The progression of V^* along a river offers a simple way to study the decline in reservoir

impact with increasing distance from the reservoir outlet. It also allows comparing reservoirs from different catchments with different hydro-meteorological regimes.

We identified all reservoirs whose operation potentially influence the water levels in the Danube during high flow conditions by combining the results from the simulation based assessment of reservoir impact and the specific retention volume.

2.3 RESULTS AND DISCUSSION

Although the main focus of this article is the analysis of reservoir impact, we have found interesting results in the evaluation of model performance. We will briefly discuss results related to the hydrological simulations in Section 2.3.1. Findings from the coupled hydrologic-hydrodynamic simulations are presented in Section 2.3.2 and an overview of the most important results from the grade based performance evaluation is given in Section 2.3.3.

2.3.1 *Hydrological simulations*

2.3.1.1 *General model performance*

All analyzed flood events originated in the southern or southeastern alpine areas. With respect to the overall hydrological model performance we observed only little differences between the three events. Averaged over all gauges, all three events had nearly identical NASH values (Table 2.5). With respect to PTDF, we observed a slightly negative average, which indicated that the timing of the simulated peak flows tended to be about 1 h too early in general. Even though the averaged PTDFs were rated "very good", standard deviations between 8 and 14 h indicated a distinct scatter (Table 2.5). Please note that with respect to our third statistic, Peak-Level-Certainty (PLCY), the most interesting results are related to its spatial distribution. We will discuss this in Section 2.3.3.

Table 2.5: Median of NASH (rather than mean due to the highly skewed distribution), average and standard deviation of Peak-Time-Difference (PTDF) for all three flood events.

Flood event	NASH (-)	Peak-Time-Difference (h)
May 1999	0.77	-2.2 ± 14
August 2002	0.82	-0.7 ± 10
August 2005	0.8	-0.9 ± 7.7
Average	0.8	-1.2 ± 10.8

The scatter between the three events was large, and high PTDFs occurred both in combination with poor and good NASH values. The two criteria thus, are independent ($r^2 < 0.01$, p-value=0.76, n=175).

Contrary to our expectations, we did not find any significant correlation between catchment size and model performance.

A general result valid for all simulations and throughout the entire basin was that all model runs overestimated the total volume by up to 10 %. This applied in particular for the tributaries Iller and Isar. The simulations at the river Iller overestimated the peak flow at the gauge NEUL (Neu Ulm, Bad Held) between 50 (May 1999) and 150 $m^3 s^{-1}$ (August 2005). The PTDF differed between -9 (May 1999) and -14 h (August 2002) at the same gauge. Hence, already at the beginning of the Bavarian Danube basin, considerable differences between simulated and observed discharges occurred. Along the Danube, between the gauges NEUL and PILZ, the NASH and PTDF statistics showed non-uniform progressions, PTDF with a parabolic shape and distinct changes at the mouths of larger rivers, NASH with a more consistent decrease (Fig. 2.4).

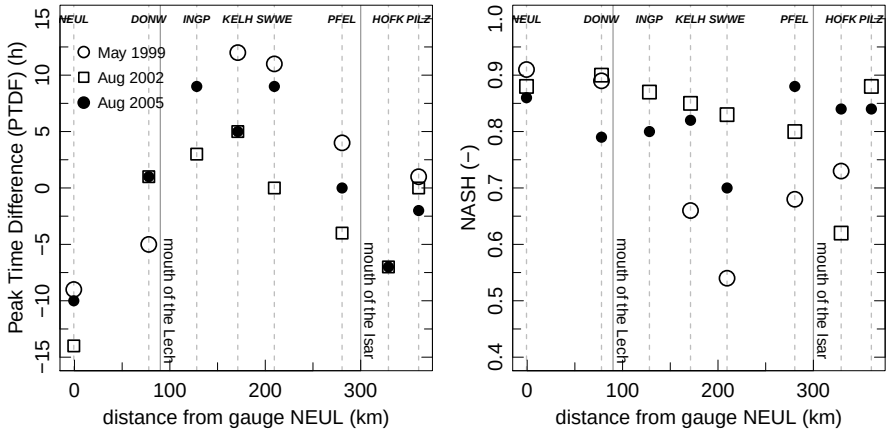


Figure 2.4: Progression of Peak-Time-Difference (PTDF) *(left)* and Nash-Sutcliffe-Efficiency (NASH) *(right)* along the Bavarian Danube for all three events starting from the gauge NEUL. Going from left to right the dotted lines mark the distance of the gauges DONW, INGP, KELH, SWWE, PFEL, HOFK and PILZ from the gauge NEUL (see also Fig. 2.1). The solid lines represent the distances of the mouths of the tributaries Lech and Isar from NEUL. One PTDF value from the gauge INGP (-22 h) and one NASH value from the gauge PILZ (-0.31), both from the flood event in May 1999 are not shown. Values for gauge INGP and the flood event 1999 were not available due to missing data.

2.3.1.2 *Model performance as a function of flood magnitude*

To evaluate model performance in terms of flood magnitude, we plotted the NASH and PTDF values for all gauges against the maximum observed flow $Q_{max,obs}$, which was normalized with the 100-year flood flow rate (HQ100) to allow for better comparison (Fig. 2.5). Both plots show two clusters of points. The first one represents basins unaffected by larger flooding (normalized $Q_{max,obs} \leqslant 0.25$). The second cluster contains gauges which have experienced floods during our periods of simulation (normalized $Q_{max,obs} \geqslant 0.25$). In general, the

higher the observed stream flow, the better the NASH (Fig. 2.5), the smaller the scatter of the data (both panels), and the greater the tendency of simulated flood peak to run ahead of the observed ones (linear regression in Fig. 2.5, left).

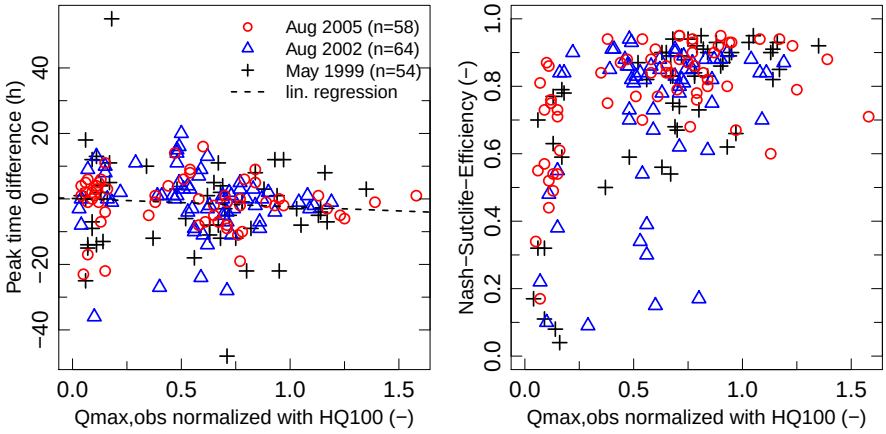


Figure 2.5: Scatterplots of the results of the hydrological model. They show Peak-Time-Difference (h) *(left)* and Nash-Sutcliffe-Efficiency (-) *(right)* for all available gauges (n) and for the three different flood events, plotted against the maximum observed discharge $(Qmax, obs)$, normalized by the 100-year discharge.

The poor results in areas with normalized $Qmax, obs < 0.25$ (Fig. 2.5) can best be explained by the previously discussed application of the models "out of their scope" rather than by physiographic characteristics of the basin. We tolerated the poor performance which occurred mainly in the northern catchments (Altmühl and Naab), as the associated flows did not significantly contribute to the regional flood wave propagation and due to our focus on the southern basins which experienced larger floods (see again Fig. 2.2), where we could analyze the potential of coordinated reservoir management.

2.3.2 *Coupled hydrological and hydrodynamic simulations*

To evaluate the potential of the hydrodynamic models, we repeated all stream flow simulations, but this time the simplified hydrological routing was replaced by the 2d-hydrodynamic model chain in the major reaches. All boundary conditions remained unchanged, and we calculated the same performance statistics as before. By comparing the results between purely hydrological (HY) and coupled (HY&HD) simulation (Fig. 2.6) it became obvious that, against our expectations, the 2d-hydrodynamic simulations did not significantly increase the overall model performance. Concerning the PTDF, neither approach outperformed the other (Fig. 2.6, left). Although average and standard deviation of the coupled simulation were better than the HY results for the events May 1999 and August 2002 (HY: -3.1 ± 11, 1 ± 7 h vs. HY&HD: -0.7 ± 12, -0.2 ± 8 h), the opposite was the case for the flood event of August 2005 (HY: $< 0.1 \pm 6$ h vs. HY&HD: -3.3 ± 8 h). In terms of the median NASH, the HY simulations even outper-

formed the HY&HD approach for the flood events of May 1999 and August 2005 (HY: 0.71, 0.84 vs. HY&HD: 0.61, 0.72). For the flood event of August 2002 both approaches yielded an identical median NASH of 0.84 (Fig. 2.6, right).

Hydrodynamic modelling results were provided by Skublics, (2014).

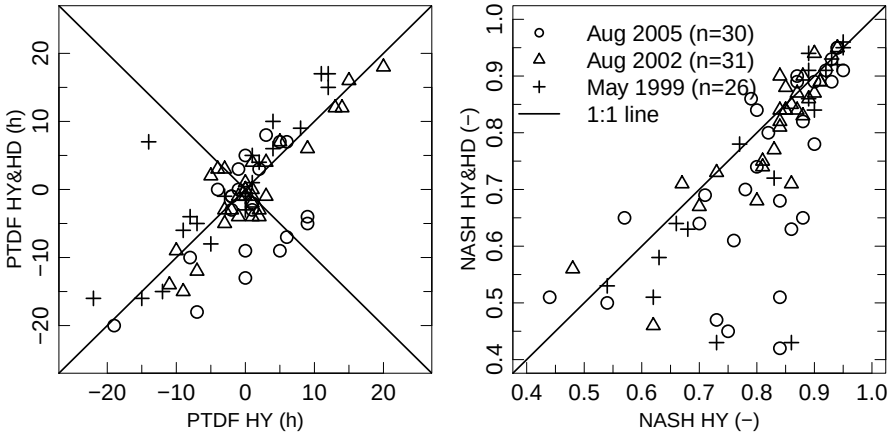


Figure 2.6: Peak-Time-Difference (PTDF) and Nash-Sutcliffe-Efficiency (NASH) of the hydrological model (HY) plotted against the results from the coupled hydrological-hydrodynamic (HY&HD) simulations. The number of considered gauges (n) is smaller than in Fig. 2.5 since it only contains gauges were where both HY and HY&HD simulations were available.

Apparently, the generally known superiority of hydrodynamic models over conceptual hydrological routing approaches did not apply in our case, and the results did not justify the large additional effort of applying the hydrodynamic models.

To analyze this unexpected behavior in more detail, we conducted an additional experiment. For this purpose we selected the stretch of the Danube between the gauges NEUL and DONW which is about 80-km long and characterized by low slopes ($\approx 0.9‰$) and extensive inundation areas. Here, flood wave propagation is subjected to substantial and complex deformations due to floodplain activation, backwater effects and meander shortcuts (Skublics, Fischer, and Rutschmann, 2009), and hence the hydrodynamic models, which, in contrast to the hydrological models, explicitly represent these effects, should clearly outperform the latter. Along this river stretch we replaced, where available, the hydrological simulations at all lateral and upstream inflow points with observations. Varying slightly among the floods, roughly 80 % of the catchment area could thus be represented by observations. We then fed the routing scheme of the hydrological model and the hydrodynamic model with these "perfect data" and compared their performance. The results clearly justified the use of the $2d - HD$ model in this river section. Even though the accuracy of both simulations increased (compared to the case with simulated lateral inflow), the NASH, PTDF and volume error of the HD model improved disproportionally more (see Fig. 2.7). Thinking backwards, this also implies that the HD model suffers disproportionally more from the lower quality lateral inflow data provided by hydrological

models than the hydrological routing schemes do. As a consequence, inaccuracies provided by a HY model may not necessarily be compensated by the subsequent use of a HD model.

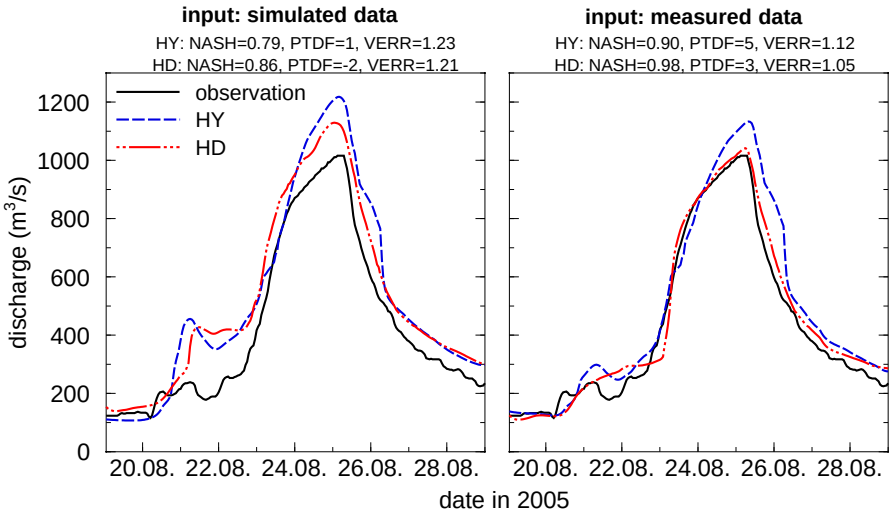


Figure 2.7: Comparison of hydrological (HY) routing (blue) and 2d-hydrodynamic (HD) modeling (red) at the end of a river stretch of the Danube of 80 km length (gauge DONW). The *left plot* shows results where data from the hydrological model was used as input. In the *right plot* measured streamflow data (covering 83 % of the catchment) was used as model input. The statistics in the top provide NASH = Nash-Sutcliffe Efficiency, PTDF = Peak Time Difference and VERR = Volume Error.

Skublics, Seibert, and Ehret, (2014) conducted additional experiments to study the interaction between both modeling approaches in more detail. Using both observed and synthetic hydrographs, they assessed the influence of errors in volume, peak-flow magnitude and peak timing on the simulation accuracy of the hydrological routing and the 2d-hydrodynamic model. They showed that equivalent modifications of the boundary conditions led to different and non-uniform reactions of the HY and the HD models. It also turned out that errors in the peak timing and/or the run time significantly affected the superposition of the flood wave of the main river stem with that of the tributaries. In their example, the resulting errors did not cancel out but (partly) even amplified each other. Moreover, the deviations were not uniform but linked to the magnitude of discharge. In summary, this analysis underlined that in order to tap the full potential of hydrodynamic models to improve simulations in large river systems, sufficiently accurate boundary conditions are required.

2.3.3 *Grade-based evaluation of model performance*

To get a meaningful overview of the spatial patterns of the performance of the hydrological models we developed a method which allowed transforming, aggregating and combining the individual criteria to a consistent relative scale. The grading procedure did not

substitute the analysis of the individual performance statistics but provided additional information which helped us to focus on areas particular sensitive to flooding and thus, to the reservoir operation. As described in Section 2.2.4.2 we converted our evaluation criteria PTDF, NASH and PLCY into grades. Aggregated over all gauges and all events, NASH, PTDF and PLCY followed a nearly identical distribution (Fig. 8). At around 20 % of the considered gauges, "satisfactory" modeling results were achieved, about 40 % of them were better ("good" or "very good"), and 40 % were worse ("sufficient" or "insufficient"). In contrast to these more or less uniform distributions, the distribution of the combined multi-criteria grade could be approximated by a Gaussian shape which had a clear maximum in the middle (i.e. overall "satisfactory" model performance). This underlines that the three performance criteria seem to be uncorrelated.

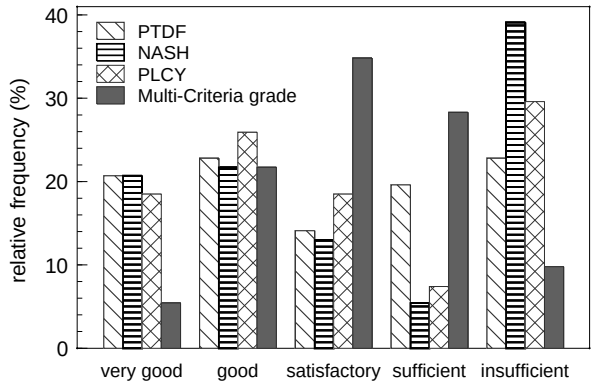


Figure 2.8: Histogram of model performance. The bars represent the relative frequency of all achieved marks based on the criteria Peak-Time-Difference (PTDF), NashSutcliffe-Efficiency (NASH) and Peak-Level-Certainty (PLCY). The bar of the multicriteria grade represents the frequency of the paired average of the three grades for all individual gauges. Data from all three flood events and all gauges were considered.

To analyze dominant spatial patterns of the performance of the HY model, we drew maps of the different criteria (Fig. 2.9). They clearly showed that the individual criteria possessed distinct spatial patterns throughout the basin and that these patterns differed among the criteria.

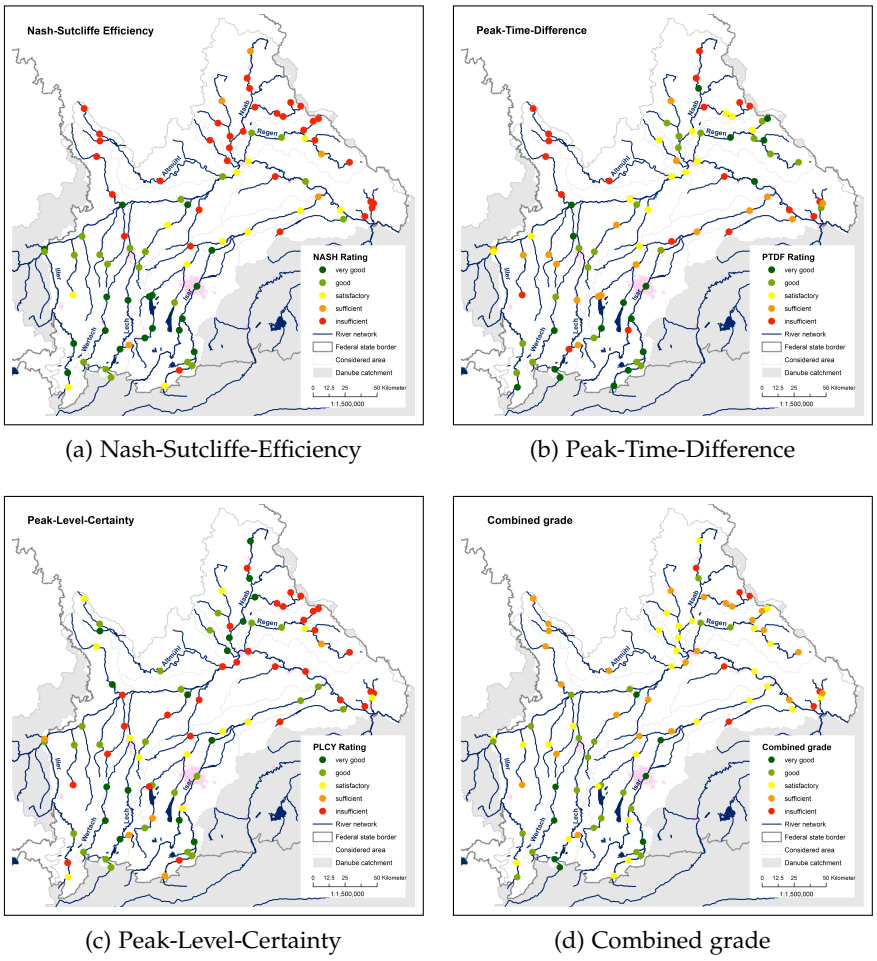
The calculated NASH values were predominately classified as "good" or "very good" in the sub-catchments of the rivers Iller, Lech, Isar and on the river sections of the upper and central Danube (Fig. 2.9, panel a). This is important since good and/or very good NASH values indicate a good representation of the observed flood peak by the simulation. This implies in turn that a precise reservoir operation should be possible within these areas (given effective reservoirs), since the uncertainty associated with the simulation is rather small compared to its accuracy. The opposite is the case in the northern basins and also in the lower reaches of the river Isar. Here we have found small NASH values which indicate higher uncertainty and in turn, that it is difficult to define an effective reservoir operation strategy. As stated

in Section 2.3.1.2, we believe that the poor model performance we have observed in these basins can best be explained by the use of the models "out of their scope" rather than by physiographic characteristics of the basin.

Regarding the PTDF statistic (Fig. 2.9, panel b), a completely different picture emerged. "Good" and "very good" results were achieved in the headwaters of the rivers Iller, Lech, Isar and Regen. This result suggests that we can expect accurate inflow simulations at all alpine reservoirs with respect to the timing, given accurate precipitation data. This finding is fundamental since it enables precise reservoir management which is mandatory for long-distance reservoir operation for flood mitigation during larger flood events. More heterogeneous results were observed at most of the smaller southern tributaries of the Danube and on most river sections of the Danube itself. The PTDFs in the northern sub-catchments were "good" on average, particularly those in the Regen basin. Poor model performance with respect to timing was found in the Altmühl and partly in the Naab catchment. The spatial patterns of the PTDF criterion approximately coincide with the spatial patterns of stream flow velocity. The northern basins Altmühl and Naab but also several of the smaller southern tributaries show comparably slow stream flow velocities ($2-4$ km h^{-1} on average) and damped high flow responses. Typically, flood waves within these basins are characterized by slow rises and recessions and a broad (and rather poorly defined) peak. This applies in particular in comparison to the larger alpine tributaries of the Danube (typical stream flow velocities differ between 7 and 14 km h^{-1}) where the opposite is the case. Comparable effects occur on rivers fed by lakes or in the very downstream sections of larger rivers. Here the significance of the PTDF criterion is reduced.

The PLCY criterion exhibited the most discontinuous behavior (Fig. 2.9, panel c). This manifests particularly along the rivers Isar and Regen, where the grades cover the entire range of values. The best results were achieved along the rivers Lech, Altmühl and Naab. For the Danube, "insufficient" results were achieved for almost the entire river section downstream of gauge DONW. The PLCY criterion can be interpreted as an indicator for areas which require particular attention in flood management, either because of poor model performance or because of local sensitivity to water level differences. Several of the gauges marked "satisfactory", "sufficient" or "insufficient" (colored yellow, orange and red in Fig. 2.9, panel c) coincide with areas that have experienced severe damages during the flood events of May 1999, August 2002, and August 2005.

Like the individual grades for NASH, PTDF and PLCY, the combined grade does not show a trend of increasing model performance with increasing catchment size (Fig. 2.9, panel d), but other patterns are visible: Grades along the rivers Iller, Lech and Isar (including their tributaries) have predominately been classified as "good" or "very good". "Satisfactory" or "sufficient" results have mainly been limited to a few head waters gauges and to the lower reaches of the river Isar. Worse overall model performance and hence potential for improve-



(a) Nash-Sutcliffe-Efficiency

(b) Peak-Time-Difference

(c) Peak-Level-Certainty

(d) Combined grade

Figure 2.9: Spatial patterns of model performance. The color coded dots represent the locations of gauges throughout the basin where results for NASH (a), PTDF (b) and PLCY (c) are available. The combined grade is shown in the bottom right panel (d).

ments is visible at several of the smaller southern tributaries of the Danube. On the Danube, good results have only been achieved upstream of the river Altmühl mouth. Beyond, overall grades declined and remained "satisfactory" or "sufficient" along the remaining Bavarian Danube.

2.3.4 *Reservoir impact assessment*

2.3.4.1 *Reservoir impact assessment based on simulations*

We discuss results from the simulation-based approach here mainly for the example of the Forggensee, as it has proved to be the most effective of the tested reservoirs. It is located on the river Lech (RID = C in Fig. 2.1), and its influence is clearly detectable throughout the entire river Lech and at several gauges of the Danube. Moreover, we found that traditional reservoir management practices with a local focus were, on the Forggensee, not necessarily in conflict with reservoir operation for regional flood mitigation. The following results were

derived from an individual analysis of the Forggensee reservoir. Observed discharge time series were used at the outlet of the Grüntensee reservoir (RID = B, river Wertach) and at the gauge DONW which is the last gauge in the Danube prior the confluence of Danube and Lech.

During the May 1999 flood, the reservoir inflow exceeded return periods of 100 years and the potential of the reservoir for flood mitigation was limited. The reservoir operation plot (Fig. 2.10, left column) and the reservoir impact plot (Fig. 2.11, left column) showed that local flood protection required the entire available retention space and that there was no room for reservoir optimization with respect to the regional gauge INGP. Already tiny modifications in the reservoir release led to an increase in water level above FIL 3 at the local gauge AUGW. The operation of the reservoir allowed a reduction (relative to the reference, see Section 2.2.5.1) of the water level of about 1 m at the local target gauge AUGW. However, the resulting water levels exceeded FIL 3, despite the significant reservoir drawdown which was initiated 48 h prior the peak inflow. At the distant gauge INGP, both local and regional optimization reduced the water levels by about 60 cm (Fig. 2.11, bottom left).

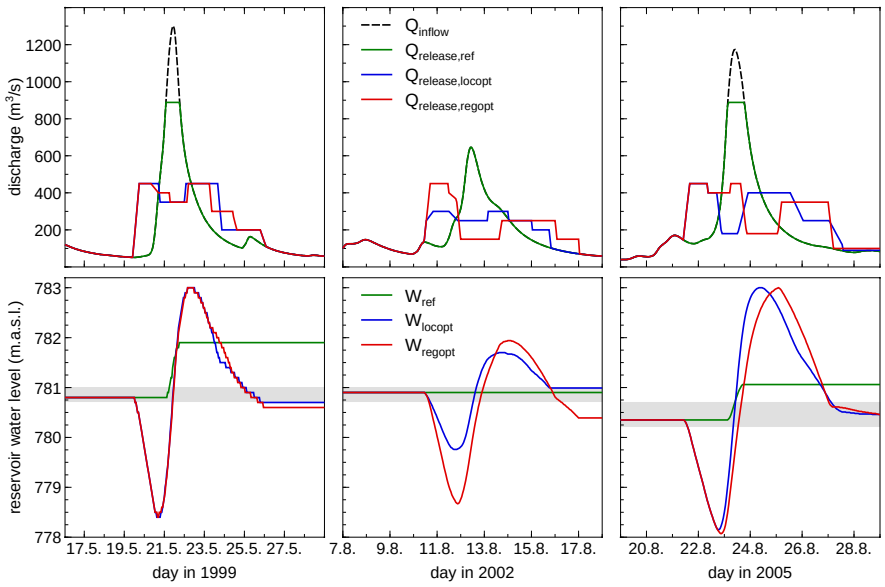


Figure 2.10: Operation of the Forggensee during flood events May 1999 (left column), August 2002 (center column) and August 2005 (right column). The upper row contains the reservoir inflow and releases for the reference scenario (ref), local (locopt) and regional (regopt) optimization. The lower row shows the corresponding water levels. The gray areas represent the standard operating water level segment, which differed between the three events.

In contrast, the potential for regional reservoir operation during the August 2005 flood was much higher. While the reservoir inflow again exceeded a 100-year flood, this time the initial lake water level was lower, leaving room for optimized operation. In the case of local optimization, the maximum water level at the local target gauge AUGW

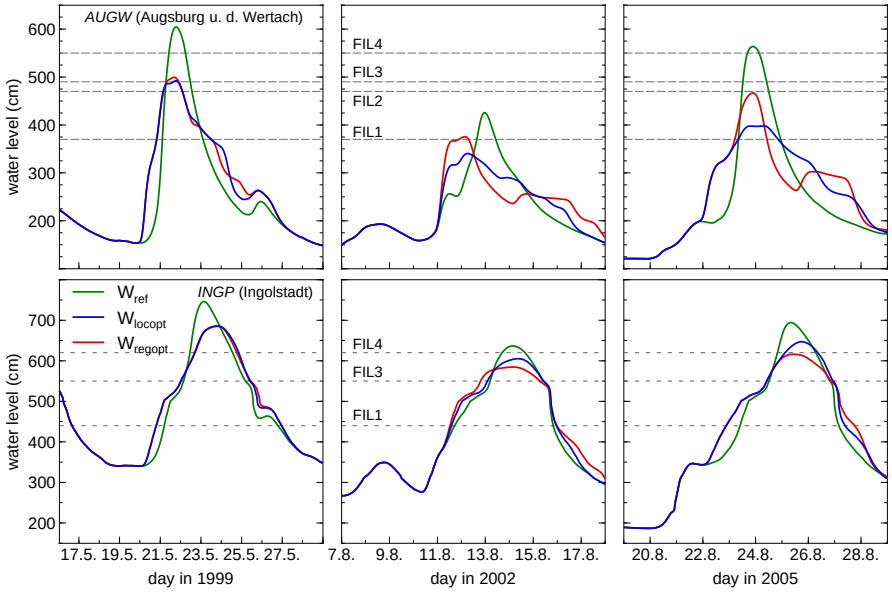


Figure 2.11: Water levels at the local gauge AUGW (upper row) and the regional gauge INGP in the Danube (lower row) as a result of the Forggensee operation. From left to right: Flood events May 1999, August 2002 und August 2005. The dashed lines mark the different flood impact levels (FIL). The colored lines represent the reference scenario (ref) in green, the local (locopt) optimization in blue and the regional (regopt) optimization in red.

was reduced by more than 2 m, and by more than 50 cm at the regional gauge INGP, compared to the reference. The latter could be reduced even further to about 1 m, when the reservoir operation was optimized for regional flood mitigation, even under the additional constraint of keeping water levels at local gauge AUGW below FIL 3. The local and regional reservoir operation strategies differed substantially, in particular with respect to timing (compare blue and red line in Fig. 2.10, top right). Regional optimization did not require a horizontal cut of the flood wave of the Lech (as for the case of local optimization), instead the highest possible discharge reduction during the falling limb was important (see red line in Fig. 2.11, top right). The offset in timing between the two optimization strategies was approximately one day during the flood event of August 2005. This indicates that local operation strategies not necessarily interfere with regional operation strategies and that it is possible to combine both local and regional flood protection issues (given sufficient retention volume).

The flood event of August 2002 takes an intermediate position. It was smaller (reservoir inflow was below a ten-year-flood), and the water levels at local gauge AUGW would have remained below FIL 2 even for the reference scenario (assuming a non-existing reservoir). Thus, local flood protection did not require any action. Nevertheless, we modified the reservoir release to test the potential of the reservoir regarding its influence on the regional flood wave propagation. Whilst the setting of the local optimization scenario was, for the above reasons, rather arbitrary, we were able to achieve a significant reduc-

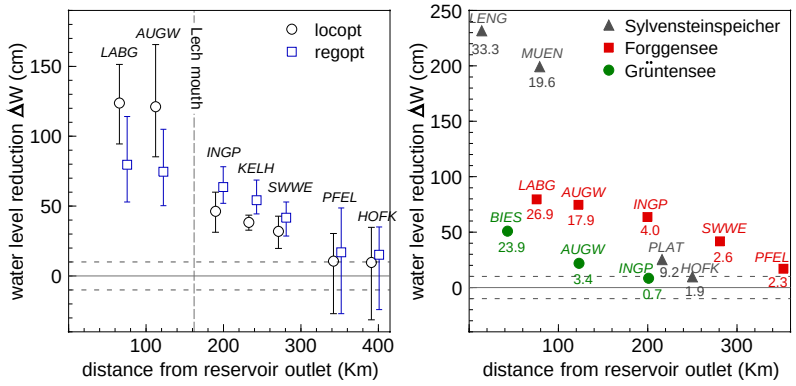
tion in water level at the regional gauge INGP through regional optimization. In the latter case we decided to initiate a much larger reservoir drawdown than required, which resulted in an artificial pre-event flood wave of tolerable height (see Figs. 10 and 11, top center), in favor of additional retention volume during the flood itself. We were thus able to throttle the reservoir release during the flood event and to reduce the resulting water level at the regional gauge INGP (Fig. 2.11, bottom center). This example illustrates the potential of the reservoir even for smaller flood events. This applies in particular since we tightened the boundary conditions in this case to a maximum tolerable water level at the local gauge AUGW to FIL 1. Thus, the results show that it can be worthwhile to consider regional flood mitigation measures even if local conditions are not critical.

The differences between the traditional "locopt" reservoir operation practice and the new "regopt" strategy become obvious when plotting the maximum possible reduction in water level against the distance from the reservoir (panel (a) in Fig. 2.12). It clearly shows that "locopt" operation ensures higher flood protection for the immediate downstream area, while "regopt" operation achieves higher water level reduction along the Danube. The average differences between the two strategies amounted to up to 30 cm at the gauge INGP. In any case, the reservoir operation was detectable over more than 300 river kilometers. Beyond roughly 350 km, the average reduction in water level was less than 10 cm, i.e. no more effect was achieved according to our cut-off criterion. The right panel in Fig. 2.12 shows the average water level reduction achieved for all three reservoirs where simulations have been made as a function of distance from the reservoir.

2.3.4.2 *Combined reservoir impact assessment based on specific retention volume and simulations*

We computed the specific retention volume (V^*, see Section 2.2.5.2) for all reservoirs and all downstream gauges in the Danube basin. It decreases with increasing distance from the reservoir as the catchment size increases (Fig. 2.13).

Based on the detailed, simulation-based information on reservoir impact (Fig. 2.12) and the easy-to-obtain specific retention volume (Fig. 2.13), we were able to formulate a simplified rule to determine the range of reservoir impact. As can be seen from Fig. 2.13, water level reductions larger than 10 cm (our cutoff criterion) only occurred when specific retention volume was larger than $1 - 2 \cdot 10^3$ m^3 km^{-2}. We set a rough but conservative limit of 1.5 m^3 km^{-2} and transferred this "translated" cutoff criterion to Fig. 2.13 (indicated by the dashed line). With this, we could approximate the range of impact for all additional reservoirs where no simulations were available. The progressions of V^* in Fig. 2.13 showed that the range of impact along the rivers differed among the reservoirs between roughly 25 and 100 river kilometers. Moreover, it became clear that none of the additional reservoirs had the potential for flood mitigation in the Danube: At the



(a) Results for the rivers Lech and Danube due to the operation of the Forggensee reservoir. The graph shows the mean, max and min values achieved for the flood events May 1999, August 2002 and August 2005. Black circles denote the traditional operation practice (locopt), blue squares the results obtained by regional optimization (regopt).

(b) Comparison of the average flood peak water level reductions in the rivers Lech, Isar and Danube due to the individual and regionally optimized operation of the reservoirs Forggensee, Sylvensteinspeicher and Grüntensee. The numbers indicate the specific retention volume V^* in (10^3 m^3 km^{-2}).

Figure 2.12: Simulated water level reductions compared to the reference scenarios. The panels show results for different reservoirs and gauges as a function of the distance of the gauge from the reservoir outlet. The gauge names are coded in italic capitals, Fig. 2.1 depicts their location.

first gauge in the Danube, V^* was always well below 1000 m^3 km^{-2} in all cases, except for the Sylvensteinspeicher ($V^* \approx 2000$ m^3 km^{-2}).

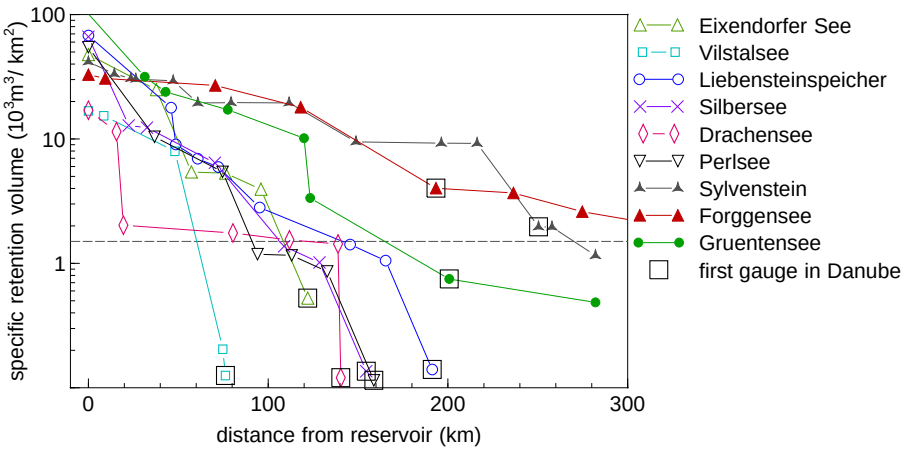


Figure 2.13: Specific retention volume (V^*) as a function of the distance from the reservoir. Each marker represents the location of a gauge. The first marker is always the reservoir itself. The encircled marker of each series represents the first gauge in the Danube. The dashed line equals the cut-off value of $V^* = 1.5$ 10^3 m^3 km^{-2}, indicating the end of measurable reservoir impact.

In summary, we determined the range of impact of nine reservoirs located on Danube tributaries to evaluate their potential usefulness for coordinated regional flood mitigation for the Danube. We performed detailed simulations for the largest three reservoirs. From this, we identified a lower limit of effectiveness expressed as a minimum specific retention volume V^* and applied this criterion to the remaining reservoirs. None of them fulfilled this requirement, in fact, only one out of the nine reservoirs (Forggensee) fulfilled it. When evaluating these results, however, it should be kept in mind that they were obtained under idealized conditions (full knowledge of rainfall, no limitations due to existing management plans or water law regulations); under real conditions, the range of impact would be even smaller.

2.4 SUMMARY AND CONCLUSIONS

The overall goal of this paper was to present and discuss methods potentially useful to set a basis for coordinated reservoir operation on the regional scale. In particular, we (i) investigated methods for regional multi-criteria evaluation of model performance, (ii) evaluated the potential of joint hydrologichydrodynamic modeling to improve model performance and (iii) applied simulation-based and simplified methods to determine the impact range of reservoirs. Finally, we investigated to what degree reservoir operation strategies optimized for local and regional points of interest were compatible.

Our investigations were done using the example of three major flood events in the basin of the Bavarian Danube (45,000 km^2) and nine reservoirs therein. Simulations in the basin were done with eight event-based hydrological (HY) models and nine 2-d hydrodynamic (HD) models covering 1300 river kilometers. The HY outputs were directly fed into the HD models without applying any kind of joint calibration.

2.4.1 *Evaluation of model performance*

Regional flood mitigation requires hydrological/hydrodynamic simulations on a large scale, potentially from many models and for many points of interest, and thus it requires, as a prerequisite, suitable methods for model evaluation. We established such an evaluation procedure with a focus on flood simulations. It combines three independent (maximum $r^2 < 0.25$) criteria which evaluate flood magnitude (NASH), peak timing (PTDF) and a criterion (PLCY) which relates simulation quality to local quality requirements. The three criteria were transformed and combined into a five-level grade system, which facilitates large scale model evaluation and detection of spatial patterns of model performance. Applied to three flood events and approximately 100 gauges in the Bavarian Danube basin, about 40 % of all simulations were rated either "good" and "very good" or "sufficient" and "insufficient". The remaining 20 % were rated "satisfactory",

which corresponds to NASH values between 0.66 and 0.75, PTDFs of $|6 - 9h|$ and PLCYs of 65-85 %. On average, simulation was "satisfactory", which corresponds to studies in comparable catchments (Mauser and Bach, 2009). The spatial patterns of model performance in the Bavarian Danube basin varied considerably among the individual criteria and the combined result, which suggests that the criteria are indeed complementary.

2.4.2 *The potential of coupled hydrological and hydrodynamic simulations*

2-d hydrodynamic models based on the shallow-water equations have a sound physical basis and require, compared to simplified hydrological routing techniques, less calibration. Adding hydrodynamic models along all major rivers in the Bavarian Danube basin to the model chain, we therefore expected simulation results to improve, especially along the Danube, where during floods large inundations occur. To our surprise, this was not the case: Recalculations of three historical flood events using hydrological routing schemes and 2-d hydrodynamic models yielded comparable errors. Within a sensitivity test, we substituted all hydrologically simulated input into an 80-km stretch of the Danube (upstream inflow and lateral tributaries) by measured stream flow data. The results indicated that especially the hydrodynamic approach profited disproportionally strong from the improved boundary conditions (or, in other words, suffered disproportionally strong from poor boundary conditions). From this we conclude that the use of 2-d hydrodynamic models is in principle superior to hydrological routing schemes, but requires boundary conditions of sufficient quality. A possible explanation for the more robust behavior of the hydrological routing schemes is that they are jointly parameterized with the hydrological model, thus allowing for compensation of model weaknesses in the runoff generation by the routing. If such parameter interactions are present (see also Kirchner, (2006)), difficulties will naturally arise if the hydrological routing is substituted by an external 2d-hydrodynamic model, since all compensatory effects will be removed. It could hence be interesting to investigate to what degree joint calibration of the hydrological rainfall-runoff schemes and the hydrodynamic models circumvents this problem.

2.4.3 *The potential of coordinated reservoir operation for regional flood mitigation*

Typically, reservoirs are designed and operated in a way to maximize (local) flood mitigation in the downstream vicinity. The question we pursued in this study was to evaluate to what degree reservoirs might, beyond local considerations, also be available for flood mitigation at more distant (regional) points of interest along larger rivers. For nine reservoirs situated on tributaries of the Bavarian Danube, we investigated this under idealized conditions, i.e. perfect knowledge of rainfall. We determined their range of impact, defined as the maximum

river length downstream of the reservoir where the flood peak water level could still be lowered by at least 10 cm. For the three largest reservoirs, Forggensee, Sylvensteinspeicher and Grüntensee we determined the range of impact with detailed hydrological simulations and could establish a relationship between the reservoirs" range of impact and specific retention volume. We then applied this relationship to the remaining reservoirs. From the nine reservoirs tested, even under the described idealized conditions, only one (Forggensee) had a detectable regional impact, which vitiated the initial idea of coordinated multi-reservoir flood mitigation in the Danube. However, an interesting finding for the Forggensee was that reservoir operation strategies optimized for local and regional flood mitigation, respectively, were not mutually exclusive but could be simultaneously pursued, provided that sufficient retention volume was available. While these results are of course highly specific to the particular reservoirs we investigated, we suggest that two reservoir-related findings of this study are more generally applicable: The first is that local and regional reservoir operation strategies are not necessarily exclusive and it may thus be worthwhile for a given reservoir to investigate the preconditions, interdependencies and limitations of these strategies. The second is that the relationship between specific retention volume and range of impact we have established is a simple tool for a quick assessment of a reservoir's capability for regional flood mitigation.

To conclude: while setting up a system for coordinated regional reservoir operation including explicit consideration of uncertainties has been beyond the scope of this study, we have presented several methods which can be used to set the basis for such as system.

Part III

DISENTANGLING TIMING AND AMPLITUDE ERRORS IN STREAMFLOW SIMULATIONS

In hydrology it is common to assess magnitude (*vertical*) errors in streamflow simulations. Timing (*horizontal*) errors are however rarely considered. In order to analyze their importance I propose a method to quantify both timing and magnitude errors. The method closely resembles the way a hydrologist would visually evaluate the agreement of observation and model output. The results show significant differences in the time-magnitude error statistics for different flow conditions (periods of low-flow and periods of rise and recession in hydrological events), which standard statistics are not able to reveal. The proposed method thus offers novel perspectives for model diagnostics and evaluation.

This study is published in *Hydrology and Earth System Science*. Part III is a reprint of:

Seibert SP, Ehret U and Zehe E (2016): Disentangling timing and amplitude errors in streamflow simulations, Hydrol. Earth Syst. Sci., 20, 3745-3763, doi:10.5194/hess-20-3745-2016.

3

PART 3: TIMING AND AMPLITUDE UNCERTAINTIES

ABSTRACT

This article introduces an improvement in the Series Distance (SD) approach for the improved discrimination and visualization of timing and magnitude uncertainties in streamflow simulations. SD emulates visual hydrograph comparison by distinguishing periods of low flow and periods of rise and recession in hydrological events. Within these periods, it determines the distance of two hydrographs not between points of equal time but between points that are hydrologically similar. The improvement comprises an automated procedure to emulate visual pattern matching, i.e. the determination of an optimal level of generalization when comparing two hydrographs, a scaled error model which is better applicable across large discharge ranges than its non-scaled counterpart, and "error dressing", a concept to construct uncertainty ranges around deterministic simulations or forecasts. Error dressing includes an approach to sample empirical error distributions by increasing variance contribution, which can be extended from standard one-dimensional distributions to the two-dimensional distributions of combined time and magnitude errors provided by SD.

In a case study we apply both the SD concept and a benchmark model (BM) based on standard magnitude errors to a 6-year time series of observations and simulations from a small alpine catchment. Time–magnitude error characteristics for low flow and rising and falling limbs of events were substantially different. Their separate treatment within SD therefore preserves useful information which can be used for differentiated model diagnostics, and which is not contained in standard criteria like the Nash–Sutcliffe efficiency. Construction of uncertainty ranges based on the magnitude of errors of the BM approach and the combined time and magnitude errors of the SD approach revealed that the BM-derived ranges were visually narrower and statistically superior to the SD ranges. This suggests that the combined use of time and magnitude errors to construct uncertainty envelopes implies a trade-off between the added value of explicitly considering timing errors and the associated, inevitable time-spreading effect which inflates the related uncertainty ranges. Which effect dominates depends on the characteristics of timing errors in the hydrographs at hand. Our findings confirm that Series Distance is an elaborated concept for the comparison of simulated and observed streamflow time series which can be used for detailed hydrological analysis and model diagnostics and to inform us about uncertainties related to hydrological predictions.

3.1 INTRODUCTION

Manifold epistemic and aleatory uncertainties make the simulation of streamflow a fairly uncertain task. The assessment of uncertainties, i.e. quantification, evaluation, and communication, is thus of great concern in decision making, model evaluation, the design of technical structures like flood protection dams or weirs, and many other issues. The quantification and evaluation of uncertainties typically involves the comparison of simulated and observed rainfall–runoff data.

For this purpose, visual hydrograph inspection is still the most widely used technique in hydrology as it allows for the simultaneous consideration of various aspects such as the occurrence of hydrological rainfall–runoff events, the timing of peaks and troughs, the agreement in shape, and the comparison of individual rising or falling limbs within an event. The main strength of visual hydrograph comparison results from the human ability to identify and compare matching, i.e. hydrologically similar parts of hydrographs ("to compare apples with apples") and particularly to discriminate vertical (magnitude) and horizontal (timing) agreement of hydrographs. Whereas the former implies that rising and falling limbs of the two time series are intuitively and meaningfully matched before they are compared, the latter refers to a joint but yet individual consideration of timing and magnitude errors. Visual hydrograph inspection is hence a powerful yet demanding evaluation technique which is still rather difficult to mimic by automated methods. Clear disadvantages of visual hydrograph inspection, however, are its subjectivity and that its application is restricted to a limited number of events.

3.1.1 *Single and multiple criteria for hydrograph evaluation*

To overcome this shortcoming, a large number of numerical criteria (Bennett et al., 2013; Dawson, Abrahart, and See, 2007; Laio and Tamea, 2007; Legates and McCabe, 1999; Nash and Sutcliffe, 1970; Pachepsky et al., 2006) have been proposed. However, each criterion typically evaluates only one or just a few hydrograph aspects and there is no "one size fits all" solution available. For this reason different attempts have been undertaken to compare expert judgement and automated criteria (Crochemore et al., 2014) and to establish model evaluation guidelines (e.g. Biondi et al., 2012; Harmel et al., 2014; Moriasi et al., 2007). Key points of related guidelines typically include the statement that the choice of the metric should depend (i) on the modelling purpose, (ii) on the modelling mode (calibration, validation, simulation, or forecast), and (iii) on the model resolution (time stepping, spatial resolution). Further, most authors recommend the combination of several, preferably orthogonal criteria, which might imply combined application of absolute and relative criteria (Willmott, 1981). Hence, within the last decade several multi-criteria approaches for model calibration and evaluation have been proposed (Boyle, Gupta, and Sorooshian, 2000; Efstratiadis and Koutsoyiannis,

2010; Gupta, Sorooshian, and Yapo, 1998; Kollat, Reed, and Wagener, 2012; Vrugt et al., 2003), which combine different performance criteria and/or evaluation against hydrological signatures such as the shape of the flow duration curve (Euser et al., 2013; Hrachowitz et al., 2014). Even approaches aiming to mimic visual hydrograph comparison were developed. These include multicomponent mapping (Pappenberger and Beven, 2004), self-organizing maps (Reusser et al., 2009), wavelets (Liu et al., 2011), the hydrograph matching algorithm (Ewen, 2011), and the "Peak-Box" approach for the interpretation and verification of operational ensemble peak-flow forecasts (Zappa, Fundel, and Jaun, 2013). Despite this considerable progress, many practical and scientific applications (Gassmann et al., 2013; Haag, Vollmer, and Heß, 2005; Kelleher, Wagener, and McGlynn, 2015; Seibert, Skublics, and Ehret, 2014; Wrede et al., 2015; Zhang et al., 2016) still rely on simple mean squared error (MSE) type distance metrics such as the long-established Nash–Sutcliffe efficiency (NASH) or the root mean squared error (RMSE) even though their shortcomings are well known (Gupta et al., 2009; Schaefli and Gupta, 2007; Seibert, 2001).

A less recognized issue of MSE-type criteria is that these compare points with identical abscissa, i.e. at the same position in time. This means that points in the observation are "vertically" compared to points in the simulation (in the following we refer to them as vertical metrics). The problem with this is that small errors in timing may be expressed as large errors in magnitude. It is obvious that neither individual criteria nor the combination of different vertical metrics within a multi-objective approach can compensate for this.

3.1.2 *Uncertainty assessment and model diagnostics – learning from model deficiencies*

Just as with performance criteria, many methods related to the quantification, visualization, and communication of uncertainties were developed in recent decades, and the value of knowledge about simulation uncertainty is now generally acknowledged. The range of methods is large and comprises manifold probabilistic and non-probabilistic approaches. Probabilistic concepts, for instance, include the total model uncertainty concept (Montanari and Grossi, 2008), methods based on Bayes' theorem (Krzysztofowicz, 1999; Krzysztofowicz and Kelly, 2000), and various ensemble techniques (Cloke and Pappenberger, 2008; Georgakakos et al., 2004; Roulston and Smith, 2003). Non-probabilistic methods include the generalized likelihood uncertainty estimation (GLUE) (Beven and Binley, 1992), possibilistic methods (Jacquin and Shamseldin, 2007), or approaches applying fuzzy-set theory (Nasseri, Ansari, and Zahraie, 2014). Uncertainty assessment is a field of ongoing research, and so far there is no generally accepted technique available. The most important points of criticism of the non-probabilistic methods are their subjectivity and their inconsistency with probabilistic approaches when these are applied to cases which can be explicitly answered using statistical approaches (Stedinger et al., 2008). On the

other hand, probabilistic approaches always rely on the assumptions of ergodicity and stationarity, which are rarely fulfilled in reality. A spin-off of uncertainty assessment is the field of model diagnostics, which ultimately aims to learn more about and from model deficiencies. Related approaches either analyse the temporal patterns of parameter identifiability (Wagener et al., 2003) or the coincidence of typical errors (Reusser et al., 2009) and parameter sensitivity (Reusser and Zehe, 2011) in streamflow simulation.

Motivated by the limitations of vertical distance metrics, Ehret and Zehe, (2011) developed the Series Distance (SD) approach. SD is not a single equation but rather a concept designed for joint but separated assessment of timing and magnitude errors in streamflow simulations, either for events in distinct periods or the entire time series. "Joint but separated" means that both the time and magnitude distances between the observed and simulated hydrographs are determined for matching pairs of points in the event, but the two distances are kept separate. Such separate treatment is for instance desirable in flood forecasting, where errors in magnitude are relevant for dike defence, whereas errors in timing are crucial for reservoir operation. The separation of timing and magnitude errors is further helpful for improving model diagnostics as they point towards different deficiencies in the model structure.

Here we present substantial improvements (Sect. 3.2) to the original approach of Ehret and Zehe, (2011), particularly the coarse-graining procedure. We furthermore introduce a heuristic approach to visualize timing and magnitude uncertainties in streamflow simulations by constructing two-dimensional uncertainty ranges in Sect. 3.3. Related to that, we provide and test several quality criteria to evaluate deterministic uncertainty ranges. The skill of uncertainty ranges is still rarely evaluated in hydrology (Franz and Hogue, 2011), and most of the available methods such as rank probability scores (Duan et al., 2007), rank histograms, or the usage of different moments of the probability density function (De Lannoy et al., 2006) were developed in climatology (Franz and Hogue, 2011; Gneiting et al., 2008). These approaches typically quantify ensemble spread and thus are probabilistic approaches to evaluate uncertainty estimation. To our knowledge only few deterministic approaches, e.g. categorical statistics such as the Brier score or contingency tables or combinations of deterministic and probabilistic approaches (Shrestha, Kayastha, and Solomatine, 2009), are available. In Sect. 3.4 we test the feasibility of the advanced SD approach in a case study and compare it to a standard benchmark error model. Section 3.5 contains the results and discussion, Sect. 3.6 the related conclusions. To foster the use of the SD approach, we publish the SD (Matlab) code, licensed under Creative Commons license BY-NC-SA 4.0, together with a ready-to-use sample data set alongside this manuscript. It is accessible via a GitHub repository `https://github.com/KIT-HYD/SerieSDistance` (Ehret and Seibert, 2016).

3.2 SERIES DISTANCE – CONCEPT AND MODIFICATIONS

SD was developed to resemble the strengths of visual hydrograph inspection in an automated procedure, which typically rests on the following premises (Ehret and Zehe, 2011):

- Hydrographs contain individual events separated by periods of low flow.
- Events are composed of rising and falling limbs or segments which are separated by peaks and troughs.
- These different parts of event hydrographs reflect different hydrometeorological processes and should be compared individually, so as to not compare apples with oranges. This is of particular importance if the simulated (*sim* in the following) and observed (*obs*in the following) hydrographs do belong to different parts of the hydrograph at the same time step t (compare black rectangle in Fig. 3.1).
- A comprehensive evaluation of the agreement of matching rising and falling limbs of two hydrographs requires consideration of both errors in timing and magnitude as this better informs us about ways to improve the model. A simulated rising limb can, for example, match perfectly with its observed counterpart with respect to values but occur systematically too early or too late, which would indicate the need to adjust model parameters related to runoff concentration and flood routing or to improve the related model components.
- A comprehensive comparison of *sim* and *obs* should also provide information on the overall agreement with respect to the occurrence of relevant events and times of low flow. This is typically expressed by contingency tables, which contain information about correctly predicted, missed, and falsely predicted events.

These criteria listed above inform about different error sources, and their individual evaluation therefore provides useful information for a targeted model improvement. As SD accounts for all of these aspects, it is not a single formula but rather a procedure which includes the following steps. For each step, the main innovations are described in detail in the sections below.

- Hydrograph preprocessing (Sect. 2.1). New: routines to create gap-free, non-negative time series and to filter irrelevant fluctuations.
- Identification and pairing of events (Sect. 2.2). New: routines to read user-specified events and to treat the entire time series as a single, long event.
- Identification, matching, and coarse-graining of segments (Sect. 2.3): New: this part has been completely reworked and now applies the coarse-graining procedure.

- Calculation of the distance between matching segments with respect to both timing and magnitude (Sect. 2.4). This is the core of SD, and it is important to note that the distances are computed between points of the hydrographs considered to be hydrologically similar. New: routines to calculate a scaled magnitude error.
- Calculation of a contingency table which counts matching, missing, and false events. No changes.

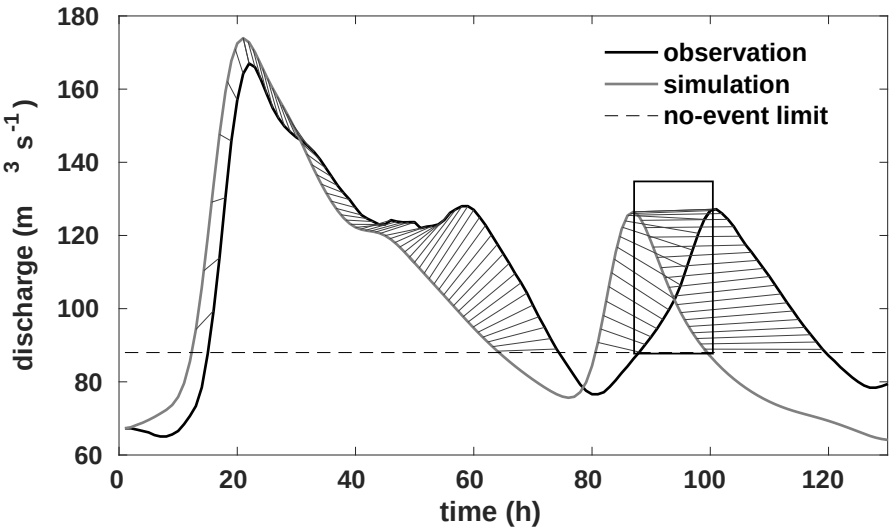


Figure 3.1: Time series of observed (black) and simulated (grey) discharge during a hydrological event. The horizontal line represents a user-specific threshold which differentiates between event and non-event periods. The light grey lines represent the Series Distance connectors linking hydrologically comparable points in the two time series. Time and magnitude distances are calculated between these points. The black rectangle highlights time steps where a part of the recession of the simulation overlaps with a rising part of the observation (figure from Ehret and Zehe, (2011)).

3.2.1 *Hydrograph preprocessing*

The application of SD usually requires some preprocessing to assure gap-free and non-negative time series of equal length; related routines are now included in the SD code. Further routines are available for the adjustment of consecutive identical values; the identification of rising and falling limbs requires non-zero gradients and for time series smoothing, which is often necessary due to the presence of sensor-related non-relevant microsegments. Smoothing is based on the Douglas–Peucker algorithm (Douglas and Peucker, 1973), which preserves extremes but filters the noise (Ehret, 2016). Preprocessing also involves the identification of segments, i.e. contiguous periods of rise or fall in the hydrograph. This is based on the slope of the hydrograph computed between two successive time steps.

3.2.2 *Identification and pairing of events*

For many aspects of hydrology such as flood forecasting or studies of rainfall–runoff transformation, it is useful to consider a hydrograph as a succession of distinct events, usually triggered by rainfall events, separated by periods of low flow. As SD is based on the concept of comparing similar parts of *obs* and *sim* hydrographs, it ideally also involves the steps of identifying events both in the *obs* and *sim* time series and then relating the resulting events between the series. On this level, the general agreement of the two series is evaluated with a contingency table, which counts the number of hits (observed events that have a matching simulated counterpart), misses (observed events without a simulated counterpart), and false alarms (simulated events without an observed counterpart). This is also the basis for the further steps of the SD procedure: only for matching pairs of *obs*–*sim* events can matching segments of rise and fall within the events be identified and the combined time–magnitude error be computed. For misses, false alarms, and periods of low flow this is not possible. For these cases, the best indicator of hydrological similarity in *obs* and *sim* is similarity in time; i.e. the distance between the observed and simulated hydrograph can be computed with a standard vertical distance measure.

The detection of events in hydrographs and their subsequent pairing, however, is not trivial and has to our knowledge not yet been solved in an automated and generalized way. The original version of SD applied a simple no-event threshold (see Fig. 3.1) which, however, often produced unsatisfactory results in the form of many non-intuitive misses or false alarms if the events peaked just above or below the threshold. To overcome these limitations, two further options are now included in SD. The first allows the reading of event start and end points and matching *obs* and *sim* events from user-provided lists. This "event mode" option allows users to apply any desired event detection method, such as those proposed by Blume, Zehe, and Bronstert, (2007), Seibert et al., (2016), or Merz and Blöschl, (2009), and is recommended if a clear distinction between events and low flow is important.

If the identification of events is either not possible or relevant, both the *obs* and *sim* time series can be treated as two single, long, matching events, and the steps of segment identification and matching as described in the next section are applied to the entire time series. Despite its simplicity, this "continuous mode" has been shown to work well in the authors' opinion after applying the SD approach to different discharge time series in both the event and the continuous mode. Shown to work well in this context means even in the continuous mode, SD linked parts of *obs* and *sim* time series that visually appeared to be matching segments within matching events. Since this is difficult to show in a simple graph or statistic, we provide the SD code and test data together with the article.

3.2.3 *Pattern matching: identification, matching, and coarse-graining of segments*

This section describes the core of the SD concept, i.e. the way to identify, within a matching pair of an observed and a simulated event, hydrologically comparable points of the hydrographs in order to quantify their distance in magnitude and time. This pattern matching procedure has been substantially improved in the new version of SD and is therefore described in detail here.

The term "hydrologically comparable" relates to how a hydrologist would visually compare hydrographs and includes several aspects and constraints. The first constraint is based on the perception that even if hydrological simulations may deviate from the observations in magnitude or timing, their temporal order is usually correct. Therefore, in SD, matching points are compared chronologically by preserving their temporal occurrence: the first point in *obs* is compared to the first in sim, the second to the second, the last to the last. Please note that this does not require the two events to be of equal length, as in SD, the hydrograph is considered a polygon from which the points to compare can be sampled by linear interpolation without restriction to its edge nodes. This is explained in detail below. The second constraint relates to the slope of the hydrograph: to ensure hydrological consistency, points within rising segments of *sim* are only compared to points in rising segments of obs, and the same applies to falling segments. This creates a problem related to the within-event variability of the two hydrographs: it is easy to imagine a case in which the number of segments in the *obs* and *sim* event differs. This can be either due to sensor-related high-frequency micro fluctuations of the observations, which can create sequences of many short rising and falling segments, or to general deviations of the simulation from the observation, such as a double-peaked simulated event while the observed event is single-peaked. In visual hydrograph evaluation, a hydrologist will detect the dominant patterns of rise and fall in the two time series and identify matching segments by doing two things: filtering out short, non-relevant fluctuations and then relating the remaining ones by jointly evaluating their similarity in timing, duration, and slope. The stronger the overall disagreement of the *obs* and *sim* event, the more visual coarse-graining will be done before the hydrographs are finally compared, while at the same time the degree of coarse-graining will also influence the hydrologist's evaluation of the hydrograph agreement: the higher the required degree of coarse-graining, the smaller the agreement.

In SD, these steps are emulated by iteratively maximizing an objective function: while increasingly coarse-graining the two events, their overall time and magnitude distance is evaluated. The final evaluation of agreement is then done on the level at which the optimal trade-off between coarse-graining and hydrograph distance occurs, i.e. where the objective function is minimal.

The procedure consists of four steps and is explained in the following sections: (1) determination of segment properties, (2) equal-

izing the number of segments in the *obs* and *sim* event, (3) iterative coarse-graining, and (4) distance computation for the optimal coarse-graining level.

1. For each segment i in the initial sequence of rises and falls of an event, its properties relevant for coarse-graining are determined: start and end time, duration ($dt(i)$), and absolute magnitude change ($dQ(i)$). From this the relative duration ($dt^*(i)$) and the relative magnitude change ($dQ^*(i)$) of each segment is calculated, i.e. its duration normalized by the total duration and its magnitude change normalized by the total sum of absolute magnitude changes of the entire event. $dt^*(i)$ and $dQ^*(i)$ are then used to determine the relative importance of each segment ($I_{SEG}(i)$) using the Euclidean distance (Eq. (3.1)). Taken together, all $I_{SEG}(i)$ of the time series sum up to 1, and segments that are relevant, i.e. that are either very long and/or include large discharge changes, receive large values of I_{SEG}.

$$I_{SEG}(i) = \sqrt{dt^{*2}(i) + dQ^{*2}(i)} \tag{3.1}$$

2. If the number of segments in the *obs* and *sim* event differs, they are *logically* equalized by removing the required number from the event with the surplus. This is done with a directed, iterative aggregation of segments: the least relevant segment (the one with the smallest value of I_{SEG}) is selected and assimilated by its two neighbouring segments. For instance, a small relevant rising segment will then be combined with its preceding and succeeding falling segment to a single, long, falling segment. For the new segment the properties are then determined; its relative importance is the sum of the previous three segments.

   It is important to note that this procedure is a purely logical assimilation: the timing and magnitude of the points in the dissolved segment remain unchanged; they are only reassigned to the new and larger segment. This also implies that the meaning of coarse-graining in the context of SD is slightly different from its meanings in statistics and thermodynamics or in upscaling (Attinger, 2003; Neuweiler and King, 2002). In the first case, coarse-graining is synonymous with the aggregation and averaging of physical quantities; in the second, it is related to the preservation of heterogeneity effects upon aggregation. In the case of SD, it means that logical ordering properties are aggregated, while the absolute values of the timing and magnitude of the data are not changed.

   Obviously, this procedure includes a false classification: the rising segment in the previous example is now hidden within a larger falling segment. This can be considered as the price of coarse-graining and can be quantified by the number of falsely classified edge nodes (n^*_{mod}) of the time series. Therefore, n^*_{mod} is a useful quantity to punish excessive coarse-graining in the objective function, Eq. (3.2).

3. With the number of segments in the *obs* and *sim* events equalized, their SD timing and magnitude distance can be computed. To this end, the first *obs* segment is compared to the first *sim* segment, the second to the second, etc. Since the segments can differ in length we here assume that for each segment pair, the appropriate number of points is evenly distributed along the segment duration and can thus be found by linear interpolation between the time series edge nodes. The first point in the *obs* segment is then connected to the first point in the *sim* segment, the second to the second, etc. For each connector its horizontal and vertical projection, i.e. length in time and magnitude, respectively, is determined (compare again Fig. 3.1), yielding the joint time and magnitude error of the particular point pair.

   In the initial version of SD, the number of points for each segment pair was found by calculating the mean of the two relative durations, I^*_{dt}, such that long-segment pairs received many points and the overall number of connector points of the time series equalled its number of edge nodes. In order to better emulate a hydrologist's perception of segment importance, in the current version of SD the number of points is determined by the mean relative importance I_{SEG} (Eq. 3.1) of a segment pair. This assigns more points to (and hence puts more emphasis on) short but steeply rising segments while still preserving the same overall number of points.

   At this point the result of the SD procedure – a two-dimensional distribution of time and magnitude errors, separately for the rising and the falling segments – is available. However, in practice the problem of non-intuitive segment matching often spoils the results. Due to the constraint of time-ordered segment matching, any minor change in monotony within a rising or a falling limb that is only present in either the *obs* or *sim* event will produce a false matching of segments. The left panel in Fig. 3.2 illustrates this problem, where the first falling segment in the observed series (labelled with "2" in a square) corrupts segment matching: in chronological terms the steep flood rise in *obs* ("3" in a square) would be compared to the second rising segment in *sim* ("3" in a circle), which is obviously wrong. In this case, the SD time and magnitude distances will be very large, while visual comparison would most likely be done as shown in the right panel of Fig. 3.2 and yield good agreement.

   We overcome this problem using iterative coarse-graining again: within the events, successively more segments are logically aggregated with their neighbours until finally the entire event consists of only two segments: one rise and one fall. Compared to the last step, in which we apply coarse-graining to either *sim* or *obs* in order to equalize the number of segments in the simulated and observed event, we here apply it simultaneously to the *obs* and *sim* event. Hence, an equal number of segments and unique segment matching is ensured. The final comparison of

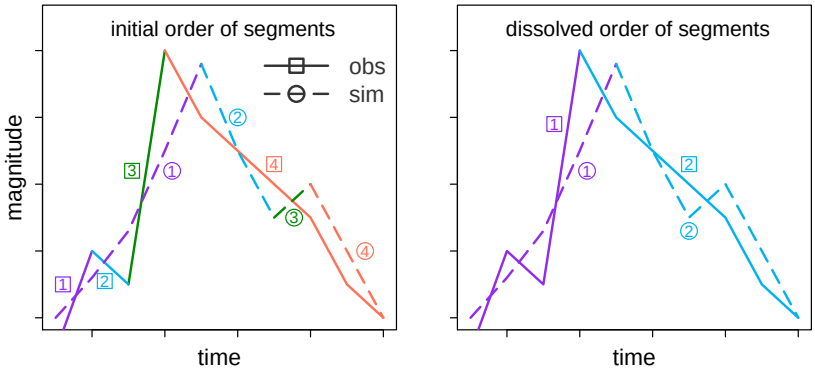


Figure 3.2: Illustration of the time-ordered matching of segments in the coarse-graining procedure. The rising and falling segments of the simulation (sim) and observation (obs) are numbered and colour-coded according to their chronological order. Series Distance compares segments with identical number/colour.

the two events is done for the coarse-graining step in which the total SD errors and the degree of coarse-graining together are small. Both requirements are considered in the coarse-graining objective function (θ). The latter consists of four criteria. The first two are as follows: (i) the number of edge nodes in falsely classified segments (n^*_{mod}) and (ii) the cumulated importance of the dissolved segments ($I^*_{SEG,cum}$). As discussed above, the false classifications inevitably occur during the aggregation of segments. Both criteria monotonically increase with the number of dissolved segments and therefore punish excessive coarse-graining. Further criteria are (iii) the SD timing ($E^*_{SD,t}$) and (iv) magnitude errors ($E^*_{SD,Q}$) summed up over all segments of the event. They are small when segments that are hydrologically similar, i.e. close in time, duration, and magnitude, are compared. As in Eq. (3.1), each criterion is first normalized to the range of [0 1] and then combined using the Euclidean distance (Eq. (3.2)):

$$\theta = \sqrt{\gamma_1 n^{*2}_{mod} + \gamma_2 I^{*2}_{SEG,cum} + \gamma_3 E^{*2}_{SD,t} + \gamma_4 E^{*2}_{SD,Q}}. \qquad (3.2)$$

Note that θ also includes weighting factors ($\gamma_1 \ldots \gamma_4$) for each criterion, which allows for a user- or time-series-specific adjustment of the objective function. Their setting is hence case-specific, with the constraint that $\gamma_1 \ldots \gamma_4$ have to sum up to unity. For example, if the temporal agreement of segments is important, the weight for $E^*_{SD,t}$ should be large. Setting $\gamma_3 = 1$ and all other weights to 0 will hence result in a vertical comparison of the time series, provided that the positions of the edge nodes are identical. The opposite case ($\gamma_4 = 1$ and $\gamma_1 = \gamma_2 = \gamma_3 = 0$) minimizes vertical deviations which leads to horizontally extended SD connectors. Large weights for either γ_1 or γ_2 will prevent any logical aggregation and the pattern matching procedure will suggest the initial conditions as the best solution.

Consequently, "extreme" parametrizations of θ are not meaningful as they will prevent the purpose of SD, which is to compare points which are hydrologically similar.

As can be seen in Fig. 3.2, dissolving a single segment can drastically change the events' overall SD time and magnitude distance. Also, as during the successive removal of segments in coarse-graining, it is impossible to predict which combination of segments dissolved in *obs* and *sim* will yield the best value of θ; thus, all possible combinations are tested and the best is kept. If, e.g., both the *obs* and *sim* event consist of 10 segments, 10×10 combinations of segment dissolutions are tested (obs_1 with sim_1, obs_1 with sim_2, etc.). The coarse-graining scheme is thus computationally demanding. The combination with the minimum θ is kept and serves as the basis for the next segment reduction step in the coarse-graining procedure.

4. Once the coarse-graining is done, the optimal value of θ is available for each reduction step, starting with the initial number of segments and ending with two. In Fig. 3.3, this is shown for a three-peak event with initially 15 segments. As can be seen in the lower right panel, the value of the objective function is initially high: here segment matching is poor and as a result SD timing errors and thus θ are high (upper left panel). After dissolving three segments, agreement is much better (lower left panel) and θ is at its minimum. Further segment aggregation does not further decrease SD errors, but now the number of falsely classified nodes increases and leads to an increase in θ (upper right panel). The interplay of the two antagonist parts of θ often leads to the occurrence of a local minimum in the coarse-graining of complex multi-peak events. The related reduction step can then be regarded as the optimal degree of coarse-graining and the final values of SD time and magnitude errors are determined based on this level. In "simple" events in which no or little coarse graining is required, the objective function values often increase fairly linearly. In any case SD time and magnitude errors are determined based upon the coarse-graining step with the smallest θ value.

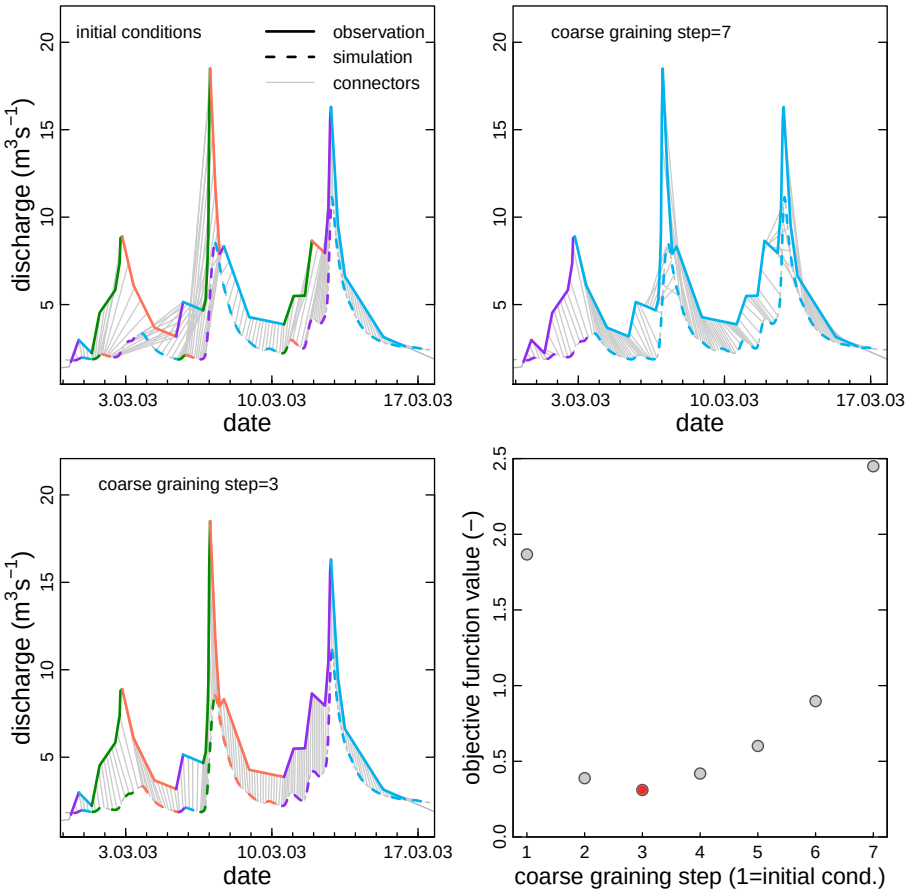


Figure 3.3: Coarse-graining steps: all plots contain data from the same multi-peak discharge event but for different levels of coarse-graining. The initial conditions (top left panel) are characterized by a large number of poorly matching simulated (dashed) and observed (solid) segments as indicated by the non-intuitively placed SD connectors (grey lines). Segments required to match according to the chronological order constraint of SD are indicated by matching colours. In the last coarse-graining step (top right panel) the connectors are placed more meaningfully but the representation of the entire event by only two segments (one rise, one fall) appears inadequately coarse. The optimal level of coarse-graining, here reached at step three, yields visually acceptable connectors while preserving a detailed segment structure (bottom left panel). This step is associated with a minimum of the coarse-graining objective function (Eq. 3.2), indicated by the red dot in the bottom right panel. Grey dots indicated the values of the objective function for all other coarse-graining steps.

3.2.4 *Modifications in the* SD *error model*

In the initial version of SD, the magnitude error ($E_{SD,Q}$) was calculated as the absolute difference between points in *sim* and *obs* linked by a Series Distance connector (c):

$$E_{SD,Q}(c) = Q_{obs}(c) - Q_{sim}(c). \tag{3.3}$$

In the current version, the magnitude error can alternatively be scaled by the mean of the connected points:

$$E^*_{SD,Q}(c) = \frac{Q_{obs}(c) - Q_{sim}(c)}{\frac{1}{2}\left(Q_{obs}(c) + Q_{sim}(c)\right)}. \tag{3.4}$$

This yields a relative and hence dimensionless expression of the vertical error ($E^*_{SD,Q}$), which facilitates the construction of uncertainty ranges of variable width (see Sect. 3). As in the first version of SD, both absolute and relative vertical error values $E^{(*)}_{SD,Q} > 0$ indicate that $Q_{obs}(c) > Q_{sim}(c)$. The calculation of Series Distance timing errors ($E_{SD,t}$) according to Eq. (3.5) remained unchanged. Error values of $E_{SD,t} > 0$ indicate that *obs* occurs later than sim:

$$E_{SD,t}(c) = t_{obs}(c) - t_{sim}(c). \tag{3.5}$$

Similar to the scaling of the vertical error, the timing error could also be scaled using, e.g., event duration. This could be helpful if the error compared to the length of the event or the average length of all events in the time series is of interest.

The application of SD timing and magnitude error models ($E_{SD,t}(c)$ and $E_{SD,Q}(c)$) makes sense where timing errors are both present and detectable, i.e. during events in which discharge is not constant in time. During low-flow conditions time offsets are, however, difficult, if not impossible to detect. Therefore, a simple one-dimensional, vertical, "standard" error model analogous to Eq. (3.3), which relates values at the same time step t, suffices here:

$$E_S(t) = Q_{obs}(t) - Q_{sim}(t). \tag{3.6}$$

Analogously to the scaled vertical SD error model in Eq. (3.4), a scaled version of the one-dimensional vertical error model ($E^*_S(t)$) was added:

$$E^*_S(t) = \frac{Q_{obs}(t) - Q_{sim}(t)}{\frac{1}{2}\left(Q_{obs}(t) + Q_{sim}(t)\right)}. \tag{3.7}$$

3.3 ERROR DRESSING: A HEURISTIC FOR THE CONSTRUCTION OF UNCERTAINTY RANGES

The SD concept can be applied to a variety of tasks such as model diagnostics, parameter estimation, calibration, or the construction of uncertainty ranges. In this section we provide one example thereof and describe a heuristic approach for the construction of uncertainty ranges for deterministic streamflow simulations. Uncertainty ranges provide regions of confidence around an uncertain estimate and are of practical relevance and a straightforward means of highlighting and of assessing magnitude and timing uncertainties of hydrological simulations or forecasts. Conceptually, uncertainty ranges should be wide enough to capture a significant portion of the observed values but as narrow as possible to be precise and, thus, meaningful. These requirements are antagonistic as large uncertainty ranges, which capture most or all observations, are usually imprecise to a degree that

makes them useless for decision-making purposes (Franz and Hogue, 2011).

The method we propose here follows the concept proposed by Roulston and Smith, (2003) and yields quantitative estimates of forecast uncertainty by "dressing" single forecasts with historical error statistics. The original approach was designed to dress ensemble forecasts; for SD it was adapted to deterministic streamflow simulations and extended from one dimension (magnitude) to two (magnitude and timing). Like statistical approaches to uncertainty assessment, error dressing is based on the fundamental assumptions of ergodicity and stationarity, i.e. the assumption that errors that occurred in the past are reliable predictors for errors in the future. In the following we first outline the regular, one-dimensional deterministic error dressing method and then describe its modifications for SD.

3.3.1 *The one-dimensional case*

Provided with a record of past streamflow observations (O_{hist}) and corresponding model simulations (S_{hist}), any valid error model such as Eq. (3.6) can be applied to calculate a distribution of historic errors. This distribution can then be sampled (Fig. 3.4, upper left panel) using a suitable strategy and the selected subset of errors can be applied to each time step of the simulation. Connecting all upper and all lower values of the dressed errors yields corresponding envelope curves (Fig. 3.4, upper right panel). For this procedure Roulston and Smith, (2003) coined the term error dressing.

The choice of the sampling strategy, however, strongly influences the statistics of the resulting uncertainty ranges and should be carefully selected. In our case, the precondition was that the approach should be extendible to two-dimensional cases to allow its later application to the error distributions of the SD approach. Therefore, we defined the sampling strategy according to the variance contribution, which is straightforward to apply for the one-dimensional case: for each point of the error distribution its relative contribution ($d\sigma_i^2$) to the *unbiased* variance of the total error distribution (σ_x^2) is calculated according to Eq. (3.8):

$$d\sigma_i^2 = \frac{(x_i - \overline{x})^2}{n\sigma_x^2} 100. \tag{3.8}$$

Here $\overline{x}$ and n denote the mean and the size of the corresponding error distribution. The usage of the unbiased variance, having n in the denominator not $n-1$, ensures that all $d\sigma_i^2$ sum up to 100. Next, all points of the error distribution are ordered by the values of $d\sigma_i^2$, and, starting with the smallest, a desired subset of all $d\sigma_i^2$, e.g. 80 % is taken from the list. This subset represents an informal probability ($p \in [0\ 1]$) as it relates to the number of observations that fall within the uncertainty range. Small values of p are associated with narrow (sharp) uncertainty ranges but at the cost of a higher portion of true values that fall outside. Contrary, high values of p cause wide (imprecise) uncertainty ranges which, however, contain most errors that

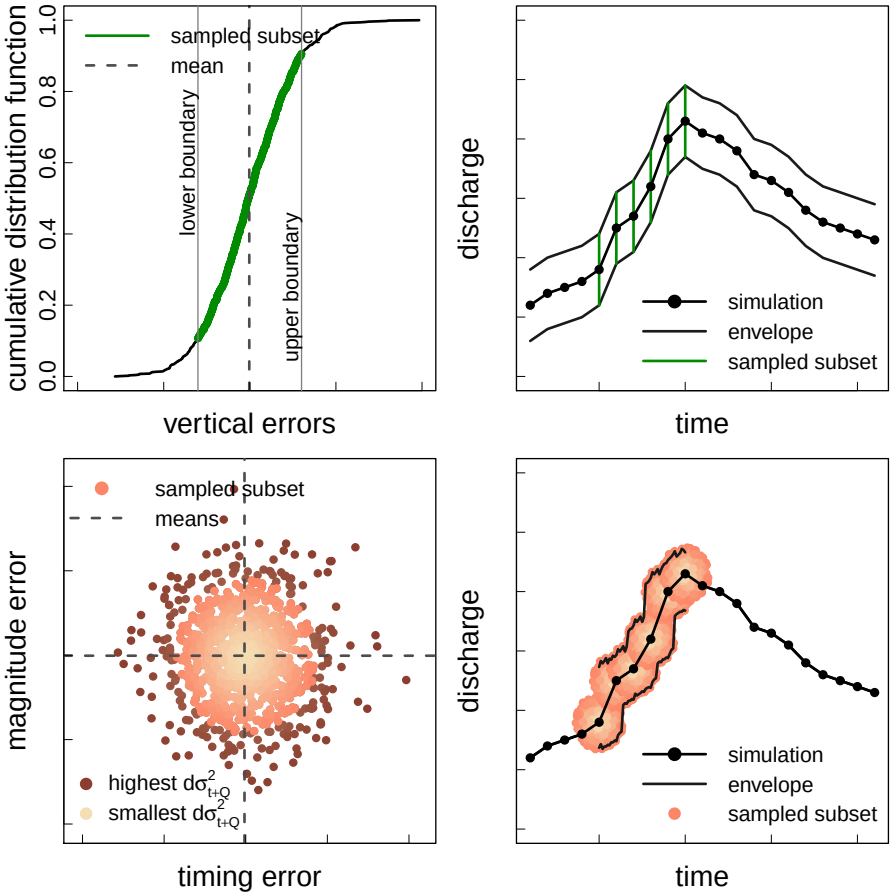


Figure 3.4: Sketch of the one- and two-dimensional error dressing method using normally distributed random numbers ($n = 1000$). The upper row panels show the one-dimensional case with an empirical cumulative distribution function of errors (*upper left panel*) and an 80 % subset thereof sampled according to increasing variance contribution. The application (dressing) of the subset of errors to a hydrograph and the construction of the corresponding envelop curves is illustrated in the *upper right panel*. The lower row panels show the same procedure for the two-dimensional case. From the two-dimensional distribution of empirical errors (*bottom left panel*) 80 % (colour-coded) are again sampled according to the combined variance contribution of both distributions (colour ramp). The *bottom right panel* contains a sketch of the two-dimensional error dressing method and the construction of envelope curves. Please note that the use of normally distributed numbers yields symmetrical samples and envelopes, which is usually not the case for real-world data, which are usually skewed.

occurred in the past. For practical applications, typically coverages of 80 to 90 % are chosen. In Fig. 3.4, top left panel, the coverage was set to $p = 0.8$.

3.3.2 *The two-dimensional case*

SD yields two-dimensional distributions of coupled errors in timing and magnitude and thus requires a two-dimensional strategy for the

sampling of error subsets and the construction of envelope curves (Fig. 3.4, lower row panels).

How does one sample from bivariate distributions of coupled errors with different units? Statistics and computational geometry offer concepts based on the ordering of multivariate data sets, such as geometric median or centre point approaches. The former provides a central tendency for higher dimensions and is a generalization of the median which, for one-dimensional data, has the property of minimizing the sum of distances. Centre points are generalizations of the median in higher-dimensional Euclidean space and can be approximated by techniques such as the Tukey depth (Tukey, 1975) or other methods of depth statistics (Mosler, 2013). Here, however, we want the errors to be centred around the mean (not around the median). Hence, we apply the same concept that we use for the one-dimensional case to SD in that we sample based on the combined contribution of each point to the total variance. Analogously to Eq. (3.8) we calculate the relative timing ($d\sigma_t^2$) and magnitude ($d\sigma_Q^2$) contribution of each point to the total variances of the corresponding distributions. Their sum yields an estimate of the combined contribution of each point to the combined variance of both error distributions:

$$d\sigma_{t+Q}^2 = d\sigma_t^2 + d\sigma_Q^2. \tag{3.9}$$

Analogously to the one-dimensional case, the points are ordered by increasing combined variance contribution $d\sigma_{t+Q}^2$, and, starting from the point with the smallest value which is close to or at the mean, a subset of errors can be extracted. The shape of the resulting subset depends on the underlying distribution of errors. Uncorrelated errors yield more or less circular or oval shapes (Fig. 3.4, lower left panel). By contrast, correlated errors yield different shapes, which is valuable for diagnostic purposes.

SD distinguishes periods of low flow, rising, and falling limbs. Hence, subsets of two 2-D error distributions (rising and falling limb) and from one one-dimensional error distribution (low flow) are calculated and applied to each point of a simulation: points of low flow are dressed with the low-flow error subset, points of rise with error subsets from rising limbs, etc. Altogether this yields a region of overlapping error ovals around a simulation (Fig. 3.4, lower right panel), which can for convenience be represented by an upper and lower envelope curve. These lines are found by subdividing the time series into time slices of length dt (the temporal resolution of the original series), centred around each edge node of series. In each time slice, the magnitude and timing of the largest and smallest error are identified. These values span the upper and lower limit of the uncertainty envelope, respectively. Using linear interpolation yields the upper and lower limits of the envelope at the points in time of the original series, which is useful to calculate evaluation statistics.

3.4 CASE STUDY

This case study, based on real-world data, serves to present and to discuss relevant aspects of SD by comparison with a benchmark error model (BM).

3.4.1 *Data and site properties*

We used discharge observations (O_{hist}) of a 6-year period (30 October 1999–30 October 2005) from gauge "Hoher Steg" (HOST), which is located in the small alpine catchment of the Dornbirner Ach River in north-western Austria. Catchment size is 113 km^2, the elevation range is 400–2000 m a.s.l., and mean annual rainfall differs between 1100 and 2100 mm yr^{-1}. For the 6-year period, hourly hydrometeorological time series ($n = 52\,633$ time steps) were used to drive an existing, calibrated conceptual water budget model of the type LARSIM (Large Area Runoff Simulation Model, gridded version, resolution = 1 km^2; Ludwig and Bremicker, 2006), which yielded acceptable simulations (S_{hist}) with a NASH of 0.78. Please note that for the discussion of the SD concept, neither the model itself nor the catchment properties are particularly relevant. The main purpose of the case study was to apply realistic data. This is also the reason why we used the entire 6-year period to both derive and apply the error distributions; i.e. we did not distinguish periods of error analysis and error application.

3.4.2 *Conceptual setup*

For the benchmark model, we derived distributions of 1-D vertical errors. We did not differentiate cases of low flow and events, which is rather simplistic but standard practice. For the SD approach we did differentiate these cases. This may be considered an unfair advantage for SD as it allows the construction of more custom-tailored uncertainty envelopes. However, as the objective of the case study is not a competition between the two approaches but a way to present interesting aspects of SD, we considered it justified. For SD, the required starting and end points of hydrological events were manually determined both in O_{hist} and S_{hist} by visual inspection. Altogether there were $n = 123$ events in each series, and they were fully matching; i.e. no missing events or false alarms occurred. The resulting contingency table is obviously trivial and therefore not discussed further here.

Both for SD and BM, we applied scaled errors ($E^*_{SD,Q}(c)$ according to Eq. (3.4) and E_{BM} according to Eq. (3.7), respectively), as we found that compared to the standard error model, they are more applicable across the usually large discharge ranges present in hydrographs. For SD, the weights $\gamma_1, \ldots, \gamma_4$ used in the objective function of the coarse-graining procedure (Eq. 3.2) were set to $\frac{1}{7}$, $\frac{1}{7}$, $\frac{5}{7}$, and 0, respectively, based on iteratively maximizing the visual agreement of segments in matching events of *sim* and *obs*. Additional studies with different data sets (not shown here) yielded similar optimal weights, which

corroborates that this is a relatively robust choice and sufficient for a proof of concept, as intended in this study. For more widespread applications, a detailed sensitivity analysis is desirable. Such an analysis is, however, difficult as several different time series, flow conditions, and rainfall–runoff events would have to be visualized and compared. Moreover, there is no robust benchmark available to which we may compare the outcome of the proposed coarse-graining procedure. For this reason we provide software such that any interested person can find out for him/herself whether the proposed method suits his or her needs or not.

Based upon SD and BM we derived empirical error distributions from the entire test period and then used them, in the same period, to construct uncertainty envelopes around the simulation S_{hist} using the error dressing approach as described in Sect. 3.3. To ensure comparability we enforced identical coverages for both approaches during the construction of the envelope curves; i.e. we made sure that the desired fraction of observations (e.g. 80 %) fell within the uncertainty envelope. For the standard error model this was straightforward: if from the 1-D distribution of errors a subset of $p = 80\,\%$ is selected and used to construct the uncertainty envelope as described in Sect. 3.3.1 for the same period of time, then by definition the number of observations within the envelope must also be 80 %. For SD, however, as a consequence of error ovals overlapping in time (Fig. 3.4, lower right panel), this is not self-evident and typically many more observations fall within the uncertainty envelope than the level p at which the subset of the 2-D error distribution is sampled. This issue was solved by iteratively sampling the error distributions at various levels of p until the desired percentage of observations (here: 80 %) fell within the uncertain envelope.

3.4.3 *Evaluation of deterministic uncertainty ranges*

The evaluation of deterministic uncertainty ranges requires methods to quantify properties such as coverage or precision. Here we propose a set of statistics which can be applied to uncertainty ranges irrespective of how they were constructed. While this ensures comparability of the SD and BM-derived ranges, it does not exploit the advantages of the SD approach, i.e. separate treatment of time and magnitude uncertainties.

1. Coverage (ϕ) is the most intuitive criterion. It quantifies the ratio of observations that fall inside the simulated uncertainty range and can take values between 0 (no single observed value included) and 1 (all observations included). ϕ can easily be obtained as the number of observations (n_{obs}) that fall inside the uncertainty range around a simulation, divided by the total length of the time series (n):

$$\phi = \frac{n_{obs}}{n}. \tag{3.10}$$

2. Precision (PRC) allows the comparison of different uncertainty ranges. PRC is the average width of the uncertainty envelope, i.e. the average difference of the upper ($UE^+(t)$) and the lower ($UE^-(t)$) envelope curve. The smaller PRC, the sharper the uncertainty range. High coverages (ϕ) typically require wide uncertainty ranges and, thus, high values of PRC. PRC has the same unit as the discharge time series.

$$PRC = \frac{1}{n}\left(UE^+(t) - UE^-(t)\right) \tag{3.11}$$

3. Finally we suggest scaling PRC by the value of the simulation according to Eq. (3.4), i.e. to express uncertainty relative to the magnitude of the simulation. PRC^* is dimensionless and decreases with decreasing width of the uncertainty range. An uncertainty range of zero width yields a PRC^* of 0. Hence, small values of PRC^* indicate high skill.

$$PRC^* = \frac{1}{n}\frac{\left(UE^+(t) - UE^-(t)\right)}{Q_{sim}(t)} \tag{3.12}$$

In the case study, we used ϕ as a means to ensure comparability rather than for comparison: coverage for both the SD and BM approach was set to $80 \pm 0.5\,\%$. For SD the required percentage of sampled errors was found by trial and error to be $p = 76\,\%$ (Table 3.3). With coverage equalized, SD and BM can be directly compared by PRC and PRC^*. High (relative) precision, i.e. small values of $PRC^{(*)}$, indicate better performance. If the evaluation of uncertainty ranges with respect to over- and undershooting is of interest, additionally the percentage of observations above or below the uncertainty range can be computed analogously to Eq. (3.10). This is for instance of interest for flood forecasters who try to minimize overshooting or water supply managers who try to minimize undershooting. For the sake of brevity, this has not been further considered here.

3.5 RESULTS AND DISCUSSION

In this section we first discuss some general aspects of the SD concept and then compare it to the benchmark approach using the case study data.

3.5.1 *Potential and limitations of the core* SD *concept*

Series Distance is an elaborate method for the comparison of simulated and observed streamflow time series. The concept allows the distinction between different hydrological conditions (low flow and rising and falling limbs) and determines joint errors in timing and magnitude of matching points within matching segments of related hydrographs. Differences in the high- and/or low-frequency agreement of the *obs* and *sim* hydrographs are considered with an iterative

coarse-graining procedure, which effectively mimics visual hydrograph comparison. This differentiated evaluation makes SD a powerful tool for model diagnostics and performance evaluation.

The challenges of SD are, however, in the details: the robust, precise, and meaningful partitioning of the hydrograph into periods of low flow and events is difficult. We tested various approaches including baseflow separation and filtering techniques (e.g. Chapman, 1999; Douglas and Peucker, 1973; Eckhardt, 2005; Perng et al., 2000), penalty functions (Drabek, 2010), fuzzy logic (Seibert and Ehret, 2012), and the methods proposed by Merz and Blöschl, (2009) and Norbiato et al., (2009). In all cases, the results were unsatisfactory when applied to a range of different flow regimes. The same applies for the matching of conjugate events in *obs* and sim. Currently, there is no robust and automated method available for any of the two cases. Possible remedies are the adaptation of any of the methods proposed above to specific conditions (Seibert et al., 2016), manual event detection, and matching, or one could treat the entire time series as a single, long event, at the expense of losing the separate treatment of low-flow cases. Within an event, the quality of the segment matching significantly determines the quality of the subsequent matching of *obs* and *sim* points and hence the quality of the SD error calculation. This challenge has been solved in a mostly very satisfactory way by the iterative coarse-graining procedure. The resulting set of matching segments and the required degree of coarse-graining is in itself a useful result which can be used for comparative hydrograph analysis.

Qualitative analyses of the weighting factors $\gamma_1 \ldots \gamma_4$ in Eq. (3.2) confirmed that these parameters emphasize different aspects of the hydrograph and thus allow for a flexible adaptation of the pattern matching procedure to different flow regimes.

Table 3.1: Qualitative description of the impact of the different weighting factors of the objective function θ (Eq. 3.2), which governs the coarse-graining procedure. Note: none of the extreme parametrizations described by the cases nos. 1–4 is meaningful as any of them prevent the comparison of hydrologically similar points.

Case	γ_1	γ_2	γ_3	γ_4	Impact
1	1	0	0	0	no aggregation of segments.
2	0	1	0	0	no aggregation of segments.
3	0	0	1	0	horizontal differences are minimized, i.e. vertical comparison.
4	0	0	0	1	vertical differences are minimized, i.e. horizontal comparison.
5	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	equal weights, compare Fig. 3.5.
6	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{5}{7}$	0	suggested default, compare Fig. 3.3 (bottom left panel).

Applied to a single event, different combinations of γ-parameters cause different segments to be identified and matched, leading to dif-

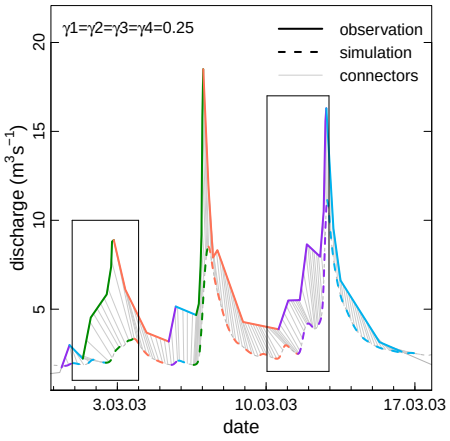


Figure 3.5: Optimal coarse-graining solution of the event depicted in Fig. 3.3 if equal weights (γ parameters) are applied to the objective function θ (Eq. 3.2). In this case the coarse-graining procedure selects different edge nodes for two segments (see black boxes) leading to slightly larger timing and smaller magnitude errors compared to the bottom left panel in Fig. 3.3.

fering SD results and aggregation steps. Overall, γ_1 and γ_2 are less sensitive than γ_3 and γ_4. Table 3.1 qualitatively summarizes the impact of the different weighting factors. Figure 3.5 provides the coarse-graining solution for the event depicted in Fig. 3.3 if θ is parametrized using equal weights (case no. 5 in Table 3.1). This plot highlights that different solutions can be acceptable and that coarse-graining remains to a certain degree arbitrary. In any case the parametrization of θ requires a visual verification as small modifications may yield different results. We found that the configuration presented in the case study (Sect. 3.4.2) which punishes large timing errors ($E^*_{SD,t}$) produces good agreement with visual coarse-graining for different events or conditions and we thus suggest it as default parametrization. A more in-depth study of the impacts of $\gamma_1 \ldots \gamma_4$ using streamflow data from different regimes and events would, however, be desirable.

The hydrograph matching algorithm (HMA) proposed by Ewen, (2011) is, to our knowledge, the only method which is similar to the SD concept in the sense that it relates elements of an observed to elements in a simulated hydrograph in an intuitive manner. Similar to SD, the HMA uses connectors ("rays") to establish these relationships. However, the manner in which these connectors are identified is different. The HMA moves chronologically through all elements of *obs* and calculates the distance to points in *sim* which are located within a defined window around the element in *obs* using a penalty function. This procedure generates a (possibly huge) matrix of penalty values. In a second step the optimal "path" through this matrix is identified, which yields the connectors. This makes the HMA computationally demanding. However, the same also applies for SD as the coarse-graining scheme may require a large number of iterations. The advantage of SD is that unique relationships of points in *obs* and *sim* are established, which is not the case for HMA. Leaving aside

these methodological finesses, we believe that for hydrological studies there is a large potential for "intuitive" distance metrics which is not yet fully exploited: in the intercomparison study of Crochemore et al., (2014) both HMA and SD closely resembled expert judgement and outperformed standard (vertical) distance metrics during high- and, for HMA, also low-flow conditions.

3.5.2 *Potential and limitations of the error dressing method*

Error dressing is a simple method and straightforward to apply. Conceptually it is very similar to statistical concepts like the total uncertainty method introduced by Montanari and Grossi, (2008) insofar as it does not distinguish between different sources of uncertainty. Unlike rigorous statistical concepts, error dressing, however, does not make any assumptions regarding the nature of the population of errors: they are directly sampled from the empirical distribution, thus avoiding the need to fit a theoretical distribution to the data. The fundamental assumption of error dressing is hence that the available sample represents the population and implies that the skill of the resulting uncertainty ranges strongly depends on the representativeness of the empirical distribution of errors. This may not be the case if records are short and/or if the available data only cover a limited range of conditions. This is, however, a frequent problem of statistical methods for uncertainty assessment (not only in hydrology), where often the extremes are of interest, although they are rare by definition (Montanari and Grossi, 2008). Further uncertainties arise from erroneous observations, which is a common problem in hydrology. These conceptual limitations lead to the fundamental question of whether it is better to profit from statistical (or heuristic) information on the basis of the stationarity assumption or to neglect it by questioning the assumption itself (Montanari, 2007). This discussion is, however, beyond the scope of this study.

The error dressing concept in the presented form does not distinguish between seasonality or different flow magnitudes as the same error distributions are applied to each rising (and/or falling) limb. More sophisticated implementations are of course possible, such as a differentiation of errors according to flow magnitudes to better capture extremes, or differentiation according to forecast lead times. The same applies for the sampling strategy: as an alternative to the method presented here based on combined variance contribution, the sampling of specific quantiles using the median as central reference or the fitting and application of any parametric function to the distribution is of course possible. A practical insight from applying the error dressing concept is that the variance-based method effectively filters outliers, which sometimes occur when errors are calculated between poorly matching segments.

A last general issue relates to the sampling from the two-dimensional error distribution. Due to the superposition of error clouds in successive time steps it is possible that errors in timing at one time step

mimic errors in magnitude at neighbouring time steps (Fig. 3.4, bottom right panel). This depends on the temporal extent of the error ovals. As a consequence, the relationship between p, which defines the size of the subset from the distribution, and coverage (ϕ) becomes non-unique. In any case it is not directly linear as in the one-dimensional case in which p equals ϕ per definition (at least for the period of calibration). Typically ϕ exceeds p in the two-dimensional case, and desired coverage rates of $\approx 80\,\%$ require us to set p to ≈ 0.65–0.75. If a specific coverage is desired, the related value of p is best found by iteration.

Altogether, the error dressing concept seems suitable for practical applications where long time series are available but more sophisticated uncertainty assessments are not feasible, either because of the required effort or because of limited knowledge of the underlying system.

3.5.3 *Case study results*

As described in Sect. 3.4.2, within the 6-year time series, altogether $n = 123$ events were manually identified in both *obs* and sim. The events matched perfectly; i.e. no missed events or false alarms occurred. This is often the case for simulations of responsive catchments where rainfall events trigger runoff events in most cases and where the precipitation time series thus carries important information about the occurrence of hydrological events. This is not necessarily the case for hydrological forecasts, especially mid- to long-term, where false precipitation events can generate false hydrological events. In the latter case, event-based information contained in the contingency table can be valuable.

The mean event durations were 146 and 154 h for *obs* and sim, respectively, and on average each event initially contained 13 (sub)peaks. The optimal level of event comparison was on average achieved after two coarse-graining steps, which reduced the number of peaks on average to four and led to average durations of 37 h for rising limbs and 109 h for falling limbs for both *obs* and sim. These statistics again bear diagnostic potential as they can be interpreted as surrogates for the mean concentration time of the catchment or as a reservoir constant and can thus be compared to other data. Generally, the matching of segments resulting from the coarse-graining procedure corresponded well with visual human reasoning (not shown). In the following we compare the error distributions and uncertainty envelopes derived from the SD and BM approach for our test case.

3.5.3.1 *Comparison of error distributions*

Altogether four error distributions were calculated: for SD two 2-D distributions (one for the rising and one for the falling event limbs) and one 1-D distribution for the low-flow conditions; for BM a single 1-D distribution of magnitude errors for the entire time series. The

distributions are shown in Fig. 3.6, corresponding statistics in Table 3.2.

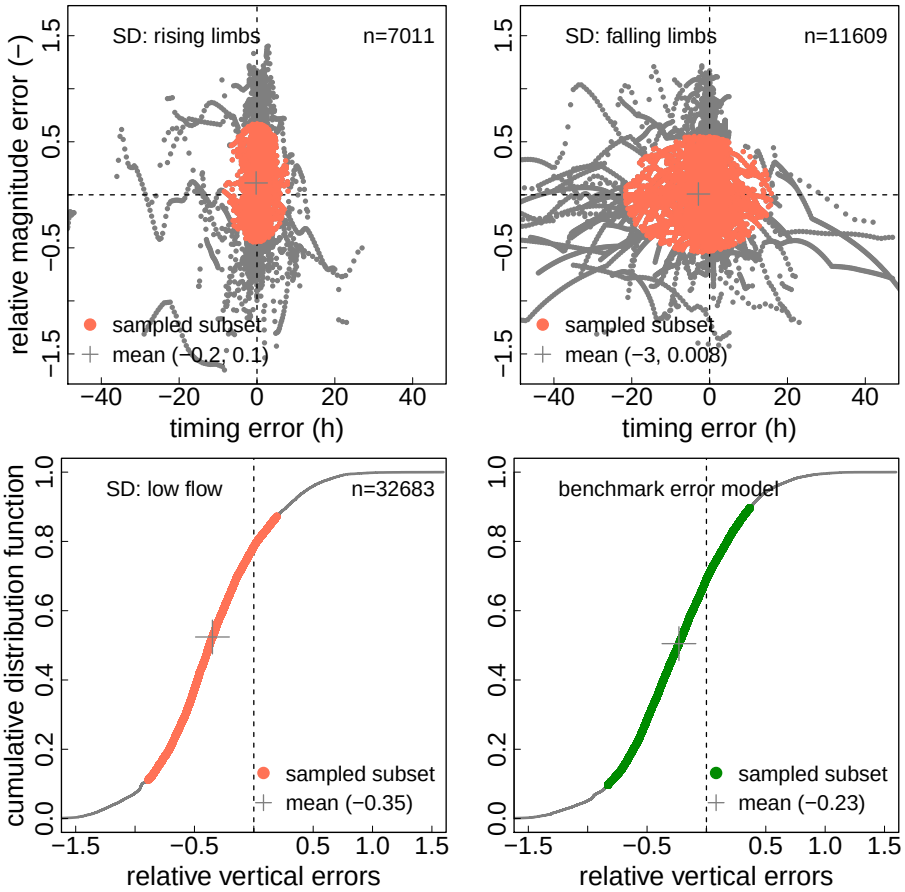


Figure 3.6: One- and two-dimensional error distributions from the case study. The upper row contains Series Distance (SD) results for the rising and falling limbs. The left panel in the lower row shows the one-dimensional SD distribution of errors for the periods of low flow. The panel in the bottom right contains the 1-D distribution of magnitude errors of the benchmark model (BM) for the entire time series. The highlighted subset represents the 80 % subset used to construct the uncertainty envelopes. Distribution statistics are provided in Table 3.2. To improve the readability of the upper two panels, we restricted their timing axes to the range [−45 45]. The number of outliers, i.e. points outside the range mean ± 3 SD (standard deviations) ([−42 36]), was < 1 % for the falling limbs and 1 order of magnitude less for the rising limbs. The dotted lines highlight the origins (all panels).

Comparing the 2-D distributions reveals distinct differences in shape: for the rising limbs the distribution is rather oval; for the falling limbs it is almost circular. This is particularly evident in the sampled subsets. The uniform spread of the errors within the oval and the circle indicates that for the data at hand, the timing and magnitude errors are largely uncorrelated but dependent upon the hydrological conditions (rise or fall). The (scaled) magnitude errors for both distributions are located between ±1.5. The magnitude biases for both distributions are relatively small and lie, according to the ranges pro-

vided by Di Baldassarre and Montanari, (2009), within the error of measurement: $SD_{Q,rise} = 0.1$ for the rising limbs, $SD_{Q,fall} = 0.008$ for the falling limbs. Note that positive magnitude biases indicate simulations that on average underestimate the observations. For timing errors, the differences are more pronounced: while for the rising limbs, timing errors are located between ± 10 h for the sampled subset and biased by -0.2 h (indicating simulations lagging behind the observations), for the falling limbs both the bias (-3 h) and the range (± 20 h) are much larger. Please note that we discuss the timing errors of the subset here rather than those of the entire sample, as the latter include few but large outliers caused by occasional poor matching of falling limbs during coarse-graining.

Table 3.2: Statistical properties of the individual Series Distance (SD) and benchmark (BM) error distributions from the case study. For the entire distribution we provide the first (1^{st} Qu.) and third quartile (3^{rd} Qu.), the mean, median, and the percentage of outliers (data points which are more than 3 standard deviations apart from the mean). For the subset we provide the sampled upper (max.) and lower (min.) boundaries. The subscripts with SD refer to errors in magnitude (Q) and timing (t) separately for the rising (rise) and falling (fall) limbs, respectively. SD_{LF} provides results for the periods of low flow.

Error	Entire distribution					Sampled subset	
Distribution	1^{st} Qu.	Mean	Median	3^{rd} Qu.	%–outlier	Min.	Max.
$SD_{Q,rise}$ (–)	-0.15	0.11	0.13	0.39	0.7	-0.44	0.67
$SD_{Q,fall}$ (–)	-0.23	0.01	0.01	0.25	0.5	-0.54	0.55
$SD_{t,rise}$ (h)	-0.50	-0.22	0.66	1.60	2.1	-8.41	7.98
$SD_{t,fall}$ (h)	-3.89	-2.87	0	1.56	2.9	-21.61	15.86
SD_{LF} (–)	-0.64	-0.35	-0.37	-0.06	0.1	-0.89	0.19
BM (–)	-0.54	-0.23	-0.24	0.09	0.1	-0.83	0.37

Together, these results confirm that different flow conditions, i.e. low-flow, rising or falling limbs of events, exhibit different error characteristics. This suggests that a differentiation between hydrological conditions can be meaningful. For instance, timing errors of the recession in the case study would be strongly underestimated by timing errors of the rising limbs, and vice versa, as depicted in the lower panel of Fig. 3.8. The comparison of 1-D distributions of the SD and BM model revealed that important error characteristics of rare events can be shadowed by frequent but often less important low-flow conditions.

3.5.3.2 *Comparison of uncertainty envelopes*

Subsets of both the SD and BM error distributions were used to construct uncertainty envelopes (UE) around the entire simulated time series S_{hist}. For better visibility of the details, only a 3-week period is shown in Fig. 3.7; the envelope statistics presented in Table 3.3, however, are based on the entire series. The percentages $p = 76\,\%$ for SD and $p = 80\,\%$ for MD of sampled errors in the subsets were selected

such that the overall coverage (ϕ) of the uncertainty envelopes was 80 % in both cases.

Compared to UE_{BM}, the UE_{SD} in Fig. 3.7 appears both smoother and more inflated. This is due to the timing component of the error model, which spreads the uncertainty envelope in time. This is particularly visible at the beginning of the events. Here, timing errors dressed to a given time step clearly extend to neighbouring time steps, representing the uncertainty about the true event start. In the case of several peaks occurring within a short time (Fig. 3.7, last event), the smoothing effect of the timing component can lead to a merging of the related uncertainty envelopes towards a single, large region. Also the difference between smaller timing errors in the rising limbs and larger timing errors in the falling limbs are visible. Partly, timing errors of the falling limb even mimic timing errors in the rising limb (compare also Fig. 3.8, lower panel). The false inflation of the uncertainty envelope due to the timing error is undesirable. The reasons for it are, however, manifold. Possible ways forward to narrow the time-inflated SD uncertainty envelope would be (i) to replace the static timing error model (Eq. (3.5)) by a relative representation, e.g. by using mean event duration, (ii) to further differentiate the error distributions, e.g. according to flow magnitude and (iii) in the consideration of the autocorrelation of the errors which is typically large in streamflow data. Of course, errors in the coarse-graining can also contribute to false inflation. In comparison, the uncertainty envelope

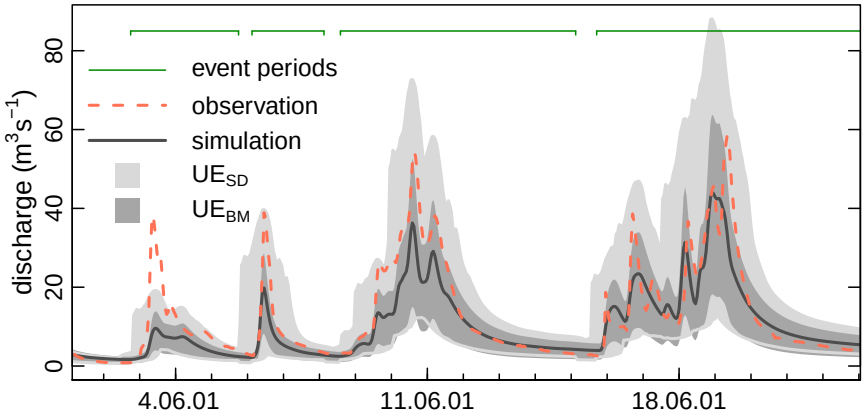


Figure 3.7: Time series detail showing the resulting one- and two-dimensional uncertainty envelopes around the historic streamflow simulation. The envelopes were derived upon Series Distance (UE_{SD}) and the benchmark approach (UE_{BM}), respectively, using error dressing. Please note that the coverage of the SD and BM envelope may differ for different subsets of the time series, like in this detail. For the entire time series, however, the coverage of BM and SD are identical.

of the BM model appears slimmer and more precise. However, due to the lack of consideration of timing uncertainties, especially during steep flood rises, the uncertainty envelopes become very narrow. Such a "vanishing" of the uncertainty envelopes implies that there are no timing errors to be expected at all (compare, e.g., the period 6–7 June

2001 in Fig. 3.7), which is deceptive, keeping in mind the SD results for the timing errors (Fig. 3.6). We thus consider this aspect a disadvantage of the one-dimensional error dressing method, especially as the timing of flood rises is often critical in hydrological applications (Seibert, Skublics, and Ehret, 2014).

The statistical evaluation of the different uncertainty envelopes (Table 3.3) confirms the visual impression: the BM uncertainty envelope outperforms SD in terms of absolute and relative precision (PRC and PRC*, respectively) given identical coverage (ϕ). On average, UE_{SD} is 3.1 $m^3\,s^{-1}$ wider than the benchmark envelope, which corresponds to a relative difference of 30 % as indicated by PRC*. This suggests that the use of the SD concept to construct uncertainty envelopes implies a trade-off between two effects: on the one hand, the explicit consideration of timing errors potentially yields better-tailored uncertainty envelopes, as apparent timing errors can be treated as such. On the other hand, if timing is not a dominant or at least substantial component of the overall error, the time-spreading effect of the SD envelope construction can lead to an undesirable inflation effect. In our case study, the latter effect apparently predominated. For hydrological forecasts based on uncertain meteorological forecasts, however, the opposite may be the case.

Table 3.3: Coverage (ϕ), precision (PRC), and relative precision (PRC*) of uncertainty envelopes. UE_{SD} and UE_{BM} denote Series Distance and benchmark error model, respectively. The last column (p) provides the percentage of sampled values of the corresponding distribution(s).

Uncertainty envelope	ϕ (−)	PRC ($m^3\,s^{-1}$)	PRC* (-)	p (%)
UE_{SD}	80.5	8.2	1.3	76
UE_{BM}	80.0	5.1	1.0	80

3.5.3.3 *Disentangling the importance of magnitude and timing errors*

To further investigate the individual effects of errors in timing and magnitude, we also applied them separately to the simulated time series. To this end we applied case-specific subsets of the error distributions – i.e. 2-D errors for rising and falling limbs and 1-D error distributions for low flow – to each point of the simulated time series just as in the previously described error dressing approaches. The difference was that we did not apply the entire error subset (oval or circle) but its projection on the time and magnitude axis, respectively. The resulting uncertainty bars therefore extend from the maximum to the minimum magnitude (upper panel) and timing (lower panel) values of the error subsets and are depicted in Fig. 3.8. For comparison we also plotted the magnitude errors of the BM approach. In this representation it becomes obvious that the error bars of the SD and BM approach show considerable differences with respect to extent and symmetry. For the magnitude error bars the deviations are

most pronounced in the rising limbs and less so in the falling limbs and during low-flow conditions. While the SD method reflects the underling characteristics of the errors, the BM method applies the same error to all cases. Constructing an uncertainty envelope from only the SD magnitude errors would yield an envelope comparable to that of BM but be more variable and have higher uncertainty towards overestimations than towards underestimations. Note that the true distribution of errors within the error bars is unknown.

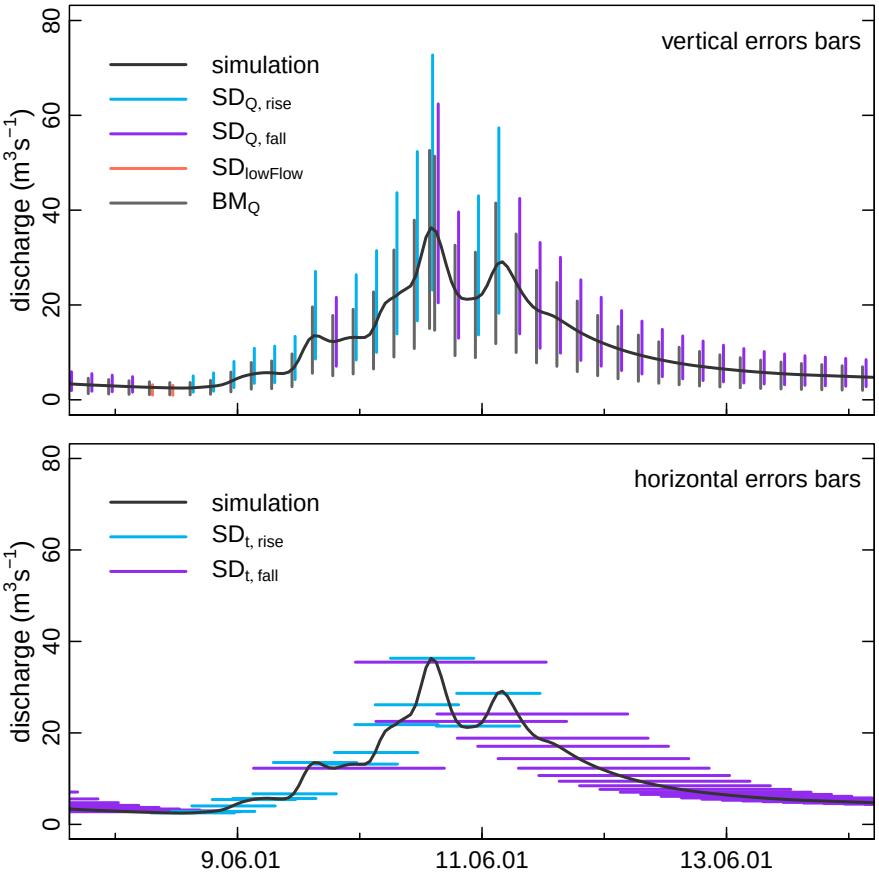


Figure 3.8: Vertical and horizontal error bars. The upper panel shows magnitude error bars (Q) for the Series Distance (SD) method and the benchmark (BM) approach. For SD different error bars are drawn for low-flow conditions and rising (rise) and falling (fall) limbs. In the BM case the same error bars are applied in all cases. The lower panel shows the corresponding timing error bars (t) of SD (not available for BM), again separately for the rising and falling limbs. To improve readability we plotted error bars only every third hour and introduced a slight time offset between SD and BM (upper panel only). Both panels show a subset of the hydrograph section depicted in Fig. 3.7 and are based on the same data.

The lower panel in Fig. 3.8 reveals that the uncertainties with respect to timing are considerable, typically during the recessions. Combining horizontal and vertical errors to construct the 2-D SD uncertainty envelope using the method described in Sect. 3.3 will inevitably cover a large region. While this is undesirable, it points towards pos-

sible alternatives to construct uncertainty ranges: rather than uniting the horizontal and vertical uncertainty components, intersecting them, i.e. to use only elements which are common to both error components,s would also be possible, for example, and most likely narrow the uncertainty envelope. Also, discharge time series usually exhibit considerable autocorrelation and so do related simulation errors. Exploiting this memory effect by time-conditioned sampling of the error distribution via a Markov process would be a further alternative to better tailor uncertainty envelopes (Montanari, Rosso, and Taqqu, 1997; Vrugt et al., 2008).

Finally, even if the SD error distributions are not used to construct uncertainty envelopes, knowledge of magnitude and timing error distributions is valuable for model diagnostics: in their approach to identifying characteristic error groups in hydrological time series Reusser et al., (2009) had to inversely infer the effect of timing errors to their signatures; SD offers a method to directly measure timing errors and thus to improve this step.

3.6 CONCLUSIONS

The main goal of this paper was to present major developments in the SD concept since its first version presented by Ehret and Zehe, (2011). These include the development of an iterative optimization procedure which effectively mimics coarse-graining of hydrographs when comparing them visually. The parameters of the inherent objective function were derived manually for this study; for more widespread applications, however, we recommend an in-depth sensitivity analysis using data from different regimes. Coarse-graining yields a set of matching segments within observed and simulated hydrological time series and the optimal degree of coarse-graining, both of which can be used as input for comparative hydrograph analysis. Further developments include the introduction of a scaled error model, which has proven to be better applicable across large discharge ranges than its non-scaled counterpart, and error dressing, a concept to construct uncertainty ranges around deterministic streamflow simulations or forecasts. Error dressing includes an approach to sample empirical error distributions by increasing variance contribution, which we extended from standard one-dimensional distributions to the two-dimensional distributions of combined time and magnitude errors of SD.

Applying the SD concept and a benchmark model (BM) based on standard magnitude errors to a 6-year time series of observations and simulations in a small alpine catchment revealed that different flow conditions (low flow and rising and falling limbs during events) exhibit distinctly different characteristics of timing and magnitude errors with respect to mean and spread. Separate treatment of timing and magnitude errors and a differentiation of flow conditions as done in SD is thus recommended in general as it preserves useful information. Exploiting these characteristics and their correlations can support targeted model diagnostics. Deeper insights can easily be

provided if the error distributions are further differentiated by discharge magnitude classes, by season, or by considering the temporal autocorrelation of errors. The latter would allow the development of a time-conditioned error sampling strategy when constructing 2-D uncertainty envelopes.

Applying the error distributions of both SD and BM to construct uncertainty ranges around the fairly accurate simulation revealed a remarkable timing uncertainty. This suggests that we commonly underestimate the role of horizontal uncertainties in streamflow simulations. For the given data, the BM-derived uncertainty ranges were in consequence visually narrower and statistically superior to the SD ranges. This suggests that the use of the SD concept to construct uncertainty envelopes according to the proposed error dressing method implies a trade-off between two effects: on the one hand, the explicit consideration of timing errors potentially yields better-tailored uncertainty envelopes, as apparent timing errors are treated as such. On the other hand, the time-spreading effect of the SD envelope construction, which essentially is the union of the time and magnitude error uncertainty ranges, can lead to an undesirable inflation. For the case study data, the latter effect predominated, while for hydrological forecasts based on uncertain meteorological forecasts the opposite may be the case. This also opens interesting avenues for new ways to construct uncertainty ranges based on the SD concept, e.g. as the intersection (rather than the union) of the two error components.

We conclude that Series Distance is an elaborate concept for the comparison of simulated and observed streamflow time series which can be used both for detailed hydrological analysis and model diagnostics. Its application, however, involves considerably more effort than standard diagnostic measures, which are typically justified if timing errors are dominant or of particular interest. More generally, we believe that for hydrological studies there is a large potential for intuitive distance metrics such as the hydrograph matching algorithm proposed by Ewen, (2011) or the SD concept, which should be further exploited as suggested by Crochemore et al., (2014).

To foster the use of the SD concept and the methods therein we publish a ready-to-use Matlab program code alongside to the manuscript under a Creative Commons license (CC BY-NC-SA 4.0). It is accessible via `https://github.com/KIT-HYD/SerieSDistance`. This repository also includes extended versions of the SD concept which we did not describe in full length here. These allow for a continuous usage of the method (no data on events required) and/or a differentiation of vertical errors according to flow magnitude.

Part IV

EXPLORING THE INTERPLAY BETWEEN STATE, STRUCTURE AND RUNOFF BEHAVIOUR OF LOWER MESOSCALE CATCHMENTS

Understanding runoff production and the underlying site-specific physiographic controls is important but rarely fully understood. In order to advance in this matter, I propose a set of diagnostic signatures for commonly available data and apply it to a small catchment inter-comparison study which uses data from the Danube Basin in Southern Germany. The suggested signatures show that different sites are *functionally* similar for base-flow generation, storm-runoff production and the seasonal water balance. Further, they highlight that biotic controls are non-stationary, that *intensity controlled* runoff formation mechanisms may be important and that variables may have higher explanatory power when they are treated as *parameter groups*.

The study is published as discussion paper in *Hydrology and Earth System Science*. Part IV is a reprint of:

Seibert SP, Jackisch C, Ehret U, Pfister L & Zehe E (2016): Exploring the interplay between state, structure and runoff behaviour of lower mesoscale catchments. Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-109, in review.

4

PART 4: INTERPLAY AMONG STRUCTURE, STORAGE AND RUNOFF

ABSTRACT

The question of how catchments actually "function" has probably caused many sleepless nights as it is still an unsolved and challenging scientific question. Here, we approach this question from the similarity perspective. Instead of comparing single physiographic features of individual catchments we explore the interplay of state and structure on different runoff formation processes, aiming to infer information on the underlying "functional" behaviour. Therefore, we treat catchments as lumped terrestrial filters and relate a set of different structure and storage descriptors to selected response measures. The key issue here is that we employ dimensionless quantities exclusively by normalizing the variable of interest by its limiting terrestrial or forcing characteristic. Specifically we distinguish extensive/additive and intensive/non-additive attributes through normalizing storage volumes by maximum storage capacities and normalizing fluxes (e.g. discharge) by permeability estimators. Moreover, we propose the normalized temporal derivative of runoff as a suitable measure to detect intensity-triggered (high frequency) runoff production.

Our dimensionless signatures evidently detect functional similarity among different sites for baseflow production, storm runoff production and the seasonal water balance. Particularly in the latter case we show that normalized double and triple mass curves expose a typical shape with a regime shift that is clearly controlled by the onset and the end of the vegetation period which we can adequately characterize by a simple temperature index model. In line with this, temperature explained 70 % of the variability of the seasonal summer runoff coefficients in 22 catchments distributed along a strong physiographic and climatic gradient in the German part of the Danube basin. The proposed non-additive response measure detected signals of high frequency intensity controlled runoff generation processes in two alpine settings. The approach, in fact edge filtering, evidently works when using “low-pass” filtered hourly rainfall-runoff data of mesoscale catchments ranging from 12 to 170 km^2.

We conclude that vegetation exerts a first order control on summer stream flow generation when the onset and termination of summer are more significantly defined by temperature than simply by the actual Gregorian day. We also provide evidence that properties describing gradients (e.g. surface topography) and resistances (e.g. hydraulic conductivities) may be much more powerful in explaining runoff response behaviour when they are treated as groups compared to their individual use. Lastly, we show that storage estimators such as the proposed normalized versions of pre-event discharge and antecedent

moisture can be valuable predictors for event runoff coefficients: For some of our test regions they explain up to 70 % of their variability.

4.1 INTRODUCTION

4.1.1 *Hydrological similarity as weak form of causality*

How close must two catchments be with respect to state and structure such that they produce runoff in a similar way? Based on the findings of Dooge, (1986) we establish our studies essentially on the expectation that hydrological systems are essentially deterministic. Hence, identical inputs of energy and rainfall will cause an identical runoff response, if two identical catchments are in the same state. This crude deterministic paradigm is, however, of low practical use, because neither the system state nor the structural setup are exhaustively observable. Hence, we can at best postulate that similarity of structure and state of a terrestrial system implies "similar" functioning (He, Bárdossy, and Zehe, 2011a; Wagener et al., 2007; Zehe et al., 2014). While such a weak form of causality may be easily defined in qualitative terms, its translation into useful similarity measures for structure, state and runoff response is far from being a straight forward exercise, particularly at the lower mesoscale.

4.1.2 *Challenges in defining structural similarity at the lower mesoscale*

Parts of the confusion stem from the inherent equifinality and non-uniqueness of most of our governing equations (Beven, 1989; Zehe et al., 2014). This is particularly true for runoff because an integrated mass flux leaving a catchment control volume is a non-unique product of a driving potential gradient and the control volume conductance (or its inverse resistance). This implies that systems that largely differ with respect to the topographical controls on the driving gradients and the pedo-geological controls on the integral conductance may produce runoff in a similar fashion (Binley and Beven, 2003; Wienhöfer and Zehe, 2014). Topographic controls and pedological controls on runoff generation must thus be interpreted as group, to be able to judge how they jointly control runoff behaviour. This requires metric data sets on topography as well as on soil water and aquifer characteristics. While the former is available as highly resolved digital elevation and landuse maps, the latter can in most cases at best be estimated using (very) coarse soil and geological maps in combination with pedo-transfer functions - not to mention the absence of data characterising preferential pathways. It is, hence, no surprise that many model-based similarity studies rely on categorical soil and landuse data and translate them into metric catchment descriptors by means of their areal share (e.g., Ali et al., 2012; Carrillo et al., 2011; Hundecha and Bárdossy, 2004; Kelleher, Wagener, and McGlynn, 2015; Merz, 2003; Sawicz et al., 2011). Notwithstanding that this approach is feasible when representing similar catchments by similar parame-

ters in conceptual models, it is way too simple to be conclusive for similarity of runoff production in the real world.

4.1.3 *Storage estimators and state estimators - how to normalize and how to achieve coherence?*

Storage estimators such as antecedent precipitation (Brocca et al., 2009; Heggen, 2001), pre-event discharge (Graeff et al., 2012; Refsgaard, 1997) or dynamic storage (Sayama et al., 2011) have been shown to be helpful to characterise storage in the catchment control volume and the related runoff proneness across scales (Tetzlaff, McNamara, and Carey, 2011). This is particularly appropriate when subsurface storage capacity controls runoff production (Struthers and Sivapalan, 2007; Struthers, Hinz, and Sivapalan, 2007a,b), which implies runoff to monotonically increase with storage and thus to be limited by additive quantities. Such quantities are rainfall depth and saturation deficit in the case of saturation excess (Dunne and Black, 1970), or rainfall depth and subsurface storage in the case of subsurface storm flow (Lehmann et al., 2007; Tromp-Van Meerveld and McDonnell, 2006).

Additive storage measures can easily be derived from either the catchment water balance or observed rainfall and discharge volumes (McNamara et al., 2011) and equally easily be upscaled, as soil and aquifer water content are additive quantities. However, absolute storage is difficult to compare between different pedolological settings as these measures require a meaningful normalization in order to be related to runoff processes. Furthermore, dynamic storage (Sayama et al., 2011) depends on the starting point of integration. As catchment inter-comparison studies should compare coherent time series, this starting point needs to be carefully chosen to ensure that integration starts at the same relate storage state. When the catchments of interest are spread across a wide topographic and climatic range, the same Gregorian day might be a very inappropriate choice, as further elaborated in section 4.2.1.

Notwithstanding the importance of storage estimators, they do not provide a full characterization of the catchment state. The latter particularly requires information on where in the catchment the water is stored (Nippgen, McGlynn, and Emanuel, 2015) and whether it is subject to strong, weak or no capillary and/ or osmotic forces. Unfortunately, soil water potentials, plant water potentials and piezometric heads are intensive state variables and thus non-additive. They can neither be determined as residuals of a balance equation nor can they easily be scaled up in an additive manner (De Rooij, 2011; Rooij, 2009; Zehe, Lee, and Sivapalan, 2006). Hence, characterization of the full system state requires comprehensive, spatially highly resolved data sets on both soil moisture and soil water potentials (Zehe et al., 2013). As these are rarely available in mesoscale catchments, similarity and catchment inter-comparison studies are challenging to work with, giving a fairly incomplete characterisation of the system state. This is

particularly unpleasant because it is the potential gradients which determine the "forces" driving water and energy fluxes (Kleidon, 2012; Zehe et al., 2014).

4.1.4 *Dimensionless response measures for and beyond capacity controlled runoff formation*

The striking success of similarity theory and scaling based on dimensionless quantities throughout a range of disciplines such as hydraulics (e.g. Reynolds, Froude, Péclet number), acoustics (e.g. Helmholtz number), chemistry (e.g. Dammköhler numbers) or micro meteorology (e.g. the Monin–Obukhov length) motivated past research for useful dimensionless quantities characterizing hydrological similarity (e.g., Bahram, Pierre, and Odgen, 1995; Berne, Uijlenhoet, and Troch, 2005; Reggiani, Sivapalan, and Hassanizadeh, 2000; Schaefli et al., 2011; Struthers, Hinz, and Sivapalan, 2007a; Woods, 2009; Woods, 2003). The Budyko curve (Budyko, 1956) is probably the most generally accepted dimensionless analysis technique in hydrology to assess similarity in the steady state water balance by plotting the evaporative fraction against a dryness index. In line with these studies we hypothesize that dimensionless state-response diagrams are suitable candidates for similarity assessment for catchment inter-comparison. Proper normalization of state and response measures means to normalize using those climate and terrestrial system properties which limit runoff production. The rationale is that one can expect these dimensionless plots to remain invariant, as long as the limiting factors remain unchanged.

Normalization in the case of capacity controlled runoff formation is straightforward as it is limited by additive quantities, essentially storage and rainfall volumes. We may hence treat catchments as lumped terrestrial filters (Black, 1997) and normalize event scale runoff by total precipitation amount, and relate this to storage estimators such as antecedent precipitation (Blume, Zehe, and Bronstert, 2007; Graeff et al., 2012; Heggen, 2001), dynamic storage (Sayama et al., 2011) or pre-event discharge (Graeff et al., 2009; Kirchner, 2009; Zehe et al., 2010). A feasible normalization of storage estimators should be based on the minimum and maximum subsurface storage volume/depth or, if this information is not available, on the storage depth in the root zone of the soil. Similarly, we may compare annual double mass curves of normalized accumulated rainfall and runoff fluxes to discriminate differences in the seasonal interplay of storage and release (Hellebrand et al., 2008; Pfister, Iffly, and Hoffmann, 2002). Although all these measures and their normalization can in principle be determined as residuals of the water balance and from available maps, the devil lies in the details as further elaborated in section 4.2.1.

Detection and normalization of intensity controlled runoff production is, however, not that straight forward (Struthers and Sivapalan, 2007). Intensity controlled runoff generation is characterized by intensive, convective rainfall forcing and a fast, high frequency stream flow

response, reflecting onset of rapid subsurface flows (Lehmann et al., 2007; Wienhöfer and Zehe, 2014) and/ or infiltration excess (Niehoff, Fritsch, and Bronstert, 2002; Zehe et al., 2005). The latter is difficult to observe *in situ* during natural forcing conditions but its occurrence is well known from many artificial rainfall simulation experiments (e.g., Fiener, Seibert, and Auerswald, 2011; Fiener et al., 2013). Intensity controlled runoff production occurs in a threshold like manner (Lehmann et al., 2007; Ruiz-Villanueva et al., 2012; Struthers and Sivapalan, 2007; Zehe et al., 2007; Zehe and Blöschl, 2004) and is neither controlled (and limited) by additive rainfall properties nor by current storage. Hortonian overland flow production is for instance controlled (and limited) by the relationship of non-additive rainfall intensity and soil infiltrability (Horton, 1939; Zehe and Sivapalan, 2014). The latter is a conglomerate of unsaturated hydraulic conductivity, and suction head as well as of the density, depth and capacity of apparent macropores (Beven and Germann, 2013). Again, none of these quantities is additive during up-scaling.

As intensity control implies i) the high frequencies to be dominant and ii) first order control of non-additive characteristics, any form of spatial and temporal data aggregation essentially implies to loose parts or even the complete signal due to low-pass filtering. There are promising options to assess highly resolved patterns of rainfall based on weather radar (e.g., Ehret et al., 2008; Kneis and Heistermann, 2008) or to estimate catchment scale patterns of biotic macropores (Palm, Schaik, and Schröder, 2013; Schaik et al., 2014). However, discharge as our best observation of runoff formation inevitably represents a convolution of distributed runoff production and concentration, which inherently implies low-pass filtering.

A cardinal question is thus on the minimum requirements for detecting intensity controlled runoff generation. Related studies often operate at relatively small scales, relying on high frequency rainfall-runoff data in combination with breakthrough or flushing of either contaminants (Gassmann et al., 2013), artificial tracers (Wienhöfer et al., 2009), sediments (Martínez-Carreras et al., 2010) or even diatoms as *smart* tracers (Klaus et al., 2015; Martínez-Carreras et al., 2015). Most "operational" data sets however do not offer these sources of extra information and are at best available at an hourly resolution and for catchment sizes $\geqslant$ 40-50 km^2. The challenge to detect intensity controlled runoff production within inter-comparison studies seems at first sight similar to the challenge to repair a watch with a monkey wrench. One way forward might be to relate temporal changes in rainfall intensities to temporal changes in runoff - which means in fact to analyse the acceleration of input and output fluxes, as further elaborated in section 4.4.4.

4.1.5 *Objectives and research questions*

While being fully aware of all the listed challenges and shortcomings of operationally available data sets, we propose and test dimension-

less measures to discriminate differences in runoff generation (storage and/ or intensity controlled) in lower mesoscale catchments. In particular, we pose three main questions:

- Question 1: How feasible is the use of dimensionless state and/or storage-response diagrams to detect differences in event scale flood production, baseflow generation and the seasonal water balance?
- Question 2: Can we detect intensity controlled runoff formation as essentially a high frequency process based on low frequency data?
- Question 3: Which structural, climatic and ecological catchment characteristics explain the differences between different catchments and among different years and which of them operate in groups?

Our study area is the Bavarian part of the Danube basin in Southern Germany, which we introduce in detail in section 4.3 together with the data and model we use. More specifically, we use an operational data set from the federal water resources management agency and standard categorial data on landscape characteristics in about 130 lower mesoscale catchments. Additionally, we apply a calibrated water budget model which covers all of the sub-basins to ensure a consistent estimation of evapotranspiration and storage estimators such as dynamic storage. Particular emphasis within our inter-comparison study is on the issues of i) proper normalization of storage estimators and fluxes, ii) assuring coherence and similar quality of associated time series and iii) on an assessment of the different storage estimators with respect to explanatory power and redundancy.

4.2 CONCEPTUAL FRAMEWORK AND CANDIDATE DIAGNOSTICS

In this section we propose a set of dimensionless "functional diagnostics", suitable for catchment inter-comparison studies across a wide range of end members. Hence, we exclusively rely on commonly available landscape properties and hydro-meteorological data.

4.2.1 *Requirements of functional diagnostics*

Useful diagnostics for runoff response and catchment state need to be sensitive to the limiting factors and allow for a normalization of the responses and state variables to i) separate meteorological from terrestrial controls and ii) to test our perception on underlying structural controls. We expect intensity controlled runoff production to occur in landscapes characterized by strong gradients, shallow and poorly developed soils, high abundances of either very coarse or fine/clay substrates, sparse surface coverage, and/or geologies which develop rift aquifers. Capacity controlled runoff generation is deemed to dominate in landscapes characterized by weak gradients, well drained

and homogeneous textured soils (without remarkable clay or skeleton contents), and medium to high degrees of surface cover over parent materials that sustain pore aquifers. Also snow dominated areas are expected to exhibit capacity controlled behaviour. At the seasonal scale we additionally need to make sure of comparing coherent time series of similar data quality.

4.2.1.1 *Normalization of states and response measures*

In our study we compare three storage estimates: a rescaled version of dynamic storage, accumulated antecedent precipitation and pre-event discharge (see details in section 4.2.2.1). All three surrogates yield estimates of absolute storage depths (L). Their normalization requires measures for storage capacity of the different catchment subsurface compartments, which should reflect both total and active storage volumes as well as the fractions of free water and capillary bounded water in soil. Estimation of these storage properties is hampered by the unknown depth of the lower boundary of the control volume and the heterogeneity of the subsurface materials (Soulsby, Tetzlaff, and Hrachowitz, 2009; Spence, 2007; Troch, Paniconi, and Emiel van Loon, 2003). We thus normalize the storage estimators using average root zone field and air capacity, which are available through national soil maps (BGR, 1995). Despite their limitation in the vertical direction, these estimates are deemed to provide an indication of the relative importance of the storage volume containing capillary bounded water, which feeds evaporation and transpiration, and free water feeding groundwater recharge and runoff production (Zehe et al., 2014).

Normalization of rainfall-runoff response and seasonally accumulated runoff is straightforward in the case of capacity controlled runoff production by means of either total rainfall depth of an event or total annual precipitation. Baseflow during dry spells (radiation driven conditions) requires a different normalization based on estimates of aquifer permeability/transmissivity as these control water release. If this information is not available, as in our case, the average soil hydraulic conductivity provides an alternative.

4.2.1.2 *Coherence and quality of integral storage measures*

Estimators of water storage such as dynamic storage (dS) (Sayama et al., 2011) depend essentially on the starting point of integration (Pfister et al., 2003). Coherence, in terms of "achieving comparability", of storage time series hence requires that integration in all catchments starts at the same relative storage amount. This could for instance be after significant dry and/or wet periods, when subsurface wetness can be deemed as being either near saturation or near the minimum. Particularly in the case of a strongly seasonal climate, distinct dry and wet periods can be useful in selecting a proper start date.

As dS is i) based on the assumption of a closed water balance and ii) calculated from (areal) estimates of precipitation and model based

estimates of evapotranspiration, related uncertainties have a direct effect on the storage estimator. A straight forward quality check of dS is to plot it against normalized accumulated precipitation for several years, using the long term annual mean precipitation for normalization. By comparing patterns of dS for time periods of potentially similar accumulated input one may detect trends, non-monotonic step changes or other inconsistencies. In our case most of the 130 datasets did not pass this benchmark test (compare section 4.3).

Finally, we face a similar challenge of "when to start" when relating integral storage measures to normalized baseflow. This is because "onset" and "duration" of the baseflow recession may have variable definitions and meanings (Blume, Zehe, and Bronstert, 2007). Furthermore, discharge at the river gauge is an aggregation of runoff production, concentration and routing along the river network. These processes cover different spatio-temporal scales which make it increasingly difficult to determine a direct relationship of baseflow behaviour to integral storage measures when moving up in scale.

4.2.2 *Candidate storage, response and intensity estimators for baseflow, runoff events and the seasonal water balance*

This section introduces normalized storage and runoff response measures, their combination into dimensionless storage/state-response diagrams as well as their statistical analysis. We distinguish among i) the generation of baseflow during radiation driven conditions, ii) rainfall-runoff events as the driven case and iii) the seasonal water balance. The latter is separated into the winter term and the vegetation period, to explore the impact of vegetation controls. Our candidate diagnostic measures for high frequency runoff processes and intensity control are introduced at the end.

4.2.2.1 *Normalized storage measures*

Firstly, we use a normalized and re-scaled version of dynamic storage (dS^*) (see Appendix A.2.2 on this aspect). dS^* is calculated as the residual of the water balance equation, using estimates of areal precipitation (P), model based estimates of evapotranspiration (E) and observed discharge (Q). As given in Equation 4.1 we use the average soil storage volume for normalization, characterised by the sum of effective field capacity (eFC) and air capacity (AC) in the root zone (τ), since metric information on aquifer capacity is not available. Estimates of eFC_τ and AC_τ are taken from the national soil map of Germany (BGR, 1995):

$$dS^*(t) = \frac{\sum_{t=1}^{T} P(t) - Q(t) - E(t)}{AC_\tau + eFC_\tau} \tag{4.1}$$

dS^* is deemed to represent the total active bulk catchment water storage and we expected it to be associated mainly with deeper storage compartments and hence to control the slower flow processes. Values of dS^* around zero indicate dry conditions whereas values

near one indicate that dynamic storage is equal to the root zone storage volume. Note that both values > 1 (e.g. during the occurrence of snow) and < 0 may occur and absolute values must not be interpreted.

The second storage estimator is chosen to better reflect near surface storage. Similar to other studies (Brocca et al., 2009; Graeff et al., 2012; Heggen, 2001) we estimate normalized antecedent moisture (θ^*), which is equal to the difference between precipitation and evaporation totals within the last seven days (T=7 days in Equation 4.2) normalized again by the average soil storage volume:

$$\theta^*(t) = \frac{\sum_{t=T-7d}^{T}(P(t) - E(t))}{AC_\tau + eFC_\tau} \tag{4.2}$$

Lastly we use a normalized specific pre-event discharge (Q^*) averaged across the last seven days:

$$Q^*(t) = \frac{\frac{1}{n}\sum_{t=T-7d}^{T} Q(t)dt}{AC_\tau + eFC_\tau} \tag{4.3}$$

The main disadvantage of Q^* is that it cannot be attributed to any specific subsurface storage compartment as it inevitably represents a combination of both, storage and release. The advantage is that it relies on the best observation we have.

4.2.2.2 *Baseflow generation during non-driven conditions*

To explore controls of catchment structure and storage on baseflow generation we relate specific baseflow depths (Q_b), normalized by the bulk average catchment hydraulic conductivity, to the different storage measures. To this end we define baseflow conditions as follows: $ET < 0.1$ mm, no occurrence of snow, $\frac{dQ}{dt} < 0$ and no input in $P > 0.1$ mm for a period of at least one, three and five days. The one- and three-day period data sets are used to visually inspect how fast Q decreases after precipitation ceases, which indicates how fast the terrestrial filter properties become dominant. Within our statistical analysis we exclusively consider stream flow data where the last input in $P > 0.1$ mm was at least five days ago. For response normalization we use the arithmetic average saturated hydraulic conductivity (Ks) of the catchment (Equation 4.4), since other estimators for bedrock permeability were not available. Ks is estimated for each catchment based on available grain size distribution using *Rosetta's* pedo-transfer functions (Schaap, Leij, and Genuchten, 2001).

$$Q_b^* = \frac{Q_b}{Ks} \tag{4.4}$$

Normalized baseflow is then related to dS^* and θ^* using the Spearman's rank correlation coefficient (ρ) and the non-parametric test of significance proposed by Best and Roberts, (1975). In the case of significant relations (p-values<0.001), we try to identify an empirical storage-baseflow relationship by fitting power laws using dS^* and θ^*

as predictors using *R* (R Core Team, 2015). The quality of these relationships are judged by comparing their root-mean-squared-error to the standard deviation of the normalized baseflow values (nRMSE).

Our approach is in line with past attempts to relate stream flow variations and drainage behaviour of hillslopes or catchments (e.g., Brutsaert and Nieber, 1977; Laurenson, 1964; Rodríguez-Iturbe and Valdés, 1979) and searches for feasible storage-baseflow relationships (Kirchner, 2009). Here, we also test the spatial consistency of these storage-discharge relationships by comparing the multiplier in the power law (as estimate of the effective catchment permeability) to the corresponding variation of Ks between the catchments.

4.2.2.3 *Event scale rainfall-runoff response*

At the scale of individual rainfall-runoff events we relate event runoff coefficients (CR_E), defined as total event quick flow volume ($\sum Q_E$) divided by total precipitation ($\sum P_E$) (Equation 4.5), to the different storage measures. To assure comparability of runoff coefficients, as recommended by Blume, Zehe, and Bronstert, (2007), and to assure a sufficiently large sample we use an automated detection of rainfall-runoff events based on a modification of the constant-k method (Blume, Zehe, and Bronstert, 2007) (details on the method are provided in appendix A.2.3).

$$CR_E = \frac{\sum Q_E}{\sum P_E} \tag{4.5}$$

We then select rainfall-runoff events with daily precipitation depth $\geqslant$ 10 mm and calculate both, coefficients of determination (Pearson) and Spearman rank correlation coefficients among CR_{Es} and the three different normalized storage estimators. Significant relationships are identified by p-values<0.001 in a two-sided t-test or the non-parametric test of Best and Roberts, (1975). These are interpreted as evidence for capacity controlled runoff production. In case the respective storage measure were uncorrelated, we test multiple regressions between CR_E and dS^*, Q^* and/or θ^*, respectively.

4.2.2.4 *Storage control on seasonal runoff generation*

To shed light on the seasonal dynamics of catchment storage and release we compile normalized annual double mass curves (nDMC) and triple mass curves (nTMC) for different hydrological years. Normalized double mass curves relate cumulated runoff ($cum.Q/\sum P$) to cumulated precipitation ($cum.P/\sum P$). The normalized triple mass curve adds cumulated evapotranspiration ($cum.E/\sum P$) as the third dimension to the plot. The rationale is to check whether the annual water balance is closed within a hydrological year, or whether the system carries stored water into the next year. The winter period and vegetation periods are separated using a temperature index model proposed by Menzel et al., (2003) and analysed separately.

The nDMCs within different catchments are compared according to a) the average and mean absolute deviations of their slopes within the

winter and vegetation period, b) the presence and onset of a regime shift marked by plateaus and c) the mean and inter-annual variation of the annual runoff coefficient (CR_{yr}). Regime shifts are further analysed based on the anti-correlation of summer and winter runoff coefficients (CR_S and CR_W) with actual annual evaporation from available water balance simulations. Finally, we attribute differences within the double and triple mass curves to a range of different (n=24) structural and climatic properties of the catchment including temperature sums, characteristics of the grain size distribution, surface cover and several others. Particularly, we test the product of topographic gradient and saturated hydraulic conductivity as an explanatory variable, as they are considered to act in concert.

4.2.2.5 *Intensity controlled runoff generation*

Our initial idea was to detect high precipitation rates as those being larger than the estimated hydraulic conductivity and to compare this to peak flow of events normalized with peak intensity of rainfall. To correct for the temporal mismatch between the maxima P and Q, we intended to employ a mean response time defined on the lag cross correlation between P and Q for each individual event (Kirchner, 2009). However, this approach did not yield clear signals due to several likely reasons. Although hot spots in rainfall intensities are known to be localised and dynamic (Fiener and Auerswald, 2009; Goodrich et al., 1995), we are left having to treat them as spatial rather uniform values due to the low density of rain gauges in our study area. Moreover, texture based estimators using the Rosetta pedo-transfer functions (Schaap, Leij, and Genuchten, 2001) remained as the only option and left us without a proper estimator of the influence of preferential pathways.

To separate high intensive rain showers from low and moderate intensive events we next calculate normalized rainfall event duration (T_E^* (h)) as the ratio of total event rain depth ($\sum P_E$) divided by the maximum observed precipitation intensity ($P_{E,max}$), for all rainfall events exceeding a threshold of 10 mm. The threshold of 10 mm h^{-1} is recommended by the German Weather Service (DWD) to detect strong rainfall events.

$$T_E^* = \frac{\sum P_E}{P_{E,max}} \tag{4.6}$$

We expect convective, high intensive and extreme rainfall events to cluster at short normalized event durations with a large total amount. Consequently, we relate the maxima in the temporal changes of discharge ($dQ_{E,max}$) and precipitation ($dP_{E,max}$) (both in mm h^{-2}) - which implies relating the acceleration of rainfall with stream flow mass. As high frequency processes are characterised by sharp peaks, we expect this normalized and dimensionless intensity change (I_E^*) (Equation 4.7) to separate intensity controlled from capacity controlled runoff production as intensity controlled conditions cluster at large

I_E^*. Note that I_E^* is an intensity measure and thus non-additive. It is independent from the runoff coefficient.

$$I_E^* = \frac{dQ_{E,max}}{dP_{E,max}} \tag{4.7}$$

Both, the normalized event duration and the normalized maximum change in stream flow are jointly analysed within scatterplots. Here, we expect intensity controlled processes to cluster around small values of T_E^* and large values of I_E^*. Additionally, we compile three-dimensional scatterplots using $\sum P$, $dP_{E,max}$ and $dQ_{E,max}$ on the x, y and z-axis respectively. Here we expect high $dQ_{E,max}$ to be associated with high $dP_{E,max}$ whereas $\sum P$ is deemed to be unimportant.

4.3 STUDY AREA AND DATASET

The feasibility of the above introduced signatures is tested by intercomparing operational data from the Bavarian Environmental Agency (LfU) for 130 catchments located in the Bavarian Danube basin ($\approx$ 45.000 km^2). In section 4.3.2 we detail the differences in the climate and physiographic setting of our test catchments and present our perception of the dominant hydrological processes. Before that, we will briefly discuss the quality of the database, which in fact was in most catchments so poor that the majority of the sites had to be excluded from the analysis.

4.3.1 *Data quality and selection of headwater catchments*

We focus on lower mesoscale catchments to minimize routing effects and select all gauged headwater sites $\leqslant 170 km^2$ within the Bavarian part of the Danube basin. For this, hourly hydro-meteorological time series from the period 01.11.1999 until 31.10.2004 are available. The data base in the resulting 130 catchments is analysed according to a set of different quality criteria. We only include catchments where i) at least one meteorological station was closer than 20 km, ii) the total absolute water balance error was smaller than 5 % , iii) the amount of missing and/or implausible meteorological data was < 5 %, and iv) where the streams are not subject to any severe regulation. This screening resulted in only 22 catchments being classified as suitable for the analysis. The sites are spread across the Bavarian part of the Danube basin (Fig. 4.1 and Appendix A.2.4).

The densest coverage in meteorological stations was for precipitation with a total number of 244 stations. The coverage of the other meteorological variables was much coarser, with 59, 55 and 43 stations for temperature, humidity and radiation, respectively. Since these numbers include stations which are up to 20 km apart from the finally selected 22 headwater catchments, we even tolerated lower densities in meteorological stations as e.g. in the Mopex data set (Duan et al., 2006; Schaake et al., 2000). The lowest densities of meteorological stations are located in the southern alpine areas and the corresponding

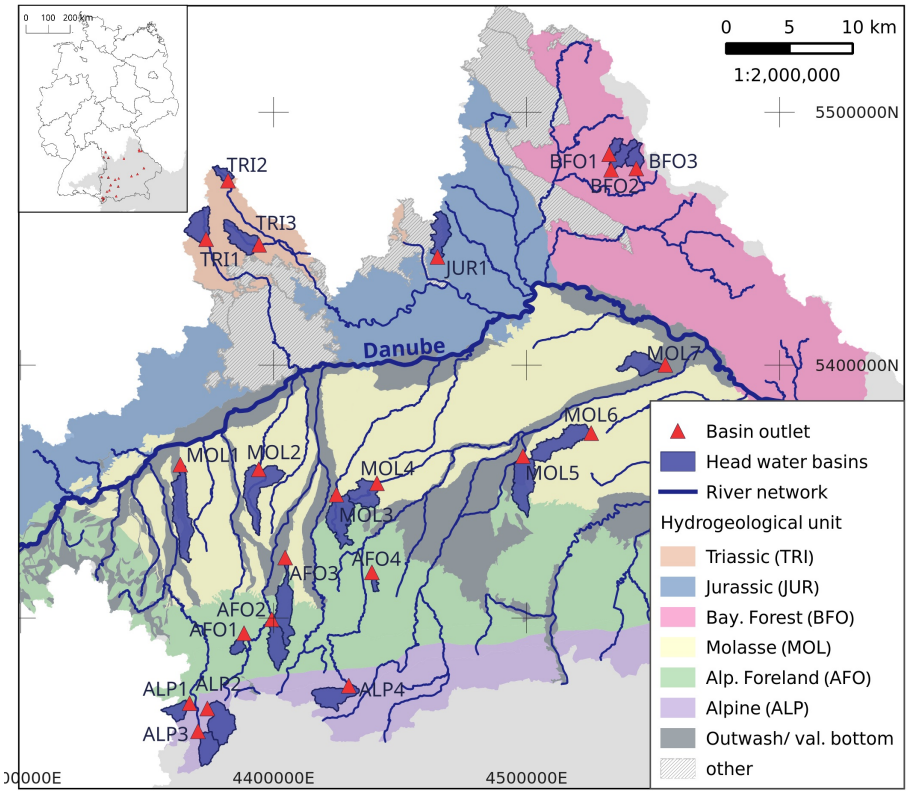


Figure 4.1: Upper Danube catchment in southern Germany with selected head water basins (blue polygons), corresponding gauges (red triangles) and major river network (blue lines). The site identifiers (IDs) refer to the corresponding (hydro)geological unit (color coded map in the background, adapted from BGR and SGD, (2015)) and a single Arabic numeral. Moving from the North-West to the South we differentiate TRI (Triassic), JUR (Jurasic), BFO (Bavarian Forest), MOL (Faulted Molasse), AFO (Alpine Foreland) and ALP (Alpine). Appendix A.2.4 provides links between the site IDs and the real gauge names. The inset in the upper left corner shows Germany's federal state boundaries, the individual head water outlets and the basin of the Danube (grey area). The grid coordinates refer to the Gauss-Kruger zone 4 projection (CRS identifier EPSG:31468).

foothills and in the north-eastern parts of the Bavarian Danube catchment. Catchment characteristics were derived from different digital map products of Germany such as for soil (scale: 1:1,000,000) (BGR, 1995), hydrogeology (scale: 1:1,500,000 and 1:500,000) (BGR and SGD, 2015; Duscher et al., 2015) and geology (scale: 1:1,000,000) (Toloczyki et al., 2006), a digital elevation model with resolution of 25 m, the CORINE land use data (as of 2006) and the official stream network provided by LfU. Last but not least, we employed the conceptual hydrological model LARSIM (Ludwig and Bremicker, 2006) as it provides consistent areal estimates of evaporation, rainfall and snow water equivalent. LARSIM has been calibrated for all study catchments and operates at hourly time steps. Evaporation is simulated using the Penman-Monteith equation. Model input is based on interpolated station data (grid point method, NOAA, 1972). Additional information

on the Bavarian part of the Danube basin, the hydro-meteorological data and the Larsim model can be obtained from Seibert, Skublics, and Ehret, (2014).

4.3.2 *Landscape setting and perceptual models of runoff generation*

Topography, landuse, geology, soil and aquifer properties are highly variable among the different headwaters of the Bavarian part of the Danube basin, as the region was unequally covered by ice during the last ice age. The remaining 22 catchments reflect the entire physiographic range (see Tables A.2 and A.3 in the appendix of part IV). This is underpinned by the large range of topographic gradients (ϕ) (Table A.2) calculated according to McGuire et al., (2005) as the flow path length from each pixel to the stream divided by the corresponding difference in height using the Whitebox geographic analysis toolbox (Lindsay, 2014). The climate gradient is also rather strong with mean annual precipitation (MAP) ranging from 600 mm in the northern sub-catchments to more than 2000 mm in the southern, alpine areas and a total average of 1000 mm. Annual potential evapotranspiration ranges from 350 to 600 mm. Both P and E regimes are characterized by distinct seasonal cycles. In some of the southern alpine areas more than 50 % of the precipitation may fall as snow.

Based on the dominant physiographic properties (referencing to Table A.2 and A.3) we grouped the 22 catchments into 5 major classes, which largely rely on (hydro)geology and detail on the expected dominant processes drawing from Peschke et al., (1999) and Schmocker-Fackel, Naef, and Scherrer, (2007), which are categorized into being capacity controlled or intensity controlled:

- The "Alpine sites" (ALP) in the very south are dominated by poorly developed and shallow soils (average root zone depth $\leqslant$ 35 cm) with high contents of skeleton and coarse material (average pore volume 110 mm, $Ks \approx 1e-6$ m s^{-1}) over highly productive fissured (partly karstified) aquifers. The surface cover is sparse with a clear dominance of forests and meadows. Rock outcrops occur, particularly above the tree line which is approximately around 1800 m.a.s.l in this environment. Catchments of this physiographic region (ALP1 ... ALP4, n=4) exhibit strong geopotential gradients (median $\phi = 0.36$) and receive about 1500 mm annual rainfall. These characteristics clearly suggest a dominance of rapid flow paths i.e. surface runoff, pipe flow/bypassing and rapid sub-surface stormflow. Here, we might thus expect high frequency, intensity controlled runoff formation, at least during extreme conditions.

- The "Triassic catchments" (TRI) (n=3) are composed of well-drained, poor to moderately developed sandy soils (mean root zone depths of 50-80 cm) with high portions of coarse material. Regosols, rendzinas, cambisols and partly podzols with rather weak vertical differentiation over calcareous sandstone

(TRI1, TRI2) and sandstone (TRI3) prevail. The parent material sustains moderately to highly productive pore and fissured aquifers. The land use is dominated by arable land (about 60 %). Long-term mean annual precipitation is around 750 to 800 mm. The median gradients within the three catchments differed slightly between 0.028 and 0.038 (-) which is an order of magnitude smaller than in the alpine areas. These characteristics suggest a perceptual model, where subsurface matrix flow dominates, and the aquifer strongly controls runoff generation. However, coarse substrates and corresponding structures may also sustain rapid flow paths and saturation excess during high intensity rainfall events.

- The faulted "molasses basin" (MOL) and adjacent transition areas belong to a heterogeneous region which hosts seven of our sites on mostly well developed, medium and deep cambisols (root zone depths > 70 cm) with high contents of aeolian sediments (silt and loess). The parent material is often composed of sheet gravel (MOL1, MOL2, MOL3, MOL5) and sedimentary rock and fluvial sediments (MOL4, MOL6, MOL7) which predominantly sustain low to moderately productive aquifers. In these catchments pore volumes are partly well above > 300 mm. The soils are fertile which promotes an intensive agricultural use. The surface topography is characterized by soft hills and U-shaped valleys. The corresponding gradients in geopotential are weak. Therefore, we expect that sub-surface capacity controlled (matrix) flow is the dominant runoff process. However, during high intensity rainfall Hortonian overland flow (due to surface crusting on arable land) and saturation overland flow due to reduced hydraulic conductivity is deemed to create a mixture of capacity and intensity controlled runoff formation.

- The catchments in the "Bavarian Forest" (BFO) (n=3) consist of loamy, partly sandy cambisols with comparably high contents of skeleton (in some areas up to 75 Vol.-%). These lie over crystalline granite and gneiss which are fractured but practically non-aquiferous rocks. The root zone depth is on average 60 cm. Forests and meadows cover 60 to 90 % of the surface. The topography is more pronounced and median gradients reach values up to 0.08. In these areas we expect that preferential flow pathways contribute significantly to runoff generation, but merely in a capacity controlled manner.

- The data set also includes four catchments from the "Alpine Foreland" (AFO1...AFO4). Like the MOL-area this region exhibits complex characteristics as it was altered by three different glacial advances (and retreats). Consequently, we observe high spatial variations in the geological parent material and thus, also in the soils, land use and hydrological characteristics (see Table and A.2 and A.3). The same applies for topography, as it is a relict of the different glacial periods. The relief, though com-

posed of similar gradients as e.g. the Triassic or Molasse sites, includes a rich variety of landforms typically found on ground, end and lateral moraines such as rolling foothills, (glacial) lakes, swamps and smaller surface water courses. Hence, there is no single dominating perceptual model on runoff formation available. Also the importance of different storage compartments cannot be estimated for this region as a whole, but needs to be evaluated individually for each site.

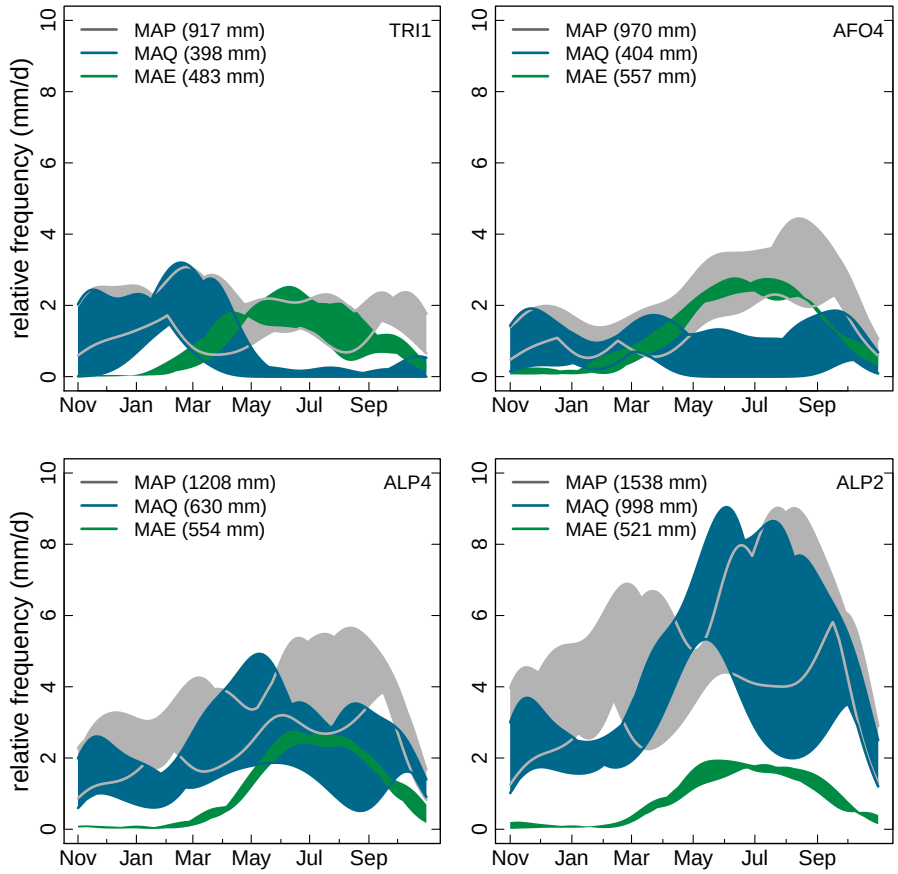


Figure 4.2: Regime curves (kernel density estimates) of observed areal precipitation (grey), discharge (blue) and calculated areal mean evapotranspiration (Penman-Monteith) (green) of four selected head water catchments. The width of the individual bands illustrates the inter-annual variation during the four year lasting period. In all cases identical kernels and bandwidths are used for variables of the same type. MAP, MAQ and MAE provide information on the four year mean annual average for P, Q and E respectively.

To further illustrate variations in the seasonal water balance, we present regime curves for four different catchments (Fig. 4.2). The catchment TRI1 (Fig. 4.2, top left) receives a fairly constant input in P throughout the year, but releases Q with a strong seasonality and pronounced minimum during summer. Compared to the other sites the inter-annual variation in E is rather large. The catchment AFO4 (Fig. 4.2, top right) in contrast shows seasonality in P but a fairly constant

output in Q. ALP4 (Fig. 4.2, bottom left) and ALP2 (Fig. 4.2, bottom right), which are both alpine sites, show a pronounced minimum in Q during February due to snow storage. ALP2 however shows a very large range in both P and Q during summer, which suggests little buffering and a high reactivity. In contrast, ALP4 has a much more damped response to P during summer and a more pronounced seasonality in ET.

4.4 RESULTS

The following section documents the performance of our diagnostics. More specifically, we present selected dimensionless storage/-state-response analyses that corroborate either their feasibility or their failure in discriminating differences in runoff behaviour in combination with the selected statistical measures introduced in section 4.2.2.2.

4.4.1 *Storage and structure control on baseflow generation*

During low flow conditions, the storage estimators dS^* and θ^* are in most cases linearly independent. Table 4.1 presents the corresponding statistics. However, significant relationships are encountered in 7 out of 22 cases, with the highest Spearman rank correlation coefficient (ρ) between dS^* and Q_b^* is 0.39 with an average of 0.07. All catchments except one have high and significant rank coefficients of determination between dS^* and Q_b^* with values ranging from 0.12 to 0.88 and an average of 0.59. Hence, dS^* seems to be a valuable predictor for low flow in a rather wide range of environmental conditions. We also find significant relationships between θ^* and Q_b^* in about 50 % of all cases but with rank correlation coefficients being all smaller than 0.28 (except for catchment AFO2, were ρ was 0.54). Hence, θ^* possesses much less predictive power.

Five catchments (MOL5, MOL4, TRI1, MOL1, BFO3, MOL7) reveal power model exponents close to 1 ± 0.3 for their estimated normalized storage-discharge relationship (Table 4.1). This corroborates a linear storage-baseflow relationship in line with Fenicia et al., (2005). For the remaining catchments, we obtain exponents clearly different from 1, suggesting a non-linear interplay of storage and baseflow production in the majority of the catchments. This finding is also supported by 2D scatterplots (Fig. 4.3) which clearly show a strongly non-linear relationship between dS^* and Q_b^* at e.g. AFO3, MOL6, MOL2, BFO1, JUR1 and other sites. The nRMSE of the estimated storage-baseflow relations was on average 0.74, with best values of 0.43 and worst values larger than 1. Hence on average, the storage-discharge relationships are a better predictor than the average Q_b^*. Furthermore, we do not find a distinct spatial pattern in the exponents as both cases (linear and non-linear) occurred throughout different geologies and climate settings.

The normalized baseflow is in fact the flow multiplied with the inverse of the conductance. Due to the gradient-flux relationship we

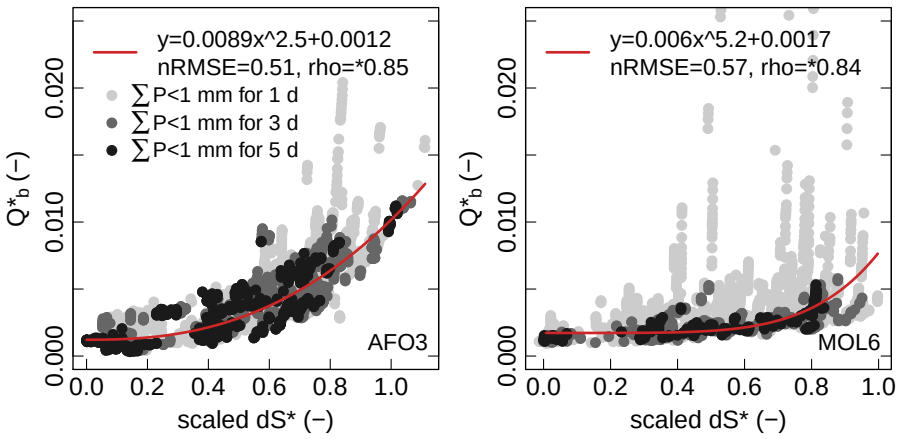


Figure 4.3: Normalized empirical storage-baseflow relationship for the catchments AFO3 (left) and MOL6 (right). dS^* and Q_b^* are calculated according to Eq. (4.1 and 4.4), respectively. The power law functions (red lines) are fitted to stream flow values where the last input in precipitation > 0.1 mm was $\geqslant$ 5d ago (statistics in the top). The quality of each relation is judged using the root-mean-square-error normalized to the standard deviation of the sample (nRMSE) and Spearman's rank coefficient of correlation (rho). Note: Fitting the model to the data required a re-scaling of dS^* to the range $[0\ \infty]$ to prevent root extraction from negative values.

thus expect the multipliers in the normalized storage discharge relationship to partly reflect the strength of the gradient driving baseflow production. In line with this we find that the median of the catchment gradients explains ($r^2 = 0.28$, $p = 0.023$) of the multipliers when leaving out two alpine sites.

4.4.2 *Storage control on rainfall-runoff response*

The total number of rainfall events within the four year period (n=14854) ranges from 334 to 859 among catchments. Omitting those influenced by snow or triggered by rainfall totals smaller than 10 mm yields 1174 rainfall events (53 events per catchment on average). The distribution of the corresponding CR_E is right skewed with a median CR_E of 0.06, a mean of 0.11 and a maximum value of 0.83 (inter-quartile range = 0.13).

Table 4.2 presents the resulting statistics of the analysis of the driven case. Significant rank correlations among the CR_E and the three storage measures dS^*, Q^* and θ^* are found in 9, 15 and 8 out of 22 cases, respectively. The corresponding average ρ values are 0.61, 0.63 and 0.52, respectively. In all cases where dS^* is significantly correlated to CR_E, Q^* is significantly correlated to CR_E as well. Basically the same applies for the comparison of Q^* and θ^*. Whenever θ^* is significantly correlated to CR_E, Q^* is significantly correlated to CR_E as well, except for the catchments ALP1 and ALP4. We may thus state that Q^* is the best predictor for CR_E with both the highest number of significant cases and the highest ρ values.

Table 4.1: Statistics of the radiation-driven case (baseflow) where the last input in precipitation > 1 mm is at least five days ago: The table contains the sample size (n), Spearmans rank coefficients of correlation (ρ) between the storage measures dS^* and θ^*, between storage measures and normalized specific stream flow depths (Q_b^*) and multiplier and exponent of the fitted non-linear model. As a quality of fit criterion for the latter we provide the root-mean-squared-error normalized by the standard deviation of the sample (nRMSE). Note: *-symbols code significant cases (p-values<0.001). As snow and frequent rainfall yielded small (n<100) and skewed samples in two alpine basins (ALP1 and ALP4) these values were set to NA.

		ρ between			non-linear model		
Site	n	dS^* & θ^*	Q_b^* & dS^*	Q_b^* & θ^*	multiplier	exponent	nRMSE
TRI1	364	*0.2	*0.67	0.01	1.8e-04	1.3	0.56
TRI2	384	-0.01	*0.41	*0.18	1.8e-07	7.2	0.89
TRI3	397	*-0.27	*0.88	-0.02	1.4e-04	1.8	0.44
JUR1	588	0.13	*0.75	-0.01	4.4e-05	2.5	0.70
BFO1	1265	0.09	*0.56	0.06	7.9e-05	2.3	0.65
BFO2	1235	0.02	*0.69	0.03	1.5e-04	1.6	0.68
BFO3	899	*0.30	*0.47	-0.01	3.7e-04	0.8	0.87
MOL1	968	*0.39	*0.84	*0.16	8.1e-03	1.1	0.61
MOL2	654	*-0.24	*0.49	0.10	8.8e-04	2.0	0.84
MOL3	447	-0.02	*0.86	-0.11	5.4e-03	1.6	0.59
MOL4	634	-0.11	*0.50	*0.26	1.7e-03	0.9	0.84
MOL5	1184	*0.18	*0.67	*0.22	2.9e-03	0.7	0.83
MOL6	213	*0.30	*0.84	*0.28	6.0e-03	5.2	0.57
MOL7	1018	*-0.28	*0.16	-0.07	1.1e-03	0.8	1.01
AFO1	250	0.16	*0.45	0.14	5.0e-03	0.3	0.95
AFO2	609	0.04	0.12	*0.54	7.8e-04	15.0	0.92
AFO3	708	*0.21	*0.85	0.07	8.9e-03	2.5	0.49
AFO4	356	0.05	*0.42	*0.22	8.8e-03	0.6	0.86
ALP1	90	NA	NA	NA	NA	NA	NA
ALP2	116	0.03	*0.66	0.19	5.5e-04	2.2	0.64
ALP3	249	0.15	*0.60	0.10	1.0e-08	6.6	0.43
ALP4	80	NA	NA	NA	NA	NA	NA
* cases	-	9	19	7	-	-	-

Furthermore, we find a rather interesting regional pattern where a distinct storage measure performs much better. Q^* is consistently not (significantly) correlated to CR_E in the four alpine sites while θ^* has significant ρ values in three of the four catchments. Consistently with this, plots of CR_E versus θ^*, thereby scaling the point size with rainfall depth, clearly corroborate the dominant influence of rainfall depth (and probably intensity) in the Alpine catchments (e.g. ALP4, Fig. 4.4, top left) (see also section 4.4.4).

A remarkable finding from the Triassic catchments is that the three catchments TRI1, TRI2 and TRI3 have the highest ρ and r^2 values between CR_E and Q^* among the entire data set, with up to 70 % of explained variance (compare Table 4.2). Relationships between CR_E

Table 4.2: Statistics of rainfall-driven conditions: Spearman rank coefficients of correlation (ρ) and Pearsons coefficient of determination (r^2) between the event runoff coeffcients (CR_E) and the three different storage measures (dS^*, θ^* and Q^*), between the storage measures themselves and, results for a multiple linear regression (equation and corresponding r^2) between CR_E and the two most explanatory (uncorrelated) storage measures. If all three storage measures were correlated significantly, both, the equation and the r^2 value were set to NA. We also provide the slope of the linear regression (b) between CR_E and Q^* for cases were the latter were correlated significantly. (*)-symbols code significant cases (p-values<0.001). Values for ALP1 and ALP4 were set to NA as snow and frequent rainfall yielded small (n<100) and highly skewed samples.

		CR_E & dS^*		CR_E & Q^*		CR_E & θ^*		dS^* & θ^*		Q^* & dS^*		Q^* & θ^*		lin reg.	multiple lin. regression	
Site	n	ρ	r^2	ρ	r^2	ρ	r^2	ρ	r^2	ρ	r^2	ρ	r^2	b	equation	r^2
TRI1	51	*0.69	*0.50	*0.81	*0.73	0.38	0.18	0.35	0.11	*0.78	*0.49	0.26	0.18	2.94	*340Q^*+0.087θ^*	0.73
TRI2	40	*0.57	*0.27	*0.82	*0.57	0.44	0.14	0.40	0.11	*0.53	0.24	0.44	*0.30	3.87	*480Q^*-0.086θ^*	0.57
TRI3	37	*0.77	*0.46	*0.88	*0.59	0.39	0.13	0.15	0.01	*0.83	*0.36	0.45	0.26	3.13	*500Q^*-0.069θ^*	0.59
JUR1	40	0.39	0.20	*0.63	*0.29	0.44	0.08	0.41	0.20	*0.63	*0.46	*0.52	*0.43	1.29	NA	NA
BFO1	67	0.30	*0.17	*0.46	*0.16	0.35	0.07	0.37	0.14	0.38	*0.27	*0.54	*0.27	1.09	91Q^*+0.034dS^*	0.22
BFO2	65	*0.51	*0.31	*0.49	*0.24	*0.41	0.09	0.38	0.14	*0.73	*0.46	*0.41	*0.26	1.22	*0.061dS^*+0.057θ^*	0.32
BFO3	52	*0.71	*0.47	*0.67	*0.46	*0.55	*0.25	0.43	0.19	*0.60	*0.43	*0.51	*0.33	1.97	*0.085dS^*+0.140θ^*	0.52
MOL1	66	*0.62	*0.23	*0.67	*0.35	0.32	0.05	0.23	0.04	*0.85	*0.63	0.29	0.03	2.68	*490Q^*+0.130θ^*	0.37
MOL2	67	0.26	0.07	*0.52	*0.19	0.18	0.03	0.11	0	*0.52	*0.24	*0.47	*0.24	1.86	NA	NA
MOL3	73	*0.49	*0.22	*0.49	*0.20	0.34	0.06	0.19	0.01	*0.73	*0.46	0.38	*0.17	1.47	*0.097dS^*+0.140θ^*	0.26
MOL4	70	0.18	0.07	*0.56	*0.20	*0.55	*0.18	0.19	0.06	*0.61	*0.32	*0.49	*0.29	1.90	NA	NA
MOL5	77	*0.55	*0.30	*0.72	*0.29	*0.56	*0.24	*0.56	*0.21	*0.74	*0.46	*0.56	*0.31	1.60	NA	NA
MOL6	40	0.42	0.20	0.49	0.24	0.30	0.08	0.25	0.04	*0.51	0.15	*0.61	*0.48	NA	NA	NA
MOL7	38	0.30	0.16	0.48	0.26	0.50	*0.28	0.15	0.01	-0.08	0	0.36	0.22	NA	0.42θ^*+300Q^*	0.37
AFO1	54	0.01	0	*0.64	*0.27	*0.57	*0.26	0.23	0.04	-0.05	0.01	*0.58	*0.53	1.65	*380Q^*-0.059dS^*	0.28
AFO2	61	0.19	0.04	*0.48	0.07	0.33	0.07	0.29	0.03	0.25	0.06	0.39	*0.24	0.75	100Q^*+0.260θ^*	0.09
AFO3	38	0.51	*0.31	0.45	0.24	0.16	0.03	0.15	0	*0.89	*0.69	0.17	0.02	NA	*0.11dS^*+0.085θ^*	0.33
AFO4	65	*0.55	*0.25	*0.58	*0.20	*0.45	0.11	0.30	0.06	*0.55	*0.30	*0.41	*0.21	1.41	NA	NA
ALP1	78	0.32	0.09	0.30	0.09	*0.38	0.12	0.34	*0.18	0.22	0.09	*0.42	*0.40	NA	0.14θ^*+0.034dS^*	0.15
ALP2	39	0.30	0.15	0.42	0.20	0.39	0.17	0.13	0.01	0.29	0.13	*0.57	*0.28	NA	31Q^*+0.04dS^*	0.26
ALP3	30	0.52	0.22	0.01	0.06	0.41	*0.34	0.30	0.13	0.24	0.14	0.41	0.19	NA	0.016dS^*+0.180θ^*	0.42
ALP4	26	0.13	0	0.23	0.15	*0.72	*0.50	0.33	0.12	0.07	0.03	0.29	0.28	NA	*0.24θ^*+1.8Q^*	0.5
* count	-	9	11	15	14	8	7	1	2	14	13	12	15	-	-	-

and θ^* are often pretty linear here, whereas that between CR_E and dS^* indicates threshold behaviour (see Fig. 4.4, top right). Also the catchments located in the Molasse area often show significant ρ values between CR_E and Q^* and the functional relationship is linear in most cases (e.g. MOL5, Fig. 4.4, bottom left). However, the relationships are clearly more noisy than in the Triassic area. The catchments located in the Bavarian forest have rather similar ρ values for both, dS^* and Q^* and the functional relationship appears linear as well (see e.g. BFO3, Fig. 4.4, bottom right).

The inter-comparison of the storage measures at the beginning of the rainfall-runoff events reveals dS^* and θ^* as being not significantly correlated, except for one catchment (MOL5) (Table 4.2). However, Q^* is significantly correlated to dS^* in 14 and to θ^* in 12 out of 22 cases, although the catchments do not coincide. The statistics also re-

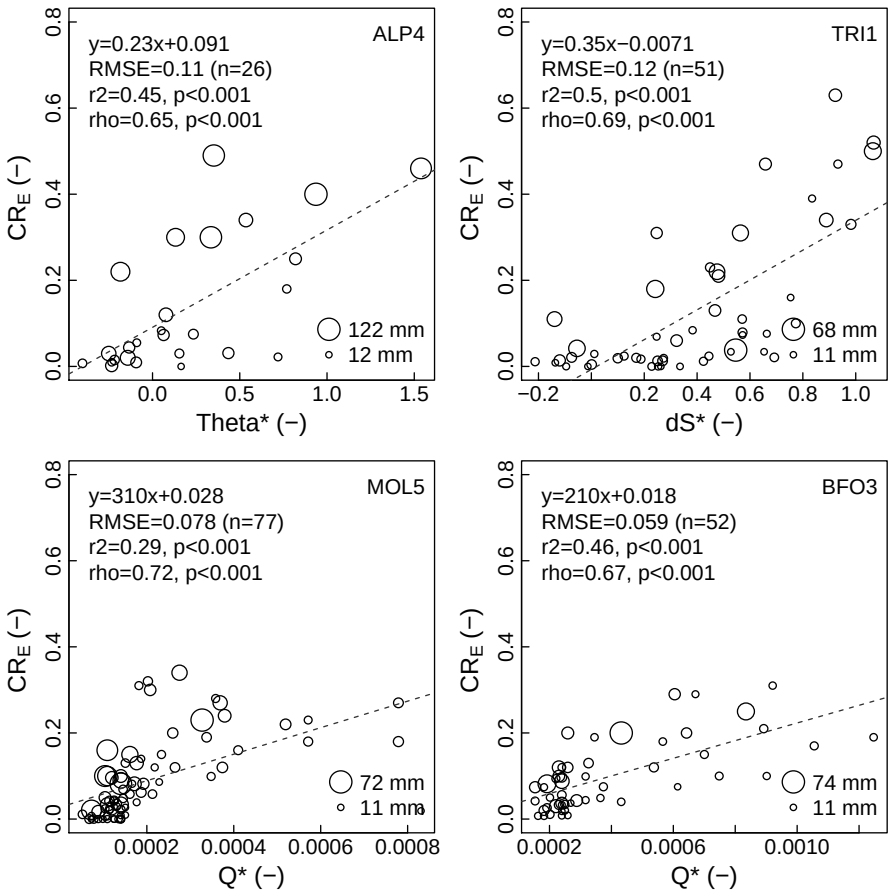


Figure 4.4: Event runoff coefficients (CR_E) plotted against the normalized storage measures θ^*, dS^* or Q^* for the sites ALP4, TRI1, MOL5 and BFO3. The point sizes are scaled according to the corresponding rain depth. Statistical information is provided in terms of a regression (dotted line), its equation, the sample size (n), the root-mean-squared-error (RMSE), Pearsons' coefficient of determination (r^2), Spearmans' rank coefficient of correlation (rho) and corresponding p-values (p).

veal that multiple linear regressions among CR_E and the two most explanatory and uncorrelated storage measures explain - at best - 5 % of additional variance (r^2) in all except three catchments (ALP3, ALP2, MOL7) compared to the univariate regressions. Hence, we conclude that the consideration of different uncorrelated storage measures does not further improve the predictability of the event runoff coefficients compared to the single best predictor.

Similar to the non-driven case we tested for significant relationships between average and median CR_E and Ks times ϕ. The resulting coefficients of determination are very small in all cases ($r^2 \leqslant 0.02$). Also the three regression slopes between CR_E and the different storage measures obtain small and insignificant coefficients of determination. However, it turned out that the median topographic gradient alone explains 31.4 % (p-value=0.0066) of the variance of the average CR_E between the catchments. We may state that gradient and resistance are conjunct during baseflow recession when the system operates close to local equilibrium conditions. During rainfall con-

ditions the gradients dominate the concert, which indicates further-from-equilibrium conditions.

4.4.3 *Seasonal interplay of storage and release*

4.4.3.1 *Normalized double mass curves*

The double mass curves are similar in all catchments in terms of a fairly linear increase in the winter period and a clear regime shift towards much flatter, partly zero slopes in the vegetation period. In fact the slopes of the nDMCs are almost constant, just parallely shifted, during the period of vegetation at many sites (e.g. at MOL2, TRI3 or BFO1, Fig. 4.5, top left, top right and bottom right, respectively). Strikingly, the onset of the vegetation period, defined by a temperature index model (Menzel et al., 2003) accurately predicts when the regime shift occurs (in terms of $cum.P/\sum P$). Moreover, temperature sums explain 70 % of the variance of the summer runoff coefficients with respect to the entire range of our physiographic setting. During winter, temperature aggregates are not significant and without predictive power (Fig. 4.5, bottom right). We may thus state that the onset of the vegetation period dominates the seasonal interplay in storage and release during the "summer" period, in all of our physiographic and climatic settings (except for the alpine region), compare Figs. in Appendix A.2.5.

The different catchments within our data set show considerable variations in the seasonal summer and winter runoff coefficients and partly also with respect to their inter-annual variation (Table 4.3). On average the seasonal winter runoff coefficients ($CR_W = 0.67$) exceed the average summer runoff coefficients ($CR_S = 0.32$) by a factor of 2, with two exceptions (ALP2 and ALP3, both alpine sites). The mean absolute deviation (mad) of the seasonal runoff coefficients are twice as large during winter ($mad_{CR_W} = 0.1$) as during summer ($mad_{CR_S} = 0.06$) (Table 4.3).

With respect to the different physiographic settings we encounter distinct seasonal and spatial patterns. During winter the highest average nDMC slopes ($CR_W = 0.8 - 0.9$) occur in the north eastern catchments (BFO1, BFO2, BFO3) which are rather densely forested, but also in the alpine ALP1 catchment. ALP4, ALP3 and ALP2, which are also alpine catchments and located on similar altitudes, show much lower winter runoff coefficients of 0.64 to 0.71 on average, probably due to storage in the snow pack. The smallest winter runoff coefficients (0.35 and 0.55) occur in MOL7 and TRI3, respectively. With respect to the inter-annual winter variance we encounter small mean absolute deviations $\leqslant 0.05$ in low lying sites of the Molasse and glacial drift areas e.g. MOL5, AFO4, AFO2, AFO1. High mean absolute deviations $\geqslant 0.15$ occur in different geologies, including the sites TRI2, BFO1, JUR1 and BFO3. Please also note that $CR_{yr} > CR_W$ in a few cases where snow exhibits a strong control on winter runoff regimes (e.g. ALP3 or ALP2, see Table 4.3). Here, fitting linear regressions to

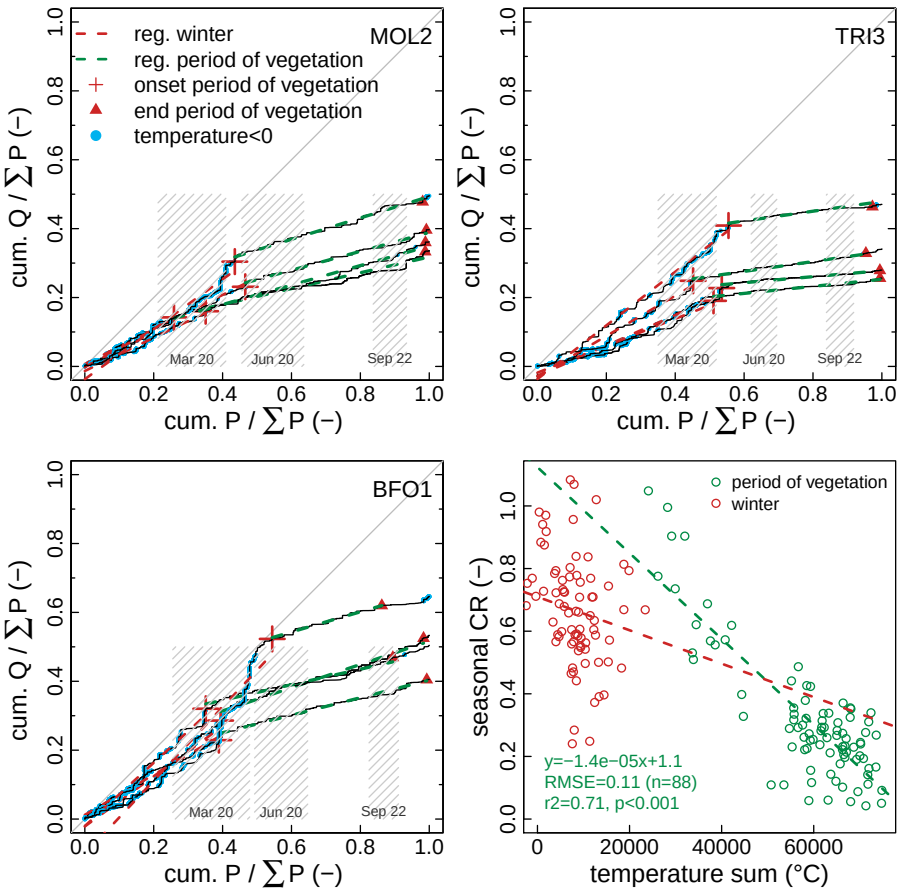


Figure 4.5: Normalized double mass curves (nDMC) for the catchments MOL2 (*top left*), TRI3 (*top right*) and BFO1 (*bottom left*) for the hydrological years 1999-2003. Onset and end of the period of vegetation are determined using a temperature index model. Regression lines are fitted to both periods (dotted lines in red/ green), their slopes are interpreted as seasonal runoff coefficients. Periods with temperatures < 0 °C are highlight in blue. Gregorian definitions for the start of spring (Mar 20th), start of summer (Jun 20th) and start of fall (Sep 22th) (hatched polygons) are added to the cum. P/ $\sum$ P plane to highlight their differences to temperature based estimates on the onset and end of the period of vegetation. Statistical properties of all nDMCs are summarized in Table 4.3. The *bottom right panel* shows seasonal hourly temperature sums (calculated for each hydrological year starting from Nov 1st) and corresponding seasonal runoff coefficients for all sites (n=22) and years (n=4). The dotted lines are regressions. Statistical information on the summer model is plotted in green. During winter there was no significant statistical relation available (r^2=0.04, p=0.062).

the double mass curves is not suitable for estimating seasonal winter runoff coefficients.

According to the statistical analyses in which we regressed 24 different variables against the slopes of the seasonal winter runoff coefficients, the most explanatory variables are sand content ($r^2 = 0.29$), median gradient (ϕ) times Ks ($r^2 = 0.22$), silt content ($r^2 = 0.22$), forest coverage ($r^2 = 0.16$), skeleton content ($r^2 = 0.15$), number of

Table 4.3: Mean seasonal winter (CR_W), summer (CR_S) and annual runoff coefficients (CR_{yr}) as indicated by the slope of regression lines fitted to the normalized double mass curves. CE_{yr} represents the mean annual evapotranspiration ratio. The inter-annual variations of these quantities within the hydrological years ('99-'03) is quantified using the mean absolute deviation which we provide by mad_{CR_W}, mad_{CR_S}, $mad_{CR_{YR}}$ and $mad_{CE_{YR}}$, respectively. All quantities are dimensionless.

Site	CR_W	CR_S	CR_{yr}	CE_{yr}	mad_{CR_W}	mad_{CR_S}	$mad_{CR_{YR}}$	mad_{CE_W}
TRI1	0.72	0.12	0.43	0.53	0.058	0.030	0.031	0.018
TRI2	0.70	0.07	0.37	0.68	0.198	0.027	0.105	0.021
TRI3	0.55	0.12	0.34	0.63	0.133	0.024	0.069	0.022
JUR1	0.73	0.25	0.48	0.56	0.150	0.017	0.054	0.015
BFO1	0.82	0.28	0.52	0.48	0.169	0.037	0.068	0.033
BFO2	0.85	0.30	0.54	0.47	0.143	0.020	0.067	0.029
BFO3	0.93	0.29	0.57	0.51	0.173	0.034	0.089	0.024
MOL1	0.60	0.24	0.38	0.62	0.103	0.040	0.021	0.023
MOL2	0.56	0.27	0.40	0.58	0.080	0.022	0.049	0.042
MOL3	0.62	0.34	0.46	0.57	0.069	0.019	0.045	0.035
MOL4	0.56	0.23	0.37	0.61	0.084	0.021	0.051	0.048
MOL5	0.69	0.22	0.41	0.58	0.055	0.028	0.026	0.023
MOL6	0.60	0.20	0.36	0.66	0.066	0.018	0.025	0.015
MOL7	0.35	0.27	0.31	0.68	0.139	0.089	0.031	0.052
AFO1	0.72	0.35	0.51	0.46	0.031	0.120	0.042	0.053
AFO2	0.68	0.34	0.49	0.48	0.036	0.099	0.043	0.052
AFO3	0.56	0.24	0.38	0.62	0.130	0.068	0.031	0.056
AFO4	0.66	0.22	0.39	0.64	0.050	0.090	0.030	0.092
ALP1	0.89	0.50	0.75	0.24	0.098	0.088	0.031	0.031
ALP2	0.71	0.82	0.83	0.17	0.067	0.161	0.047	0.015
ALP3	0.64	0.84	0.86	0.18	0.082	0.109	0.025	0.017
ALP4	0.66	0.53	0.64	0.33	0.064	0.066	0.032	0.051
mean	0.67	0.32	0.49	0.51	0.10	0.06	0.046	0.035

frost days ($r^2 = 0.14$), effective field capacity ($r^2 = 0.13$) and absolute sum of negative temperatures ($r^2 = 0.12$). All other variables have coefficients of determination $r^2 \leqslant 0.10$. In several multiple linear regressions based on the above mentioned variables the best result is achieved for a combination of ϕ times Ks, forest cover and absolute sum of negative temperatures (multiple $r^2 = 0.30$, p-value<0.001). Active storage estimates (dS), summer temperature sums and length or end of the period of vegetation from the previous hydrological year do not help to improve the prediction of the actual CR_W. The key finding in this analysis is that ϕ times Ks yields a $r^2 = 0.22$, whereas two variables alone only explain 0.02 and 0.08 % of the variance in the CR_W, respectively. This corroborates that surrogates for gradients and resistances act jointly and that their impact is detectable even at the lower mesocale.

The summer season is characterised by an opposite spatial pattern compared to the seasonal winter runoff coefficients. The highest sea-

sonal $CR_S \geqslant 0.8$ is found in the snow-dominated alpine catchments of ALP3 and ALP2. The smallest CR_S with values between 0.07 and 0.12 are encountered at the Triassic sites (TRI3, TRI2, TRI1). It is also important to note here that several low-lying sites the CR_S shows very little inter-annual variance as indicated by mean absolute deviations $\leqslant 0.03$ (e.g. MOL5, TRI3, TRI2, MOL6, MOL2, MOL4, JUR1, MOL3 and others) (Table 4.3). In these catchments the slopes of the nDMCs are fairly constant throughout different hydrological years indicating a very strong control of evapotransiration on the water balance during summer. At these sites the curves of the nDMCs in summer have nearly identical slopes and are simply shifted in parallel depending on the onset of vegetation activity.

We may hence state that normalized double mass analyses are powerful tools for discriminating seasonal differences in the interplay of storage and release among mesoscale catchments. However, they do not provide insights into the reasons for inter-annual variations. In several cases we observed inter-annual variations in $CR_{yr} \geqslant 0.1$, which could stem from variations of P or ET or a carry over of water storage into the next year. To provide more insights we introduce normalized triple mass curves by adding $cum.E/\sum P$ as a third dimension.

4.4.3.2 *Normalized triple masse curves*

Conceptually we usually assume that the change in storage tends to zero within a single hydrological year. Hence, we assume that large inter-annual variations in the rainfall-runoff ratio CR_{YR} coincide with large inter-annual variations in the evapotranspiration ratio (CE_{YR}). To evaluate this assumption on our data set we construct normalized triple mass curves and calculate the mean absolute deviation for both, CR_{YR} and CE_{YR}. Within our sample we find several catchments where the mean absolute deviation in the evapotranspiration ratio ($mad_{CE_{yr}}$) is rather similar to the mean absolute deviation in the annual runoff coefficients ($mad_{CR_{yr}}$) e.g. MOL5, MOL2, ALP1, MOL4, or MOL1 (see examples in Fig. 4.6, upper row). However, we also find several sites where $mad_{CR_{yr}}$ clearly exceeded $mad_{CE_{yr}}$ e.g. TRI3, TRI2, BFO1, JUR1, BFO3 or BFO2 (compare Fig. 4.6, lower row). This may be attributed to a carry over of water storage feeding runoff formation (blue water) between the hydrological years, indicating inter-annual memory (under the assumptions of a closed control volume). Only in two catchments (AFO4 and MOL7) $mad_{CR_{yr}}$ is substantially smaller than the corresponding $mad_{CE_{yr}}$. This can be explained by a carry over of water into neighbouring years, feeding ET (green water).

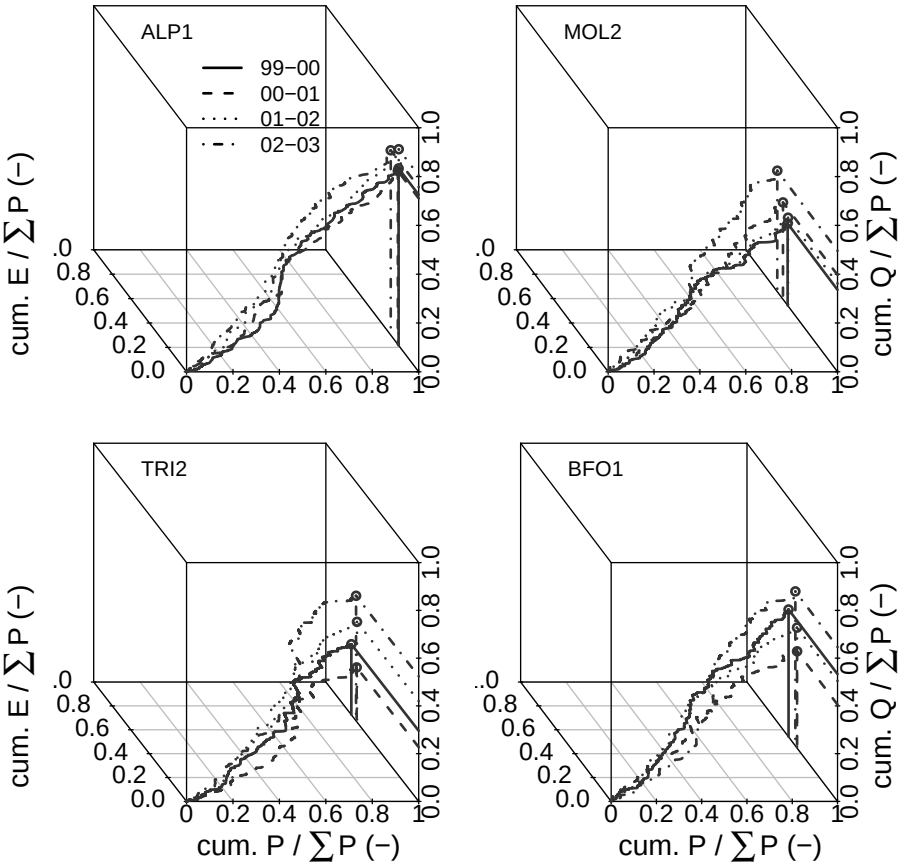


Figure 4.6: Normalized triple mass curves (nTMCs) from the catchments ALP1, MOL2, TRI2 and BFO1. Each plot contains data from four different hydrological years ('99-'00, '00-'01, '01-'02 and '02-'03) which are coded using different line styles. In the *upper row* (sites ALP1 and MOL2) the inter-annual variations in cum. $Q/\sum P$ are rather identical to the inter-annual variations in cum. $E/\sum P$. At TRI2 and BFO1 (*lower row*), the inter-annual variations in cum. $Q/\sum P$ are much larger than the inter-annual variations in cum. $E/\sum P$. Corresponding statistics of the normalized double and triple mass curves are summarized in Table 4.3.

4.4.4 *Intensity controlled runoff formation*

4.4.4.1 *Data evidence in Alpine catchments*

Strikingly, we also find signatures of intensity controlled runoff in two Alpine catchments (ALP1 and ALP2). This is illustrated in Fig. (4.7) which compares two flood events from site ALP2 caused by rather similar totals of rainfall (244 and 200 mm) and identical event runoff coefficients ($CR_E = 0.58$). The rainfall intensities as well as the discharge peaks in the right panel are however twice as large compared to the left panel, yielding considerable differences in the normalized temporal intensity changes ($I_E^* = 0.08$ vs. $I_E^* = 0.32$). Similar rainfall-runoff dynamics with strong temporal changes in P which are followed by strong increases in Q are observed during many events at site ALP1.

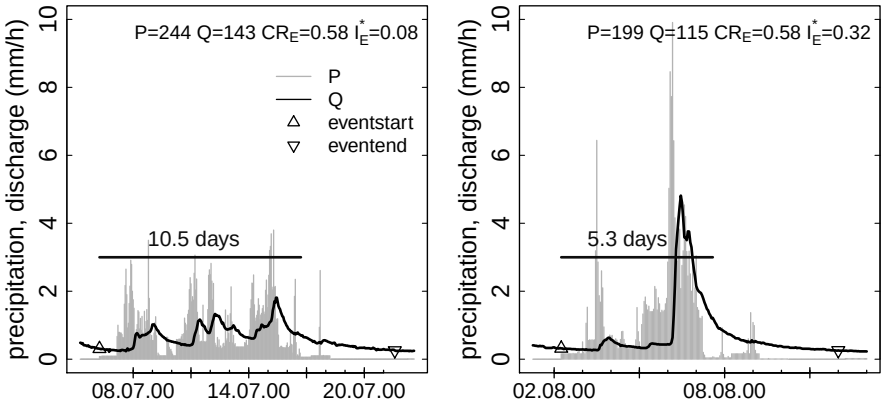


Figure 4.7: Two storm events from the alpine catchment ALP2 with almost similar total amounts of precipitation (P) and discharge (Q), identical runoff coefficient (CR_E), but with different duration and thus, intensities. The latter is reflected in different I_E^* (Eq. 4.7) which corresponds to the normalized maximum temporal change in intensity. In the first case (*left panel*) we expect that capacity controlled processes to dominate the runoff generation. In the second case (*right panel*) we assume that the steep rising limb and the high peak discharge are caused by intensity controlled runoff formation processes.

In these catchments the highest normalized temporal intensity changes indeed cluster at small normalized event durations (Fig. 4.8, left panel). The same scatterplot for MOL2 (Fig. 4.8, right panel) reveals normalized temporal intensity changes to spread equally across all event durations. The three-dimensional scatterplots of normalized temporal runoff changes against total precipitation and maximum intensity (Fig. 4.8, lower row), reveal clearly that large runoff changes coincide partly with high intensities and small rainfall totals. We may, hence, state the proposed signatures are feasible for detecting high frequency runoff even within low frequency data sets in mesoscale catchments. To illustrate that this is of more than academic importance we present a comparative model exercise.

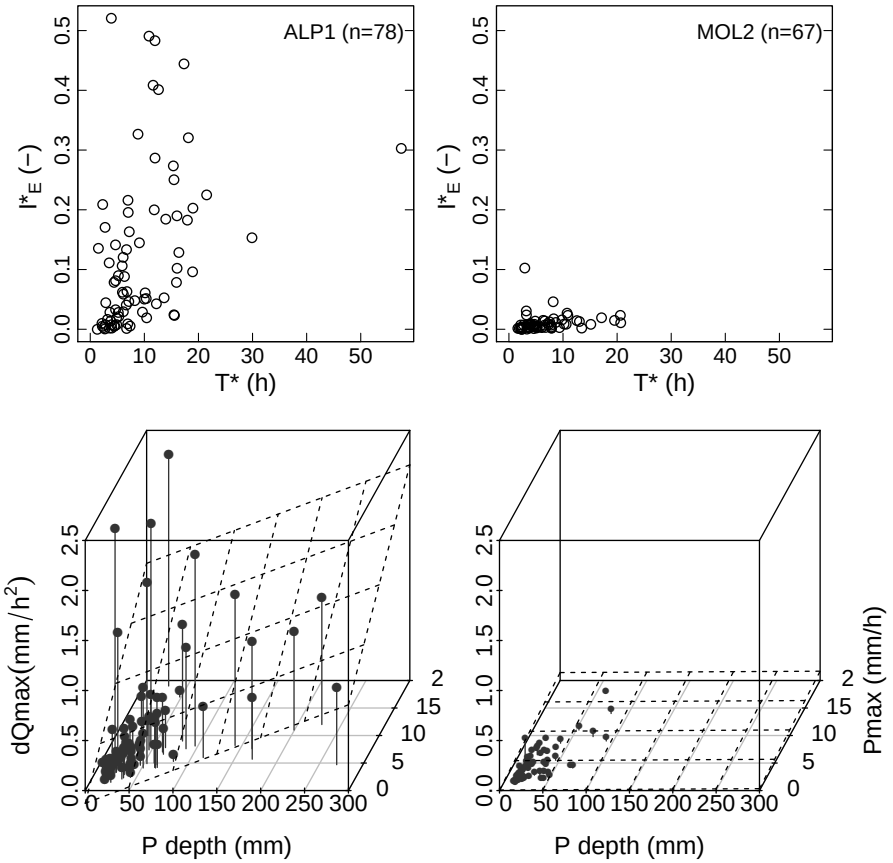


Figure 4.8: Diagnostics for the detection of intensity controlled conditions. The panels in the *upper row* show scatterplots of the normalized event duration T_E^* (Eq. 4.6) which is plotted against the normalized temporal intensity changes (I_E^*) (Eq. 4.7). The *lower row* shows corresponding three-dimensional scatterplots with event rain depth, maximum observed rain intensity and the maximum of the temporal derivative of observed discharge $(dI_{Q,max})$ are plotted on the x, y and z-axis respectively. The inclined plane (dotted) represents the plane of a multiple linear regression. The left column shows a conclusive case (site ALP1) where the frequent occurrence of intensity controlled runoff formation processes is likely. The right column shows inconclusive results from the site MOL2.

4.4.4.2 *Explorative modelling for intensity limited runoff formation*

We compare two different model concepts to further elaborate the feasibility of our diagnostics for detecting intensity control. Our comparison shall particularly highlight the errors we might expect when simulating intensity controlled runoff formation with models relying on capacity controlled runoff formation with respect to the events depicted in Fig. 4.7. Specifically, we compare the HBV beta store (Bergstroem, 1976) with a Green and Ampt approach (G&A) using the solution of Peschke, (1985), as typical concepts for capacity and intensity controlled runoff formation. Both runoff generation concepts are implemented in R (R Core Team, 2015) and combined with a simple linear reservoir, whereas surface runoff is allowed to bypass the

latter in the case of G&A. Both implementations are then fitted to observed stream flow data in an event based mode. Here we optimize the maximum storage depth ($SMax$), beta parameter (β) and the reservoir constant (k_{res}) in the case of HBV using a simulated annealing algorithm in combination with the root-mean-squared-error as objective function. The G&A approach is parametrized based upon a Rosetta Schaap, Leij, and Genuchten, 2001 estimate of Ks and a literature value for the suction head (psi) at the wetting front (Maidment, 1993). The parameters of the linear reservoir ($SMax$ and k_{res}) are adopted from the HBV optimization to ensure identical conditions.

During both events depicted in Figure 4.9 the HBV type setup outperforms G&A, when being judged on the Nash-Sutcliffe-Efficiency (NASH) criterion. During intensity controlled conditions the HBV bucket concept however clearly fails to reproduce the high runoff frequencies in terms of the slope of the rising limb and in the peak discharge (compare black box in Fig. 4.9, right panel). A closer look at the right panel reveals, however, that G&A matches the magnitude of peak discharge (which is important for flood warning) much better than the beta store model. The slightly worse NASH value is because the timing error in peak occurrence is punished, which is a well known deficiency of the NASH statistical analysis (Seibert, Ehret, and Zehe, 2016).

This exercise suggests that we are much better off in capturing sharp peaks of high frequency, intensity controlled runoff formation processes when using model concepts that are sensitive to the controlling, intensive properties compared to model concepts which do not account for this issue. Though this is essentially not new, data-driven diagnostics which assist in deciding whether intensity controlled processes need to be considered in hydrological modelling are novel and rarely available.

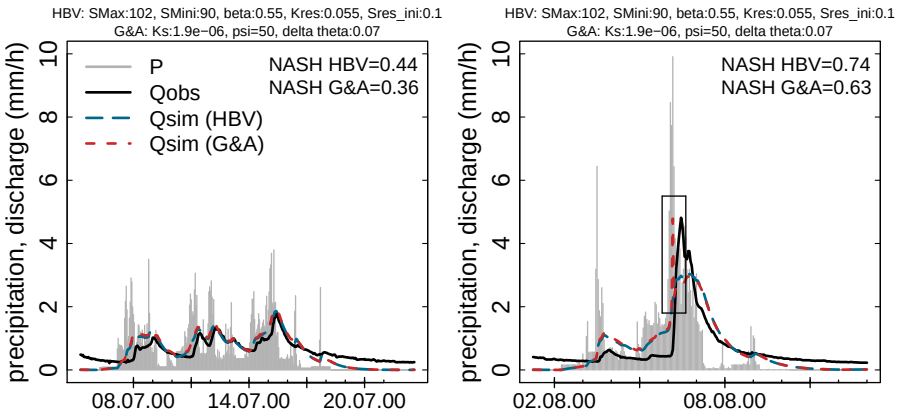


Figure 4.9: Simulated storm hydrographs from a capacity controlled event (*left*) and a (partly) intensity controlled event (*right*). Simulations use a HBV type betastore (red line) and a Green and Ampt (G&A) approach (blue line). Both concepts are combined with a linear reservoir. For the HBV type we optimize maximum storage depth (SMax), beta parameter (beta) and reservoir constant (kres) in an event based mode using a simulated annealing algorithm (initial fillings of the betastore SMini and of the linear reservoir Sres, ini are insensitive to the peak flow simulation accuracy). G&A is parametrized based upon a Rosetta (Schaap, Leij, and Genuchten, 2001) estimate of Ks and a literature value for the suction head (psi) at the wetting front (Maidment, 1993). The parameters of the linear reservoir (SMax and kres) are adopted from the HBV optimization. As (statistical) reference for the model performance we provide the Nash-Sutcliffe-Efficiency (NASH) criterion. Statistics of the rainfall-runoff events are provided in Fig. 4.7.

4.5 DISCUSSION AND CONCLUSIONS

In this study we propose various dimensionless diagnostics to characterize differences in terrestrial runoff production of catchments at the seasonal and the event scale. Particular emphasis is on a) their suitable normalization and b) on the question whether low-passed rainfall-runoff data from mesoscale catchments still bear detectable signals of high frequency, intensity controlled runoff production. As benchmark we use operational rainfall-runoff data from 22 catchments spread across a wide range of physiographic and climate conditions in the Bavarian part of the Danube catchment.

4.5.1 *Normalized double mass curves discriminating seasonal runoff behaviour*

Normalized double mass curves turn out to be an easy-to-compute, yet very powerful means to detect similarity and differences in the seasonal water balance. In our case their general shape is (invariantly) characterized by a linear increase in the winter period and a regime shift with small near to zero slopes when vegetation starts to control the water balance. The onset and duration of this regime shift could be predicted very well by a simple temperature index model (Menzel et al., 2003). In line with this, temperature explains 70 % of the variability of the summer runoff coefficients within the 22 catchments. It is noteworthy that the usual (Gregorian) definition of spring and fall onset are of little help to predict the regime shift here. We hence conclude that vegetation exerts first-order control on stream flow generation in "summer", while onset and end of the summer (i.e. the vegetation period) is defined by temperature conditions rather than simply by the Gregorian day. This finding is important as it suggests that phenological data (and corresponding surrogates) provide valuable information which is mostly not included in standard hydrological data (or at least hardly considered). We further conclude that any assessment of the "pure abiotic controls" of the catchment water balance should be restricted to (snow free) periods of the dormant season.

The variability of winter runoff coefficients is generally much less predictable by the available structural and climatic descriptors. Also the different storage estimators are of little use. The most interesting finding is that the rather coarse estimates of the catchment soil hydraulic conductivity and the median gradient operate indeed as a group and that their impact is even detectable at lower mesoscale sites: their product explains 22 % of the variance in the winter slopes, while either of the values itself is an insignificant predictor. Expressing runoff by the product of an effective gradient and a control volume conductance, requires the system to operate close to local thermodynamic equilibrium conditions. We hence conclude that this is at least partly the case at the seasonal scale.

Also normalization of the double mass curve is straight forward. Here we use annual precipitation totals of the respective hydrological year. The advantage is that both axes are normalized to one. The disadvantage is that the same amount of relative accumulated rainfall does not correspond to the same total rainfall amount. This would require us to normalize precipitation and discharge with the long term mean annual precipitation at the prize, that the maximum ordinates would then not be constrained to one.

We also provide evidence that normalized triple curves are well suited for explaining inter-annual variability of annual runoff coefficients either by different accumulated evaporation totals or by carry over in storage. The drawback of this signature is that it needs a calibrated water balance model for its calculation, due to the required estimates of E, while the double mass curve relies exclusively on standard observables.

4.5.2 *Pre-event discharge as best predictor for capacity controlled runoff production*

At the event scale we compare plots of event runoff coefficients against the three partly independent storage estimators i) normalized dynamic storage dS^*, ii) accumulated normalized discharge Q^* and iii) the normalized difference between antecedent precipitation and evaporation, θ^*. Generally Q^*, though it is the less sophisticated measure, does clearly outperform the other storage measures in terms of the explained variances and with respect to the number of catchments with significant rank correlations. Yet the comparison of different storage measures reveals regionally specific dominances. Particularly for the alpine catchments, Q^* is insignificant while antecedent wetness explains most of the variance within three out of four alpine catchments. This is in line with our perception that these basins are composed of shallow soils and we thus expect a higher importance of the near surface storage.

Another interesting finding is that the median topographic gradient explains 31 % of the variability of the mean catchment event runoff coefficient averaged over all 22 catchments. In fact we find the same result for the 90 % quantiles of the runoff coefficients and when using the square root of the topographic gradient. We hence conclude that this correlation reflects fairly well the strong dependence of rainfall totals and intensity on elevation (and the gradient) rather than the influence of the topographic gradient as a force for driving runoff concentration during rainfall driven conditions.

In comparison to Q^* and θ^* and with regard to the effort of its derivation, normalized dynamic storage provides little additional value. Thereupon, dS^* is significantly correlated to Q^* in 14 out of 22 catchments, although it is to be expected that parts of these correlations are spurious (Kenney, 1982; Pearson, 1987), as both, Q^* and dS^* are calculated upon the same variable. Similarly, we argue that correlations between dS^* and response measures such as event runoff coefficients

are more difficult to evaluate as they may also involve (non-trivial) spurious fractions because both variables are again calculated based on discharge and precipitation. Partly, this applies for Q^* and θ^* as well, but certainly to a lower degree. We hence conclude that dS^* is of limited use for explaining rainfall driven runoff formation compared to the well known pre-event discharge and the well known antecedent precipitation. We furthermore conclude that these different storage measures characterise storage in different depths, and that event runoff production is controlled by different storage compartments in different regions.

Our results corroborate that the event runoff coefficient is a useful and easy to calculate normalized response measure to discriminate capacity controlled runoff formation. However, it fails to discriminate high frequency runoff production processes as underlined by our findings in the alpine catchments ALP1 and ALP2. We also conclude that normalization of different storage estimators is generally helpful to compare their sensitivity ranges, even when relying on such simple estimates as the root zone storage depth or the pore volume therein, as done here.

4.5.3 *Heterogeneous performance of storage-baseflow relations*

The occurrence of 19 (out of 22) significant relations between dS^* and Q_b^* confirms that dS^* is a meaningful storage measure for the prediction of low flow conditions under a range of different (humid) physiographic settings. We also found specific storage-baseflow relationships by fitting power laws, with their exponents and factors being sensitive to changes in the physiographic setting. This finding is in line with the results of Shaw and Riha, (2012) who derived storage-discharge relationships for several catchments of up to 6400 km^2 in size using an adaptation of the method proposed by Kirchner, (2009).

A closer look reveals however that the estimated storage-baseflow relations are only partly of convincing quality. In many places they are rather noisy despite the fact that the nRMSE suggests predictive power. This corroborates that visual inspection of such relations is indispensable before using them for instance to parametrize regional baseflow production in a model. Parts of the noise in the relations most likely arise from the inherent data uncertainty. However, normalization of dS^* by the root zone storage volume and of Q_b by K_s is also error prone. K_s is at best a surrogate for the aquifer conductance and thickness. The soil map (BGR, 1995) suggests that e.g. BFO1 and TRI1 have identical K_s values (according to Rosetta estimates (Schaap, Leij, and Genuchten, 2001)). However, the hydrogeological map (Duscher et al., 2015) reveals that the aquifer underlying BFO1 is composed of virtually all non-aquiferous fissured rock, whereas the subsurface of TRI1 hosts low to moderately productive pore aquifers (compare Table A.3). The alpine sites show similar contradictions as the smallest K_s values coincide with the highest productive aquifer types (still Table A.3). A more meaningful structural normalization of

Q_b^* would thus require estimates of aquifer transmissivity. We hence conclude that the identification of catchments with similar baseflow production was not feasible with the proposed approach.

Within our attempts to explain differences in the power law multiplier based on catchment characteristics, we derive ϕ based upon surface topography and not, as ideally required, upon bedrock topography. Nevertheless, we find that a considerable portion of the variability is explained by the topographic gradient, in line with the flux–gradient–conductance relation. Given the large number of significant relationships between Q_b^* and dS^* we conclude that dS^* is a feasible predictor for baseflow production within a rather wide range of physiographic settings. This supports our initial assumption that dS^* characterises deep storage.

4.5.4 *"Edge filtering" of low passed data to detect high frequent runoff processes*

The proposed intensity signature detects evidence for high frequency intensity controlled runoff generation in two alpine catchments within the available low frequency data sets. The key is to "edge-filter" both rainfall and discharge data by taking their temporal derivatives and then to normalize the maximum runoff change by the maximum in precipitation change. This response measure separates the available rainfall-runoff events into subsamples with high normalized temporal intensity changes clustering at small event durations. We conclude that the approach we present is an easy-to-apply technique to test for the occurrence of intensity controlled runoff generation processes. This is also relevant for hydrological modelling as we show that a wrong conceptualization of intensity controlled runoff by a capacity controlled model approach might imply that the model misses the flood peak (even though it has a good NASH statistic). Thus, data-driven signatures on high frequency intensity controlled runoff generation can assist in the conceptualization of hydrological models or serve as structural benchmarks. We hence conclude that high frequency runoff production might play a much more prominent role in lower mesocale flood production than we usually conclude from analysing hourly (or even lower) resolution data sets. Therefore, we recommend that operational data should be recorded and stored with at least 5 min resolution, as this might reveal high frequency processes operating even more frequently than we expect.

4.5.5 *Conclusion and Outlook*

Overall, we recommend the following signatures as suitable to discriminate differences in terrestrial runoff production in lower-mesoscale catchments:

- Normalized double mass curves for the seasonal water balance,

- the event runoff coefficient in relation to pre-event discharge for capacity controlled runoff formation,
- the event duration in combination with the normalized intensity change for detecting high frequency processes.

The onset and termination of the vegetation period are useful to explain differences in the summer water balance. We also argue that gradients and conductances - and hence their underlying controls - are not independent if one attempts to explain functional differences by differences in the physiographic and climate setting. However, despite the good explanation we found for differences in the summer water balance, we were not able to robustly link functional similarity to structural similarity, based upon the properties available within our (operational) data set.

This brings us to our last conclusion which is founded on the dilemma of quantity vs. quality. On the one hand, inter-comparison studies require large sample sizes to include a sufficient number of end-members and to avoid type I errors (false hits). This makes widely available operational data sets indispensable (at least for the moment). On the other hand, we need accurate and sufficiently resolved data beyond rainfall and runoff to avoid type II errors (false negatives). Such data are (at least for the moment) only included in "research" data sets which are fairly limited in number and spatial distribution. To increase confidence in the proposed signatures we suggest they be applied to i) to a larger number of catchments and ii) to a (nested) set of small and densely instrumented catchments with homogeneous geological setup.

Part V

SYNTHESIS AND DISCUSSION

In this part I summarize the key findings I obtained in the individual studies with respect to the guiding research questions I posed in the introduction. I further discusses the different topics focusing on the relations between them and finish with a brief synthesis.

5

PART 5: SYNTHESIS AND DISCUSSION

5.1 SUMMARY

This chapter briefly summarizes the major achievements with respect to guiding questions posed in the chapters 1.2, 1.3 and 1.4 in the introduction, respectively:

Which reservoirs do have a regional impact on flood mitigation (Q 1.1) and are regional operation strategies conformable with local flood mitigation (Q 1.2)? The impact of a reservoir on a flood wave is the higher, the larger its retention volume is in relation to the volume of the flood wave. The propagation of flood waves through domains of several ten-thousands of square kilometers is however a complex process as the flood wave is non-linearly altered through tributaries, activation of floodplains, or the operation of water power plants and flood protection reservoirs, among others.

For the Bavarian part of the Danube basin we found that only one out of nine reservoirs, the Forggensee which is located in the upper reaches of the river Lech, had a significant regional impact on the Danube during high flow conditions. Re-calculations of three historic flood events suggest that the regional impact of this reservoir is significant (partly >50 cm at spots located more than 250 km downstream of the reservoir outlet, Fig. 2.12). Even more important the results show that local (*traditional*) reservoir management practices are not necessarily in conflict with reservoir operation for *regional* flood mitigation. In our case the two reservoir operation strategies exposed an offset in timing of 20-40 h (Fig. 2.10). These findings suggest that coordinated and/or regional operation of reservoirs in large-scale river basins has great potential to improve flood mitigation.

Is large-domain hydrological modelling accurate enough to allow for a regional operation of reservoirs? (Q 1.3)? Recalculating three historic flood events revealed a heterogeneous picture of model accuracy throughout the Bavarian part of the Danube basin. Expressed in grades and aggregated over all gauges ($n \approx 90$), flood events and the different statistics, model performance was rated "satisfactory" at around 20 % of all cases. The remaining gauges were classified either better ("good" or "very good"), or worse ("sufficient" or "insufficient") with nearly identical portions of fairly 40 %. With respect to the individual statistics we found the best Nash-Efficiencies (> 0.75) at the highest observed stream flow rates suggesting an increase in model performance with flow rate. The latter is most likely due to the calibration of the operational models on the one-in-one-hundred year flood. Peak timing errors were slightly biased negatively and more precise at higher flow ates than for lower streamflow values. In absolute terms we observed average timing errors $\geqslant \pm 12$ h in more

than one-fifth of the evaluated gauges and distinct spatial differences across the investigation area (compare Fig. 2.9 top right panel and, Table 2.5). Although we found good and satisfactory results at many of the southern tributaries in terms of the NASH criterion, all model runs overestimated the total volume by up to 10 % which can be crucial for reservoir operation. This points out that volume errors are not captured well by the Nash-Efficiency. The accuracy in timing was heterogeneous and should be improved in several reaches. This applies in particular for the (central) Danube where the simulated peaks were (almost consistently) too late as depicted in Fig. 2.4. Here, the performance in peak timing was rated "satisfactory" or worse in eight out of ten gauges (compare Fig. 2.9, top right), implying absolute Peak-Time-Differences between the simulated and observed time series of at least 6-9 h. Comparing this to the width of the window in time available for the regional operation of the Forggensee reservoir, which is 20-30 h (Fig. 2.10) makes clear that timing uncertainties are large and that accurate information on the timing and associated uncertainties is of particular relevance for reservoir operation. This is particularly obvious for the central Danube where large (timing) errors (Fig. 2.4) coincide with the highest potential water level reductions due to regional reservoir operation (Fig. 2.12, left panel).

The poorest model performance among all considered gauges was observed in the northern sub-catchments of the Danube. We attributed this to a poor model parameterization, i.e. the estimation of event runoff coefficients (compare chapter 2.3.1.2) and not to physiographic properties of the basin. The specific individual roles of the model parameterization, that of the data and that of the physiographic properties, remain however unknown and open for future research.

To what extent does the coupling of hydrological and 2d-hydrodynamic models improve the simulation accuracy (Q 1.4)? Contrary to our expectations the substitution of hydrological (HY) flood routing schemes by 2d-hydrodynamic (HD) models did not significantly increase the simulation accuracy in the major reaches. The resulting performance of the coupled model was identical, partly even worse, compared to the results achieved by the hydrological model alone. The reason was that the lateral inflow data which were provided by the hydrological model was of too poor quality due to the overestimation in volume mentioned above. Additional analyses from Skublics, Seibert, and Ehret, (2014) showed that HD models can suffer disproportionally from low quality boundary conditions, compared to hydrological routing schemes. It even turned out that inaccurate boundary conditions can cause errors that do not cancel out but amplify. The potential of hydrodynamic models hence stands and falls with the accuracy of the boundary conditions provided. Inaccuracies of simplified hydrological routing schemes may not necessarily be compensated if they are substituted by a HD model. In any case a joint calibration of the different methods is suggested. Effects of (rapid) mobilization of groundwater into the channel due to groundwater riding (Cloke et al., 2006) were not relevant in this context.

How to emulate (human reasoning in) visual hydrograph inspection (Q 2.1)? In chapter 3 I present an advancement of the Series Distance (SD) approach for an improved discrimination and visualization of timing and magnitude uncertainties in streamflow simulations. The method emulates visual hydrograph comparison by distinguishing periods of low-flow and periods of rise and recession in hydrological events. Within these periods, it calculates the distance of two hydrographs not between points of equal time, but between points that are considered hydrologically comparable. The latter refers to a chronological comparison of points belonging either both to two corresponding rising or falling segments of the hydrograph. The identification of suitable points is however not trivial and requires a pattern matching procedure. In visual hydrograph evaluation, a hydrologist detects the dominant patterns of rise and fall in the two time series and identifies matching segments by doing two things: Filtering out short, non-relevant fluctuations and then relating the remaining by jointly evaluating their similarity in timing, duration and slope. The stronger the overall disagreement of the observed and simulated event, the more visual *coarse-graining* will be done before the hydrographs are finally compared, while at the same time the degree of *coarse-graining* will also influence the hydrologist's evaluation of the hydrograph agreement: The higher the required degree of *coarse-graining*, the smaller the agreement. One of the key advancements I present in this context is the emulation of this process by an automated *coarse-graining* procedure that determines the optimal level of generalization when comparing two hydrographs. Technically this is solved by iteratively maximizing an objective function.

What is the role of timing uncertainties in hydrological streamflow simulations (Q 2.2), how to consider them in the construction of uncertainty envelopes and what is their impact on the region of confidence (Q 2.3)? I assessed the role of timing errors in event-based simulations considering about 90 gauges from the Danube catchment (reservoir study) and based upon continuous simulations using data from the gauge *Hoher Steg* which is located in Vorarlberg, Austria (Series Distance case study). Key findings obtained in the reservoir study were:

1. The horizontal component of the error is not negligible in absolute terms whilst highly relevant for the operation of flood protection reservoirs (see left panel in Fig. 2.4 and Fig. 2.10).

2. Simple error statistics like the Peak-Time-Difference (PTDF) are a step ahead in the quantification of timing errors but have several limitations: These include the fact that the PTDF statistic is fairly unspecific to timing errors in either the rising or falling limb of the hydrograph as only the peaks of the two time series are compared. For the same reason the PTDF statistic is error-prone as the comparison rests on two single (here hourly) observations only, whereas the nature of rainfall-runoff events may be complex (e.g. multi-peak events or flash floods) with durations varying between periods of time < 1 h and several days.

> Last, the significance of the PTDF criterion is also reduced in catchments with a dampened flood response. This applies e.g. for streams which are fed by lakes or for groundwater dominated basins. Here the rising and falling limbs can be flat so that the crest of the flood wave is not well defined.

The most important results from the Series Distance evaluation are that the proposed method offers elaborated techniques for the comparison of simulated and observed stream flow time series. These can be used for detailed hydrological analysis, model diagnostics and to inform about uncertainties related to hydrological predictions.

Applied to the case study SD reveals that different flow conditions (low flow, rising and falling limbs during events) exhibit distinctly different time-magnitude error characteristics with respect to mean and spread (see Fig. 3.6). Further, we find a remarkable timing uncertainty in a continuous simulation from the gauge Hoher Steg (see Fig. 3.8), though the simulation is considered fairly accurate when judged upon the NASH criterion (NASH=0.78). This suggests that the role of timing uncertainties is underestimated.

Based upon the SD results it was also possible to consider timing errors in the construction of uncertainty envelopes. I used 2-dimensional error dressing for this purpose. The results show that the horizontal error component *inflates* the region of confidence compared to traditional approaches which only consider the magnitude component of the error. This suggests that the combined use of time and magnitude errors to construct uncertainty envelopes implies a trade-off between the added value of explicitly considering timing errors and the associated, inevitable time-spreading effect in the related uncertainty ranges. Which effect dominates depends on the characteristics of timing errors in the hydrographs at hand.

Are dimensionless state-response and forcing-response diagrams feasible to characterize and to detect differences in event scale runoff production, baseflow generation and the seasonal water balance (Q 3.1)? Attempting to better understand the interplay of state, structure and forcing and runoff generation on different time scales I propose a set of diagnostics, i.e. dimensionless state(forcing)-response plots. These are tailored for baseflow production (radiation driven case), event runoff generation (rainfall driven case) and the seasonal water balance. In the former two cases (compare Figs. 4.3 and 4.4) I employ different storage surrogates including the well-known pre-event discharge (Q^*) and antecedent moisture (θ^*) which I normalize using root zone storage capacity to predict runoff responses. Runoff responses are baseflow (radiation driven case), which I normalize using saturated hydraulic conductivity *(Ks)* and, event runoff coefficients (rainfall driven case), i.e. quickflow divided by total rainfall depth.

The suggested diagnostics proved useful as they explain up to 70 % of the variability in the runoff responses at some of our test sites $(n = 22)$. This applies although the applied methods rest on (partly) strong assumptions and fairly coarse, operational data. The comparison of different catchments further reveals distinct spatial patterns,

suggesting regional differences in the importance of different storage compartments. Among the compared storage measures (normalized) pre-event discharge was the most important predictor for event runoff response in terms of the explained variance and with respect to the number of catchments which exposed significant correlations, though it is the less sophisticated measure (see Tables 4.1 and 4.2). On the seasonal scale normalized double (and triple) mass curves prove to be an easy-to-compute, yet very powerful means to detect similarities and differences in the partitioning of rainfall into evapotranspiration and discharge across scales. In our case their general shape is (invariantly) characterized by a linear increase in the winter period and a regime shift with small, near-to-zero slope when vegetation starts to control the water balance.

Compare also the double mass curves reported by Hellebrand et al., 2008; Jackisch, 2015; Pfister, Iffly, and Hoffmann, 2002.

Is it possible to detect evidence for intensity controlled runoff formation based upon (hourly aggregated) operational data (Q 3.2)? Intensity controlled runoff generation is characterized by intensive, convective rainfall forcing and a fast, highly intensive stream flow response, reflecting onset of rapid subsurface flows and/ or infiltration excess. Intensity controlled runoff production hence occurs at time scales of minutes (Blöschl and Sivapalan, 1995) and in a threshold-like manner (Lehmann et al., 2007; Struthers and Sivapalan, 2007; Zehe and Blöschl, 2004). It is neither controlled (or limited) by additive rainfall properties nor by current storage. The key to detect intensity controlled runoff formation mechanisms is hidden in the *high frequencies* of the involved signals. As operational data are often hourly aggregates and hence poorly resolved, i.e. *low-passed* from the perspective of intensity controlled processes, *edge-filtering* proved to be a meaningful technique to carve out traces of these mechanisms from operational time series. Therefore, I propose to combine the normalized temporal intensity changes (Eq. 4.7) and the normalized event duration (Eq. 4.6) into a diagnostic signature. It evidently detects signals of high frequency, intensity controlled runoff generation in at least two Alpine catchments (compare Fig. 4.8).

Which structural, climatic and ecological catchment characteristics explain the differences between different catchments and among different years and do any of them operate in groups (Q 3.3)? My investigations suggest that vegetation exerts a much stronger control on seasonal stream flow generation than formerly expected. In my sample of (humid) catchments I find that onset and termination of the period of vegetation trigger significant regime-shifts in seasonal runoff production across a range of different physiographic settings and scales (compare Fig. 4.5 and Table 4.3). Temperature based estimates of the vegetation period thereby prove to be simple to derive and more robust than Gregorian definitions such as the beginning of spring or that of the hydrological year. Interpreting temperature sums as bulk surrogates for ecological controls I find that they explain 70 % of the variability of the average seasonal summer runoff coefficients.

The results further suggest that it can be helpful to identify and to jointly evaluate the components that govern our dynamical laws and

which ultimately drive every flux (Zehe et al., 2014). These components are *gradients* (e.g. surface/bedrock topography) and *resistances* (e.g. hydraulic conductivity). Both, during radiation-driven conditions (see chapter 4.4.1) and during the dormant season (see chapter 4.4.3.1), the product of saturated hydraulic conductivity (Ks) times the median topographic gradient (ϕ) explains a higher portion of the variance in the runoff response than the two parameters alone. This corroborates that the use of *parameter groups* can be meaningful, at least at certain time scales. I conclude that the *gradient-flux-resistance* relationship should be further explored and that the proposed diagnostics should be applied to a larger number of catchments and to a (nested) set of small and densely instrumented catchments with homogeneous geological setup.

5.2 DISCUSSION AND OUTLOOK

The reservoir study involved the application of distributed hydrological and hydrodynamic models to a large spatial-domain and confronted me with *parameter estimation*, *model evaluation* and the need for a better (distributed) *process understanding*. Parameter estimation was required in the estimation of event runoff coefficients and baseflow rates. Thorough evaluation techniques were relevant to relate reservoir impact to model accuracy and to judge spatial model performance. Explaining the latter called for a deeper understanding of the runoff production mechanisms and their physiographic controls. The development of novel evaluation techniques (chapter 3) and data-driven process diagnostics (chapter 4) opened up several exciting and challenging research questions (Q 2.1 ... 3.3). Partly I could answer these questions, partly aspects thereof remain open for future research. In this chapter I address noteworthy issues related to the evaluation of closeness and the development of process diagnostics. I further name potential future research avenues and highlight links between the different topics.

Additional background information on (operational) modelling aspects is provided in appendix A.1.1.

5.2.1 *New avenues in assessing closeness*

My research introduces new methods for the assessment of closeness that is fundamental for closing the *learning cycle* (see chapter 1.1). Basically, I suggest two new alternatives: The first relates to the *traditional* way of evaluating hydrological models by comparing simulated model outputs against observations. Therefore I improved the Series Distance (SD) method which is able to quantify errors on the ordinate and on the abscissa. The second alternative proposes the use of *functional* signatures. These can be used for model/ process diagnostics but also for evaluation. First I elaborate the former.

In hydrology it is common to assess magnitude (*vertical*) errors in streamflow simulations. However, timing (*horizontal*) errors are rarely considered. Since time is a property which we can measure with high precision, timing errors in the measurement of streamflow are in deed

negligible. Temporal deviations of the simulation from the observation are however not negligible as these may point towards deficiencies in the parametrization or in the structural form of the infiltration, runoff concentration and/ or routing schemes. The development of methods which disentangle timing and amplitude errors and which explicitly quantify timing errors are thus highly relevant for both, practice and science.

In environmental modelling a wealth of literature on performance evaluation exists (Bennett et al., 2013). Most of the statistics which are used in hydrology judge closeness in a purely *vertical sense* however, meaning that only errors on the ordinate are evaluated. In this respect, a high effort has been undertaken to formulate vertical metrics in various ways such that they are sensitive to different ranges of the ordinate. A well known example for this is the Nash-Sutcliffe-Efficiency which emphasizes deviations in the higher magnitude ranges due to the squaring of errors. A common approach to better evaluate deviations in the lower magnitude ranges is to apply the Nash-Efficiency to the (log) transformed time series.

Approaches which analogously judge model performance with respect to *temporal* aspects are rarely reported and/or are fairly simplistic. Examples for the former include seasonal evaluations as proposed by He et al., (2015) or the consideration of variables like the *time evolution of snow heights* or the *timing of snowmelt-induced spring runoff* as suggested by Hingray et al., (2010). Examples for the latter include the use of (over)simplified criteria such as the Peak-Time-Difference (chapter 2.2.4.1) or the consideration of the lag time between event rainfall and simulated vs. observed runoff. While this is in principle a meaningful parameter its quantification is difficult and estimates based upon the maximum lag of the cross-correlation function (Kirchner, 2009; Yilmaz, Gupta, and Wagener, 2008) are rather coarse. A more elaborate alternative is the assessment of error groups as proposed by Reusser et al., (2009). In this approach the authors indirectly infer the effect of timing errors on a selection of vertical error statistics, which have been summarized into different *error response groups*, by means of synthetic errors. The Series Distance (SD) concept I present in chapter 3 offers a method to directly measure timing errors and thus, to avoid such detours.

The second avenue I propose is the (additional) consideration of *functional* signatures in the model evaluation process. This is in line with the seminal papers of Hrachowitz et al., (2013), Wagener et al., (2007), and Yilmaz, Gupta, and Wagener, (2008). While the potential of signatures for evaluation and classification is widely accepted, a commonly agreed upon definition on what *meaningful* signatures are, is not yet available. In many recent studies signatures are defined as specific characteristics of the hydrograph such as autocorrelation, slope of/ or bias in the flow duration curve (or different segments thereof), rising limb density or peak distribution (Euser et al., 2013). Others even consider flow statistics such as mean, variance, skewness or the coefficient of variation as signatures (Ley et al., 2011). In any case, an obvious shortcoming of such a definition of signatures is

Further examples for the application of signatures (and associated uncertainties) are provided by Casper et al., (2012), McMillan et al., (2014), Olden and Poff, (2003), Pfannerstill, Guse, and Fohrer, (2014), and Westerberg and McMillan, (2015), among others.

that only the *output* of the system, here observed and/or simulated streamflow is considered. An evaluation of *behavioural consistency* as proposed by Yilmaz, Gupta, and Wagener, (2008) however requires the consideration of the entire *input-state-output behaviour* of a catchment. This is precisely what the signatures proposed in chapter 4.2.2 do, in that they jointly consider at least two components of the *input-state-output* triple. For this reason I consider them as *functional* or *behavioural signatures*. The need to characterize and to evaluate *behaviour* is of high relevance for both science and practice and has also been emphasized by McMillan et al., (2011a), McMillan et al., (2014), and Schaefli et al., (2011).

Due to the promising results I obtained in the application of signatures and due to the importance of timing uncertainties I observed in streamflow simulations, I conclude that available model evaluation guidelines (e.g. Biondi et al., 2012; Harmel et al., 2014; Moriasi et al., 2007) should be extended with respect to two general aspects: i) timing errors should be considered in model evaluation and ii) the use of signatures should be become common practice in the evaluation processes.

5.2.2 *Improving the understanding of runoff production at different time scales*

Knowledge of the dominating runoff production mechanisms and their corresponding physiographic controls is important and highly relevant for various issues including the modelling of large spatial domains or catchment inter-comparison studies. Unfortunately these mechanisms are often not well understood. In consequence, modellers are confronted with the difficulty that it is (almost) impossible to gain a deeper understanding of the runoff production mechanisms and their spatio-temporal variations with reasonable effort. To overcome this limitation I propose a set of diagnostic signatures (compare chapter 4.2.2) which are based on commonly available data. The proposed diagnostics relate two components of the *input-state-output* triple and include, at least in parts, a structural normalization. Consequently, the signatures allow the assessment of *runoff behaviour* more thoroughly than signatures which are derived based upon *output* only. The normalization fosters a signature based comparison of different sites.

Several of the results obtained in chapter 4 are noisy. Parts of the noise can likely be explained by the inherent uncertainty of the (large number of) aspects that had to be considered in the development of the proposed diagnostics. This includes normalization, catchment scale storage estimation, to derive structural (subsurface) properties, the automated detection of rainfall-runoff events in continuous time series and/or poorly resolved data e.g. to detect intensity controlled runoff production mechanisms. Several of these issues are hard to grasp, difficult to quantify and/or not yet fully understood. In this respect noisy results are not surprising. It is perhaps noteworthy that

the proposed methods suggest functional similarity among different sites for baseflow production, storm runoff production and the seasonal water balance, despite that fact that they were applied to mesoscale catchments. To increase confidence in the proposed methods I suggest applying them to a (nested) set of small and densely instrumented research catchments with preferably homogeneous geological setup. Further, I suggest the testing of different normalization schemes. A successful validation of the introduced methods would open up new avenues to learn about functional similarity in a data-driven way.

Please also refer to the final revised manuscript of Seibert et al., (2016) which will be published after this thesis and include details which are not mentioned here.

To highlight some links between the different chapters I first want to recall the necessity of estimating event runoff coefficients for the event-based rainfall-runoff models in the reservoir study (compare chapter 2.2.2.1). These were required as input parameters. In the reservoir study I estimated them empirically by optimizing them for many events and sub-catchments and then determining areal averages for different pre-event discharge classes. While this procedure relies on discharge as well, it does not account for spatial variations in the importance of this predictor or the *kind* in which discharge is used as predictor. With respect to the results obtained in the chapter on "storage control on rainfall-runoff response" (chapter 4.4.2) I conclude that much better predictions of the event-runoff coefficients would have been possible by considering different predictors and techniques, at least in certain areas. Open for future research is a more thorough assessment of the time of integration/summation in the proposed pre-event discharge and/or near surface storage estimators (Eq. 4.3 and 4.2). It may introduce a high degree of subjectivity as pointed out by Heggen, (2001) and Graeff et al., (2012) and I expect, that it can be defined in a more meaningful way.

A possible application of the proposed signatures on "storage and structure control on baseflow generation" (chapter 4.4.1) is the direct use of (empirical) storage-discharge relationships in hydrological modelling or to estimate initial conditions for it. The latter is particularly relevant for event-based rainfall-runoff models where wrong state estimates can introduce a significant uncertainty. This was also the case in the LARSIM models (Ludwig and Bremicker, 2006) which I applied in the reservoir study. Partly I had to estimate the initial states based upon static and thus, rough estimates of average conditions. As several of the storage-discharge relationships were noisy it could be worth revising the approach with regard to the suggestions of Rupp and Selker, (2006) who assess noise in dQ/dt-recession plots, a method which was initially proposed by Brutsaert and Nieber, (1977).

Further links between the proposed diagnostic signatures on runoff production and the assessment of closeness can easily be established in that the proposed signatures are included in the evaluation process. All of the proposed methods can be calculated based upon simulated discharge data (Q) as well, in that the observed Q in Eqs. 4.1, 4.3, 4.4, 4.5 and 4.7, respectively is replaced by its simulated pendant. Com-

paring simulated and observed state(forcing)-response plots allows us to identify (dis)similarities in model performance.

A promising benchmark for model evaluation can also be derived from the double mass curves (chapter 4.4.3.1 and appendix A.2.5). I suggest that the reproduction of the tipping point (regime shift) can serve as a powerful (timing) benchmark. This is particularly relevant for models which are used for climate change studies where coping with the (most likely) nonstationary role of biotic controls (Milly et al., 2008) is of utmost importance.

The regime shift which I observed across different scales and physiographic settings is also visible in the findings obtained by Pfister et al., (2003) and Jackisch, (2015). The wide-spread occurrence of such distinct patterns raises the general question whether the importance of phenological controls are adequately acknowledged and whether (hydrological) models represent the phenological cycle in sufficient detail. The parameterization of controlling variables such as albedo, leaf-area-index, stomata and cuticula resistance which govern plant/-canopy roughness/resistance still prevails in the form of static look-up tables which have typically been derived within a certain hydro-climatic setting (Zehe et al., 2001). This also applies for so called *physical* models such as Catflow (Maurer, 1997; Zehe et al., 2001) or WaSiM-ETH (Schulla, 1997). For further reading please refer to Loritz et al., (2016) who conduct additional studies on this topic and, among others, confront modelled ET estimates with sap-flow data and both Gregorian and temperature based estimates for the beginning of the vegetation period.

Fundamental research further remains to be undertaken in detecting intensity controlled runoff production, in the search for other, more meaningful storage descriptors and in developing proper normalization strategies. The latter is an important topic for the use of the diagnostics in inter-comparison studies. Also the double (and triple) mass curves should be subject to further research. The DMCs can in essence be regarded as an analogy to dimensionless tracer breakthrough curves which are widely used in soil physics to assess transport and adsorption properties (Hillel, 2004; Jury and Horton, 2004). In soil physics however different normalization schemes are used. Accordingly I suggest the exchange of total precipitation on the abscissa by the available pore volume (despite that associated uncertainties are high). This way, mass input is expressed in terms of storage volume. A further alternative to normalize the abscissa would be to use a data-driven estimate for the maximum potential evaporation such as net solar radiation divided by the latent heat of vaporization (assuming the entire incoming energy would be consumed for evaporation). This would separate cold from warm years.

5.3 SYNTHESIS

The two overarching goals of modelling are the *ability to predict* and to gain a *deeper understanding* of the system under consideration.

The first part of my dissertation deals with hydrological modelling in large spatial domains. Therefore I employ distributed models from the Bavarian flood forecasting agency which are operationally used for flood forecasting. A vital aspect in the application of any *prediction* tool is to know about its precision. For this purpose, I revised the Series Distance (SD) (Ehret and Zehe, 2011) method which assess timing and magnitude errors in streamflow simulations. Among others, I extended the SD procedure by a *coarse-graining*, i.e. pattern matching, procedure. Contrary to the initial version of SD, the method does now efficiently mimic visual hydrograph inspection in an automated way. The SD method can hence be applied to continuous time series.

In the context of *predictions* it is important to recall that *forecasting* implies extrapolating from one (potentially unobserved) system state to another (potentially unobserved) system state. As Reusser, (2010) pointed out, this can only be done in a meaningful way if the model represents the bio-physical processes in a realistic way. At the end of the day this requires *behavioural modelling* (Schaefli et al., 2011) or as Yilmaz, Gupta, and Wagener, (2008) or Clark et al., (2008) put it, *structural consistency*. The introduced set of diagnostics or *functional signatures* seeks to enhance our understanding on the bio-physical processes in that it assesses catchment runoff production at different time scales and with consideration of the physiographic controls. Although the results from a small catchment inter-comparison study are noisy in parts, they clearly suggest that different catchments are *functionally* similar for base-flow generation, storm-runoff production and/or the seasonal water balance.

Together these topics close a learning cycle in that the precision of an available model is quantified and new hypotheses towards possible model improvements are derived.

Part VI

APPENDIX

A

APPENDIX

A.1 APPENDIX OF PART II

A.1.1 *Hydrological processes and man-made water regulation issues in the Bavarian part of the Danube basin*

The need and potential of hydrological models for flood mitigation is beyond question. Flood forecasting is however a difficult business as the focus is on extreme and thus rare events. In the *reservoir study* I faced additional challenges which are related to the large spatial domain of the study area and which are caused by the high physiographic variability and the large number of man-made water regulation and management issues. In the following I provide background information on selected issues thereof.

The Danube basin which is described in the chapters 2.2 and 4.3 covers very heterogeneous landscapes, ranging from alpine areas over glacial outwash sequences towards flat and low-lying regions south of the Danube. These continue north of the Danube as partly karstified areas and/ or regions which are built up of sand-stone or crystalline rock. Consequently, these regions host a large number of different catchments which in turn expose a range of *locally* specific processes that may however be relevant for the *regional* modelling of hydrological processes. To name a few important ones:

- *Exfiltration of stream flow into adjacent ground water bodies*: This occurs particularly at the lower river courses of the larger southern tributaries of the Danube (see Fig. 2.1). In these areas the streams expose very small hydraulic gradients and cross regions which host extensive and permeable aquifers. Although the significance of this process is well-known in a qualitative sense (particularly during high flow conditions), quantitative estimates of the amount of stream flow which exfiltrates and the corresponding dynamics are unknown.

- *Activation of flood plains:* Depending on the degree of flooding vast areas are flooded alongside the Danube and its major tributaries. The activation of flood plains, related backwater effects and meander shortcuts are a widespread, case specific and complex hydrodynamic process which dampen, delay and/or deform the flood wave. Solely for the Danube, Skublics, (2014) estimated a total active (natural) retention volume of $220 \cdot 10^6$ m^3 by means of hydrodynamic modelling within the river section between the cities of Neu-Ulm and Straubing, assuming a flood wave with a 100-year return period. Historically, a total volume of approximately $489 \cdot 10^6$ m^3 was available in the same reach.

- *Subsurface processes:* South of the Danube between the rivers Iller and Isar extensive gravel fields prevail. The *Iller-Lech-Schotterplatten* and adjacent areas cover several thousand square kilometers and coincide approximately with the *light green* region covered by the model with the MID 3 in Fig. 2.1. This area is characterized by very weak hydraulic gradients and extensive, permeable aquifers. In consequence it hosts practically no streams and surface water bodies and it is clear that the action does not take place on the surface but in the fluid mechanics of the subsurface.

- *Intensity controlled runoff production*: Strong topographic gradients, shallow and poorly developed soils, sparse surface coverage and a dominance of karst and rift aquifers suggests that intensity controlled runoff formation process are of high importance at many of the southern Alpine landscapes. As these areas receive large amounts of precipitation they are important source areas for floodings in the Bavarian part of the Danube basin.

The processes described impact runoff formation, concentration and the propagation of flood waves, i.e. routing. Even though the occurrence of the processes described above is spatially limited, it is clear that they may have a *regional* impact and that they should hence be considered in large-domain modeling exercises such as the reservoir study. The same applies for a large number of man-made water regulation and management issues. In the Bavarian part of the Danube basin these include for instance:

- *Chains of water barrages*: Alone for the river Lech, a total number of 26 water barrages exist. 17 and 16 further barrages are located within the Bavarian parts of the rivers Danube and Inn, respectively. Partly these are operated according to strict operation plans, partly not, at least this is what hydrodynamic evaluations suggest (Skublics et al., 2013).

- *Regulation and re-distribution of water*: The stream network in the Bavarian part of the Danube basin is subject to severe regulation. Solely the hydrological model of the sub-catchment of the river Isar comprises more than 80 individual intersects which route water (mainly threshold-based) from one point to another in order to account for power plant operation, provision of cooling water, and others. Further examples of severe regulation can be found in the Altmühl basin where water is constantly transferred from the Danube Basin to the Rhine basin, to minimize low-flow impacts within the Regnitz River. Inland water transportation, the operation of different channels and in particular, drainage for agricultural reasons promote a fairly intensive, widespread and complex regulation of the terrestrial water cycle.

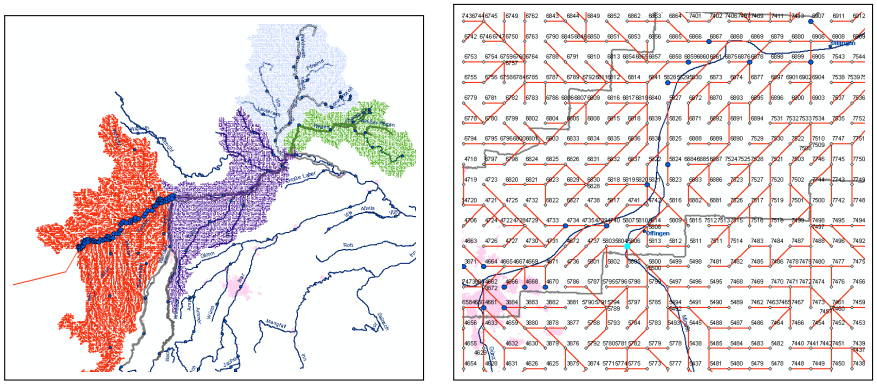
- *Manual operation of reservoirs and flood retention basis*: In total, 13 larger reservoirs and retention basins with a total retention volume of $\approx 127 \cdot 10^6$ m^3 are distributed across the Bavarian part

of the Danube basin (another five retention basins are currently in the planning stage). They are operated according to operation guidelines but manually and on demand. This makes it very difficult to incorporate clear operation rules into any (operational) model.

Physiographic and man-made processes impact the terrestrial water cycle in various ways and highlight that high-resolution modelling of large-domains requires consideration of a multitude of natural and artificial processes. This makes modelling a fairly complex business. In respect of the above it is hence not surprising that the operational models at hand include (or require, depending on the perspective) a large number of technical quick-fixes and improvisation. These include for instance dead-end intersections that route water threshold based into the *nowhere*, to mimic losses into the ground water body (despite that violates the mass balance). Others bypass (partly larger) river sections to emulate effects caused by the presence of water power plant operations or ground water effects. To compensate for mass balance issues, LARSIM (Ludwig and Bremicker, 2006) provides a scaling factor for areal precipitation ("KG" Faktor) implying that discharge measurements enjoy a much higher degree of confidence than estimates of (areal) precipitation. Last, a set of different assimilation techniques allows pulling simulated time series on top of the observed e.g. by updating state variables whenever required.

These examples illustrate our difficulties when coping with processes that are not yet well understood and demonstrate what operational *solutions* may look like. The latter are justified for operational purposes as these *kinds* of models are tailored to *predict* and not to *improve understanding*. Further, operational purposes require feasible solutions for a reasonable effort. This is in line with Gupta, Wagener, and Liu, (2008) who point out that operational practice is more concerned with the *accuracy* (unbiasedness) and *precision* (minimality of uncertainty) of the model simulation or forecast, than in the correctness of the structural form of the model. Models which seek to improve the understanding of the system under investigation however need to represent the bio-physical processes in a realistic way. This is a vital requirement for *distant* extrapolations where *forcasting* implies extrapolating from one (potentially unobserved) system state to another (potentially unobserved) system state (Reusser, 2010).

A.1.2 *Structure of the LARSIM model*



(a) Overview (b) Detail

Figure A.1: Screenshot of the grid based (size = 1 Km^2) structure of the LARSIM model (Ludwig and Bremicker, 2006).

A.2 APPENDIX OF PART IV

A.2.1 *Physiographic site properties*

	topography				% land use coverage						hydro-meteorology				
ID	A [km^2]	elev [m]	ϕ [-]	$\phi \cdot Ks$ [m/s]	infr	arab	past	frst	wet	rock	MAP [mm]	$\overline{P}$ [mm]	$\overline{Q}$ [mm]	$\overline{CR}$	var.c(Q)	sFDC
TRI1	88	481	2.8e-02	1.0e-06	0.04	0.54	0.20	0.21	0	0	802	919	0.045	0.43	2.00	-1.46
TRI2	26	460	2.5e-02	9.8e-07	0	0.60	0.12	0.29	0	0	707	801	0.033	0.36	2.20	-1.94
TRI3	93	468	3.8e-02	8.3e-07	0.02	0.62	0.07	0.30	0	0	738	829	0.032	0.33	1.80	-0.99
JUR1	90	518	7.2e-02	2.0e-06	0.01	0.59	0.15	0.26	0	0	833	839	0.046	0.48	0.83	-0.78
BFO1	25	620	6.9e-02	2.5e-06	0	0.35	0.06	0.60	0	0	889	933	0.054	0.51	0.89	-0.61
BFO2	64	635	6.1e-02	1.3e-06	0.02	0.39	0.12	0.47	0.01	0	893	920	0.055	0.53	0.91	-0.64
BFO3	58	624	7.9e-02	1.2e-06	0.01	0.24	0.21	0.55	0	0	908	825	0.052	0.56	1.10	-0.76
MOL1	166	543	1.1e-02	1.6e-08	0.07	0.42	0.29	0.23	0	0	889	973	0.042	0.38	0.83	-0.63
MOL2	163	515	4.0e-02	6.9e-08	0.03	0.28	0.37	0.32	0	0	901	1010	0.045	0.39	0.90	-0.36
MOL3	163	558	1.4e-02	2.0e-08	0.05	0.69	0.10	0.15	0	0	933	1100	0.057	0.45	0.64	-0.27
MOL4	97	517	2.5e-02	3.9e-08	0.04	0.77	0.03	0.15	0	0	888	1016	0.042	0.36	0.93	-0.50
MOL5	133	473	2.0e-02	4.4e-08	0.05	0.81	0.05	0.09	0	0	883	1016	0.047	0.40	1.00	-0.57
MOL6	146	484	4.2e-02	7.8e-08	0.02	0.79	0.04	0.15	0	0	856	721	0.029	0.35	1.60	-0.58
MOL7	87	379	2.4e-02	3.4e-08	0.02	0.73	0.01	0.24	0	0	744	733	0.026	0.31	1.10	-0.71
AFO1	45	840	3.4e-02	3.6e-08	0.01	0.11	0.27	0.62	0	0	1388	1243	0.073	0.51	1.90	-1.32
AFO2	95	777	4.0e-02	9.1e-08	0.01	0.11	0.55	0.31	0.01	0	1292	1466	0.083	0.50	1.30	-0.76
AFO3	136	751	2.2e-02	5.2e-08	0.05	0.12	0.62	0.22	0	0	1198	1015	0.045	0.38	0.72	-0.75
AFO4	12	688	2.9e-02	3.2e-08	0.07	0.32	0.31	0.29	0	0	1114	1024	0.047	0.40	1.20	-1.03
ALP1	47	1279	3.3e-01	1.8e-06	0.00	0.05	0.50	0.45	0	0	2212	2662	0.230	0.75	1.40	-1.14
ALP2	127	1433	4.0e-01	7.7e-07	0.01	0.02	0.55	0.28	0	0.15	2315	2526	0.240	0.83	1.10	-0.83
ALP3	76	1539	5.1e-01	7.6e-07	0.01	0.01	0.59	0.18	0	0.21	2438	2181	0.210	0.86	0.89	-1.07
ALP4	114	1270	4.2e-01	5.1e-07	0.01	0.01	0.25	0.61	0.02	0.09	1826	1684	0.120	0.64	1.00	-0.49

Figure A.2: Physiographic catchment properties in terms of topography, landuse and hydro-meteorology. The columns contain site identifier (ID), catchment size A, mean catchment elevation above sea level (elev), median gradient (ϕ), median topographic gradient times average saturated hyd. conductivity ($\phi \cdot Ks$), relative land coverage ratios for infrastructure (infr), arable land (arab), pasture (past), forest (frst), wetlands (wet) and rock outcrops (rock), the 30 year mean annual precipitation (MAP), four year mean annual precipitation ($\overline{P}$), discharge ($\overline{Q}$), runoff coefficient ($\overline{CR}$), streamflow coefficient of variation (var.c(Q)) and the slope of the flow duration curve between the 33 and 66% percentiles (sFDC). $\overline{CR}$, var.c(Q) and sFDC are dimensionless.

SITE	τ [cm]	eFC_τ [mm]	AC_τ [mm]	FC_τ [mm]	TPV_τ [mm]	clay [%]	silt [%]	sand [%]	skel [%]	Ks [m/s]	n soils	APR
TRI1	68.3	73.9	47.1	262.3	309.4	42.7	9.3	48.6	1.8	3.6e-05	3	2
TRI2	67.1	71.0	49.8	248.4	298.2	39.9	9.0	51.7	1.9	3.9e-05	3	1
TRI3	76.1	94.2	59.2	216.9	276.2	23.3	18.0	58.4	2.5	2.2e-05	4	1/2
JUR1	26.2	32.5	37.0	71.0	108.0	39.2	16.8	43.5	3.7	2.8e-05	6	1
BFO1	60.0	74.0	55.5	110.4	165.9	7.5	9.5	83.2	4.0	3.6e-05	5	4
BFO2	54.8	74.5	40.8	124.0	164.8	12.4	17.0	71.1	3.6	2.1e-05	6	4
BFO3	58.0	67.8	36.4	125.3	161.7	14.3	18.7	67.6	3.8	1.5e-05	5	4
MOL1	85.1	144.8	44.6	330.9	375.5	26.4	58.8	14.4	1.1	1.5e-06	8	1/2/3
MOL2	78.9	146.1	44.3	312.1	356.4	22.3	56.5	21.8	1.4	1.7e-06	4	2
MOL3	86.5	151.3	55.4	290.7	346.1	22.0	48.5	28.9	1.9	1.5e-06	8	2
MOL4	89.2	132.5	46.2	323.9	370.2	22.7	57.2	21.0	1.2	1.6e-06	4	2
MOL5	88.2	138.3	54.3	288.8	343.0	19.0	46.5	34.8	1.9	2.2e-06	12	1/2
MOL6	88.3	129.5	47.7	308.3	355.9	21.0	54.0	24.7	1.3	1.9e-06	3	2
MOL7	90.0	167.1	47.5	328.6	376.1	23.0	66.8	11.2	1.0	1.4e-06	3	2/3
AFO1	100	153.0	76.0	322.5	398.5	21.0	39.0	40.0	2.0	1.1e-06	1	3
AFO2	74.0	149.6	57.0	285.8	342.8	19.8	39.5	40.8	2.2	2.3e-06	5	3
AFO3	79.4	138.4	61.8	266.0	327.8	19.3	33.7	47.1	2.5	2.4e-06	8	3
AFO4	80.0	142.2	59.5	283.4	342.9	20.8	38.0	41.2	1.8	1.1e-06	2	2
ALP1	55.5	90.2	38.5	209.3	247.8	15.8	23.7	60.5	2.5	5.5e-06	2	2/3
ALP2	28.1	42.2	18.8	108.0	126.7	28.2	28.8	44.5	4.2	1.9e-06	5	1/2
ALP3	24.1	35.2	16.0	92.7	108.7	29.4	28.7	42.4	4.4	1.5e-06	5	1/2
ALP4	28.9	41.1	18.7	110.7	129.4	37.5	24.4	39.3	4.3	1.2e-06	5	1

Figure A.3: Soil properties of the selected headwater catchments: Root zone depth (τ), effective field capacity ($e\mathrm{FC}_\tau$), air capacity (AC_τ), field capacity (FC_τ), total pore volume (TPV_τ), and contents of clay, silt, sand and skeleton (skel). FC, eFC, AC and TPV refer to τ. nsoils gives information on the total number of different soils classes within the individual sites. Average saturated hydraulic conductivity (Ks) was estimated based on grain size data Schaap, Leij, and Genuchten, 2001. The soil properties are weighted means (areal share) the national soil map of Germany (BGR, 1995). We also provide aquifer productivity classes (APR) from the international hydrogeological map of Europe (Duscher et al., 2015). The categorical APR values 1, 2, 3 and 4 indicate dominance of highly productive, low to moderately productive conditions, dominance of locally aquiferous rocks and non-aquiferous rocks.

A.2.2 *Re-scaling and consistency of integrative storage measures*

To ensure that two catchments at least potentially store the same amount of water and start with a similar storage amount, we define the starting point of integration of dS^* using seasonal criteria. Therefore, we first plot dS^* against accumulated annual precipitation normalized by the long-term mean annual precipitation (MAP) (Fig. A.4). This helps to compare states with similar potential accumulated input. Next, we re-scale the ordinate such that the origin corresponds to the mean of the local periodic minima, assuming that the soil moisture is near the permanent wilting point at these times. This way we gain a dimensionless estimator for the total active bulk catchment water storage. Values in dS^* of around zero indicate dry conditions whereas values around 1 indicate that dynamic storage is equal to the root zone storage volume. Note that both values > 1 (e.g. during the occurrence of snow) and values < 0, may occur and that absolute values must not be interpreted. Please note, that we encountered significant trends and erratic fluctuations in dS^* in the majority of all sites.

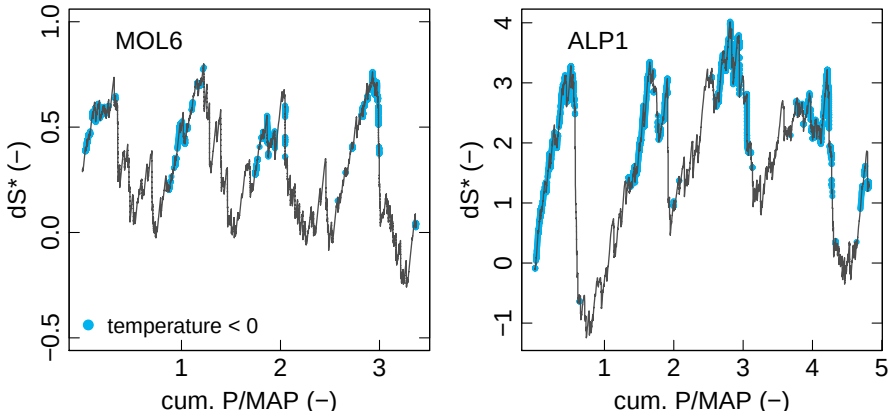


Figure A.4: Coherent normalization of integrative storage measures. The plots show examples from the sites MOL6 and ALP1 for the same four year period (11/1999-10/2003). Re-scaled and normalized dynamic storage dS^* (Eq. 4.1) is plotted on the ordinate, the abscissa shows cumulated precipitation divided by the long term mean annual precipitation (MAP). Please note the differences between the two sites in the scaling of the axes.

A.2.3 *Automated delineation of rainfall driven events*

Comparability of runoff coefficients requires essentially an automated detection of rainfall-runoff events in continuous time series to pool enough events into a statistically analyzable sample. The concept and interpretation of runoff coefficients (CR) on both event and annual time scales is old and dates back to Sherman, (1932). Up to now CRs are frequently used as diagnostic variables to describe response properties and runoff generation (compare e.g., Capell et al., 2012; Gra-

eff et al., 2012; Merz, Blöschl, and Parajka, 2006; Merz and Blöschl, 2009; Pearce, Stewart, and Sklash, 1986, and many others). However, CRs are not defined consistently (e.g. total runoff over total precipitation vs. total quick flow over total precipitation) and the literature describes a range of different methods for the detection of the start and end of an event which are required for the separation of the slow flow component as illustrated by Blume, Zehe, and Bronstert, (2007).

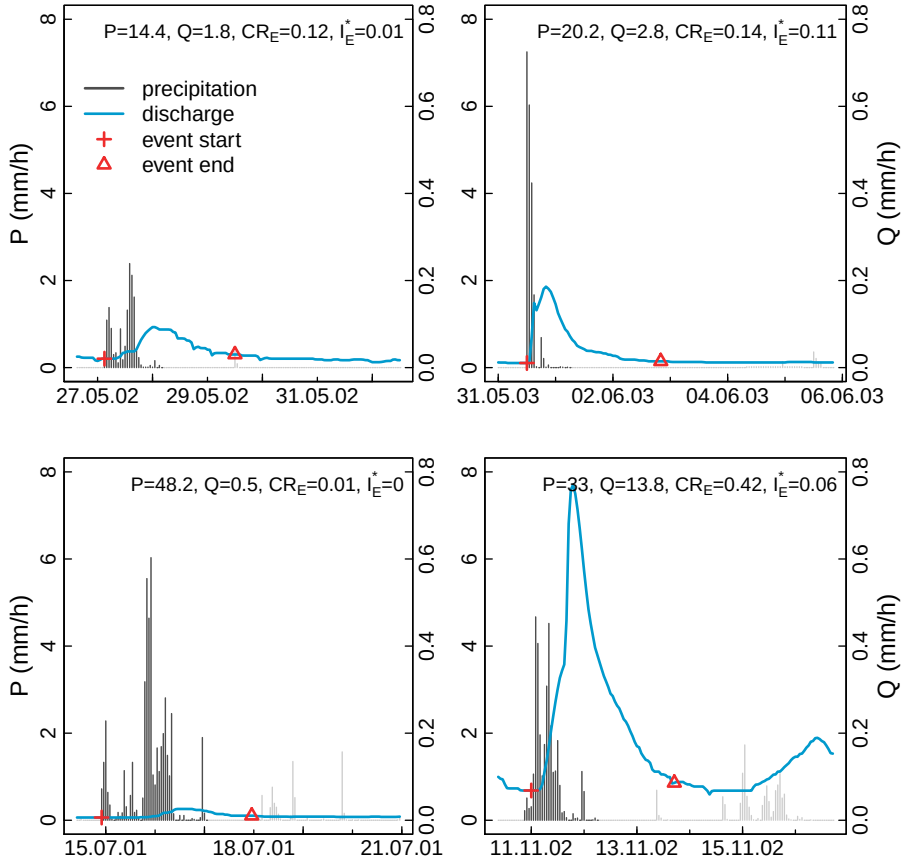


Figure A.5: Automated detection of rainfall-runoff events. The plots show results from four selected events from the site TRI3. The examples are selected from different seasonal periods and illustrate different temporal dynamics of forcing and response. The statistics in the top provide event specific totals of rainfall (P) and quickflow (Q), the event runoff coefficient CR_E (Eq. 4.5) and the normalized temporal intensity change I_E^* (Eq. 4.7).

We extensively tested various approaches including baseflow separation and filtering techniques (e.g., Chapman, 1999; Douglas and Peucker, 1973; Eckhardt, 2005; Perng et al., 2000), penalty functions (Drabek, 2010), fuzzy logic (Seibert and Ehret, 2012), and the methods proposed by Merz and Blöschl, (2009) and Norbiato et al., (2009). However, the results of these methods were usually unsatisfactory when applied to a range of different regimes of precipitation and stream flow. In the end we adapt and recombine different existing techniques and detect rainfall-runoff events based upon the following principles: First, we select *rainfall events* as subsequent periods of liquid rainfall (maximum up to 6 h of rain free period are tolerated)

with at least 10 mm of daily rain depth (compare also Fig. A.5). Given these periods, we identify the corresponding *discharge events* starting with the maximum flow rate. Between the latter and the beginning of rainfall we search for the first point in time where $dQ/dt > 0$ holds true for five subsequent time steps which we define as the start of the discharge event. Starting from the peak flow we next define the end of the discharge event using the constant-k method proposed by Blume, Zehe, and Bronstert, (2007). Due to missing convergence of this approach in 20 - 40 % of all cases we combine it with additional cut-off criteria (e.g. threshold exceedance, beginning of next rainfall event, and others). Missing convergence often results from varying rainfall intensities throughout the event. In our data set we observed that the occurrence of multiple peaks and troughs within a "single" event is more often the rule rather than the exception. Upon request, a program code for the automated detection of rainfall-runoff events in hourly time series which is written in R (R Core Team, 2015) can be obtained from the author.

A.2.4 *Linkage between site identifiers and gauge names*

Table A.1: Link table that relates the site identifiers (ID) introduced in section 4.3 to the corresponding gauge and stream names. Gauge locations are provided in Gauß-Krüger zone 4 coordinates (GKR and GKH).

ID	Gauge	Stream	GKR	GKH
TRI1	Reichenbach (REIB)	Wörnitz	4373327	5449863
TRI2	Binzwangen (BINZ)	Altmühl	4381996	5473002
TRI3	Bechhofen (BECH)	Wieseth	4394270	5447640
JUR1	Holnstein (HOLN)	Unterbürger Laber	4464800	5442860
BFO1	Gartenried (GART)	Murach	4532661	5483477
BFO2	Untereppenried (UEPR)	Ascha	4533425	5477338
BFO3	Tiefenbach (TIEF)	Bayerische Schwarzach	4543360	5477800
MOL1	Roth (ROTR)	Roth	4363140	5360723
MOL2	Fleinhausen (FLEI)	Zusam	4394141	5358887
MOL3	Mering (MERI)	Paar	4424840	5348870
MOL4	Odelzhausen (ODZH)	Glonn	4440860	5353360
MOL5	Appolding (APPO)	Strogen	4498575	5364071
MOL6	Dietelskirchen (DIKI)	Kleine Vils	4525540	5373175
MOL7	Wallersdorf (WALR)	Reißingerbach	4554850	5400160
AFO1	Unterthingau (alt) (UTHI)	Kirnach	4388313	5294058
AFO2	Hörmanshofen (HOER)	Geltnach	4399272	5299593
AFO3	Buchloe (BUCH)	Gennach	4404574	5323974
AFO4	Herrsching (HERR)	Kienbach	4438860	5318140
ALP1	Gunzesried (GZRI)	Gunzesrieder Ach	4366798	5266382
ALP2	Reckenberg (RECK)	Ostrach	4373822	5264305
ALP3	Oberstdorf (OBTR)	Trettach	4370128	5255320
ALP4	Oberammergau (OAMM)	Ammer	4429723	5273332

A.2.5 *Normalized double mass curves of remaining sites*

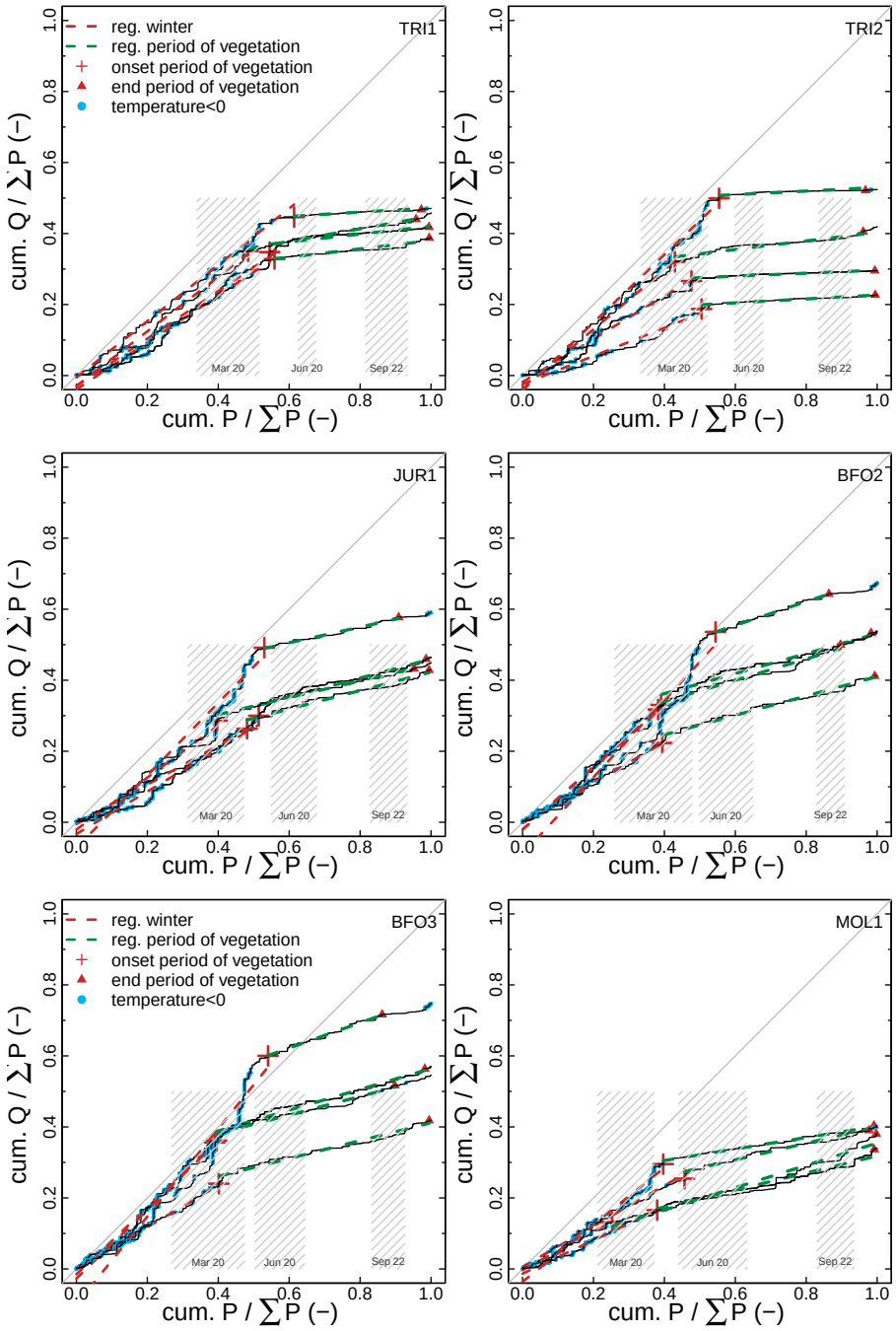


Figure A.6: Normalized double mass curves for the catchments TRI1, TRI2, JUR1, BFO2, BFO3 and MOL1.

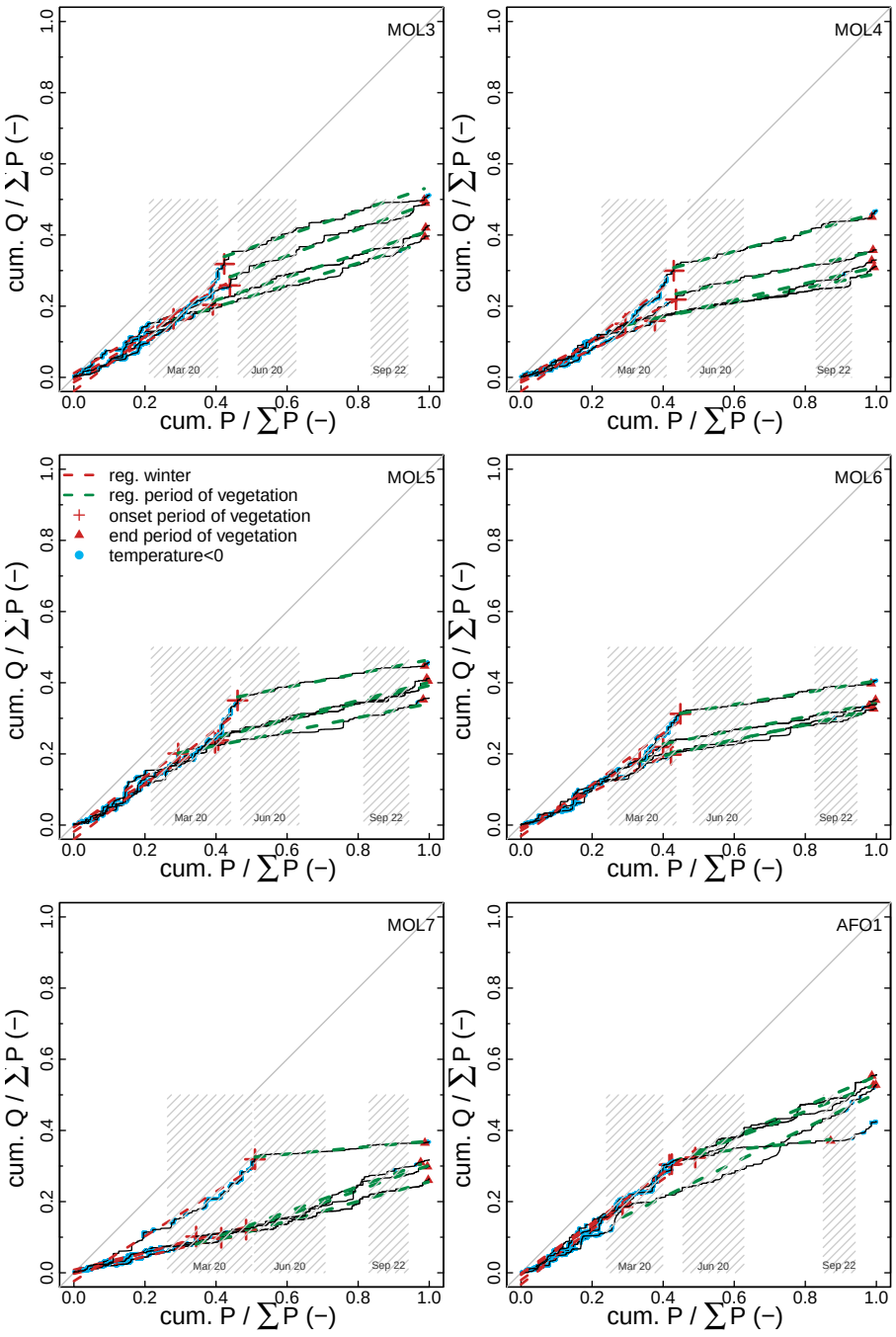


Figure A.7: Normalized double mass curves for the catchments MOL3, MOL4, MOL5, MOL6, MOL7, AFO1.

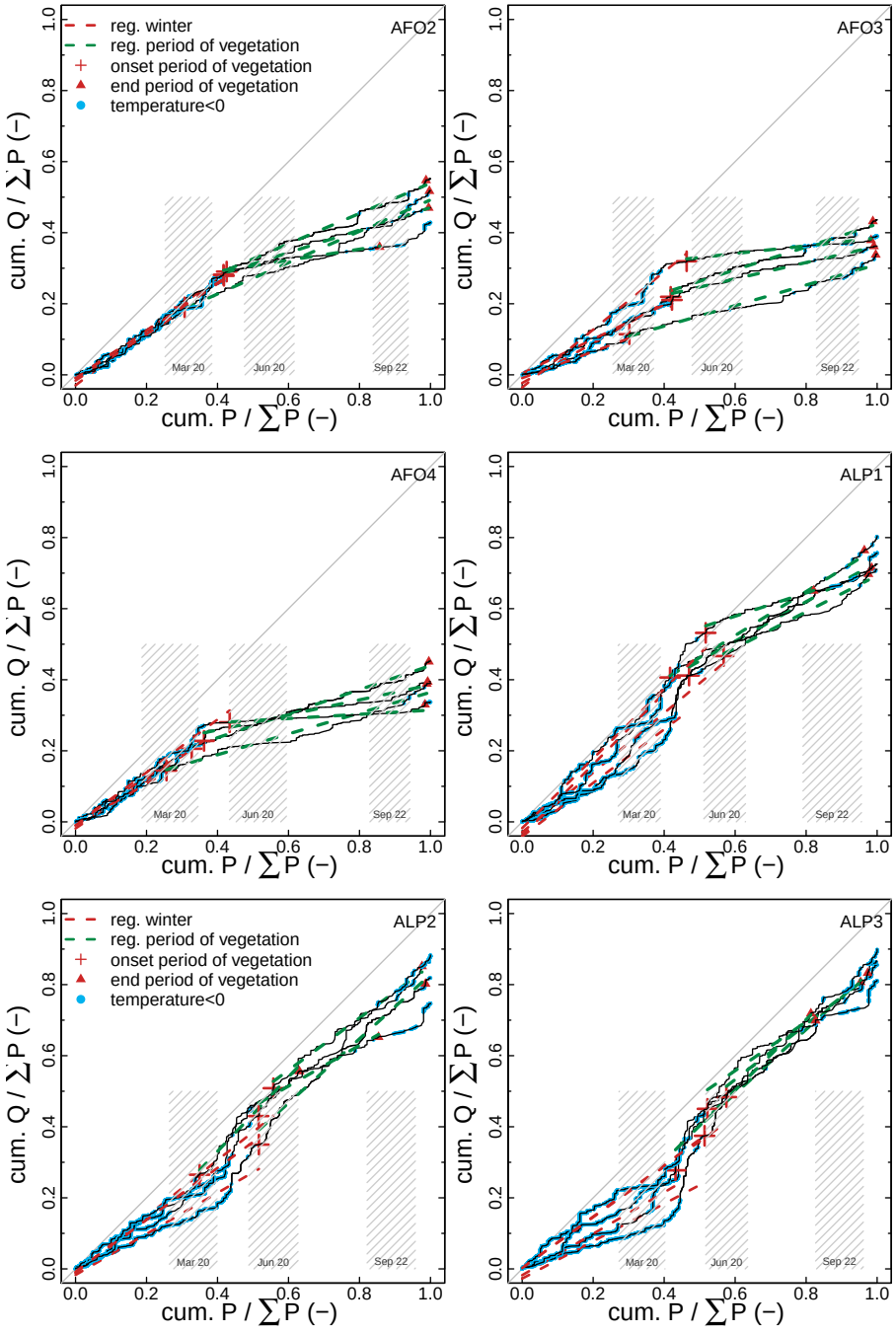


Figure A.8: Normalized double mass curves for the catchments AFO2, AFO3, AFO4, ALP1, ALP2, ALP3.

Part VII

BIBLIOGRAPHY AND BACK MATTER

BIBLIOGRAPHY

Ali, Geneviève, Doerthe Tetzlaff, Chris Soulsby, Jeffrey J. McDonnell, and René Capell (2012). "A comparison of similarity indices for catchment classification using a cross-regional dataset." In: *Advances in Water Resources* 40, pp. 11–22. DOI: 10.1016/j.advwatres.2012.01.008.

Attinger, Sabine (2003). "Generalized coarse graining procedures for flow in porous media." In: *Computational Geosciences* 7.4, pp. 253–273. DOI: 10.1023/B:COMG.0000005243.73381.e3.

BGR (1995). *Bodenübersichtskarte von Deutschland 1:1,000,000 (BÜK1000). Federal Institute for Geosciences and Natural Resources (BGR)*. Hannover.

BGR and SGD (2015). *Hydrogeological spatial structure of Germany (HYRAUM). Digital map data v3.2. German Federal States Geological Surveys (SGD) and Federal Institute for Geosciences and Natural Resources (BGR).* Hannover.

BLfU (2002). *Hochwasser im August 2002. Gewässerkundlicher Dienst Bayern. Bayerisches Landesamt für Wasserwirtschaft (Herausgeber und Verlag)*. Tech. rep. München.

BLfU (2003). *Hochwasser Mai 1999 Gewässerkundliche Beschreibung. Bayerisches Landesamt für Wasserwirtschaft (Herausgeber und Verlag)*. Tech. rep. München.

BLfU (2007). *Gewässerkundlicher Bericht Hochwasser August 2005. Bayerisches Landesamt für Wasserwirtschaft (Herausgeber und Verlag)*. Tech. rep. München.

BLfU (2011). *Gewässerkundlicher Monatsbericht Januar 2011 – Hochwasser. Bayerisches Landesamt für Umwelt (Herausgeber und Verlag)*. Tech. rep. München, pp. 1–9.

BMU (2002). *Hydrologischer Atlas von Deutschland. Bundesministerium für Umwelt, Naturschutz und Reaktorsicherheit (BMU) (Eds.)* Universität Freiburg. ISBN: 3000056246.

Bahram, S., J. Y. Pierre, and F. L Odgen (1995). "Similarity in catchment response: 1. stationary rainstorms." In: *Water Resources Research* 31.6, pp. 1533–1541.

Bao, Hongjun and Linna Zhao (2012). "Development and application of an atmospheric-hydrologic-hydraulic flood forecasting model driven by TIGGE ensemble forecasts." In: *Acta Meteorologica Sinica* 26.1, pp. 93–102. DOI: 10.1007/s13351-012-0109-0.

Bennett, Neil D et al. (2013). "Characterising performance of environmental models." In: *Environmental Modelling & Software* 40.0, pp. 1–20. DOI: http://dx.doi.org/10.1016/j.envsoft.2012.09.011.

Bergstroem, S (1976). *Development and application of a conceptual runoff model for Scandinavian catchments*. Tech. rep. Norrkoping: Swedish Meteorological and Hydrological Institute (SMHI), Report RHO 7.

Berne, A., R. Uijlenhoet, and P. A. Troch (2005). "Similarity analysis of subsurface flow response of hillslopes with complex geometry." In: *Water Resources Research* 41.9, pp. 1–10. DOI: 10.1029/2004WR003629.

Best, D.J. and D.E. Roberts (1975). "The Upper Tail Probabilities of Spearman's Rho." In: *Journal of the Royal Statistical Society. Series C* 24.3, pp. 377–379. DOI: 10.2307/2347111.

Beven, K. and Mj Kirkby (1979). "A Physically Based, Variable Contributing Area Model of Basin Hydrology." In: *Hydrological Sciences Bulletin Vol. 24, No. 1, p 43-69, March 1979. 16 fig, 1 tab, 36 ref* 24.1, pp. 43–69.

Beven, Keith (1989). "Changing ideas in hydrology - The case of physically-based models." In: *Journal of Hydrology* 105.1-2, pp. 157–172. DOI: 10.1016/0022-1694(89)90101-7.

Beven, Keith and Andrew Binley (1992). "The future of distributed models: model calibration and uncertainty prediction." In: *Hydrological Processes* 6, pp. 279–298. DOI: 10.1002/hyp.3360060305.

Beven, Keith and Peter Germann (2013). "Macropores and water flow in soils revisited." In: *Water Resources Research* 49.6, pp. 3071–3092. DOI: 10.1002/wrcr.20156.

Biancamaria, Sylvain, Paul D. Bates, Aaron Boone, and Nelly M. Mognard (2009). "Large-scale coupled hydrologic and hydraulic modelling of the Ob river in Siberia." In: *Journal of Hydrology* 379.1-2, pp. 136–150. DOI: 10.1016/j.jhydrol.2009.09.054.

Binley, Andrew and Keith Beven (2003). "Vadose zone flow model uncertainty as conditioned on geophysical data." In: *Ground Water* 41.2, pp. 119–127. DOI: 10.1111/j.1745-6584.2003.tb02576.x.

Biondi, Daniela, Gabriele Freni, Vito Iacobellis, Giuseppe Mascaro, and Alberto Montanari (2012). "Validation of hydrological models: Conceptual basis, methodological approaches and a proposal for a code of practice." In: *Physics and Chemistry of the Earth* 42-44, pp. 70–76. DOI: 10.1016/j.pce.2011.07.037.

Black, Peter E. (1997). "Watershed Functions." In: *Journal of the American Water Resources Association* 33.1, pp. 1–11. DOI: 10.1111/j.1752-1688.1997.tb04077.x.

Blöschl, G and Murugesu Sivapalan (1995). "Scale issues in hydrological modelling: A review." In: *Hydrological Processes* 9, pp. 251–290. DOI: 10.1002/hyp.3360090305.

Blume, Theresa, Erwin Zehe, and Axel Bronstert (2007). "Rainfall-runoff response, event-based runoff coefficients and hydrograph separation." In: *Hydrological Sciences Journal* 52.5, pp. 843–862. DOI: 10.1623/hysj.52.5.843.

Böhm, O. and K.-F. Wetzel (2006). "Flood history of the Danube tributaries Lech and Isar in the Alpine foreland of Germany." In: *Hydrological Sciences Journal* 51.5, pp. 784–798. DOI: 10.1623/hysj.51.5.784.

Boorman, D B, J M Hollis, and A Lilly (1995). *Hydrology of soil types: a hydrologically-based classification of the soils of United Kingdom.* Tech. rep. 126, Institut of Hydrology, Wallingford, p. 137.

Boyle, D. P., H. V. Gupta, and S. Sorooshian (2000). "Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods." In: *Water Resources Research* 36.12, pp. 3663–3674. DOI: 10.1029/2000WR900207.

Bradley, A. A., P. J. Cooper, K. W. Potter, and T. Price (1996). "Floodplain mapping using continuous hydrologic and hydraulic simulation models." In: *Journal of Hydrologic Engineering* 1.2, pp. 63–68. DOI: 10.1061/(asce)1084-0699(1996)1:2(63).

Bravo, J. M., D. Allasia, a. R. Paz, W. Collischonn, and C. E. M. Tucci (2012). "Coupled Hydrologic-Hydraulic Modeling of the Upper Paraguay River Basin." In: *Journal of Hydrologic Engineering* 17.5, pp. 635–646. DOI: 10.1061/(ASCE)HE.1943-5584.0000494.

Brocca, L., F. Melone, T. Moramarco, and V. P. Singh (2009). "Assimilation of Observed Soil Moisture Data in Storm Rainfall-Runoff Modeling." In: *Journal of Hydrologic Engineering* 14.2, pp. 153–165. DOI: 10.1061/(ASCE)1084-0699(2009)14:2(153).

Brutsaert, Wilfried and John L. Nieber (1977). "Regionalized drought flow hydrographs from a mature glaciated plateau." In: *Water Resources Research* 13.3, pp. 637–643. DOI: 10.1029/WR013i003p00637.

Budyko, M I (1956). "The heat balance of the earth's surface." In: *Teplovoĭ balans zemnoĭ poverkhnosti.English* 2.4, 259 p.–. DOI: 10.1080/00385417.1961.10770761.

Bühner, M (2011). *Einführung in die Test- und Fragebogenkonstruktion*. Addison-Wesley Verlag, pp. 917–931. ISBN: 1853467960. DOI: 10.1126/science.1247727.

Capell, R., D. Tetzlaff, A. J. Hartley, and C. Soulsby (2012). "Linking metrics of hydrological function and transit times to landscape controls in a heterogeneous mesoscale catchment." In: *Hydrological Processes* 26.3, pp. 405–420. DOI: 10.1002/hyp.8139.

Carrillo, G., P. A. Troch, M. Sivapalan, T. Wagener, C. Harman, and K. Sawicz (2011). "Catchment classification: Hydrological analysis of catchment behavior through process-based modeling along a climate gradient." In: *Hydrology and Earth System Sciences* 15.11, pp. 3411–3430. DOI: 10.5194/hess-15-3411-2011.

Casper, M. C., G. Grigoryan, O. Gronz, O. Gutjahr, G. Heinemann, R. Ley, and A. Rock (2012). "Analysis of projected hydrological behavior of catchments based on signature indices." In: *Hydrology and Earth System Sciences* 16.2, pp. 409–421. DOI: 10.5194/hess-16-409-2012.

Chapman, Tom (1999). "A comparison of algorithms for stream flow recession and baseflow separation." In: *Hydrological Processes* 13.5, pp. 701–714. DOI: 10.1002/(SICI)1099-1085(19990415)13:5<701::AID-HYP774>3.0.CO;2-2.

Clark, Martyn P., Andrew G. Slater, David E. Rupp, Ross a. Woods, Jasper a. Vrugt, Hoshin V. Gupta, Thorsten Wagener, and Lauren E. Hay (2008). "Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models." In: *Water Resources Research* 44, pp. 1–14. DOI: 10.1029/2007WR006735.

Cloke, H L and F Pappenberger (2009). "Ensemble flood forecasting: A review." In: *J. Hydrol.* 375.3-4, pp. 613–626. DOI: DOI:10.1016/j.jhydrol.2009.06.005.

Cloke, H.L., M.G. Anderson, J.J. McDonnell, and J.-P. Renaud (2006). "Using numerical modelling to evaluate the capillary fringe groundwater ridging hypothesis of streamflow generation." In: *Journal of Hydrology* 316.1-4, pp. 141–162. DOI: 10.1016/j.jhydrol.2005.04.017.

Cloke, Hannah L. and Florian Pappenberger (2008). "Evaluating forecasts of extreme events for hydrological applications: An approach for screening unfamiliar performance measures." In: *Meteorological Applications* 15.1, pp. 181–197. DOI: 10.1002/met.58.

Collischonn, Walter, Daniel Allasia, Benedito C Da Silva, and Carlos E M Tucci (2007). "The MGB-IPH model for large-scale rainfall—runoff modelling." In: *Hydrological Sciences Journal* 52.5, pp. 878–895. DOI: 10.1623/hysj.52.5.878.

Crochemore, L (2011). *Evaluation of hydrological models: Expert judgement vs Numerical criteria*. Tech. rep. Cemagref, Hydrosystems and Bioprocesses Research Unit, final-year internship report, p. 62.

Crochemore, Louise, Charles Perrin, Vazken Andréassian, Uwe Ehret, Simon P Seibert, Salvatore Grimaldi, Hoshin Gupta, and Jean-Emmanuel Paturel (2014). "Comparing expert judgement and numerical criteria for hydrograph evaluation." In: *Hydrological Sciences Journal* 60.3, pp. 402–423. DOI: 10.1080/02626667.2014.903331.

Cunge, J. A. (1969). "On the subject of a flood propagation computation method (Musklngum Method)." In: *Journal of Hydraulic Research* 7.2, pp. 205–230. DOI: 10.1080/00221686909500264.

Dawson, C. W., R. J. Abrahart, and L. M. See (2007). "HydroTest: A web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts." In: *Environmental Modelling and Software* 22.7, pp. 1034–1052. DOI: 10.1016/j.envsoft.2006.06.008.

Dawson, C W, R J Abrahart, and L M See (2010). "Hydro Test: Further development of a web resource for the standardised assessment of hydrological models." In: *Environmental Modelling & Software* 25.11, pp. 1481–1482. DOI: 10.1016/j.envsoft.2009.01.001.

De Lannoy, G. J. M., Paul R. Houser, Valentijn R N Pauwels, and Niko E C Verhoest (2006). "Assessment of model uncertainty for soil moisture through ensemble verification." In: *Journal of Geophysical Research Atmospheres* 111.10, pp. 1–18. DOI: 10.1029/2005JD006367.

De Rooij, G. H. (2011). "Averaged water potentials in soil water and groundwater, and their connection to menisci in soil pores, field-scale flow phenomena, and simple groundwater flows." In: *Hydrology and Earth System Sciences* 15.5, pp. 1601–1614. DOI: 10.5194/hess-15-1601-2011.

Di Baldassarre, G. and A. Montanari (2009). "Uncertainty in river discharge observations: a quantitative analysis." In: *Hydrology and Earth System Sciences Discussions* 6.1, pp. 39–61. DOI: 10.5194/hessd-6-39-2009.

Dooge, J C I (1986). "Looking for hydrological laws." In: *Water Resources Research* 22.9S, 46S–58S.

Douglas, D H and T K Peucker (1973). "Algorithms for the reduction of the number of points required to represent a digitized line or its caricature." In: *Cartographica: The International Journal for Geographic Information and Geovisualization* 10.2, pp. 112–122. DOI: 10.3138/FM57-6770-U75U-7727.

Drabek, Ulrike (2010). "Anwendungsbezogene Aspekte der operationellen Durchflussvorhersage." PhD thesis. Institut für Wasserbau und Ingenieurhydrologie, Technische Universität Wien. ISBN: 9783852341132.

Duan, Q. et al. (2006). "Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops." In: *Journal of Hydrology* 320.1-2, pp. 3–17. DOI: 10.1016/j.jhydrol.2005.07.031.

Duan, Q., N. K. Ajami, X. Gao, and S Sorooshian (2007). "Multi-model ensemble hydrologic prediction using Bayesian model averaging." In: *Adv.Water Resour.* 30, pp. 1371–1386.

Dunne, Thomas and Richard D. Black (1970). "An Experimental Investigation of Runoff Production in Permeable Soils." In: *Water Resources Research* 6.2, pp. 478–490. DOI: 10.1029/WR006i002p00478.

Duscher, Klaus, Andreas Günther, Andrea Richts, Patrick Clos, Uta Philipp, and Wilhelm Struckmeier (2015). "The GIS layers of the "International Hydrogeological Map of Europe 1:1,500,000" in a vector format." In: *Hydrogeology Journal* 23.8, pp. 1867–1875. DOI: 10.1007/s10040-015-1296-4.

Eckhardt, K. (2005). "How to construct recursive digital filters for baseflow separation." In: *Hydrological Processes* 19.2, pp. 507–515. DOI: 10.1002/hyp.5675.

Efstratiadis, Andreas and Demetris Koutsoyiannis (2010). "One decade of multi-objective calibration approaches in hydrological modelling: a review." In: *Hydrological Sciences Journal* 55.1, pp. 58–78. DOI: 10.1080/02626660903526292.

Ehret, U (2016). "Structogram: A method to describe structuredness and complexity of data sets." In: *submitted to Mathematical Geosciences*.

Ehret, U. and S. P. Seibert (2016). "The Series Distance matlab code." In: *GIT repository, available at: https://github.com/KIT-HYD/SeriesDistance*. DOI: 10.5281/zenodo.60356.

Ehret, U. and E. Zehe (2011). "Series distance - An intuitive metric to quantify hydrograph similarity in terms of occurrence, amplitude and timing of hydrological events." In: *Hydrology and Earth System Sciences* 15.3, pp. 877–896. DOI: 10.5194/hess-15-877-2011.

Ehret, Uwe, Jens Götzinger, Andras Bardossy, and Geoffrey G.S. Pegram (2008). "Radar-based flood forecasting in small catchments, exemplified by the Goldersbach catchment, Germany." In: *International Journal of River Basin Management* 6.4, pp. 323–329. DOI: 10.1080/15715124.2008.9635359.

Euser, T., H. C. Winsemius, M. Hrachowitz, F. Fenicia, S. Uhlenbrook, and H. H G Savenije (2013). "A framework to assess the realism

of model structures using hydrological signatures." In: *Hydrology and Earth System Sciences* 17.5, pp. 1893–1912. DOI: 10.5194/hess-17-1893-2013.

Ewen, J (2011). "Hydrograph Matching Method for Measuring Model Performance." In: *Journal of Hydrology* 408.1-2, pp. 178–187.

Fenicia, F., H. H. G. Savenije, P. Matgen, and L. Pfister (2005). "Is the groundwater reservoir linear? Learning from data in hydrological modelling." In: *Hydrology and Earth System Sciences Discussions* 2.4, pp. 1717–1755. DOI: 10.5194/hessd-2-1717-2005.

Fiener, P and K Auerswald (2009). "Spatial Variability of rainfall on a sub-kilometer scale." In: *Earth Surface Processes and Landforms* 34, pp. 848–859.

Fiener, P., S. P. Seibert, and K. Auerswald (2011). "A compilation and meta-analysis of rainfall simulation data on arable soils." In: *Journal of Hydrology* 409.1-2, pp. 395–406. DOI: 10.1016/j.jhydrol.2011.08.034.

Fiener, P., K. Auerswald, F. Winter, and M. Disse (2013). "Statistical analysis and modelling of surface runoff from arable fields in central Europe." In: *Hydrology and Earth System Sciences* 17.10, pp. 4121–4132. DOI: 10.5194/hess-17-4121-2013.

Fischer, M (2008). "Ungesteuerte und gesteuerte Retention entlang von Fließgewässern." PhD thesis. Berichte des Lehrstuhls und der Versuchsanstalt für Wasserbau und Wasserwirtschaft, Bd. 119, Technischen Universität München.

Flügel, W.-A. (1996). "Hydrological Response Units (HRUs) as modeling entities for hydrological river basin simulation and their methodological potential for modeling complex environmental process systems." In: *Die Erde* 127, pp. 42–62.

Flügel, Wolfgang-Albert (1995). "Delineating hydrological response units by geographical information system analyses for regional hydrological modelling using PRMS/MMS in the drainage basin of the River Bröl, Germany." In: *Hydrological Processes* 9.3-4, pp. 423–436. DOI: 10.1002/hyp.3360090313.

Franz, K. J. and T. S. Hogue (2011). "Evaluating uncertainty estimates in hydrologic models: Borrowing measures from the forecast verification community." In: *Hydrology and Earth System Sciences* 15.11, pp. 3367–3382. DOI: 10.5194/hess-15-3367-2011.

Gassmann, M., C. Stamm, O. Olsson, J. Lange, K. Kümmerer, and M. Weiler (2013). "Model-based estimation of pesticides and transformation products and their export pathways in a headwater catchment." In: *Hydrology and Earth System Sciences* 17.12, pp. 5213–5228. DOI: 10.5194/hess-17-5213-2013.

Georgakakos, Konstantine P., Dong Jun Seo, Hoshin Gupta, John Schaake, and Michael B. Butts (2004). "Towards the characterization of streamflow simulation uncertainty through multimodel ensembles." In: *Journal of Hydrology* 298.1-4, pp. 222–241. DOI: 10.1016/j.jhydrol.2004.03.037.

Gneiting, Tilmann, I Stanberry Larissa, Eric P Grimit, Leonhard Held, and Nicholas A Johnson (2008). "Assessing probabilistic forecasts

of multivariate quantitites, with an application to ensemble pprediction of surface winds." In: *Test* 17.2, pp. 211–235.

Goodrich, David C., Jean-Marc Faurès, David a. Woolhiser, Leonard J. Lane, and Soroosh Sorooshian (1995). "Measurement and analysis of small-scale convective storm rainfall variability." In: *Journal of Hydrology* 173.1-4, pp. 283–308. DOI: 10.1016/0022-1694(95)02703-R.

Gottschalk, Lars (1985). "Hydrological regionalization of Sweden." In: *Hydrological Sciences Journal* 30.1, pp. 65–83. DOI: 10.1080/02626668509490972.

Graeff, T., E. Zehe, T. Blume, T. Francke, and B. Schröder (2012). "Predicting event response in a nested catchment with generalized linear models and a distributed watershed model." In: *Hydrological Processes* 26.24, pp. 3749–3769. DOI: 10.1002/hyp.8463.

Graeff, Thomas, Erwin Zehe, Dominik Reusser, Erika Lück, Boris Schröder, Gerald Wenk, Hermann John, and Axel Bronstert (2009). "Process identification through rejection of model structures in a mid-mountainous rural catchment: Observations of rainfall-runoff response, geophysical conditions and model inter-comparison." In: *Hydrological Processes* 23.5, pp. 702–718. DOI: 10.1002/hyp.7171.

Grayson, Rodger B., Andrew W. Western, Francis H. S. Chiew, and Günter Blöschl (1997). "Preferred states in spatial soil moisture patterns: Local and nonlocal controls." In: *Water Resources Research* 33.12, p. 2897. DOI: 10.1029/97WR02174.

Grigg, David (1965). "The Logic of Regional Systems." In: *Development* 55.3, pp. 465–491.

Gupta, Hoshin V., Thorsten Wagener, and Yuqiong Liu (2008). "Reconciling theory with observations: Elements of a diagnostic approach to model evaluation." In: *Hydrological Processes* 22.18, pp. 3802–3813. DOI: 10.1002/hyp.6989.

Gupta, Hoshin V., Harald Kling, Koray K. Yilmaz, and Guillermo F. Martinez (2009). "Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling." In: *Journal of Hydrology* 377.1-2, pp. 80–91. DOI: 10.1016/j.jhydrol.2009.08.003.

Gupta, Hoshin Vijai, Soroosh Sorooshian, and Patrice Ogou Yapo (1998). "Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information." In: *Water Resources Research* 34.4, p. 751. DOI: 10.1029/97WR03495.

Haag, I, S Vollmer, and S Heß (2005). *Aufstellung eines Wasserhaushaltsmodells für das Einzugsgebiet der Iller (Erläuterungsbericht)*. Tech. rep. Wasserwirtschaftsamt Kempten, (unpublished), p. 82.

Harmel, R. D., R. J. Cooper, R. M. Slade, R. L. Haney, and J. G. Arnold (2006). "Cumulative uncertainty in measured streamflow and water quality data for small watersheds." In: *Asabe* 49.3, pp. 689–701.

Harmel, R. D., P. K. Smith, K. W. Migliaccio, I. Chaubey, K. R. Douglas-Mankin, B. Benham, S. Shukla, R. Munoz-Carpena, and B. J. Robson (2014). "Evaluating, interpreting, and communicating perfor-

mance of hydrologic/water quality models considering intended use: A review and recommendations." In: *Environmental Modelling and Software* 57, pp. 40–51. DOI: 10.1016/j.envsoft.2014.02.013.

He, Y., A. Bárdossy, and E. Zehe (2011a). "A catchment classification scheme using local variance reduction method." In: *Journal of Hydrology* 411.1-2, pp. 140–154. DOI: 10.1016/j.jhydrol.2011.09.042.

He, Y., A. Bárdossy, and E. Zehe (2011b). "A review of regionalisation for continuous streamflow simulation." In: *Hydrology and Earth System Sciences* 15.11, pp. 3539–3553. DOI: 10.5194/hess-15-3539-2011.

He, Z. H., F. Q. Tian, H. V. Gupta, H. C. Hu, and H. P. Hu (2015). "Diagnostic calibration of a hydrological model in a mountain area by hydrograph partitioning." In: *Hydrology and Earth System Sciences* 19.4, pp. 1807–1826. DOI: 10.5194/hess-19-1807-2015.

Heggen, Richard J. (2001). "Normalized Antecedent Precipitation Index." In: *Journal of Hydrologic Engineering* 6.5, pp. 377–381. DOI: 10.1061/(ASCE)1084-0699(2001)6:5(377).

Hellebrand, Hugo, Rudi Van Den Bos, Lucien Hoffmann, Jerôme Juilleret, and Laurent Pfister (2008). "The potential of winter stormflow coefficients for hydrological regionalization purposes in poorly gauged basins of the middle Rhine region." In: *Hydrological Sciences Journal* 53.916417941, pp. 773–788. DOI: 10.1623/hysj.53.4.773.

Hillel, D. (2004). *Introduction to Environmental Soil Physics*. San Diego, California: Academic Press, Elsevier, p. 493.

Hingray, B., B. Schaefli, A. Mezghani, and Y. Hamdi (2010). "Signature-based model calibration for hydrological prediction in mesoscale Alpine catchments." In: *Hydrological Sciences Journal* 55.6, pp. 1002–1016. DOI: 10.1080/02626667.2010.505572.

Hohenrainer, J, M Hunger, G Moretti, C Elpers, and B Huth (2009). *Aufstellung eines Wasserhaushaltsmodells für das Einzugsgebiet des Lech (Erläuterungsbericht)*. Tech. rep. (unpublished), Wasserwirtschaftsamt Kempten, p. 82.

Horton, Robert E. (1939). "Analysis of runoff-plot experiments with varying infiltration-capacity." In: *Transactions, American Geophysical Union* 20.4, p. 693. DOI: 10.1029/TR020i004p00693.

Hrachowitz, M et al. (2013). "A decade of Predictions in Ungauged Basins (PUB) a review." In: *Hydrological Sciences Journal-Journal Des Sciences Hydrologiques* 58.6, pp. 1198–1255. DOI: 10.1080/02626667.2013.803183.

Hrachowitz, M., O. Fovet, L. Ruiz, T. Euser, S. Gharari, R. Nijzink, J. Freer, H. H G Savenije, and C. Gascuel-Odoux (2014). "Process consistency in models: The importance of system signatures, expert knowledge, and process complexity." In: *Water Resources Research* 50.9, pp. 7445–7469. DOI: 10.1002/2014WR015484.

Hundecha, Yeshewatesfa and András Bárdossy (2004). "Modeling of the effect of land use changes on the runoff generation of a river basin through parameter regionalization of a watershed model."

In: *Journal of Hydrology* 292.1-4, pp. 281–295. DOI: 10.1016/j.jhydrol.2004.01.002.

Jackisch, C. (2015). "Linking structure and functioning of hydrological systems. How to achieve necessary experimental and model complexity with adequate effort." PhD thesis. Institut für Wasser und Gewässerentwicklung, Bereich Hydrologie, Karlsruher Institut für Technologie (KIT). DOI: 10.5445/IR/1000051494.

Jacquin, Alexandra P. and Asaad Y. Shamseldin (2007). "Development of a possibilistic method for the evaluation of predictive uncertainty in rainfall-runoff modeling." In: *Water Resources Research* 43.4, W04425. DOI: 10.1029/2006WR005072.

Jury, W. A. and R. Horton (2004). *Soil Physics*. Hoboken, New Yersey: Wiley, p. 384.

Kelleher, C., T. Wagener, and B. L. McGlynn (2015). "Model-based analysis of the influence of catchment properties on hydrologic partitioning across five mountain headwater sub-catchments." In: *Water Resour. Res.* 51.6, pp. 4109–4136. DOI: DOI:10.1002/2014WR016147.

Kenney, Bernard C. (1982). "Beware of spurious self-correlations!" In: *Water Resources Research* 18.4, pp. 1041–1048. DOI: 10.1029/WR018i004p01041.

Kim, Jongho, April Warnock, Valeriy Y. Ivanov, and Nikolaos D. Katopodes (2012). "Coupled modeling of hydrologic and hydrodynamic processes including overland and channel flow." In: *Advances in Water Resources* 37, pp. 104–126. DOI: 10.1016/j.advwatres.2011.11.009.

Kirby, M.J. (1975). "Hydrograph modelling strategies." In: *Progress in Physical and Human Geograph*. Ed. by R. Peel. London: Heinemann, pp. 69–90. ISBN: 9780435356255.

Kirchner, James W. (2006). "Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology." In: *Water Resources Research* 42.3, 5 pp. DOI: 10.1029/2005WR004362.

Kirchner, James W. (2009). "Catchments as simple dynamical systems: Catchment characterization, rainfall-runoff modeling, and doing hydrology backward." In: *Water Resources Research* 45.2, pp. 1–34. DOI: 10.1029/2008WR006912.

Klaus, Julian, Erwin Zehe, Martin Elsner, Juliane Palm, Dorothee Schneider, Boris Schröder, Sibylle Steinbeiss, Loes van Schaik, and Stephanie West (2014). "Controls of event-based pesticide leaching in natural soils: A systematic study based on replicated field scale irrigation experiments." In: *Journal of Hydrology* 512, pp. 528–539. DOI: 10.1016/j.jhydrol.2014.03.020.

Klaus, Julian, Carlos E. Wetzel, N. Martinez-Carreras, Luc Ector, and Laurent Pfister (2015). "A tracer to bridge the scales: On the value of diatoms for tracing fast flow path connectivity from headwaters to meso-scale catchments." In: *Hydrological Processes* 29.25, pp. 5275–5289. DOI: 10.1002/hyp.10628.

Kleidon, Axel (2012). "How does the earth system generate and maintain thermodynamic disequilibrium and what does it imply for the future of the planet?" In: *Philosophical Transactions of the Royal*

Society A: Mathematical, Physical and Engineering Sciences 370.1962, pp. 1012–1040. DOI: 10.1098/rsta.2011.0316.

Kneis, D. and M. Heistermann (2008). "Bewertung der Güte einer Radar-basierten Niederschlagsschätzung am Beispiel eines kleinen Einzugsgebiets." In: *Hydrologie und Wasserbewirtschaftung* 53.3, pp. 160–171.

Knijff, J. M. van der, J. Younis, and A. P. J. De Roo (2010). "LISFLOOD: a GIS-based distributed model for river basin scale water balance and flood simulation." In: *Int. J. Geogr. Inf. Sci.* 24.2, pp. 189–212. DOI: 10.1080/13658810802549154.

Kollat, J. B., P. M. Reed, and T. Wagener (2012). "When are multiobjective calibration trade-offs in hydrologic models meaningful?" In: *Water Resources Research* 48.3, W03520. DOI: 10.1029/2011WR011534.

Krzysztofowicz, Roman (1999). "Bayesian forecasting via deterministic model." In: *Risk Analysis* 19.4, pp. 739–749. DOI: 10.1023/A:1007050023440.

Krzysztofowicz, Roman and Karen S. Kelly (2000). "Hydrologic uncertainty processor for probabilistic river stage forecasting." In: *Water Resources Research* 36.11, p. 3265. DOI: 10.1029/2000WR900108.

Labadie, John W. (2004). "Optimal Operation of Multireservoir Systems: State-of-the-Art Review." In: *Journal of Water Resources Planning and Management* 130.2, pp. 93–111. DOI: 10.1061/(ASCE)0733-9496(2004)130:2(93).

Laio, F. and S. Tamea (2007). "Verification tools for probabilistic forecasts of continuous hydrological variables." In: *Hydrology and Earth System Sciences* 11.4, pp. 1267–1277. DOI: 10.5194/hessd-3-2145-2006.

Laurenson, E. M. (1964). "A catchment storage model for runoff routing." In: *Journal of Hydrology* 2, pp. 141–163. DOI: 10.1016/0022-1694(64)90025-3.

Leavesley, G H (1973). "A mountain watershed simulation model." PhD thesis. Fort Collins: Colorado State University, p. 174.

Legates, David R. and Gregory J. McCabe (1999). "Evaluating the use of 'goodness-of-fit' measures in hydrologic and hydroclimatic model validation." In: *Water Resources Research* 35.1, pp. 233–241. DOI: 10.1029/1998WR900018.

Lehmann, P, C Hinz, G McGrath, H J Tromp-van Meerveld, and J J McDonnell (2007). "Rainfall threshold for hillslope outflow: an emergent property of flow pathway connectivity." In: *Hydrology and Earth System Sciences* 11.2, pp. 1047–1063. DOI: 10.5194/hessd-3-2923-2006.

Ley, R, M C Casper, H Hellebrand, and R Merz (2011). "Catchment classification by runoff behaviour with self-organizing maps (SOM)." In: *Hydrology and Earth System Sciences* 15.9, pp. 2947–2962. DOI: 10.5194/hess-15-2947-2011.

Li, Hongyi, Murugesu Sivapalan, and Fuqiang Tian (2012). "Comparative diagnostic analysis of runoff generation processes in Oklahoma DMIP2 basins: The Blue River and the Illinois River." In:

Journal of Hydrology 418-419, pp. 90–109. DOI: 10.1016/j.jhydrol.2010.08.005.

Lindsay, J. B. (2014). "The Whitebox Geospatial Analysis Tools project and open-access GIS," in: *Proceedings of the GIS Research UK 22nd Annual Conference*, p. 8. ISBN: 5198244120. DOI: 10.13140/RG.2.1.1010.8962.

Liu, Yuqiong, James Brown, Julie Demargne, and Dong Jun Seo (2011). "A wavelet-based approach to assessing timing errors in hydrologic predictions." In: *Journal of Hydrology* 397.3-4, pp. 210–224. DOI: 10.1016/j.jhydrol.2010.11.040.

Loritz, Ralf, Sibylle K. Hassler, Conrad Jackisch, Niklas Allroggen, Loes van Schaik, Jan Wienhöfer, and Erwin Zehe (2016). "Picturing and modelling catchments by representative hillslopes." In: *Hydrology and Earth System Sciences Discussions*, pp. 1–56. DOI: 10.5194/hess-2016-307.

Ludwig, K. (1978). "Systematische Berechnung von Hochwasser-Abflussvorgängen mit Flussgebietsmodellen." PhD thesis. Mitteilungen des Institutes für Hydrologie, Wasserwirtschaft und landwirtschaftlichen Wasserbau Bd. 44. Technischen Universität Hannover, pp. 263–462.

Ludwig, K. (1982). "The program system FGMOD for calculation of flood runoff processes in river basins." In: *Z. f. Kulturtechnik und Flurbereinigung* 23, pp. 25–37.

Ludwig, K. and M. Bremicker (2006). *The Water Balance Model LARSIM*. Tech. rep. 22. Institut für Hydrologie der Universität Freiburg i.Br.

Maidment, D. (1993). *Handbook of Hydrology*. New York, NY: McGraw-Hill Education, p. 1424. ISBN: 0824754735.

Marka, Ole, Sutat Weesakula, Chusit Apirumanekula, Surajate Boonya Aroonneta, and Slobodan Djordjević (2004). "Potential and limitations of 1D modelling of urbanflooding." In: *Journal of Hydrology* 299.3-4, pp. 284–299. DOI: 10.1016/j.jhydrol.2001.08.014.

Markstrom, S. L., L. E. Hay, and M. P. Clark (2016). "Towards simplification of hydrologic modeling: identification of dominant processes." In: *Hydrology and Earth System Sciences Discussions*, pp. 1–33. DOI: doi:10.5194/hess-2015-508.

Martínez-Carreras, N., Andreas Krein, Francesc Gallart, Jean F. Iffly, Laurent Pfister, Lucien Hoffmann, and Philip N. Owens (2010). "Assessment of different colour parameters for discriminating potential suspended sediment sources and provenance: A multi-scale study in Luxembourg." In: *Geomorphology* 118.1-2, pp. 118–129. DOI: 10.1016/j.geomorph.2009.12.013.

Martínez-Carreras, N., C. E. Wetzel, J. Frentress, L. Ector, J. J. McDonnell, L. Hoffmann, and L. Pfister (2015). "Hydrological connectivity inferred from diatom transport through the riparian-stream system." In: *Hydrology and Earth System Sciences* 19.7, pp. 3133–3151. DOI: 10.5194/hess-19-3133-2015.

Matott, L S, J E Babendreier, and S T Purucker (2009). "Evaluating uncertainty in integrated environmental models: A review of con-

cepts and tools." In: *Water Resour. Res.* 45, pp. 1–14. DOI: 10.1029/2008WR007301.

Maurer, Thomas. (1997). "Physikalisch begründete zeitkontinuierliche Modellierung des Wassertransports in kleinen ländlichen Einzugsgebieten." PhD thesis. Mitteilungen des Instituts für Hydrologie und Wasserwirtschaft, Bd. 61, Universitat Fridericiana zu Karlsruhe (TH).

Mauser, Wolfram and Heike Bach (2009). "PROMET - Large scale distributed hydrological modelling to study the impact of climate change on the water flows of mountain watersheds." In: *Journal of Hydrology* 376.3-4, pp. 362–377. DOI: 10.1016/j.jhydrol.2009.07.046.

McCarthy, G T (1938). "The Unit Hydrograph and Flood Routing." In: *Conf. of the North Atlantic Div. U. S. Engineer Department*. New London, Conneticut.

McGuire, K. J., J. J. McDonnell, M. Weiler, C. Kendall, B. L. McGlynn, J. M. Welker, and J. Seibert (2005). "The role of topography on catchment-scale water residence time." In: *Water Resources Research* 41.5, pp. 1–14. DOI: 10.1029/2004WR003657.

McMillan, Hilary K., Martyn P. Clark, William B. Bowden, Maurice Duncan, and Ross A. Woods (2011a). "Hydrological field data from a modeller's perspective: Part 1. Diagnostic tests for model structure." In: *Hydrological Processes* 25.4, pp. 511–522. DOI: 10.1002/hyp.7841.

McMillan, Hilary, Bethanna Jackson, Martyn Clark, Dmitri Kavetski, and Ross Woods (2011b). "Rainfall uncertainty in hydrological modelling: An evaluation of multiplicative error models." In: *Journal of Hydrology* 400.1-2, pp. 83–94. DOI: 10.1016/j.jhydrol.2011.01.026.

McMillan, Hilary, Myriam Gueguen, Elisabeth Grimon, Ross Woods, Martyn Clark, and David E. Rupp (2014). "Spatial variability of hydrological processes and model structure diagnostics in a 50 km 2 catchment." In: *Hydrological Processes* 28.18, pp. 4896–4913. DOI: 10.1002/hyp.9988.

McNamara, James P., Doerthe Tetzlaff, Kevin Bishop, Chris Soulsby, Mark Seyfried, Norman E. Peters, Brent T. Aulenbach, and Richard Hooper (2011). "Storage as a Metric of Catchment Comparison." In: *Hydrological Processes* 25.21, pp. 3364–3371. DOI: 10.1002/hyp.8113.

Melsen, L. A., A. J. Teuling, P. J. J. F. Torfs, R. Uijlenhoet, N. Mizukami, and M. P. Clark (2016). "HESS Opinions: The need for process-based evaluation of large-domain hyper-resolution models." In: *Hyrology and Earth System Sciences* 20.3, pp. 1069–1079. DOI: 10.5194/hess-20-1069-2016.

Menzel, Annette, Gert Jakobi, Rein Ahas, Helfried Scheifinger, and Nicole Estrella (2003). "Variations of the climatological growing season (1951-2000) in Germany compared with other countries." In: *International Journal of Climatology* 23.7, pp. 793–812. DOI: 10.1002/joc.915.

Merz, R. (2003). "A process typology of regional floods." In: *Water Resources Research* 39.12, pp. 1–20. DOI: 10.1029/2002WR001952.

Merz, R., G. Blöschl, and J. Parajka (2006). "Spatio-temporal variability of event runoff coefficients." In: *Journal of Hydrology* 331.3-4, pp. 591–604. DOI: 10.1016/j.jhydrol.2006.06.008.

Merz, Ralf and Günter Blöschl (2004). "Regionalisation of catchment model parameters." In: *Journal of Hydrology* 287.1-4, pp. 95–123. DOI: 10.1016/j.jhydrol.2003.09.028.

Merz, Ralf and Günter Blöschl (2009). "A regional analysis of event runoff coefficients with respect to climate and catchment characteristics in Austria." In: *Water Resources Research* 45.1, pp. 1–19. DOI: 10.1029/2008WR007163.

Mikhailova, M. V., V. N. Mikhailov, and V. N. Morozov (2012). "Extreme hydrological events in the Danube River basin over the last decades." In: *Water Resources* 39.2, pp. 161–179. DOI: 10.1134/S0097807812010095.

Milly, P C D, Julio Betancourt, Malin Falkenmark, Robert M Hirsch, Z. W. Kundzewicz, Dennis P Lettenmaier, and Ronald J Stouffer (2008). "Stationarity Is Dead: Whither Water Management?" In: *Science* 319.5863, pp. 573–574. DOI: 10.1126/science.1151915.

Montanari, Alberto (2007). "What do we mean by 'uncertainty'? The need for a consistent wording about uncertainty assessment in hydrology." In: *Hydrological Processes* 21.6, pp. 841–845. DOI: 10.1002/hyp.6623.

Montanari, Alberto and Giovanna Grossi (2008). "Estimating the uncertainty of hydrological forecasts: A statistical approach." In: *Water Resources Research* 44.12, pp. 1–9. DOI: 10.1029/2008WR006897.

Montanari, Alberto, Renzo Rosso, and Murad S. Taqqu (1997). "Fractionally differenced ARIMA models applied to hydrologic time series: Identification, estimation, and simulation." In: *Water Resources Research* 33.5, p. 1035. DOI: 10.1029/97WR00043.

Moriasi, D.N., J.G. Arnold, M.W. Van Liew, R.L. Binger, R.D. Harmel, and T.L. Veith (2007). "Model evaluation guidelines for systematic quantification of accuracy in watershed simulations." In: *Transactions of the ASABE* 50.3, pp. 885–900. DOI: 10.13031/2013.23153.

Mosler, Karl (2013). "Depth statistics." In: *Robustness and Complex Data Structures*. Ed. by C. Becker, R. Fried, and S. Kuhnt. Berlin Heidelberg: Springer-Verlag, pp. 17–35. ISBN: 978-3-642-35493-9. DOI: 10.1007/978-3-642-35494-6.

NOAA (1972). *National Weather Service River Forcast System Forecast Procedures*. Tech. rep. NWS-Hydro-14, NOAA Technical Memorandum. US Department of Commerce, Washington, D.C.

Nash, J. E. and J. V. Sutcliffe (1970). "River flow forecasting through conceptual models part I - A discussion of principles." In: *Journal of Hydrology* 10.3, pp. 282–290. DOI: 10.1016/0022-1694(70)90255-6.

Nasseri, M., A. Ansari, and B. Zahraie (2014). "Uncertainty assessment of hydrological models with fuzzy extension principle: Evaluation of a new arithmetic operator." In: *Water Resources Research* 50.2, pp. 1095–1111. DOI: 10.1002/2012WR013382.

Nester, Thomas, Robert Kirnbauer, Dieter Gutknecht, and G. Blöschl (2011). "Climate and catchment controls on the performance of regional flood simulations." In: *Journal of Hydrology* 402.3-4, pp. 340–356. DOI: 10.1016/j.jhydrol.2011.03.028.

Neuweiler, I. and P.R. King (2002). "Coarse graining of the solute concentration probability distribution for advective transport in porous media." In: *Proceedings of the 14. International Conference on Computational Methods in Water Resources*. Ed. by S. M. Hassanizadeh, R. J. Schotting, W. G. Gray, and G. F. Pinder. Delft: Elsevier Science Publishers B.V., pp. 1147–1154.

Nicholas, A. P. and C. A. Mitchell (2003). "Numerical simulation of overbank processes in topographically complex floodplain environments." In: *Hydrological Processes* 17.4, pp. 727–746. DOI: 10.1002/hyp.1162.

Niehoff, Daniel, Uta Fritsch, and Axel Bronstert (2002). "Land-use impacts on storm-runoff generation: Scenarios of land-use change and simulation of hydrological response in a meso-scale catchment in SW-Germany." In: *Journal of Hydrology* 267.1-2, pp. 80–93. DOI: 10.1016/S0022-1694(02)00142-7.

Nippgen, Fabian, Brian L. McGlynn, and Ryan E. Emanuel (2015). "The spatial and temporal evolution of contributing areas." In: *Water Resources Research* 51.6, pp. 4550–4573. DOI: 10.1002/2014WR016719.

Norbiato, Daniele, Marco Borga, Ralf Merz, Günther Blöschl, and Alberto Carton (2009). "Controls on event runoff coefficients in the eastern Italian Alps." In: *Journal of Hydrology* 375.3-4, pp. 312–325. DOI: 10.1016/j.jhydrol.2009.06.044.

Nujic, M. (2003). *HYDRO AS-2D. Ein zweidimensionales Strömungsmodell für die wasserwirtschaftliche Praxis*. Tech. rep. Rosenheim.

Ogden, F. L., H. O. Sharif, S. U S Senarath, J. A. Smith, M. L. Baeck, and J. R. Richardson (2000). "Hydrologic analysis of the Fort Collins, Colorado, flash flood of 1997." In: *Journal of Hydrology* 228.1-2, pp. 82–100. DOI: 10.1016/S0022-1694(00)00146-3.

Olden, J D and N L Poff (2003). "Redundancy and the choice of hydrologic indices for characterizing streamflow regimes." In: *River Research and Applications* 19.2, pp. 101–121. DOI: 10.1002/rra.700.

Opan, M (2011). "Real-Time Optimal Operation of Multiple Reservoirs System." In: *Teknik Dergi* 22.2, pp. 5359–5385.

Pachepsky, Yakov, Andrey Guber, Diederik Jacques, Jiri Simunek, Marthinus Th. Van Genuchten, Thomas Nicholson, and Ralph Cady (2006). "Information content and complexity of simulated soil water fluxes." In: *Geoderma* 134.3-4, pp. 253–266. DOI: 10.1016/j.geoderma.2006.03.003.

Paiva, Rodrigo C D, Walter Collischonn, and Diogo Costa Buarque (2013). "Validation of a full hydrodynamic model for large-scale hydrologic modelling in the Amazon." In: *Hydrological Processes* 27.3, pp. 333–346. DOI: 10.1002/hyp.8425.

Pall, Karin and Georg A Janauer (2003). "Impoundment Hoechstaedt (river-km 2538.0-2530.8) in the upper reach of the Danube River in Germany." In: *Archiv f. Hydrobiol. Suppl. Vol. Large Rivers* 147.1-2, pp. 55–64.

Palm, Juliane, N. Loes M B van Schaik, and Boris Schröder (2013). "Modelling distribution patterns of anecic, epigeic and endogeic earthworms at catchment-scale in agro-ecosystems." In: *Pedobiologia* 56.1, pp. 23–31. DOI: 10.1016/j.pedobi.2012.08.007.

Pappenberger, F, K J Beven, N M Hunter, P D Bates, B T Gouweleeuw, J Thielen, and A P J Roo (2005). "Cascading model uncertainty from medium range weather forecasts (10 days) through a rainfall-runoff model to flood inundation predictions within the European Flood Forecasting System (EFFS)." In: *Hydrology and Earth System Sciences* 9.4, pp. 381–393. DOI: 10.5194/hess-9-381-2005.

Pappenberger, Florian and Keith J. Beven (2004). "Functional classification and evaluation of hydrographs based on Multicomponent Mapping (Mx)." In: *International Journal of River Basin Management* 2.2, pp. 89–100. DOI: 10.1080/15715124.2004.9635224.

Pappenberger, Florian, Patrick Matgen, Keith J. Beven, Jean Baptiste Henry, Laurent Pfister, and Paul Fraipont (2006). "Influence of uncertain boundary conditions and model structure on flood inundation predictions." In: *Advances in Water Resources* 29.10, pp. 1430–1449. DOI: 10.1016/j.advwatres.2005.11.012.

Pearce, A. J., M. K. Stewart, and M. G. Sklash (1986). "Storm Runoff Generation in Humid Headwater Catchments: 1. Where Does the Water Come From?" In: *Water Resources Research* 22.8, pp. 1263–1272. DOI: 10.1029/WR022i008p01263.

Pearson, Karl (1987). "Mathematical Contributions to the Theory of Evolution. - On a Form of Spurious Correlation Which May Arise When Indices Are Used in the Measurement of Organs." In: *Proceedings of the Royal Society of London* 60, pp. 489–498. DOI: 10.1098/rspl.1896.0076.

Pegelvorschrift (1991). *Anlage D Richtlinie für das Messen und Ermitteln von Abflüssen und Durchflüssen*. Länderarbeitsgemeinschaft Wasser (LAWA) und Bundesministerium für Verkehr (BMV) (Eds.)

Pelletier, Jon D. and Craig Rasmussen (2009). "Geomorphically based predictive mapping of soil thickness in upland watersheds." In: *Water Resources Research* 45.9, n/a–n/a. DOI: 10.1029/2008WR007319.

Perng, C.-S., H. Wang, S.R. Zhang, and D.S. Parker (2000). "Landmarks: a new model for similarity-based pattern querying in time series databases." In: *Proceedings of 16th International Conference on Data Engineering (Cat. No.00CB37073), San Diego, CA*. Pp. 33–42. DOI: 10.1109/ICDE.2000.839385.

Peschke, G (1985). "Zur Bildung und Berechnung von Regenabfluss." In: *Wissenschaftliche Zeitschrift der Technischen Universität Dresden* 34.4, pp. 195–200.

Peschke, G, C Etzenberg, G Müller, J Töpfer, and S Zimmermann (1999). *Das wissensbasierte System FLAB - ein Instrument zur rechnergestützten Bestimmung von Landschaftseinheiten mit gleicher Abflussbilung*. Tech. rep. Zittau: Internationales Hochschulinstitut Zittau, IHI-Schriften, BD. 10, p. 122.

Pfannerstill, Matthias, Björn Guse, and Nicola Fohrer (2014). "Smart low flow signature metrics for an improved overall performance

evaluation of hydrological models." In: *Journal of Hydrology* 510, pp. 447–458. DOI: 10.1016/j.jhydrol.2013.12.044.

Pfister, L, J F Iffly, and L Hoffmann (2002). "Use of regionalized stormflow coefficients with a view to hydroclimatological hazard mapping." In: *Hydrological Sciences Journal-Journal Des Sciences Hydrologiques* 47.3, pp. 479–491. DOI: 10.1080/02626660209492948.

Pfister, L., G. Drogue, A. El Idrissi, J. Humbert, J.F. Iffly, P. Matgen, and L. Hoffmann (2003). "Predicting peak discharge through empirical relationships between rainfall, groundwater level and basin humidity in the Alzette river basin (grand-duchy of Luxembourg)." In: *Journal of Hydrology and Hydromechanics* 51.3, pp. 210–220.

R Core Team (2015). *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*. Vienna.

Raje, Deepashree and P. P. Mujumdar (2010). "Reservoir performance under uncertainty in hydrologic impacts of climate change." In: *Advances in Water Resources* 33.3, pp. 312–326. DOI: 10.1016/j.advwatres.2009.12.008.

Ramirez, J. A. (2000). "Prediction and Modeling of Flood Hydrology and Hydraulics." In: *Inland Flood Hazards: Human, Riparian and Aquatic Communities*. Ed. by E Wohl. Cambridge University Press, pp. 1–52.

Refsgaard, J C (1997). "Validation and intercomparision of different updating procedure for real time flood forecasting." In: *Nordic Hydrology* 28.2, pp. 65–84.

Reggiani, Paolo, Murugesu Sivapalan, and S. Majid Hassanizadeh (2000). "Conservation equations governing hillslope responses: Exploring the physical basis of water balance." In: *Water Resources Research* 36.7, p. 1845. DOI: 10.1029/2000WR900066.

Reusser, D. E. and E. Zehe (2011). "Inferring model structural deficits by analyzing temporal dynamics of model performance and parameter sensitivity." In: *Water Resources Research* 47.7, W07550. DOI: 10.1029/2010WR009946.

Reusser, D. E., T. Blume, B. Schaefli, and E. Zehe (2009). "Analysing the temporal dynamics of model performance for hydrological models." In: *Hydrology and Earth System Sciences* 13.7, pp. 999–1018. DOI: 10.5194/hess-13-999-2009.

Reusser, Dominik E (2010). "Combining smart model diagnostics and effective data collection for snow catchments." PhD Thesis. University of Potsdam, Germany.

Rimböck, Andreas, Marc-Daniel Heintz, Karin Henning, Thomas Henschel, Wolfgang Kleber-Lerchbaumer, Uwe Kraier, Gabriele Merz, Martin Schmid, Gudrun Seidel, and Frank Wilhelm (2014). *Hochwasserschutz, Aktionsprogramm 2020plus*. Tech. rep. München: Bayerisches Staatsministerium für Umwelt und Verbraucherschutz, p. 56.

Rodríguez-Iturbe, Ignacio and Juan B. Valdés (1979). "The geomorphologic structure of hydrologic response." In: *Water Resources Research* 15.6, pp. 1409–1420. DOI: 10.1029/WR015i006p01409.

Rooij, GH (2009). "Averaging hydraulic head, pressure head, and gravitational head in subsurface hydrology, and implications for

averaged fluxes, and hydraulic conductivity." In: *Hydrology and Earth System Sciences* 13.7, pp. 1123–1132. DOI: 10.5194/hess-13-1123-2009.

Rösl, G, H J Rosemann, and J Stockenhuber (1984). "Nachweis der Speicherwirkungen auf den Hochwasserabfluß in der bayerischen Donau bis Kelheim mit Hilfe eines Flußgebietsmodells." In: *Internationales Symposion Interpraevent*. Villach, pp. 197–215.

Roulston, M. S. and L. A. Smith (2003). "Combining dynamical and statistical ensembles." In: *Tellus, Series A: Dynamic Meteorology and Oceanography* 55.1, pp. 16–30. DOI: 10.1034/j.1600-0870.2003.201378.x.

Ruiz-Villanueva, V., M. Borga, D. Zoccatelli, L. Marchi, E. Gaume, and U. Ehret (2012). "Extreme flood response to short-duration convective rainfall in South-West Germany." In: *Hydrology and Earth System Sciences* 16.5, pp. 1543–1559. DOI: 10.5194/hess-16-1543-2012.

Rupp, D E and J S Selker (2006). "Information, artifacts, and noise in dQ/dt-Q recession analysis." In: *Advances in Water Resources* 29.2, pp. 154–160.

Santhi, C., P. M. Allen, R. S. Muttiah, J. G. Arnold, and P. Tuppad (2008). "Regional estimation of base flow for the conterminous United States by hydrologic landscape regions." In: *Journal of Hydrology* 351.1-2, pp. 139–153. DOI: 10.1016/j.jhydrol.2007.12.018.

Sawicz, K., T. Wagener, M. Sivapalan, P. A. Troch, and G. Carrillo (2011). "Catchment classification: Empirical analysis of hydrologic similarity based on catchment function in the eastern USA." In: *Hydrology and Earth System Sciences* 15.9, pp. 2895–2911. DOI: 10.5194/hess-15-2895-2011.

Sayama, Takahiro, Jeffrey J. McDonnell, Amod Dhakal, and Kate Sullivan (2011). "How much water can a watershed store?" In: *Hydrological Processes* 25.25, pp. 3899–3908. DOI: 10.1002/hyp.8288.

Schaake, J., Q. Duan, M. Smith, and V. Koren (2000). *Criteria to select basins for hydrologic model development and testing*. 15th Conference on Hydrology (P1.8), AMS, January 9-14. Long Beach, CA, pp. 10–14.

Schaap, M. G., F. J. Leij, and M. T. van Genuchten (2001). "Rosetta: A computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions." In: *Journal of Hydrology* 251.3-4, pp. 163–176. DOI: 10.1016/S0022-1694(01)00466-8.

Schaefli, B., C. J. Harman, M. Sivapalan, and S. J. Schymanski (2011). "HESS Opinions: Hydrologic predictions in a changing environment: Behavioral modeling." In: *Hydrology and Earth System Sciences* 15.2, pp. 635–646. DOI: 10.5194/hess-15-635-2011.

Schaefli, Bettina and Hoshin V. Gupta (2007). "Do Nash values have value?" In: *Hydrological Processes* 21.15, pp. 2075–2080. DOI: 10.1002/hyp.6825.

Schaik, Loes van, Juliane Palm, Julian Klaus, Erwin Zehe, and Boris Schröder (2014). "Linking spatial earthworm distribution to macro-

pore numbers and hydrological effectiveness." In: *Ecohydrology* 7.2, pp. 401–408. DOI: 10.1002/eco.1358.

Scherrer, Simon and Felix Naef (2003). "A decision scheme to indicate dominant hydrological flow processes on temperate grassland." In: *Hydrological Processes* 17.2, pp. 391–401. DOI: 10.1002/hyp.1131.

Schmocker-Fackel, P, F Naef, and S Scherrer (2007). "Identifying runoff processes on the plot and catchment scale." In: *Hydrology and Earth System Sciences* 11.2, pp. 891–906. DOI: 10.5194/hess-11-891-2007.

Schulla, J (1997). "Hydrologische Modellierung von Flussgebieten zur Abschätzung der Folgen von Klimaäderungen." PhD thesis. Zürcher Geographische Schriften, Bd. 69, ETH Zürich, p. 161. DOI: 10.3929/ethz-a-001763261.

Seibert, J (2001). "On the need for benchmarks in hydrological modelling." In: *Hydrological Processes* 15.6, pp. 1063–1064. DOI: 10.1002/hyp.446.

Seibert, S P and U Ehret (2012). "Detection of flood events in hydrological discharge time series (EGU2012-5924)." In: *Geophysical Research Abstracts*. Vol. 14. 5, pp. 14–15.

Seibert, S. P., D. Skublics, and U. Ehret (2014). "The potential of coordinated reservoir operation for flood mitigation in large basins - A case study on the Bavarian Danube using coupled hydrological-hydrodynamic models." In: *Journal of Hydrology* 517.0, pp. 1128–1144. DOI: 10.1016/j.jhydrol.2014.06.048.

Seibert, S P, C Jackisch, L Pfister, U Ehret, and E Zehe (2016). "Exploring the interplay between state, structure and runoff behavior of lower mesoscale catchments." In: *Hydrol. Earth Syst. Sci. Discuss.* Pp. 1–51. DOI: 10.5194/hess-2016-109.

Seibert, Simon Paul, Uwe Ehret, and Erwin Zehe (2016). "Disentangling timing and amplitude errors in streamflow simulations." In: *Hydrology and Earth System Sciences* 20, pp. 3745–3763. DOI: 10.5194/hess-20-3745-2016.

Shaw, Stephen B. and Susan J. Riha (2012). "Examining individual recession events instead of a data cloud: Using a modified interpretation of dQ/dt-Q streamflow recession in glaciated watersheds to better inform models of low flow." In: *Journal of Hydrology* 434-435.0, pp. 46–54. DOI: 10.1016/j.jhydrol.2012.02.034.

Sherman, L K (1932). "Streamflow from rainfall by the unit graph method." In: *Eng. News Rec* 108, pp. 501–505.

Shrestha, D. L., N. Kayastha, and D. P. Solomatine (2009). "A novel approach to parameter uncertainty analysis of hydrological models using neural networks." In: *Hydrology and Earth System Sciences* 13.7, pp. 1235–1248. DOI: 10.5194/hess-13-1235-2009.

Skublics, D. (2014). "Großräumige Hochwassermodellierung im Einnzugsgebiet der bayerischen Donau." PhD thesis. Berichte des Lehrstuhls und der Versuchsanstalt für Wasserbau und Wasserwirtschaft, Bd. 131, Technische Universität München.

Skublics, D, M Fischer, and D Rutschmann (2009). "Numerical investigation on natural flood retention at the Bavarian Danube." In:

33rd International Association of Hydraulic Engineering & Research (IAHR) Congress 9.-14. Aug 2009 Vancouver. Canada. ISBN: 978-90-78046-08-0.

Skublics, D, S P Seibert, and U Ehret (2014). "Abbildung der Hochwasserretention durch hydrologische und hydrodynamische Modelle unter unterschiedlichen Randbedingungen. Sensitivitätsanalyse am Donauabschnitt zwischen Neu-Ulm und Donauwörth." In: *Hydrologie und Wasserbewirtschaftung* 3.58, pp. 178–189. DOI: 10.5675/HyWa_2014,3_2.

Skublics, D., Simon P Seibert, Uwe Ehret, P. Rutschmann, and E. Zehe (2013). *Flussgebietsweite operationelle Steuerung der Abflüsse im Extrembereich*. Tech. rep. München: TU München, Karlsruher Institut für Technologie, p. 234.

Smith, M B, K P Georgakakos, and X Liang (2004). "The distributed model intercomparison project (DMIP)." In: *Journal of Hydrology* 298, pp. 1–3. DOI: 10.1016/S0022-1694(04)00353-1.

Soares, Alexandre Kepler, Dídia I. C. Covas, and Luisa Fernanda R. Reis (2011). "Leak detection by inverse transient analysis in an experimental PVC pipe system." In: *Journal of Hydroinformatics* 13.2, p. 153. DOI: 10.2166/hydro.2010.012.

Soulsby, C., D. Tetzlaff, and M. Hrachowitz (2009). "Tracers and transit times: windows for viewing catchment storage?" In: *Hydrological Processes* 23. DOI: 10.1002/hyp.7501.

Spence, C. (2007). "On the relation between dynamic storage and runoff: A discussion on thresholds, efficiency, and function." In: *Water Resources Research* 43.12, pp. 1–11. DOI: 10.1029/2006WR005645.

Stedinger, Jery R., Richard M. Vogel, Seung Uk Lee, and Rebecca Batchelder (2008). "Appraisal of the generalized likelihood uncertainty estimation (GLUE) method." In: *Water Resources Research* 44, W00B06. DOI: 10.1029/2008WR006822.

Struthers, I and M Sivapalan (2007). "A conceptual investigation of process controls upon flood frequency: role of thresholds." In: *Hydrology and Earth System Sciences* 11.4, pp. 1405–1416. DOI: 10.5194/hess-11-1405-2007.

Struthers, Iain, Christoph Hinz, and Murugesu Sivapalan (2007a). "Conceptual examination of climate-soil controls upon rainfall partitioning in an open-fractured soil II: Response to a population of storms." In: *Advances in Water Resources* 30.3, pp. 518–527. DOI: 10.1016/j.advwatres.2006.04.005.

Struthers, Iain, Christoph Hinz, and Murugesu Sivapalan (2007b). "Conceptual examination of climate-soil controls upon rainfall partitioning in an open-fractured soil II: Response to a population of storms." In: *Advances in Water Resources* 30.3, pp. 518–527. DOI: 10.1016/j.advwatres.2006.04.005.

Szolgay, J (2004). "Multilinear discrete cascade model for river flow routing and real time forecasting in river reaches with variable wave speed." In: *Hydrological Risk: Recent advances in peak river flow modelling, prediction and real-time forecasting - Assessment of the impacts of land-use and climate changes*. Ed. by A Brath. European Sci-

ence Foundation Workshop, Bologna (Italy), 24-25 October 2003, pp. 271 –284.

Tetzlaff, Doerthe, James P. McNamara, and Sean K. Carey (2011). "Measurements and modelling of storage dynamics across scales." In: *Hydrological Processes* 25.25, pp. 3831–3835. DOI: 10.1002/hyp.8396.

Thielen, J., J. Bartholmes, M.-H. Ramos, and A. de Roo (2009). "The European Flood Alert System - Part 1: Concept and development." In: *Hydrology and Earth System Sciences* 13, pp. 125–140. DOI: 10.5194/hess-13-125-2009.

Tian, Fuqiang, Hongyi Li, and Murugesu Sivapalan (2012). "Model diagnostic analysis of seasonal switching of runoff generation mechanisms in the Blue River basin, Oklahoma." In: *Journal of Hydrology* 418-419, pp. 136–149. DOI: 10.1016/j.jhydrol.2010.03.011.

Tilmant, A., Q. Goor, and R. Kelman (2011). "Optimal Multipurpose-Multireservoir Operation Model with Variable Productivity of Hydropower Plants." In: *Journal of Water Resources Planning and Management* 137.February, pp. 258–267. DOI: 10.1061/(ASCE)WR.1943-5452.0000117.

Todini, E. (2007). "A mass conservative and water storage consistent variable parameter Muskingum-Cunge approach." In: *Hydrology and Earth System Sciences* 11, pp. 1645–1659. DOI: 10.5194/hess-11-1645-2007.

Toloczyki, M., P. Trurnit, A. Voges, and H. Wittekindt (2006). *Geological Map of Germany 1:1,000,000 (GK1000)*. Federal Institute for Geosciences and Natural Resources (BGR), Hannover.

Troch, Peter A, Claudio Paniconi, and E Emiel van Loon (2003). "Hillslope-storage Boussinesq model for subsurface flow and variable source areas along complex hillslopes: 1. Formulation and characteristic response." In: *Water Resour. Res.* 39.11, p. 1316. DOI: 10.1029/2002wr001728.

Tromp-Van Meerveld, H. J. and J. J. McDonnell (2006). "Threshold relations in subsurface stormflow: 1. A 147-storm analysis of the Panola hillslope." In: *Water Resources Research* 42.2, n/a–n/a. DOI: 10.1029/2004WR003778.

Tukey, J (1975). "Mathematics and Picturing Data." In: *Proceedings of the 1974 Congress of Mathematicians*. Ed. by R James. Vol. 2. Vancouver, pp. 523–531.

Vrugt, Jasper A., Cajo J. F. ter Braak, Martyn P. Clark, James M. Hyman, and Bruce A. Robinson (2008). "Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation." In: *Water Resources Research* 44.12, pp. 1–52. DOI: 10.1029/2007WR006720.

Vrugt, Jasper a., Hoshin V. Gupta, Luis A. Bastidas, Bouten Willem, and Soroosh Sorooshian (2003). "Effective and efficient algorithm for multiobjective optimization of hydrologic models." In: *Water Resources Research* 39.8, pp. 1–19. DOI: 10.1029/2002WR001746.

Wagener, T, N Mcintyre, M J Lees, H S Wheater, and H V Gupta (2003). "Towards reduced uncertainty in conceptual rainfall-runoff

modelling: Dynamic identifiability analysis." In: *Hydrological Processes* 17, pp. 455–476. DOI: 10.1002/hyp.1135.

Wagener, Thorsten, Murugesu Sivapalan, Peter Troch, and Ross Woods (2007). "Catchment Classification and Hydrologic Similarity." In: *Geography Compass* 1.4, pp. 901–931. DOI: 10.1111/j.1749-8198.2007.00039.x.

Wagener, Thorsten, Murugesu Sivapalan, Peter A. Troch, Brian L. McGlynn, Ciaran J. Harman, Hoshin V. Gupta, Praveen Kumar, P. Suresh C Rao, Nandita B. Basu, and Jennifer S. Wilson (2010). "The future of hydrology: An evolving science for a changing world." In: *Water Resources Research* 46.5. DOI: 10.1029/2009WR008906.

Weglarczyk, Stanislaw (1998). "The interdependence and applicability of some statistical quality measures for hydrological models." In: *Journal of Hydrology* 206.1-2, pp. 98–103. DOI: 10.1016/S0022-1694(98)00094-8.

Westerberg, I. K. and H. K. McMillan (2015). "Uncertainty in hydrological signatures." In: *Hydrology and Earth System Sciences* 19.9, pp. 3951–3968. DOI: 10.5194/hess-19-3951-2015.

Wienhöfer, J. and E. Zehe (2014). "Predicting subsurface stormflow response of a forested hillslope-the role of connected flow paths." In: *Hydrology and Earth System Sciences* 18.1, pp. 121–138. DOI: 10.5194/hess-18-121-2014.

Wienhöfer, J., K. Germer, F. Lindenmaier, A. Färber, and E. Zehe (2009). "Applied tracers for the observation of subsurface stormflow at the hillslope scale." In: *Hydrology and Earth System Sciences* 13, pp. 1145–1161. DOI: 10.5194/hess-13-1145-2009.

Williams, J R (1969). "Flood Routing with Variable Travel Time or Variable Storage Coefficients." In: *Transactions of the ASAE* 12.1, pp. 100–103. DOI: 10.13031/2013.38772.

Willmott, C. J. (1981). "On the validation of models." In: *Physical Geography* 2.2, pp. 184–194.

Winter, Thomas C. (2001). "the Concept of Hydrologic Landscapes." In: *Journal of the American Water Resources Association* 37.2, pp. 335–349. DOI: 10.1111/j.1752-1688.2001.tb00973.x.

Woods, Ross A. (2009). "Analytical model of seasonal climate impacts on snow hydrology: Continuous snowpacks." In: *Advances in Water Resources* 32.10, pp. 1465–1481. DOI: 10.1016/j.advwatres.2009.06.011.

Woods, Ross (2003). "The relative roles of climate, soil, vegetation and topography in determining seasonal and long-term catchment dynamics." In: *Advances in Water Resources* 26.3, pp. 295–309. DOI: 10.1016/S0309-1708(02)00164-1.

Wrede, Sebastian, Fabrizio Fenicia, Nuria Martinez-Carreras, Jerome Juilleret, Christophe Hissler, Andreas Krein, Hubert H G Savenije, Stefan Uhlenbrook, Dmitri Kavetski, and Laurent Pfister (2015). "Towards more systematic perceptual model development: A case study using 3 Luxembourgish catchments." In: *Hydrological Processes* 29.12, pp. 2731–2750. DOI: 10.1002/hyp.10393.

Yazdi, J. and S. A A Salehi Neyshabouri (2012). "Optimal design of flood-control multi-reservoir system on a watershed scale." In:

Natural Hazards 63.2, pp. 629–646. DOI: 10.1007/s11069-012-0169-6.

Yilmaz, Koray K., Hoshin V. Gupta, and Thorsten Wagener (2008). "A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model." In: *Water Resources Research* 44.9, n/a–n/a. DOI: 10.1029/2007WR006716.

Zappa, Massimiliano, Felix Fundel, and Simon Jaun (2013). "A 'PeakBox' approach for supporting interpretation and verification of operational ensemble peak-flow forecasts." In: *Hydrological Processes* 27.1, pp. 117–131. DOI: 10.1002/hyp.9521.

Zehe, E., H. Lee, and M. Sivapalan (2006). "Dynamical process upscaling for deriving catchment scale state variables and constitutive relations for meso-scale process models." In: *Hydrology and Earth System Sciences* 10, pp. 981–996. DOI: 10.5194/hess-10-981-2006.

Zehe, E. and M. Sivapalan (2014). "Threshold behaviour in hydrological systems as (human) geo-ecosystems: manifestations, controls, implications." In: *Hydrology and Earth System Sciences* 13, pp. 1273–1297. DOI: 10.5194/hess-13-1273-2009.

Zehe, E., T. Maurer, J. Ihringer, and E. Plate (2001). "Modeling water flow and mass transport in a loess catchment." In: *Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere* 26.7-8, pp. 487–507. DOI: 10.1016/S1464-1909(01)00041-7.

Zehe, E., H. Elsenbeer, F. Lindenmaier, K. Schulz, and G. Blöschl (2007). "Patterns of predictability in hydrological threshold systems." In: *Water Resources Research* 43.7, pp. 1–12. DOI: 10.1029/2006WR005589.

Zehe, E., T. Graeff, M. Morgner, A. Bauer, and A. Bronstert (2010). "Plot and field scale soil moisture dynamics and subsurface wetness control on runoff generation in a headwater in the Ore Mountains." In: *Hydrology and Earth System Sciences* 14.6, pp. 873–889. DOI: 10.5194/hess-14-873-2010.

Zehe, E., U. Ehret, T. Blume, A. Kleidon, U. Scherer, and M. Westhoff (2013). "A thermodynamic approach to link self-organization, preferential flow and rainfall-runoff behaviour." In: *Hydrology and Earth System Sciences* 17.11, pp. 4297–4322. DOI: 10.5194/hess-17-4297-2013.

Zehe, E. et al. (2014). "HESS Opinions: From response units to functional units: A thermodynamic reinterpretation of the HRU concept to link spatial organization and functioning of intermediate scale catchments." In: *Hydrology and Earth System Sciences* 18.11, pp. 4635–4655. DOI: 10.5194/hess-18-4635-2014.

Zehe, Erwin and Günter Blöschl (2004). "Predictability of hydrologic response at the plot and catchment scales: Role of initial conditions." In: *Water Resources Research* 40.10, pp. 1–21. DOI: 10.1029/2003WR002869.

Zehe, Erwin, Rolf Becker, András Bárdossy, and Erich Plate (2005). "Uncertainty of simulated catchment runoff response in the presence of threshold processes: Role of initial soil moisture and precipitation." In: *Journal of Hydrology* 315.1-4, pp. 183–202. DOI: 10.1016/j.jhydrol.2005.03.038.

Zehetmair, Swen, Jürgen Pohl, Katharina Ehrler, Britta Wöllecke, Uwe Grünewald, Sabine Mertsch, Reinhard Vogt, and Yvonne Wieczorrek (2008). "Hochwasservorsorge und hochwasserbewältigung in unter-schiedlicher regionaler und akteursbezogener ausprägung." In: *Hydrologie und Wasserbewirtschaftung* 52.4, pp. 203–211.

Zhang, Y. Y., Q. X. Shao, A. Z. Ye, H. T. Xing, and J. Xia (2016). "Integrated water system simulation by considering hydrological and biogeochemical processes: model development, with parameter sensitivity and autocalibration." In: *Hydrology and Earth System Sciences* 20.1, pp. 529–553. DOI: 10.5194/hess-20-529-2016.

OWN PUBLICATIONS

PEER-REVIEWED INTERNATIONAL PUBLICATIONS

Seibert S. P., Jackisch C., Pfister L., Ehret U. & Zehe E. (2016): Exploring the interplay between state, structure and runoff behavior of lower mesoscale catchments. Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-109, in review.

Seibert S. P., Ehret U. and Zehe E. (2016): Disentangling timing and amplitude errors in streamflow simulations, Hydrol. Earth Syst. Sci., 20, 3745-3763, doi:10.5194/hess-20-3745-2016.

Seibert S. P., Skublic D. & Ehret U. (2014): The Potential of coordinated reservoir operation for flood mitigation in large basins - a case study at the Bavarian Danube using coupled hydrological-hydrodynamic models. J. Hydrol., 517, 1128-1144, doi: 10.1016/j.jhydrol.2014.06.048.

Crochemore L., Perrin C., Andreassian V., Ehret U., *Seibert S. P.*, Grimaldi S., Gupta H. & Paturel J.-E. (2014): Comparing expert judgement and numerical criteria for hydrograph evaluation. Hydrol. Sci. J.-J. Sci. Hydrol., 60, 3, 402-423, doi: 10.1080/02626667.2014.903331.

Skublics D., *Seibert S. P.* & Ehret U.(2014): Abbildung der Hochwasserretention durch hydrologische und hydrodynamische Modelle unter unterschiedlichen Randbedingungen. Sensitivitätsanalyse am Donauabschnitt zwischen Neu-Ulm und Donauwörth. Hydrologie und Wasserbewirtschaftung 3 (58). 178-189, doi: 10.5675/HyWa_2014,3_2.

Fiener P., *Seibert S. P.* & Auerswald K. (2011): A compilation and meta-analysis of rainfall simulation data on arable soils. J. Hydrol., 409, 395-406, doi: 10.1016/j.jhydrol.2011.08.034.

OTHER PUBLICATIONS

Seibert S. P. & Ehret U. (2014): Series Distance – ein innovatives Gütemaß zur Bewertung der Simulationsgüte hydrologischer und hydraulischer Modelle. In: „Hochwasser und kein Ende – Statusberichte, aktuelle Vorhaben, neue Planungswerkzeuge". Beiträge zur Fachtagung am 3. und 4. Juli 2014 in Obernach. pp. 145-158. ISBN 978-3-943683-06-6.

Skublics D. & *Seibert S. P.* (2012): Einzugsgebietsweite Abbildung der Hochwasserwellenbeeinflussung steuerbarer und nicht steuerbarer Rückhaltemaßnahmen - Im Einzugsgebiet der bayerischen Donau, Tagungsband Wasserbausymposium 12.-15. September 2012, Graz, Österreich, ISBN:978-3-85125-230-9.

Seibert S. P., Auerswald K., Fiener P., Disse M., Martin W., Haider J., Michael A. & Gerlinger K. (2011): Surface runoff from arable land - a homogenized data base of 726 rainfall simulation experiments. DOI: 10.1594/GFZ.TR32.2.

CONFERENCE CONTRIBUTIONS

Seibert S. P. & Zehe E. (2016): Unraveling ecological and abiotic controls on seasonal runoff dynamics at lower mesoscale catchments (oral contribution).

European Geophysical Society General Assembly, Vienna, Geophysical Research Abstracts Vol. 18 (EGU2016-13423).

Seibert S. P., Loritz R. & Zehe E. (2015): Functional hydrologic signatures to detect first order runoff formation mechanisms on the headwater scale (oral contribution). European Geophysical Society General Assembly, Vienna, Geophysical Research Abstracts Vol. 17 (EGU2015-13512).

Loritz R., Weiler M. & *Seibert S. P.* (2015): Data mining methods for predicting event runoff coefficients in ungauged basins using static and dynamic catchment characteristics (poster). European Geophysical Society General Assembly, Vienna, Geophysical Research Abstracts Vol. 17 (EGU2015-11072).

Seibert S. P., Zehe E., Ehret U. & Jackisch C. (2014): Signatures aS. P.roxies for hydrological functioning (oral contribution). European Geophysical Society General Assembly, Vienna, Geophysical Research Abstracts Vol. 16 (EGU2014-10140).

Seibert S. P., Ehret U. & Skublics D. (2014): Erlaubt eine flussgebietsweit abgestimmte Speicherbewirtschaftung eine Optimierung des Hochwassermanagements in großen Einzugsgebieten? Ein Fallbeispiel aus dem Bayerischen Donaueinzugsgebiet. LARSIM Anwendertreffen, 18.-19.04.2014, Karlsruhe.

Ehret U. & *Seibert S. P.* (2013): Series Distance – a metric for the quantification of hydrograph errors and forecast uncertainty, simultaneously for timing and magnitude (oral contribution). American Geophysical Union. Fall Meeting 2013. (H11K-03).

Seibert S. P. & Ehret U. (2012): Detection of flood events in hydrological discharge time series (poster). European Geophysical Society General Assembly, Vienna, Geophysical Research Abstracts Vol. 14 (EGU2012-5924).

Seibert S. P. & Ehret U. (2011): Joint consideration of timing and amplitude uncertainties in hydrological simulation and forecasting (oral contribution). European Geophysical Society General Assembly 2011, Vienna, Geophysical Research Abstracts Vol. 13 (EGU2011-2153).

Seibert S. P., Fiener P. & Auerswald K. (2010): Problems and solutions in the meta analysis of rainfall simulation data (oral contribution). European Geophysical Society General Assembly, Vienna. Geophysical Research Abstracts Vol. 12 (EGU2010-8388).

Seibert S. P., Fiener P. & Auerswald K. (2010): How to derive Horton infiltration parameters from rainfall simulation data (poster). European Geophysical Society General Assembly, Vienna. Geophysical Research Abstracts Vol. 12 (EGU2010-8423).

ACKNOWLEDGMENTS

This dissertation was written during challenging times and personal family circumstances. However, the opportunity to concentrate on exciting topics, to work in a creative environment and most importantly to work with many inspiring people was an important source of power for me which enabled me to carry on. For this I feel deeply grateful.

First and foremost I thank my supervisors *Erwin Zehe* and *Uwe Ehret*. I felt honored that you accepted me into your PhD program. You supported (most of) my ideas and gave me the space to find my own way. You never lost trust in my abilities even during challenging times and gave me more support than I could have asked for. It is in particular our inspiring discussions, *Erwin*, that I appreciated most. Science is about generating and exploiting ideas, about developing hypotheses and about rejecting them when they turn out to be invalid. You were an outstanding mentor in this regard and collaborating with you was a great pleasure for me. Thank you for giving me the opportunity to learn so much. *Uwe*, you organized and supervised two of the projects I've been working on. You were always there to answer my questions and helped me whenever I needed assistance. Most importantly, you brought me back on track by providing structure whenever confusion took over. Last but not least I thank Laurent Pfister who reviewed my thesis and provided the Attert data.

Special thanks also go to my former PhD colleague *Conrad Jackisch*. Conrad, you were an excellent room mate and provided sound support in many different respects. My colleagues *Daniel Skublics* from TUM as well as from KIT *Jan Wienhöfer, Martin Helms, Malte Neuper, Ralf Loritz, Petra Gabelmann, Simon Höllering, Lucas Reid, Jürgen Ihringer, Martijn Westhoff, Ulrike Scherer, Giorgia Fosser* and *Julian Klaus* contributed in many different ways to my dissertation and I feel sorry that I could but little return to you during the last years.

None of this would have been possible without the guidance of *Karl Auerswald* from TUM who recognized my talents and paved my way into science many years ago.

In addition to those colleagues many other people were involved in the different projects. The flood mitigation chapter would not have been possible without the help of *Franz-Klemens Holle, Martin Schmid, Stefan Laurent, Katja Moritz, Karl-Heinz Daamen, Natalie Stahl* and *Alfons Vogelbacher* from the Bavarian Environmental Agency (LfU) and the different water authorities (WWA), who shared site specific knowledge and provided profound assistance in various issues related to data and modelling. *Ingo Haag, Nicole Henn* and *Mario Böhm* from HYDRON and *Manfred Bremicker* from the Environmental Agency of Baden-Württemberg cooperated fruitfully in the development of the water budget model for the Danube and provided their outstanding

support. In this connection I further gratefully acknowledge funding and data provision by the Bavarian Environmental Agency (LfU) and the German (DWD) and Austrian weather services (ZAMG).

The improvement of the Series Distance (SD) method would not have been possible without the users of SD who provided valuable feedback and constructive criticism. *Tilmann Gneiting* from the Heidelberg Institute for Theoretical Studies (H-ITS) offered valuable discussions on the error dressing concept and *Clemens Mathis* from Wasserwirtschaft Vorarlberg provided the case study data and the hydrological model.

I also acknowledge the help of the anonymous reviewers in the peer-review process whose valuable comments helped to improve the quality of my publications. Furthermore I thank the open-source software community for maintaining so many archives and forums - these are extensive stocks of knowledge which enabled me to accomplish my thesis using transparent and non-commercial software products. I further acknowledge the support for open access publishing by the Deutsche Forschungsgemeinschaft (DFG) and the Open Access Publishing Fund of Karlsruhe Institute of Technology (KIT).

Last but not least I thank my family and particularly my wife *Toni*. I do know now challenges in life that make writing a dissertation a comparatively easy job. You made it possible that we mastered both. I am very proud of you and incredibly happy to have you at my side. Thank you for being there! I also want to express my special thanks to my parents *Else* and *Anton Seibert* and to my parents-in-law *Elfie* and *Hans Kuhn*. Your contribution was behind the scenes but enormous and you took much more load from my back than I could have asked for. It was your help and support that made me realize that I am not standing alone but can count on my family standing by me. Thank you four your continuous loving.

DECLARATION

Eidesstattliche Versicherung gemäß §6 Abs. 1 Ziff. 4 der Promotionsordnung des Karlsruher Instituts für Technologie für die Fakultät für Bauingenieur-, Geo- und Umweltwissenschaften:

1. Bei der eingereichten Dissertation zu dem Thema *Flood mitigation, model uncertainty and process diagnostics* handelt es sich um meine eigenständig erbrachte Leistung.

2. Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keiner unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtlich oder sinngemäßaus anderen Werken übernommene Inhalte als solche kenntlich gemacht.

3. Die Arbeit oder Teile davon habe ich bislang nicht an einer Hochschule des In- oder Auslands als Bestandteil einer Prüfungs- oder Qualifikationsleistung vorgelegt.

4. Die Richtigkeit der vorstehenden Erklärungen bestätige ich.

5. Die Bedeutung der eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung sind mir bekannt.

Ich versichere an Eides statt, dass ich nach bestem Wissen die reine Wahrheit erklärt und nichts verschwiegen habe.

Karlsruhe, im Mai 2016

COLOPHON

Typesetting of the thesis was done in LaTeX using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*".

I use Linux and prefer free and open-source software over commercial products. *R* was used for various tasks including data management, statistical evaluations and the making of graphs, maps, and technical drawings. Geographical analyses were predominantly done using the *GIS* programs and libraries implemented in *QGIS*, *GRASS* and *Whitebox*. *(My)SQL* relational database systems and corresponding front-ends like *HeidiSQL* were used for data management. Last but not least the paper management software *Mendeley* and different editors including *notepad++* and *sublime* proved to be versatile and powerful tools. Commercial software was partly required in the first two parts of this thesis. This involved *ESRI's ArcGIS*, *Matlab* and *MS Windows/ Office*.