RUHR-UNIVERSITÄT BOCHUM

RUB

SCHRIFTENREIHE
HYDROLOGIE/WASSERWIRTSCHAFT

Application of Machine Learning enhanced Agent-based Techniques in Hydrology and Water Resource Management

von Benjamin Mewes

2019

LEHRSTUHL FÜR HYDROLOGIE, WASSERWIRTSCHAFT
UND UMWELTTECHNIK

31

Die vorliegende Arbeit wurde von der Fakultät
für Bau- und Umweltingenieurwissenschaften als Dissertation angenommen.

Doktorarbeit eingereicht am:	09.01.2019
Tag der mündlichen Prüfung:	16.05.2019

Berichter:

Prof. Dr. rer. nat. habil. Andreas Schumann, Ruhr-Universität Bochum
Prof. Dr.-Ing. Markus König, Ruhr-Universität Bochum
Prof. Dr. Chaopeng Shen, Pennsylvania State University, United States of America

Universitätsstraße 150, 44801 Bochum
Tel. +49 (0) 234 32-24693, Fax. -14153

ISSN 0949-5975

Herausgeber: Prof. Dr. rer. nat. habil. Andreas Schumann

Kurzfassung

Die vorliegende Arbeit untersucht die Anwendung adaptiver agenten-basierter Ansätze in Hydrologie und Wasserwirtschaft. Im Informationszeitalter werden neuartige Ansätze benötigt, um aus der Menge an verfügbaren Daten neue Informationen zu gewinnen. Dies spielt vorrangig beim Umgang mit großen Datenmenge eine Rolle, die häufig mit dem Begriff der Big Data in Verbindung gebracht werden. Dabei wurden in der Vergangenheit verschiedene Ansätze entwickelt, die jedoch diametral unterschiedlich zu betrachten sind. Auf der einen Seite stehen die reinen datengestützten Auswertungsansätze, zu denen u.a. stochastische Anwendungen und maschinelles Lernen gehören, auf der anderen Seite stehen wissensbasierte Prozessmodelle, wie physikalische Prozessmodelle oder die auf Regeln aufbauende Agenten-basierte Modellierung. An der Schnittmenge finden sich konzeptionelle Modelle, die beide Seiten in Ansätzen vereinen, ohne jedoch von den jeweiligen Stärken zu profitieren.

In dieser Arbeit werden anhand von Beispielen aus der Hydrologie und Wasserwirtschaft exemplarisch die Ansätze des maschinellen Lernens sowie der Agenten-basierten Modellierung als Vertreter beider Analyseansätze behandelt und in einem konzeptionellen Ansatz vereint. Dabei zeigen sich besonders die Vorteile des maschinellen Lernens bei der Anwendung auf große Datensätze zur Ereignisseparation von Abflussdaten. Des Weiteren helfen informations-gestützte Ansätze die Vorhersagekraft von Daten zu beurteilen und können somit bei der Interpolation von schwer zu ermittelnden Messwerten dienen, wie das Beispiel der Vorhersage von Tracer Messungen aus Abflussdaten zeigt. Zudem erlauben die informationsbasierten Gütekriterien eine objektive Interpretation der interpolierten Daten, die im weiteren Verlauf zur Modellierung Karst-hydrologischer Systeme genutzt werden können. Da die Auswahl eines lernenden Algorithmus nicht eindeutig vorab bestimmbar ist, wird für jedes Fallbeispiel, frei nach dem Free-Lunch-Theorem, eine Auswahl an Ansätzen angewendet und getestet. Somit kann festgestellt werden, dass für die Ereignisseparation Support Vector Machines und Extreme Learning Machines die sinnvollsten Anwendungen sind, um zu guten Ergebnissen zu kommen. Neuronale Netzwerke hingegen sind nicht für die Ereignisseparation einzusetzen, zeigen aber gute Ergebnisse im Bereich der Vorhersage von Tracer Konzentrationen aus Abflussdaten.

Im Gegensatz zu den daten-gestützten Methoden des maschinellen Lernens steht die Agenten-basierte Modellierung für die regelbasierte Modellierung. Hierbei stehen autonome Softwareprogramme im Fokus, die anhand eines definierten Regelwerkes Entscheidungen treffen und sich untereinander abstimmen müssen. Zudem befinden sich die autonomen Einheiten im Austausch mit ihrer Umwelt, wodurch sich ein komplexes Zusammenspiel zwischen den Modellkomponenten ergibt. Das Regelwerk, nach denen sich die Agenten verhalten und ihre Aktionen koordinieren, muss vorab definiert werden. Somit ist die Modelltechnik interessant für Anwendungsfälle, in denen grundlegende Prinzipien verstanden, aber noch nicht in jedem Detail durchdrungen wurden. Der Einsatz Agenten-basierter Modelle in Hydrologie und Wasserwirtschaft beschränkte sich in der Vergangenheit oft auf sozio-hydrologische Modelle, in denen die Agenten einzelne Akteure im wasserwirtschaftlichen System darstellen. In dieser Arbeit wird gezeigt, dass ein Agenten-basiertes Modell ein sinnvoller Ansatz ist, um komplexe, physische Systeme räumlich und zeitlich differenziert zu betrachten. Hierzu wurde ein Modellframework geschaffen, das die Bewegung von Wasser durch die Bodenzone darstellt. Die autonomen Agenten zeigten dabei beobachtbare Verhaltensmuster von Wasser auf, die ansonsten nicht in Modellen abbildbar sind, wie die Altersstruktur des fließenden Wassers sowie Austausch- und Kontaktzeiten zwischen unterschiedlichen Wasseragenten. Neben der Agenten-basierten Modellierung ist auch die Agenten-basierte Klassifikation eine sinnvolle Bereicherung des Methodensets zur Aufarbeitung und Analyse hydrologischer oder hydrologisch relevanter Daten unter

Berücksichtigung von Expertenwissen. Anhand eines räumlichen Beispiels aus Nebraska, USA wird die Tauglichkeit agenten-basierter Klassifikation zur Identifikation von bewässerter Landwirtschaft aus spektralen Fernerkundungsdaten aufgezeigt. Die komparative Studie enthüllte, dass die agenten-basierte Klassifikation vollständigere Klassifikationsergebnisse produziert als pixel-basierte Gegenstücke. Im Vergleich zur objekt-basierten Klassifikation entfällt zudem eine genauere Parametrisierung des Segmentationsalgorithmus, da die Parametrisierung der Agenten-basierten Klassifikation lediglich eine Voruntersuchung darstellt. Trotz aller Vorzüge bringt die Agenten-basierte Klassifikation weitere Nachteile mit sich, die sich besonders im Bereich der Regelwerke manifestieren. Pixel-basierte Ansätze zeigen eine weniger stark ausgeprägte Abhängigkeit vom definierten Regelwerk sowie der Qualität der Eingangsdaten, da Fehler in diesem Ansatz auf den fehlerhaften Bereich eines oder weniger Pixel beschränkt sind.

Abschließend werden beide Auswertungsansätze im Bereich der adaptiven Agenten-basierten Modellierung zusammengeführt. Hierbei handelt es sich um einen Ansatz, der die Vorzüge des maschinellen Lernens mit den Vorteilen der Agenten-basierten Modellierung kombiniert. Hierbei werden Schwellenwerte, die die Aktionen der Agenten auslösen, mittels maschinellem Lernen an die vorherrschende Situation angepasst.

Insgesamt zeigt diese Arbeit anhand einer Vielzahl an unterschiedlichen Fallbeispielen aus Hydrologie und Wasserwirtschaft mögliche Anwendungsfelder neuartiger Modellierungs- und Analysetechniken. Neben den Verbesserungen wird ein Fokus auf Hindernisse und Fallstricke im Bereich Big Data und Maschinellem Lernen gelegt. Durch die freien Strukturen neigen die Modelle und Algorithmen zur Überanpassung und zu einer mangelnden Übertragbarkeit der Ergebnisse. Somit wird, falls möglich, jede Fallstudie im Vergleich zu anerkannten Methoden komparativ aufgebaut, um die Verbesserungen durch die neuartigen Analysemethoden zu quantifizieren und qualitativ zu beschreiben. Des Weiteren wird für jedes Fallbeispiel eine Strategie zur Übertragung der erlangten Ergebnisse auf andere Gebiete und Probleme präsentiert. Die Hydrologie kann mittels der hier vorgestellten Methoden neuartige Erkenntnisse aus existierenden Daten gewinnen. Zudem wird der Einbezug von Expertenwissen in daten-gestützte Anwendungen vereinfacht und formalisiert.

Abstract

This thesis introduces the application of machine learning enhanced adaptive agent-based techniques in hydrology and water resource management. In the age of increased data availability the deduction of information from the data is the major task for any data scientist. Therefore, novel approaches were developed to extract information either through da-ta-driven approaches like machine learning or the alternative direction, knowledge-based systems. One of the most recent knowledge-driven approaches is the so-called agent-based modelling where autonomous programs decide on their actions based on a predefined rule set. Conceptual models, like the common HBV model (Lindström et al., 1997), are located at the intersection of both modelling worlds without incorporating the advantages of either approaches. This thesis proposes a preliminary solution which profits from both approaches in a water resource management model.

Current hydrological models rarely make use of the increased data availability because the applied core concepts originate from times where computational power was limited or non-existed. Moreover, the ingestion of data from different sources remains difficult. In order to answer our more complex research questions with a focus on linkages between processes in the environment and society, novel more dynamic approaches are required. Using case studies from hydrology and water resource management, both modelling approaches are explained and introduced. Machine learning shows its merits in combination with big data archives. Therefore, its applicability on flood event separation from large datasets of runoff was tested. Here, Support Vector Machines and Extreme Learning Machines showed the highest performance. Thus, information-theory based criteria of performance could be applied to judge the information content of data and the quality of interpolation results from machine learning. Tracer measurements were taken from several different karst springs in France to predict tracer concentrations from discharge. In contrast to the flood event separation in Bavarian catchments, the tracer prediction of French karst springs did not show any preference towards any algorithm but the information content of the available data sets was quantified by the information-theory based criteria.

Contrary to data-driven machine learning approaches, knowledge-driven techniques like agent-based computing require a general understanding of the modelled pattern. Formerly, this approach was limited in application to decision making problems in social sciences and behavioral biology. The ability of agent-based models to capture physical systems in hydrology was shown by the creation of a modelling framework to describe the movement of water through the soil via water agents. In contrast to social systems, the general rule sets are well known and yet does the chaotic nature of the agent-based modelling approach reveal insights into system internals that would have remain hidden else? The chaotic nature in the agent-based model originates from the outcome of these models: Due to a high number of individual objects, the final outcome is more than the sum of individual decisions but the result from the interplay between the autonomous entities setting up the mod-el. In comparison to a classical numerical storage model, an agent-based model delivered comparable results. Moreover, this case study indicates behavior of water that could not be modelled by numerical model approaches like the interaction of spatially explicit water particles and the age structure of water in a soil column.

Not only were the modelling capabilities of agent-based computing tested, but also the classification capabilities of the approach. Therefore, an agent-based classification scheme was set up for the delineation of irrigated agriculture in Nebraska, USA, from spectral re-mote sensing data. Here, it could be shown that the agent-based classification delivers a more complete set of classified structures than its pixel contravenes. Additionally, it de-creases the influence of segmentation parameters in the object-based image classification. Nevertheless,

the agent-based classification still imposes several problems that are linked with the rule set that defines the interactions.

As a keystone bridging both worlds of modelling, the adaptive agent-based modelling is finally introduced. Therefore, a simplified irrigation model was set up based upon the historic Balinese water temple scheme. In contrast to a non-learning agent-based model, an adaptive agent-based model incorporating a machine learning approach was able to improve yields in a changing environment thus showing potential for self-improving models.

Overall, this thesis reveals benefits and disadvantages of unique data analysis approaches through various case studies from hydrology and water resource management. Next to improvements and information gain from existing big data archives, like the enhanced flood event separation by machine learning, limits and hurdles are discussed. Here, overfitting and a lack of transferability are the major sources of problems. If possible, for any case study a comparative analysis is shown where the novel approach is compared to an established approach to judge any improvement.

The methods presented in this thesis, help to extract novel information from existing data. Moreover, the incorporation of expert knowledge in data-driven approaches is simplified and formalized.

Preface

Artificial Intelligence (AI) is the core element in many new developments in science and technology. It is impressive how high the expectations in cognitive functions of computers are. The discussions, e.g. in the field of autonomous driving, show that the way from analytical AI to humanized AI, which would require cognitive, emotional, and social intelligence, will be long and difficult. But also the application of analytical AI is a challenge for hydrology and water management as it demands a cognitive representation of problems and is based on learning from data and experience to inform decisions. Machine learning, based on data analyses and/or experiments, is one very important component of most AI applications. A way to integrate this derived knowledge into models (and finally into decision making processes) are agent-based models, where components ("agents") simulate simultaneous operations and are able to interact with multiple other agents to simulate and predict the appearance of complex phenomena.

Benjamin Mewes thesis is focused on the application of such adaptive agent-based approaches in hydrology and water management. In his scientific work, the adaptation of systems is based on methods of machine learning, which are analyzing large data sets to derive rules and defining the functions of the agents. One substantial component in his research work is the comparison of different algorithms of machine learning with respect to their requirements to get input information and their ability to provide new information for the derivation of rules. In agent-based models, the interaction of model components is controlled by encapsulated, autonomous software units, whereby each of these agents has a strategy to achieve pursuing a specific goal.

This requires a problem-orientated machine learning system, which is applying the right method for the current problem and (more complicated) for the information requirements of specific types of agents. Since the agents interact with each other, the various steps to reach the overall goal have to be defined in advance and the behavior of each agent has to be adapted according to the current situation and the predefined knowledge by rules which was derived from the machine learning. In this way, the interplay of data mining and agent-based modeling became the methodological focus of this thesis. The complexity of the research work results from three components:

- the variety of hydrological and water management problems, determining the required information,
- the selection of appropriated machine learning tools to provide such information from data and
- the integration of this information into agent-based models.

In this way, Benjamin Mewes demonstrated with this thesis a high level of methodological competence in hydrology and water management but also in modern IT. With three meaningful applications he demonstrated the potential of adaptive agent-based modeling in conjunction with machine learning approaches. This required a deep understanding of the hydrological and water management components to develop the appropriated knowledge base for the agent-based models which represent these components. In this sense, the thesis is an interdisciplinary approach which deserves special appreciation. I hope that many other scientists will benefit from the experience published by Benjamin Mewes here and in several other publications, even if this would shorten the duration of novelty of his results. However, this is the fate of many very modern approaches in a dynamic field of research.

Prof. Dr. Andreas Schumann

Content

Abbreviations- and symbols

Abb.	Description
aABM	Adaptive Agent-based Model
AB	Agent-based
ABC	Agent-based Classification
ABM	Agent-based Model
ACC	Accuracy
ANN	Artificial Neural Networks
BaIM	Balinese Irrigation Model
CART	Classification and Regression Tree
CDL	Crop Data Layer
cmf	Catchment Modelling Framework
COHYST	Cooperative Hydrological Study
Cov	Temporal Coverage of an estimated and an observed event
EEA	European Environmental Agency
ELM	Extreme Learning Machines
Eq	Equation
ETM	Enhance Thematic Mapper
ETRF	Evaporative Fraction
GA	Green-Ampt Infiltration Model
GAMA	GIS Agent-based Modelling Architecture
GEE	Google Earth Engine
GI	Green Index
GP	Genetic Programming
GPU	Graphic Processing unit
HBV	Hydrologisk Buråns Vattenavdelning
HYShare	Share of farmers who grow High-Yielding rice varieties
IPA	Integrated Platform for Agent-based modelling
ID3	Iterative Dichotomiser 3
MI	Mutual Information
ML	Machine Learning
MVR	Mean Volume Ratio of an estimated and an observed event
NDVI	Normalized Vegetation Index
NGI	Normalized Green Index
RMSE	Root Mean Squared Error
SC	Soft Computing
SRTM	Shuttle Radar Topography Mission
SVM	Support Vector Machine
TOA	Top Of Atmosphere

Thrs(D)	Threshold of plant Disease that triggers immediate harvest of rice
VG	Van Genuchten
WS0	Mean Water Storage in time interval 0
$\overline{c_T}$	Mean tracer concentration
est	estimated
obs	observed

1 Introduction

In times of availability of big data archives and powerful computational technology, a variety of modelling techniques become more and more important. To find information in these data sets, data mining methods are common tools to investigate internal structures of the data sets to preprocess the data before the actual modelling takes place. The effort research has to invest on investigating and preprocessing the data is relative to the amount of available data. So, in the recent years Machine Learning (ML) has become a powerful tool in terms of data science (Kelleher et al., 2015). ML algorithms try to identify patterns in the data and then replicate the learned pattern to an unknown data set. ML fits well to predict missing data and to classify data sets where the relations within the data are not fully known and can thus not be defined a-priori. Often, the nature of hydrological systems is not fully known. Consequently, the rules leading to a desired result can't be described and data driven methods like ML are suitable tools to investigate the data set. Consequently, ML could fill the gap and lead to better results than classical approaches, e.g. to separate single flood events from continuous time series of runoff (Fig. 1). Known filters (Chapman, 1999) or recession-based approaches require local customizations and are rarely transferable (Mei and Anagnostou, 2015). ML-based approaches on the other hand could be used to detect patterns in runoff to separate flood events without the need for manual corrections and customizations. Baseflow recession helps to find the end of a runoff event, but it does not help to identify the beginning of the event. Thus, either an extension of the baseflow recession approach has to be formulated, or a different methodology to separate the flood event from the time series of runoff has to be found.

As the information content of data is of relevance in any data-driven ML approach, a deeper analysis of the underlying data has to be conducted before setting up ML schemes. An investigation of the information content sheds light on the hidden structure of the information in the data but it requires knowledge on the true reference data. Therefore, a different application has to be found than flood event separation, because a true reference data of trustworthy flood events is not available. Information rich applications, like the tracer analysis for the analysis of flood composition and the related catchment analysis (Garvelmann et al., 2017), on the other hand offer true reference data in terms of measurable tracer concentrations. Often, in tracer-based hydrology, the available time series of tracer measurements are too short to work with. So, the idea evolved to predict the hard to measure variable (the tracer concentration) with an easy to measure variable (the runoff) as it is shown in Fig. 1. The information content analysis is applied to judge the ML prediction of the tracer concentrations.

Because of the unknown internal structure of the data and the relations within the data ML-based approaches are known as bottom-up techniques. The bottom of the model, the data defines the starting point to understand and replicate the system. In the hydrological community this approach of modelling is often accused of a black-box behavior with a lack of interpretability in real world applications (Shen et al., 2018). The opposite of bottom-up approaches are top-down modelling techniques. Under this umbrella term rule-based techniques, like Agent-based computation are collected. Here, the rules of the interaction between the model components have to be defined beforehand. These rule-dependant approaches are suitable to investigate assumptions on the interplay of model components and resulting patterns. Hydrologically, these top-down modelling techniques fit well with highly dynamic situations like the flow of water through a porous soil matrix. From a bottom-up perspective, the emergent behavior could also be modelled with deep learning approaches (Shen, 2018). This approach was discarded because the interpretability of the internal states of a deep learning network is often complex or even impossible, so this phenomenon was modelled with a rule-based agent-based model. Although the general physical rules are known, the definitions of rule sets within hydrological models often remain incomplete and thus traditional models do not lead to observed behavior of water in the vadose zone. Furthermore, model conceptions like HBV (Lindström et al., 1997) or cmf (Kraft et al., 2011) don't allow the analysis of certain aspects of the water like the age distribution or exchange of solutes in the soil. Therefore, an Agent-based model (ABM) is setup to describe the behavior of water in the soil and to model the reaction of a soil column to an infiltration input. The water is represented through autonomous software agents that interact with their environment according to a defined rule set. With this approach, the age distribution of water remains intact and interpretable in the here presented Agent-based model for the transport of water through the porous soil matrix which is a major advantage against numerical soil water models like Hydrus 1D (Simunek et al., 2005). Next to modelling, Agent-based methods can be used to classify data by incorporating the aforementioned advantages of the fully connected top-down Agent-based approach (Blaschke et al., 2013; Hofmann et al., 2015). Here, irrigated agriculture is identified in Nebraska, USA, from spectral input data with Agent-based classification techniques (ABC) (Fig. 1). In contrast to traditional pixel-based remote sensing image interpretation techniques, the Agent-based classification considers similar pixels as objects that are able to alter and to interact. Thus, the topography of the objects and the temporal development of the objects remains intact. Hence, spatial patterns can be investigated by a more complete classification result that takes advantage of the extended classification abilities of software agents which know about their environment and

their neighbors.

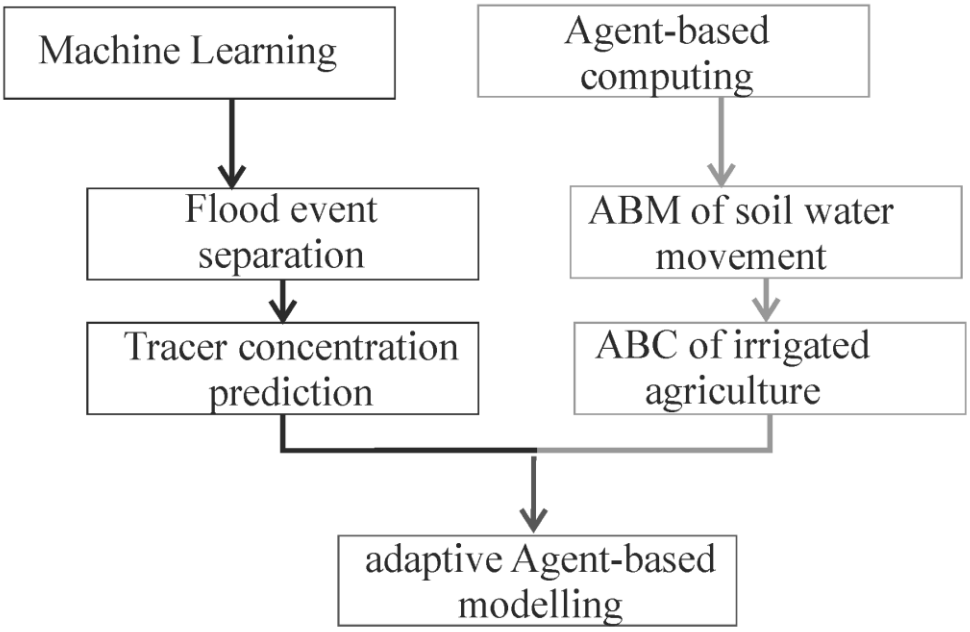


Fig. 1: Methodical evolution of an adaptive Agent-based model for water resource management combining the advantages of rule-based Agent-based computing and data-driven Machine Learning as a flow chart in this thesis.

To overcome limitations of both approaches (abstract results, uninterpretable internal states of ML and limits of rule sets in ABM), a combination of both worlds is a promising modelling technique: the adaptive Agent-based model. Here, ML is used to adapt thresholds in behavioral rules in Agent-based models. This leads to a highly adaptive modelling approach which alters the interplay of model components dynamically at runtime. Through this approach, water resource management problems and the social development of strategies to adapt to changing environmental conditions can be modelled like the evolution of water distribution strategies. In this thesis, the example of medieval Bali with its antique water-focused temple religion is taken (Lansing, 2007). This religion, or cult, was created to regulate the distribution of water among the farmers and state the societal rules and mindset how to treat topics like inequalities in water distribution, cropping schemes and pest-control. The model reveals the benefits of a water-based religion that was developed to distribute water in rice farming communities. The interactions between the farmers and a coordinating are modelled to maximize the yield and to minimize to impact of pests that could harm the rice harvest. ML is applied to investigate the current hydrological situation and adapt the systematic behavior of the community towards drought, flooding and a destruction of harvest by pest. Traditional models cannot be adapted at runtime and are thus not able to model social adaption to rapid environmental changes. It can be shown that the adaptive ML-enhanced Agent-based model of the Bali irrigation system delivers higher yields and thus outperforms

the non ML-enhanced model. Consequently, the extension of the Agent-based modelling technique by ML methods represents a fundamental improvement for the application of Agent-based methods in the modelling of socio-hydrology and water resource management.

1.1 Overview and brief explanation of techniques applied in this study

The approaches presented here origin from information technology and its related sciences. Therefore, a general introduction with a definition is given before the actual case studies are presented. In the following section main terms are explained and definitions given. For a better readability some terms will later be abbreviated. Nevertheless, all used abbreviations are collected in the 'Abbreviations and Symbols' section of the thesis.

1.1.1 Big Data

The term "Big Data" is one of the major keywords in many data driven studies in the recent years (Mauro et al., 2016). Nevertheless, the definition of big data is vague and not obvious (Mohammed et al., 2016). The following selection of V-words is widely used to describe big data sets: Velocity (of transmission), Volume (of data), Value (information derived from the big data set), and Veracity (the trust towards the data set) (Mauro et al., 2016). Big data and the linked analysis of these archives is part of the data-information-wisdom hierarchy that states that information, not the data per se, is the core fundament of any big data analysis (Rowley, 2007; Mauro et al., 2016). The technological aspect of big data archives is crucial because the sheer amount of data has to be stored, analyzed and in case delivered to the user. Following Moore's law, the capacity of computing resources doubles every 18 to 24 months, but so does the amount of data (Hilbert and López, 2011). So, big data mirrors the race between the amount of available data and suitable approaches to extract data from the archives and leaves gaining new information from existing data as the major advantage of big data (Chaney et al., 2018). Thus, big data archives and novel analysis approaches, based on information gathered from these archives, are of great interest for any research community.

1.1.2 Machine Learning

Machine Learning is the hypernym for computer programs that extract patterns from data sets. Hereby, the patterns are not defined by the researcher, but characteristics of the patterns are identified by the machine and then transferred to a new problem. To find these patterns a ML algorithm has to be defined which adapts to a given set of training data. This represents the learning character of ML: Expertise from available data is taken to fit a black-box algorithm and then transferred to new data. ML has seen a variety of applications in hydrology, from rainfall-runoff-modelling (Solomatine and Dulal, 2003; Chen and Adams, 2006; Yu et al., 2017), estimations of soil moisture (Coopersmith et al., 2014), tracer concentration modelling in small rivers (Piotrowski et al., 2007), forecasting streamflow (Talei et al., 2010; He et al., 2014), evapotranspiration estimation (Tabari et al., 2012) and many more (Raghavendra and Deka, 2014).

ML can be divided into three branches: supervised learning, un-supervised learning and reinforcement learning (Kelleher et al., 2015). Supervised learning requires the definition of training data. The training data must contain the minimum amount of data to capture all relevant structures within the data without under- or overrepresentation of a target (Han and Kamber, 2010). The target of the ML approach can either be a class (after a classification) or a continuous variable (after a regression). Contrary, unsupervised learning does not require training data but an idea of what kind of information is required from the data. Applying an unsupervised learning algorithm for a classification requires a number of individual differentiable entities, like classes, or a measure of distance or similarity. Reinforcement learning consists of a reward-punishment system while training. This means that after each model run the chosen strategy is checked as to whether it leads to better or worse results. Variations leading to better results are repeated more often in this learning process than those variations leading to a worse performance (Goodfellow et al., 2016; Shen et al., 2018). This approach is mostly used for maximization and minimization processes without a known global optimum. The forward-backward modelling type is often referred as deep learning where the structure of the model is rather implicit than explicit.

1.1.3 Agent-based Modelling

A rather novel approach in modelling is agent-based modelling, originating from social sciences and biology. Here, encapsulated software units build the modelled system and act under certain constraints and boundary conditions autonomously (Macal and North, 2010; North, 2014). Agent-based models require a defined rule set for each agent as well as global boundary conditions. This modelling technique allows the investigation of the interplay of model components, and the resulting patterns and is perfectly suited for modelling problems at the interface between human activities and natural systems (Gunkel, 2005). For example, agent-based modelling is part of a planning tool which cities like Cologne use for their flood evacuation system developed by topoCare GmbH. Here, the agents are limited to human workers that help to construct mobile dams and distribute goods among the workers.

All of the mentioned approaches are presented in detail in the related subsections of this thesis. For each subsection a hydrological problem and the specific application was set up. The overall aim of the thesis is to show the marriage between both columns of the work: the data-driven black-box approaches of ML and the white-box approaches with the clearly defined rule sets in ABM. Therefore, the general applicability of both columns is shown first and in the last part of the work a combined model is presented.

1.2 Main problem

Hydrological research requires data and adequate modelling techniques to find solutions to real world problems. Google's Earth Engine stores more than a peta-byte of environmental data in a virtual data warehouse: the cloud (Gorelick et al., 2017). The database includes data from the last 40 years with new data coming every day. In this time span of its existence, a large volume of spatio- and temporal-explicit data was collected in the Google Earth Engine.

Alphabet does not manage and collates this enormous amount of data alone, the European Environmental Agency ESA, the US-American National Aeronautics and Space Administration NASA, the US geological survey USGS and many more also provide environmental data at a high spatio- and temporal resolution. To analyze this data and to gain new insights is a major task for data science in environmental sciences. Nevertheless, the advantages of big data archives is on the downside, limited by the processing capability of the researchers. Especially in areas where not all significant rules of interplay are known, e.g. for flow in the vadose zone and the interaction of water with the soil matrix or the separation of flood events from continuous time series of runoff, it would be a massive improvement to let the machine search for patterns in data which was the major motivation to investigate the possibilities of ML (Samuel, 1959). Moreover, coupled models with combinations of different systems, like human societies and natural systems could reveal further information on future development like climate change or alterations in vulnerable ecosystems.

Generally, the evolution of information theory, ML and big data are driving forces for the evolution in industry and science. All terms are synonyms for the democratization of knowledge: Computational power became cheap and accessible so that knowledge from data is not a privilege of few but in the hands of many. Thus great economic and scientific interest is in the application of these approaches. The German gross domestic product could be increased by 11.3% until the year 2030 solely by artificial intelligence and big data in all branches (Kirschniak, 2018). Companies like Alphabet and Facebook are dedicated to generate money from data and the analysis of private data for advertisement. Also in environmental science data is a driving factor. In 1998 the European Union passed the regulation act EC 1376/2006 to share environmental data for the application in decision making in environmental matters (EEA 2018). By 2018, most of the European member states shared their data via the open portal of the European Environmental Agency. The USA and Canada share their data on national databases that are also free to access for any interested user. Long-term state-owned satellite-observation programs like Landsat, Sentinel-2 or SPOT are free to access and store billions of gigabyte of spatial and temporal data. Recently, European administrations started to share their environmental data, like measured runoff at gauges with the EEA data base and national databases. To access these massive amounts of data, novel approaches in data mining, modelling and hypothesis building are required that overcome burdens and limits of established approaches. Whether these approaches are data- or knowledge-driven is not obvious and depends on the available data, the quality of the data and the nature of the question to the data. This means, that the choice of method depends highly on the desired information to get from the data to find a solution for the given problem.

In the hydrological community the availability and the democratization of data led to many studies investigating the application of data-driven approaches in data mining, modelling and decision making (Chaney et al., 2018). As progress on the computer science front was made, the different approaches took their time to diffuse into hydrology. The complementary avenue (Shen et al., 2018) introduced the path hydrological modelling has to take, if system understanding moves its focus from top-down to bottom-up, trying to understand a hydro-

logical system by the structures within the data. Furthermore, the data-driven bottom-up approach shows a lower bias in choosing model components than the traditional way of model development where a defined structure is fitted to the data.

"With data, opportunities arise" (Shen et al., 2018) – but challenges as well

The availability of large data sets created new opportunities to model complex environmental situations. The interpretation of ML results is often not trivial, and only rarely transferable to similar yet different problems. Moreover, the information content of data moves into the spotlight of research (Solomatine and Ostfeld, 2008). Consequently, the information content is of higher importance than the pure quantity of data. Because novel data mining and modelling approaches are data-driven, noisy or information-poor data negatively influence the performance. The information content limits of the explanatory power of data, which then influences the choice of approach for further investigation as well as reliability of the generated information from the analysis. The No-Free-Lunch-Theorem states that there is no one-solution-fits-all approach but that an algorithm that fits well with a certain problem will also be able to solve a similar problem with a degraded performance (Wolpert and Macready, 1997). Following the No-Free-Lunch-Theorem, a detailed analysis of structurally different approaches has to be considered before a setup can be chosen prior to the actual analysis process (Ho and Pepyne, 2002; Wolpert and Macready, 1997). To find the approach that degrades the performance less and requires less work for the implementation, is the major task in data-driven science or data science (Han and Kamber, 2010). The process to choose a suitable setup requires the abstraction of a complex problem to fit a variable model structure. The influence of the expert is controversial in the scientific discussion: While some authors clearly state that data-driven approaches support the scientist (Solomatine and Ostfeld, 2008), others claim that any influence of the expert in the data-driven process is some level of bias (Shen et al., 2018). So, one has to find a balance between these two poles: The expert is of importance as long as the structure of the data-driven approach is hidden. Knowledge-based approaches are per-se biased by the researcher, because of the a-priori formulated rule sets.

Two major branches of modelling approaches exist, black-box and white-box methods (Fig. 2). While the prior tries to link in- to output without adding knowledge to the system function, the latter puts the description of the system by rules and constraints in the foreground. White-box modelling represents the knowledge-driven modelling approach. Black-box modelling is a metaphorical understanding of data-driven approaches. The data is transformed in a box to a known outcome, without strict limitations of the fitting within the box.

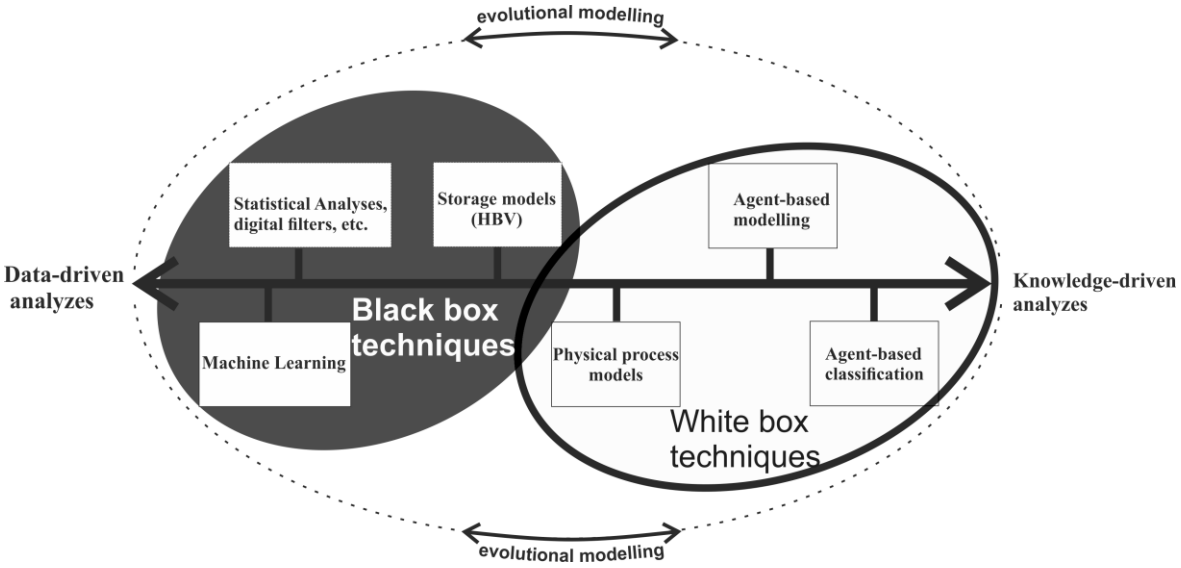


Fig. 2: The diametric dimensions of modelling approaches: black-box data-driven approaches and knowledge-driven white-box techniques. Evolutional, or adaptive modelling combines the advantages of both worlds in a novel approach that is unique in hydrology and water resource research.

In hydrology and water resource management, the application of Machine Learning and big data has an extensive history. Nevertheless, the attitude towards results from data-driven approaches is reserved and often derogative (Shen et al., 2018; Solomatine and Ostfeld, 2008). The black-box character of data-driven approaches a bone of contention in discussions about the usability of the results, because of the limited interpretability of the internal structures in applied sciences like hydrology. Data-driven structures are hard to interpret and sometimes not reproducible when the input data changes. Hence, a detailed analysis of the resulting algorithm has to be given in addition to performance and accuracy measures.

Deep learning (Goodfellow et al., 2016) tries to move Machine Learning from black-box to a more comprehensible data structure where the found structure of the model is in the spotlight of any analysis. The inner core of the adapted model remains intact and only the outer layers are adapted to the specific problem. Because of the relative youth of this approach, deep learning is just about to find suitable applications in hydrology and water resource management (Shen, 2018). The alternation of the data mining process between white box and black box modelling at runtime could be aggregated to the term evolutional modelling (Fig. 2) and could be achieved by deep learning structures. Agent-based models (ABMs) on the other hand represent a variant of white-box-modelling with clearly stated rules (Macal and North, 2010). Nevertheless, ABMs allow the identification of unforeseeable interactions between model components and model outcome because of autonomous software units that try to fulfill their goal under the same environmental conditions and thus lead to dynamic results (Mewes and Schumann, 2018b). The strictly formulated rule sets are a major burden for ABMs. Additionally, ABMs are often criticized because of their limited generality, transferability and the limited application to real world data (Bruch and Atwell, 2015). Although, big data archives require dynamic data mining approaches like agent-based computation, usage is rare because the strict definition of rules hinders their application as well as the

computational demand to run these models.

1.3 Aim of this work

In this thesis the fundamentals of Machine Learning and agent-based modelling are presented to introduce adaptive agent-based modelling as an alternative analysis approach in hydrology and water resource management. By a series of consecutive case studies a combined approach utilizing methods from both ends of the modelling sphere (Fig. 2) is developed: The adaptive agent-based modelling. By mixing both worlds, a highly dynamic modelling and analyzing technique is created. Therefore, a variety of problems is discussed in order to show promises, opportunities as well as traps and burdens of these novel modelling approaches.

This study covers the complementary avenue by Shen (2018) for an application of learning modelling approaches in hydrology and water resource research. The complementary avenue states that big data archives, powerful Machine Learning approaches and knowledge-driven interrogative approaches need to be combined to derive new information from the novel possibilities available in data science to profit from the development of those tools. Hence, application-related studies of the new tools are needed that cover the comparison and the combination of the aforementioned tools and approaches. The diametric dimensions of the two different modelling approaches are represented by the two outer columns of the arc (Fig. 3). On the one side the possibilities introduced by Machine Learning are discussed while on the other side the virtues of agent-based modelling are presented. The differences of both worlds are bridged by the keystone of the combined approach, the so called adaptive Agent-based modelling (aABM). Here, aABM combines the strengths of both columns. Nevertheless, the application of ML and ABM requires conceptual understanding of both modelling hemispheres. Hence, the thesis will be presented as the construction of an arc. Each side of the arc will be constructed individually with the keystone as the final bridging element between both sides.

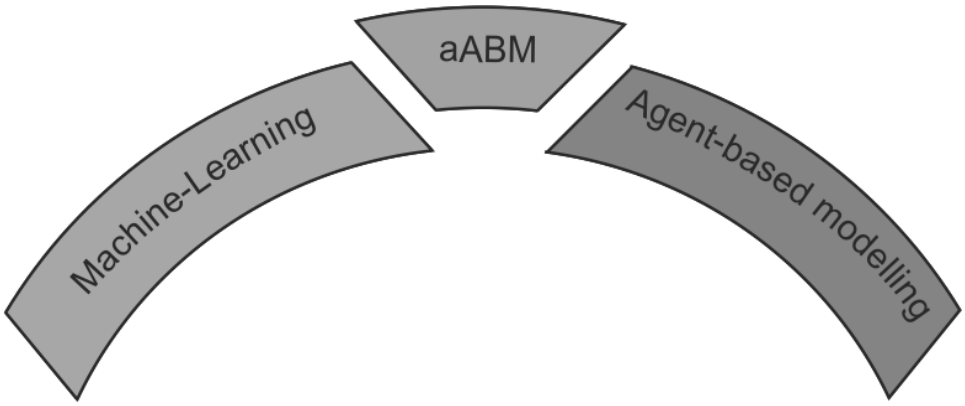


Fig. 3: Arc of data-driven and knowledge-driven modelling presented in this work, bridged by the keystone of the combined approach, the adaptive agent-based modelling.

For each of the techniques applied a short introduction and two hydrological case studies were conducted. It starts in the second chapter with fundamentals of ML and an application of a comparative ML framework to separate flood events from time series of discharge by the application of established machine-learning algorithms. It was shown that a comparative analysis of ML results allows a regionalization of trained machines which is the first step towards deep learning and adaptive dynamic models. To keep results comparable, each of the techniques was compared to the results from manual separation of events. This case study revealed that ML helps to ease the separation of rule sets. It depends on the training data which patterns are found in the data. Apart from a single approach nearly every ML algorithm was able to detect these patterns in chunks of runoff data.

In the third chapter of the thesis, information-theory based criteria were introduced to judge the information content of data. These criteria helped to choose the most suitable ML approach to predict natural tracer signatures from discharge in karstic environments. Moreover, the pre-analysis of the input and training data allows to judge interpolated time series without having data for comparison. Results show that it was more complex to predict tracer concentrations from runoff than separate runoff events. The complexity was expressed by the non-existing preference towards an algorithm but a dependency of region, data and tracers. Nevertheless, it can be shown that by a combined approach of different learning strategies the ML algorithms are able to predict tracer concentrations to a certain level.

The forth chapter of the study is separated into two parts. Here, agent-based modelling is introduced in its fundamentals. A preliminary framework to model the flow of soil water is presented to underline the applicability of agent-based models for physical hydrological problems. The chapter reveals technique-specific problems like the dependency of performance to a chosen scheduling method. In contrast to storage-based conceptual models, ABM consists of a multitude of autonomous objects with an explicit spatial setting. So, scheduling has an influence on which water is allowed to flow first. This problem was overcome in storage models by intelligent numerical approaches. Here, the simultaneity of processes requires new ideas to schedule the behavior of the ABM. Overall, results are very promising in comparison to a known hydrological modelling framework.

The fifth chapter aims to promote another new application of agent-based computation: Agent-based image analysis of remotely sensed images. For a well-investigated region, the state of Nebraska, USA, the capability of agent-based image classification was investigated. It could be shown that ABC was able to use fuzzy knowledge in the interpretation of unknown scenes and remote sensing images that otherwise could not be used in the process of image interpretation. For the delineation of irrigated agriculture from spectral remote sensing data, the approach is a good fitting especially if the temporal resolution of the data is low. First results show that the approach is highly sensible to the quality of the input data but generally it is able to compete with traditional image classification approaches.

The sixth chapter presents the synthesis of the previous chapters implementing a deep learning architecture with ABMs under the premises of big data, adaption dynamics and pattern recognition in structured data with less strict formulated rules of agent behavior. Therefore,

a simplified irrigation model from medieval Bali was implemented based on the findings of Lansing (2007). The simplified deep learning approach in a multi-layered ABM was used to adopt thresholds to changing environmental conditions.

In the final conclusion and outlook future developments and planned research are described and briefly outlined. The here presented adaptive agent-based modelling is a promising new way to combine both modelling techniques and to profit from both approaches. aABM may overcome conceptual limits especially of the ABM. Meanwhile the strength of the ML approach is captured by the dynamic adoption to altering conditions. The introduced case studies are revisited and possible new applications are developed from the experience from the complementary approach.

2 Machine Learning applications in hydrology and water resource management

Learning algorithms are the logical cause of the magnitude of available data. From its very beginning, ML was seen as a prospect to lower programming work in data-rich environments to gain information from the available data (Samuel, 1959). The term ML itself vanished because of the low availablity computational resources in the time of its first appearance in the 1950s - 1960s. With the dawn of powerful cloud computing environments, ML reappeared in the shadow of soft-computing approaches. Soft-computing (SC) is an umbrella term in information theory that covers computational approaches that analyze data with a low vulnerability towards imprecision and uncertainty to gain a robust, low-cost result (Zadeh, 1996). It unifies keywords like machine-learning, fuzzy sets and probabilistic reasoning (Bonissone, 1997). The SC approaches are not per se competitive but rather form partnerships of distinct methods to solve the questions that the researcher posed (Zadeh, 1996). ML evolved in the past decade from the shadow of soft-computing and is now a field of research within the computational engineering and the statistical and mathematical faculties (See et al., 2007; Solomatine and Ostfeld, 2008).

2.1 Fundamentals of Machine Learning

Within the SC approaches Machine Learning (ML) has evolved as one of the most promising tools to retrieve information from large data sets in an automated manor (Goodfellow et al., 2016; Kelleher et al., 2015; Han and Kamber, 2010). ML is basically "*a […] program [or algorithm] that learns from experience E with respect to some task T and some performance measure P, if its performance on T as measured by P improves with experience E"* (p.2 , Mitchell, 2010). So, ML algorithms adapt a variable structure to some known results in order to rebuild the general problem by the given data. ML algorithms either classify unknown targets, or solve a regression problem for the estimation of a continuous variable. ML approaches are more feasible approaches and often less cost-intensive than traditional modelling approaches (Domingos, 2012). Using data-driven approaches is of great importance if a) the data base is large enough to cover the relevant phenomena and b) it is hard to build purely knowledge-driven models (Solomatine and Ostfeld, 2008; Shen et al., 2018). To conduct the classification or to solve the regression problem, the chosen algorithm has to be fitted to the problem (Kelleher et al., 2015). Here, the correct question to the data has to be

found: What is the key information required from the data? Most relevant in the process of setting up a ML model is the generalization of the problem. It is highly unlikely that the exact pattern of the training data is repeated in the validation or testing data, so the focus has to lie on the characteristics of the pattern (Domingos, 2012). The characteristics have to be derived from the data and thus need to be objectified. Taking the separation of runoff events from a continuous time series of runoff as an example, possible characteristics identified by the ML could be the steepness of recession curves, the peak flow or many other combinations that may not be obvious from literature but remain hidden in the data until the ML approach discovers them. The question limits the obtainable results as well as the choice of algorithm and the minimal amount of data used for training.

For the training of the algorithm, various strategies are available: unsupervised and supervised training. The combination of both strategies, the semi-supervised learning is not treated in this thesis. A rather novel addition to this choice is the reinforcement learning that represents a problem specific trial and error scheme with gratification and punishment (Goodfellow et al., 2016; Shen et al., 2018). While the supervised learning strategy requires a set of examples to train the algorithm to the problem, analyzes the unsupervised strategy and the existence of natural breaks in the data. Un-supervised classification approaches require a clearly stated idea of what should be found in the data. Due to the highly complex nature of hydrological problems, e.g. of flood types or irrigation strategies, the exact number of distinguishable groups within the data cannot be defined a-priori. Reinforcement learning approaches that make use of interrogative techniques to improve their system understanding require a lot of training data and scenarios. Moreover, the fitting of a deep learning model core is work intensive and should be avoided for simpler problems (Bengio, 2009).

For most of the problems presented in this work, a supervised learning strategy was applied. This means that a trustable set of training data was required to fit the algorithms (presented in the following subsections). The increase of training data in combination with a random selection of samples from the training data simulated the growing data base and the unknown additional information content by the newly added sample. Kelleher et al. (2015) group the available ML approaches into four different families: information-based, similarity-based, probability-based and error-based learning. In this thesis, a focus was on two of these four families: information-based and error-based learning techniques. None of the probability-based approaches were included due to the hardly objective choice of a suitable probability density function for continuous regression tasks in the specific field of flood event separation. Moreover, the subjectivity to specify a measure of similarity hindered us from using a similarity-learning approach for the problems in this study. In the following sections, the chosen set of ML approaches is presented. All presented approaches were implemented using Python Scikit-Learn library (Pedregosa et al. 2011).

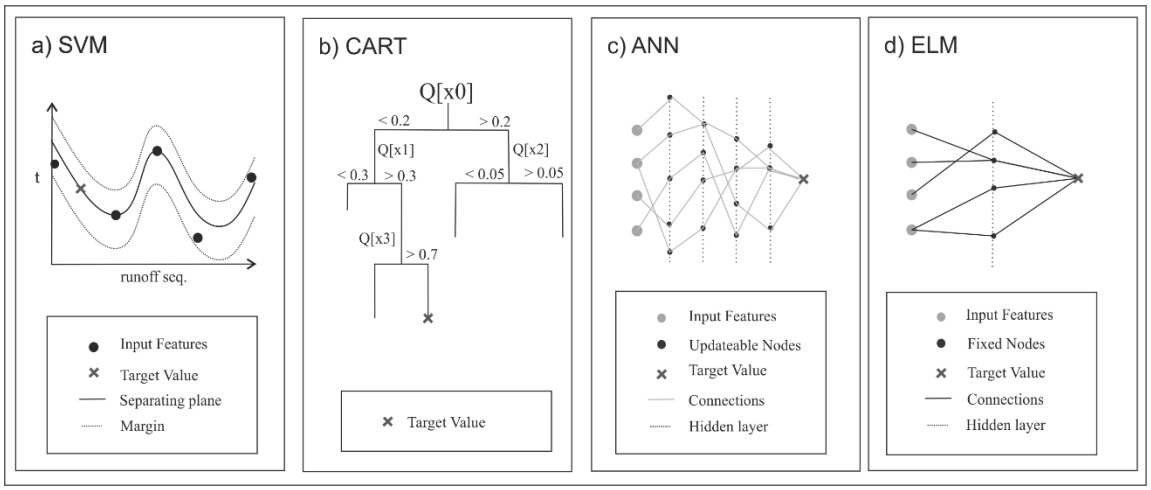


Fig. 4: Applied ML-based approaches in this study, covering a) the SVM, b) the CART-based regression tree, c) the multi-layered ANN and d) the forward propagating ELM. The schemes show examples of how the approaches are used to solve hydrological problems. The SVM creates a separating hyperplane, while the CART represents a cookbook to follow. ANN and ELM are networks of neural nodes that alter an information on its way through the network towards the desired target.

2.1.1 Support Vector Machine

A Support Vector Machine (SVM) originates from the family of error-based learning methods. It tries to find a decision boundary to either classify the unknown target into a group of predefined classes or to solve the regression task (Cortes and Vapnik, 1995). The fitting of the algorithm takes place in a high dimensional space, where the margin between the nearest features to the boundary (the so-called support vectors) is maximized (Fig. 4a). In the higher dimensional domain the hyperplane is adjusted for the vectors closest to the separating hyperplane. By retransformation the higher dimensional result is transferred back into the dimension of origin.

A SVM model is defined as in Eq. (2.1), where q is a descriptive feature, $[d_1 \ldots d_s]$ are support vectors and w_0 is the first weight of the decision boundary, α describes a set of parameters that is optimized while fitting the hyperplane and t_i represents the unknown target. The product is a Lagrange multiplier with the parameters α, d and w_0 that leads to a constrained quadratic optimization problem which has to be solved (Kelleher et al., 2015).

$$M_{\alpha,w0}(q) = \sum_{i=1}^{s} (t_i \times \alpha[i] \times (d_i \cdot q) + w_0) \tag{2.1}$$

In most cases the input features have to be mapped to a higher dimensional space, as the input data does not solve the regression task in the original feature space or the separation is not obvious. Therefore, a kernel function migrates the data into a higher dimension until the regression task becomes solvable. In this study, a radial-base-function (RBF) kernel is used to transform a non-separable problem into a higher dimension in order to minimize training time (Chang et al., 2010). Other kernels, like a linear or a sigmoid kernel, vary the results only slightly, so to keep results simple, only the results with a RBF kernel are shown here.

For more in-depth information on SVM see Vapnik (2013).

To solve the SVM, a set of parameters has to be set. In this case the penalty term C and the term ε, for defining the margin that defines the border after which no penalty is given because of classification errors, need to be determined. *C* is set to 0.1 in order to lower the influence of a misclassification and have a smaller margin of the hyperplane. The parameter ε describes the distance from an actual value not causing error in the fitting process to the dividing hyperplane. It can be set to 0.1 because of the normalization of the input. Moreover, the error is further lowered if the hyperplane is too narrow. The parameters remains constant over the fitting and are not target of further improvement or optimization.

2.1.2 Classification And Regression Tree (CART)

Regression trees represent a guide book to make decisions in classifications in form of a branched tree. The information tree follows the ID3 (Iterative Dichotomiser 3) method where a tree of nodes with certain decisions of the character of an element *x* is analyzed to eventually be classified as a member of a class T_i. At each node a specific characteristic of *x* leads to either the next node or finally the leaves that represent the target classes. As shown in Fig. 4b the element x0 is checked whether it is smaller than 0.2 that leads to the next node where element x1 of a sample is checked whether it is higher than 0.7 leading to the result that the target has the value of class "red" in this example. The here applied information-based Decision Tree Regressor originates from the CART family (Breiman et al., 1984), and was also derived from the ID3 algorithm (Quinlan, 1986; Kelleher et al., 2015). Regression trees are built node per node with a successive reduction of regression error between the estimate and the true value (Fig. 4b). As a limiting parameter the maximum depth can be defined. This describes the number of nodes from the root to the leaves. Here, the depth is set to 1/10 of the available input features. The data point showing the highest impact on the reduction of error in the regression becomes the root of the tree. The other branches follow the hierarchy of error reduction leading to a further ramification of the tree. Finally, the branches result in leaves which are the regression result or the final labels assigned to the unknown object. To quantify the reduction of regression error, often the Root Mean Squared Error between target and estimated value is used. Generally, the decision tree acts like a guidebook to segment the input to data, leading to the desired regression outcome.

Neither boosting nor bagging are implemented in the CART. The maximum depth of the tree is set to half the number of input values to avoid overfitting. The minimum depth is 3 taking in short runoff windows all values into account. Generally, a CART like algorithm is part of the complementary investigation because its structure is easily interpretable like a guide book through the data.

2.1.3 Artificial Neural Network

The Artificial Neural Network (ANN) is a classification and regression approach that is in-

spired by the structure of the human brain (Goodfellow et al., 2016). Input features and targets are connected within a network through a number of hidden layers and nodes (Fig. 4c). This means that a sequence of input features is dissolved into single parts and guided through the network. The influence of the input features is varied through various hidden layers and nodes. On each hidden layer an activation function maps the input feature to the next layer. The connections between layers and nodes have updateable weights that control the mapping of the input feature to the desired outcome (Haykin, 1999; He et al., 2014). Whether a neural node is activated or not is controlled by the activation function. Once a threshold is reached the node actively influences the outcome by its weight. The weights of connections and nodes are updated during the process of fitting the network to the data. As the update information is just passed backward through the network, the process is called backpropagation.

The ANN used in this case-study is based on a multi-layer perceptron using a stochastic gradient descent for optimization (Goodfellow et al., 2016). As the number of hidden layers cannot be determined unbiased and has to be estimated by trial and error, the initial setup of the ANN was kept simple in the case study, limited to 1 hidden layer with 1,000 neurons. A detailed analysis of the influence of the ANN geometry can be found in Sec 2.5.4. As the default activation function the Rectified Linear Unit function was chosen. Like in a CART tree every input value has a certain influence on the regression outcome. The weights per connection and node lead to the final output layer and vary the numerical influence of each input feature on the results.

As activation function between the layers, the choice fell on the *tangens hyperbolic* (*tanh*) function *tanh* as it is a rather smooth s-shaped activation function for small data sets (Ingrassia and Morlini, 2005). Other common activation function like the rectified linear unit (ReLU) can be ruled out due to the non-linear data hydrological data represent.

2.1.4 Extreme Learning Machine (ELM)

An Extreme Learning Machine is a special form of an ANN with fixed node weights from input to the hidden layer (Fig. 4d). The name “extreme” originates from the learning speed of this approach. Due to the limitation of updated weights, the ELM learns faster than comparable MLPCs. The connections from the nodes of the hidden layer to the outcome are updateable (Huang et al., 2004). The weights of the nodes are estimated randomly in the first iteration of fitting and remain the same for the whole process of fitting. The simplification of the update process results in faster learning while the regression output remains stable in comparison with the ANN and big data sets. The mathematical definition of an ELM is given in Eq. (2.2) where the number of hidden nodes is N, the activation function $g(x)$ for each sample x and the weight β of the vector connecting the hidden node and the output node (Baraha and Biswal, 2017)

$$f(x) = \sum_{i=1}^{N} \beta_i g_i(x_i) \tag{2.2}$$

As activation function *g(x)* also the *tanh* function is utilized. By the reduction to only one layer, an ELM trains faster than an ANN and has the smallest training error (Baraha and Biswal, 2017).

2.2 The No-Free-Lunch-Theorem – addressing the problem of model choice

The plethora of available ML approaches results in a non-obvious choice for the relevant problem. The *No-Free-Lunch-Theorem* (Wolpert and Macready, 1997; Ho and Pepyne, 2002) addresses exactly this problem: If an arbitrary method performs well on a certain type of problems then it will achieve a degraded performance on the remaining problems. For example if a ML algorithm is able to separate runoff events on an hourly data base then the algorithm will also be able to separate events on a daily data base but with a degraded performance. Eventually, this theorem states that all approaches will perform with a degraded performance and a varying amount of training effort to achieve the performance (Schumacher et al., 2001). Wolpert's No-Free-Lunch theorem states that no algorithm can beat random guessing in cases that the data is uniformly drawn from all mathematically possible functions which is not the case in real world problems (Wolpert and Macready, 1997; Domingos, 2012). In the quintessence, there is no overall preference towards a single approach but the choice heavily depends on the question and the available data. Hence, multiple different approaches have to be tested and compared especially in the data-driven approaches where the solution might be biased by the data choice and less by expert decisions.

For the application presented in this thesis, the No-Free-Lunch-Theorem means that for any real world application, like hydrological or water resource management questions, a set of different approaches has to be considered and compared in terms of their respective ability to solve the question and retrieve the desired information from the data. Thus, a set or even a combination of soft computing approaches might be a suitable way to derive the desired information solely from the given data and a general recommendation cannot be given (Solomatine and Ostfeld, 2008).

2.3 Machine Learning based temporal flood event separation

In the following section the aforementioned approaches were tested for their ability to separate single runoff events from continuous time series of runoff. Flood event analysis is a widely used approach in hydrology to characterize the reaction of a watershed to a rainfall event (Maidment, 1993; Blume et al., 2007). Although a great variety of tools exist to extract single runoff events from continuous time series of runoff, none of them is applicable in all cases: manual separation requires heavy workload (Blume et al., 2007; Hall, 1968), tracer-based methods on the other hand need large databases that are often not available (Klaus and McDonnell, 2013). Although the latter allows a separation of the runoff into single components that can be linked to distinguishable processes within the catchment (Weiler et al.,

2018), high costs and the lack of applicability on historical data sets are downsides of this method. Recession-based approaches often require manual correction or computational intense recalibration (Tallaksen, 1995; Hammond and Han, 2006; Mei and Anagnostou, 2015). Hydrologists defined several rule sets to separate flood events manually (e.g. Furey and Gupta, 2001), which reduces the flood event separation to a pattern recognition problem. Thus, the intention was born to automate the flood event separation by ML. A similar, yet not comparative approach for baseflow separation was conducted using ANNs (Corzo and Solomantine, 2007) or digital filters (Chapman, 1999). None of these studies investigated the influence of the algorithm choice on the outcome or the amount of training data needed to achieve suitable results.

2.3.1 Adaption of ML algorithms for flood event separation

With the help of the ML algorithms, the beginning and the end of a runoff event were estimated. The ANN and the ELM both consisted of a single hidden layer with a *tanh* activation function and 1,000 neurons to map the input values to the desired outcome. The SVM was fitted automatically with a RBF kernel to map the problem to a higher dimensionality. Other kernels were tested but were rejected due to worse separation results or longer computational time. The ML approaches tried to identify the markers of the beginning and the end of the flood event. The training data and the true data for reference were derived from manual separation. The ELM and the ANN required at least three training storm events to converge, whereas SVM and CART did deliver results with only one sample. So, the minimum amount of samples for all approaches was set to 10 runoff sequences to avoid underfitting.

ML approaches require input data of the same length, so a window of runoff with the desired event has to be cut from the complete time series. Therefore, the complete time series of runoff was divided into windows of the same length with the peak of the event as the center. For the case study in this work a window length of 200h with the peak as the center was chosen (Uhlemann et al., 2010; Nied et al., 2014). Within the 200h the beginning of the event as well as the main part of the recession curve in small- and medium-scale catchments should be covered. Both time steps of interest, the beginning and end of an event, were represented through markers (subdivided into t_{Start} and t_{End}). Consequently, both t_{Start} and t_{End} varied between 0 and 200. All 200 runoff values from the snippet were taken as input values. Apart from the cutting of the time series into chunks, no further pre-selection of data took place.

2.3.2 Data choice and pre-processing of runoff data

For this case study, hourly runoff data from ten different Bavarian catchments were taken from 1961 until 2015. Those events leading to the highest monthly discharge were considered as the events of interest. The runoff was normalized by the gauge specific mean discharge. The target for the automatic separation by the ML approaches are the temporal markers of the begin and the end of the flood event. Events with missing data were eliminated from the data base. As mentioned before, the training data was derived in the shape of windows of 200h length from the highest daily discharges from the ten different catchments.

The number of events was nearly even for each catchment leading to ca. 40-50 events in the database per catchment.

In order to determine the needed amount of training data for a successful learning of the ML, algorithm the training data set was constantly increased from 10 - 95% incrementally by 1% of the available data. Three major basins are the organizational units in which the ten catchments are located (Fig. 5). The three major basins: Regen, Main and Iller represent different natural units with varying sizes, mean altitudes and physio-geographical properties (Tab. 14, see appendix A).

The Iller basin is the highest of all catchments in this study with a mean catchment height > 1.500 m a.s.l. The catchment is located in the northern Alps of Bavaria that consist mainly of calcit rock (Landesamt für Umwelt Bayern, 2017). With a size of 35.6 km² it is smaller than the other catchment of the Iller basin: Immenstadt-Zollbruecke. Located in the lower parts of the Bavarian pre-alpine area it covers 724 km². 110 events are related to both catchments of which Immenstadt-Zollbruecke comprises 60 and Birgsau 50 flood events.

In the Regen basin, the size of the four catchments varies between 115.9 km² (Lohmannmuehle) and 2590.4 km² (Marienthal). The catchments are located in the geographical unit of the Donau-Isar gravel plains. Because of the river bed geology in gravel plains, intermediate storage is higher than in the alpine catchments of the Iller basin. The gradient of height in these catchments stretches from the East to the West with the highest mid-mountain range at the eastern border in catchments Lohmannmuehle and Zwiesel. In total, 242 runoff events are available for the Regen catchments. The variation in numbers of events per catchment is low, because the minimum number of events is 56 (Kothmaissling) and the maximum reaches up to 65 (Lohmannmuehle).

The rivers of the Main catchment do not discharge into the Danube like the rivers in the other two main basins but eventually in the river Rhine and the North Sea. The four catchments range in a size between 11.1 km² (Friedersdorf) and 165.3 km² (Lohr). While the eastern parts still have mid-mountain heights of > 1.000 m a.s.l, are the western parts located in the lower parts of Bavaria. The natural units in which the catchments are located in vary between plains (Friedersdorf, Lohr, Untersteinach) and hillsides (Gampelmuehle).

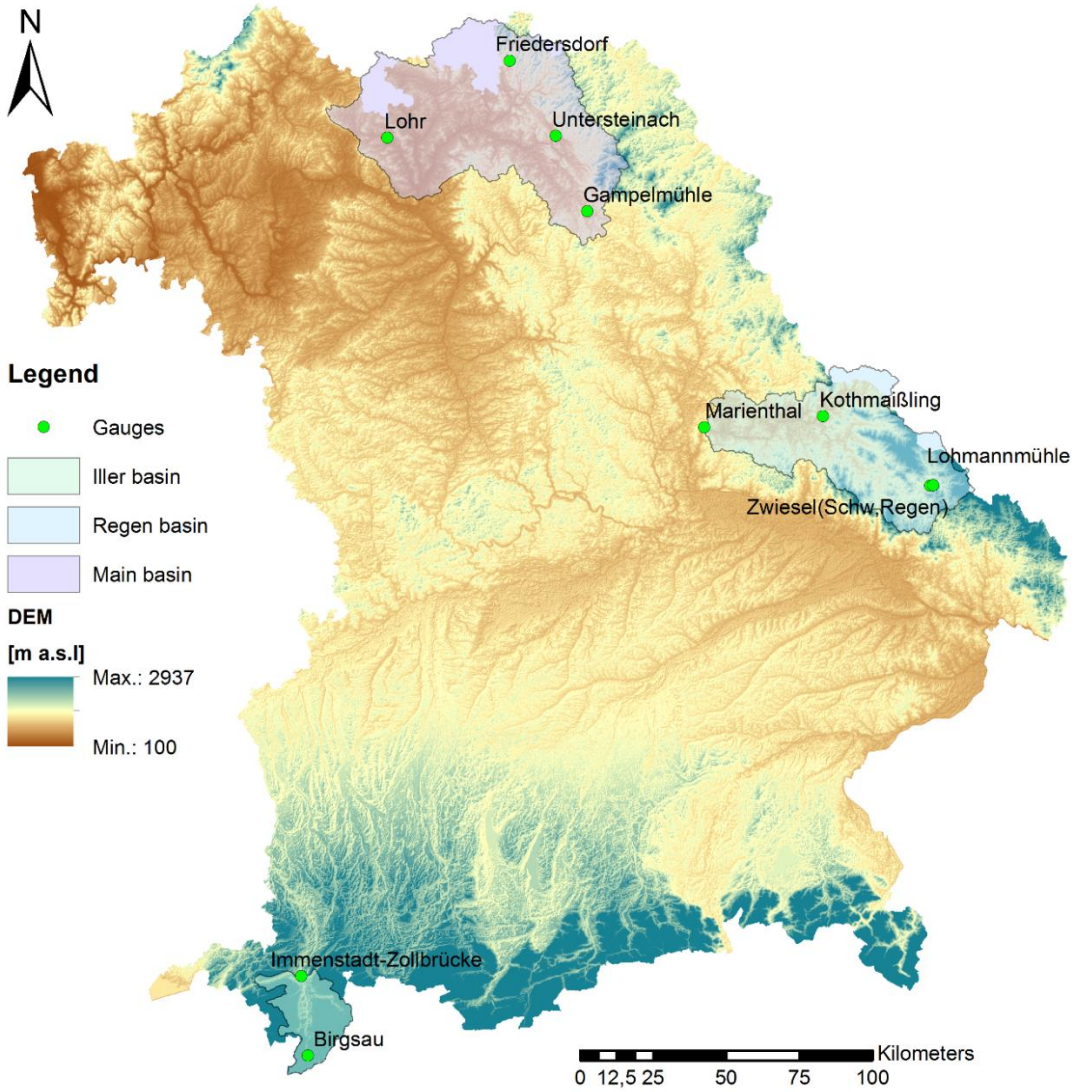


Fig. 5: Map of Bavaria with topography from SRTM data (Jarvis et al., 2008) and the location of the 10 catchments that were considered in this study. The catchments are grouped into three major basins (Iller, Regen, Main) with different geographical characteristics.

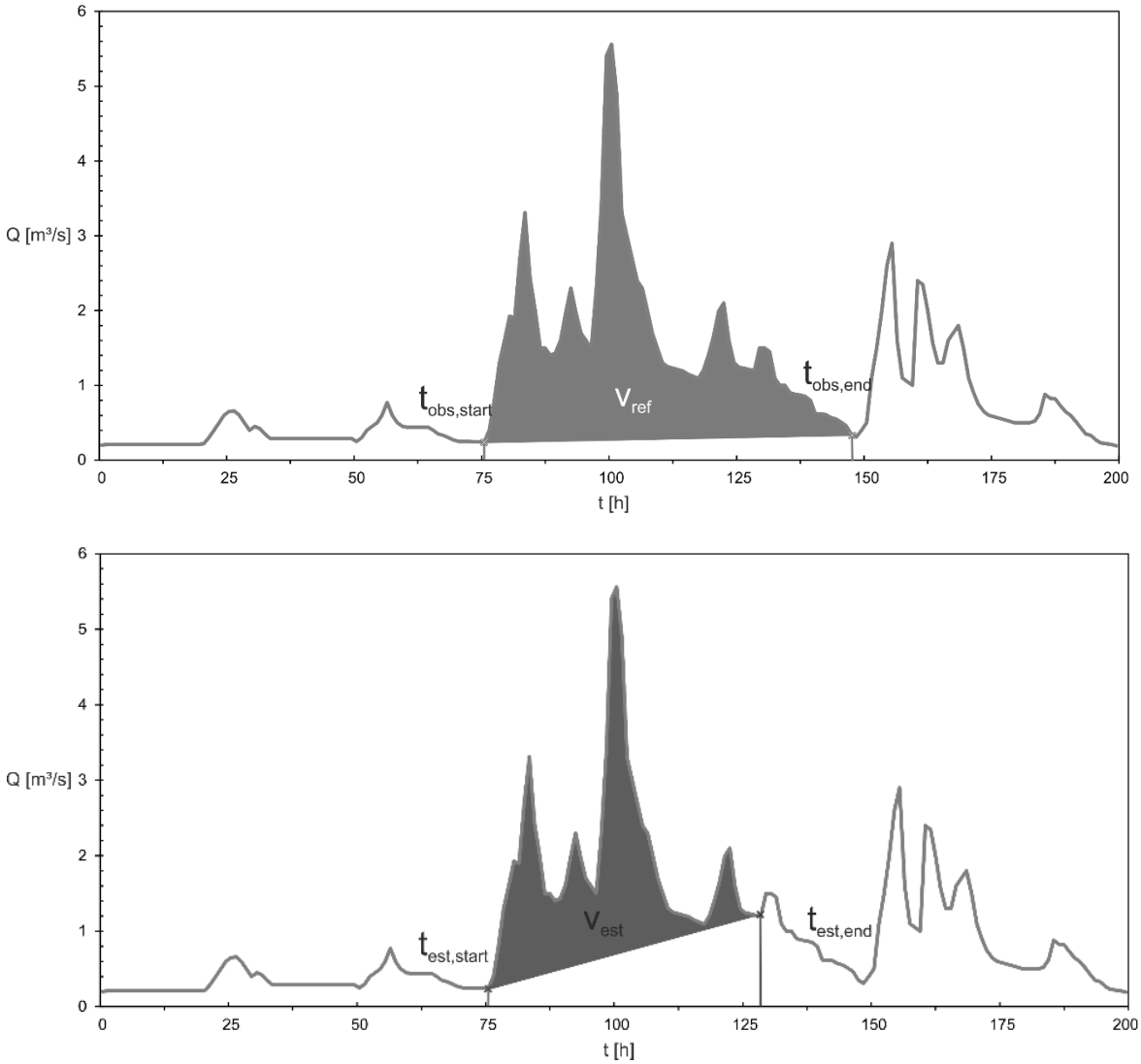


Fig. 6: Hydrograph separation problem, covering a flood event with a single peak embedded in a continuous time series of discharge. The blue marked area represents the manually separated reference event, whereas the red marked event is estimated by the machine-learning algorithm. The straight line connected markers delineate the baseflow from direct runoff describing the catchments response to an event

The time steps t_{start} and t_{end} represent the target variables of the four chosen ML approaches (Fig. 6). Due to their continuous nature, the ML approaches solved a regression problem but not the classification problem as the target is not a per se defined class but a numerical value representing the time step. In the next step the performance metrics have to be defined to judge the ML performance (Domingos, 2012).

2.3.3 Manual separation rules for training data

As input feature and single explanatory variable the complete window of runoff data was taken. A manual separation of the flood events acted as reference. The reference events were derived by the rules by Furey and Gupta (2001):

1. The beginning of an event is the starting point of the rising limb around the peak. The end is characterized trough recurrence to the recession of the hydrograph before the event.
2. Discharge at the ending point Q_2 has to be bigger than discharge at starting point Q_1 in order to separate an event as a single event.
3. In case that a subset has more than one peak (local maxima), all peaks belong to one event if the following local minima $Q_{2,i}$ is higher than $Q_{2,i+1}$ where i denotes the index of peak in the subset. In case that $Q_{2,i+1}$ is equal or higher than $Q_{2,i}$ the subset is divided into multiple events.

Knowledge on the domain (here the catchment specific runoff characteristics) added more information which cannot be modelled by traditional approaches where expert knowledge is often hard or impossible to incorporate without losing transferability of the model (Solomatine and Ostfeld, 2008). Naturally, every choice of a reference value for separation adds a bias to the result, but this bias is systematical within the application. This systematic bias only limits the significance of the determined preferred algorithm for this specific choice of algorithms but not the general applicability of ML approaches for flood event separation. An unbiased general classification scheme for flood event types would significantly lower the bias of the reference value.

2.3.4 Performance metrics to judge separation quality

To measure the performance of the ML-based separation, two characteristics of the hydrograph were chosen: the volume of the direct runoff and the temporal coverage of the ML separated events with the benchmark data, derived from the manually separated events (Fig. 6). To judge the volume error of the separated events, the RMSE of volume was calculated, where N denotes the number of events, v is the volume divided into volume of observed event v_{obs} and the volume of the estimated event v_{est}:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (v_{\mathrm{obs}} - v_{\mathrm{est}})^2}{N}} \tag{2.3}$$

To show the direction of volume error, the Mean Volume Ratio (MVR) was used, which utilized the same notation of variables like in Eq. (2.3). The optimum value of MVR is 1.0, where the mean volume of events separated by ML algorithms and manually separated events was equal:

$$MVR = \frac{\sum_{i=1}^{N} \frac{v_{est}}{v_{obs}}}{N} \tag{2.4}$$

The temporal coverage Cov is a criterion that relies on a position comparison between true event and estimated event. The time intervals (positions) covered by separated and true event were analyzed Eq. (2.4), where Int denotes the intervals of either true or estimated event divided by the number of true intervals only regarding the amount of positions which belong in both scalars:

$$Cov = \frac{\#\left\{i \in (\mathrm{Int}_{est} \cup \mathrm{Int}_{true})\right\}}{N_{Int_{True}}} \tag{2.5}$$

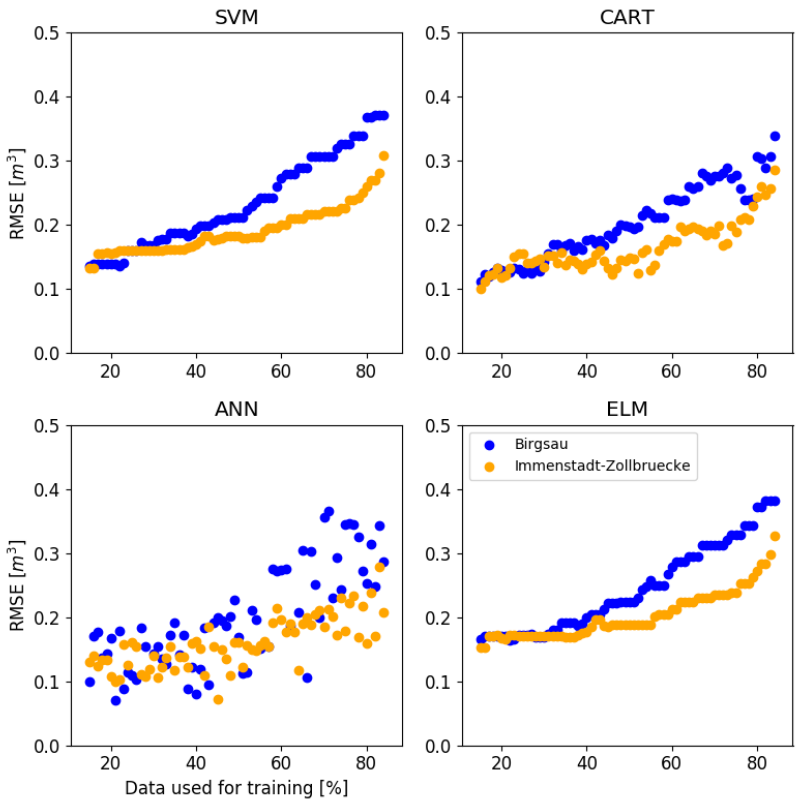
For example, the true event starts at interval 95 and ends at interval 102 and the estimated event starts at interval 96 and ends at interval 103, the relative coverage comprises intervals [96, 97, 98, 99, 100, 101, 102] which would result in a relative coverage of 0.875 showing that 87.5% of all intervals in the estimated event are also covered in the true event.

Among the investigated ML approaches, the preferred approach was determined. Therefore, it was investigated which approach delivers the best performance measures with the smallest amount of training data. So, in case that e.g. SVM was able to separate events with a similar volume (MVR converging to 1.0) with the lowest amount of training data and the temporal coverage is close to 100% using SVM, the choice of approach was obvious. If, in a different catchment, the RMSE was lowest and MVR close to 1.00 using the ELM and only 15% of the available training data, but Cov was highest using SVM, the preferred algorithm was ELM, because of the focus on the volume as the main characteristic of interest. Consequently, a relative ranking facilitating the choice of the approaches was derived.

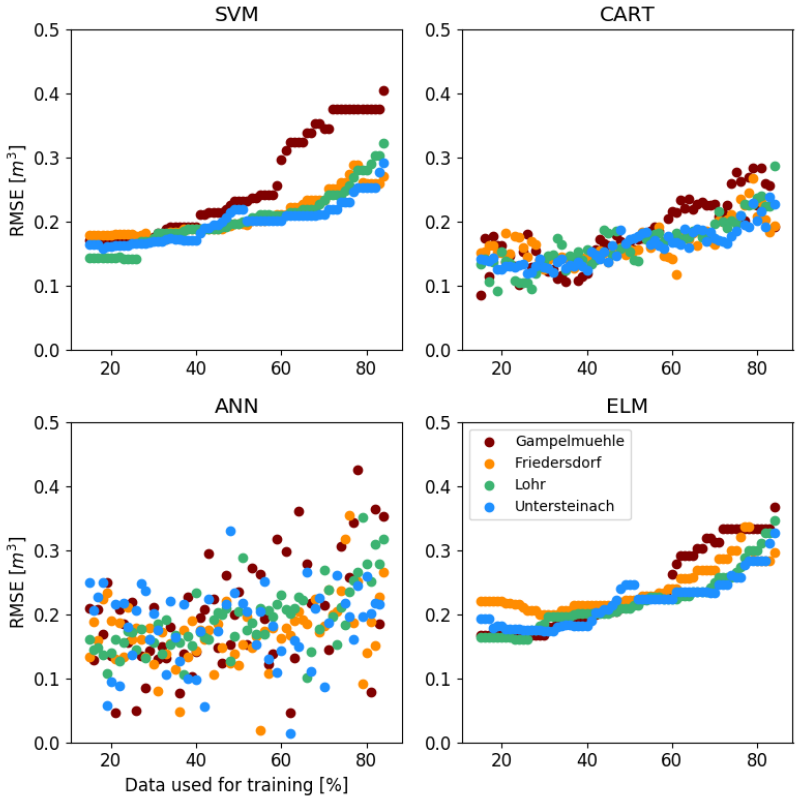
2.4 Separation results

2.4.1 Individual machines per catchment

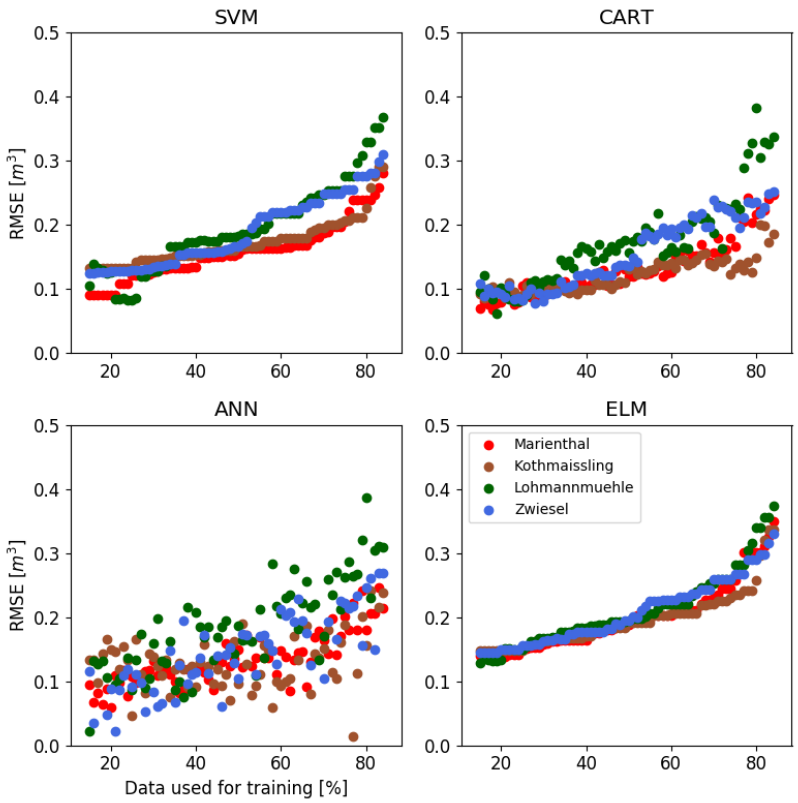
Considering the RMSE of the separated volume in contrast to the manual separation, the general development showed that the volumetric error increased the more data was used for training. This rise of error was observable over all approaches (SVM, CART, ANN and ELM), yet the shape is different (Fig. 7). Moreover, the catchments located in the Main basin (Lohr, Friedersdorf, Untersteinach and Gampelmuehle) showed with RMSE ~ 0.2 – 0.3 m³ a lower volumetric error than those in the Iller basin (Birgsau and Immenstadt-Zollbruecke), that have a final RMSE ~ 0.35 – 0.4 m³. Generally, the RMSE was low over all approaches regardless of the region or the chosen approach. No preference towards any method became obvious although in some regions the RMSE has risen like a hockey-stick (e.g. Immenstadt-Zollbruecke) whereas in other regions the RMSE rose linearly (e.g. Birgsau or Zwiesel). Catchment Gampelmuehle behaved differently than all the other catchments, reaching a plateau of stable error using ~ 65% of the available data for training.


SVM
CART
ANN
ELM
RMSE [m^3]
Data used for training [%]
Birgsau
Immenstadt-Zollbruecke

a)


SVM
CART
ANN
ELM
RMSE [m^3]
Data used for training [%]
Gampelmuehle
Friedersdorf
Lohr
Untersteinach

b)

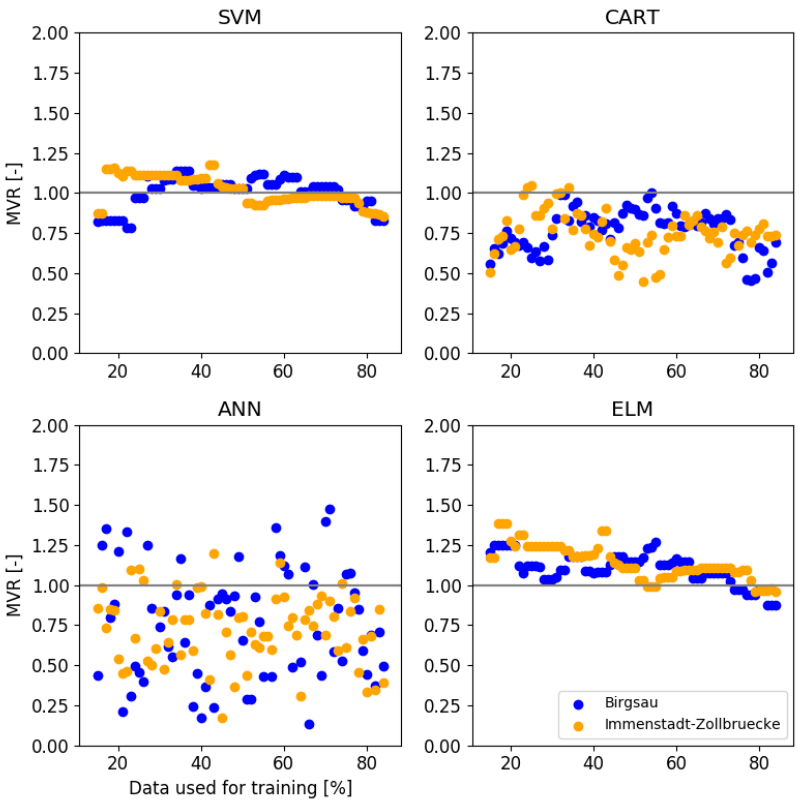


c)

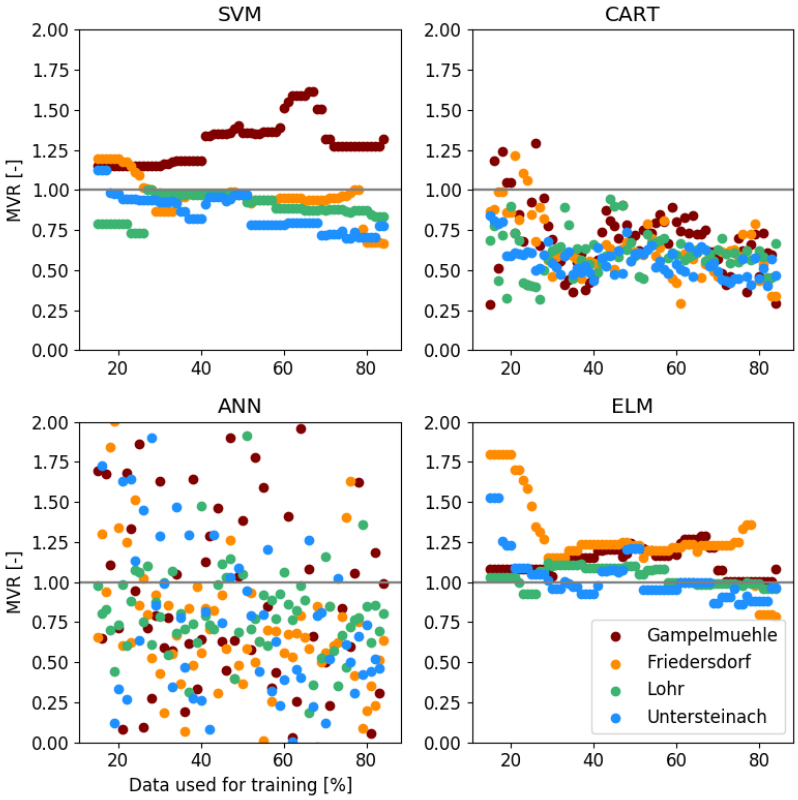
Fig. 7: RMSE of volume covering all approaches in all catchments (a,b,c). The more data is used for training, the higher the RMSE of volume gets. Most catchments show a hockey-stick behavior. By the final level of error, a choice cannot be made among the approaches.

In contrast to the evaluation of separated events by RMSE, a preference towards SVM and ELM was visible for the evaluation by MVR (Fig. 8). Apart from catchment Gampelmuehle, the SVM improved the results with a MVR close to the optimum of 1.0 using mostly 20-30% of the available data for training. CART generally underestimated the volume of the automatically separated events. The ANN was not able to deliver stable results in any of the catchments while the ELM with its special structure was able to score stable and good MVR scores but with a higher demand for training data.

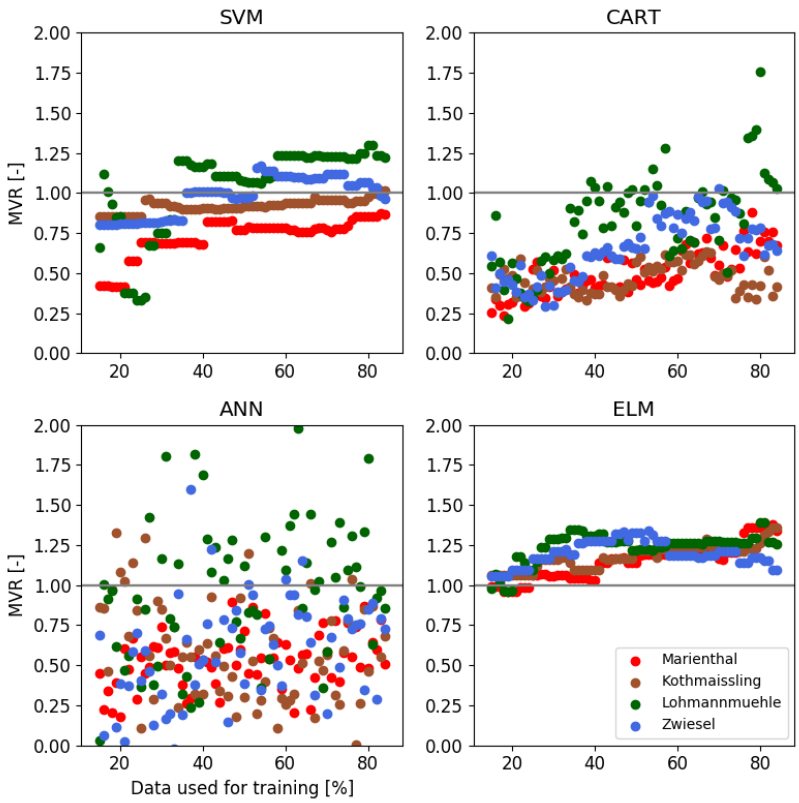
Considering the Cov of the automatically separated events, again SVM and ELM delivered the best results covering >60% (SVM) respectively > 80% (ELM) of the true event (Fig. 9). Again CART scores lower values than SVM or ELM. In the Main catchments (Marienthal, Kothmaissling, Lohmannmuehle and Zwiesel) the CART results improved, whereas in the Regen catchments (Gampelmuehle, Friedersdorf, Lohr and Untersteinach) deteriorated the more data is used for training. Again, the ANN did not show any influence of the amount of training data on the performance metric.


SVM
CART
ANN
ELM
MVR [-]
Data used for training [%]
Birgsau
Immenstadt-Zollbruecke

a)

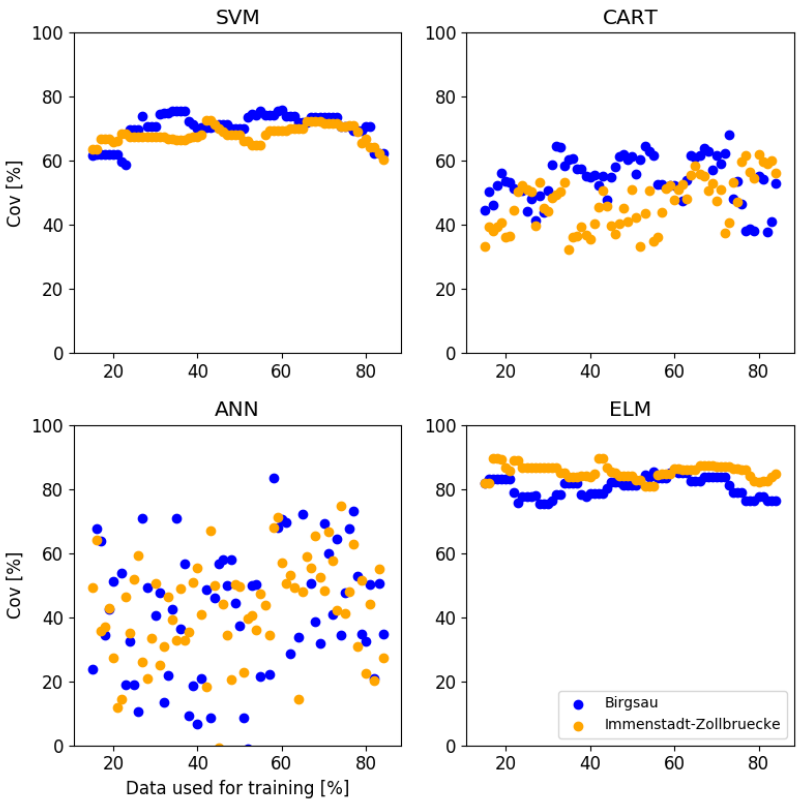

SVM
CART
ANN
ELM
MVR [-]
Data used for training [%]
Gampelmuehle
Friedersdorf
Lohr
Untersteinach

b)

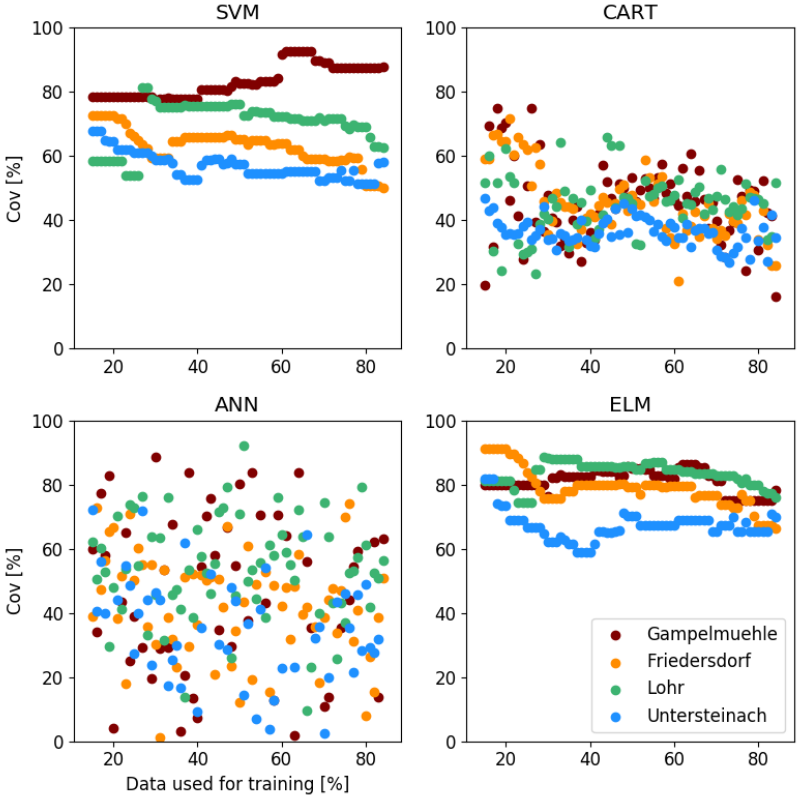


c)

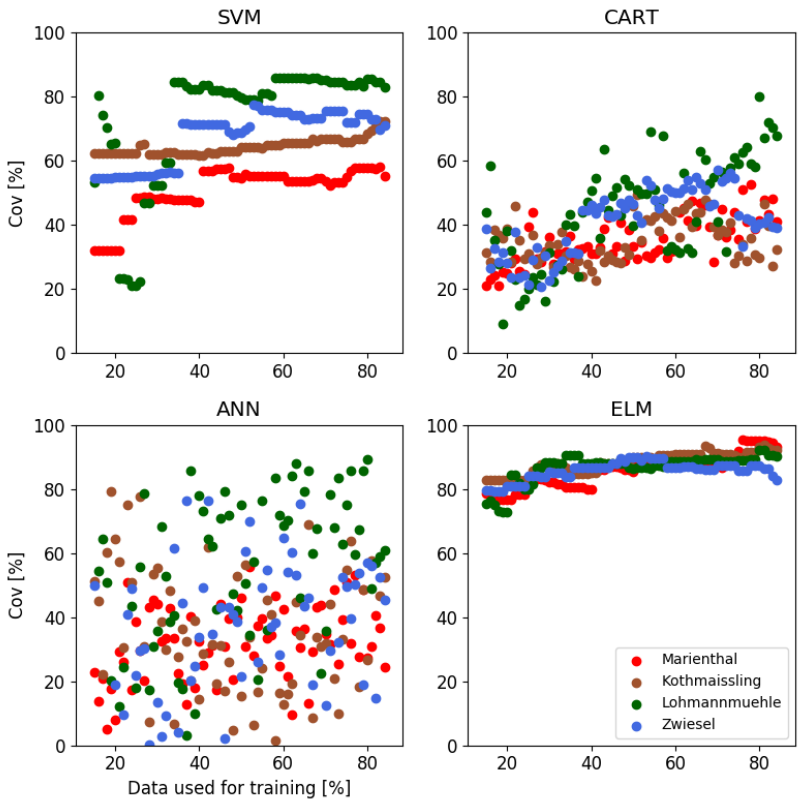
Fig. 8: Mean Volume Ratio of all ML approaches (a,b,c) over all catchments. Differences in terms of volume estimation capability become obvious, favoring SVM and ELM.


SVM
CART
ANN
ELM
Cov [%]
Data used for training [%]
Birgsau
Immenstadt-Zollbruecke

a)


SVM
CART
ANN
ELM
Cov [%]
Data used for training [%]
Gampelmuehle
Friedersdorf
Lohr
Untersteinach

b)



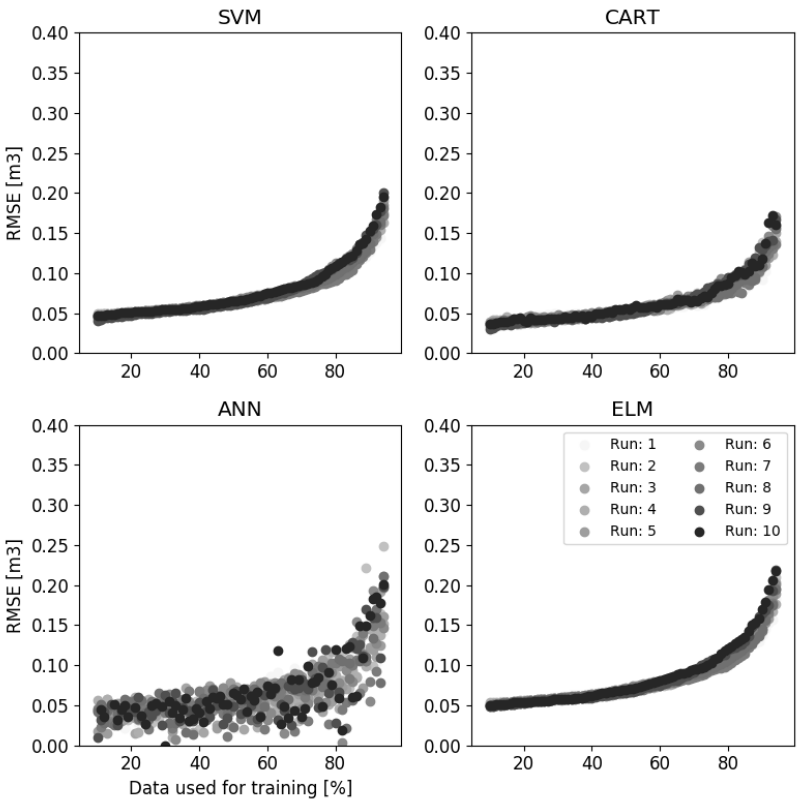
c)

Fig. 9: Coverage (Cov) of all catchments using all four approaches (a,b,c). 100% coverage marks the optimum value. ELM seems to be the method of choice over all catchments scoring Cov values > 80% with less than 20% of available data used for training.

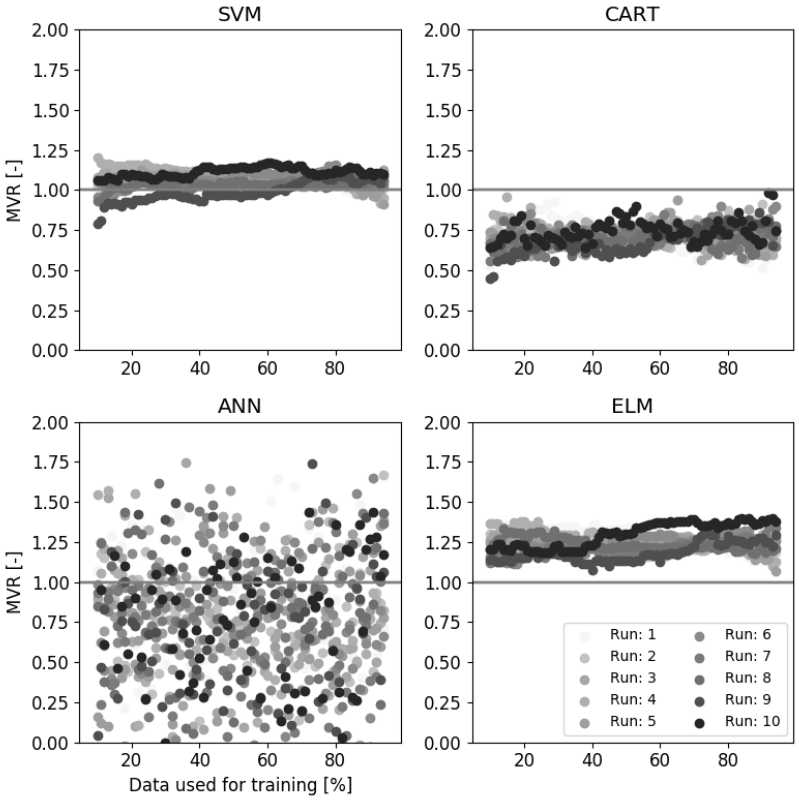
2.4.2 Separation results of global machine

Next to the individual fitting of a machine per catchment, the applicability of a global machine was tested as well. Therefore, all available training data was merged to a global data set and resampled 10 times to shuffle the composition of the data for each individual run and to avoid a selection bias. This merge resulted in a data base of 537 manually separated events. The separation results of the global machine were in line with the results from those of the individual machines, strengthening the preference towards SVM and ELM (Fig. 10). CART again underestimated the volume. ANN did not deliver stable results in any of the runs. The resampling revealed only slight variations within the runs. Hence, one can assume that the variation within the composition of data has only slight influence on the results of the separation. The application of a globally trained machine showed that one machine is sufficient to separate flood events automatically with only few (< 10 manually separated) training events.

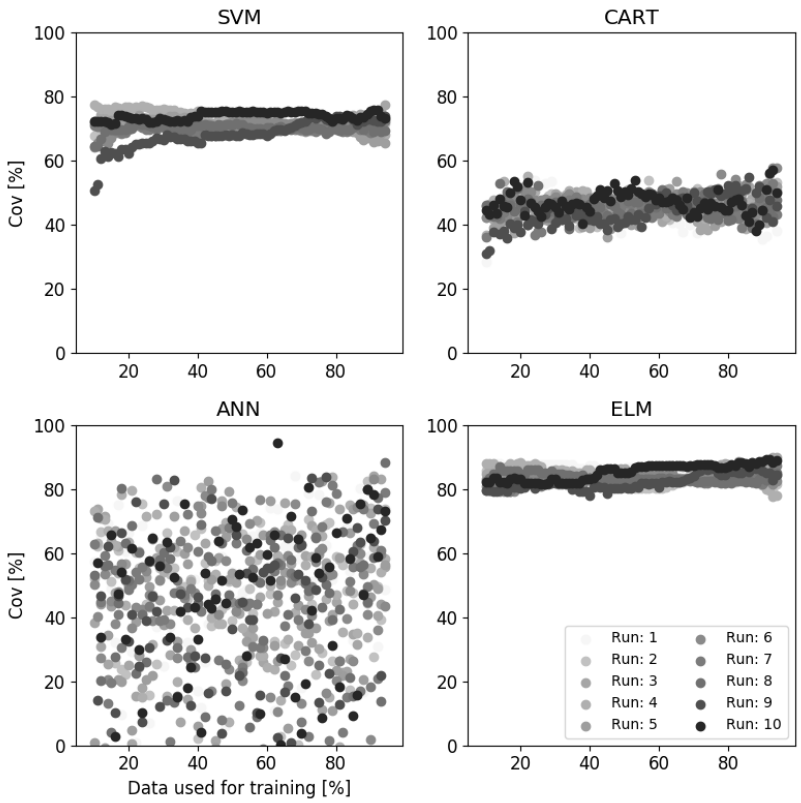
Because of the low number of required events for training, one can assume that not all catchments are represented in the training data. Hence, the global machine was able to separate flood events in catchments where no dedicated information is available.


SVM
CART
ANN
ELM
RMSE [m3]
Data used for training [%]
Run: 1
Run: 2
Run: 3
Run: 4
Run: 5
Run: 6
Run: 7
Run: 8
Run: 9
Run: 10

a)


SVM
CART
ANN
ELM
MVR [-]
Data used for training [%]
Run: 1
Run: 2
Run: 3
Run: 4
Run: 5
Run: 6
Run: 7
Run: 8
Run: 9
Run: 10

b)



c)

Fig. 10: Separation results from global machine for all three performance measures (a,b,c). The training data was resampled 10 times to avoid selection bias.

2.5 Discussion of automatically separated events

Generally, ML based approaches were able to automatically separate single runoff events from continuous time series of runoff. Yet, as mentioned in other ML studies in hydrology the choice of the ML approach has an impact on the quality of the results and the so-derived conclusion whether the chosen ML strategy suits the problem or not (Solomatine and Ostfeld, 2008; Raghavendra and Deka, 2014; Shortridge et al., 2016).

Returning to the initial research question: Are ML algorithms able to detect patterns in runoff and are they able to separate flood events from continuous time series of discharge? Considering all catchments individually, one can see that two approaches stand out: SVM and ELM. The preference of algorithm per catchment could be found in Tab. 1. Both delivered stable results in terms of MVR and Cov with only a small training data set. On the third place was the CART algorithm that was influenced stronger by the training data. CART either got better or worse, whereas ELM and SVM nearly showed constant performance using small training data sets (comprising less than 10 events per catchment). Due to their similar structure ELM and SVM played a head-to-head (Liu et al., 2012): SVM performed better in terms of MVR whereas ELM scored higher Cov values. Consequently, ELM estimated the events longer than SVM because Cov reached 100% if all true time steps were also covered by the

automatically separated event. This finding was underlined by studies showing the capability of SVMs in hydrological applications (Raghavendra and Deka, 2014; Tabari et al., 2012; Yu et al., 2017)

If the ML overestimates the length, Cov might stay at 100% but MVR would be higher and less close to the optimum. An overestimation of the coverage has less influence on the separation performance, because the volume of the event is slightly overestimated. Thus, the MVR is the most important ratio to judge the goodness of ML choice as the RMSE shows only slight variations. The results from the global machine underlined the findings from the individual machines: SVM and ELM tend to be the most suitable tools for this specific case. Additionally, one can derive from the merged data set that a globally trained machine is able to separate flood events automatically, even in catchments not represented in the training data.

Tab. 1: Ranking of ML algorithms per catchment

No.	Name	Main basin	1.Choice	2.Choice	3.Choice
1	Birgsau	Iller	SVM	ELM	CART
2	Immenstadt-Zollbruecke	Iller	SVM	ELM	CART
3	Friedersdorf	Main	SVM	ELM	CART
4	Gampelmuehle	Main	ELM	SVM	CART
5	Lohr	Main	ELM	SVM	CART
6	Untersteinach	Main	SVM	ELM	CART
7	Kothmaissling	Regen	SVM	ELM	CART
8	Lohmannmuehle	Regen	ELM	SVM	CART
9	Marienthal	Regen	ELM	SVM	CART
10	Zwiesel	Regen	ELM	SVM	CART

2.5.1 Comparison of ML derived events with recession-based flood events

To judge the separation performance of the ML algorithms a comparison is conducted with Blume's constant k approach (Blume et al., 2007). Blume's constant-k method tries to fit multiple linear functions to the hydrograph to determine the change of the slope k. Thus, this approach analyzes the change of slope to determine the end of a stormflow event. To fit the linear functions a suitable window width for the slope analysis has to be found. This window length does not have any measurable components and is therefore a variable to be calibrated. The calibration of Blume's constant-k method was conducted with the same events that were used for training of the ML algorithms. The width of the observing window Δt was estimated

by a Monte-Carlo simulations with 10,000 cases where the length of the time window and the sampling of training data is varied.

In order to identify the beginning of the event the rising of the hydrograph is analyzed for the first dip of discharge that causes the sequent rise of discharge to the peak. As the constant-k approach requires the fitting of multiple regressions for small time-steps, a global parameter set for all catchments was estimated to reduce computational time and to show the general ability of the method for regionalization. Calibration was conducted by comparing the separated volume and the temporal mismatch with the manually separated events. The constant-k method was applied to four different catchments (chosen randomly from all Bavarian watersheds): Gampelmuehle, Lohr, Untersteinach and Friedersdorf.

Using the constant-k method with a global parameter set, the overall volume was separated with a reasonable performance resulting in a ratio of 1.1 - 2.0, but the amount of data used for training did not reflect any increase in performance (Fig. 11). The mean temporal mismatch of end and beginning of an event was variable. The Cov is worse than for the ML derived approaches as one can see in the median temporal mismatch between the events (Fig. 12). Especially in catchment Friedersdorf the median temporal mismatch accounts for 100h. Using a window length of 200h the temporal mismatch already covers half of the observed period. With a median temporal mismatch of ca. 18h catchment Gampelmuehle shows the lowest temporal mismatch. Overall, the Cov and MVR is worse for the recession-based events than for the ML derived events. So, the ML-based approaches are a suitable tool for the separation of flood events from time series of discharge and outperform traditional recession-based approaches like the here presented constant-k approach by far. What becomes obvious is, that the constant-k approach shows severe errors in the quantity of the estimated event. This is mainly due to the fact, that recession-based approaches rarely hit the start of the storm runoff event. But for analyzing floods, especially extreme floods, the behavior of the flood wave at the beginning is of importance. Thus, the recession-based approaches, like the here discussed constant-k, are not suitable and more complete results like those from ML are required.

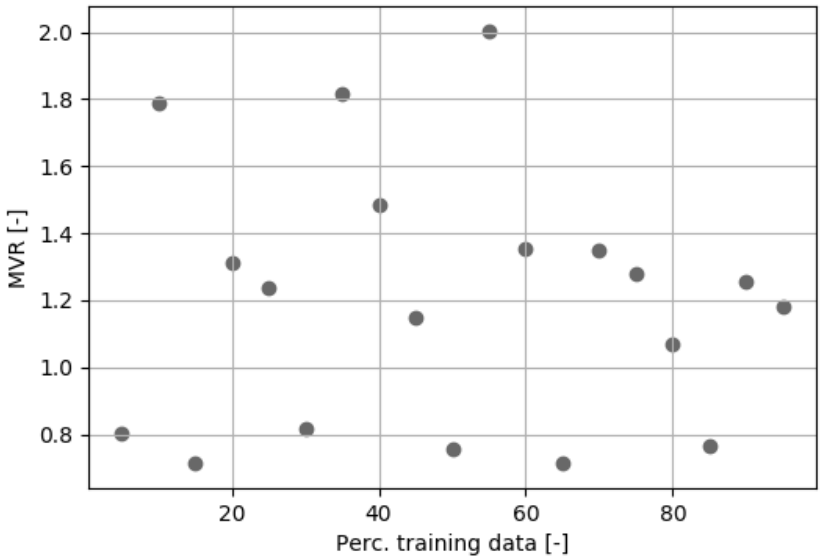


Fig. 11: MVR of constant-k derived events in comparison to manually derived flood events. One can see that the events overestimate the volume massively.

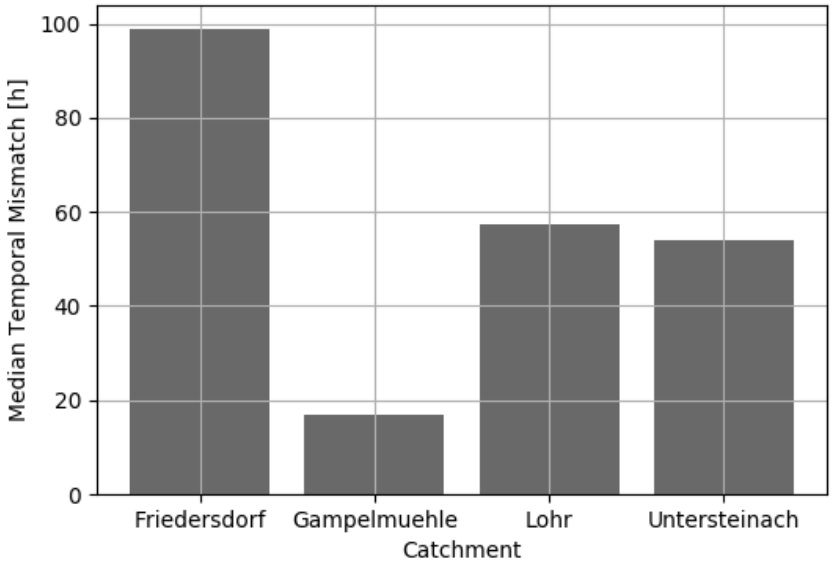


Fig. 12: Median temporal mismatch of constant-k derived events. The highest mismatch becomes visible in catchment Friedersdorf where more than 100 h difference of event length can be observed.

2.5.2 Spatial patterns of algorithm preference

Due to the different ranking of algorithms per catchment (Tab. 1), one can suspect a spatial pattern of the preferred ML algorithm. In the alpine basin of the river Iller, both catchments, Birgsau and Immenstadt-Zollbruecke prefered the SVM for flood event separation as indicated by the red marked area (Fig. 13). In the Main basin catchments on the other side, no clear preference became visible in the spatial pattern. Two headwater catchments (Friedersdorf and Gampelmuehle) prefered different algorithms (SVM and ELM), while similar sized catchments like Untersteinach and Gampelmuehle also showed a different preference towards the most promising ML approach for the task of flood event separation.

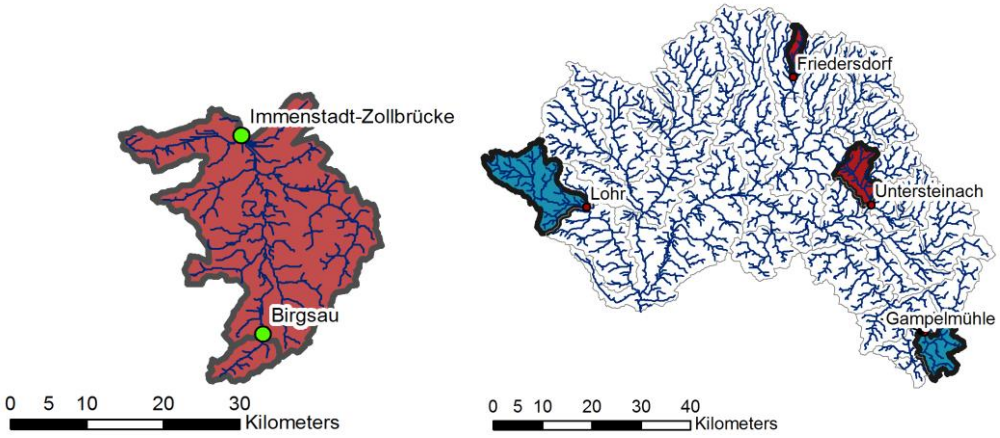


Fig. 13: Preferred algorithm in the Iller basin and the Main basin catchments (red equals SVM, blue equals ELM).

Most of the Regen basin catchments prefered ELM over SVM (apart from catchment Kothmaissling). Again, the information concerning the relative location of the catchment or its size did not have influence on the preference towards any algorithm. Also a direct neighborhood between the catchments did not automatically lead to a shared preference as it could be seen for Marienthal and Kothmaissling (Fig. 14). Interestingly, Lohmannmuehle, Zwiesel and Marienthal are nested catchments and have an East-West flow direction, that showed a preference towards the ELM algorithm. So, one can conclude that there is no obvious relationship between the spatial patterns or preference of algorithm and the spatial catchment characteristics. But one can assume that nested catchments might have a similar preference towards the most suitable ML algorithm. The reason for similarities and dissimilarities must be hidden in the pattern of data and the respective information content of the data used for training.

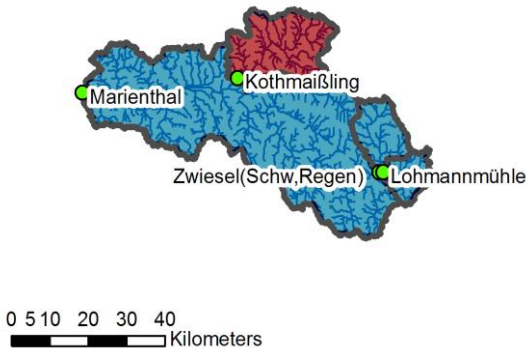


Fig. 14: Preferred choice of ML algorithm in the Regen basin catchments (red equals SVM, blue equals ELM).

2.5.3 Uncertainty induced by window length

The induced uncertainty by the chosen window length of 200h was discussed by a brief discussion on the most frequently chosen start and end points. If the window is too narrow, most of the chosen start and end points will be located at the fringes of the window. Therefore, a CART tree was deconstructed revealing the most important nodes (is equal to the topmost nodes) of the decision tree thus revealing which runoff values are most important for the estimation of the start and the end (Fig. 15). One can clearly see that most of the important runoff values were located within 30h around the peak in each direction, this equals a window length (WL) of 30 steps. Most of the second-most important nodes were located within 50h before and 50h after the peak. As the nodes acted as guides through the tree to the leaves, the target of the guides had to be evaluated as well. A detailed investigation of the final position of the markers revealed that in 0% of all cases the beginning was set to time step 0. In 2 – 6% of all cases was the final marker set at time step 200. Hence, the chosen length of 200h has no negative influence on the decision structure and the relevant information is located around the peak. For the beginning or the end even runoff values on the respective temporal opposite of the peak are of importance, here the pattern is not obvious for the researcher. This shows the need to take all 200 runoff values into account as the information content for the ML algorithms is not necessarily easy to predict beforehand.

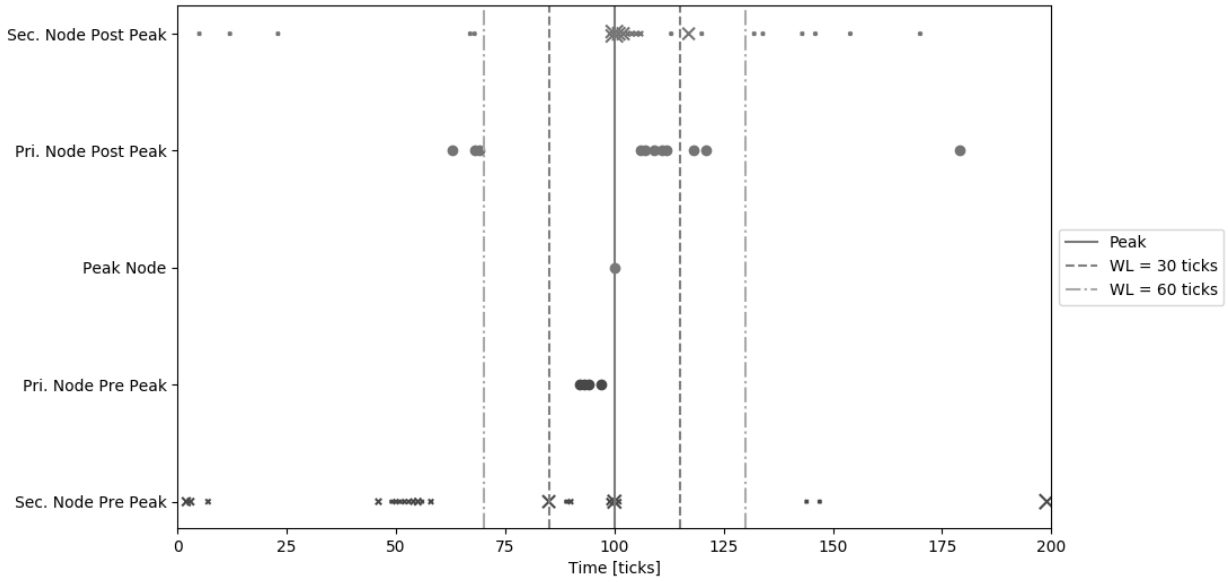


Fig. 15: Positions of the two most important runoff values for the determination of the start and end. The most important information for the separation is located within 60 time steps from the center.

2.5.4 Performance issues of ANN in flood event separation

The ANN did not show stable behavior in anye of the catchments investigated here although the number of hidden layers and neurons was the same as in the well-performing ELM. In order to exclude the geometry from the list of possible errors, the same task was conducted with varying numbers of neurons (500 – 10,000) and layers (1 - 4). All of the tested configurations showed unstable behavior as shown by the MVR over all runs (Fig. 16). Although, less neurons and more layers (configuration L4 / N500 with four hidden layers and 500 neurons each) showed the most stable results over all configurations, the performance was worse than for any other approach presented in this case study. A completely simplified ANN with only 1 Layer and 200 neurons did not converge with an acceptable learning rate in 10,000 iterations and was consequently not presented here.

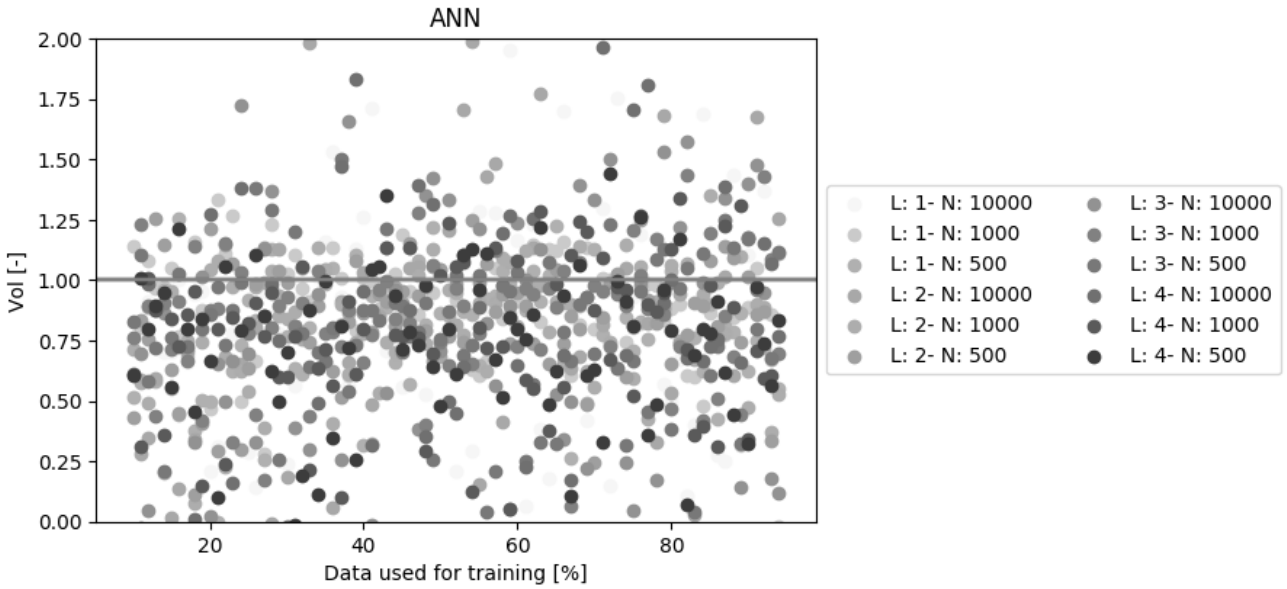


Fig. 16: Uncertainty induced by different ANN geometries.

Overall setups with a lower number of layers, in this case all L1 runs, showed a lower spread of volume error, whereas the complex models showed a higher spread around the optimal value of 1.0 (Fig. 16). Heteroscedasticity did not apply here, because the results shown are independent from the training data. One can derive from this investigation that the ANN with the chosen general structure of backpropagation was not able to detect any patterns in the input data it can link to the output. The ELM, which is a special case of the ANN with forward propagation on the other hand, was able to achieve good scores for the specified problem.

Because the ANN has proven its general applicability in hydrological research, the specific problem of flood event separation was not well suited or the structure of the input data with chunks of time series data did not match the expected structure. In future research a derivative of the ANN specialized on time series analysis will be tested. Furthermore, the MLPT setup of the ANN will again be tested in a different hydrological context to analyze whether the disappointing performance is problem or data-structure specific.

As the structure of the algorithm reveals no hints on the performance of ML algorithms, the reason for the different performances must be found in the underlying data. Therefore the information content and the structure of the information has to be investigated to judge the performance of the ML applications.

3 Information-theory based criteria for data mining

As it became evident in the previous chapter, catchments share similarities not necessarily visible in the Meta information that one can derive from physio-geographical characteristics but supposedly in the internal structure of the data. This means that the information content and the interdependencies within data reveal patterns that ML approaches detect and adapt on. Hence, detailed knowledge of the information content of a data set is crucial for data mining. To measure the information content of data, the principle of Shannon's entropy was developed (Shannon, 1948; Gong et al., 2014; Kelleher et al., 2015). The principle of entropy is a broadly accepted measure of information content within data and originates from signal processing. The information content is expressed by a probability P that the data x contains new information on a given data set. In the context of big data the information content is crucial to save computational time and storage. To find information in data, data mining approaches have been developed. Data mining approaches deliver answers to the questions that are related to big data: How much of the data has to be stored, which data is redundant in terms of information, what variability can be expressed by the data stored?

For any data-driven approach the information content of the data is relevant. Redundant data confuse the ML algorithms because the noise/information ratio gets worse. Consequently, prior to any ML application the information content has to be determined and compared among the available variables and the target value. As the information content can only be investigated if a reliable truth data for comparison is available, a different field of research in hydrology has to be found than flood event separation as the reference data is per se biased and not measurable. Therefore, the underlying data of tracer hydrology is investigated. Here, the natural and chemical tracers allow an investigation of subterraneous hydrological processes on a catchment scale. From the interplay of various tracer signatures, catchment reaction, e.g. to stormflow, can be revealed and explained.

3.1 Tracer prediction in karstic environments by ML approaches

Tracer measurements are often the only way to separate streamflow into amounts per origin. This deeper understanding is often part of a model calibration strategy, especially in highly dynamic environments like karst springs. Single tracers are not always the key to success for a comprehensive understanding of the underlying natural system (Garvelmann et al., 2017; Lee and Krothe, 2001). Hence, pairs of tracers are often compared in order to relate their

dynamics to the dynamics of the underlying system (Mahler et al., 2009; Mudarra and Andreo, 2011; Hartmann et al., 2016; Hartmann et al., 2017; Klaus and McDonnell, 2013). Their dynamics capture the system's behavior and reveal changing interactions of often hidden interactions. Therefore, the joint analysis of combined tracer measurements is a commonly used tool to investigate the interplay that lead to characteristic streamflow compositions (Garvelmann et al., 2017). The choice of tracers for the investigation in this thesis fell on SO_4^{2-} and NO_3^-. While NO_3^- is an indicator for fast water fluxes from the shallow surface (Mahler et al., 2009), SO_4^{2-} is an indicator for slow geogenic contributions from the phreatic subsurface (Hartmann et al., 2017).

As mentioned before, tracer-based methods are among the most widely used techniques to separate streamflow and describe the underlying processes with relatable chemical measurements. As the reproducibility of tracer measurements in situ is impossible, the need for a ML approach to fill gaps or recreate time series of tracer measurements evolves as it allows the prediction of something hard to measure with something easy to measure. Further downsides of tracer-based methods are relatively high costs for permanent measurement setups as well as an increased manual workload to analyze in-situ measured probes. So, the question is whether runoff data, which is available in data bases for more regions than tracer measurements, is able to predict tracer concentrations. This presumes that runoff data has equally the same continuous entropy as the mutual information between tracer and runoff and hopefully between pairs of tracers. The main idea is that runoff values and their dynamics contain enough information to predict the tracer dynamics from the derived patterns (Mewes et al., 2018).

The runoff was taken as daily values from Banque Hydrologique (Eaufrance, 2018b) and tracer measurements of both tracers are derived from ADES (Eaufrance, 2018a) with a sub daily temporal resolution. As ML approaches like SVM and CART are not able to process time series data well, snippets from the time series of runoff were taken (Fig. 17). Per pair of tracer measurement, a snippet was taken, thus the total number of snippets exceeds 1,200 pairs of concentration measurements. The optimal length of the window is unknown. Therefore, a variation of lengths was tested and performance measures like the RMSE were presented as a boxplot covering the complete range of window length.

For the prediction, the ML algorithms were meant to predict the normalized tracer concentration. The main aim was to predict complete time series of tracer measurements from runoff. Therefore, the most suitable ML strategy had to be found. To find the most appropriate approach, two different strategies of prediction were applied: The univariate prediction and the multivariate prediction. The primer fitted a ML algorithm per tracer (resulting in the equal number of machines and tracers to be predicted) while the latter trained one machine for both tracers. In cases that dependencies between the two tracer data sets were observable in the data, one could reveal those dependencies with that analysis. The approach to combine ML approaches for a better prediction capability is referred as complementary approach (Solomatine and Ostfeld, 2008). Here, different sets of machines were combined into a Meta

machine to achieve the best result. For example, if a tracer A' can be predicted by the runoff and the tracer B can be predicted in a better way by the runoff and the information about tracer A, resulting ML setup should train a univariate machine for target A' and a complementary multivariate machine for target B'. Furthermore, by MI and the comparison of learning strategies the direction of information flow could be examined. If tracer B' can be predicted in a better way by incorporating the information from runoff and tracer A, the information flows from tracer A to tracer B.

For cross validation the available data was again divided into training and testing data. To visualize the impact of the amount of training data, the share of training data was incrementally increased by 1% from 5% - 95%. To avoid bias by the ordering, the training data sets were resampled ten times. Once drawn, the pair of tracer measurement was not returned, so doubles in the training data are excluded.

The CART tree was neither boosted nor bagged but the maximum depth was given by half the window length and a minimum of 1 nodes. So, not all runoff values of the snippet were considered in the CART tree and overfitting was prevented. The SVM used all available input data with a RDF Kernel and the penalty term $C = 0.1$ and $\varepsilon = 0.1$ which defined the margin where an error in fitting the hyperplane to the supporting vector was not punished by a penalty. Like in the flood event separation, the ANN and the ELM represented both the neural networks. The latter is a special version of the former, as documented in the section on ML approaches in Section 2.1.4. The ANN and the ELM consisted of one layer and the same number of neurons as half the length of the input window of runoff, the minimum amount of neurons is three. This would consider all input values of the runoff series as relevant for the ML prediction. Smaller windows of input data caused convergence errors and led to erroneous interpretations of underfitted values.

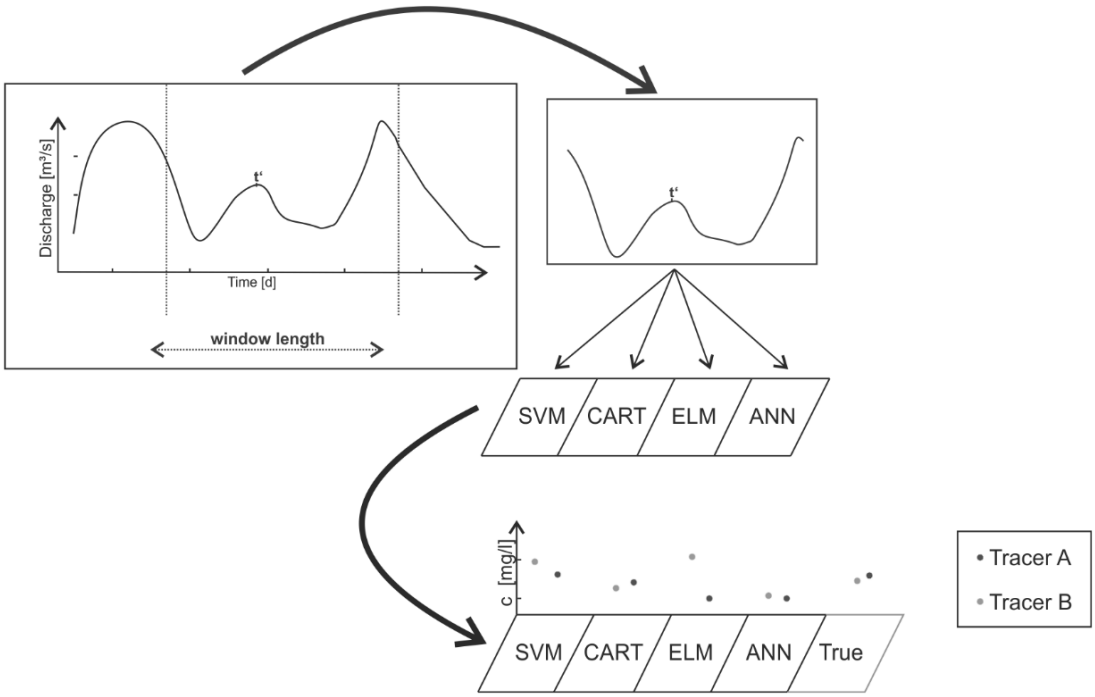


Fig. 17: Schematic application of machine-learning for tracer concentration prediction by windows of runoff.

As mentioned before, the length of the snippet of runoff values from the time series was varied in order to test which length of runoff information was required for tracer prediction. The smallest possible window had a length of three days with a maximum of half a year, which reflected the medium response time of slow catchments. The reference point for the tracer measurement was located at the center *t'* of the runoff snippet (Fig. 17). For each predicted concentration the outcome of the prediction was compared to the true measured tracer concentration. In contrast to the flood event separation, a real-world true value of the tracer concentration was available. Consequently, the information content of the input data as well as the shared information with the output were calculated. As memory plays a role in the understanding of a catchment's reaction, the different windows of input data are tested (lasting up to 180 days).

3.2 Definition of entropy and mutual information

To measure the information content in the tracer data a suitable measure of information content has to be found and defined. The Shannon entropy measures the information content of data by its impurity (Kelleher et al., 2015). For a classification problem, the entropy H is defined by the chance of a sample x_d to be of one of the given classes $\{x_1, \ldots x_{Nt}\}$ with $P(x_n)$ as the probability that $X_t = x_n$ with a sample length N and the data set X:

$$H(\mathrm{X}) = \sum_{n=1}^{l} -P(x_n) \times \log_2(P(\mathrm{x}_n)) \tag{3.1}$$

The above presented Shannon entropy measures the information content in *bit* according to the base 2 of the logarithm in Eq. (3.1). Other units are possible by different bases of the logarithm, but are not necessary for the applications presented in this thesis.

The main problem of Shannon's entropy model is the limitation to discrete target variables. In the flood event separation no classification into defined types is conducted because no true reference classification exists. Because of the limitation of the Shannon entropy to discrete classes in the data, the concept of entropy can be extended to measure the continuous entropy $h(X_c)$:

$$h(X_c) = -\int_{\Omega} f(x) \log_2 f(x) dx \tag{3.2}$$

Here, $f(x)$ is the probability density function of the continuous sample X_c and Ω is the defined domain of X_c (Gong et al., 2014). In order to measure the shared, or Mutual Information, between two data sets the concept of entropy is again extended to the conditional entropy (Thomas and Cover, 2006; Sharma, 2000). In contrast to the classical concept of entropy the explanatory power of Mutual Information (MI) has a direction. The MI gives insight on the information flow between variables. Between two variables x and y MI is defined as:

$$MI = \int\int f_{x,y}(x,y) \log_2 \left[\frac{f_{x,y}(x,y)}{f_x(x) f_y(y)} \right] dxdy \tag{3.3}$$

In this definition, $f_x(x)$ and $f_y(y)$ are marginal probability density functions of x and y. The joint probability density function of x and y is given by $f_{x,y}(x,y)$ for a sample with the length N. Following Sharma (2000) Eq. (3.3) can be approximated by:

$$MI = \frac{1}{N} \sum_{i=1}^{N} \log_2 \left[\frac{f_{x,y}(x_i, y_i)}{f_x(x_i) f_y(y_i)} \right] \tag{3.4}$$

In this approximation the probability density functions represent the same sample of data (Sharma, 2000; Fernando et al., 2009). Due to empirical character of hydrological problems, a kernel estimator is used to retrieve the respective densities without the need to fit a known probability density function (Fernando et al., 2009).

To analyze the information content of the runoff and the tracers and to finally judge the prediction capabilities of a ML-based approach for tracer concentration prediction, the MI of the chosen tracers and the continuous entropy of the input window of runoff are calculated and compared.

3.3 Data base

As mentioned before, the data base comprised runoff data from seven French springs. The measurements were taken on a daily basis and cover a length from 11 to more than 50 years (Tab. 2). The measurement interval of the tracer concentrations varied from catchment to catchment and were not equidistant. In catchment Baget 24 pairs of tracer measurements (SO_4^{2+} and NO_3^-) are recorded, whereas in Source du Lez 300 pairs of tracer measurements are available. The mean Pearson correlation between both tracers SO_4^{2-} and NO_3^- showed a strong correlation with r = 0.67 over all catchments. Because of the varying number of available tracer measurements, the training data was resampled 10 times per catchment to lower the influence of single measurements (Eaufrance, 2018a, 2018b).

Tab. 2: Overview of used data for tracer prediction in karstic springs based on runoff data

Source	Length of daily runoff measurements	Tracer measurements SO_4^{2-} and NO_3^-
Baget	1968 – 2015	24
Durzon	1996 – 2016	154
Fontaine de Vaucluse	1966 – 2016	51
Fontbelle	2004 - 2015	194
Source de Fontestorbes	1965 - 2015	43
Source de la Touvre	1980 - 2016	125
Source du Lez	1987 - 2016	300

The runoff was normalized by the catchment specific mean. Moreover, both tracer concentrations were also normalized by the individual catchment specific mean. Detailed information on the measurement setup was not provided by Eaufrance. As mentioned before, the complete time series of runoff was cut into sequences with the pair of tracer measurement in the center.

3.4 Performance metrics for tracer concentration prediction

Like in the previous ML application a set of performance metrics was applied to identify well performing combinations of data sets and ML structures. In order to show the general prediction performance of both tracers SO_4^{2-} and NO_3-, the Root Mean Square Error (RMSE) for observed and estimated tracer measurements was applied, which becomes 0 for a perfect prediction. To calculate RMSE for the tracer content, c_T as the tracer concentration was divided into true and estimated, and a *N* is number of samples:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (c_{T_{true}} - c_{T_{est}})^2}{N}} \tag{3.5}$$

For both tracers, the RMSE is determined individually and presented as a mean of both. As the RMSE has no information of the direction on error, the average concentration ratio $\overline{c_T}$ tells about the general strength and direction of the error.

$$\overline{c_T} = \frac{1}{N} \sum_{i=1}^{N} \frac{c_{Test}}{c_{Tobs}} \tag{3.6}$$

By the sign of the mean concentration ratio the direction of over- and underestimation can be revealed. As the dynamics of the tracer pair is of interest, the accuracy of the relative ranking of both tracers is investigated. Therefore, the overall accuracy is calculated:

$$Acc = \frac{pos_{True}}{(pos + neg)} + \frac{neg_{True}}{(pos + neg)} \tag{3.7}$$

Here, the relative positioning of both tracers is calculated and its accuracy shows whether the interplay between the two tracers is captured or not. Here, the index tells about whether the estimated ranking is true or not. Positive (*pos*) and negative (*pos*) reveal which tracer is relatively higher than the other. A high *Acc* means that the ranking of both tracers is always captured in the right manor. The lowest possible *Acc* = 0 where no ranking reflects the correct situation between the two tracers.

3.5 Entropy and mutual information of the investigated tracer data sets

The continuous entropy of the complete time series of runoff from all seven springs was rather low, resulting in a low information content of the runoff data with only 1-2.5 bit (Fig. 18). The more data was used, the already low information content was even lowered to less than 1bit. The MI between the tracers varied. In some catchments (e.g. Baget, Durzon and Source du Lez) it diminished at the same pace, the more data was used until some kind of turning point where suddenly the information content increased to more than 25 bit. The maximum of more than 200bit information seemed to be an outlier of the data and the according situation of measurements should be excluded from further analysis. Nevertheless, the maximum MI among the two tracers reached a plateau at 25 bit -30 bit which was 15-20 times higher than the continuous entropy of the runoff. Missing data in the curves of entropy and MI resulted from data samples where the density estimation of the mutual information did not converge, thus leading to *NaN*.

In most cases, more than 70% of the available tracer measurements were needed to exceed the information content of the runoff (e.g. Durzon, Fontaine de Vaucluse, Fontbelle, Source de la Touvre and Source du Lez). In Baget the decoupling of both information contents

started earlier at about 25% of the available data. In total numbers: In Baget only 5 relatively unique tracer measurements were sufficient to capture the information variability, whereas in Fontbelle more than 135 pairs of tracer measurements were required to exceed the relatively low information content of the runoff.

A detailed analysis of these five tracer measurements was not feasible due to the historic character of the data set. Additionally, detailed information on the measurement setup, the environmental situation or any possible human influence on the results was not stored in the database and thus not accessible for researchers without detailed in depth knowledge.

Although the mean continuous entropy and MI were used, one can see that the charts are ragged (e.g in Source de Fontestorbes). This means that single events still had a large influence on the information content. These events had to be included in training data to get the maximum information from the data. So, a training data set was chosen that contains the maximum amount of information by the lowest number of events in the sample which is known as the maximum entropy approach (Berger et al., 1996; Brodley, 2004; Phillips et al., 2004). So, the absolute number of measurements is not the main driver of performance but the information content of the individual measurement. The lower the number of available tracer concentration measurements, the more ragged the MI chart and the earlier a plateau of information content was reached (e.g. catchment Baget, Source de Fontestorbes). Catchments with a higher number of available tracer measurements reached the plateau later with bigger training data sets.

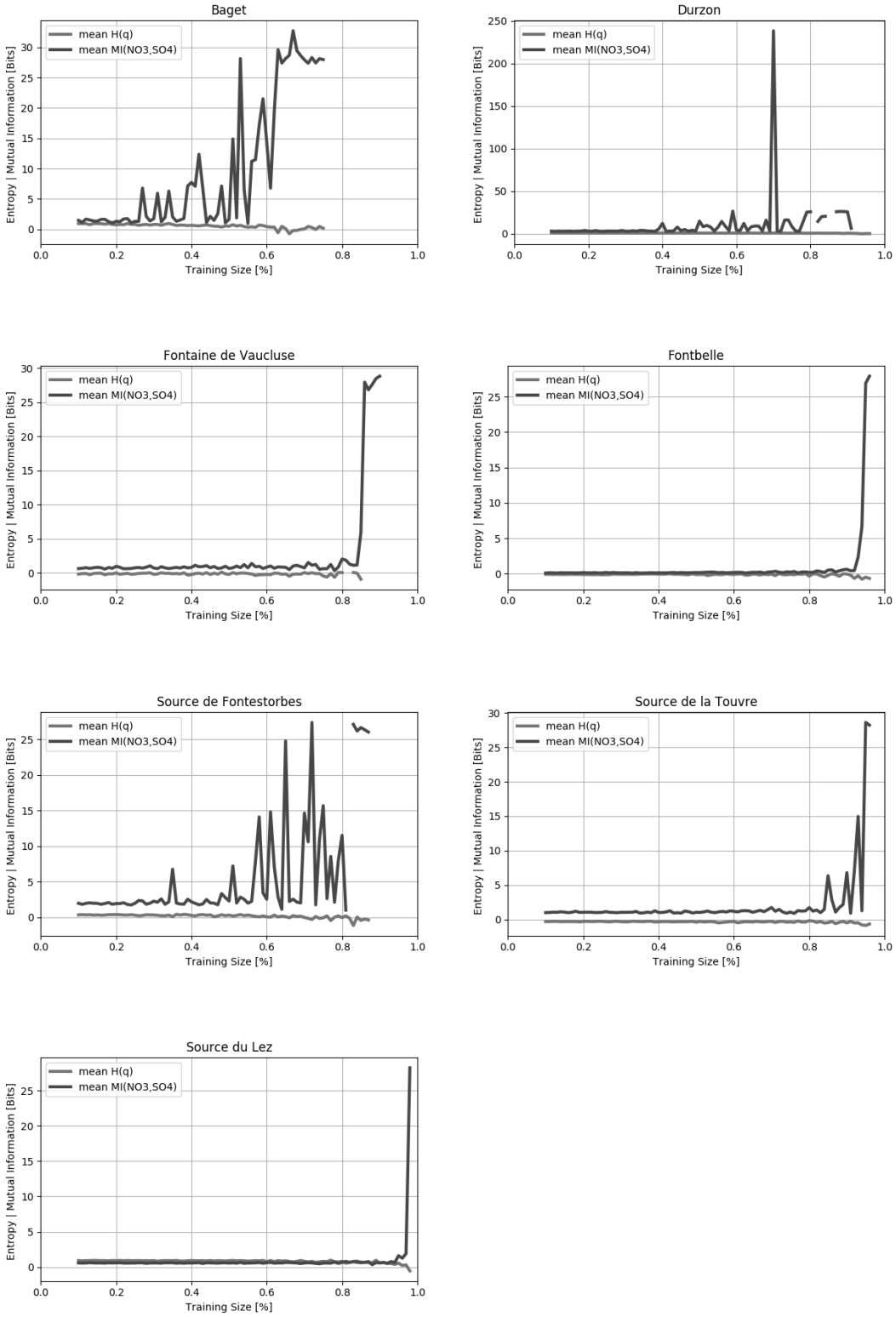


Fig. 18: Mean continuous entropy and mutual information between NO_3^- and SO_4^{2-}. The overall maximum of the mutual information is at about 25 – 30 bit, while the continuous entropy does not exceed 2.5 bit. The lower the number of available tracer measurements (compare Tab. 2), the more ragged is the mutual information graph and the earlier a plateau is reached.

In those cases, that MI was higher than the continuous entropy of the runoff (e.g. in catchment Baget with more than 20% of the training data, or in catchment Source de la Touvre with more than 85% of the available data as training data), one could assume that the tracers share more information about the system dynamics than it could be found in the runoff data. Hence, the potential of tracer concentration prediction by runoff alone is limited to system behavior visible in the runoff data. Moreover, some dynamics should not become visible in the ML-based prediction as the information content of the runoff data was too low to explain all processes in detail.

3.6 Validation of tracer concentration prediction

As mentioned before, two different training strategies were applied: The univariate and the multivariate approach. In the univariate approach for each tracer a machine was trained. Contrary, in the multivariate approach one algorithm was fitted to predict both tracers as a multi output with the same algorithm. To use knowledge on the information content of the catchments, only as much training data was used where the continuous entropy of the runoff and the MI of tracers have equal levels. Both training strategies were compared due to their capability in prediction and, even more important, to show a direction of information flow between the tracers. As one can see in the box plot covering different time spans of runoff input data, no clear preference towards an algorithm could be stated. It is rather region dependent as to which algorithm performed best as one can read from the mean concentration ratio $\overline{c_T}$ (Fig. 19). In all four approaches, the box plot revealed that SO_4^{2-} had a higher variance than NO_3^- with the whiskers showing the deviation from the optimum of 1.0 in $\overline{c_T}$.

In contrast to the stream flow separation all investigated approaches delivered stable results. SO_4^{2-} seemed to cause more problems in the prediction than NO_3^-. Over all catchments and approaches NO_3- was closer to the optimum of 1.0. Generally, one can see that the tracer concentration in some catchments can be predicted in a better way than in others: well performing catchments were Fontaine de Vaucluse, Fontbelle, Source de Fontestorbes and Source du Lez. Of these catchments, Fontbelle could not be predicted by CART. In the other catchments no preference towards any algorithm could be observed.

The prediction in catchment Baget tended to underestimate both tracer concentrations. Generally, the results in this catchment revealed a broad variance. Thus, the window length of runoff had a higher influence on the overall ML performance. Here, CART showed the best performance, especially considering the multivariate strategy with $\overline{c_T}$ values close to the optimum of 1.0. Apart from the SVM, Source du Lez showed the opposite behavior: The variance of results was low, close to the optimum.

Quite contrasting, the SO_4^{2-} concentration in catchment Durzon was overestimated by factor 3 while the NO_3^- concentration was underestimated by factor 2. This contrary behavior was also observable in catchment Source de la Touvre. The choice of algorithm only influenced

the magnitude of this contrary behavior.

While the multivariate strategy ameliorated the prediction results of NO_3^- it deteriorated the prediction capabilities of SO_4^{2-}. The SVM did not show a preference towards any of the different learning strategies. Contrary, a tendency got obvious in CART, ELM and ANN. This means that prediction of NO_3^- could be improved by incorporating the other tracer and the prediction of SO_4^{2-} cannot retrieve any useful information from that additional data.

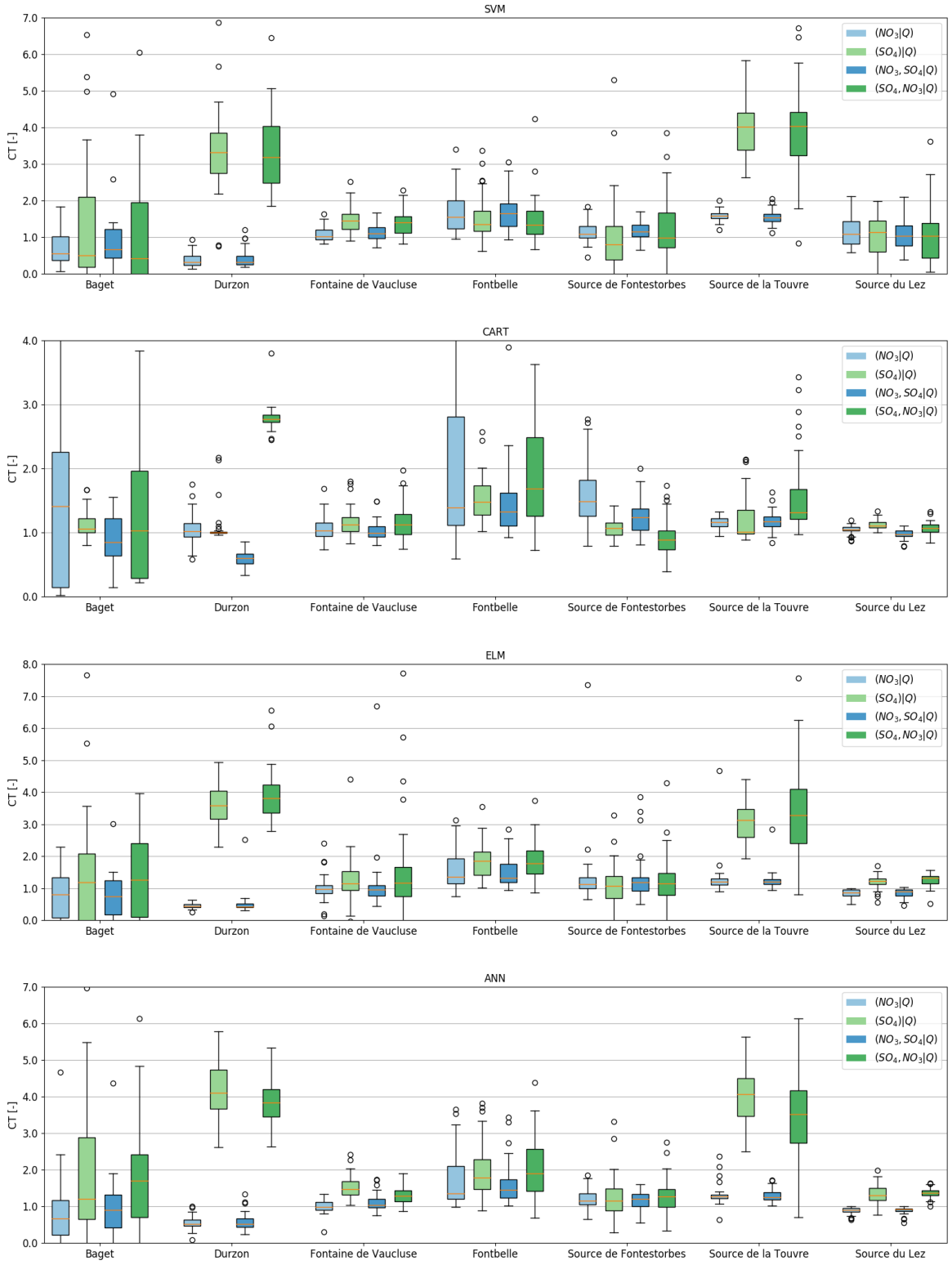


Fig. 19: $\overline{c_T}$ of SVM, CART, ELM and ANN. The variability shows the performance according to the applied type of training data. While the most catchments show good results regardless of the applied machine with only slight variations in the influence of amount of training data. In some catchments both tracers cannot be predicted, like Baget, in other the prediction of a single tracer was heavily biased, like SO_4^{2-} in Source de la Touvre. In some cases the multivariate approach (NO, SO | Q) performs better than the univariate algorithm (NO | Q) and (SO| Q).

Next to the mean concentration ratio $\overline{c_T}$, the RMSE was also an important performance metric for judging ML capabilities in tracer prediction (Fig. 20). The findings from the RMSE emphasized the results derived from the analysis of $\overline{c_T}$. Again, catchment Baget showed the highest variability in the validation. Fontaine de Vaucluse, Fontbelle and Source de la Touvre scored low RMSE values over all four approaches. This performance metric underlined the findings from before. It is rather a catchment specific tendency and not a preference towards the specific algorithm that finally makes the choice towards a ML setup. Catchments that performed well with SVM for example, also performed well using ANN, ELM and CART.

There were some exceptions to that rule: Source du Lez seemed to cause problems using a SVM. This underperformance was also visible in the CART application but to a lower degree and more interestingly limited to NO_3^-. Fontaine de Vaucluse revealed a higher mean of RMSE and a higher variance in error using an ELM. Using CART, the multivariate strategy deteriorated the prediction performance, while using the other ML approaches the performance delivered similar results. The catchments Durzon and Source du Lez showed opposite error behavior as before. The prediction of SO_4^{2-} resulted in a lower error, while the prediction of NO_3^- was worse than for the other tracer. Here, the bias of the performance measure became obvious and highlighted the need for a comparative analysis of performance measures. Each performance measure added a bias to the result because any measure focusses on different details. In future research, the different preferences of the ML approach might be used to regionalize karstic systems as the tendencies towards algorithms, the learning behavior and the relation between MI of tracers and continuous entropy of runoff might share distinct patterns of information that are yet hidden.

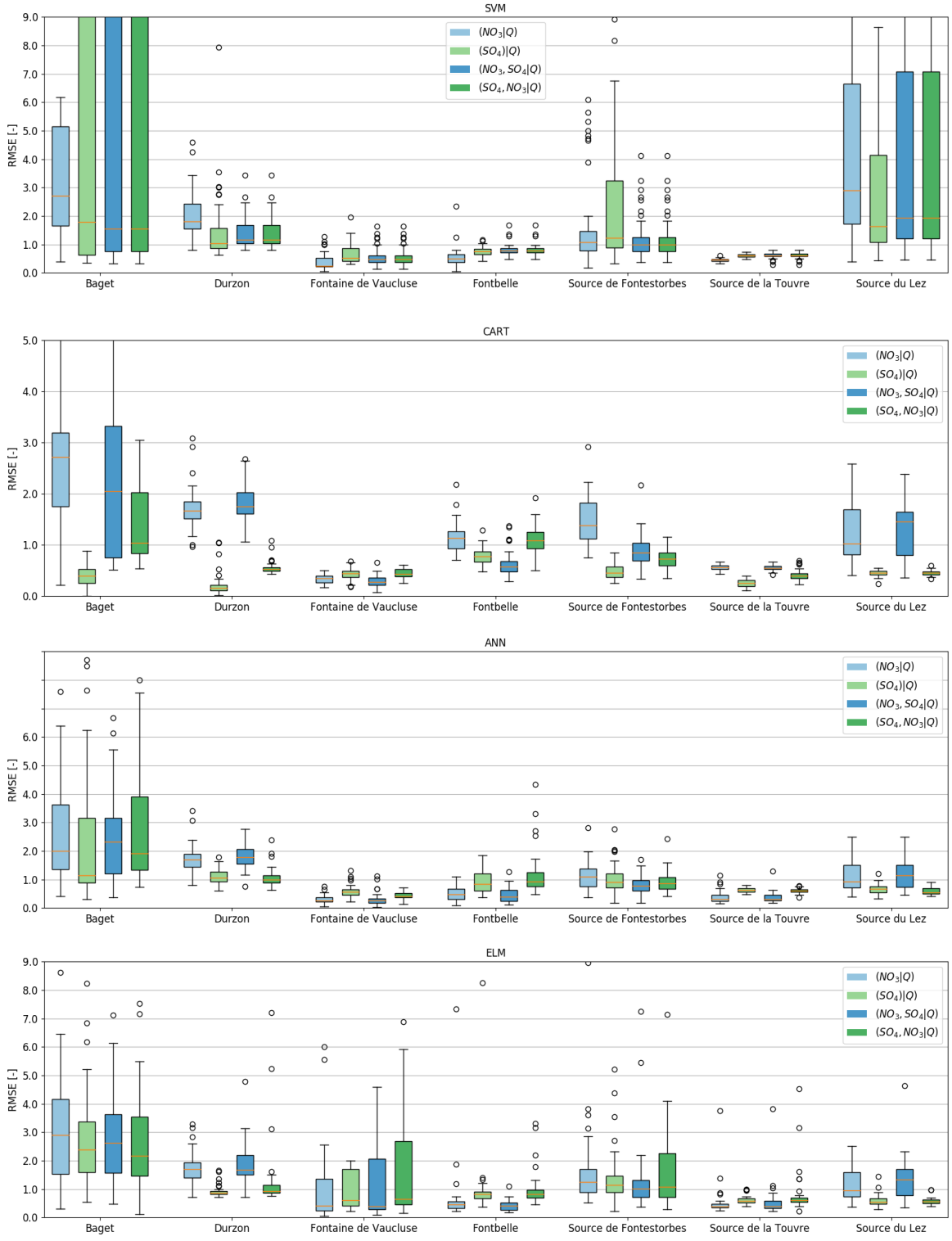


Fig. 20: RMSE of SVM, CART, ELM and ANN for univariate and multivariate algorithms. The variability shows influence of learning threshold on the development of RMSE in the catchments. The RMSE follows the results from $\overline{c_T}$ showing that the error relates to the average tracer concentration, revealing that in some catchments tracers cannot be predicted like catchment Baget. The choice of the machine has only low influence on the error and depends on the region.

The last performance measure to check in the process of validation was the accuracy (*Acc*) of the pair of tracers (Fig. 21). Like in the results before, no general preference towards any of the aforementioned ML algorithms became visible. It was merely a question of the catchment whether the *Acc* was high or low. Over all catchments and approaches values of *Acc* > 0.5 have been achieved which means that in more than 50% of the cases the relative ranking of predicted tracers was modelled correctly. Interestingly, the SVM in catchment Baget, previously with the highest variety of RMSE and $\overline{c_T}$, scored the maximum of *Acc* > 0.8, with outliers of 1.0. But again, the variability was higher than in any other catchments.

With the *Acc* analysis, the difference between the two strategies, uni- and multivariate, became even more obvious. The most interesting catchments were Durzon and Source de la Touvre. Using the SVM in both catchments the multivariate strategy improved the overall *Acc*. In contrast, using CART the multivariate strategy massively deteriorated the performance. Over all approaches it holds true that those catchments that show low variations in the other performance metrics also show low variations in prediction capability but do not reach similar heights as the performance catchments with a high variation. Overall, this analysis showed the need for a complementary prediction framework that allows to predict combinations of tracers with their specific strategy.

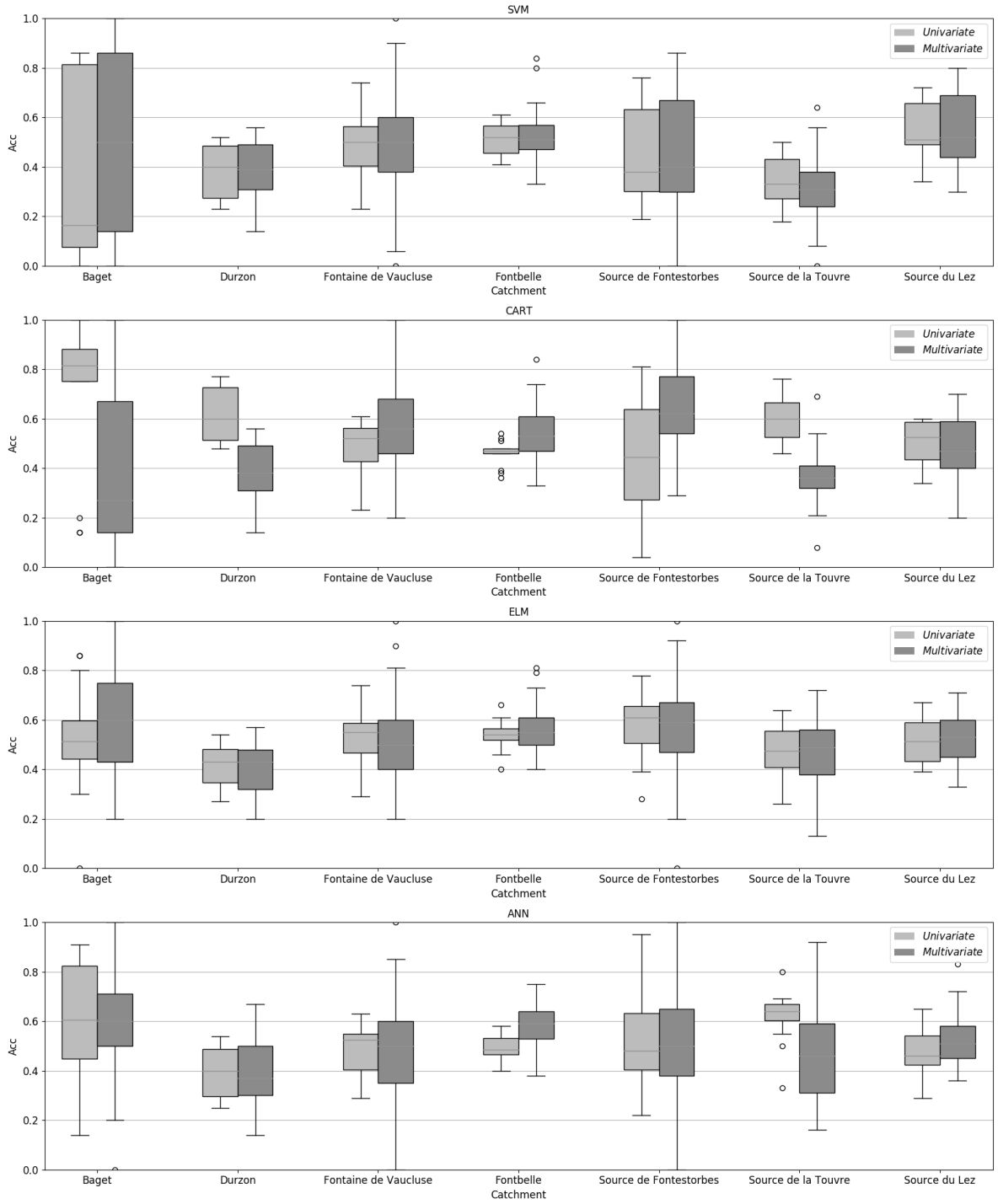


Fig. 21: *Acc* of tracer prediction defined as the correctly predicted relative ranking between both tracers.

3.6.1 Influence of window length on prediction capacity

Keeping the aforementioned results from the individual performance measures in mind, the emphasized influence of window length on the prediction capacity, and the impact of the window length on the performance were investigated and discussed. As catchment Baget turned out to be an interesting catchment in terms of performance variance, the four ML application in this catchment were investigated on their window length dependency (Fig. 22). One can see that very short window length of only ±1d around the tracer measurement lead to good results, especially for NO_3^-. For SO_4^{2-} long window ranges (> ±60) led to similar good results. This was coherent with the meaning assigned to both tracers, one for short-term changes and fast components while the other one more or less focusses on long-term changes and slow phreatic evaporation processes. These temporal patterns were overlapping in the input data, hence could only be detected by the ML algorithm if the window size was adapted. Too long, or too short windows confused the ML algorithms in the search for a pattern. Erroneous results were caused by non-convergence or underfitting as one can see for SO_4^{2-} predicted by ELM and ANN. Here, the performance got better the longer the input windows were.

In contrast to the results from Baget, the window length showed different behavior for predicting the tracers in catchment Source de la Touvre (Fig. 23). Generally, SO_4^{2-} was overestimated, apart from using an ELM. Contrasting to the assumptions that SO_4^{2-} requires longer windows as it represents long-term variations of the system, the overestimation got worse the longer the window was.

Consequently, for the application of ML for prediction of tracer concentrations, a combined strategy had to be developed. The strategy must contain a data-specific definition of training data. Even if the structure or the problem is similar and even if the input data is the same, the strategy on how to use the data in the ML approach has to be overthought and adapted to the specific problem. A future application of the window length investigation might be the hypothesis test whether the tracer can be linked of assumed natural processes with a defined time span. Here, the mutual information between tracer and process information can be analyzed and links to the conceptual model can be established. Moreover, a data-driven regionalization of karstic catchments can be conducted.

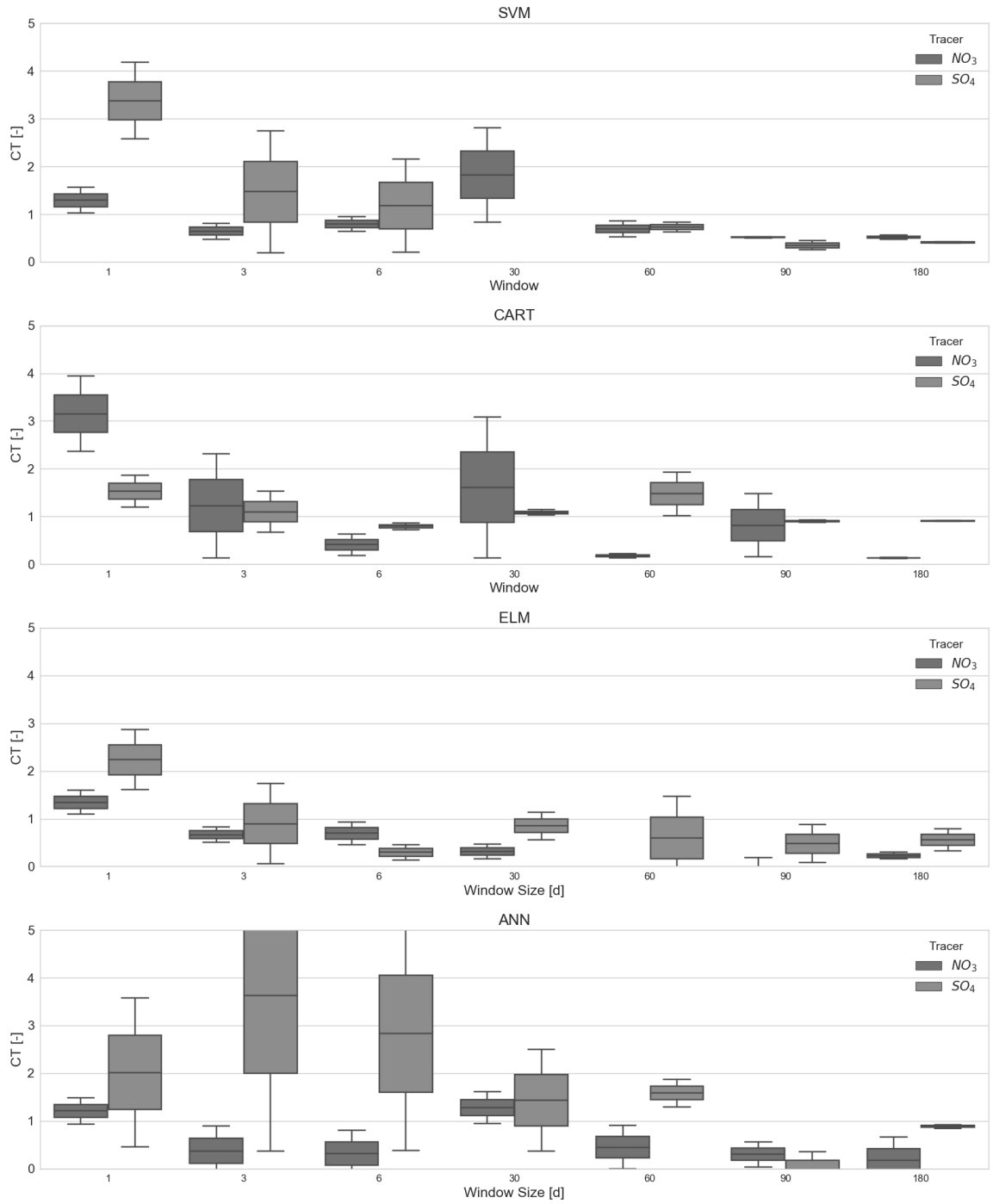


Fig. 22: Dependency of chosen window length on $\overline{c_T}$ at catchment Baget. The different preferred window length for good performance in tracer concentration prediction underlines the different meanings assigned to the tracers. SO_4^{2-} can be predicted in better way the longer the input data window is, while NO_3^- reaches the best performance values using small windows

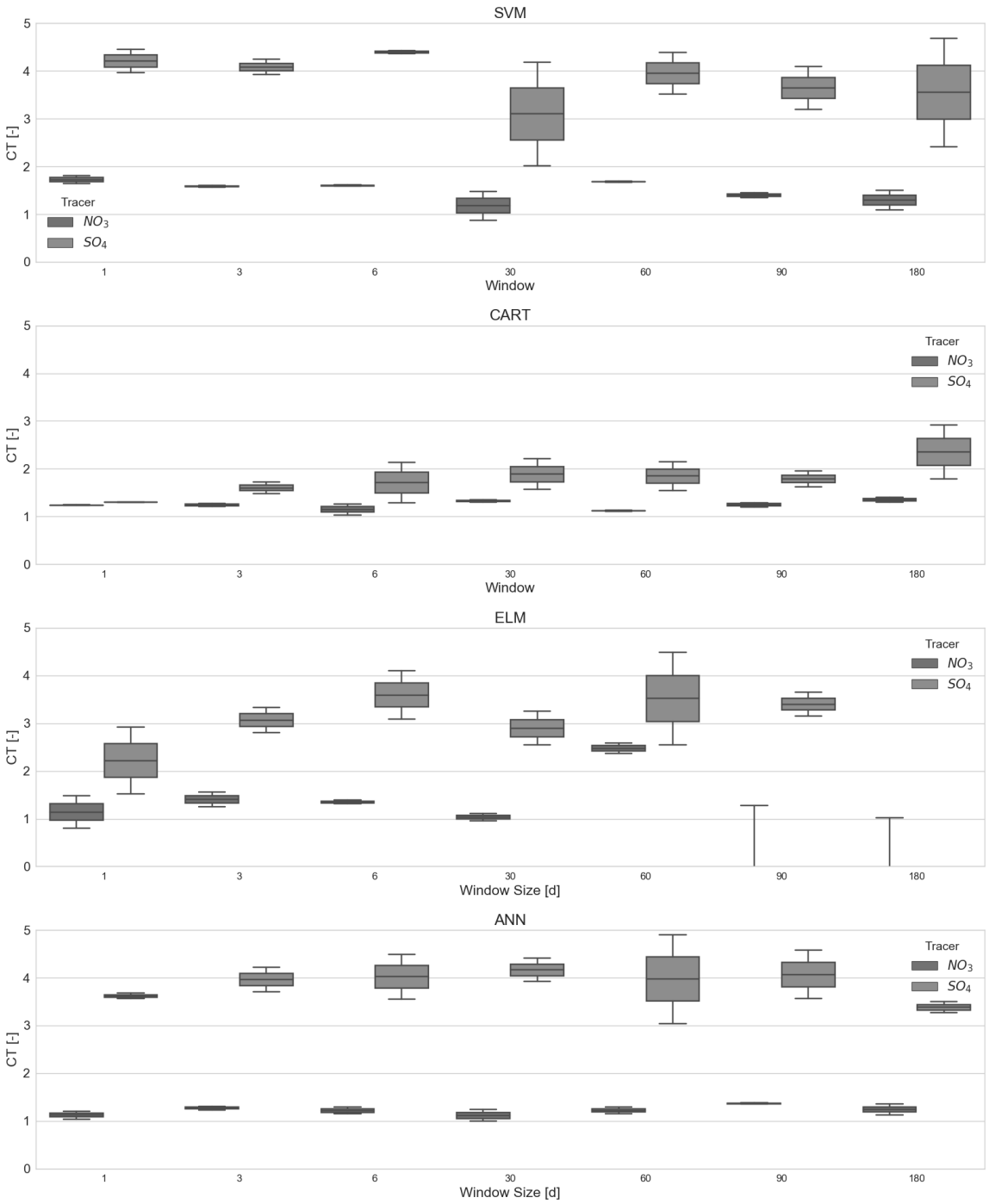


Fig. 23: Influence of window length on performance of $\overline{c_T}$ results in catchment Source de la Touvre. SO_4^{2-} is generally overestimated and quite surprisingly worsens the longer the window of input data is.

3.6.2 Meaning of entropy for tracer prediction

Comparing the entropy, the MI and the results of the ML predictions of tracer concentrations, one can remark a certain pattern. Catchments, like Baget, that had a fast rise of MI in contrast to a stable, low continuous entropy showed a greater variability of results and a heavier dependency on the appropriate window choice (Fig. 18, Fig. 19, Fig. 20, Fig. 22). The higher

the information content, the higher the possibility to achieve a good *Acc* as well (Fig. 21). As there was no clear tendency towards any algorithm, the *No-Free-Lunch-Theorem* has been proven in practice: The choice of the algorithm is a matter of both, the problem and the data. Apparently, the prediction of tracer signatures in karstic environments depends on the structure and the information content of the data. High entropy data sets with only few, but unique structures, like from Baget may deliver more variable results, whereas data sets from catchments with a lower information content like Source de la Touvre show less variable results, but are furthermore limited to lower possibilities for prediction.

This became also visible in the influence of the window length. High-entropy data sets showed a dependency between the results on the window length as patterns in data may overlap for short-term processes or may be confused in terms of slow long-term processes (Fig. 22). Even though the term of high-entropy might be misleading in the context of similar final levels of information in all catchments (Fig. 18), the development of entropy is different, which became visible in the results as mentioned before. Consequently, one can state that the 24 tracer measurements in Baget capture more information than the 300 in Source du Lez. Apparently, similar hydrological situations were measured, that do not create a strong pattern in both tracers. From this definition, catchments Baget and Source de Fontestorbes (43 tracer measurements) can be marked as information-rich data sets (Tab. 2). To a certain level also the data from Durzon can be counted as information-rich, although the information content > 200 bit seems to be erroneous as it just peaked and did not reach a stable level.

3.6.3 Interpolation quality of ML approaches

One of the major improvements of a trained ML algorithm is the ability to interpolate complete time series of tracer concentrations. Even if the amount of tracer measurements is limited, a ML approach should be able to guess the tracer concentration based on the collected experience. Although it is not possible to create new extreme constellations, the known range of tracer concentrations should be reproducible as long as the pattern of their origin is detectable in the training data.

For catchment Fontbelle, an interpolation was conducted to create a continuous time series of tracer measurements by an ANN trained with 20% of available events (drawn randomly). One can see that the interpolated SO_4^{2-} concentration is limited using the univariate strategy (Fig. 24). Using the multivariate strategy (multiple tracers predicted with one machine) the prediction reached a variability similar to the measured values. In total, it seems as if the univariate strategy predicted a damped time series, whereas the multivariate strategy captured the evolution of the actual measurement better.

The prediction of NO_3^- on the other hand showed only marginal differences to each other. The multivariate strategy was more variable than the univariate strategy. This was comparable to the performance measures where both strategies deliver similar results (Fig. 19, Fig. 20).

As an explicit choice of the ML approach was not possible due to the qualitative comparison, expert knowledge and qualitative comparison of interpolated time series could be used to judge the goodness of fit for each interpolated time series. Although the qualitative comparison seemed to be less valuable than the quantitative analysis, it helped to determine the performance of the ML approaches and the appropriateness of fit. One can conclude that the information content of the measured runoff reflects the interpolation quality of the ML approaches.

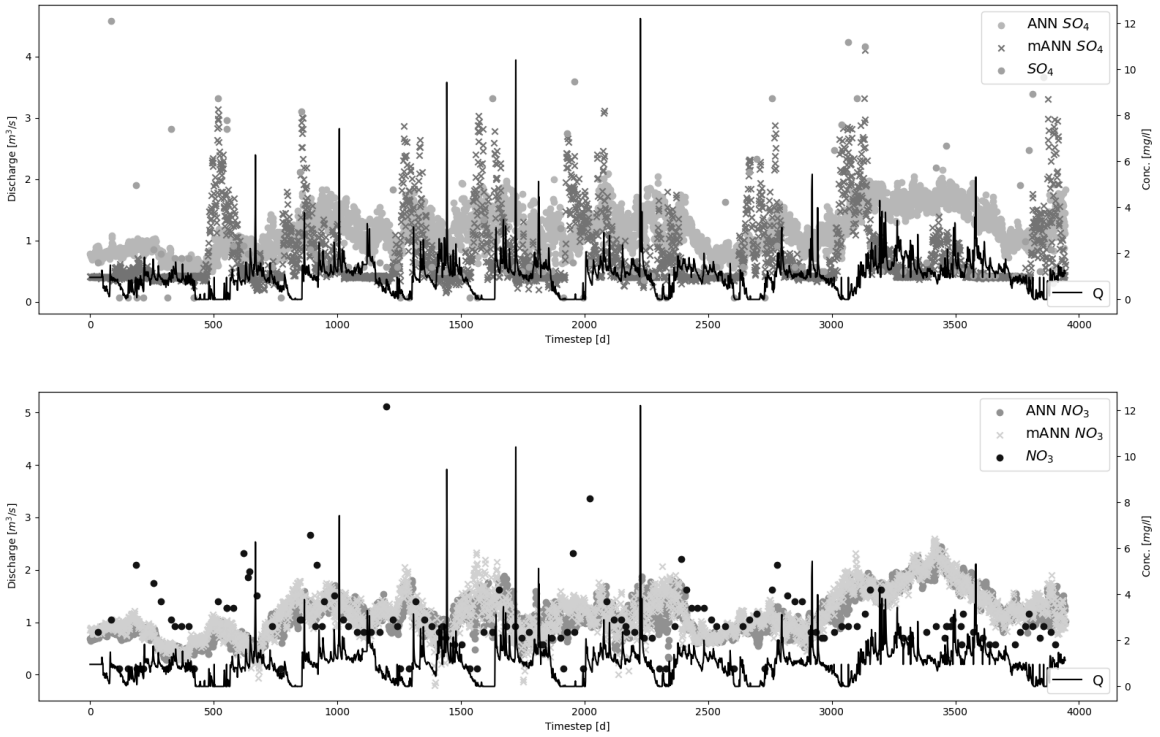


Fig. 24: Interpolated time series of NO_{3-} and SO_4^{2-}. As predicting ML algorithm an ANN trained with 20% of the available tracer measurements was taken. The multivariate learning (mANN) strategy allows an interpolation of SO_4^{2-} closer to the range of measured data

3.7 Concluding remarks on entropy-based data mining

In this case study, the application of entropy and information-theory based applications of data mining in the context of ML-based interpolation methods was discussed. With the continuous entropy and the MI the information structure of tracer measurements and complete time series of runoff could be investigated. Moreover, hidden links like the information flow between NO_3^- and SO_4^{2-} were revealed.

The gap between the information in the runoff data from the Eaufrance database and the available pairs of tracer measurements explained the limited interpolation ability. The point where the continuous information from runoff and the MI of tracer diverged, the runoff was not able to describe the variability of the tracer measurements and thereby limited the ability

to describe the complete system. Although the information content is a fuzzy measure that is hard to link to physical measures, the knowledge about the information content in hydrological data is crucial for modelling purposes. For the interpretability of ML-derived results, the investigation of the information content as a part of data mining and data processing is a further step forward as it offers a quantitative measure for the limits of the prediction capabilities that the trained algorithm.

The window size of the input data confirmed the thesis on the represented processes. So, this application could be of use to confirm process hypothesis and further strengthen the interpretation of tracer-based methods. Here, the need for complementary approaches in ML setups is further underlined by the information flow and the demand for different learning strategies.

3.8 Conclusion of ML-enhanced approaches in hydrology

As a conclusion from the data-driven side of the modelling spectrum (Fig. 3), one can say that ML approaches help to investigate catchments and their internal structure purely based on the available data. Black-box approaches help to develop a deeper understanding of systems behavior on integral measurements, like runoff.

As shown in Sec. 2.3 ML approaches could be used to facilitate pattern recognition in data, and replicate expert knowledge without exact mathematical definition. Hence, expert knowledge becomes transferable at low costs. Following the No-Free-Lunch-theorem (Sec. 2.2) it was explained, why different approaches have to be considered: As the *a-priori* choice is often impossible, several different approaches have to be tested. All algorithms should eventually solve the problem, but the amount of work, in this case data, is different. For the flood event separation a preference towards SVM and ELM was shown. Less than 10 runoff events are needed to score sufficient performance measures even when the algorithm is applied to a large data base (Sec. 2.4). This feature even holds true for increasing data sets which means that some key requirements of Big Data are achieved: volume and velocity!

The case studies revealed that some structures might not be appropriate for the specified problem or require further adaption (see ANN in Sec. 2.4.1 & 2.4.2). So, it is not automatically said that ANN are not able to separate flood events, but, and this is the major point, other algorithms can perform at least equally well with less work needed.

Due to the low number of required training data for stable results, one can assume that it is not the total number of events but the quality of information in the data is the main trigger for a successful ML application. A measure for information quality, the Shannon entropy, was presented in Sec. 3.2.

To investigate the possible use of the entropy model in ML approaches in hydrological applications, the prediction of tracer signatures in karstic environments by runoff data was tested (Sec. 3.1). For this type of measurement a ML-based interpolation method would

eliminate certain limitations of tracer-based methods. Due to high costs and low reproducibility of tracer measurements, continuous time series of tracer measurements are rare. So, the need for an elaborated interpolation method to fill the gaps based on known system dynamics (like runoff) would be of interested. The analysis reveals that runoff has a lower information content than the here presented pair of tracers share among each other. Furthermore, the total number of tracer measurements has a negative influence on the prediction capabilities. Especially in catchments with high numbers of available measurements more training data is required to include information-rich tracer combinations. As soon as the catchment is either dynamic or the tracer measurements capture these dynamics by chance, the prediction applicability is limited. Nevertheless, more than 50% of the relative ranking can be predicted which is, for applications of tracer measurements in some model calibration strategies, an advantage.

Thanks to the test of uni- and multivariate learning ML strategies some light was shed on the direction of information flow and the interdependence of information in data. SO_4^{2-} can be predicted more accurately by multivariate ML approaches than NO_3^-. The shared information helps to predict one tracer whereas it deteriorates the prediction performance of the other. Thus, one can conclude that for this tracer combination, a complementary modelling framework would be an appropriate toolset. A univariate ML approach would predict NO_3- while the multivariate approach predict SO_4^{2-}. A further ramification of the setup is not analyzed but different lengths of the input windows might also improve the performance of the prediction. This analysis of the influence of window length might be used for future regionalizations of karstic catchments where geomorphological classification schemes reach their limits due to complex subterraneous structures.

From these findings, a ML framework for the application in hydrology can be formulated comprising the choice of the approach and the selection of data. The framework has six columns that all have to be answered before considering a ML approach in hydrology.

1. Define the question to the data. Provide a defined truth to compare to.
2. Analyze the information content of the data. Do in- and output data have the same information content?
3. Choose as much data as needed but at least enough for possible training in order to avoid confusion of the algorithm evoked by the decreased information/noise ratio.
4. Choose a set of ML approaches and test those against performance metrics that suit the problem.
5. Perform a qualitative and quantitative analysis of the results.

ML based approaches are interesting either for large data sets that would require heavy workload to analyze manually or for problems with a complex and yet unknown structure. This qualifies for some of the requirements posed to big data: Velocity and Variability. Any newly incoming set of data can be analyzed and processed according to the training data.

Nevertheless, this study shows the need for expert knowledge in terms of approach selection

and the pre-processing of the input data. Especially for the qualitative analysis and interpretation of results expert knowledge and the ability to judge results for a specific question is of interest.

4 Emerging systems modelling by Agent-based models

The aforementioned black-box approaches like ML delivered promising results, but apart from CART none of the approaches was interpretable by the researcher. To overcome the explaining limits of black-box, approaches like the highly specialized genetic programming (GP) can be applied (e.g. Wang, 1991; Parasuraman et al., 2007). Alternatively, the modelling scheme shifts to white-box modelling approaches. White-box approaches share the characteristic that conceptual knowledge of the researcher is taken into account. The knowledge-driven characteristics are achieved by the definition of rule sets for the model component. They are often applied if the general rules in a system are known but the exact interactions are either hard to describe or unknown (Bruch and Atwell, 2015). In the past decade social sciences, psychology and behavioral ecology developed numerous approaches to model systems that consist of patterns where the rules that lead to the pattern are not completely known. By adding more and more rules, the interaction of model components and the rules can be examined. Often the absolute performance is not of interest, but the pattern created by the model is in the focus of research. To model these systems from this bottom-up perspective, expert knowledge has to be utilized.

In order to incorporate expert-knowledge into the process of data analysis, the focus has to be shifted from black-box applications to knowledge-driven approaches like agent-based computing. From a theoretical perspective, agent-based computing is often classified as soft computing (Bruch and Atwell, 2015). In those applications the quantitative information steps back behind the qualitative information which has its origin in the fuzzy nature of soft computing approaches. Soft computing approaches utilize knowledge of a system's behavior without the exact mathematical definition. They rely heavily on logical definitions and the connections with if…then… statements (Sivanandam, 2011). These logical conditions with thresholds are also the fundament of agent-based modelling which is a recently developed common rule-driven modelling approach for complex interaction models.

4.1 Fundamentals of Agent-based models

To model emerging system behavior of complex interacting systems, often agent-based models are applied (ABM). Agents are encapsulated, autonomous software units (Gunkel, 2005; Macal and North, 2010). Each agent follows a strategy to achieve a goal (Jennings, 2000). Therefore, certain rules are assigned how to behave in their specified environment. These rules represent the boundary conditions of the models. As the agents have a goal and a strategy, the actions allowing to achieve a goal have to be defined beforehand (Macal and

North, 2010; Lempert, 2002; North, 2014). The agent also has certain attributes that constitute its internal states and may trigger certain behavioral patterns.

Through sensors the agent is able to sense the environment and to communicate with other agents (Fig. 25). The agents do not only have sensors and actors for their environment but also for their neighborhood (Blaschke et al., 2013; Hofmann, 2017; Hofmann et al., 2016). The definition of the environment remains complex and has to be discussed in detail with the specific problem in mind. ABM consist of a magnitude of autonomous agents that build the entire model. By this, the interactions among the agents create a dynamic within the model that cannot be modelled by traditional methods (van Parunak et al., 1998) and is therefore of great interest to model problems that have a highly dynamic internal composition of states (Mewes and Schumann, 2018b).

The multitude of autonomous agents also creates new problems that are out of focus of conceptual models. Here, the scheduling is harder to determine as there is no obvious order of agent action. A detailed problem-specific scheduling analysis is discussed in Sec 4.3.3.

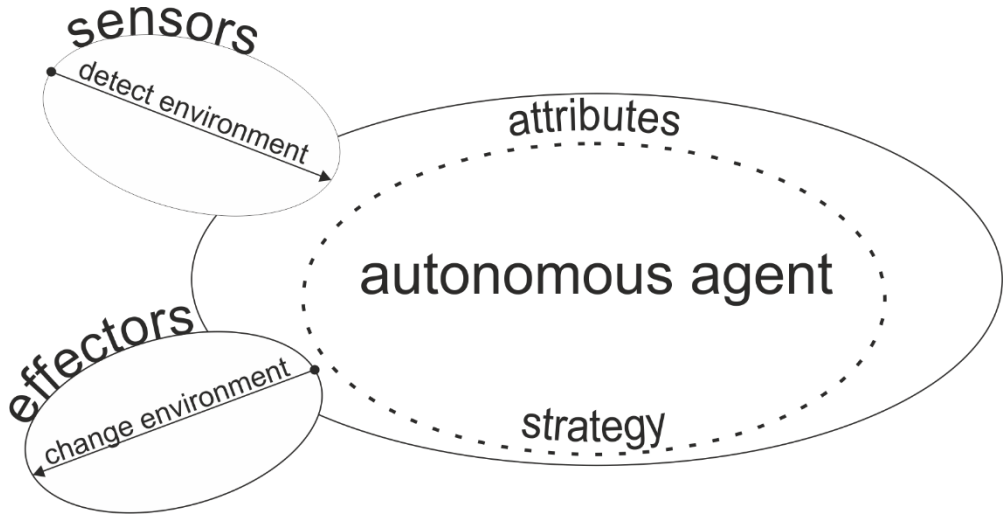


Fig. 25: Scheme of an autonomous software agent with sensors and actors, here referred as effector. After: Hofmann et al. (2016)

An ABM consists of the experiment describing the modelled scenario, in the case study a soil column filled with a certain type of soil, and the unifying global agent that collects all parameters for the complete model as time step length and ending condition (Fig. 26). In ABM time steps are called ticks, because like a clock at any tick the model states have to be reevaluated and recalculated.

The actual agents may be of static nature, for example Meta agents act as representation units, or are dynamic. The dynamic agents alter their shape, their size and their internal state (Mewes and Schumann, 2018a). They represent the core of any emerging ABM and their actions and rules are the fundament of each agent-based model. In contrast to cellular automata, the agents have alternating positions and shapes (Shao et al., 2015; Parsons and Fonstad, 2007; Macal and North, 2010). By inter- and intra-class communication information flows in this model. By this communication structure among autonomous software agents

the emerging character of ABM evolves. These connections may reveal strengths and downsides of theoretical found relationships in a complex natural system.

ABMs are complex computer programs that manage a plethora of objects that decide individually what actions to take. Hence, a software environment has to be chosen that allows the organization of multiple objects (Abar et al., 2017). The choice fell on GIS Agent-base modelling architecture (GAMA), an open-source ABM framework based on Java with the ability to be extended through hooks and add-ins (Taillandier et al., 2012; Taillandier et al., 2014). GAMA already provides tools and routines to cover spatial data, either as vector or raster objects. This high number of individual software objects increases the computational demand of ABM in contrast to equation-based modelling. Therefore, approaches for computation on graphical processing units (GPU) were developed to reduce the computational demand on the main central processing unit (Crooks et al., 2008; Wang et al., 2013a). To get a surplus of performance, the model steps have to be cut into single pieces. Every sub-step again is allocated to the GPU or a multithreaded central processing unit. Thereby, a high number of calculations, like the determination of agent states, is parallelized and thus the computational time of the model decreased. An improvement of performance is only achieved when the sub-steps are of the same nature and require only a low transfer of data via the system's bus unit. Else, loading times from the main working storage to the local graphical storage would take longer, than the GPU calculations save on computational time by calculating faster (Aaby et al., 2010).

In contrast to traditional hydrological models and classification approaches, the validation of ABMs requires different techniques and proxies to make results comparable. As the agents represent different types of actors in a system, the validation strategy has to be adapted to the specific model use. Either pattern comparison can be used as a validation approach which requires a measure of similarity that can be applied to the patterns. Alternatively, a proxy value can be installed to compare the results with those from a traditional modelling approach. If the proxy value is chosen as a reference one has to keep in mind that the choice of the proxy value again is a bias and needs to be discussed.

In terms of big data and the related analysis approaches, ABM covers the variability of data, and is also able to detect and visualize the changing relations among the agents to increase the information gain from the data. Computational time is the limiting factor when analyzing data with a high update velocity but the aforementioned parallelization methods will improve computational time increasing the applicability of ABM in big data contexts.

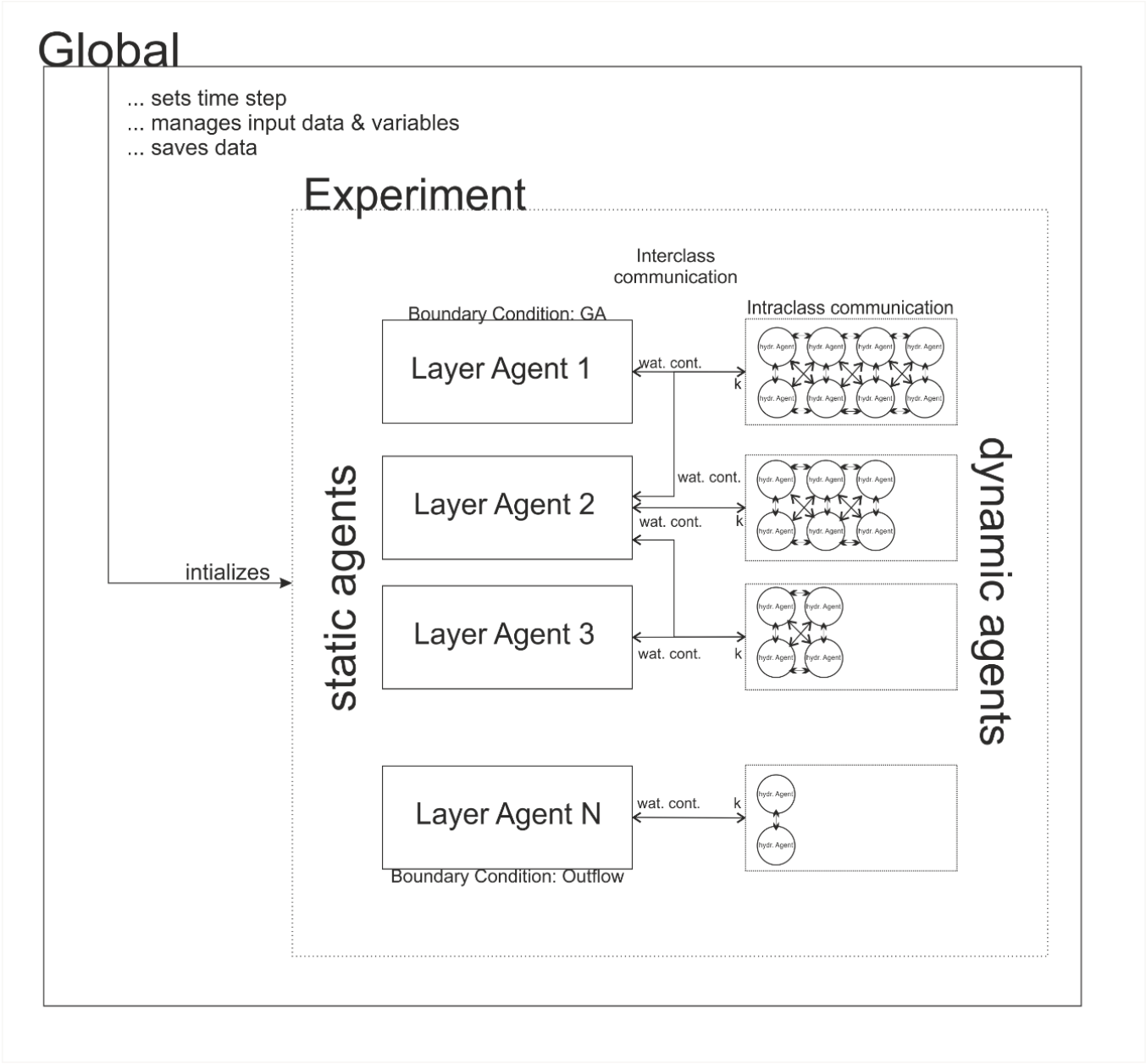


Fig. 26: Conceptual scheme of an ABM comprising the global agent, the experiment and the actual agents.

4.2 Applications of ABM in hydrology and water resource management

Originating from social sciences and behavioral modelling (Cernuzzi et al., 2005; Centarowicz et al., 2010; Kofler et al., 2014; Troy et al., 2015; van der Vaart et al., 2016), first applications in physical-based hydrological modelling and water resource managements appeared in the current decade (Reaney, 2008; Rakotoarisoa et al., 2014; Servat, 2000; Shao et al., 2015; Wang et al., 2013a; Folino et al., 2006; Grashey-Jansen and Timpf, 2010). Three of those applications do not count as ABM, as they merely rely on cellular automata (Folino et al., 2006; Parsons and Fonstad, 2007; Shao et al., 2015) while the other apply agent-based

modelling. Yet, the studies lack of a general strategy in using ABM to answer physical questions in hydrology and water resource management.

ABM applications in water resource management often represent decision makers in policy modelling. The socio-economic part is modelled by agents with a conceptional hydrological model in the background (Bithell and Brasington, 2009; Troy et al., 2015; Mashhadi Ali et al., 2017; O'Connell, 2017). For socio-economic agents rules are derived from sociological system understandings. Thus, these coupled models are more fitted to scenario-based models, where validation is hard and sometimes impossible. ABMs that model hydrological systems are sparse in literature and mostly site-specific models (Servat, 2000; Reaney, 2008; Rakotoarisoa et al., 2014). In these studies the rules and boundary conditions are derived from well-known physical relationships. Hence, these models are able to be validated as long as a proxy is used that translates the agents into volume or fluxes. Like in different scientific fields, hydrological ABMs are only rarely transferred to different yet similar problems. This might be due to the bottom-up style of this modelling technique that is extremely linked to the problem, but also linked with a lack of a framework for ABMs in hydrology.

The lack of a general modelling framework and a conceptual scheme might have hindered the rise of the modelling technique in the past. In this study, a fundamental idea is proposed for the usage of this modelling technique as well in hydrology as in water resource management. As an initial framework for the application of ABM in physical hydrology, a framework for soil water movement and soil interactions is developed, called Integrated Platform for Agent-based modelling (IPA) which is published by Mewes and Schumann (2018b).

4.3 Framework development of an ABM for soil water movement and in-soil interactions

Following the requirements of an ABM framework (Sec. 4.1), the class of the dynamic agents, the class of static agents and the global agent, that controls the modelling experiment as a meta agent, are introduced (North, 2014; Macal and North, 2010). In order to set up an ABM for soil water modelling, some principal thoughts on the nature of software agents, their interaction with their environment and eventually the constitution of the model environment have to be given (Crooks et al., 2008). Software agents are encapsulated entities with a defined boundary and attributes, that follow rules to fulfil their goal (Macal and North, 2010; Blaschke et al., 2013). Agents interact with their environment through actors and interpret their environment through sensors, whose rules of interaction have to be defined *a-priori* by the modelling expert (Hofmann et al., 2015; Macal and North, 2010). The environment acts in form of a defined number of static agents that comprise all hydrologic agents within their spatial and temporal extent. All actions and interactions are coupled and lead to emergent system dynamics: Agent A decided to perform action I, which hinders Agent B to perform action II but leads B to perform action III and eventually force the environmental layer agent to influence Agent A's future decision. The IPA framework handles all agent

classes, the general composition of the modelled system and global model behavior and is designed to manage agents in a dynamic way to allow the composition of large scale ABM models through the underlying GAMA architecture in headless mode to save computational time (Taillandier et al., 2012; Boulaire et al., 2015).

In the IPA framework, the global agent manages all static agents (the layers) and dynamic agents (the hydrologic agents).The static agents get information from the hydrologic agents, e.g. how much water is already stored within the layer. Conversely, the layers share information on physical properties to the hydrologic agents. They require this information to calculate their movement speed based on the environmental parameters. Through the knowledge about the hydrologic agents inside the layer, the boundary conditions for each layer are checked. In contrast to the aforementioned inter-class communication between layer and agents, the intra-class communication of hydrologic agents is crucial for the decision of movement. In dilemma situations, e.g. in case that the target pore space is already covered, the intra-class communication is used to solve that dilemma situation by negotiating the different states of the hydrologic agent (Fig. 26).

In contrast to classical, equation-based modelling approaches, the amount of parameters for tuning is smaller (van Parunak et al., 1998) but the amount of computational time is higher, which results in a demand for parallelized computation either on GPUs or on high performance systems (Wang et al., 2013b). As mentioned before, the analysis of ABM outcome differs from the analysis of conceptual storage based models because an integral measure of e.g. runoff is rarely a measure. Neither is the fitting of a measurand the target of the ABM but to replicate the pattern that leads to the measurand. So, the pattern-oriented analysis is to prefer, especially in case of spatially-distributed ABM (Grimm et al., 2005).

4.3.1 Dynamic agents: hydrologic agents

4.3.1.1 *Class description of hydrologic agent*

Hydrologic agents are carriers of a constant amount of water w that defines their mass (represented as grey circles in Fig. 27). All agents carry the same amount of water, but their spatial extent is different because of changing environmental characteristics. Here, the spatial extent of the hydrologic agents is determined by a circle with radius r of an agent A that is influenced by the surrounding porosity Φ_E. So, the size of the hydrologic agent may change during its path through the soil column although its mass remains the same, due to a change in the porosity. For future applications the density ρ of the carried water is also included (but here set to 1).

$$r_A = \frac{w}{\Phi_E \rho} \tag{4.1}$$

The influence $I_{hA,L}$ of each hydrologic agent on the static layer agents can be quantified by the radial area that a hydrologic agent covers in relation to the complete area of a layer of the hydrologic agent (4.2):

$$I_{hA,L} = \frac{A_{hA,L}}{A_{hA}} \tag{4.2}$$

Here, $A_{hA,L}$ is the area of layer L covered by Agent hA, and A_{hA} is the area of the hydrologic agent hA. This influence can reach a maximum of 1 if the hydrologic agent covers only one layer, or smaller splits with a sum of 1 per agent, if it covers multiple layers. This influence is used to calculate the saturation of layers and the surrounding porosity Φ_E of hydrologic agents in the next time step. The saturation of layers Sat_L is calculated by the contributed amount of water w_{hA} of the agents located within the layer ($h_{A,0}$ ... $h_{A,N}$) weighted by the influence I_{hA} and the total pore volume of the layer *V*.

$$Sat_L = \frac{\sum_{hA,0}^{hA,N} I_{hA,i} w_{hA,i}}{V} \tag{4.3}$$

In order to analyze the possible future location of the agent, a cone-shaped view shed is constructed (light grey cones in Fig. 27). The view shed has a larger extent than the area of influence, although its length also depends on the radius r. Moreover, the saturated percolation speed of the agent in its environment given by the hydraulic conductivity *ks* and the chosen model time step Δt determine the view shed. This can be seen as a tool for numerical integration in the discretized model environment (Servat, 2000), as it shows the maximum distance the agent can travel within the next time step. The cone is constructed with an angle of φ =45 ° and the maximum distance *d* in Eq. (4.4) and saturated hydraulic conductivity denoted as *ks*.

$$d = \sqrt{r \cdot ks \cdot \Delta t} \tag{4.4}$$

The calculation of the view shed is influenced by Darcy's law incorporating the hydraulic conductivity. Hence, *ks* represents the possible step width and the time step of the model. As the agent has a spatial extent, the radius has to be considered as well. The angle φ is a parameter to include the variability of pathfinding due to different grain sizes in the soil structure. This model setup consist of an 1D soil column and the gradient is limited to one direction. The angle as well is limited to 45° in direction of the gravitational gradient. This angle is chosen because in the 1D case this angle represents the possible range of direction of movement without a substantially changed gradient. The direction and the speed of movement define the pathfinding algorithm of IPA.

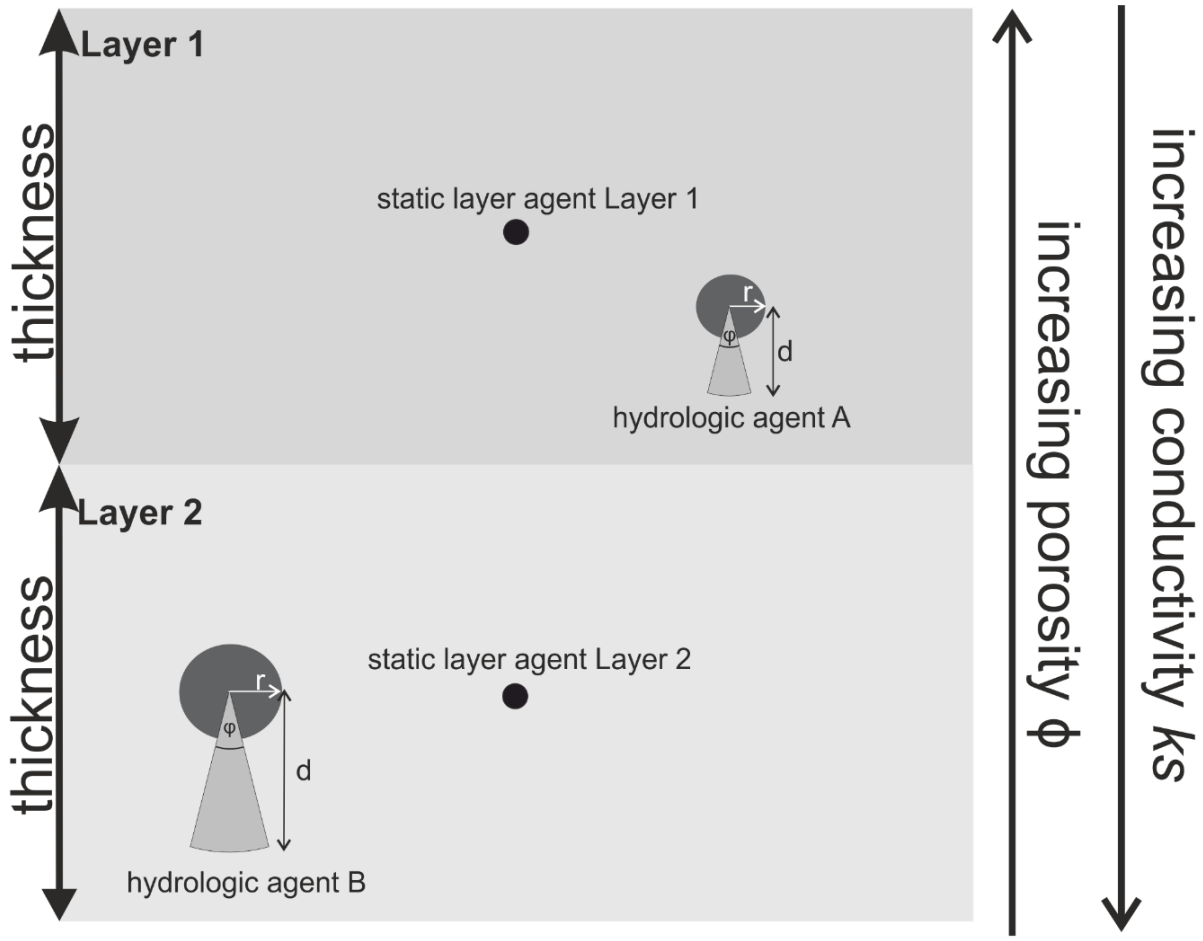


Fig. 27: IPA scheme with two layers with decreasing porosity per depth and two hydrologic agents.

The pathfinding is gradient-based and respects the entity of the hydrologic agents. If the available pore space in the target area is covered, no further movement is allowed and thus the agent is hindered from moving there. If that is the case, the agent tries to find a different target that lies within its direction of movement. Within an angle of 90° the gradient remains intact and thus allows the agent to take an indirect, hence longer way that reduces the distance to pass.

4.3.1.2 *Rule set for hydrologic agents*

The hydrologic agent has to decide whether to move or not to move. Once it has decided whether it shall move, the direction of movement has to be considered. In the case study, the rules of movement are defined by physical laws of soil water movement, which can be seen as a trade-off between the vertical forces of gravity ψ_G and matrix potential ψ_M that holds water against gravity. These forces are known as the driving potentials (van Genuchten, 1980). The osmotic potential ψ_O is neglected, which reduces the decision of each agent for movement to:

$$\psi_H = \psi_M - \psi_G \tag{4.5}$$

ψH equalling 0 means no downward movement with the potential gradient takes place, whereas $\psi H > 0$ leads to capillary rise. $\psi H < 0$ results in further deeper percolation of the agent with the speed k of the agent determined by a soil depended retention curve (van Genuchten, 1980; DBG Arbeitsgruppe Kennwerte des Bodengefüges, 2009). The speed k is the actual hydraulic conductivity that is higher than the saturated hydraulic conductivity ks. In case of an infiltration front moving through a wet soil, the matrix potential ψ_M can also be in the same sign with the gravitational potential ψ_G. In case that the future location of the agent at $ti + \Delta t$ is already occupied, the agent tries to find another route following its gradient of potential. Thus the running order, e.g. schedule of hydrologic agents is of importance, which is discussed later in Sec. 4.3.3. If no other route is possible, the particular agent's movement is suspended for this time step, or tick as it is called in agent-based modelling. The speed of the movement is given by the k-value of the surrounding area, which itself depends on the predominant moisture of the environment which is calculated by van Genuchten's model (van Genuchten, 1980). This model links soil moisture, the predominant potentials ψ_H, ψ_M and ψ_G with the saturated hydraulic conductivity *ks* and the hydraulic conductivity *k* which is higher under saturated conditions. The physical soil properties used to calculate van Genuchten's model (VG) are given in Tab. 15 in the appendix A.

4.3.2 Static agents: Layer agents

4.3.2.1 *Class description of layer agent*

As stated before, the layers act as static observing agents that survey all dynamic hydrologic agents that belong to their layer (Fig. 27). To each static layer agent a corresponding rectangular area is assigned, later on referenced as the layer. So, the global environment is discretized according to the available data on porosity and layer extents. The total volume of the modelled system is subdivided into a number of single layers. The corresponding soil moisture per layer is calculated by the sum of internal agents' carried water. As the detection of hydrologic affiliation to a specific layer is vulnerable to numerical artefacts from abrupt changes, the calculated soil moisture is smoothed by a univariate spline with a fifth degree. This spline was found to fit the characteristics of soil water content well, but still needs refinement as shown in the detailed analyses Sec. 4.4.5. Each layer controls whether the movement of agents is possible, such that problematic situations, e.g. over-saturation of layers, are avoided. With this the layer is like an internal boundary condition for the decision making process of the hydrologic agents. The interaction between hydrologic agents and the layer agents is bidirectional: Not only corresponding amounts of water carried by hydrologic agents alter layer processes but also the alteration of soil moisture content is coupled with future agent's decision due to the influence of the soil retention curve on speed and direction of movement (Grashey-Jansen and Timpf, 2010).

4.3.2.2 *Rule set for layer agents*

Static layer agents have various duties: they create hydrologic agents, monitor the soil moisture and oversee that all hydrologic agents act within the boundary conditions. For creation

of the hydrologic agents, an infiltration model has to be used. This can be a potential-based agent model, or as in this case a Green-Ampt (GA) approach of infiltration leading to the general assumption of a continuous movement of the infiltration front in the matrix. So, the infiltration in the upmost layer represents the upper boundary condition of the model. In this framework an environmental layer can be assigned with a GA infiltration which offers an approximation of GA fairly easy to compute (Ali et al., 2016).

Ali et al. (2016) presented an approximation to Green-Ampt where $F(t)$ represents the cumulative infiltration, S the sorption parameter defined by Ali et al. (2016), ks the saturated percolation velocity depending on the soil and t^* is a dimensionless infiltration time Eq. (4.6).

$$F(t)=\frac{S^2}{2ks}\left\{-1-\left\{\frac{t^*+\ln\left(1+t^*+\frac{\sqrt{2t^*}}{1+\frac{\sqrt{2t^*}}{6}}\right)}{\frac{1}{1+t^*+\frac{\sqrt{2t^*}}{1+\frac{\sqrt{2t^*}}{6}}}-1}\right\}\right\} \tag{4.6}$$

The cumulative infiltration *F(t)* is transformed into an actual infiltration rate *f(t)* which sets the number of hydrologic agents at a normally distributed random starting position in the up most layer Eq. (4.7). For the calculation of the infiltration into the soil column, the parameters given in Tab. 15 in the appendix are used to calculate the Green-Ampt infiltration in each time step.

$$f(t)=\left(F(t)-F(t-1)\right)/\Delta\mathrm{t} \tag{4.7}$$

The mass of the newly generated agents is fixed at a certain amount. This knowledge is of importance once IPA is able to run on either graphic card accelerated systems or on parallel computing platforms like cloud-based services, because memory allocation and data streaming between processing units becomes the bottleneck of performance and have to be formalized beforehand (S. Rybacki et al., 2009; Kofler et al., 2014). In contrast to the upper boundary of the model within the IPA framework, the lower boundary is defined by an outflow rate that relies on the ks value of the lowest layer. Once the centroid of the agent, given as the center of the circular shaped agent, has left the system, it dies and the carried amount of water accounts as outflow.

In IPA all layers or agents of the environment collect information about processes that take place within their extent and along their boundaries. In order to assign a weight depending on the distance of the hydrologic agents to the center of the layer, a density kernel approach

is applied to assign weights to each agent to smooth results and reduce numerical and graphical artefacts for the integration of all agent movements that are highly variable and are dependent on the simulated situation.

4.3.3 Global agent setup

Creating ABMs requires a planned scheme of the running order of processes, actions and actors (e.g. hydrologic agents and global agents) of the model. Thus, the unifying global agent, that combines model parameters, hydrologic agents and the observing layer agents, acts as the controlling unit of the whole model (Macal and North, 2010). This global agent controls the time and acts as an organizing agent, because it monitors the initialization of the model (at the beginning of the simulation), and asks hydrologic agents to register their layer belonging. Moreover, the global agent is able to force the observing agents to recalculate their state in terms of their current storage.

4.3.4 Model framework for comparison: cmf

For comparison purpose, a single column model , created in the cmf framework (Kraft et al., 2011) has been used. cmf was chosen because it offers an open framework for spatially-distributed process-based modelling (like solving the Richard's equation for unsaturated flow). Moreover, the general structure of cmf allows a spatial discretization and spatial modelling and can thus be seen as a possible benchmark for a hydrological agent-based modelling framework like IPA.

In the presented model setup, a single cmf cell, subdivided into ten layers with a depth of 10cm each, was used. The uppermost layer was connected by a GA infiltration process and a constant head of water available for infiltration upon surface. Transportation of water within the cmf soil column was calculated using the Richards equation for unsaturated flow and Darcy's law for saturated flow. The soil retention curve was modelled with the help of the van-Genuchten (VG) model. The outlet of the soil column was defined as a free boundary where water can exfiltrate from the system.

4.3.5 Model setup and parametrization of environment

For comparison of the general ability of the usage of IPA for the simulation of soil water movements, a simple synthetic scenario was created. A soil column with a height and a width of 1m (complete volume of 1 m^3) was used as the model setup. The soil column was divided into 10 single layers with a constant thickness of 10 cm. This column was filled with a homogenous sand soil (mS) with soil parameters given in Tab. 15 in the appendix A. All parameters applied in the VG model influence the calculation of *k* by the potential gradient and the saturation of the environment, whereas the GA parameters only affect the upper boundary condition. The chosen time step was 1h in order to reduce computational time, keeping in mind that a time step this long is vulnerable to numerical integration problems. By the reduction of time step length, less steps for the hydrologic agents had to be calculated. All

layers were equally pre-filled with soil water, resulting in a layer specific soil moisture θ of 20 % of available pore space. Although being different in their internal structure, the IPA model and the cmf model (Kraft et al., 2011) shared exactly the same model setup and parametrization. Infiltration was fed by an initial head of water of 1.0 m at the surface. The number of ticks was set to 90 which means that both models calculated 90 h or 3.75 days of infiltration and soil water movement. The model time was set long enough (3.75 days) to allow a deeper movement of the infiltration front through a set of layers, without the head of water on the surface becoming 0. In both models, the time step is chosen as 1h to increase comparability of results and remains constant in all figures and applications.

4.3.6 Performance measures

To estimate the quality of IPA representation of the water movement within the soil column, a suitable measure of performance had to be found. Both models deliver time series of their current states, the layer specific soil moisture θ. Here, the choice to measure the performance of IPA fell on the r² value.

$$r^2 = \frac{\sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} \tag{4.8}$$

Where on the one hand $\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ denotes the explained variation through the model and $\sum_{i=1}^{n} (y_i - \bar{y})^2$ on the other hand depicts the total variation, and *n* as the number of given samples in both models. A value of $r^2 = 1$ would show a perfect model fit, whereas an $r^2 = 0$ would mean that no variation is explained through the new model. This measure of determination, or in this case, measure of error helps to quantify the relation between known model structure and new modelling approach without the need to specify the differences in model structure in detail and is suitable measure of performance for model comparison.

4.4 Comparison results of IPA and cmf

To judge IPAs general modelling capabilities, the synthetic experiments were compared for both model types. As said before, cmf was considered as the true estimate. Both models showed no numerical errors over the simulation. The volume error of both models in the end with values of about 1%, which are considered neglectable. With a runtime of 1.1 min on an i5 with 2.2 GHz, 8GB RAM IPA computed only slightly slower than the numerical cmf model that needed about 30s on the same setup. Running IPA in headless mode without graphical output, the computational time was reduced to 48s. Further reduction of computational time could be archived by an outsourcing of the pathfinding to the graphical computation unit.

4.4.1 Experiment: homogenous soil column

Comparing both model results, it became obvious that results were not identical, yet the dynamics were similar (Fig. 28). The development of soil moisture in the layers followed the same pattern. Saturation reached a similar level for the first three layers, while the velocity of saturation was different in IPA from the cmf results. Layer 1 did not saturate as fast as in cmf, but movement from Layer 1 to Layer 2 started earlier in IPA. In cmf, soil water movement from the uppermost layer to the next lower layer started after approximately 7 h while the agent-based model triggered movement of hydrologic agents immediately after the first hour of simulation. After 70 h both models showed saturation in the first layer, so both models reached the same final stable state. In both models the layers were nearly completely saturated at a soil moisture of about 32.8%. The transport from Layer 2 to Layer 3 started in both models 21h after the beginning of the simulation. Meanwhile, cmf showed a numerically smoother behavior than IPA, while the general system behavior is similar as one can see it in the variation of soil moisture in all layers in IPA, although some numerical oscillations in the soil moisture of Layer 2 became visible.

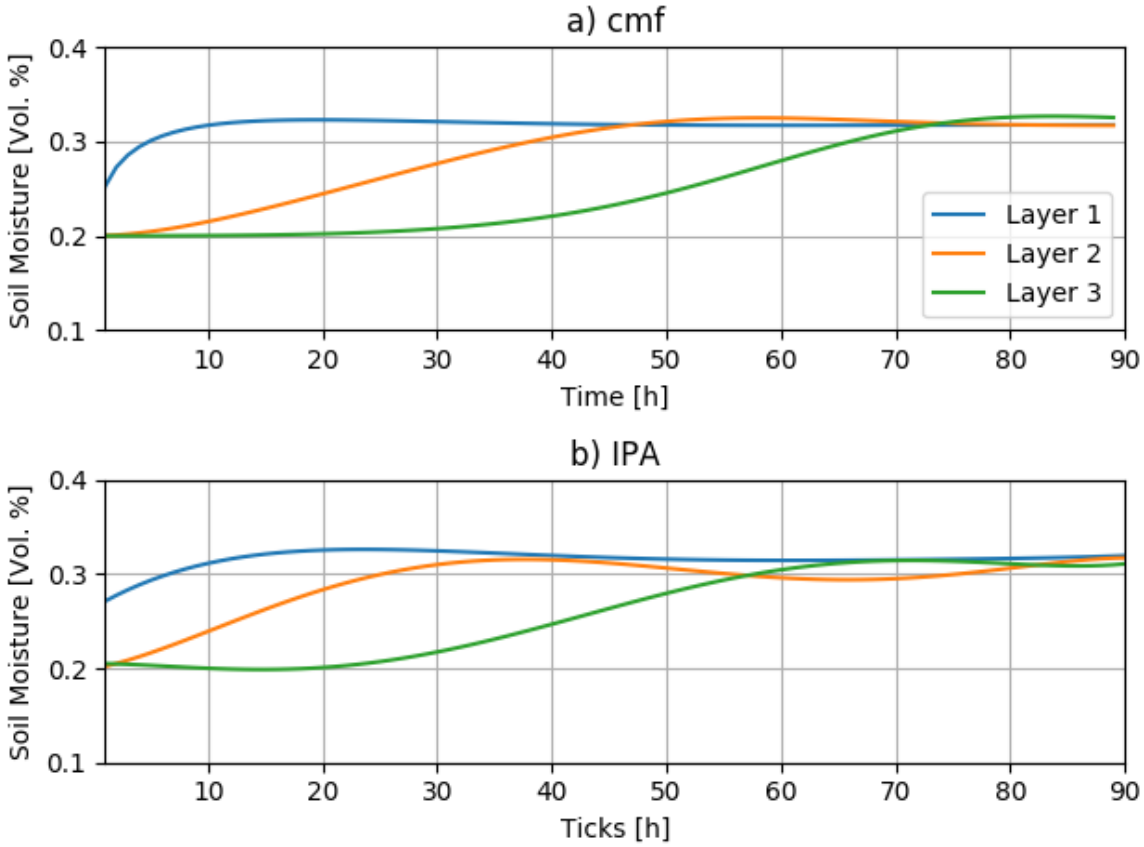


Fig. 28: Comparison of soil moisture development of the upmost three layers with a homogenous soil in the column.

To express the accordance between IPA and cmf for this run, the corresponding r^2 value was used. Here the mean r^2 value of the upper three layer scored $r^2 = 0.80$. The standard deviation of both models is slightly different 0.039 % (cmf) to 0.045 % (IPA) while the mean values

of soil moisture were the same (Tab. 16 in the appendix A).

4.4.2 Experiment: soil column with heterogeneous soil

As stated before, the synthetic case was extended to a more complex situation of two heterogeneous soil types. In order to show the general ability of IPA to model complex systems, the 1D soil column was packed with two different soils leading to the problem of a boundary between two different types of soils with different physical properties. The geometry and the discretization of the grid for the cmf-model remained the same, but the topmost layer consists of Su2 (a weakly silty sand) instead of mS. Su2 was chosen because although it has different physical characteristics, it is still a relative to the original mS soil with a lower share of sand but a higher share of clay. This change of soil type affects highly the process of infiltration and the transition between Layer 1 and Layer 2. None of the other layers was changed, so the ability of IPA to simulate with its pathfinding algorithm (as introduced in Sec. 4.3.1.1) and its suspension of movement approach was tested in regard to the added layer transition between Su2 and mS.

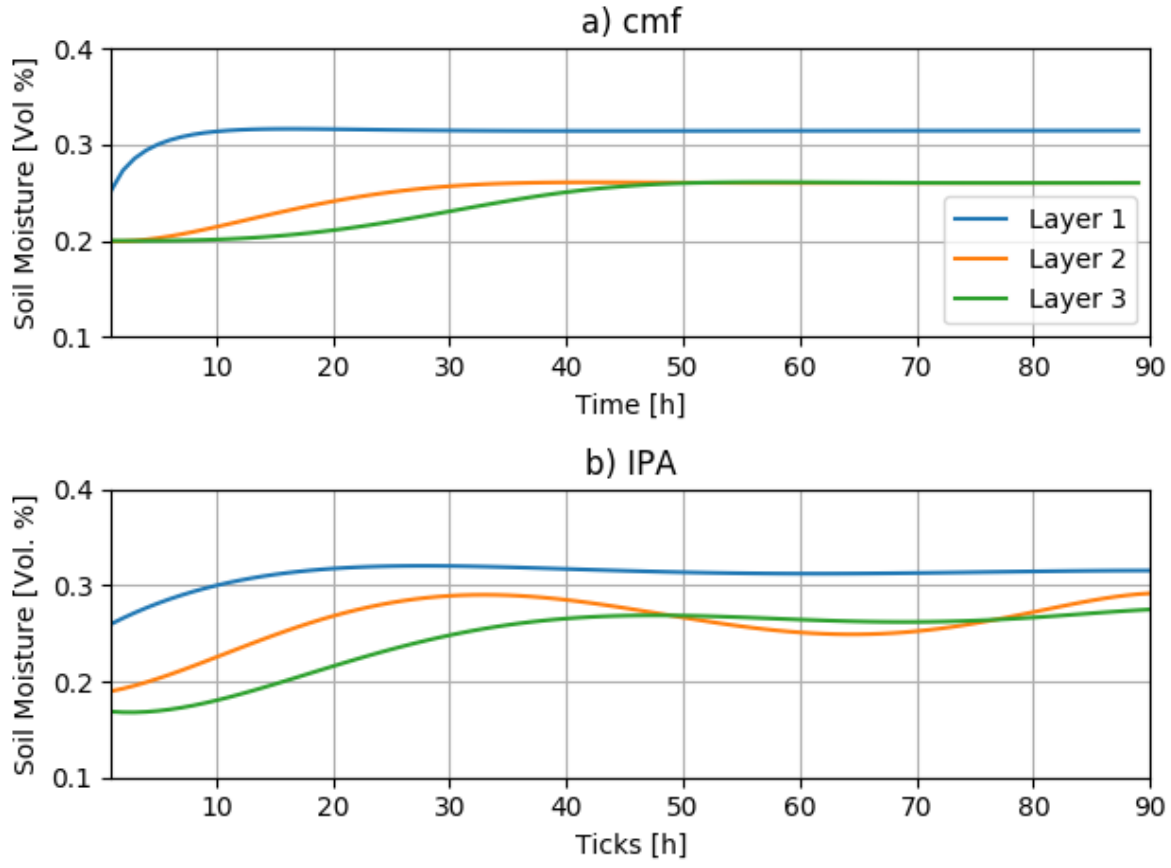


Fig. 29: Comparison of soil moisture development of the upmost three layers with a transition boundary between Layer 1 and Layer 2.

Again, both models showed a similar, yet slightly different behavior (Fig. 29). Transport from Layer 1 to Layer 2 started immediately as did the movement of water between Layer 2 and Layer 3 in IPA. Saturation of Layer 1 in IPA is reached slower than in cmf but the result

after 40h of simulation is a stable system with comparable saturation near full saturation in all layers, although the general IPA behavior was less smooth than cmf. IPA showed slightly higher saturation of about 27% in contrast to 26 % for Layer 2 and 3 in cmf. The saturation of the infiltration Layer 1 shows for both models exactly the same values. The r^2 value scores 0.71, indicating a high correlation between the outcomes of both models, even though the dynamics between Layer 2 and 3 differ from those in cmf. This could be related to the dilemma of spatiality in the agent-based model as all hydrologic agents have a certain shape and it is likely that this shape had a significant influence on the model outcome. The slightly higher saturation, might be the cause from the boundary conditions that the global surveying agents has to check to avoid oversaturation.

4.4.3 Influence of model scheduling

As mentioned before, scheduling of agent actions is a sensible question in agent-based modelling. Especially in the context of parallelization, the question on groups of agents that update their state simultaneously. Here, three different methods for scheduling were implemented:

- Random calling of agents, that calls agents randomly by chance
- Energy-based scheduling, that allows agents with higher gradients to move first
- Age-based scheduling, allow a movement according to the age (either young first, or old first)

In order to test the influence of the scheduling approaches on the representation capacity of IPA, a test with the same setup presented in the precedent study, was performed (Fig. 30). Random calling of agents was the easiest way to use: Every tick the running order of hydrologic agents was determined randomly. It could be seen that a random scheduling led to huge smoothing errors because the energy gradient of each agent (the current state of the agent) was not taken into account. Deeper layers showed more fluctuations of soil moisture as it could be seen from tick 80 - 90. To overcome this random approach, an energy-based approach was developed: Those agents with the highest energy-gradient were allowed to move first, which resulted in smoother results with less numerical fluctuations. This was the case, because the advantage of hydrologic agents with a high potential energy limited conflicts between slow and fast moving agents. Moreover, it helped to moderate conflicts in pathfinding through a clearly defined regulation which agents had priority in moving first, trying to get their potentials in balance. Last but not least, an age-based way to organize the running order was implemented. The age was anticipated by the name, because the unique names of all agents were not reused as soon as an agent has left the system but originate from consecutive numbering during creation. In the test, old water was allowed to move first, so the scheduling was in a decreasing order. This approach had some problems with the distribution of old water, because the names of those agents that represent old water were rather similar because they had been created during initialization.

The correlation coefficients showed that all types of scheduling had less impact in Layer 1 than in Layer 2, because all methods did have a high correlation among each other (Tab. 18, Tab. 19). However, correlation coefficients for Layer 2 showed that the energy based approach and the age-based approach had high correlation but struggled less with numerical artefacts like Random Calling. The soil moisture of the upmost layer was in all three cases nearly identical, which shows once more the dependency of the state of the infiltration-affected layer from the chosen infiltration model. From this analysis, it was clear that the energy-based approach seemed to be the best fitting approach. These scheduling approaches may be of interest for upcoming application of IPA because technically this scheduling is the major impact factor on the decision making processes of the hydrologic agents, as one can see in the analysis, and can be used for hypothesis testing for the behavior of water.

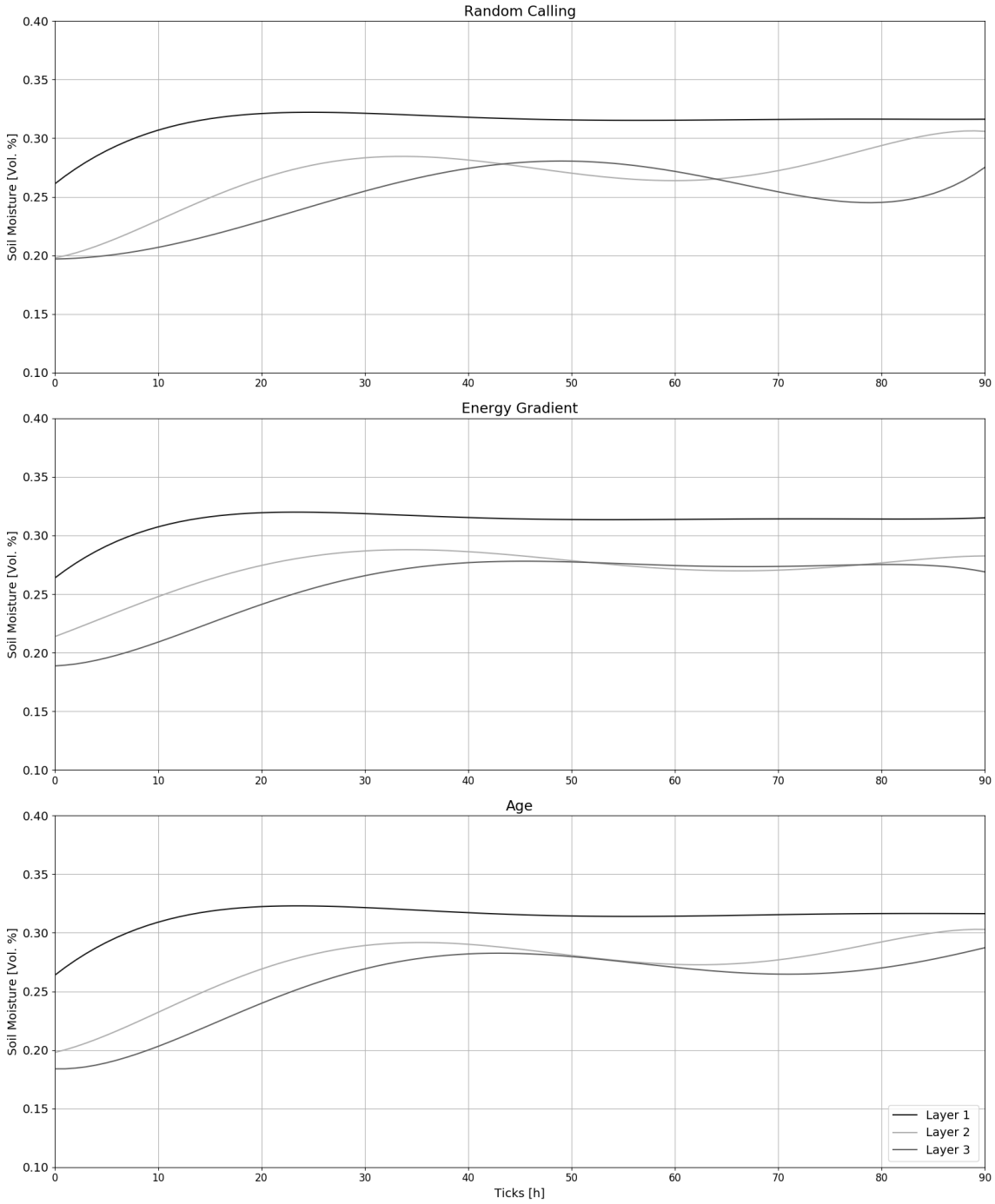


Fig. 30: Analysis of different scheduling methods for soil column with two different soils.

4.4.4 Impact of randomly chosen starting point of hydrologic agents after creation

Another spatial agent-based modelling specific problem was the starting position for the hydrologic agents within the system. The process of infiltration describes the spatial transition of water from the surface to the soil matrix. Therefore, one can assume that each hydrologic agent is located with its complete shape in the topmost layer somewhere near the upper boundary. The x-coordinates within this layer were chosen randomly around the top of the layer, but always deep enough in the soil such that its shape was completely within the layer.

In order to verify the assumptions and to show the impact of different starting positions one can show the influence of the chosen starting position for the same model set up with 20 runs. The starting position was chosen by a random normal distribution with $\overline{x} = \overline{Wd}$ and $\sigma = p \cdot \overline{Wd}$ while p was varied from 0.1 to 0.9 in 20 steps and $\overline{Wd}$ was the width of layer, in the case study 1m. As one can clearly see, soil moisture in the uppermost layer was only affected by infiltration because calculated soil moisture was nearly constant without any visible influence of the choice of starting position, which makes sense as the hydrologic agent is always located completely within it's the infiltration layer (Fig. 31).

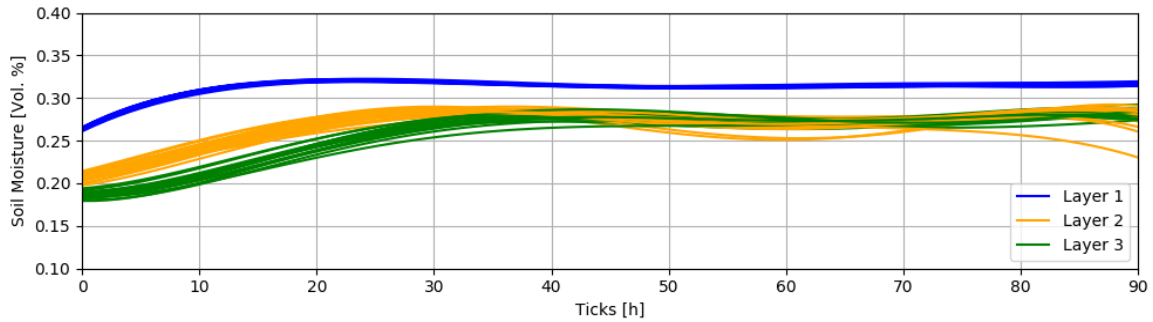


Fig. 31: Influence of randomly chosen starting point. Calculated with 20 runs and a model setup with two different soil types.

Thus the relevant layers were the deeper layers 2 and 3. Both showed slight variations that look like numerical oscillations which makes sense as the smoothing affects the calculation of layer soil moisture, because the starting position affected the speed and the pathfinding of the hydrologic agents. The maximum difference between the estimated soil moistures per layer was at 3 % for Layer 2 and 3 and at 0.5% at Layer 1 for the 20 runs. Yet, the variance in soil moisture was visible, so, a multi-run of n runs should solve the problem and consequently a mean of these n runs reducde this numerical artefact effectively that were introduced by the random starting point of the agent (Fig. 32).

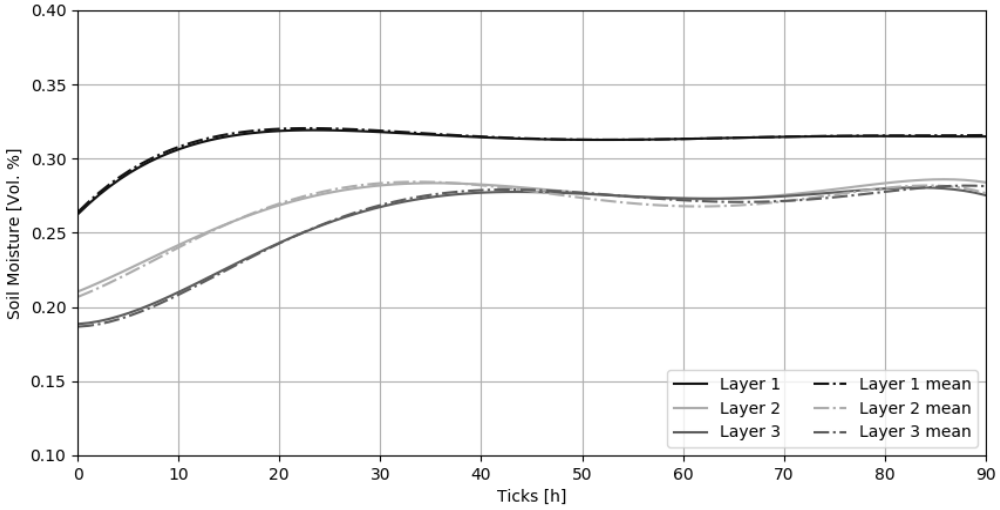


Fig. 32: Mean resulting soil moisture after 20 runs to reduce effects of randomly chosen starting position.

4.4.5 Weight assignation: From univariate, fitted spline towards more comprehensible methods

In the first step for each hydrologic agent a weight for influence on the layer was assigned by a fitted univariate spline with degree 5 in order to smooth the numerical artefacts from the calculation of the layer affiliation of hydrologic agents. As univariate splines fit well, but interpretation and transfer to other applications is difficult, a density kernel estimator with a simple logarithmic distance function to assign a weight, where w_i denotes the weight of the specific hydrologic agent i at distance d_l from the layer l with whom it has a spatial intersection Eq. (4.9) was chosen. This distance is normalized by the maximum possible distance that a hydrologic agent centroid may have with a corresponding layer agent at distance ld, which is defined as the maximum of layer depth or the most far away located agent that still corresponds to the layer's moisture:

$$w_i = \frac{\ln\left(d_i\right)}{\max\left(d_l, ld\right)} \tag{4.9}$$

Hydrologic agents lose their influence on the layer moisture with increasing distance from the static layer agent representing the layer center (Fig. 33). Implementing a new weight assignation in IPA removed the demand for smoothing the soil moisture per layer. Results for the two-layered synthetic case showed that the approach is promising, although it is not fully usable because numerical artefacts still appear (especially in Layer 2), where fluctuations around the correct mean soil moisture for this layer occurred with the relative strong variation of about 5 % of soil moisture (Fig. 34). Layer 1 was modelled better with less fluctuations, the soil moisture raises faster, maximum soil moisture is as well modelled correctly and showed only little numerical oscillation. In Layer 1 the layer affiliation of each agent was only relevant during the transition from the layer of origin to the target layer. The

overall r^2 value was lower with 0.62 than for the spline smoothing, but mean moisture was nearly the same (Tab. 3).

Tab. 3: Statistical parameters from model comparison between cmf and IPA with kernel-based weight determination

Model	Std [%]	Mean [%]	r^2
cmf	0.033	0.27	0.62
IPA	0.042	0.26	

The general interpretation that the model showed similar dynamics was supported by an r^2 value that as higher than 0.5, but yet the standard deviation for both modelling approaches was much higher with 0.042 % regarding 0.0332 % for cmf or 0.0363 % for spline-based IPA. The kernel-based weighting approach looks promising, as the chosen function is easier to interpret than a univariate spline. But, for future applications, this weight determination has to be improved and might be as well part of a study regarding different distance weighting functions and the construction of a method to quickly find the appropriate function.

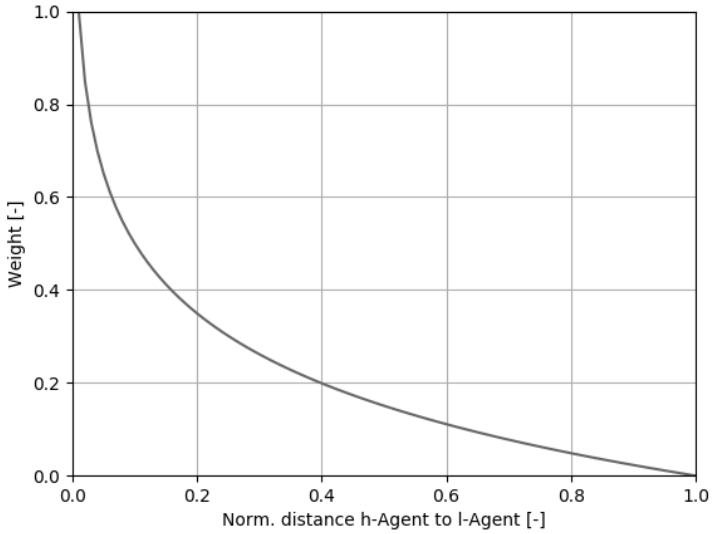


Fig. 33: Decreasing weight with increasing distance of agent's centroid to layer centroid.

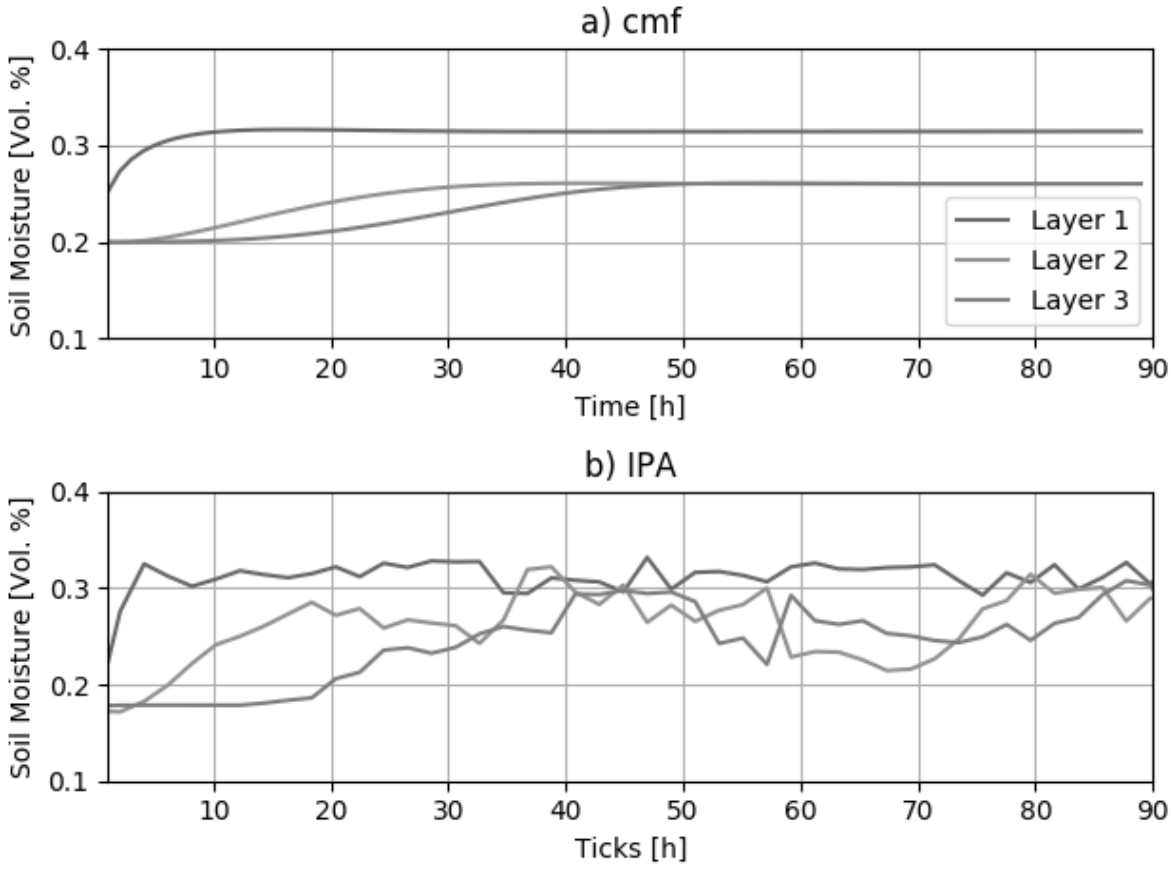


Fig. 34: Modelled soil moisture without spline smoothing but logarithmic kernel weight assignation.

4.5 Conclusion of ABM in physical hydrological models

ABM is in the context of alternative modelling and data mining approaches a promising addition to classical modelling techniques. The approach fits well for complex hypothesis testing, especially for physically-based process models. Here, the rules are given by physics and have to be applied to a construct of entities of water. The construct of an encapsulated water entity requires more refinement, but the idea of having water agents transporting a certain mass through a system interacting with their environment has its advantages for chemical modelling as well models that include density changes of the water during its course through the model. So, the concept of entities of water is one of the main limitations of the modelling approach because it migrates the problem of defining one entity of water from the catchment- or layer-sized scale down to the scale of particles but remains nevertheless a downside for this type of modelling.

It was shown that the proposed approach of agent-based modelling could be used for answering detailed physical-based hydrological questions like the movement of soil water through a soil column. From comparison with a conceptional cmf-based model, one can

conclude that an agent-based approach performs well. Furthermore, its results are comparable to those of a classical Richard's model with Green-Ampt infiltration set up within the cmf framework which is found to be a suitable environment for modelling complex, spatially distributed hydrological situations on a physical basis. The comparison revealed some further tasks as problems arise from the agent-based modelling dogma: The smoothing for the calculation of layer moisture needs further refinement, as a spline requires too many degrees of freedom for the task of assignation of weights to each single hydrologic agent (Servat, 2000). A different kernel function is required (instead of a univariate spline) for better explanatory power of the smoothing process that is needed to compare the highly discretized hydrologic agents with the rather rugged layers in cmf. Moreover, the scheduling needs more refinement, especially in terms of age-based scheduling that still has a high random component, as it became clear during the analysis of different scheduling techniques. From the case study presented here, the age and energy-gradient scheduling methods for hydrologic agents are the most promising approaches, depending on the research question.

The computational time of the IPA model is slightly higher than for the cmf model. The required computation time could be further lowered by running the framework headless, which could be a suitable approach for multi-run optimization approaches like Monte-Carlo simulations. The advantages of ABM compensate for the disadvantages of the approach presented before. For example, the possibility to implement different scheduling approaches is a useful approach to test hypothesis of water movement in the soil column based on its age or energy-gradient.

In future research, a focus will lie on possible age distributions for hydrologic agents that represent old water. Through that, one can ameliorate the suspension process that hinders a hydrologic agent from moving in favor of another hydrologic agent that blocks the route along the gradient of forces. So, the commonly observed phenomenon of residual old water or pushed out old water due to fresh water intrusion can be modelled. In fact, an age-based scheduling also allows finer modelling of hydro-chemical and small-scale pedophysical processes that occur during the transport of water through the soil matrix than common storage models. Another interesting usage of such a refined age-based scheduling is the residence time of water within a coarse rock glacier where melt water is released during rather short melting periods and the water draining from these rock glaciers shows different signatures of age, proving that some water refreezes during the melt and its drain is delayed to later melting periods. Potential applications of spatially distributed hydrological agent-based models are numerous and IPA might be a suitable framework to answer more complex hydrological questions by adding new rule sets. For the modelling of macro-pore effects on the soil water movement, the principle of agent-based modelling can be interesting: The fastest hydrologic agents wet the surrounding matrix and allow the following water to use the macro-pore as a short-cut through the matrix until the pore is filled. Through the existence of other hydrologic agents the macro-pore is filled and travel speed is lowered which results

in an alternative pathfinding through the matrix because the potential gradient allows a dispersion from the pore to the matrix. Next to residence times, the contact between entities and thus the exchange of attributes between the agents might be another application of the AB approach. Here, the suspension of chemical tracers and the respective degradation can be modelled on a spatial and temporal explicit scale.

Overall, one can say that IPA and generally agent-based hydrological process models are at their beginning. In times of big data and a plethora of highly resolved data, this new modelling approach can be of use for those questions where system behavior can easier be described with dynamic agent-based models than with stiff storage-based models. Within the examples it was shown that the new modelling approach is as good (or as bad) as traditional, storage-based models (like those created in the cmf model) but offer the variability of extension by rules and different scheduling routines. Even at this stage, an agent-based process model offers a great variability in model design for future research questions as it is able to depict the changing dynamics of model components like in nutrient transport models or complex rock glacier models with changing internal model constellations. By that, the aforementioned variable inner structure of agent-based models extends the modeler's capabilities to describe those systems.

5 Computational intelligence in data mining by agent-based classification

Apart from novel modelling approaches, big data also requires novel data mining techniques like classification of unknown data sets (Chaney et al., 2018; . In the upcoming section, the merits of agent-based classification (ABC) and the advantages of soft-computing are investigated. Therefore, the pixel-wise classification approaches that are common in the hydrological or remote sensing application of image classification has been compared to the novel approach.

Overall, the agent in its core concept remains the same as in agent-based modelling. The spatial extent is given by the image object. So, a major disadvantage of the IPA is not present in the ABC application. Similar to its modelling sibling, the image object agent has sensors and actors for its environment and is embedded in a topology of other image object agents. The relations in the neighborhood are known and required for the communication among the image object agents. The scheduling does not pose a problem for this formulation of the ABC framework. In contrast to the ABM, where all actions are taking place on the same action map, all actions in the ABC are parallelized and split onto multiple maps. The best improvement by the ABC decides the changes on the general map.

5.1 Fundaments and origin of agent-based classification

Next to ABM, Agent-based classification (ABC) approaches have become a matter of research (Chen et al., 2018a; Chen et al., 2018b). Especially in the interpretation of remote sensing images ABC is of interest and might utilize the principles of agent-based computing for better classification results (Hay et al., 2005; Blaschke et al., 2013; Borna et al., 2014; Peña et al., 2014; Hofmann et al., 2016; Hofmann, 2017). Here, image object agents are the successors of the object-based image analysis, where a remote sensing scene is not interpreted pixel-wise but on the basis of similar objects. In the traditional, pixel-wise image classification each pixel is analyzed on its spectral information (Tso and Mather, 2009; Lillesand et al., 2014).

In remote sensing, image interpretation is the common method to retrieve a desired information from remotely sensed data. As sensors do not directly measure the information, like evapotranspiration, ground-true data is used to link known information with spectral or radiometric signals. Patterns in the data between the different signals can either be detected by natural breaks, the so called unsupervised classification, or by regions of interest that act as

training regions with a manually assigned class (Lillesand et al., 2014).

In the traditional approaches for each pixel a class is assigned. Even though neighboring pixels may have the same class, the pixels do not represent an object but a collection of isolated and blind individual pixels. In the object-based image classification, pixels that share similar characteristics are merged by a segmentation algorithm to similar objects. These so-created objects have a geospatial extent and are embedded in a neighborhood of other objects. This network or topology of objects can be incorporated in the analysis and adds information to the image interpretation process that else could not be used or derived from the existing data (Lang et al., 2014; Chen et al., 2018a; Wang et al., 2018). The objects are then classified by the classification scheme.

ABC takes object-based classification as fundamental (Hofmann et al., 2016). This means that pixels with similar characteristics are combined as meaningful objects. This combination is performed by a segmentation algorithm that segments the remote sensing scene into separated objects. These objects are the initial image object agents. The image object agents share the fundamentals with the hydrologic agents from Sec. 4. They are embedded objects that try to achieve a goal by choosing independently from a pre-defined set of actions. The image agents have a major advantage over the pixel-wise image classification and the object-based classification: The agents communicate among each other and allow therefore an optimization of the classification within the structure of image objects.

Like ABM an ABC requires a software environment that allows the organization of autonomous software units. As an additional requirement one can add the support of stacked remote sensing images, the so-called scenes. Moreover, the geospatial topology of objects has to remain intact, even if the shape or the position of the object is altered. Furthermore, the objects have to work on different layers to evaluate the outcome of different options. In this study, the choice as the environment fell on eCognition Developer (eCognition Developer, 2014). eCognition allows to segment images and maintain the network of segmented objects.

5.2 Delineation of irrigated agriculture in Nebraska with ABC

For the purpose of demonstrating the advantages of ABC, the novel approach was applied to delineate irrigated agriculture in Nebraska, USA. The application of ABC was hence focused on water resource management. Irrigation is a major component in the hydrological circle and widely influenced by human activities and decisions (FAO, 2012). Yet, spatial information on irrigation is sparse, although the information is crucial for the development of global water usage models and their respective application in the calculation of scenarios (Siebert et al., 2010; Hoogeveen et al., 2015; Salmon et al., 2015; Meier et al., 2017;). Therefore, many remote sensing studies cover crop identification and the existence of irrigation, especially in data sparse regions (Debats et al., 2016; Boyaci et al., 2017; Pun et al., 2017)

To demonstrate advantages and possible downsides of ABC in contrast to established classification approaches, a comparative study showed the major improvements by the approach as presented and was published in Mewes and Schumann (2018a). To compare the fundamentally different approaches, a shared theoretical background with similar classes and labelling strategy needs to be developed (Berhane et al., 2018). Hence, the choice fell on a fuzzy labelling scheme that assigns each target (either pixel, object or image agent) a class with a certain membership μ. Eventually, each object has a membership value for each possible class. The maximum membership μ defines the resulting class of the target. This approach is applicable to all presented examples. Moreover, it fits well to the fuzzy nature of object- and agent-based classification (Benz et al., 2004; Belgiu et al., 2014a; Hofmann, 2017).

5.2.1 Study region and reference data

The plains of Nebraska, USA, were chosen as region for an initial application due to the available reference data on spatial distribution of irrigated agriculture from COperative Hydrological STudy map, COHYST (Center for Advanced Land Management Information Techniques, 2005). From the complete state of Nebraska, two example regions were chosen with low amounts of urban land cover and a high share of agriculture from the CropDataLayer (CDL, Johnson and Mueller, 2010). As spectral input, Landsat 5 Enhanced Thematic Mapper (ETM) remote scenes were chosen. The scenes were stitched for the region of Nebraska, divided into quarterly observations of three months to guarantee a cloud-free image covering the time-span from January 2005 to December 2007. So, the data was split into mean images covering the seasons 1 – 4 covering a time span of 3 months starting with season 1 in January and ending with season 4 in December.

The choice fell on Landsat 5 TOA data due to the reference data which was updated the last time in 2005, when Landsat 7 already failed and Landsat 8 was not yet launched. Landsat 5 delivers spectral information on seven different bands with a 30 m spatial resolution:

Tab. 4: Wavelength and bands of Landsat 5 platform (Lillesand et al., 2014)

Band	Wavelength [µm]
Band 1 - Blue	0.45 - 0.52
Band 2 - Green	0.52 - 0.60
Band 3 - Red	0.63 - 0.69
Band 4 - Near Infrared (NIR)	0.77 - 0.90
Band 5 - Short-wave Infrared	1.55 - 1.75
Band 6 - Thermal Infrared	10.40 - 12.50
Band 7 - Short-wave Infrared	2.09 - 2.35

As one can see in the true-color image (Fig. 35), both regions showed signs of the existence of irrigated agriculture. At the pivot irrigation plots, one can assume that the time when the scene was captured by the sensor, this land was either vegetated (presented in green) or barren (brownish colors). On the northern and eastern border of Region B, erroneous input data are visible where the mosaicking of different Landsat scenes and paths was conducted. Urban areas were not visible in the true-color images, a cross-check with landuse data also revealed no further human use in the investigated regions.

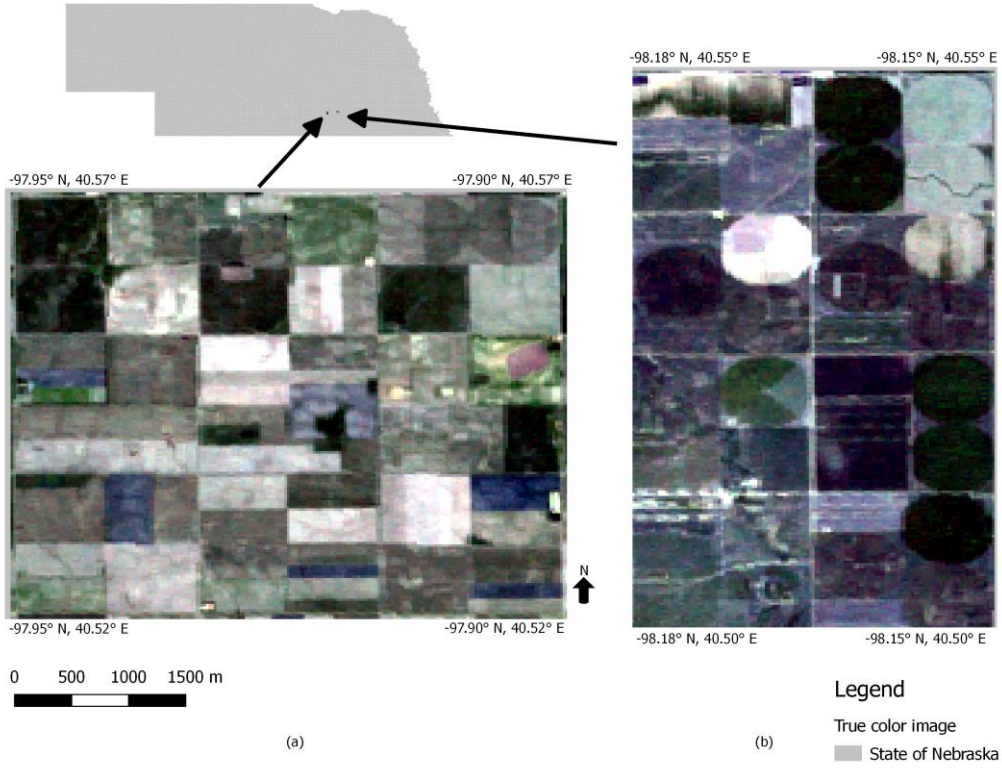


Fig. 35: True-color image of both regions investigated. Agricultural objects become visible in each region, represented through homogenous, clearly differentiable spatial objects, like pivot irrigation.

As computational time for the agent-based approach is high, only two relatively small subsets with a high variety of crops grown and different irrigation techniques were selected. Region A was chosen manually according to the obvious existence of radially irrigated plots and Region B by a stratified random sample that comprises the highest variability of grown crops from CDL in Nebraska (Johnson and Mueller, 2010). Moreover, the choice was restricted to regions without urban influence to lower the chance of confusion. The size of the subsets is similar (Tab. 5). Both subsets contained irrigated and non-irrigated areas, which qualified them for the investigation of the presented workflow.

Tab. 5: Size of investigated regions (in pixel) in Nebraska with a pixel resolution of 30m

Size Region A	Size Region B
142 x 219	219 x 165

All spectral data was gathered from Google Earth Engine (GEE), which is a freely available remote sensing cloud-based archive comprising also data like CDL directly from the provider without any other database system for queries and data pre-processing. All named indices were calculated in the GEE IDE (Integrated Development Environment).

5.2.2 Spectral indices for the identification of irrigated agriculture

For the classification of irrigated agriculture, several indices from spectral data were calculated: The NDVI (Normalized Difference Vegetation Index) and NGI (Normalized Green Index) were computed from corrected top of atmosphere (TOA) reflectance values captured by the Landsat 5 Enhanced Mapper. To cover the seasonal variability for each season a cloud-free mean Landsat image covering 3 months was mosaicked.

The NDVI is a normalized index Eq. (5.1), which reveals the vegetation activity and is computed through the visible red band ρ_{red} and the near infrared band ρ_{nir} ranging from -1 to 1 (Lillesand et al., 2014). The NDVI is a well-known index to describe the plant activity and health. Due to the choice of widely available bands from infrared and near infrared, the NDVI can be calculated from nearly any spectral sensor. This lowers the bias by the index and emphasizes the universality of the classification approach comparison.

$$NDVI = \frac{\rho_{nir} - \rho_{red}}{\rho_{nir} + \rho_{red}} \tag{5.1}$$

The NGI is an index more robust to identify irrigation as it takes the green activity into account (Pun et al., 2017). It is the product of the Green Index (GI) and the aforementioned NDVI (Boyaci et al., 2017). The choice of the NGI as an additional spectral index puts a higher weight on the green activity of the plant. This can be used to distinguish between irrigated plants that grow within their potential in contrast to plants that suffer from drought and a water gap in the growing season.

$$NGI = NDVI \cdot GI \tag{5.2}$$

Finally, the GI is defined through the band ratio where ρ_{nir} again is the near infrared band and ρ_{green} stands for the green band:

$$GI = \frac{\rho_{nir}}{\rho_{green}} \tag{5.3}$$

This relation between the near infrared (or thermal infrared) and the green band of the sensor allows to judge the plant activity in the leaf. It covers the chlorophyll activity and the resulting transpiration of the plant (Ozdarici-Ok et al., 2015).

5.2.3 Fuzzy classification scheme

As mentioned before, the labelling was conducted via a one-to-many fuzzy classification scheme that was built by an ontology (Arvor et al., 2013; Belgiu et al., 2014b; Andrés et al., 2017). An ontology is a kind of guide book combining the information given by the objects

and the chosen spectral indices and ancillary information. The result is the membership μ to the respective class if the classification scheme is described in a fuzzy manor (Benz et al., 2004). The maximum μ defines the class of the object or agent.

The class *Agr* was likely to be assigned if the shape of the target object was either round or rectangular, due to the human influence on the shape of the agriculturally used plot (Fig. 36). Moreover, the object must have exhibited NDVI values > 0. The class *Irr* was assigned once that the target had a high similarity to class *Agr* but also showed higher NGI values and the distance to a neighboring irrigated plot was small. The barren soil class *Barren* had characteristically low NDVI values but the shape of the object had an anthropogenic character like objects of class *Agr* or *Irr*. The membership to the class *Barren* decreased near NDVI = 0 where the vegetated part of the index began.

In this approach, the class *Irr* comprises a characteristic, limited to the agent-based approach: The distance to the next object labelled as *Irr*. This characteristic is limited to agent-based classification, as other agents can only be identified as *Irr* after a first iteration. In this case, one can assume that irrigated plots are located near to other irrigated plots. So, the membership to the class *Irr* increased the lower the distance to the next irrigated plot was. This proximity criterion was restricted to the object- or agent-based approach because the distance to the next object was merely dominated by the grid structure of the pixels.

These membership functions are adjustable to the problem and require expert knowledge of the objects to classify the target region: e.g. the NDVI curve for vegetation depends on the investigated region with its variety of crops and cropping techniques. Here the knowledge-based characteristic of agent-based computing becomes obvious again. Without the expert knowledge on the behavior of classes the classification scheme could not be attributed.

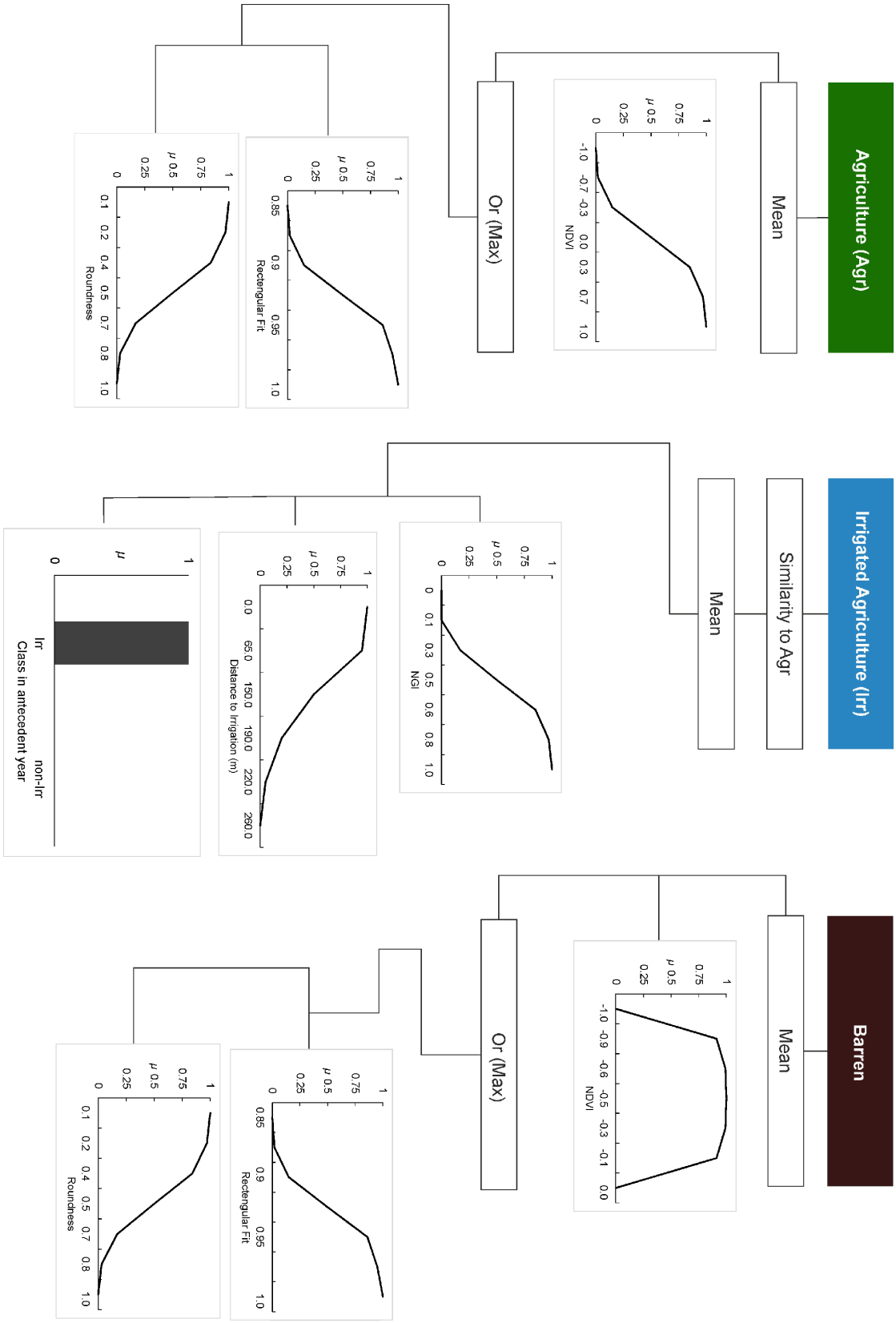


Fig. 36: Ontology of membership functions for classes non-irrigated Agriculture (*Agr*), irrigated Agriculture (*Irr*) and Barren.

5.2.4 Object-based classification for the delineation of irrigated agriculture

As mentioned before, object-based image classification requires a segmentation of the scene into meaningful objects. Hence, an appropriate algorithm for segregation is required. The choice fell to the multiresolution segmentation algorithm implemented in eCognition Developer 9.3 (eCognition Developer, 2014). This multiresolution segmentation algorithm (MRS) starts with a single pixel and merges neighboring pixels until the threshold value of shape is reached or the neighboring pixels are not similar in specified characteristic like a mean band value or the texture (Baatz and Schäpe, 2000). The parametrization of the segmentation process remains a problem that is often solved by trial and error (Hay et al., 2005) or specialized tools. The parameters used for the multiresolution segmentation were estimated by the estimation of scale parameter (ESP) approach (Drăguţ et al., 2010) and fitted a set of segmentation parameters for the whole scene to lower the classification error (Tab. 6).

Tab. 6: Segmentation parameters from ESP

Scale	Shape	Compactness
5	0.1	0.5

As spectral input, NDVI and the NGI were applied to perform a segmentation for each seasonal image. This segmentation finally led to 200 – 500 image objects per scene that shared similar NDVI and NGI values. The classification was conducted through the ontology defined before and lead to labelled objects. To use the NDVI and the NGI in the MRS with its integer band weights, the NDVI and the NGI were normalized to an 8 bit integer covering the values from 0 - 255. In case of the NDVI the minimum value -1 was normalized to 0, whereas the former maximum was normalized to 1. For a better readability and comparability of the later applied values, this normalization was removed again after the classification.

5.2.5 Agent-based classification for delineation of irrigated agriculture

Agent-based classification extends the previously shown concept of object-based image classification by a variable communication between the objects to improve the classification results (Hofmann et al., 2015). Through negotiation among each other, the agents choose the action that leads to the largest improvement of the classification. For example, the agents merge with promising neighboring objects or grow if the membership to the target class is increased (Fig. 37, Fig. 38). The main feature that the agent-based classification focusses on is the shape of the object. Therefore, the rectangular fit and the roundness of the object are included in the classification scheme. The rectangular fit follows a sigmoid function with the minimum at 0.0 and the maximum at 1.0. Meanwhile, the roundness behaves reversely to the rectangular fit with the maximum at 0.0 and the minimum at 1.0.

Each image object agent is embedded in the topology of neighbors and tries to maximize its class membership μ to one of the classes. To retain the topology is one of the major tasks of the software environment eCognition. After each agent alteration step, the topology of the

scene had to be reevaluated and the neighborhoods remapped. Without knowledge of neighborhood connections between the agents, the action maps for growing and merging cannot work and the relation of the agents is important to save computational time. In order to lower the computational demand only direct neighbors of the agent are asked for merger. Also the environment for growing and shrinking is limited to 10 pixels from each border to lower the number of possibilities per iteration. A wider area for growing and shrinking would demand tremendously more storage, as each pixel in each direction means 4-8 additional pixels.

One of the major advantages of agent-based image classification is the implication of the history and an adaption of behavioral rules towards environmental changes. The primer advantage will be analyzed in this study. This is of special interest for the delineation of irrigated agriculture because the cropping scheme may alter within a short period of time but the spectral input might not directly deliver enough data to identify the area as irrigated.

For example: In year 1 the spectral information gave clear hints that region A is irrigated. The grown crop is irrigated when drought occurs, so in the following, wetter year the irrigation is not used to save money. ABC should gather this information from the agents: What are the general conditions, what happened in the last period of measurement, etc? This information is used in the classification for year 2. Here, the region A is still classified as irrigated but with a lower class membership that will further decrease until valid information from the remote sensing data is integrated.

To use information about the past of the image object agents, a single slot spatio-temporal memory is included. Once that a scene is classified by the agent-based classification approach, the agents are stored for the next season and the next year as a shapefile comprising the shape and the assigned class. This shapefile is used instead of the segmentation to create a first set of image object agents.

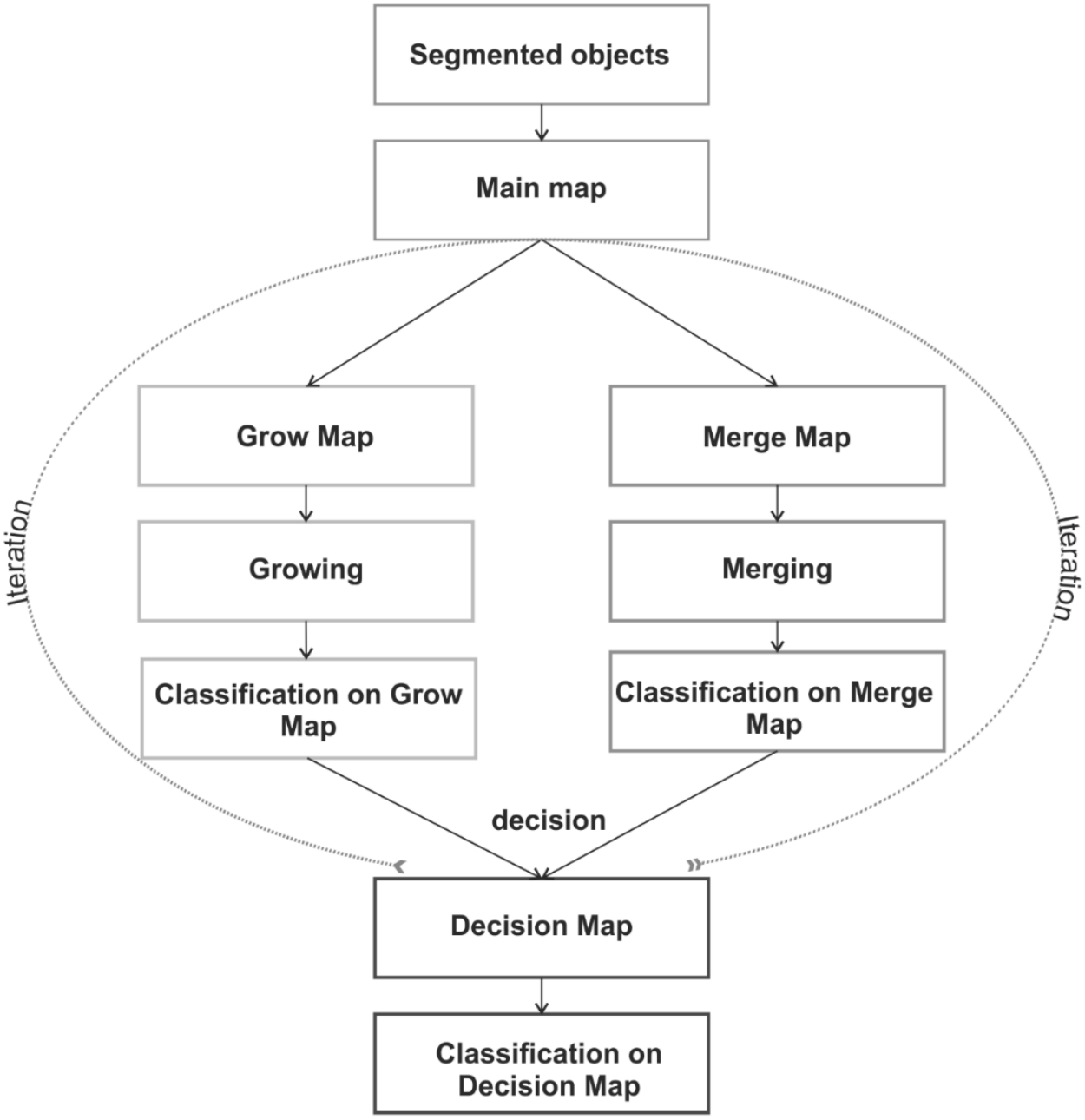


Fig. 37: Flowchart of object manipulation in agent-based image analysis. The agent-based object alteration is part of an iterative process on two separated maps covering both decision maps.

In the first time step, the image objects originating from the segmentation process are the basis for the agent-based classification. They form the main map upon which two separate action maps are created: the grow map and the merge map (Fig. 37). For each action the outcome in terms of improvement of maximum class membership are calculated. On the final decision map the best choice of all actions by the maximum improvement is performed. This alteration of image agents is repeated until no further improvement of maximum class belonging is measurable or the maximum number of five iterations is exceeded. The number

of iterations is limited to five to keep computational time low. A higher number of iterations may improve the results but each iteration doubles the amount of required storage because of the two individual action maps.

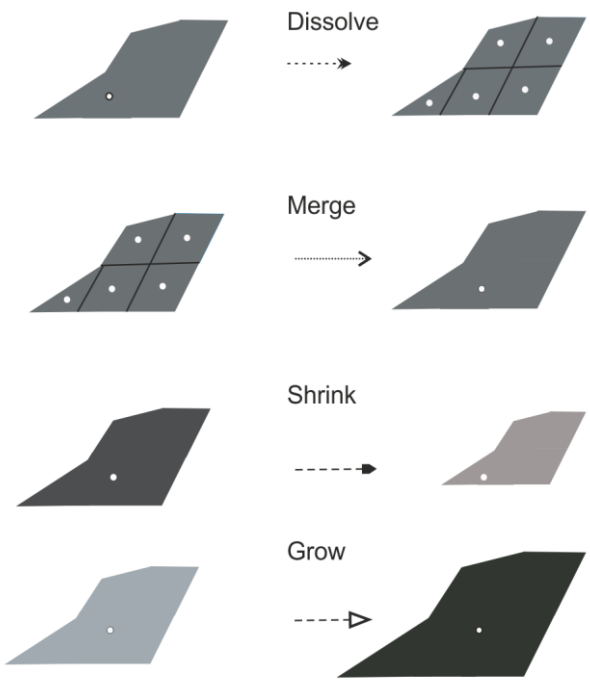


Fig. 38: Defined agent actions that allow alteration of structure to improve classification outcome and their respective effect on the changed structure and topology of objects.

5.2.6 Accuracy measure

To measure the performance of the approaches in comparison to real-world data the classification accuracy was calculated by a confusion matrix (Tso and Mather, 2009). This matrix comprises four components: true positive (TP), false positive (FP), true negative (TN) and false negative (FN). In this case study, *positive* means a classification as *Irr*, whereas *negative* stands for a classification of any other class. So, e.g. an object identified as *Irr* by the agents or in object-based classification that is also in reality an irrigated object is counted as a true positive. On the other hand, if an object of any other class is identified but the object is irrigated agriculture in reality, the result counts as a false negative. This strict formulation is softened by the extension that *Barren* objects are also counted as *Irr*, because the irrigation specific characteristics might be invisible after sowing or harvest when agricultural plots are barren. In the following sections, this way of accounting is referred as "soft" in contrast to the aforementioned "strict" formulation.

The accuracy was computed through the number of elements that are correctly identified as the target class, in this case *Irr*, and the total true number of objects known as *Irr*. In order to keep results comparable, all objects are dissolved into pixels, which means that the total number N of all elements remains the same for all approaches. For example, Acc_{TP} (5.4)

describes the accuracy of correctly identified irrigated objects: N_P, the number of objects correctly identified as *Irr*, divided by the true number of irrigated objects N_{KP}. This measure of accuracy is also computed for the cases TN, FP, and FN.

$$Acc_{TP} = \frac{N_P}{N_{KP}} \tag{5.4}$$

This measure of accuracy was chosen because it does not affect any of the classification approaches because only the respective results from the methods are compared. Moreover, this measure of accuracy is valid for all utilized approaches and is not limited to any of the methods. Furthermore, in this accuracy measure the different counting methods can easily be implemented, allowing the comparison of hard and soft similarity definitions.

5.2.7 Comparison between ABC and traditional image interpretation approaches

The main focus was laid on the results from season 2 that covers the period April – June because that is the season when most of the crops are planted and the harvest season has not yet started. All crops grown should be in the green phase of their respective phenological development. So, one can assume that most agricultural plots should be covered by crops. The visual comparison of the three approaches in Region A showed that already the pixel-based approach resulted in interpretable patterns that reminded of real-world patterns in the ground true data (Fig. 39 and Fig. 35).

The objects were homogenized by the agent-based classification where several objects were merged and the shape altered to increase the maximum class membership. So overall, the ABC results showed a more complete classification of the scene. In contrast to the pixel-based approach, the network of objects became visible in the ABC results. For this time step no history was available for the agents, hence no information from the genesis of the agent could be taken into account. In the pixel-based classification the class *Other Landuse* was not assigned, whereas in both object- and agent-based classification this class was used. Some of the circular objects were classified as *Barren* because of low NDVI values and very likely low plant activity when the scene was captured by the sensor. In comparison to the ground true information on irrigation only few were misclassified as *Agr* (non-irrigated agriculture). Apparently, the NDVI was high enough to suggest a classification as *Agr*, but the NGI was too low to alter the maximum class membership in favor of *Irr*.

As mentioned before, two different methods to determine the *Acc* the ‘soft’ and the ‘strict’ formulation were applied. The strict formulation counted only assigned irrigated agriculture as irrigated agriculture, while the soft formulation also counted assigned barren regions as irrigated agriculture. The strictly formulated accuracy achieved by the pixel-based approach for the detection of irrigated agriculture scored a higher *Acc* value than the soft formulation (Tab. 7). The object-based approach scored slightly weaker results for the detection of irrigation using the strict formulation but improved for the soft formulation. Interestingly, both

techniques using the object-based approach identified more irrigation pixels correctly (object-based strict 94.9%, agent-based strict 96.1%) than the purely pixel-based approach (64.6%) formulation (Tab. 7). The improvement of the iterations in the agent-based classification was only slight and reduced the accuracy by only 0.9%. Allowing class *Barren* to count as *Irr* reduced the number of classes that form the negative fraction of the accuracy and consequently increases the indicator for TN by 11.5% - 19.3% (Tab. 7). The agent- and object-based classifications showed their advantages in the completeness of the retrieved classes, but the accuracy only slightly increased. The overall accuracy could be improved by a more detailed classification scheme. Without a more detailed classification strategy, the advantages of the agent- and object-based classification were only noticeable in the completeness of the derived image objects because the applied data did not allow enough degrees of freedom in this classification scheme.

Tab. 7: Accuracy in Region A shown for each of the three approaches through an error matrix. The same fuzzy classification scheme is applied to all three different approaches. Results from the agent-based approach are shown after one iteration and five iterations. For each formulation of the accuracy rule, the results are presented individually. The object- and agent-based approach improve the correct identification of irrigated objects by 30 % – 32.4 %.

Classification methods	**Accuracy evaluation indices**			
	TP	**TN**	**FP**	**FN**
Pixel-based, strict	0.646	0.814	0.354	0.186
Pixel-based, soft	0.442	0.997	0.558	0.003
Object-based, strict	0.949	0.624	0.051	0.376
Object-based, soft	0.654	0.817	0.346	0.183
Agent-based, strict	0.961	0.622	0.039	0.378
Agent-based, soft	0.646	0.814	0.354	0.186
Agent-based 5x, strict	0.970	0.700	0.030	0.300
Agent-based 5x, soft	0.671	0.815	0.329	0.185

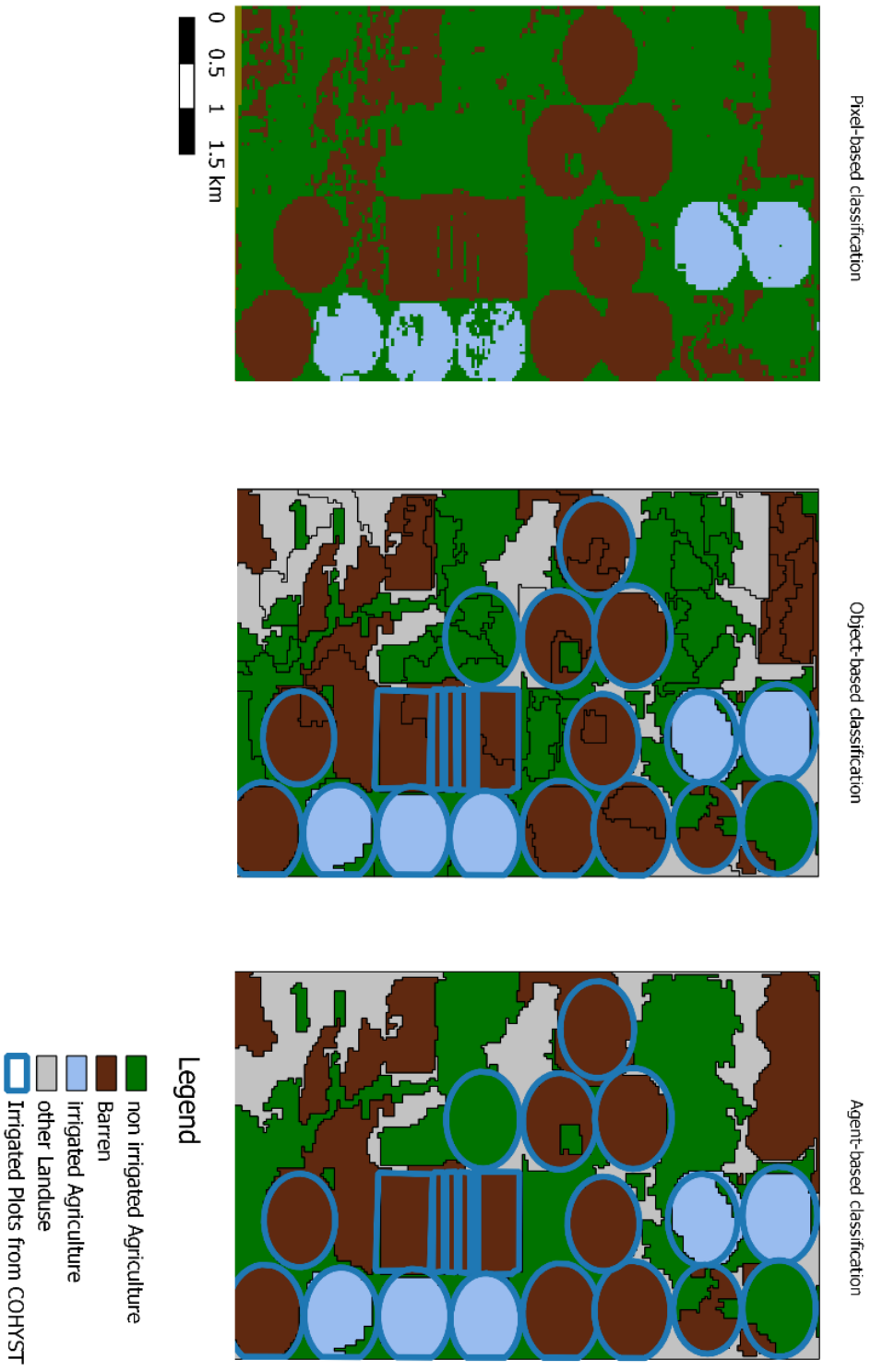


Fig. 39: Classification results from pixel-, object- and agent-based classification in Region A in Season 2 (May – June). The pixel-based approach already shows a pattern of pixels that is close to the ground-true information on irrigation. Meaningful objects are created by object- and agent-based classification, that both improve identification accuracy. Some irrigated areas are classified as Barren, which is a result from low plant activity when the scene was captured by the sensor.

In Region B (also in Season 2) the accuracy values for the identification of irrigated areas using the object- and agent-based approach were lower than in Region A (Tab. 8). Here, the object- and agent-based approaches were not able to identify any irrigated areas. Formulating the accuracy the soft method, the pixel-based approach assigned the class *Irr* in only 42.1% of all cases correctly, whereas the object-based approach had a true positive identification quota of 99.0%. This result had to be related to the TN value, showing that less than 60% of the other classes were assigned correctly. Agent-based classification on the other hand assigned the class *Irr* in 74.5% of all cases correctly with the soft formulation. The lack of identified irrigated areas in the object- and agent-based classification was a result from the low mean NGI values of the segmented objects. While single pixels showed NGI values high enough to justify a classification as *Irr*, the objects had a lower mean NGI and were thus classified as *Barren* or *Agr*.

Tab. 8: Accuracy in Region B (Season 2) shown for each of the three approaches through an error matrix. The same fuzzy classification scheme is applied to all three different approaches. The strict formulation of accuracy shows that neither object- nor agent-based classification are able to identify any irrigated area.

Classification methods	**Accuracy evaluation indices**			
	TP	**TN**	**FP**	**FN**
Pixel-based, strict	0.935	0.597	0.065	0.403
Pixel-based, soft	0.421	0.597	0.579	0.403
Object-based, strict	0.000	0.594	0.000	0.406
Object-based, soft	0.99	0.594	0.010	0.406
Agent-based, strict	0.000	0.594	0.000	0.406
Agent-based, soft	0.745	0.594	0.255	0.406
Agent-based 5x, strict	0.000	0.594	0.000	0.406
Agent-based 5x, soft	0.745	0.594	0.255	0.406

Due to the weak results of all approaches in the Region B in Season 2, the focus was shifted the next quarter of the year, Season 3 covering July – September, when harvest was just about to start for many crops, like winter wheat, potatoes and sunflowers grown here. The shift towards the next quarter shows slightly improved pixel-based accuracies, although the identification of *Irr* delivers worse results. The identification of non-irrigated areas improved using the strict formulation (Tab. 9). The results of the pixel-based approach using the soft formulation remained the same more or less. Both, the object- and the agent-based approach are finally able to successfully identify irrigated areas although the performance of both approaches stays behind those from Region A. Here, the identification rate of irrigation reaches

47.5% at maximum whereas the non-irrigated areas are classified in approximately 60% of the cases.

Tab. 9: Accuracy of all approaches applied in Region B in season 3 that covers July - September. Here, the object-based methods are able to identify irrigated agriculture with the strict formulation. The iterations in agent-based image classification show nearly no influence on the results.

Classification methods	Accuracy evaluation indices			
	TP	TN	FP	FN
Pixel-based, strict	0.624	0.737	0.376	0.263
Pixel-based, soft	0.439	0.713	0.561	0.287
Object-based, strict	0.412	0.574	0.588	0.426
Object-based, soft	0.482	0.613	0.518	0.387
Agent-based, strict	0.475	0.612	0.525	0.388
Agent-based, soft	0.418	0.579	0.582	0.421
Agent-based 5x, strict	0.475	0.612	0.525	0.388
Agent-based 5x, soft	0.418	0.579	0.582	0.421

The visual comparison revealed that both, the objects and the image agents, presented a more homogenous pattern in contrast to pixel-based classification. Again, the pixel-based approach did not assign the class *Other Landuse*, whereas the object- and agent-based approaches struggled with the classification of objects close to the border (Fig. 40). Most of the objects close to the fringes were classified as *Other Landuse* which was in this case a sign of erroneous input data. The misclassification in this scene became obvious which led to an underestimation of class *Irr* and an overestimation of class *Other Landuse*. Consequently, the object- and agent-based classifications only worked in the center of this specific subset. In contrast to results from Region A, the agent-based approach substantially improved the results from object-based classification and increased the irrigated areas towards a more realistic pattern.

Using information from the previous period in the following year, the results showed that the agent-based classification performed best in terms of the classification of irrigated and non-irrigated agriculture (only shown with the strict formulation, Tab. 10, Region A Season 2). While the accuracy of the correctly identified irrigated agriculture stays low at 53.7%, the identification of all other classes remains high, 61.5%. Pixel- and object-wise classification with a new segmentation showed higher accuracies for the identification of irrigated agriculture (91.2%, respectively 99%) but failed on the identification of all *Other Landuse* classes (45.2%, respectively 42.3%). So, the overall accuracy using the historic information

by agent-based classification allowed at least a classification accuracy of >50% which was better than in any other approach shown here.

Tab. 10: Accuracy of all approaches applied in Region B in season 2 incorporating the memory of the agent's classification of the antecedent year. Agent-based classification delivers the overall best accuracy in detection of irrigated agriculture.

Classification methods	**Accuracy evaluation indices**			
	TP	**TN**	**FP**	**FN**
Pixel-based,	0.912	0.452	0.088	0.548
Object-based	0.990	0.423	0.010	0.577
Agent-based	0.537	0.615	0.463	0.385

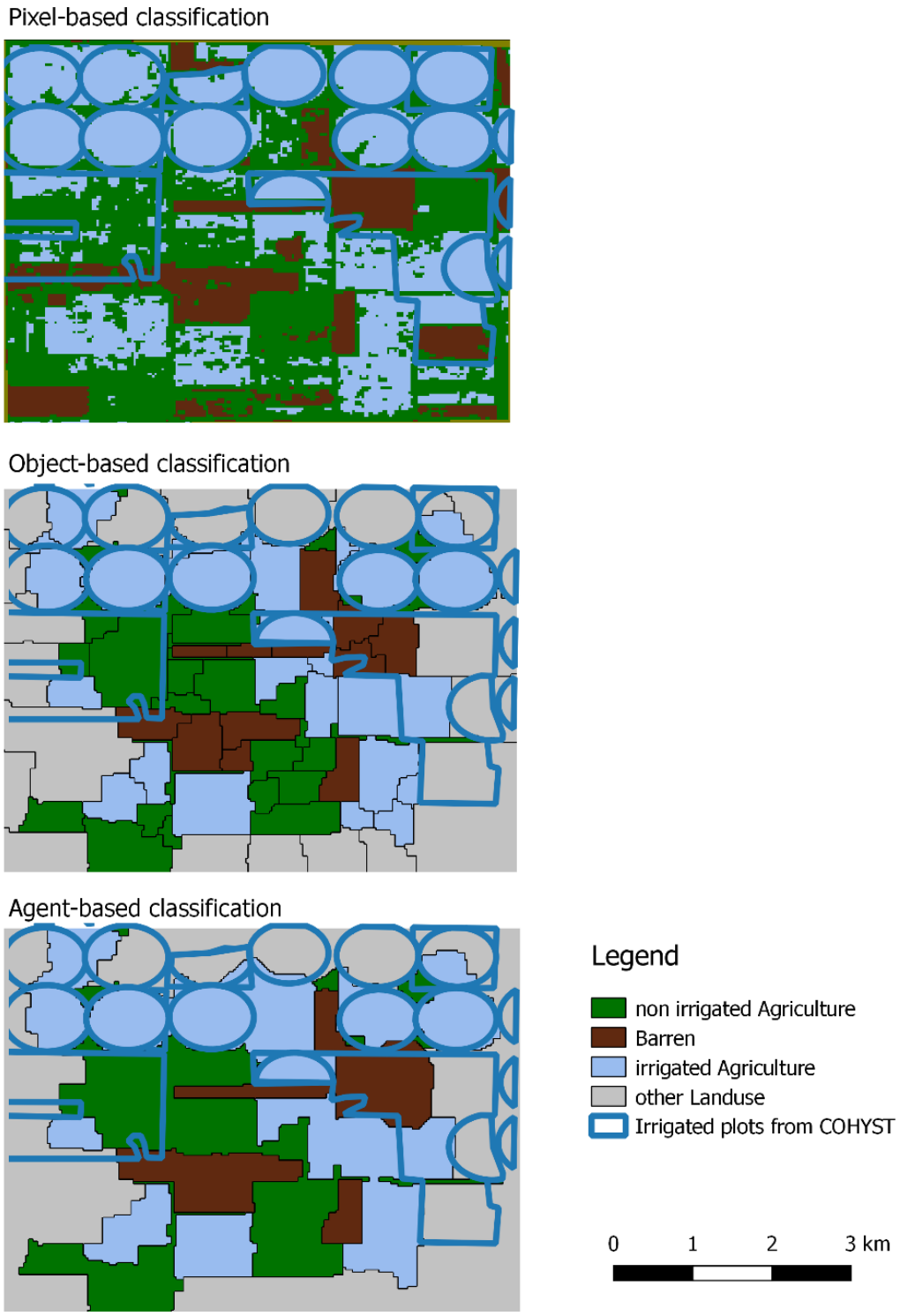


Fig. 40: Classification results from pixel-, object- and agent-based classification in Region B in Season 3 (July – September). In contrast to region A, objects and agents close to the border of the scene are erroneous whereas the pixel-based classification delivers a suitable and close to the ground-true data classification of the scene.

5.3 Concluding remarks on the application of ABC in hydrological remote sensing

Agent-based approaches add a dynamic component to models and data-driven analysis strategies. Through the manifoldness of independent software components and the emergent character of their interplay, AB methods offer a variable tool set for hydrological and water resource management problems. In this extension, the AB modelling concept was extended to interpret remote sensing scenes. With the focus placed on object and agent-based methods for remote sensing analyses with special regards to the delineation of irrigated agriculture. It was clear that both novel methods delivered similar or better results to the classic pixel-based approaches. Although pixel-based classification resulted in patterns also comparable to the observed real-world structures, only classification originating from object- and agent-based approaches created coherent and meaningful objects, which are required for spatial hydrological or water resource modelling. Generally, the concept of agent-based image classification ameliorates the promising results from purely object-based classification. Hence, one can propose agent-based image classification as a suitable tool for the pre-processing of remote sensing data in hydrological or water resource management models, due to the more meaningful representation of the real-world problem.

The classification framework is still in its initial state and might thus be a subject for further improvements. Specifically the knowledge on which a crop is grown on the object of interest might be a trigger for improved results. Moreover, one can see that the agent-based image classification ameliorates all object-based results, even with less than 5 iterations (Region B), which is a strong argument for pro agent-based classifications. Adaptive image agents that alter their rule set in classification and not just their shape would reduce the impact of expert knowledge in unsupervised learning (Hofmann et al., 2016; Hofmann, 2017). Meaning, the initial classification scheme could be adapted towards a more versatile and dynamic classification strategy, especially in data sparse regions.

Future research is planned on a refined database system that stores the image objects. By adding the history of the agent the robustness of the agent-based classification approach was strengthened. Moreover, the so modified classification process might reveal temporal changes, either seasonal (Season 2: *Barren*, Season 3: *Irr*) or long term (land use changes, cropping technique changes, etc.) more easily. Hence, the application of agent-based classification for the interpretation of spatio-temporal varying systems is of advantage. Here, machine-learning algorithms can be used to improve the rulesets and add even more information from the data to the classification process.

Together, the results are very promising, especially with regard to water resource modelling, where the knowledge about the spatial distribution of irrigated agriculture has a leverage on prediction capacity. Furthermore, hydro-metrological data sets should be included to further incorporate neighborhood relations between assumed irrigated plots and non-irrigated plots to withdraw information from the general availability of water and the NDVI and NGI values

of the plots. Furthermore, it is planned to fully incorporate the SEBS evapotranspiration dataset (Su and Su, 1988) to use the identification scheme by Boyaci et al. (2017) by incorporating the evaporative fraction (ETRF) in the classification scheme.

6 Adaptive agent-based modelling

As mentioned before, ABMs are used to model types of coupled systems with anthropogenic influence, originally in social sciences, later in ecology and finally in hydrology and water resource management (Ng et al., 2011; Mashhadi Ali et al., 2017; O'Connell, 2017; Gunkel, 2005; Lempert, 2002; Bithell and Brasington, 2009; Bouziotas et al., 2017; Mewes and Schumann, 2018b). ABMs are often used, despite that can be highly empiric and specialized to describe the situation in a defined context. Consequently, a transfer to a similar but different problem is limited due to the hard-formulated behavioral rules of the agents (Bruch and Atwell, 2015).

In the times of big data archives delivering information with high temporal and spatial resolution, this dynamic modelling approach might shed some light onto hidden patterns in the data and yet unknown relations between actions and causes that lead to observed patterns. The rules of interaction and behavior for the ABM have to be defined a-priori. This turned out to be a non-obvious task (see Sec. 4 and Sec. 5). Data-driven approaches for the creation of behavioral rules are rare; most of the applications depend on manual rule definition.

Although ML approaches are widely used for data analysis in terms of big data and pattern recognition, these learning algorithms are not popular to derive or to adapt behavioral strategies in ABMs. ML is the hypernym for programs that detect patterns in data and relate them in an algorithm with a known output. So, one could consider the ML as program that learns from a set of known data to predict a target value or to classify a set of unknown data.

Changing the behavioral rules of agents would represent a simple adaption strategy, which the agents might be able to adapt to a changing environment. Moreover, unforeseen relations can be revealed by the path of alteration presented during the course of the model. The alteration of existing rules is the primer step to advanced reinforcement learning that would also allow a creation of behavioral rules at run time of the model. Right now, there are no frameworks and only few concepts to incorporate ML for alteration of behavioral rules in agent-based models. As a primary step, evolutionary games can be used to reach a farmer specific optimization by behavioral adaption (Janssen, 2007; Lansing et al., 2017). Nevertheless, these evolutionally games are not suitable to change a model at runtime.

In the case study, a simple agent-based irrigation model with a data driven machine-learning adaption strategy is presented. The adaption strategy allows changing behavioral rules to maximize the global yield at the runtime of the model. As a reference model for irrigation Lansing's model of Balinese water temples is taken, that was originally set up as a cellular automata model with no communication between the isolated cells (Lansing, 2007). In the

agent-based model, the isolated cells are replaced by autonomous software units that represent the farmers and the temple. To reduce the amount of parametrization and to replicate the learning effect of the system, ML is used to identify classes of situations. This classification is later used in the adaptive agent-based model to identify the current system and to adapt the agents' behavioral rules to it. This does not mean that the rules are fundamentally changed in their fundaments, but that the thresholds are adapted that trigger a certain behavior. With synthetic runs, the general applicability of a ML approach as an adaption strategy is shown. Furthermore, the implications added by adaptive agent-based modelling and ML for adaption strategies are discussed.

6.1 Methods

6.1.1 Balinese water temple cult

The anthropologist Lansing (2007) found that the medieval Asian culture developed a water-based faith to distribute water, plan the sowing of rice and synchronize the date of harvest in order to improve the rice yield from the harvest on the island Bali. Regardless the limited possibilities of their time, an effective pest control mechanism was achieved by a collective harvest: harvest all rice before any disease could severely damage the plants in order to starve out all pests. The society developed a water-based religion organized by local temples to distribute the water among the farmers. Furthermore, to reduce the vulnerability towards pest and maximize to communal rice yield, the harvest was synchronized. As a result of this approach, the rice pests were killed by withdrawing the host from the parasite. Lansing (2007) showed by a cellular automata model that this strategy led to the highest possible rice yields that were possible by the contemporary technological methods without the need for chemical herbicides and fungicides. Nevertheless finding how this model was optimal the evolution of this strategy remains unclear as written historical documents from this time are not available.

6.1.2 Lansing's Balinese irrigation model

Lansing's model describes the Balinese irrigation system as a set of a temple with a varying number of connected farmers. The temple defines the quantity of water distributed to the farmers and synchronizes the harvest in order to minimize the loss in yield that is caused by pests. In the model, the global pest level is given by the variable D that increases exponentially. If D hits the maximum value of 1.0, all the harvest is lost to pest and the communal yield equals 0. Generally, the temple has the choice between two different rice varieties: a normal yielding rice plant and a high yielding variety. The high-yielding rice variety was emphasized to double the expected yield, whereas the normal rice plant yields 1/12 [t] per month. The yield grows linearly over the growing season for both varieties. If the required water can not be delivered, the plant suspends growth until the water demand can be covered.

The expected yields are purely synthetic values influenced by those values presented in Lansing (2007). The growing season of rice starts with the sowing in March and lasts until the harvest in September.

Next to the higher vulnerability towards pests, the high-yielding rice varieties require more water to mature in contrast to the normal yielding rice plants: 0.5 water units for normal yielding rice, 0.8 water units for high yielding varieties. To feed this water demand, the temple distributes the water from a storage to the farmers. The storage S_{temple} was defined as a minimum function with a maximum capacity set to 2 water units:

$$S_{temple} = \min(S_i, 2) \tag{6.1}$$

The water supply to the temple from a hypothetic river was modelled by a synthetic time series of runoff that offers a set of 11 different sequences of monthly runoff with similar characteristics. The water supply followed in its main core a sinus-shaped function with temporal peaks from flood and drought periods. The occurrence of those events is adjusted over the year to create different situations for the temple and the farmers. Because of local climate and the geographical location, the dry phases occur mostly before the monsoon season in June and July. Moreover, some variations in the magnitude led to shortages or overflow of water. In cases where the temple stores less water than the farmers demand, the amount of water is distributed a) fairly among all or b) with preference towards the high-yielding crop farmers (HY). The overall goal of temple and farmers is to maximize the global rice yield.

The development of *D* is given by a simple sigmoid function defined in the domain [0,1] where *t* equals the month in the year:

$$D(t) = \frac{1}{1 + c \cdot e^{-t}} \tag{6.2}$$

The vulnerability of the grown rice varieties towards the disease depends on the grown rice variety. High yielding varieties are exposed with a higher coefficient *c* (doubling the growth of pests) than normal yielding varieties where *c* equals 1. If *D* reaches the global maximum of 1.0 at any farm, the complete harvest is lost to pest. Next to a higher disease vulnerability, the water demand for high yielding crops is higher than for normal yielding crops as well. The disease threshold Thrs(D) triggers the communal harvest and is also set by the temple. If $D \geq Thrs(D)$ the temple releases a signal to the farmers to start the harvest. This threshold represents a certain risk for the community if the share of farmers with high-yielding rice varieties is too high and a high threshold is taken, the risk of a complete loss of higher than in a conservative strategy with a lower threshold. But, if the threshold is too low, the expectable yield decreases because harvest starts too early. So, in general the Thrs(D) is characterized by a tradeoff between yield surplus and the risk of loss.

In the model defined by Lansing, the temple plans once a year the cropping pattern in the upcoming growth season (i.e. the distribution of normal and high-yielding rice varieties) and

the starting point of sowing. By the cropping scheme, the temple can adapt to changing environmental conditions, i.e. if the temple priests expect dry or a wet season. At a certain point in the year, the sowing starts and the rice plants remain in the system until a collaborative harvest starts. Lansing showed with a cellular automata model that this method of pest control delivered the best results amongst a variety of randomly picked cropping schemes.

The cellular automata approach is the predecessor of agent-based modelling, resulting in a set of isolated grid cells. The grid cells do not communicate with each other, leaving the communication uni-directional in contrast to the bi-directional communication of agent-based modelling. Feedback between farmer and water temple could not be modelled. Hence, the new Balinese model was extended from the cellular automata model to a fully dynamic agent-based model.

6.1.3 Balinese Agent-based irrigation model

The Balinese Agent-based irrigation model (BAim) consist of a supervising agent that controls the boundary conditions like the maximum amount of water that can be stored in the temple, initializes all agents and controls the model time. It supervises the two different classes of agents included in the model: the farmer and the temple agent. The model itself is divided into two different layers that are separated but influence each other. The volumetric layer L1 where the water supply is added to the temples storage, the delivered water and the demand are calculated. The second layer is the communication layer L2 where the signal of harvest is shared among the farmers and the mean disease level in the system is communicated.

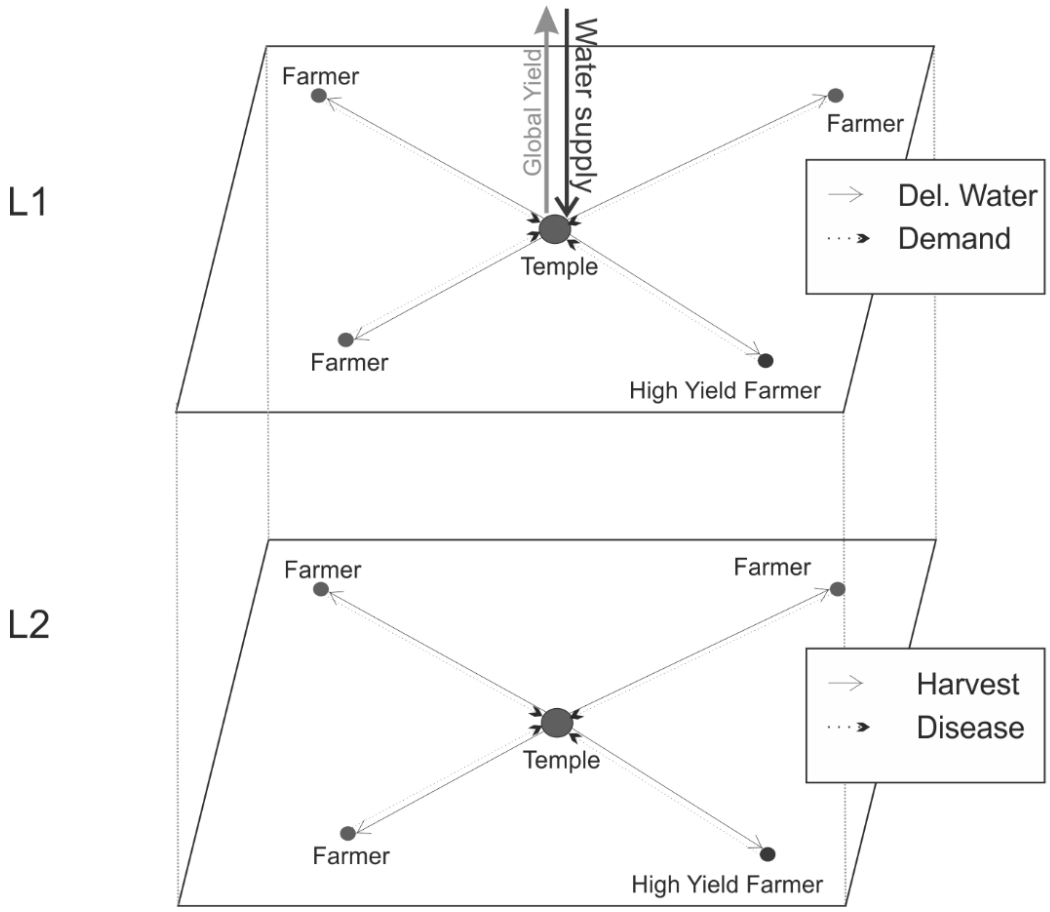


Fig. 41: Conceptual scheme of Balinese ABM with two different layers, the communication layer L2 and the volumetric stream layer L1.

In this example, four farmers belong to one temple, and build the collective decision making and cropping system. Conceptually, the temple is a decisional entity in this model. In this synthetic use-case only one temple is modelled, further temples require an additional communication layer between the temples to rotate and shift the cultivation pattern in terms of cropping strategies.

6.1.3.1 *Class description Temple*

As mentioned before, the temple sets the boundary conditions for the upcoming season. The temple sets the share of farmers who grow high-yielding rice plants, determines the Thrs(D) and supervises the level of *D*. The share of high-yield rice growing farmers (HYShare) and Thrs(D) are set with different strategies. In the non-learning ABM, this threshold is either set a-priori or changed randomly without taking the current environmental situation into account. After setting these parameters the temple becomes a passive component of the ABM and supervises the global level of *D*.

Once Thrs(D) is globally reached, the collective harvest starts following an active signal emitted from the temple. The signal to harvest is sent from the temple to all farmers as soon as the threshold for maximum plant disease is reached. All farmers obey and harvest the rice immediately regardless of phenological status of the rice plant.

A further value stored by the temple is the mean water storage per three month. The mean water storage is therefore calculated for the months January – March, April - June, July – September, October – December. A lower mean water storage in the temple signifies for a dryer year and thus for a lower expectable yield.

6.1.3.2 *Class description Farmer*

The farming agent awaits its signals to sow and to harvest from the temple. On the communication layer, the farmers queue their water demand. After they get the signal to harvest, the rice is harvested and added to the global yield. The current disease status of the farm is shared with the temple where the maximum disease of all farmers is taken as the global *D*. The farm has no opportunity to save water, everything has to be consumed directly. After the harvest, the farming agent resets its disease level to 0.0 until growing season starts again.

6.2 Adaptive Balinese Agent-based irrigation model

The disease threshold and the share of high-yield farmers remained a parameter for the non-learning ABM that had to be defined *a-priori*. Setting these parameters requires intensive manual analysis of the situation, a calibration function or a working ML scheme to judge the condition the agents are in. The non-learning trial and error approach considers the result from antecedent year as the result to beat in the actual year. By increasing or lowering the number of high-performance farmers and the threshold of acceptable disease the system tries to adapt to a changing environment without any kind of learning. If the temple recognizes the current period as a low flow cycle where a higher threshold may increase the yield, the temple decides to increase the threshold until the target class is reached.

ML has the advantage to ingest new information during the runtime of the model: once a set of training data is established, the algorithm can be further ameliorated by any new information coming into the system. In this case study, a k-means algorithm created a classification of environmental situations with four resulting clusters. In contrast to the aforementioned approaches likes SVM, ANN and ELM, ML required less input data for training and was thus not limited to a certain size of training data. Therefore, a ML approach has to be applied that delivers stable results with limited training data and less parameters to maintain some degrees of freedom while the model is so structured to be interpreted by the researcher. In contrast to the case studies presented in Sec. 2 & 3, the here presented problem is not a regression problem but a classification problem which is another argument for the chosen k-mean algorithm.

The k-means algorithm falls into the category of unsupervised learning where for each sample the closest cluster is assigned. The k-means algorithm requires a number of cluster seeds. As one could imagine four possible ways for the temple to actively adapt to the environmental conditions to maximize the communal yield, four clusters were searched (Tab. 11). These four reactions comprised the increase or decrease of Thrs(D) and the variation of HYshare.

So, the reactions were simplified to two main groups of actions: alter Thrs(D) and adapt HYShare.

Tab. 11: Possible reactions of temple to a changing environment

Rule Adapt 1	Rule Adapt 2	Rule Adapt 3	Rule Adapt 4
Increase Thrs(D)	Decrease Thrs(D)	Increase HYShare	Decrease HYShare

To reduce the number of dimensions of the k-means, the available parameters were reduced to 3, originating from the number of clusters minus the target variable. For the k-means the three parameters that shared the most information regarding the target variable by the MI were taken (Tab. 12). Therefore, MI between the target, the global rice yield and all available parameters derived from the model was calculated. By the MI values, one is able to identify the three most informative parameters. The calculation of mutual information follows the same notation as in Sec. 3.2 with Eq. (3.4) in a higher dimensional case.

Thrs(D) showed the highest MI with the global yield with an average of 1 bit of information. HYShare still accounted for 0.6 bit while the third parameter, WS0, had a mean of 0.11 bit of shared information with the global yield.

Tab. 12: Highest mutual information between possible parameters and global yield

Parameter	Threshold D	Share of HY farmers	Mean Water Storage 0
MI [bit]	1	0.6	0.11

Hence, the parameters Thrs(D), HYShare, the mean water storage in the temple before sowing (WS0) are clustered to find groups that control the expectable yield. The cluster centers are found by the smallest quadratic euclidian distance from an element x (that represents the current situation of WS0, HYShare and Thrs(D)) to the chosen center μ:

$$J = \sum_{i=1}^{k} \sum_{x_j \in S_i} \| x_j - \mu_i \|^2 \tag{6.3}$$

As training data, 1,000 synthetic runs were calculated with the simple alteration from the non-learning ABM scheme to create input data for the k-means. The resulting cluster centers were exported to the GAMA environment. In case that one of three variables of the k-means was not utilized, for example the share of HY farmers due to a fixed number of HY farmers, the dimension was left out reducing the complexity of the domain for the algorithm.

6.3 Results

Analyzing the synthetic runs, three parameters were used to find cluster centers to adapt the temples strategy to maximize the global rice yield: the dimensionless Thrs(D), HYShare and WS0 (Fig. 42). These three parameters were found to have the highest mutual information on the global yield (Tab. 12) and resulted in the following cluster centers (Tab. 13). The four

clusters showed a broad range of expectable yields from different behaviors and environmental situations. Class 1 (red dots in Fig. 42) represented a low flow situation with low disease threshold, whereas class 2 (green dots) was located in a low flow situation with a high threshold delivering a higher expectable global yield. Class 3 (blue dots) was similar to class 1 in wetter conditions with a mean water supply of 0.22 (Tab. 13). Class 4 (gold dots) was the cluster with the highest global yield and required a mean water supply of 0.26, a high percentage of HY farmers and a higher threshold for diseases. For visualization purposes, the predictor high yield farmers was left out of Fig. 42, as the initial run of the model did not consider this number of HY farmers.

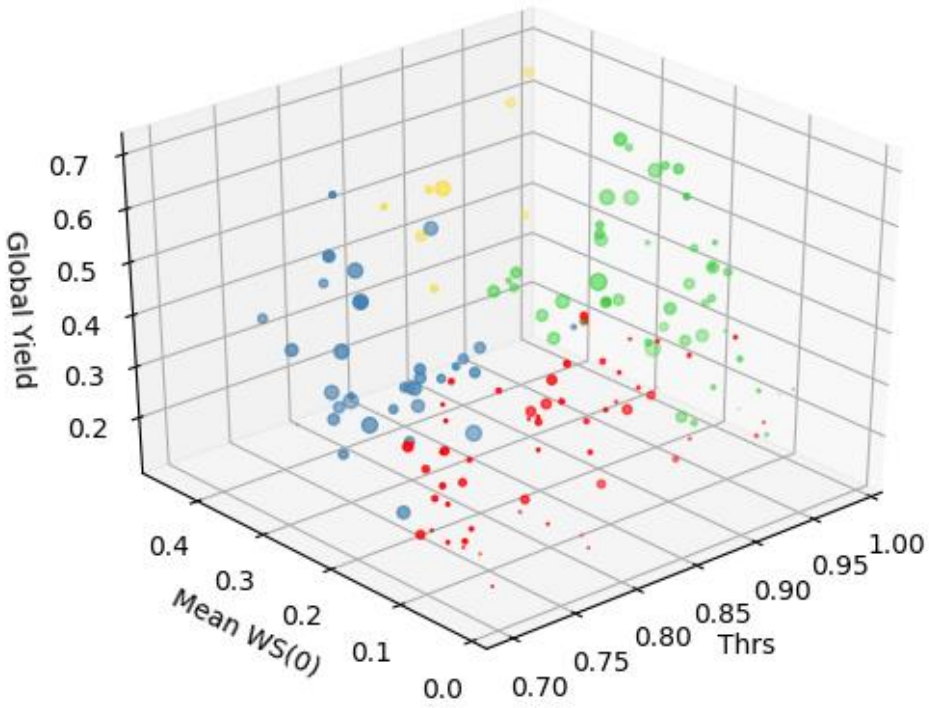


Fig. 42: Cluster Analysis with global yield [% of possible yield], threshold of disease [-] and the mean water supply [water units] in the sowing season. One can see the separation into four different clusters. Two of those clusters show higher expectable global yields: Class 2 (green) and class 4 (gold). Class 1 (red) represents a low flow situation with a low disease threshold and thus a risk-averse strategy. Class 3 (blue) is an intermediate class with relatively high water supply and an expected medium rice harvest.

Tab. 13: Cluster centers with the target variable, global yield, the predictors mean disease threshold, the share of high-yield farmers and the current water supply situation

Predictor	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Mean Thrs (D)	0.76	0.91	0.77	0.93
Mean Water Supply	0.10	0.42	0.22	0.26
Mean Global Yield	0.27	0.49	0.43	0.55
Share of High-Yield Farmers	0.25	0.75	0.5	0.75

To highlight out the adaption capacity of the newly developed adaptive ABM (aABM), the model with learning adaption strategy was compared to the model without learning adaption. Considering only the parameter with the highest mutual information in the adaption process, the disease threshold, the expectable global yield has risen from 0.34 t to 0.41 t over a period of 100 years. This means that the global yield increased by 32% using the single parameter ML driven adaption strategy. Meanwhile, the computational time increased from 43sec to 1min 12sec. With an expected optimal outcome (with roughly 50% loss due to diseases) of 1.7 t, only 20% - 25% of the expected yield could be generated. Here, the number of HY farmers was restricted to one of four. This means that the share of HY farmers is not considered in the k-means and remained fixed. The aABM delivered stable and slightly better results, because the variance stayed at the same low level for both strategies at about 0.014 t.

Taking all three parameters and the complete cluster analysis into account, the ML adaption strategy exceeded the non-learning strategy by far. The achievable maximum (due to the higher number of high-yield farmers, expecting a 50% availability of water) was increased to 2.3 t. Overall, the multi-adaption strategy reached 53% or a mean 1.217 t of rice per year. In 73% of the 1,000 synthetic cases the number of HY farmers changed in the analysis phase, the threshold of disease was altered in 67% of all cases. This means, the environmental conditions forced an adaption of behavior over time.

To show the influence of the timing, the decisional meeting was shifted from the end of the year to three months before sowing. Here, the mean global yield reached only 1.001 t which means that 43.5% of the possible yield was gathered. The computational time for the multiple adaption time increased to 1min 45sec. Here, the global yield was lower than for the decisional meeting after 12 months and computational time was slightly higher.

Evolving the cellular automata approach of the original Lansing model moved the model towards an ABM irrigation model. Here, social interactions of actors and processes that occur in knowledge-based processes like the adaption of traditional irrigation techniques and strategies to changing environmental conditions were connected. The results from this case

study revealed that the agent-based model profits from the learning layer. The original cellular automata model would not be able to alter behavioral patterns at the runtime from knowledge gained at model time because the communication between the model components are not existent.

6.4 Concluding remarks on adaptive agent-based models

Agent-based modelling with a data driven adaption strategy has shown its ability to model complex coupled systems. ML has proven its potential for learning strategies in adaptive agent-based models that often required massive abstraction or pre-analysis of appropriate behavioral rules and thresholds that trigger certain reaction patterns. By incorporating a ML enhanced adaption strategy, an optimization of agent behavior could be achieved. Furthermore, this adaption strategy was describable without additional rules or thresholds. Hence, ML adds intelligence to a prior limited system.

The lack of learning strategies was identified as a massive downside of agent-based models that requires the application of ML to reduce the manual workload (Abar et al., 2017; Kavak et al., 2018; Lamperti et al., 2018). Especially in the case of the Lansing Bali model ML helped to deduct the adaption strategy and could be used to increase the understanding of the farmers and priests, their aims and goals, as well as their developed adaption strategies (Janssen, 2007; Lansing et al., 2017). For future agent-based models covering socio-hydrological systems, this implementation of learning might bring improvements as soon as aspects like knowledge and history of agents has influence on the model structure and outcome, like for vulnerability in the context of flood (Du et al., 2017), decision making processes (Gunkel, 2005; Mashhadi Ali et al., 2017) and generally water resource management in a changing environment (O'Connell, 2017). Instead of an adaption of thresholds, GP should be applied to adapt the formulation of rules in the learning phase.

Again, scheduling impact has proven to be hard to determine as it has major influence on the outcome of the agent-based model. Like in the agent-based framework for the modelling of soilwater flow IPA (Mewes and Schumann, 2018b), the Balinese Agent-based Irrigation (BAIm) model scores different values in terms of optimal outcome with different scheduling methods. This reveals scheduling as a sensitive parameter that differs in its sensitivity from stiff equation based models. In future research focus has to be put on the scheduling of agents in these models to further understand the process of determination of the most appropriate scheduling method for the specific question.

Additionally, this learning strategy allows to model ignorance of agents. If an ignorant agent ignores the advice from the temple and acts differently a new set of training data is created that can again be used to fit a ML algorithm. This leads to competing strategies and could end in a competition of learning strategies without explicitly describing the differences in strategies. Hence, a ML adaption strategy can be the first step towards a fully adaptive agent-based model with behavioral rule set development by reinforcement learning.

7 Summary

In the beginning of the thesis the hypothesis was raised that with more data, novel analysis and modelling approaches are needed in hydrology and water resource management to cope with the rising availability of data and the demand for new insights from existing data. Especially with the dawn of the big data era, methods that create new information from available data are highly interesting for the hydrologic community. In the past decade, data and information proved to be the fuel of future scientific evolution in data-rich environments. The complementary approach stated by Shen (2018) requires the combination of big data archives, powerful ML approaches and the application of interrogative approaches that combine the virtues of the novel information theory based improvements to gain new insights from existing data.

The case study revealed that hydrologic pattern recognition by ML works fine for flood event separation. The application of ML approaches lowered the demand for developing dedicated algorithms to perform this task. Pre-processing the data, like cutting a continuous time series into chunks allows the application of algorithms like SVM for time series problems. The Multi-Layer-Perceptron did not deliver any suitable solutions for this problem. A detailed analysis of the applied structure of the neural network revealed, that the setup is not the reason for the comparatively weak results. Interestingly, a similar yet different approach, the ELM did deliver good results in the flood event separation. Together with the SVM, ELM seemed to be the most suitable approach for separating flood events from continuous time series of runoff. Regional disparities between the preferred catchment specific algorithms were explained by the *No-Free-Lunch-Theorem*: Once an algorithm has proven its capability, it will solve similar problems with a degraded performance. There was no one optimal choice of tools, especially in data-driven approaches. The preference of algorithm could be a data-driven related method for regionalization of catchments or hydrologic systems. The reason for the regional similarities and dissimilarities remained unknown and is of interest for future research. As the algorithms do not have a dedicated understanding of hydrological processes and catchment structures, one might assume that the information content of the runoff data in the investigated chunks is the key to understand the patterns of regionalization.

As most measures of information content are only available for discrete data, a measure for continuous data had to be found. Another limiting factor to describe the information content in the flood event separation problem was the missing true reference data. There was no measurable ground-true separation only referable proxy-methods. Hence, the success of the approach is always biased by the choice of the reference data. So, a different problem has to be found to develop the measure of information content in continuous data. This is the case

for tracer-based methods.

Tracer-based methods are a common tool to understand catchments and their subterraneous states. A major advantage of this proxy data is that we have a conceptual understanding of single natural or artificial tracers and their interpretability in terms of catchment analysis. Moreover, more than one measurable size can be investigated. As a true reference data set was available, one can express the information content of an explaining variable, the runoff, towards a target variable, the tracer concentration. Hereby, the MI, the shared information content between two tracer concentrations, was applied to express the relation between two variables on an information-scale. With the objective to predict tracer concentrations from snippets of the runoff dynamics, it could be shown that the information content of runoff is limited in terms of a differentiable description of the hydrological system. The information content of runoff was lower than the shared information between tracer signature and runoff requires. Consequently, the prediction capability of the ML-approaches lacked in performance. Nevertheless, a time series of tracer concentrations could be predicted. Again, the determination of the preferred algorithm was not obvious and remains problem-specific. In contrast to the flood event separation, the ANN provides results that are similar to the other approaches. Here, the complementary approach of modelling becomes obvious. While nitrate was more predictable with a specialized machine, in most cases sulfate scored better performances with a multivariate machine predicting both tracer concentrations at the same time. So, one can conclude that sulfate shares information with nitrate, which is a directed information between the target variables that becomes obvious through this study.

All of the presented ML approaches are black boxes with some constraints. The structure of the system linking in- and output is found by the algorithm without any further expert knowledge of the researcher. This black box character is well suited if the predominant rules that drive the system are unknown or its relations are not clearly describable. So, a white-box approach should be evaluated to reveal patterns in hydrological systems. ABMs are suitable white-box approaches to model patterns in complex system. As shown before, the case studies of ABMs comprise complex networks of agents in power grids, computer networks and many more. Traditionally used in social science and biology, ABMs can be used to describe natural systems as well. In physical hydrological models the rules of interaction are well documented. Nevertheless, the modelling techniques require simplification of these rules to build a model. Showing the movement of soil water through the matrix with an ABM revealed an interaction of the water entities that are described as balls with a certain mass and spatial extent with their specific environment. This simplification was caused by the massive computational demand of this approach, because any agent acts autonomously. Comparing the IPA model with a soil water model in the storage-based cmf model framework, one can see that the ABM is able to capture similar dynamics as the traditional model. Some of the subterraneous processes are hard to replicate in storage based lumped models without an allocation of a unique ID towards an entity of water. Technically, this modelling technique could further evolve by outsourcing certain parts of the calculation from the main

processor to the GPU. Now, computational time is the main limit next to the conceptual problems of scheduling (as presented by different types of scheduling the agents) and the fixed rule sets that control the agents' behavior.

Apart from modelling AB is also an appropriate method to improve classification strategies. ABC includes soft parameters, history and expert knowledge in the classification process. In the presented case study, the delineation of irrigated agriculture from remote sensing data was shown. The image object agents were formed by an ontology that was set up to describe the world of the agents by certain attributes. The image object agents also acted under a rule set to achieve a certain goal: To maximize their belonging to a certain class. This belonging is expressed by the fuzzy membership μ which remains a free parameter in the classification strategy. Thresholds and the defined set of actions of the agents limit the variability of the approach. Similar to ABM, ABC demand a lot of computational time. The inner structure of specialized action maps for each action and a collective iteration deciding which action delivers the best results allows to run ABC on high-performance parallel computers or dedicated GPUs. In terms of interpretability, ABC outscores the traditional pixel-wise classification strategies. Although the rule set of classification in the case study did not allow major improvements in accuracy of the approach, the completeness of the derived classes was higher than in any other approach.

8 Conclusions

Current hydrological modelling and data mining strategies often fail when applied to big data archives or to data sets where the structure of the data is unknown. Here, ML and AB methods show their advantages. As shown in this thesis, ML is able to reproduce patterns on hydrological data without the exact mathematical formulation of the laws leading to the reference data. The data-driven character of the ML application (Fig. 2) allows searching for patterns in the data. In contrast to the understanding that more data contains more information and is consequently fruitful for the modelling expert, the case studies revealed that the information content of the data is of higher importance than the total amount of available data. Hence, the investigation of the information content of the data has to be conducted to avoid over- or underfitting of data-driven approaches. Shannon's entropy model could be a key tool to understand the information content of the data and to judge the results from ML-based approaches as the tracer prediction case study has shown. The concept of entropy and MI is helpful to reveal information-rich data-sets and improves pure data-driven applications.

In contrast to ML techniques, AB approaches represent the knowledge-driven pole of the modelling spectrum, where the rules are expressed in a fixed manner (Fig. 2). Apart from the ML applications, the AB case studies have clearly shown the merits of these techniques in hydrology and water resource management. The incorporation of fuzzy information in spatio-temporal context of objects (like in the ABC example of Nebraska) increases the range of information to be derived from remote sensing data. Informal knowledge and fuzzy relations between objects in hydrological contexts can be modelled and information revealed. Without the ABC application, the historical development of irrigated agriculture wouldn't be detectable from spectral remote sensing data without additional information. The completeness of the ABC derived classification results exceeds the traditional approaches. Likewise, in ABM, the proposed modelling technique allows the modelling of spatially distributed complex hydrological situations. Here, the autonomous software units form patterns that are comparable to those measured in field or under laboratory conditions in a soil column, e.g. the accumulation of water at transition layers. Hence, AB methods allow to investigate natural and socio-hydrological systems in all its facets.

Nevertheless, the advantages of AB and ML are isolated without the combined approach of aABM presented in this thesis. The aABM is the keystone that links both modelling worlds (Fig. 3). The here presented aABM approach overcomes the traditional limits of both modelling paradigms and allows the adaption of the core fundaments of the ABM to the data. Data- and knowledge-driven modelling merge, to profit from both techniques and to gain

insight from the existing data. Furthermore, socio-hydrological models can describe the genesis of apparently naturally developed strategies. Again, the information content and the MI of the input data and the desired target are crucial for aABMs that converge within an acceptable time and deliver interpretable results.

In summary, applications in hydrology and water resource management profit from both: ML and AB. The combination of both approaches leads to a promising modelling technique that could result in un unique insights in existing data, a path to access big data archives and to get the most from the increasing amount of existing data from environmental sensors.

9 Outlook

Like ABM the fixed rules hinder the general application of ABC. This is the justification for the final part of the thesis. aABM combines both approaches and is a simple form of deep learning. By ML the ABM of the Balinese irrigation model BaIM can adapt their thresholds that trigger the harvest to recent environmental conditions. Although the ML component of the model is rather simple (represented by a *KNN*-algorithm fed by 1,000 synthetic runs), the additional layer helps to double the expected yield. Moreover, the complexity of the model itself is increased without losing further degrees of freedom. The higher the decisional power of the ML component was, the better the learning agents scored in terms of rice yield. By this combination of approaches, the advantages of both black-box data-driven adaption and white-box interpretability of internal states and structures are joint in the aABM. Although limitations from ABM and ABC are inherited, this highly dynamic and adaptive approach allows an ABM to calibrate without extensive Monte-Carlo runs or solution space investigations. Using the MI between the target variable and the multitude of possible explaining variables, the set with the highest shared amount of information can be determined. This helps to regularize the ML approach and to avoid overfitting. In addition to a reduced amount of calibration, aABMs allow to model scenarios in which the agents face unknown situations and alter their behavior. So, to model systems at the interface between social and natural systems, aABMs are an appropriate technique for scenario building and evaluation. Furthermore, aABMs allow the comparison of modelling strategies and to analyze how the autonomous agents agree on the optimal strategy. Now, strategy finding requires a high number of changes, because the agents can just adapt thresholds and are not able to alter rules at their core.

Changing rules and equations based on data-driven decisions is the first step towards deep learning and reinforcement learning. A deep learning rule adaption in agent-based computing lowers the problems of applying ABMs to real-world data because the work-intensive definition of rule sets and the calibration of threshold parameters diminishes. In deep learning the internal structure of the ML approach remains intact while the outmost layer is adapted to each new situation. Consequently, a found deep learning structure could be used not only for regionalization but also for a transfer of knowledge over regions and scientific problems. For the deep learning structure also techniques like genetic programming (GP) could be used to create interpretable internal structures based on the training data. In deep learning the information content is crucial, because of the transferable core of the ML. Again, complementary modelling could be a method to improve modelling capabilities by using

specialized approaches within the ML approach. The combination of data-driven ML approaches with agent-based techniques could improve the acceptance in the scientific community as well as the transferability of scientific research into application.

Reinforcement learning describes a deep learning variety that allows the algorithm to adapt its internal structure to new training data using a reward-punishment-approach. The algorithm follows a trial and error method to improve its structure. Improvements are rewarded whereas the declination of performance is punished. Consequently, patterns leading to a reward are more often repeated than those leading to a punishment. Here, the core of the ML approach remains intact and only the last layer is adapted to the problem. This could increase the performance and the adaptability of algorithms faced with unknown situation where thresholds might not hold and novel solutions that are non-existent in the training data should be tested and evaluated. Successful applications of these reinforcement learning approaches cover strategy detection in videogames, autonomous cars and medicine. A combination with ABM should substantially improve AB modelling techniques and lower the risk of over specialization like in the thresholds in ABC.

The second major improvement in aABM and ML would be the extensive application of GP. GP is the umbrella term for algorithms that try to find mathematical solutions to a given problem on the basis of symbolic programming. Symbolic programming on the other hand is a kind of computational linguistics deciphering the given problem into a tree of symbolic interactions. This tree can recursively be solved and used as guidebook for the specific case. GP is highly specialized on the given problem but in combination with ABM the derived GP trees might give insight into an unknown system. Traditionally, GP algorithms are applied either for problems with unknown relations between the components or if an approximation is well enough.

With aABM strategies in changing environments can be modelled and strategies of strategy-finding can be compared. For this approach each agent follows the goal to interpret the administrative rules for personal advantage. Additionally, the aABM allows to compare strategies at run-time. So, if a strategy is found to work for an individual although it disobeys the administrative rules, the individual can switch to that strategy. The aABM is an evolution of two very different modelling approaches combining the advantages of both methods. The aABM allows to model complex hydrological systems from physical questions to socio-hydrological questions and overcomes the limits of AB and ML.

10 References

Aaby, B.G., Perumalla, K.S., Seal, S.K., 2010. Efficient Simulation of Agent-based Models on multi-GPU and Multi-core Clusters, Proceedings of the 3rd International ICST Conference on Simulation Tools and Techniques. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), ICST, Brussels, Belgium, Belgium.

Abar, S., Theodoropoulos, G.K., Lemarinier, P., O'Hare, G.M.P., 2017. Agent Based Modelling and Simulation tools: A review of the state-of-art software. A review of the state-of-art software. Computer Science Review 24 (24), 13–33.

Ali, S., Islam, A., Mishra, P.K., Sikka, A.K., 2016. Green-Ampt approximations. A comprehensive analysis. Journal of Hydrology 535, 340–355.

Andrés, S., Arvor, D., Mougenot, I., Libourel, T., Durieux, L., 2017. Ontology-based classification of remote sensing images using spectral rules. Computers & Geosciences 102, 158–166.

Arvor, D., Durieux, L., Andrés, S., Laporte, M.-A., 2013. Advances in Geographic Object-Based Image Analysis with ontologies. A review of main contributions and limitations from a remote sensing perspective. Integration of Geodata and Imagery for Automated Refinement and Update of Spatial Databases 82, 125–137.

Baatz, M., Schäpe, A., 2000. Multiresolution segmentation—an optimization approach for high quality multi-scale image segmentation. In: J. Strobl, T. Blaschke, G. Griesebner (Editor), Angewandte Geographische Informations-Verarbeitung XII,. Wichmann Verlag, Karlsruhe, pp. 12–23.

Baraha, S., Biswal, P.K., 2017. Implementation of activation functions for ELM based classifiers, 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET). IEEE, pp. 1038–1042.

Belgiu, M., Hofer, B., Hofmann, P., 2014a. Coupling formalized knowledge bases with object-based image analysis. Remote Sensing Letters 5 (6), 530–538.

Belgiu, M., Tomljenovic, I., Lampoltshammer, T., Blaschke, T., Höfle, B., 2014b. Ontology-Based Classification of Building Types Detected from Airborne Laser Scanning Data. Remote Sensing 6 (2), 1347–1366.

Bengio, Y., 2009. Learning Deep Architectures for AI. Foundations and Trends® in Machine Learning 2 (1), 1–127.

Benz, U.C., Hofmann, P., Willhauck, G., Lingenfelder, I., Heynen, M., 2004. Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GIS-ready information. Integration of Geodata and Imagery for Automated Refinement and Update of Spatial Databases 58 (3–4), 239–258.

Berger, A.L., Della Pietra, V.J., Della Pietra, S.A., 1996. A maximum entropy approach to natural language processing. Computational linguistics 22 (1), 39–71.

Berhane, M.T., Lane, R.C., Wu, Q., Anenkhonov, A.O., Chepinoga, V.V., Autrey, C.B., Liu, H., 2018. Comparing Pixel- and Object-Based Approaches in Effectively Classifying Wetland-Dominated Landscapes. Remote Sensing 10 (1), 46.

Bithell, M., Brasington, J., 2009. Coupling agent-based models of subsistence farming with

individual-based forest models and dynamic models of water distribution. Environmental Modelling & Software 24 (2), 173–190.
Blaschke, T., Hay, G.J., Kelly, M., Lang, S., Hofmann, P., Addink, E., Queiroz Feitosa, R., van der Meer, F., van der Werff, H., van Coillie, F., Tiede, D., 2013. Geographic Object-Based Image Analysis – Towards a new paradigm. ISPRS Journal of Photogrammetry and Remote Sensing 87 (100), 180–191.
Blume, T., Zehe, E., Bronstert, A., 2007. Rainfall—runoff response, event-based runoff coefficients and hydrograph separation. Hydrological Sciences Journal 52 (5), 843–862.
Bonissone, P.P., 1997. Soft computing. The convergence of emerging reasoning technologies. Soft Computing 1 (1), 6–18.
Borna, K., Moore, A.B., Sirguey, P., 2014. Towards a vector agent modelling approach for remote sensing image classification. Journal of Spatial Science 59 (2), 283–296.
Boulaire, F., Utting, M., Drogemuller, R., 2015. Dynamic agent composition for large-scale agent-based models. Complex Adaptive Systems Modeling 3 (3), 1–23.
Bouziotas, D., Ertsen, M., Bouziotas, D., Ertsen, M., 2017. Socio-hydrology from the bottom up: A template for agent-based modeling in irrigation systems. Hydrol. Earth Syst. Sci. Discuss., 1–27.
Boyaci, D., Erdogan, M., Yildiz, F., 2017. Pixel-versus object-based classification of forest and agricultural areas from multiresolution satellite images. Turkish Journal of Electrical Engineering & Computer Sciences 25 (1), 365–375.
Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and Regression Trees. Taylor & Francis.
Brodley, C. (Ed.), 2004. Proceedings of the Twenty-First International Conference on Machine Learning. ACM.
Bruch, E., Atwell, J., 2015. Agent-Based Models In Empirical Social Research. Sociological methods & research 44 (2), 186–221.
Centarowicz, K., Paszyński, M., Pardo, D., Bosse, T., La Poutré, H., 2010. Agent-based computing, adaptive algorithms and bio computing. ICCS 2010 1 (1), 1951–1952.
Center for Advanced Land Management Information Techniques, 2005. COHYST. Platte River Cooperative Hydrology Study. https://calmit.unl.edu/2005. Accessed December 15, 2017.
Cernuzzi, L., Cossentino, M., Zambonelli, F., 2005. Process models for agent-based development. Agent-oriented Software Development. Engineering Applications of Artificial Intelligence 18 (2), 205–222.
Chaney, N.W., van Huijgevoort, M.H.J., Shevliakova, E., Malyshev, S., Milly, P.C.D., Gauthier, P.P.G., Sulman, B.N., 2018. Harnessing big data to rethink land heterogeneity in Earth system models. Hydrology and Earth System Sciences 22 (6), 3311–3330.
Chang, Y.-W., Hsieh, C.-J., Chang, K.-W., Ringgaard, M., Lin, C.-J., 2010. Training and testing low-degree polynomial data mappings via linear SVM. Journal of Machine Learning Research 11 (4), 1471–1490.
Chapman, T., 1999. A comparison of algorithms for stream flow recession and baseflow separation. Hydrol. Process. 13 (5), 701–714.
Chen, G., Weng, Q., Hay, G.J., He, Y., 2018a. Geographic object-based image analysis (GEOBIA). Emerging trends and future opportunities. GIScience & Remote Sensing 55 (2), 1–24.
Chen, J., Adams, B.J., 2006. Integration of artificial neural networks with conceptual models in rainfall-runoff modeling. Journal of Hydrology 318 (1-4), 232–249.
Chen, Y., Zhou, Y.'n., Ge, Y., An, R., Chen, Y., 2018b. Enhancing Land Cover Mapping through Integration of Pixel-Based and Object-Based Classifications from Remotely

Sensed Imagery. Remote Sensing (10), 77.
Coopersmith, E.J., Minsker, B.S., Sivapalan, M., 2014. Using similarity of soil texture and hydroclimate to enhance soil moisture estimation. Hydrol. Earth Syst. Sci. 18 (8), 3095–3107.
Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine Learning 20 (3), 273–297.
Corzo, G., Solomantine, D., 2007. Baseflow separation techniques for modular artificial neural network modelling in flow forecasting. Hydrological Sciences Journal 52 (3), 491–507.
Crooks, A., Castle, C., Batty, M., 2008. Key challenges in agent-based modelling for geo-spatial simulation. GeoComputation: Modeling with spatial agents 32 (6), 417–430.
DBG Arbeitsgruppe Kennwerte des Bodengefüges, 2009. Bodenphysikalische Kennwerte und Berechnungsverfahren für die Praxis / Fachgebiete Bodenkunde, Standortkunde und Bodenschutz, Inst. für Ökologie. TU Berlin, Selbstverl., Berlin.
Debats, S.R., Luo, D., Estes, L.D., Fuchs, T.J., Caylor, K.K., 2016. A generalized computer vision approach to mapping crop fields in heterogeneous agricultural landscapes. Remote Sensing of Environment 179, 210–221.
Domingos, P., 2012. A few useful things to know about machine learning. Communications of the ACM 55 (10), 78–87.
Drăguţ, L., Tiede, D., Levick, S.R., 2010. ESP. A tool to estimate scale parameter for multiresolution image segmentation of remotely sensed data. International Journal of Geographical Information Science 24 (6), 859–871.
Du, E., Cai, X., Sun, Z., Minsker, B., 2017. Exploring the Role of Social Media and Individual Behaviors in Flood Evacuation Processes. An Agent-Based Modeling Approach. Water Resour. Res. 53 (11), 9164–9180.
Eaufrance, 2018a. ADES. Portail nationale d'Accès aux Données sur les Eaux Souterraines. http://www.ades.eaufrance.fr/. Accessed June 18, 2018.
Eaufrance, 2018b. Banque Hydro. http://hydro.eaufrance.fr/. Accessed June 18, 2018.
eCognition Developer, T., 2014. 9.0 User Guide. Trimble Germany GmbH: Munich, Germany.
FAO, 2012. Coping with water scarcity. An action framework for agriculture and food security. Food and Agriculture Organization of the United Nations, Rome.
Fernando, T.M.K.G., Maier, H., Dandy, G., Maier, H.R., Dandy, G.C., 2009. Selection of input variables for data driven models. An average shifted histogram partial mutual information estimator approach. Journal of Hydrology 367 (3-4), 165–176.
Folino, G., Mendicino, G., Senatore, A., Spezzano, G., Straface, S., 2006. A model based on cellular automata for the parallel simulation of 3D unsaturated flow. Parallel Computing 32 (5–6), 357–376.
Furey, P.R., Gupta, V.K., 2001. A physically based filter for separating base flow from streamflow time series. Water Resour. Res. 37 (11), 2709–2722.
Garvelmann, J., Warscher, M., Leonhardt, G., Franz, H., Lotz, A., Kunstmann, H., 2017. Quantification and characterization of the dynamics of spring- and stream water systems in the Berchtesgaden Alps with a long-term stable isotope dataset // Quantification and characterization of the dynamics of spring and stream water systems in the Berchtesgaden Alps with a long-term stable isotope dataset. Environmental Earth Sciences 76 (22), 766.
Gong, W., Yang, D., Gupta, H.V., Nearing, G., 2014. Estimating information entropy for hydrological data. One-dimensional case. Water Resour. Res. 50 (6), 5003–5018.
Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep learning. MIT Press.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google Earth Engine. Planetary-scale geospatial analysis for everyone. Remote Sensing of Environment (202), 18–27.

Grashey-Jansen, S., Timpf, S., 2010. Soil Hydrology of Irrigated Orchards and Agent-Based Simulation of a Soil Dependent Precision Irrigation System. Advanced Science Letters (3), 259–272.

Grimm, V., Revilla, E., Berger, U., Jeltsch, F., Mooij, W.M., Railsback, S.F., Thulke, H.-H., Weiner, J., Wiegand, T., DeAngelis, D.L., 2005. Pattern-oriented modeling of agent-based complex systems: lessons from ecology. Science (New York, N.Y.) 310 (5750), 987–991.

Gunkel, A., 2005. The application of multi-agent systems for water resources research– Possibilities and limits. Master Thesis, Albert-Ludwigs-Universität Freiburg, Freiburg, Germany.

Hall, F.R., 1968. Base-Flow Recessions—A Review. Water Resour. Res. 4 (5), 973–983.

Hammond, M., Han, D., 2006. Recession curve estimation for storm event separations. Journal of Hydrology 330 (3), 573–585.

Han, J., Kamber, M., 2010. Data mining. Concepts and techniques. Elsevier, Amsterdam.

Hartmann, A., Barberá, J.A., Andreo, B., 2017. On the value of water quality data and informative flow states in karst modelling. Hydrol. Earth Syst. Sci. 21 (12), 5971–5985.

Hartmann, A., Kobler, J., Kralik, M., Dirnböck, T., Humer, F., Weiler, M., 2016. Model-aided quantification of dissolved carbon and nitrogen release after windthrow disturbance in an Austrian karst system. Biogeosciences (1), 159–174.

Hay, G.J., Castilla, G., Wulder, M.A., Ruiz, J.R., 2005. An automated object-based approach for the multiscale image segmentation of forest scenes. International Journal of Applied Earth Observation and Geoinformation (4), 339–359.

Haykin, S., 1999. Neural networks. A comprehensive foundation. Prentice Hall, Upper Saddle River.

He, Z., Wen, X., Liu, H., Du, J., 2014. A comparative study of artificial neural network, adaptive neuro fuzzy inference system and support vector machine for forecasting river flow in the semiarid mountain region. Journal of Hydrology 509, 379–386.

Hilbert, M., López, P., 2011. The world's technological capacity to store, communicate, and compute information. Science (332), 60–65.

Ho, Y.C., Pepyne, D.L., 2002. Simple Explanation of the No-Free-Lunch Theorem and Its Implications. Journal of Optimization Theory and Applications 115 (3), 549–570.

Hofmann, P., 2017. A Fuzzy Belief-Desire-Intention Model for Agent-Based Image Analysis, Modern Fuzzy Control Systems and Its Applications. InTech.

Hofmann, P., Andrejchenko, V., Lettmayer, P., Schmitzberger, M., Gruber, M., Ozan, I., Belgiu, M., Graf, R., Lampoltshammer, T.J., Wegenkittl, S., 2016. Agent based image analysis (ABIA)-preliminary research results from an implemented framework.

Hofmann, P., Lettmayer, P., Blaschke, T., Belgiu, M., Wegenkittl, S., Graf, R., Lampoltshammer, T.J., Andrejchenko, V., 2015. Towards a framework for agent-based image analysis of remote-sensing data. International Journal of Image and Data Fusion 6 (2), 115–137.

Hoogeveen, J., Faurès, J.-M., Peiser, L., Burke, J., van de Giesen, N., 2015. GlobWat – a global water balance model to assess water use in irrigated agriculture. Hydrology and Earth System Sciences 19 (9), 3829–3844.

Huang, G.-B., Zhu, Q.-Y., Siew, C.-K. (Eds.), 2004. Extreme learning machine: a new learning scheme of feedforward neural networks.

Ingrassia, S., Morlini, I., 2005. Neural Network Modeling for Small Datasets. Technometrics 47 (3), 297–311.

International Conference on Information Communication and Embedded Systems (ICICES) (Ed.), 2016.

Janssen, M.A., 2007. Coordination in irrigation systems. An analysis of the Lansing–Kremer model of Bali. Agricultural Systems 93 (1-3), 170–190.

Jarvis, A., Reuter, H., Nelson, A., Guevara, E., 2008. Hole-filled seamless SRTM data V4. http://srtm.csi.cgiar.org. Accessed October 31, 2018.

Jennings, N.R., 2000. On agent-based software engineering. Artificial Intelligence 117 (2), 277–296.

Johnson, D.M., Mueller, R., 2010. The 2009 Cropland Data Layer. PE&RS, Photogrammetric Engineering & Remote Sensing 76 (11), 1201–1205.

Kavak, H., Padilla, J.J., Lynch, C.J., Diallo, S.Y., 2018. Big data, agents, and machine learning. Towards a data-driven agent-based modeling approach, Proceedings of the Annual Simulation Symposium. Society for Computer Simulation International, Baltimore, Maryland, pp. 1–12.

Kelleher, J.D., Mac Namee, B., D'Arcy, A., 2015. Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. MIT Press.

Kirschniak, C., 2018. Impact of Artificial Intelligence in Germany. PWC Press.

Klaus, J., McDonnell, J.J., 2013. Hydrograph separation using stable isotopes. Review and evaluation. Journal of Hydrology 505, 47–64.

Kofler, K., Davis, G., Gesing, S., 2014. SAMPO: an agent-based mosquito point model in OpenCL, Proceedings of the 2014 Symposium on Agent Directed Simulation. Society for Computer Simulation International, Tampa, Florida, pp. 1–10.

Kraft, P., Vaché, K.B., Frede, H.-G., Breuer, L., 2011. CMF: A Hydrological Programming Language Extension For Integrated Catchment Models. Environmental Modelling & Software 26 (6), 828–830.

Lamperti, F., Roventini, A., Sani, A., 2018. Agent-based model calibration using machine learning surrogates. Journal of Economic Dynamics and Control 90, 366–389.

Landesamt für Umwelt Bayern, 2017. Naturräumliche Haupteinheiten. https://www.lfu.bayern.de/natur/naturraeume/index.htm. Accessed February 27, 2018.

Lang, S., Hagenlocher, M., Pernkopf, L., Kienberger, S., Tiede, D., 2014. Object-based multi-indicator representation of complex spatial phenomena. South-Eastern Eur J Earth Observation Geomatics 3, 625–628.

Lansing, J.S., 2007. Priests and programmers. Technologies of power in the engineered landscape of Bali. Princeton University Press, Princeton, N.J.

Lansing, J.S., Thurner, S., Chung, N.N., Coudurier-Curveur, A., Karakaş, Ç., Fesenmyer, K.A., Chew, L.Y., 2017. Adaptive self-organization of Bali's ancient rice terraces. Proceedings of the National Academy of Sciences 114 (25), 6504–6509. http://www.pnas.org/content/pnas/114/25/6504.full.pdf.

Lee, E.S., Krothe, N.C., 2001. A four-component mixing model for water in a karst terrain in south-central Indiana, USA. Using solute concentration and stable isotopes as tracers. Chemical Geology (179), 129–143.

Lempert, R., 2002. Agent-based modeling as organizational and public policy simulators. Proceedings of the National Academy of Sciences of the United States of America (99), 7195–7196.

Lillesand, T., Kiefer, R.W., Chipman, J., 2014. Remote sensing and image interpretation. John Wiley & Sons.

Lindström, G., Johansson, B., Persson, M., Gardelin, M., Bergström, S., 1997. Development and test of the distributed HBV-96 hydrological model. Journal of Hydrology 201 (1-4), 272–288.

Liu, X., Gao, C., Li, P., 2012. A comparative analysis of support vector machines and extreme learning machines. Neural Networks 33, 58–66.

Macal, C.M., North, M.J., 2010. Tutorial on agent-based modelling and simulation. Journal of Simulation 4 (3), 151–162.

Mahler, B.J., Garner, B.D., Mahler, B.J., Garner, B.D., 2009. Using nitrate to quantify quick flow in a karst aquifer. Ground Water 47 (3), 350–360.

Maidment, D.R. (Ed.), 1993. Handbook of hydrology. McGraw-Hill, New York, NY.

Mashhadi Ali, A., Shafiee, M.E., Berglund, E.Z., 2017. Agent-based modeling to simulate the dynamics of urban water supply. Climate, population growth, and water shortages. Sustainable Cities and Society 28, 420–434.

Mauro, A. de, Greco, M., Grimaldi, M., 2016. A formal definition of Big Data based on its essential features. Library Review 65 (3), 122–135.

Mei, Y., Anagnostou, E.N., 2015. A hydrograph separation method based on information from rainfall and runoff records. Journal of Hydrology 523, 636–649.

Meier, J., Zabel, F., Mauser, W., 2017. Extending global irrigation maps – going beyond statistics. Hydrol. Earth Syst. Sci. Discuss. 2017, 1–16.

Mewes, B., Oppel, H., Hartmann, A., 2018. Information based machine-learning for tracer signature prediction in karstic environments. Water Resour. Res., Currently under review.

Mewes, B., Schumann, A., 2018a. An agent-based add-on for object-based image analysis for the delineation of irrigated agriculture from remote sensing data. International Journal of Remote Sensing, Accepted for publication.

Mewes, B., Schumann, A.H., 2018b. IPA (V1). A framework for agent-based modelling of soil water movement. Geosci. Model Dev. Discuss. 2018 (11), 2175–2187.

Mitchell, T.M., 2010. Machine learning. McGraw-Hill, New York, NY.

Mohammed, A.F., Humbe, V.T., Chowhan, S., 2016. A review of big data environment and its related technologies. In: International Conference on Information Communication and Embedded Systems (ICICES) (Editor).

Mudarra, M., Andreo, B., 2011. Relative importance of the saturated and the unsaturated zones in the hydrogeological functioning of karst aquifers. The case of Alta Cadena (Southern Spain). Journal of Hydrology 397 (3-4), 263–280.

Ng, T.L., Eheart, J.W., Cai, X., Braden, J.B., 2011. An agent-based model of farmer decision-making and water quality impacts at the watershed scale under markets for carbon allowances and a second-generation biofuel crop. Water Resour. Res. 47 (9), 3.

Nied, M., Pardowitz, T., Nissen, K., Ulbrich, U., Hundecha, Y., Merz, B., 2014. On the relationship between hydro-meteorological patterns and flood types. Journal of Hydrology 519, 3249–3262.

North, M.J., 2014. A theoretical formalism for analyzing agent-based models. Complex Adaptive Systems Modeling (2), 3.

O'Connell, E., 2017. Towards Adaptation of Water Resource Systems to Climatic and Socio-Economic Change. Water Resources Management 31 (10), 2965–2984.

Ozdarici-Ok, A., Ok, A., Schindler, K., 2015. Mapping of Agricultural Crops from Single High-Resolution Multispectral Images. Data-Driven Smoothing vs. Parcel-Based Smoothing. Remote Sensing (7), 5611–5638.

Parasuraman, K., Elshorbagy, A., Si, B.C., 2007. Estimating Saturated Hydraulic Conductivity Using Genetic Programming. Soil science society of America journal 71 (6),

1676–1684.
Parsons, J.A., Fonstad, M.A., 2007. A cellular automata model of surface water flow. Hydrological Processes 21 (16), 2189–2195.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825-2830.
Peña, M.J., Gutiérrez, A.P., Hervás-Martínez, C., Six, J., Plant, E.R., López-Granados, F., Peña, J., Gutiérrez, P., Plant, R., 2014. Object-Based Image Classification of Summer Crops with Machine Learning Methods. Remote Sensing 6, 5020–5041.
Phillips, S.J., Dudík, M., Schapire, R.E., 2004. A maximum entropy approach to species distribution modeling. In: C. Brodley (Editor), Proceedings of the Twenty-First International Conference on Machine Learning. ACM, pp. 655–662.
Piotrowski, A., Wallis, S.G., Napiórkowski, J.J., Rowiński, P.M., 2007. Evaluation of 1-D tracer concentration profile in a small river by means of Multi-Layer Perceptron Neural Networks. Hydrol. Earth Syst. Sci. 11 (6), 1883–1896.
Pun, M., Mutiibwa, D., Li, R., 2017. Land Use Classification. A Surface Energy Balance and Vegetation Index Application to Map and Monitor Irrigated Lands. Remote Sensing 9 (12), 1256.
Quinlan, J.R., 1986. Induction of decision trees. Machine Learning 1 (1), 81–106.
Raghavendra, N.S., Deka, P.C., 2014. Support vector machine applications in the field of hydrology: A review. Applied Soft Computing 19, 372–386.
Rakotoarisoa, M.M., Fleurant, C., Amiot, A., Ballouche, A., Communal, P.Y., Jadas-Hécart, A., La Jeunesse, I., Landry, D., Razakamanana, T., 2014. Agents-based modelling for hydrological surface processes on a small watershed (Layon, France). International Journal of Geomatics and Spatial Analysis / Revue Internationale de Géomatique 24 (3), 307–333.
Reaney, S.M., 2008. The use of agent based modelling techniques in hydrology. Determining the spatial and temporal origin of channel flow in semi-arid catchments. Earth Surf. Process. Landforms 33 (2), 317–327.
Rowley, J., 2007. The wisdom hierarchy. Representations of the DIKW hierarchy. Journal of information science 33 (2), 163–180.
S. Rybacki, J. Himmelspach, A. M. Uhrmacher, 2009. Experiments with Single Core, Multi-core, and GPU Based Computation of Cellular Automata, Advances in System Simulation, 2009. SIMUL ’09, pp. 62–67.
Salmon, J.M., Friedl, M.A., Frolking, S., Wisser, D., Douglas, E.M., 2015. Global rain-fed, irrigated, and paddy croplands: A new high resolution map derived from remote sensing, crop inventories and climate data. International Journal of Applied Earth Observation and Geoinformation 38, 321–334.
Samuel, A.L., 1959. Some studies in machine learning using the game of checkers. IBM J. Res. Dev. 3 (3), 210–229.
Schumacher, C., Vose, M.D., Whitley, L.D. (Eds.), 2001. The no free lunch and problem description length. Morgan Kaufmann Publishers Inc.
See, L., Solomantine, D., Abrahart, R., Toth, E., 2007. Hydroinformatics. Computational intelligence and technological developments in water science applications—Editorial. Hydrological Sciences Journal 52 (3), 391–396.
Servat, D., 2000. Modélisation de dynamiques de flux par agents. Application aux processus de ruisellement, infiltration et érosion. Dissertation, Université Pierre et Marie Curie, Paris.

Shannon, C.E., 1948. A mathematical theory of communication. The Bell System Technical Journal 27 (3), 379–423.
Shao, Q., Weatherley, D., Huang, L., Baumgartl, T., 2015. RunCA: A cellular automata model for simulating surface runoff at different scales. Journal of Hydrology 529 (3), 816–829.
Sharma, A., 2000. Seasonal to interannual rainfall probabilistic forecasts for improved water supply management. Part 1—A strategy for system predictor identification. Journal of Hydrology 239 (1), 232–239.
Shen, C., Laloy, E., Albert, A., Chang, F.-J., Elshorbagy, A., Ganguly, S., Hsu, K.-L., Kifer, D., Fang, Z., Fang, K., Li, D., Li, X., Tsai, W.-P., 2018. HESS Opinions. Deep learning as a promising avenue toward knowledge discovery in water sciences. Hydrology and Earth System Sciences Discussions 2018, 1–21.
Shortridge, J.E., Guikema, S.D., Zaitchik, B.F., 2016. Machine learning methods for empirical streamflow simulation: a comparison of model accuracy, interpretability, and uncertainty in seasonal watersheds. Hydrol. Earth Syst. Sci. 20 (7), 2611–2628.
Siebert, S., Burke, J., Faures, J.M., Frenken, K., Hoogeveen, J., Döll, P., Portmann, F.T., 2010. Groundwater use for irrigation – a global inventory. Hydrol. Earth Syst. Sci. 14 (10), 1863–1880.
Simunek, J., van Genuchten, M.T., Sejna, M., 2005. The HYDRUS-1D software package for simulating the one-dimensional movement of water, heat, and multiple solutes in variably-saturated media. University of California-Riverside Research Reports 3, 1–240.
Sivanandam, S.N., 2011. Principles of soft computing. Wiley, New Delhi.
Solomatine, D.P., Dulal, K.N., 2003. Model trees as an alternative to neural networks in rainfall—runoff modelling. Hydrological Sciences Journal 48 (3), 399–411.
Solomatine, D.P., Ostfeld, A., 2008. Data-driven modelling. Some past experiences and new approaches. J Hydroinform 10 (1), 3–22.
Su, B., Su, Z., 1988. The Surface Energy Balance System (SEBS) for Estimation of Turbulent Heat Fluxes. Hydrology and Earth System Sciences 6 (1), 85–100.
Tabari, H., Kisi, O., Ezani, A., Hosseinzadeh Talaee, P., 2012. SVM, ANFIS, regression and climate based models for reference evapotranspiration modeling using limited climatic data in a semi-arid highland environment. Journal of Hydrology 444, 78–89.
Taillandier, P., Grignard, A., Gaudou, B., Drogoul, A., 2014. Des données géographiques à la simulation à base d'agents: application de la plate-forme GAMA. Cybergeo : European Journal of Geography (671).
Taillandier, P., Vo, D.-A., Amouroux, E., Drogoul, A., 2012. GAMA: A Simulation Platform That Integrates Geographical Information Data, Agent-Based Modeling and Multi-scale Control. In: N. Desai, A. Liu, M. Winikoff (Editors), Principles and Practice of Multi-Agent Systems: 13th International Conference, PRIMA 2010, Kolkata, India, November 12-15, 2010, Revised Selected Papers. Springer, Berlin, Heidelberg, pp. 242–258.
Talei, A., Chua, L.H.C., Wong, T.S.W., 2010. Evaluation of rainfall and discharge inputs used by Adaptive Network-based Fuzzy Inference Systems (ANFIS) in rainfall–runoff modeling. Journal of Hydrology 391 (3), 248–262.
Tallaksen, L.M., 1995. A review of baseflow recession analysis. Journal of Hydrology 165 (1), 349–370.
Thomas, J.A., Cover, T.M., 2006. Elements of information theory. Wiley New York, NY, USA.
Troy, T.J., Konar, M., Srinivasan, V., Thompson, S., 2015. Moving sociohydrology forward: a synthesis across studies. Hydrology and Earth System Sciences 19 (8), 3667–

3679.
Tso, B., Mather, P.M., 2009. Classification Methods for Remotely Sensed Data. CRC Press, Boca Raton.
Uhlemann, S., Thieken, A.H., Merz, B., 2010. A consistent set of trans-basin floods in Germany between 1952–2002. Hydrol. Earth Syst. Sci. 14 (7), 1277–1295.
van der Vaart, E., Johnston, A.S.A., Sibly, R.M., 2016. Predicting how many animals will be where: How to build, calibrate and evaluate individual-based models. Ecological Modelling 326, 113–123.
van Genuchten, M.T., 1980. A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. Soil science society of America journal 44 (5), 892–898.
van Parunak, H.D., Savit, R., Riolo, R.L., 1998. Agent-based modeling vs equation-based modeling. A case study and users' guide. Lecture notes in computer science 1534, 10–25.
Vapnik, V., 2013. The nature of statistical learning theory. Springer science & business media.
Wang, J., Rubin, N., Wu, H., Yalamanchili, S., 2013a. Accelerating Simulation of Agent-Based Models on Heterogeneous Architectures, Sixth Workshop on General-Purpose Computation on Graphics Processing Units (GPGPU-6).
Wang, J., Rubin, N., Wu, H., Yalamanchili, S., 2013b. Accelerating Simulation of Agent-Based Models on Heterogeneous Architectures, Sixth Workshop on General-Purpose Computation on Graphics Processing Units (GPGPU-6).
Wang, Q.J., 1991. The genetic algorithm and its application to calibrating conceptual rainfall-runoff models. Water Resour. Res. 27 (9), 2467–2471.
Wang, X., Liu, S., Du, P., Liang, H., Xia, J., Li, Y., 2018. Object-Based Change Detection in Urban Areas from High Spatial Resolution Images Based on Multiple Features and Ensemble Learning. Remote Sensing 10 (2).
Weiler, M., Seibert, J., Stahl, K., 2018. Magic components - why quantifying rain, snow- and icemelt in river discharge isn't easy. Hydrological Processes 32 (1), 160-167.
Wolpert, D.H., Macready, W.G., 1997. No free lunch theorems for optimization. IEEE transactions on evolutionary computation 1 (1), 67–82.
Yu, P.-S., Yang, T.-C., Chen, S.-Y., Kuo, C.-M., Tseng, H.-W., 2017. Comparison of random forests and support vector machine for real-time radar-derived rainfall forecasting. Journal of Hydrology 552, 92–104.
Zadeh, L.A., 1996. Soft computing and fuzzy logic, Fuzzy Sets, Fuzzy Logic, And Fuzzy Systems: Selected Papers by Lotfi A Zadeh. World Scientific, pp. 796–804.

Appendix A: Tables

Tab. 14: Catchment characteristics comprising the geographical unit of the catchment (after Landesamt für Umwelt Bayern, 2017), the basin area, av. flow and the number of events covered in the database.

No.	Name	River	Geogr.Unit	Basin area	Av. Flow	No. Events
				[km^2]	[m^3/s]	[-]
1	Birgsau	Stillach	Northern Calc. Alps	35.6	2.1	50
2	Immenstadt-Zoll-bruecke	Iller	Northern Calc. Alps	724	30.5	60
3	Friedersdorf	Buchbach	Donau-Iller-Lech plain	11.1	0.2	47
4	Gampelmuehle	Ölschnitz	Oberpfalz-Obermain hills	62.2	0.5	44
5	Lohr	Baunach	Keuper-Lias-plain	165.3	1	43
6	Untersteinach	Lower Steinach	Donau-Iller-Lech plain	137.4	2	45
7	Kothmaissling	Chamb	Donau-Isar gravel plains	405	4.3	56
8	Lohmannmuehle	Small Regen	Donau-Isar gravel plains	115.9	2.8	65
9	Marienthal	Regen	Donau-Isar gravel plains	2590.4	37.7	62
10	Zwiesel	Black Regen	Donau-Isar gravel plains	293.4	8.1	59

Tab. 15: Soil physical parameters for Green Ampt and van Genuchten model.

Parameter	Description	Su2	mS
Qr [Vol. %]	Residual water content	0	0
Qs [Vol. %]	Saturated water content	0.3786	0.3886
α [-]	VG-Parameter	0.20387	0.216188
η [-]	VG/GA Parameter	1.2347	1.3533
l [-]	VG Parameter	-3.339	-0.579
k0 [mm/d]	Initial hydraulic con-ductivity	285.5	507.5
ks [mm/d]	Saturated hydraulic conductivity	185.0	375.0

Tab. 16: Statistical parameters from model comparison between cmf and IPA for homogenous soil

Model	Std [%]	Mean [%]	r^2
cmf	0.045	0.29	0.80
IPA	0.039	0.29	

Tab. 17: Statistical parameters from model comparison between cmf and IPA for inhomogeneous soil

Model	Std [%]	Mean [%]	r^2
cmf	0.033	0.27	0.71
IPA	0.036	0.28	

Tab. 18: Correlation coeffecient between scheduling methods for Layer 1

	Random Calling	Energy Gradient	Age
Random Calling	1	0.73	0.53
Energy Gradient	0.73	1	0.66
Age	0.53	0.66	1

Tab. 19: Correlation coefficient between scheduling methods for Layer 2

	Random Calling	Energy Gradient	Age
Random Calling	1	0.9	0.63
Energy Gradient	0.9	1	0.86
Age	0.63	0.86	1

Tables

Figures

Bisher erschienene Hefte der Schriftenreihe HYDROLOGIE / WASSERWIRTSCHAFT

Heft	Jahr	Autor und Titel
1	1983	**Klatt, Peter** Vorhersage von Hochwasser aus radargemessenem und prognostiziertem Niederschlag
2	1983	**Scheider, Klaus** Modell zur gleichzeitigen Erzeugung von Tagesabflussdaten an mehreren Stellen eines Einzugsgebietes
3	1984	**Strübing, Gert** Satellitendaten als Basis der Bestimmung von monatlichen Abflüssen für wasserwirtschaftliche Planungen
4	1985	**Harboe, Ricardo** Optimaler Betrieb wasserwirtschaftlicher Verbundsysteme mit Speichern und anderen Anlagen
5	1986	**Tegtmeier, Ulrike** Wasserwirtschaftliche Projektbewertung – Methoden und Anwendungsbeispiele
6	1987	**Richter, Karl Gerd** Vergleichende hydrologische Untersuchungen des Hochwasserablaufes in Testeinzugsgebieten mit unterschiedlicher Bebauungsdichte
7	1989	**Salas, Edgar** Anwendung der Bayesschen Theorie auf wasserwirtschaftliche Planungen mit hydrologischen Datenreihen
8	1990	**Vogt, Roland** Stauraumverlandung – Naturmessung und Computersimulation
9	1992	**Tiedt, Michael** Freizeitnutzung als Komponente der Wasserwirtschaftlichen Projektbewertung
10	1993	**Gyau-Boakye, Philip** Filling Gaps in Hydrological Runoff Data Series in West-Africa Ergänzung lückenhafter Abflussreihen in West-Afrika
11	1993	**Schumann, Andreas H.** Der Einfluss von Veränderungen der Umweltbedingungen und sozio-ökonomischer Faktoren auf Hydrologie und Wasserwirtschaft
12	1993	**Fett, Werner** Die Nutzung räumlich hoch aufgelöster Gebietsinformationen für die Simulation von Hochwasserganglinien in humiden Mittelgebirgslandschaften

Heft	Jahr	Autor und Titel
13	1994	**Papadakis, Ioannis** Berechnung historischer Abflüsse mit Hilfe multispektraler und multitemporaler digitaler Satellitenbilder
14	1995	**Schultz, G.A. (Hrsg.)** Verfügbarkeit von Wasser Beiträge zur 8. wissenschaftlichen Tagung des DVWK vom 22.-23.03.1995 an der Ruhr-Universität Bochum
15	1996	**Su, Zhongbo** Remote Sensing Applied to Hydrology: The Sauer River Basin Study Fernerkundung angewandt in der Hydrologie: Die Sauer-Einzugsgebiets-Studie
16	1997	**Wolbring, Frank** Wissensbasierte Methoden für den Betrieb von Talsperren
17	1998	**Hornbogen, Martin** Die Planung von Wasserversorgungssystemen auf der Basis des Nachhaltigkeitsprinzips
18	2002	**Schumann, Andreas H. (Hrsg.)** Proceedings Workshop HydroGIS NRW 2002 23.05.2002 Ruhr-Universität Bochum
19	2003	**Quirmbach, Markus** Nutzung von Wetterradardaten für Niederschlags- und Abflussvorhersagen in urbanen Einzugsgebieten
20	2006	**Brass, Carsten** Betriebsoptimierung von Talsperrensystemen mittels Stochastisch Dynamischer Programmierung (SDP) unter Berücksichtigung veränderlicher Ziele und Randbedingungen
21	2006	**Dietrich, Jörg** Entwicklung einer Methodik zur systemanalytischen Unterstützung adaptierbarer Entscheidungsprozesse bei der integrierten Flussgebietsbewirtschaftung
22	2006	**Gattke, Christian** Modellvergleiche zur Untersuchung struktureller Unsicherheiten – Anwendung objektorientierter Methoden in der hydrologischen Modellierung
23	2009	**Schumann, Andreas H. (Hrsg.)** Verbundvorhaben Entwicklung integrativer Lösungen für das operationelle Hochwassermanagement am Beispiel der Mulde – Abschlussbericht

Heft	Jahr	Autor und Titel
24	2009	**Schumann, Andreas H. (Hrsg.)** Integrative Nutzung des technischen Hochwasserrückhalts in Poldern und Talsperren am Beispiel des Flussgebiets der Unstrut
25	2009	**Klein, Bastian** Ermittlung von Ganglinien für die risikoorientierte Hochwasserbemessung von Talsperren
26	2010	**Wisser, Dominik** Modeling of Irrigation and Reservoirs in Regional and global Water Cycles
27	2013	**Nijssen, David** Improving spatiality in decision making for river basin management
28	2015	**Schulte, Markus** Anwendung von Copula-Modellen in der Hochwasserstatistik zur Planung technischer Rückhaltemaßnahmen
29	2016	**Tyralla, Christoph** Identifikation und Reduktion struktureller Unsicherheiten in hydrologischen Modellen
30	2019	**Oppel, Henning** Entwicklung eines selbstkalibrierenden Niederschlags-Abfluss Modells auf Basis der geomorphologischen Einheitsganglinie und Methoden des Machine Learning
31	2019	**Mewes, Benjamin** Application of Machine Learning enhanced Agent-based Techniques in Hydrology and Water Resource Management

Lehrstuhl für Hydrologie, Wasserwirtschaft und Umwelttechnik
Ruhr-Universität Bochum, 2019

Universitätsstraße 150, 44801 Bochum
Tel. +49 (0234) 32 - 24693, Fax. - 14153

ISSN 0949-5975