Orhan D. Tanrikulu, Uwe Ehret, Ingo Haag, Ralf Loritz & Ute Badde

Untersuchungen zum Potenzial maschineller Lernverfahren für die hydrologische Simulation und Vorhersage am Beispiel von LSTM und LARSIM in Baden-Württemberg

Exploring the potential of machine learning methods for hydrological simulation and forecasting using LSTM and LARSIM in Baden-Württemberg

Die Simulation und Vorhersage von Abfluss in Gewässern ist eine zentrale Aufgabe der hydrologischen Modellierung. Dazu finden prozessbasierte, konzeptionelle und datenbasierte Methoden Anwendung, wobei vor allem für letztere in den letzten Jahren eine rasante Entwicklung stattgefunden hat. In diesem Kontext ist es das Ziel dieser Studie, das Potenzial von Long Short-Term Memory (LSTM) Modellen für die langfristige Abflusssimulation und für die kurzfristige Abflussvorhersage zu untersuchen. Dazu werden LSTM-Modelle für vier Pegel in Baden-Württemberg auf Basis hydro-meteorologischer Eingangsgrößen aufgestellt und Abflusssimulationen über vier Jahre bzw. 72 Stunden in die Zukunft reichende Abflussvorhersagen für Hochwasser auf Basis gemessener meteorologischer Antriebsdaten erstellt. Diese werden mit Abflussmessungen an den Pegeln und mit den Ergebnissen des etablierten prozessbasierten Wasserhaushaltsmodells LARSIM (Large Area Runoff Simulation Model) verglichen. Die Ergebnisse dieser Studie zeigen, dass die wichtigsten Einflussgrößen für die Simulationsgüte der LSTMs i) die zeitliche Tiefe des Netzwerks, ii) die Wahl der Eingangsgrößen und iii) die Wahl der Zielfunktion für das Training sind. Für i) stellten sich 3 Monate als bester Kompromiss zwischen Simulationsgüte und Rechenzeit heraus, für ii) verbesserte die Hinzunahme aggregierter Zeitreihen zusätzlich zu den stündlichen Eingangsdaten die Simulationsgüte, und für iii) verbesserte die Erweiterung des mittleren quadratischen Fehlers mit einer zusätzlichen Gewichtung des höchsten aufgetretenen Fehlers die Simulationsgüte vor allem für hohe Abflüsse. Weitere Modifikationen der Eingangsgrößen wie Quantiltransformation und stratifizierte Datenauswahl ergaben keine Verbesserungen. Die Langfristsimulationen mit den LSTMs sind sehr gut (Nash-Sutcliffe Effizienzen von 0,84 bis 0,90) und qualitativ vergleichbar mit denen von LARSIM. Auch hinsichtlich der Wiedergabe von gemessenen Hochwasserscheiteln erzielen die LSTMs sehr gute Ergebnisse, die qualitativ mit denen von LARSIM vergleichbar sind.
Für die Vorhersage von Hochwasser wurden zwei LSTM-Varianten aufgestellt: Rekursive LSTMs, die ihre eigene Vorhersage des vorigen Zeitschritts als Eingangsgröße nutzen, und multi-LSTMs, bei denen für jede Vorhersagetiefe ein separates LSTM trainiert wird. Es zeigte sich, dass rekursive LSTMs wegen des Effekts der Fehlerfortpflanzung weniger robust sind als multi-LSTMs. Mit den multi-LSTMs wurden auch bei den Hochwasservorhersagetests ähnlich gute Ergebnisse erzielt wie mit LARSIM. Auf Basis dieser Ergebnisse wird dargelegt, das Potenzial von maschinellen Lernverfahren für die hydrologische Simulation und Vorhersage weiter auszuloten und auszuschöpfen, zum Beispiel durch die Kombination der Stärken prozessbasierter und datenbasierter Modelle in hybriden Systemen.
Schlagwörter: Maschinelles Lernen, LSTM, LARSIM, Simulation, Vorhersage

The simulation and forecasting of catchment runoff is a central task of hydrological modelling. For this purpose, process-based, conceptual and data-based methods are used, and especially the latter method has shown rapid development in recent years. In this context, the aim of this study is to investigate the potential of Long Short-Term Memory (LSTM) models for long-term simulation and for short-term forecasting of catchment runoff. For these purposes, LSTM models are trained for four gauging stations in Baden-Württemberg using hydro-meteorological input variables, and applied for four years of continuous runoff simulation and 72-hour short-term flood forecast tests based on measured meteorological driving data. The model results are compared with measurements from gauging stations and with the results of the well-established process-based water balance model LARSIM. The results of this study show that the most effective parameters for the LSTM's simulation quality are i) input sequence length, ii) choice of input variables, and iii) choice of the objective function for training. With respect to i), 3 months turned out to be the optimum trade-off between simulation quality and training time, with respect to ii) using aggregated time series data in addition to the hourly input data improved the simulation quality, and with respect to iii) a modified mean squared error with an additional weighting of the highest error improved the simulation quality especially for high discharges. Further modifications and manipulations of the input variables such as quantile mapping and stratified sampling did not result in any improvements. Long-term simulations by LSTMs are very good (Nash-Sutcliffe efficiencies of 0.84 to 0.90) and comparable in quality to LARSIM. The LSTMs also achieve very promising results in simulating observed flood peaks, comparable in quality to the LARSIM simulations.
Two LSTM variants were established for short-term discharge forecast: Recursive LSTM, which uses its own streamflow simulation from the previous time step as an input, and multi-LSTM, in which separate LSTMs are trained for each forecast depth. It was shown that the Recursive LSTMs are less robust than the multi-LSTMs, mainly due to the effect of error propagation. With the multi-LSTMs short-term forecast tests achieved results comparable in quality to those of LARSIM. Based on these results, we propose to further explore and investigate the potential of machine learning methods for hydrological runoff simulation and forecasting, for example by combining the strengths of process-based and data-based models in hybrid systems.
Keywords: Machine Learning, LSTM, LARSIM, simulation, forecast

1 Einleitung

Die Simulation und Vorhersage von Abfluss in Gewässern ist schon immer eine wichtige Aufgabe der Hydrologie. Die ersten Modelle dazu beruhten auf einfachen Regressionen (MULVANY, 1851), mit der zunehmenden Verfügbarkeit von Messdaten, Prozessverständnis und der Rechenkapazität von Computern wurden sie weiterentwickelt zu konzeptionellen, prozessbasierten, und datenbasierten Modellen. Konzeptionelle und prozessbasierte Modelle beruhen hauptsächlich auf der Kombination von physikalischem Prozesswissen und messbaren Strukturgrößen in festen Systemarchitekturen. Datenbasierte Modelle hingegen lernen Beziehungen zwischen Eingabedaten und Zielvariablen hauptsächlich aus gemessenen Zeitreihen relevanter Eingangs- und Ausgangsgrößen. Die Werkzeuge für den Lernprozess datenbasierter Modelle werden heutzutage meist unter dem Begriff Machine Learning (ML) zusammengefasst. ML umfasst Methoden, mit denen ein Computer mit Hilfe von Algorithmen und statistischen Methoden ohne explizite Anweisungen durch einen Nutzer Muster in Daten erkennen und für verschiedene Zwecke nutzbar machen kann (XU und LIANG, 2021). Künstliche neuronale Netze (artifical neural networks (ANN)) spielen dabei wegen ihrer Flexibilität inzwischen eine zentrale Rolle. ANNs basieren auf Arbeiten von MCCULLUCH und PITTS (1943), die neuronale Netzwerke als Rechenarchitekturen vorschlugen, HEBB (1949), der dafür einen Lernalgorithmus entwickelte, und ROSENBLATT (1958), der das Perceptron einführte, eine spezielle neuronale Netzwerkarchitektur, die die Basis moderner ANNs bildet. Eine breite Nutzung erfuhren ANNs schließlich mit der Entwicklung effizienter Trainingsmethoden (RUMELHART et al., 1986) und leistungsfähiger Computer. In der Hydrologie kommen ANNs seit Anfang der neunziger Jahre zur Anwendung (KANG et al., 1993). Die Qualität der Ergebnisse war vor allem dadurch limitiert, dass hydrologische Zeitreihen stark autokorreliert sind, solche zeitlichen Abhängigkeiten aber durch ANNs nicht direkt abgebildet werden können (RUMELHART et al., 1986). Rekurente neuronale Netzwerke (RNN), eine Weiterentwicklung von ANNs, können zeitliche Abhängigkeiten abbilden, allerdings hauptsächlich für kurze Zeiträume. Langzeitabhängigkeiten, wie zum Beispiel der Einfluss von winterlichem Schneefall auf Abflüsse im Frühjahr, werden von RNNs "vergessen". Dieses Problem wurde durch die Entwicklung von Long Short-Term Memory Netzwerken (LSTM) durch HOCHREITER und SCHMIDHUBER (1997) weitgehend behoben. Long Short-Term Memory Netzwerke haben sich daher für die ML-basierte hydrologische Modellierung rasch als Standard etabliert. KRATZERT et al. (2018) zeigten am Beispiel von 241 Einzugsgebieten aus dem CAMELS-US Datensatz (NEWMAN et al., 2015; ADDOR et al., 2017), dass LSTM-Modelle eine ähnlich gute, tägliche Abflusssimulation ermöglichen wie das konzeptionelle hydrologische Modell SAC-SMA. Speziell im Hinblick auf die Simulation von Niedrigwasser kommen SAHOO et al. (2019) zu ähnlichen Ergebnissen.

HU et al. (2018) zeigten anhand einer Simulation von 98 Hochwasserereignissen im Einzugsgebiet des Fen-Flusses, dass LSTMs sich auch für die Hochwasservorhersage eignen und bessere Ergebnisse liefern als ein konzeptionelles hydrologisches Modell. Diese Ergebnisse werden unterstützt durch NEVO et al. (2022), die zeigten, dass LSTMs erfolgreich für die operationelle Hochwasservorhersage in großen Flusseinzugsgebieten eingesetzt werden können, und HUNT et al. (2022), die LSTMs erfolgreich für die Mittelfrist-Hochwasservorhersage nutzten.

In diesem Kontext ist es das Ziel dieser Studie, das Potenzial von LSTM-Modellen für die Langfristsimulation (im folgenden "Simulation") und die Kurzfristvorhersage (im folgenden "Vorhersage") von Abfluss an Gewässerpegeln in Baden-Württemberg zu untersuchen, und durch den Vergleich mit den Ergebnissen eines etablierten prozessbasierten Modells zu bewerten. Dazu werden LSTMs für vier Pegel in Baden-Württemberg aufgestellt. Die Pegel wurden so ausgewählt, dass sie bzgl. Einzugsgebietsgröße und Abflusscharakteristik ein breites Spektrum abdecken. Bei der Simulation werden vierjährige Abflusszeitreihen mit gemessenen Antriebsdaten simuliert. Bei der Vorhersage werden 72-stündige Abflussvorhersagen erstellt, ebenfalls unter Verwendung gemessener Antriebsdaten. Es handelt sich daher um Vorhersagetests unter der idealen Randbedingung, dass das Wetter im Vorhersagezeitraum bekannt ist. Der Grund für die Wahl dieser idealen Randbedingung ist dadurch bedingt, dass dadurch die Qualität der Wettervorhersage keinen Einfluss auf die Qualität der Abflussvorhersage hat, und somit die Qualität der hydrologischen Modelle direkter beurteilt werden kann.

Für die Simulation und die Vorhersage werden jeweils eigene LSTM-Modelle aufgestellt, und die Ergebnisse mit Messdaten und Ergebnissen des Wasserhaushaltsmodells LARSIM verglichen (BREMICKER, 2000; LEG, 2023). Für die Simulationen erfolgt der Vergleich in einem zusammenhängenden Zeitraum von vier Jahren für die Vorhersagen für eine Auswahl der jeweils höchsten an jedem Pegel gemessenen Hochwasserereignisse. Für die Beurteilung der Simulations- bzw. Vorhersagequalität wird das Gütemaß Nash-Sutcliffe Effizienz (NSE) verwendet.

Der Artikel ist wie folgt strukturiert: In Kapitel 2 werden die verwendeten Daten, Methoden und Modelle vorgestellt, und die durchgeführten Experimente beschrieben. In Kapitel 3 werden die Ergebnisse der Experimente vorgestellt und diskutiert. In Kapitel 4 werden die wesentlichen Ergebnisse zusammengefasst und Schlussfolgerungen gezogen.

2 Daten und Methoden

2.1 Hydrometeorologische Daten

LSTM-Modelle wurden für insgesamt vier Pegel in Baden-Württemberg aufgebaut. Die entsprechenden LARSIM-Modelle wurden unverändert von der LUBW übernommen. Die Lage der Pegel, ihre Einzugsgebiete und die wichtigsten Fließgewässer sind in Abbildung 1 dargestellt, hydrologische Kenngrößen der Pegel sind in Tabelle 1 aufgelistet. Die Pegel liegen in den Einzugsgebieten der Enz und der Rems und wurden so ausgewählt, dass sie ein breites Spektrum von Gebietsgrößen und Abflusscharakteristika abdecken. Die Gebietsgrößen reichen von 76 km^2 (Haubersbronn/Wieslauf) 418 km^2 (Schorndorf/Rems), die spezifischen mittleren Abflüsse reichen von 11,7 $ls^{-1}km^{-2}$ (Haubersbronn/Wieslauf) bis 24,8 $ls^{-1}km^{-2}$ (Lautenhof/Enz), und das Verhältnis zwischen mittlerem und höchstem Abfluss im Messzeitraum reicht von 23,6 (Lautenhof/Enz) bis 92,2 (Schwäbisch Gmünd/Rems). Für fast alle Pegel standen 24-jährige Zeitreihen stündlicher Messwerte zur Verfügung (siehe Tab. 1), lediglich für Schwäbisch Gmünd/Rems standen nur 18 Jahre zur Verfügung.

Für LARSIM und die LSTM-Modelle wurden gemessene Zeitreihen von Niederschlag, Lufttemperatur und Globalstrahlung als Antrieb genutzt, für LARSIM zusätzlich noch Luftfeuchte, Wind-

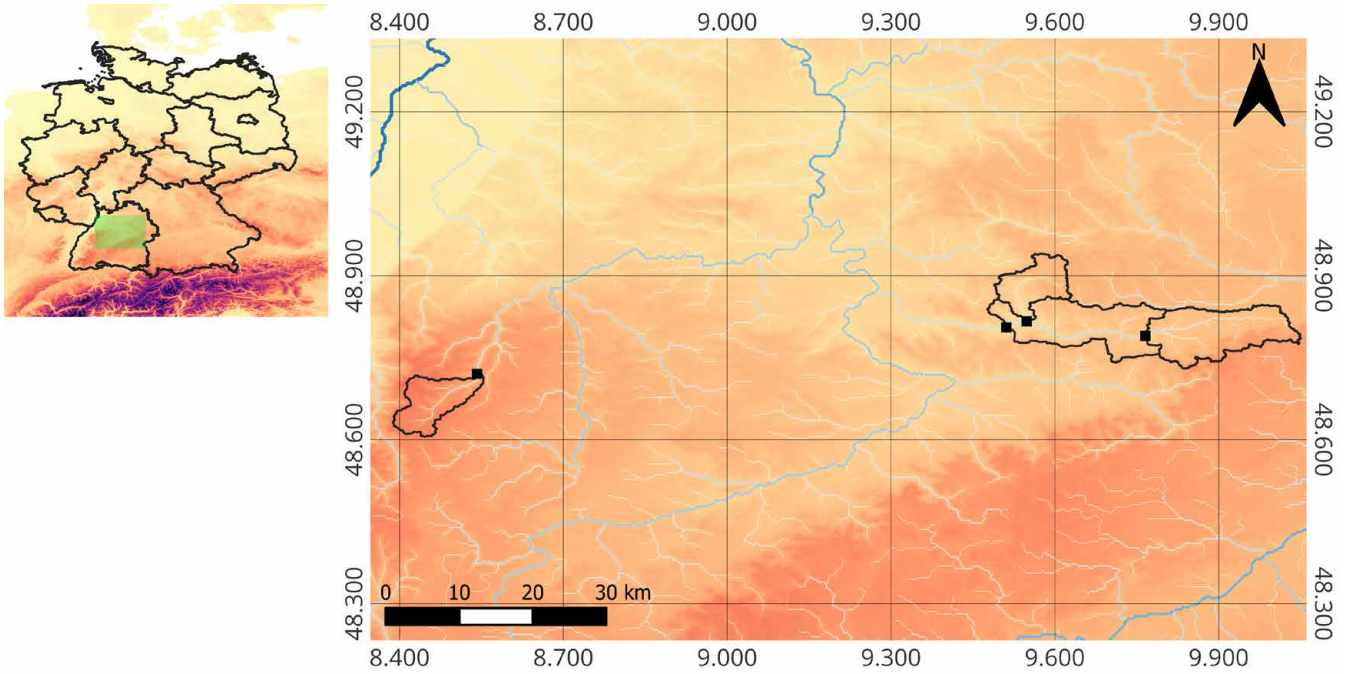

Abbildung 1
Lage der vier Untersuchungspegel und ihrer Einzugsgebiete in Deutschland und Baden-Württemberg. Hintergrundkarte: HYDROSHEDS (https://www.hydrosheds.org/products/hydrosheds).
Location of the four gauging stations and their catchment areas in Germany and Baden-Württemberg. Background map: HYDROSHEDS.

geschwindigkeit und Luftdruck. Grundsätzlich können für die LSTM-Modelle eine beliebige Anzahl von Antriebsdaten genutzt werden, allerdings steigt damit die Gefahr des "Overfittings" (siehe Kap. 2.2) und der Aufwand für Modelltraining und -betrieb. Für die hier aufgestellten LSTM-Modelle zeigten sich die genannten drei Eingangsdaten als bester Kompromiss zwischen Umfang und Informationsgehalt. Die als punktuelle Stationsmessdaten vorliegenden Zeitreihen wurden nicht direkt als Eingangsdaten für die Modelle verwendet. Vielmehr wurden mithilfe eines in LARSIM integrierten modifizierten Rasterpunktverfahrens Zeitreihen der Einzugsgebietsmittel erstellt und als Input für die Modelle verwendet. Die Methoden zur Ermittlung der Einzugsgebietsmittelwerte sind in der LARSIM-Dokumentation (LEG, 2023) in Kapitel 3.2.4 detailliert beschrieben. Es ist zu beachten, dass die LARSIM-Einzugsgebiete eine Größe von jeweils 1 km² haben (siehe Kap. 2.3), die der LSTMs entsprechen den Gebietsgrößen in Tabelle 1. Die räumliche Auflösung der Eingangsdaten ist daher für LARSIM höher als für die LSTMs, da die Einzugsgebiete aber relativ klein sind, ist der daraus resultierende Informationsverlust aber als relativ gering anzusehen.

2.2 LSTM Struktur und Training

Die folgende kurze Einführung in die Struktur von LSTMs ist mit leichten Modifikationen entnommen aus KRATZERT et al. (2021). Weitere Details sind dort, und in KRATZERT et al. (2018) nachzulesen.

"Das LSTM gehört zur Familie der rekurrenten neuronalen Netze. Dies sind neuronale Netze, die Eingabedaten in sequenzieller Reihenfolge verarbeiten. Eine spezielle Eigenschaft von LSTMs ist, dass sie dedizierte interne Speicher besitzen, um Informationen für lange Zeit speichern zu können. Zusätzlich verfügen LSTMs über eine Reihe von sogenannten Gates. Diese kontrollieren in jedem Zeitschritt (a) welche Informationen aus dem Speicher gelöscht werden, (b) was für neue Informationen aus den Eingabedaten in den internen Speicher hinzugefügt werden, und

Tabelle 1
Übersicht über die hydrologischen Eigenschaften der vier Untersuchungspegel und -gebiete, und die verfügbaren Messzeitreihen.
Overview of the hydrological properties of the four gauges and catchments, and the available measurement time series.

Pegelname	Gewässer	Gebietsgröße [km²]	Mittlerer Abfluss im Messzeitraum [m³/s]	Maximaler Abfluss im Messzeitraum [m³/s]	Messzeitraum
Lautenhof	Enz	84,5	2,1	49,6	01.05.1996 – 31.05.2021
Schwäbisch Gmünd	Rems	168,6	2,0	184,4	01.01.2003 – 31.05.2021
Haubersbronn	Wieslauf	76,7	0,9	76,2	01.05.1996 – 31.05.2021
Schorndorf	Rems	418,2	5,2	225,8	01.05.1996 – 31.05.2021

(c) aus welchen Informationen des aktuellen Speichers die Vorhersage gewonnen werden kann".

Die Anzahl der Schichten ("layers") und der zeitlichen Tiefe ("input sequence length" ISL) des LSTM-Netzwerks bilden wichtige strukturelle Parameter, die zusammen mit den Werten aller freien Modellparameter mit Hilfe von vorhandenen Eingangs- und Zielgrößen durch Optimierung bestimmt werden.

LSTMs sind durch ihre inhärente Flexibilität sehr gut geeignet, beliebige Muster in Daten zu erkennen und sie für Aussagen über andere Daten zu nutzen. Diese Flexibilität bringt allerdings im Vergleich zu prozessbasierten Modellen die erhöhte Gefahr des "Overfittings" mit sich, d. h. dass, statt allgemein gültige und robuste Zusammenhänge zwischen den Daten zu lernen, lediglich der Trainingsdatensatz "auswendig gelernt" wird (MAIER et al., 2023). Diese Gefahr kann durch eine geschickte Nutzung der vorhandenen Daten verkleinert werden. Dazu wird ein Teil der Daten direkt für die Bestimmung der Modellparameter durch Optimierung genutzt ("Parameter Training", entspricht der Kalibrierung in der hydrologischen Literatur). Das so optimierte Modell wird in einem separaten Zeitraum angewendet und die Modellgüte bewertet. Die Modellgüte in diesem Zeitraum steuert die Wahl globaler Modellparameter ("Hyperparameter" wie z. B. die Art der Netzwerks, die Anzahl der Netzwerkschichten und -knoten, oder die Wahl der Zielfunktion für die Parameteroptimierung). Dieser Schritt wird als "Hyperparameter Training" bezeichnet und mehrmals iterativ durchlaufen. Er ist in der hydrologischen Literatur am ehesten mit "Modellselektion" gleichzusetzen. Da die Wahl der Hyperparameter auf Basis von Daten erfolgt, die nicht für das Training genutzt wurden, wird die Gefahr des Overfittings reduziert. Außerdem werden weitere Techniken wie Dropout, L1- und L2-Regularisierung verwendet, um ein Overfitting zu verhindern (FROCHTE, 2020).

Das finale Modell, mit festgelegten Werten für Hyperparameter und Parameter, wird abschließend anhand eines bis dahin unter Verschluss gehaltenen Datensatzes final beurteilt ("Validierung"). Dafür werden üblicherweise ca. 20 % der vorhandenen Daten verwendet (JOSEPH, 2022). Die Modellgüte im Validierungszeitraum gibt Aufschluss darüber, wie gebrauchstauglich das Modell ist, d. h. wie gut es bei neuen Situationen funktioniert. Es sei darauf hingewiesen, dass die hier gewählte Terminologie auf hydrologische Verhältnisse angepasst ist. In der englischsprachigen Fachliteratur zu Machine Learning wird für den hier "Hyperparameter Training" genannten Schritt häufig der Begriff "Validation" verwendet, und für den hier "Validation" genannte Schritt der Begriff "Testing". Details dazu, und zum Aufbau und Training von neuronalen Netzwerken sind ausführlich und übersichtlich in MAIER et al. (2023) dargestellt.

2.3 LARSIM

Das Wasserhaushaltsmodell LARSIM (Large Area Runoff Simulation Modell) wurde von BREMICKER (2000) entwickelt. Mit dem Modell werden Interzeption, Schneedynamik, Evapotranspiration, Bodenwasserhaushalt und Abflussbildung auf der Ebene räumlich hoch aufgelöster Hydrotope mit jeweils gleicher Landnutzung und ähnlichen Bodeneigenschaften prozessorientiert simuliert. Abflusskonzentration und Abflussfortpflanzung in den Gewässern werden auf der Ebene von Teileinzugsgebieten prozessorientiert simuliert. Eine Übersicht der Modellstruktur ist in LEG (2023), Abbildung 3-1 zu sehen. Im hier verwendeten LARSIM-Modell für das Neckareinzugsgebiet bestehen die Teileinzugsgebiete aus einen regelmäßigen Raster der Größe 1 km² (LUCE et al., 2006). Das Modell wird durch eine LARSIM-Entwicklergemeinschaft gepflegt und kontinuierlich koordiniert weiterentwickelt (BREMICKER et al., 2013; LEG, 2023). LARSIM-Modelle werden für zahlreiche Fragestellungen angewandt, wie z. B. die Analyse von Auswirkungen des Klimawandels auf den Wasserhaushalt (STAHL et al., 2016; THIREL et al., 2019), die Analyse vergangener Ereignisse (LUDWIG et al., 2023), die Bemessung und Maßnahmenplanung (BREMICKER et al., 2013; HAAG et al., 2023), die Analyse von Sturzfluten in der Folge lokaler Starkregen (HAAG et al., 2022) sowie die Simulation und Vorhersage von Niedrigwasser und Wassertemperaturen (HAAG & LUCE, 2008; ISHIKAWA et al., 2021; HAAG et al., 2023).

Ein zentraler Anwendungsbereich von LARSIM-Wasserhaushaltsmodellen ist jedoch die operationelle Hochwasservorhersage. Entsprechend werden räumlich hoch aufgelöste LARSIM-Wasserhaushaltsmodelle in zahlreichen Bundesländern, unter anderem in Baden-Württemberg, sowie in der Schweiz und Österreich standardmäßig für die Hochwasservorhersage und Hochwasserwarnung angewandt (BREMICKER et al., 2013; LEG, 2023). LARSIM ist somit ein etabliertes und weit verbreitetes prozessbasiertes Hochwasservorhersagemodell, das sich gut als Referenz für die hier untersuchten LSTM-Modelle eignet.

2.4 Experimente

In diesem Abschnitt wird beschrieben, wie LARSIM und die LSTM-Modelle für die Simulationen und Vorhersagen aufgestellt und betrieben wurden, und auf welche Weise sie mit Messdaten und untereinander verglichen wurden. Wie in Kapitel 1 bereits erwähnt, werden bei der Simulation vierjährige Abflusszeitreihen mit gemessenen Antriebsdaten simuliert. Bei der Vorhersage werden 72-stündige Abflussvorhersagen erstellt, ebenfalls unter Verwendung gemessener Antriebsdaten. Es handelt sich daher um Vorhersagetests unter der idealen Randbedingung, dass das Wetter im Vorhersagezeitraum bekannt ist. Der Grund für die Wahl dieser idealen Randbedingung ist, dass dadurch die Qualität der Wettervorhersage keinen Einfluss auf die Qualität der der Abflussvorhersage hat, und somit die Qualität der hydrologischen Modelle besser beurteilt werden kann.

2.4.1 LSTM-Modelle

Für jedes Untersuchungsgebiet wurden jeweils eigene LSTM-Modelle aufgestellt, die unabhängig voneinander betrieben wurden. Das stellt einen Unterschied zu prozessbasierten Flussgebietsmodellen dar, bei denen der in einem Oberliegergebiet simulierte oder vorhergesagte Abfluss über das Gewässernetz stromabwärts weitergereicht wird und damit in Unterliegergebieten als Eingangsgröße zur Verfügung steht. Für die Simulation und die Vorhersage wurden jeweils eigene LSTM-Modelle trainiert, um eine bestmögliche Anpassung an die jeweilige Aufgabe und Datenverfügbarkeit zu ermöglichen. Entgegen der üblichen Vorgehensweise beim Aufbau von Machine Learning Modellen (siehe Kap. 2.2) wurden die verfügbaren Daten nicht in drei sondern nur in zwei Zeiträume aufgeteilt: Das Parameter Training der Modelle erfolgte mit Daten aus dem Zeitraum 1. Mai 1996 bis 30. Spetember 2016 (> 20 Jahre), das Hyperparameter Training erfolgte mit Daten aus dem Zeitraum 1. Oktober 2016 bis 31. Mai 2021 (> 4 Jahre). Der Grund für diese Wahl lag in der verbesserten Vergleichbarkeit mit den LARSIM-Ergebnissen: LARSIM wurde mit Daten aus dem gesamten Zeitraum kalibriert

(siehe Kap. 2.4.2), ein Vergleich mit LSTM Validierungsergebnissen hätte daher einen einseitigen Vorteil für LARSIM dargestellt. Die LSTM-Hyperparameter werden in Kapitel 2.4.1.1 näher erläutert, und in Kapitel 2.4.1.2 werden zwei LSTM-Modellvarianten für die Vorhersage – rekursive und multi-LSTM – vorgestellt und ihre jeweiligen Stärken und Schwächen beschrieben.

2.4.1.1 Allgemeiner Modellaufbau

Alle in dieser Studie erstellten LSTM-Modelle bestehen aus zwei Schichten. Zwei-Schicht LSTMs bieten einen guten Kompromiss zwischen Flexibilität und Optimierbarkeit (HEATON, 2008) und werden sehr häufig verwendet (KRATZERT et al., 2018). Für die Bestimmung der optimalen Anzahl von Basiseinheiten wurden die Varianten {20, 30, 64, 80, 100, 120, 128, 256} pro Schicht untersucht. Dabei erwies sich die Kombination 128-64 als optimal und wurde für alle Modelle verwendet.

Neben der Anzahl der Schichten und Basiseinheiten ist die input sequence length ISL (siehe Kap. 2.2) einer der wichtigsten LSTM-Hyperparameter. Die ISL bezeichnet die Länge des im Netzwerk abgebildeten Zeitfensters. In Abbildung 2 (a) ist dieses Zeitfenster exemplarisch als grünes Rechteck im gesamten Parameter Trainingszeitraum abgebildet, in Abbildung 2 (b) ist das Zeitfenster im Detail dargestellt. Um Abfluss zu einem Zeitpunkt t_0 zu simulieren, nutzt das LSTM alle Eingangsdaten X im Zeitfenster t = [t_0-1, … t_0-1-ISL]. Für hydrologische Modelle wird in der Literatur meist ein Zeitfenster von einem Jahr empfohlen (KRATZERT et al., 2018; GAUCH et al., 2021; FRAME et al., 2022; NEARING et al., 2022), um dem Modell Zugang zu allen Daten des hydrologischen Jahresgangs zu geben. Dabei handelt es sich allerdings meist um Tageswertmodelle, womit der Datenumfang begrenzt bleibt und bei der (rechenintensiven) LSTM-Optimierung keine Rechenzeitprobleme erzeugt. Im Rahmen der hier vorgestellten Arbeiten wurden aber Stundendaten verwendet, was mit den für diese Studie zur Verfügung stehenden Rechenressourcen eine Verkürzung des Zeitfensters auf drei Monate erforderlich machte. Um den damit verbundenen Informationsverlust über den hydrologischen Jahresgang wie z. B. saisonübergreifende Effekte von Schnee, Grundwasser und Bodenfeuchte zu kompensieren, wurden zusätzlich zu den Niederschlagsdaten in Stundenauflösung auch noch Zeitreihen aggregierter Niederschläge (Tages-, Wochen-, Monats-, und 3-Monats-Summen) erzeugt und als Modellantrieb verwendet.

Ein weiterer wichtiger Hyperparameter ist die Wahl der für die Optimierung verwendeten Zielfunktion. Die Zielfunktion misst den Abstand zwischen der Modellsimulation und den zugehörigen Beobachtungswerten, dieser wird im Verlauf des Parameter-Trainings durch Parametervariation iterativ minimiert. Für Optimierungsaufgaben existieren eine Vielzahl von Zielfunktionen, die allgemein gebräuchlichste (auch in ML) ist der mittlere quadratische Fehler MSE. Im MSE werden große Abweichungen überproportional gewichtet und damit bei der Optimierung bevorzugt verkleinert (WANG et al., 2022). Für die hydrologische Modellierung mit dem Ziel einer guten Wiedergabe hoher Abflüsse bietet sich daher der MSE als Zielfunktion an, und wurde zunächst auch für die LSTM-Optimierung verwendet. Dabei zeigte sich, dass Hochwasserabflüsse – trotz ihrer großen Werte – aufgrund ihrer Seltenheit in der Zielfunktion zu wenig Gewicht haben, so dass die Wiedergabe von Hochwasserabflüssen nicht zufriedenstellend war. Um große Fehler bei großen Abflüssen noch stärker zu berücksichtigen wurde schließlich ein erweiterter MSE-Wert entwickelt und verwendet (siehe Gl. 1).

$$MSE^{+} = MSE + k\,(\max[(obs - sim)^2]) \qquad (1)$$

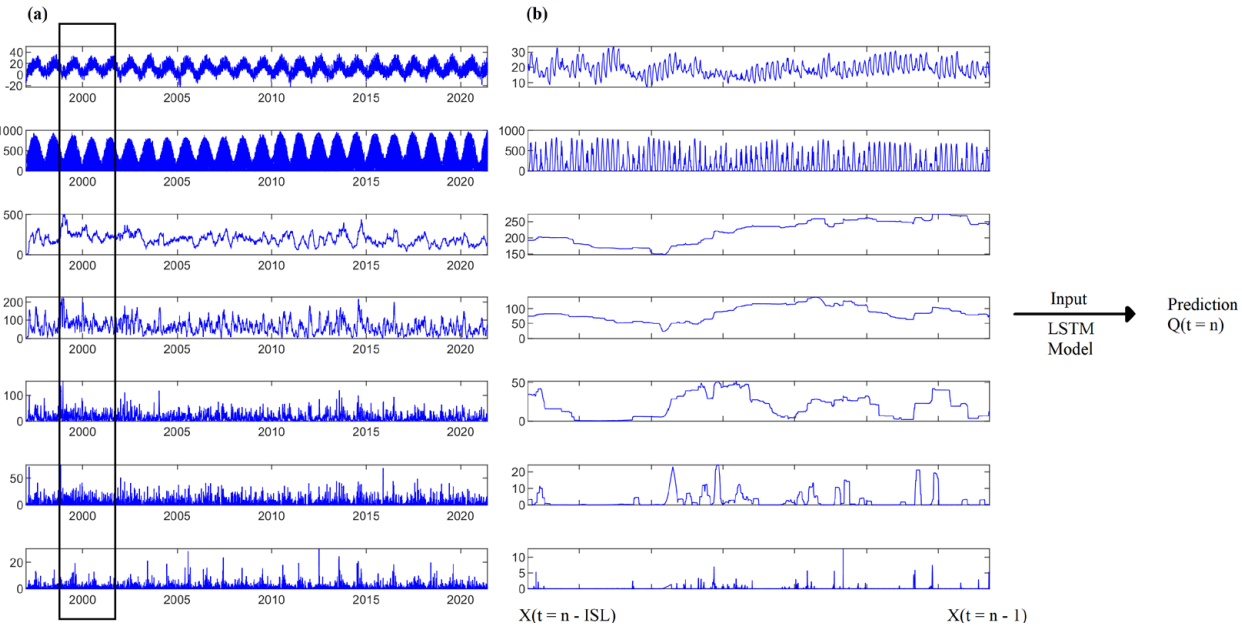

Abbildung 2
Übersicht zur Datennutzung beim LSTM-Training: (a) Darstellung aller Daten für das Parameter Training und der Länge einer Trainingssequenz; (b) Exemplarische Darstellung einer Trainingssequenz im Detail.
Use of data for LSTM Training: (a) Representation of all data for the parameter training and the length of a training sequence; (b) Exemplary representation of a training sequence length in detail.

Dabei wird der MSE um den größten aufgetretenen Fehler ergänzt, der durch den Faktor k variabel gewichtet werden kann. Durch Variation von k im Wertebereich [0,5] wurde schließlich als bester Wert k = 0,2 bestimmt. In Abbildung 3 sind exemplarisch für den Untersuchungspegel Lautenhof/Enz Streudiagramme für gemessene gegen LSTM-simulierte Abflüsse gezeigt, wobei das LSTM einmal mit MSE optimiert wurde (obere Reihe), und einmal mit MSE⁺ (untere Reihe). Insbesondere für den Trainingszeitraum (linke Spalte) zeigt sich für den MSE⁺ eine deutlich realitätsnähere Abbildung hoher Abflüsse als für MSE. Auch für den Testzeitraum liegen die MSE⁺-basieren Simulationen insgesamt näher an den Messungen.

Neben MSE und MSE⁺ wurde als Zielfunktionen auch noch NSE getestet, der allerdings keine Verbesserungen gegenüber MSE⁺ erbrachte. Zusätzlich zu den genannten Anpassungen wurde noch untersucht, inwieweit eine Transformation und Erweiterung der Eingangsdaten die Modellgüte verbessert. Untersucht wurden die Quantiltransformation und die stratifizierte Auswahl der Eingangsdaten, beide mit dem Ziel einer stärkeren Gewichtung hoher Abflusswerte, und die Hinzunahme einer Abflusskategorisierung in drei Klassen auf Basis der Abflussdauerlinie jedes Pegels. Keine dieser Maßnahmen führte zu einer Verbesserung der Modellgüte, so dass sie nicht weiter verwendet wurden. Darüber hinaus wurde untersucht, ob ein regionales LSTM-Training im Vergleich zu lokalem Training Vorteile bietet. Beim regionalen Training wird zuerst ein gemeinsames regionales LSTM mit den Daten aller Untersuchungsgebiete trainiert; anschließend wird das regionale LSTM jeweils einzeln mit lokalen Daten aus den Gebieten nachtrainiert und damit lokal verfeinert. Regional trainierte LSTMs sind durch die Größe und hohe Bandbreite des regionalen Trainingsdatensatzes üblicherweise robuster als ausschließlich lokal trainierte Modelle (KRATZERT et al., 2019). Im Rahmen dieser Studie wurden beide Varianten untersucht, und, da die Unterschiede in der Modellqualität nur gering waren (in der Größenordnung ± 0,05 NSE), wurden aus Gründen der Einfachheit in der Folge lokal trainierte Modelle verwendet. Der positive Effekt des regionalen Modelltrainings war vermutlich wegen der geringen Anzahl von nur vier beteiligen Gebieten klein. Für Untersuchungen mit vielen Gebieten bietet sich aber regionales Training an.

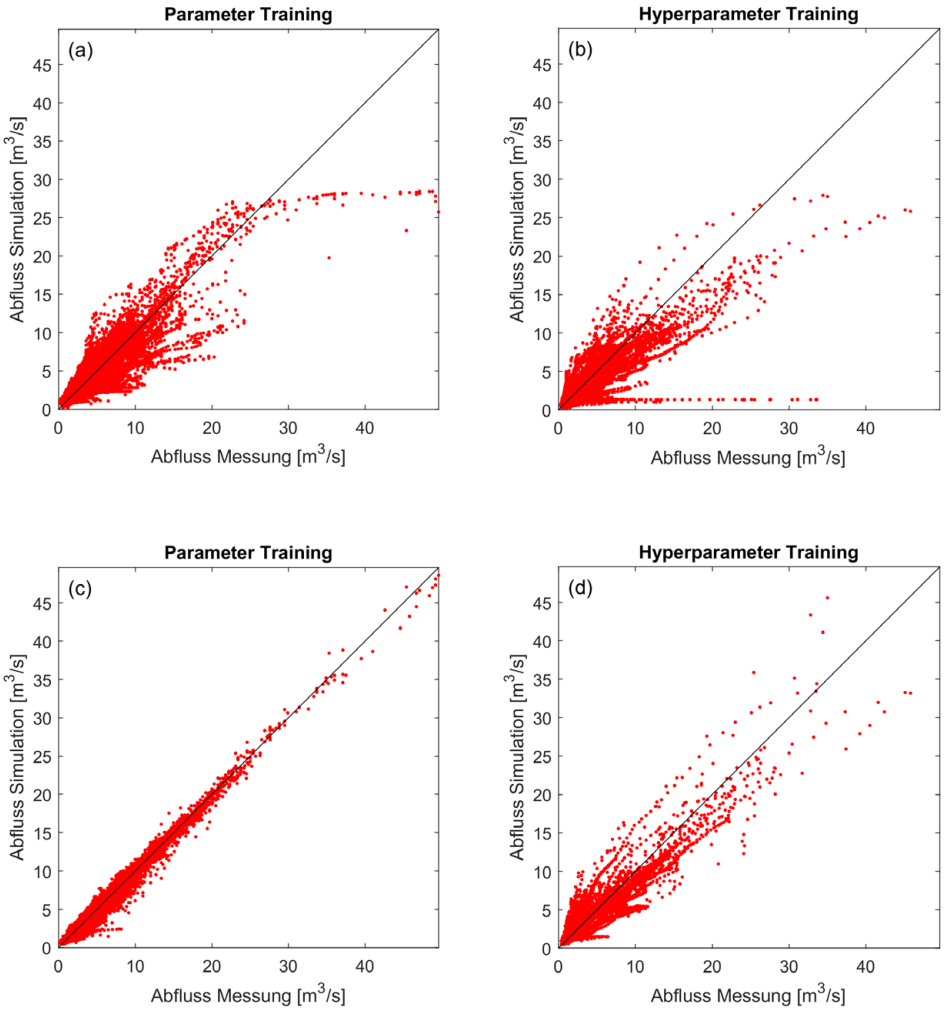

Abbildung 3
Ergebnisse für LSTM-Parameter- und Hyperparameter-Training mit Zielfunktion MSE (a und b) und mit Zielfunktion MSE⁺ (c und d) am Pegel Lautenhof/Enz.
Results for LSTM parameter and hyperparameter training with objective function MSE (a and b) and with objective function MSE⁺ (c and d) at gauge Lautenhof/Enz.

2.4.1.2 Modellvarianten für die Vorhersage

Bei der operationellen hydrologischen Vorhersage liegt der Fokus auf der korrekten Vorhersage von Abflusswerten in der nahen Zukunft, üblicherweise im Zeitraum von Stunden bis Tagen, während in der Langfristsimulation der Fokus auf der korrekten Wiedergabe des hydrologischen Jahresgangs liegt. Aufgrund dieser verschiedenen Zielsetzungen unterscheiden sie sich auch bzgl. der Verfügbarkeit und der Relevanz ihrer Eingangsdaten. Während für die Langfristsimulation Messwerte saisonal variierender Größen wie Lufttemperatur, Schneehöhen oder Grundwasserstände wichtig sind, sind für die Kurzfristvorhersage aktuelle Abflussmessungen die wichtigsten Indikatoren des aktuellen Gebietszustandes, und damit auch der Abflussbildung für die nahe Zukunft (NEARING et al., 2022). Abflussmessungen sind für operationelle Vorhersagen meist bist zum Vorhersagezeitpunkt verfügbar, und können daher als Eingangsgrößen für das Training von Vorhersagemodellen genutzt werden. Für die in der Zukunft liegenden Anwendungszeiträume liegen diese Daten allerdings nicht vor, im Gegensatz zu anderen Eingangsgrößen wie Niederschlag oder Temperatur, für die meist Vorhersagen von Wetterdiensten zur Verfügung stehen. Dieses Problem kann für LSTM-Vorhersagemodelle auf zwei Arten gelöst werden, beide wurden im Rahmen dieser Studie untersucht: Rekursive LSTMs verwenden auch im Vorhersagezeitraum Abflusswerte als Eingangsgröße, und erzeugen diese selbst durch ein schrittweises (rekursives) Voranschreiten in der Zeit (Abb. 4, linke Spalte).

Dies entspricht der Anwendung prozessbasierter Modelle für die Hochwasservorhersage. Für LSTMs mit ihrer hohen Flexibilität und Sensibilität hinsichtlich der Trainingsdaten entsteht aus dieser Vorgehensweise allerdings ein Problem: Das LSTM wird mit gemessen Abflüssen trainiert, aber mit selbst erzeugten Abflüssen betrieben. Wenn diese Datenkollektive voneinander abweichen ("Bias" der Abflussvorhersagen), kann das zu einer selbstverstärkenden Fehlerfortpflanzung und damit einer schnell sinkenden Vorhersagequalität führen (LAMB et al., 2016). Interessanterweise kann dieses Problem verkleinert werden, indem man die für das Training verwendeten Abflussdaten "verrauscht",

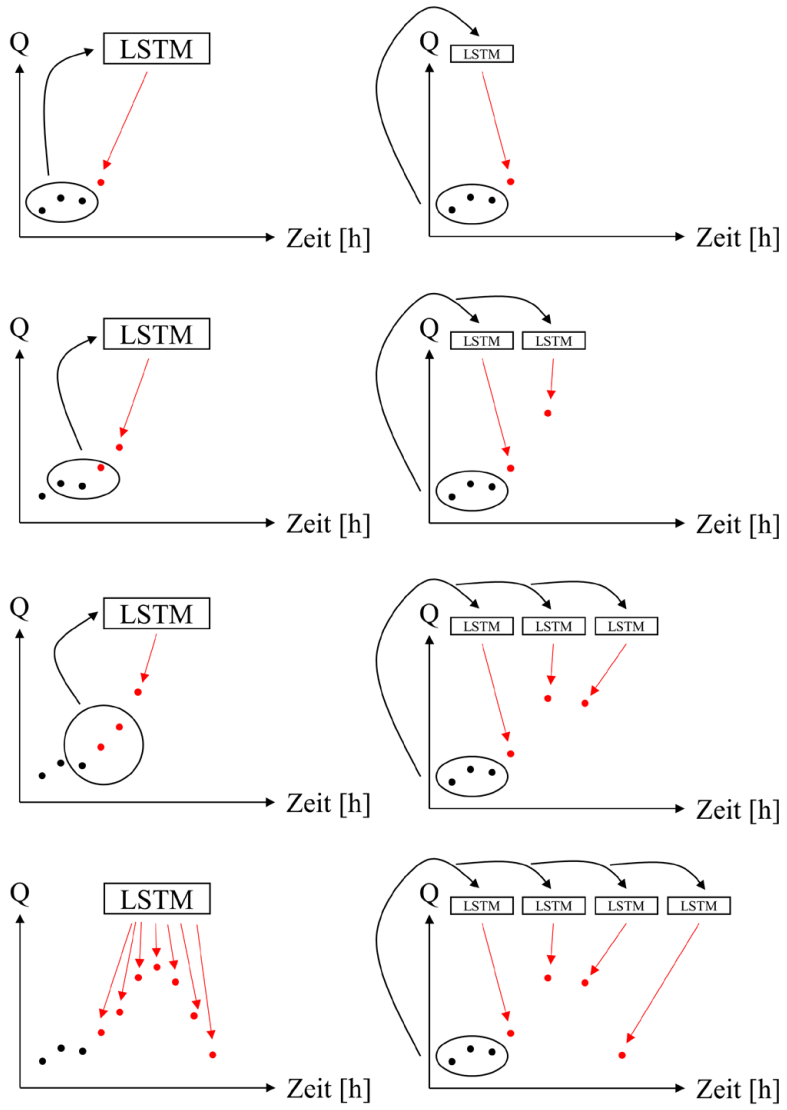

Abbildung 4
Arbeitsablauf des rekursiven LSTM (linke Spalte) und des multi-LSTM (rechte Spalte). LSTM-Ausgaben sind als rote Punkte gekennzeichnet.
Workflow of the recursive LSTM (left column) and the multi-LSTM (right column). LSTM outputs are marked as red points.

d. h. sie mit einem kleinen, zufälligen Fehler in der Größenordnung des Vorhersagebias beaufschlagt. Damit lernt das LSTM von Daten, die den später bei der Modellanwendung verfügbaren Daten eher entsprechen als die ursprünglichen Messdaten (NEARING et al., 2022. Im Rahmen dieser Studie wurden die für das Training genutzten Abflussmessungen zeitschrittweise unabhängig mit einem gleichverteilten multiplikativen Fehlerterm im Wertebereich [0,8, 1,2] verrauscht. Der Wertebereich – ein Hyperparameter – wurde iterativ aus dem Wertebereich [0,5, 1,5] heraus durch Maximierung der Vorhersagegüte bestimmt. In Abbildung 5 ist dazu ein exemplarisches Beispiel zu sehen.

Für die Vorhersagetests wurden neben den rekursiven LSTMs auch noch eine zweite LSTM-Variante – multi-LSTMs – untersucht. Dabei wird für jede gewünschte Vorhersagetiefe im Vorhersagezeitraum jeweils ein separates LSTM trainiert, jeweils mit denselben meteorologischen Messungen und den Abflussmessungen bis zum Vorhersagezeitpunkt, und mit jeweils allen meteorologischen Vorhersagen bis zur jeweiligen Vorhersagetiefe im Vorhersagezeitraum. Der Arbeitsablauf eines multi-LSTM Ensembles ist in Abbildung 4, rechte Spalte dargestellt. Da die einzelnen Modelle des multi-LSTMs voneinander unabhängig operieren, entfällt das Problem der Fehlerfortpflanzung, dafür ist der Trainingsaufwand im Vergleich zum rekursiven LSTM höher, da statt nur einem viele LSTMs trainiert werden müssen. Um den Trainingsaufwand zu begrenzen, wurden daher im Rahmen dieser Studie nicht für jede Stunde der gewünschten Vorhersagetiefe (72 Stunden) LSTMs erstellt, sondern nur für die Vorhersagetiefen {1, 3, 6, 9, 12, 18, 24, 30, 36, 45, 54, 63, 72} Stunden, und dazwischen linear interpoliert.

2.4.2 LARSIM

Für diese Studie wurden LARSIM-Modelle von der Hochwasservorhersagezentrale (HVZ) der Landesanstalt für Umwelt Baden-Württemberg (LUBW) zur Verfügung gestellt. Diese Modelle stehen flächendeckend für Baden-Württemberg zur Verfügung und werden wie in Kapitel 2.3 beschrieben für unterschiedliche Fragestellungen genutzt, insbesondere aber für die operationelle Hochwasservorhersage. Die landesweiten LARSIM-Modelle wurden anhand von ca. 200 Pegeln mit Messdaten des Zeitraums 1. November 1997 bis 9. Juni 2021 (> 23 Jahre) manuell kalibriert (MORETTI et al., 2022). Die Kalibrierung erfolgte nach den Maßgaben des entsprechenden Leitfadens für operationelle LARSIM-Modelle (HAAG et al. 2021). Dabei werden die Modelle nicht ausschließlich für die bestmögliche Wiedergabe des Abflusses am Pegel mithilfe einer einzelnen Zielfunktion (z. B. MSE, NSE) optimiert. Vielmehr werden neben den Ganglinien am Pegel unter anderem auch hydrologische Hauptwerte wie MNQ und MHQ, der realistische Verlauf der Bodenfeuchte, die Nachbildung des Base-Flow-Index als Maß für die Tiefenversickerung und die räumliche Konsistenz der Modellparameter einbezogen. Durch diese multikriterielle Kalibrierung erhält man in der Regel eine wesentlich robustere Parameterwahl als durch die reine Minimierung des MSE oder die Maximierung der NSE. Durch die Kalibrierung wird also nicht der minimale MSE oder die maximale NSE erreicht, sondern robuste, räumlich konsistente Modelle, die in der Regel auch eine gute Extrapolationsfähigkeit aufweisen. Es

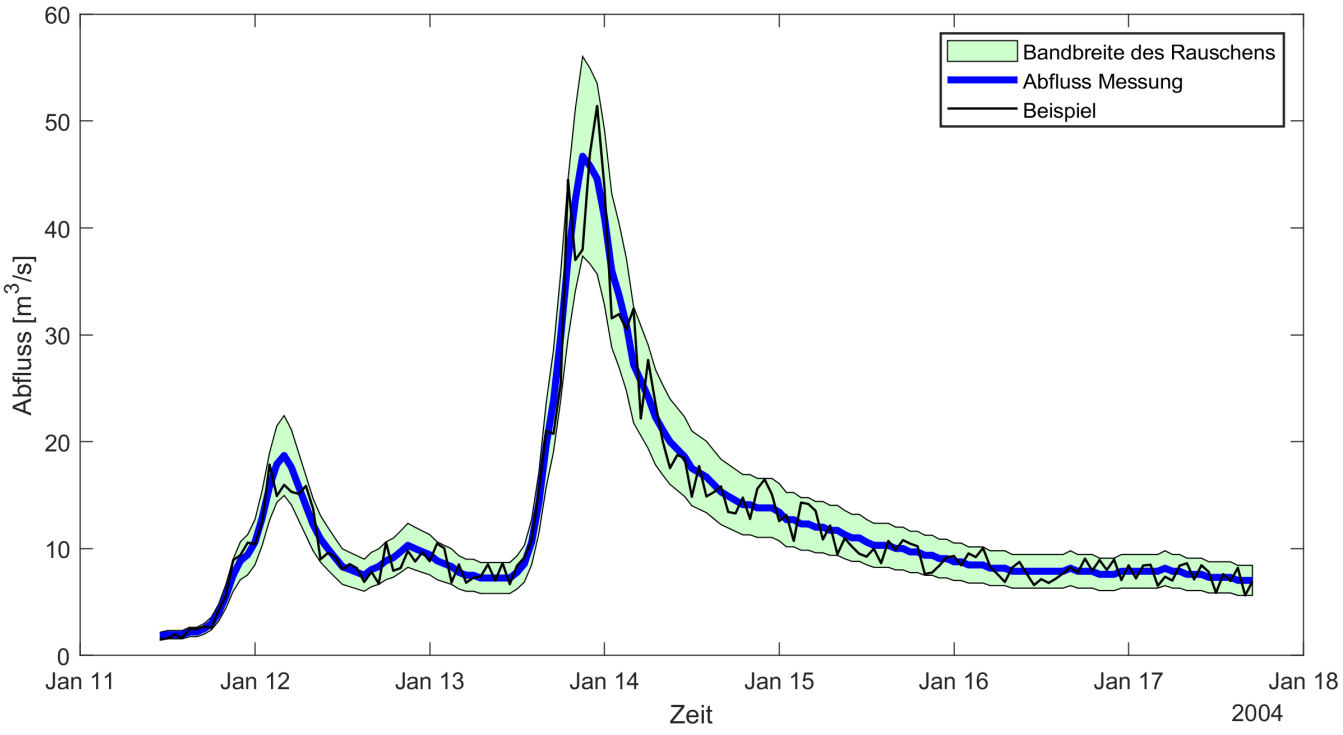

Abbildung 5
Beispiel für das Verrauschen von Eingangsdaten für das Training des rekursiven LSTMs. Die blaue Linie zeigt Original-Messwerte, das grüne Band den möglichen Wertebereich des Rauschens mit Faktoren aus dem Wertebereich [0,8, 1,2], die schwarze Linie zeigt daraus beispielhaft eine Realisation.
Example of noisy input data for training of the recursive LSTM. The blue line shows original measurement values, the green band shows the possible value range of the noise with factors from the value range [0.8, 1.2], the black line shows an example of actual noisy input used in the training.

ist zu beachten, dass der für den Vergleich zwischen LARSIM und LSTM gewählte Zeitraum 1. Oktober 2016 bis 31. Mai 2021 innerhalb des LARSIM-Kalibrierungszeitraums liegt. Der Grund dafür ist, dass noch keine neueren Daten zur Verfügung standen. Wie in Kapitel 2.4.1 erwähnt wurden daher die LARSIM-Ergebnisse mit denen des LSTM-Hyperparameter-Trainings verglichen. Die quantitative Vergleichbarkeit zwischen LARSIM- und LSTM-Modellen ist also dadurch etwas eingeschränkt, dass LARSIM nicht streng hinsichtlich der bestmöglichen Nachbildung des Abflusses am Pegel optimiert wurde, und der vierjährige Vergleichszeitraum bei der Anpassung beider Modelle unterschiedlich berücksichtigt wurde. Kleinere Unterschiede in numerischen Gütemaßen sollten also nicht überinterpretiert werden. Trotz dieser Einschränkung kann der Vergleich mit den LARSIM-Ergebnissen zweifellos zur qualitativen Bewertung der LSTM-Modelle herangezogen werden.

2.4.3 Bewertung der Modellgüte

Für die Bewertung der Modellgüte von LARSIM- und der LSTM-Modelle bei den Simulationen wird die Nash-Sutcliffe Effizienz (NSE) zwischen simulierten und gemessenen Abflüssen an den vier Untersuchungspegeln über den gesamten Testzeitraum berechnet. Darüber hinaus wird auch die korrekte Wiedergabe von Hochwasserereignissen untersucht. Dazu werden für jeden Pegel die jeweils sechs größten Hochwasserereignisse im Testzeitraum ausgewählt und der jeweilige Scheitelabfluss bestimmt. Für Haubersbronn/Wieslauf traten in diesem Zeitraum nur drei relevante Hochwasserereignisse auf, so dass nur diese untersucht wurden. Aus den simulierten Abflussganglinien wird ebenfalls der zugehörige Scheitelabfluss bestimmt und mit den Messwerten verglichen: Weicht der simulierte vom gemessenen Scheitelabfluss um mehr als ± 25 % ab, so wird das Ereignis als unter- bzw. überschätzt kategorisiert, alle anderen Fälle gelten als korrekt.

Für die Bewertung der Vorhersagegüte werden für die ausgewählten Hochwasserereignisse stündliche 72-Stunden-Vorhersagen mit LARSIM und den LSTMs gerechnet. Dabei liegt der erste Vorhersagezeitpunkt jeweils kurz vor Ereignisbeginn, der letzte kurz nach Erreichen des Abflussscheitels, dazwischen liegen sie in stündlichem Abstand. Dann wird separat für jeden Pegel und jede Vorhersagetiefe, aber gemeinsam für alle Vorhersageläufe eines Modells, der NSE zwischen Vorhersage und Messwert bestimmt.

3 Ergebnisse und Diskussion

3.1 Simulation

Für alle Untersuchungspegel sind die NSE- und KGE-Werte im gesamten Vergleichszeitraum für LARSIM und LSTM in Tabelle 2 aufgelistet. Da sich die Ergebnisse für NSE und KGE sehr ähneln, werden im Folgenden nur die NSE-Werte diskutiert. Im Mittel über alle Pegel liegt die Simulationsgüte für LARSIM bei 0,81 und für LSTM bei 0,85. Beide Modelle können damit als sehr gut bezeichnet werden. Die einzige Ausnahme bildet die LARSIM-Simulation für den Pegel Haubersbronn mit einem NSE-Wert von 0,69. Der Grund für diese Abweichung ist ein einzelnes Hochwasserereignis, das deutlich unterschätzt wurde. Beim direkten Vergleich zwischen LARSIM und LSTM zeigt sich, dass LSTM an drei von vier Pegeln eine bessere Simulationsgüte aufweist (fett markierte Werte in Tab. 2), die Unterschiede aber nur gering sind. Wenngleich die kleinen NSE-Unterschiede vor dem Hintergrund der unterschiedlichen Optimierung der beiden Modelle nicht überbewertet werden sollten (siehe Kapitel 2.4.2), kann festgehalten werden, dass die LSTM ähnlich gute Simulationsgüten erreichen wie LARSIM.

Zum weiteren Vergleich sind in Abbildung 6 für alle Untersuchungspegel Streudiagramme zwischen den Messwerten und den LARSIM- bzw. LSTM-Simulationen im gesamten Vergleichszeitraum abgebildet (rechte Spalte). Auch hier zeigt sich eine gute Wiedergabe der Messwerte durch die Simulationen im gesamten Abflussspektrum (die Punkte liegen meist nahe an der ersten Winkelhalbierenden), lediglich bei Haubersbronn ist das einzelne durch LARSIM stark unterschätzte Hochwasser als horizontale Punktreihe zu erkennen. In der linken Spalte sind exemplarische Zeitreihen gemessener und simulierter Abflüsse abgebildet, aus denen hervorgeht, dass die beobachtete Abflussdynamik inklusive Hoch- und Niedrigwasserphasen sowohl durch LARSIM als auch LSTM ähnlich gut abgebildet wird.

Für die Beurteilung der Simulationsgüte bei Hochwasser sind in Tabelle 3 für insgesamt 21 Hochwasserscheitel – die jeweils sechs höchsten Hochwasser an drei Pegeln und die drei höchs-

Tabelle 2
Simulationsgüte (Nash-Sutcliffe Effizienz NSE und Kling-Gupta Effizienz KGE) für LARSIM und LSTM im gesamten Vergleichszeitraum. Für jeden Wert ist der im Vergleich LARSIM-LSTM höhere (= bessere) Wert fett gekennzeichnet.
Simulation quality (Nash-Sutcliffe efficiency NSE and Kling-Gupta efficiency KGE) for LARSIM and LSTM in the entire comparison period. For each catchment, the higher (= better) value in the LARSIM-LSTM comparison is marked in bold.

Pegel	LARSIM NSE	LSTM NSE	LARSIM KGE	LSTM KGE
Lautenhof/Enz	0,84	**0,89**	0,72	**0,85**
Schwäbisch Gmünd/Rems	**0,86**	0,80	**0,92**	0,72
Haubersbronn/Wieslauf	0,69	**0,84**	0,81	**0,82**
Schorndorf/Rems	0,87	**0,90**	0,91	**0,92**

Tabelle3
Trefferquote von LARSIM und LSTM für Scheitelabflüsse von 21 ausgewählten Hochwasserereignissen im Vergleichszeitraum (Je 6 Ereignisse pro Pegel, mit Ausnahme von Haubersbronn mit nur 3 Ereignissen).
Hit rate of LARSIM and LSTM for peak discharges from 21 selected flood events in the comparison period (6 events per gauge, except Haubersbronn with only 3 events).

			LSTM		
			Unterschätzung	Korrekt	Überschätzung
LARSIM		**Gesamt**	**7**	**12**	**2**
	Unterschätzung	**8**	2	6	0
	Korrekt	**13**	5	6	2
	Überschätzung	**0**	0	0	0

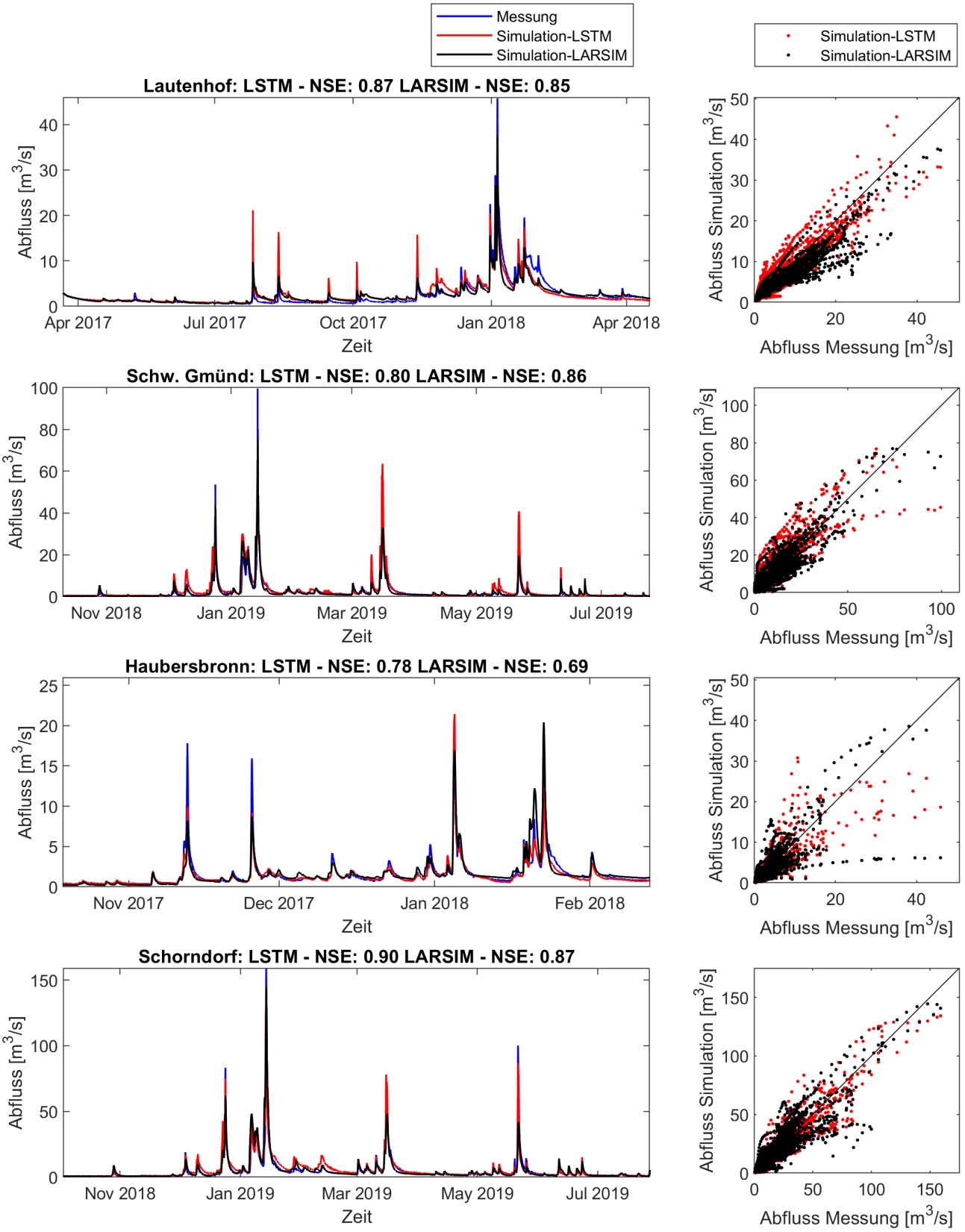

Abbildung 6
Linke Spalte: Exemplarische Abflussganglinien der vier Untersuchungspegel. Messwerte (blau), mit LARSIM simuliert (schwarz) und mit LSTM simuliert (rot). Rechte Spalte: Abfluss-Streudiagramm mit Messwerten gegen LARSIM-Simulationen (schwarz) und gegen LSTM-Simulationen (rot) für den gesamten Vergleichszeitraum.
Left column: Example of streamflow hydrographs of the four catchments. Measurements (blue), simulated with LARSIM (black) and simulated with LSTM (red). Right column: Scatter plot of measured streamflow vs. LARSIM simulations (black) and LSTM simulations (red) for the entire comparison period.

ten Hochwasser am Pegel Haubersbronn – die LARSIM- und LSTM- Simulationen der Scheitelabflüsse als kategorische Kontingenztabelle abgebildet (vgl. Kap. 2.4.3). Nach dieser Klassifikation simuliert LARSIM 13 der Scheitelabflüsse korrekt, unterschätzt acht und überschätzt keinen. Es zeigt sich also eine mehrheitlich korrekte Reproduktion der Scheitelabflüsse und eine Tendenz zur Unterschätzung. Bei LSTM liegt der Anteil korrekt simulierter Scheitelabflüsse mit 12 etwas niedriger als bei LARSIM. Wie bei LARSIM ist auch bei LSTM der Anteil unterschätzter Fälle (sieben) größer als der Anteil überschätzter Fälle (zwei). Es zeigt sich also sowohl für LARSIM als auch LSTM eine mehrheitlich korrekte Reproduktion der Scheitelabflüsse.

3.2 Vorhersage

3.2.1 Rekursives LSTM vs. multi-LSTM

In Kapitel 2.4.1.2 wurden zwei mögliche LSTM-Modellvarianten für die Vorhersage vorgestellt, rekursive LSTM und multi-LSTM. Um zu entscheiden, welche der Varianten verwendet werden soll, wurden beide Modelle an allen Pegel trainiert und anschließend Vorhersagetests für die jeweils sechs höchsten Hochwasserereignisse pro Pegel durchgeführt. Für Haubersbronn wurden nur die höchsten drei Hochwasserereignisse genutzt (siehe Kap. 2.4.3). Daraus wurden NSE-Werte für jede Vorhersagetiefe berechnet (siehe Kap. 2.4.3). Für das rekursive LSTM lagen die NSE-Werte über alle Vorhersagetiefen im Bereich [0,3, 0,6], für das multi-LSTM lagen alle über 0,75. Das multi-LSTM ist also deutlich besser und wurde für die weiteren Vorhersagetests verwendet. Das Ergebnis zeigt, dass auch das künstliche Verrauschen der Abflussmessungen das Problem der Fehlerfortpflanzung nicht vollständig beheben konnte. Zur Verdeutlichung sind dazu in Abbildung 7 für ein exemplarisches Hochwasserereignis am Pegel Lautenhof/Enz und einem Vorhersagezeitpunkt kurz vor dem Beginn des eigentlichen Hochwasserereignisses die Vorhersageganglinien beider Modellvarianten dargestellt. Für die ersten Vorhersageschritte sind beide Vorhersagen noch sehr ähnlich, dann aber unterschätzt das rekursive LSTM den Abflussanstieg und behält durch Fehlerfortpflanzung diese Unterschätzung im weiteren Verlauf bei.

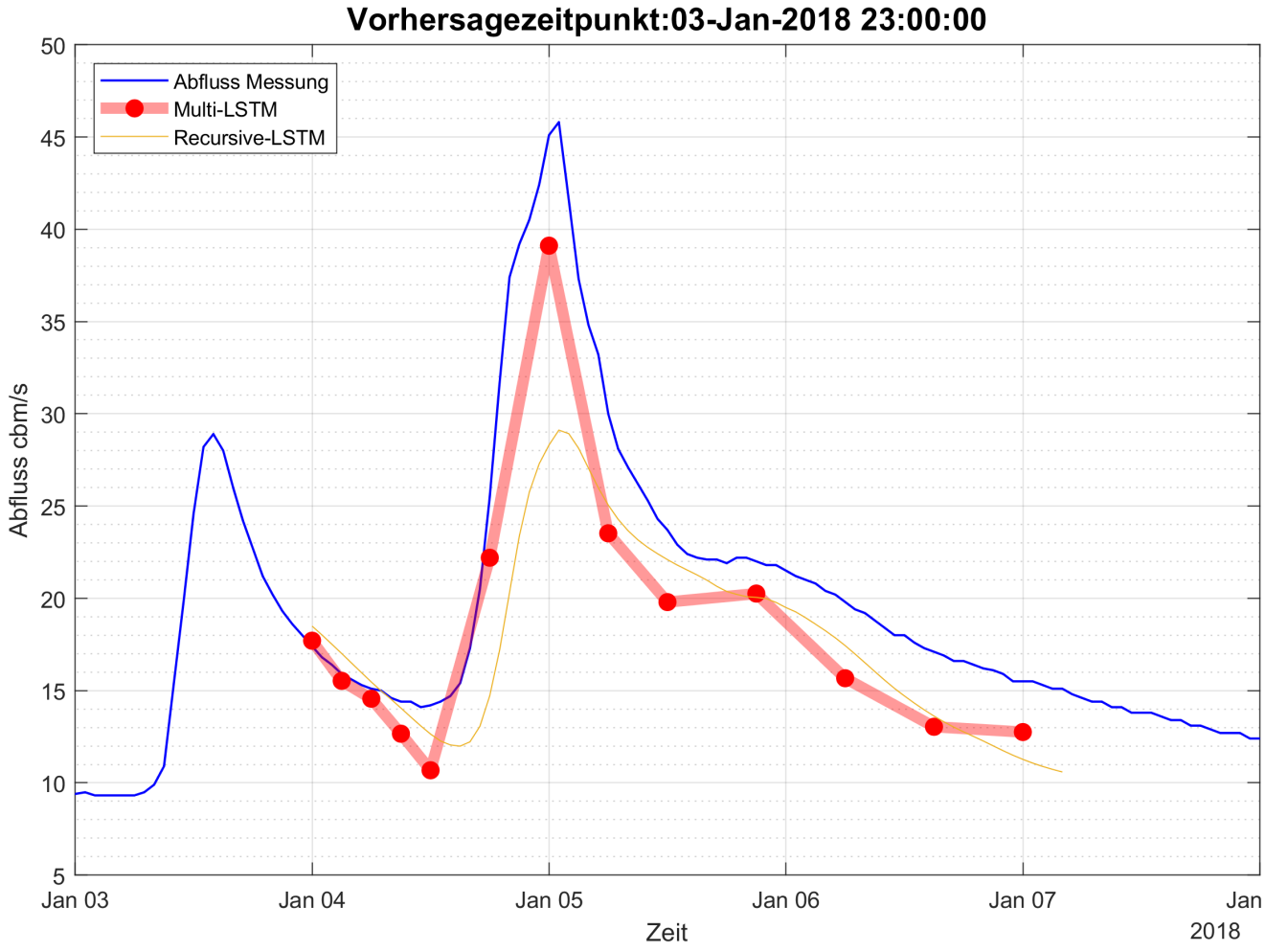

Abbildung 7
Vorhersageganglinien für ein Hochwasserereignis am Pegel Lautenhof/Enz, erstellt mit rekursivem LSTM (gelb) und multi-LSTM (rot, die Punkte markieren die berechneten Vorhersagetiefen, die breite Linie markiert eine lineare Interpolation).
Forecast hydrographs for a flood event at gauge Lautenhof/Enz, produced by recursive LSTM (yellow) and multi-LSTM (red, the dots mark the predicted forecast depths, the wide line is a linear interpolation).

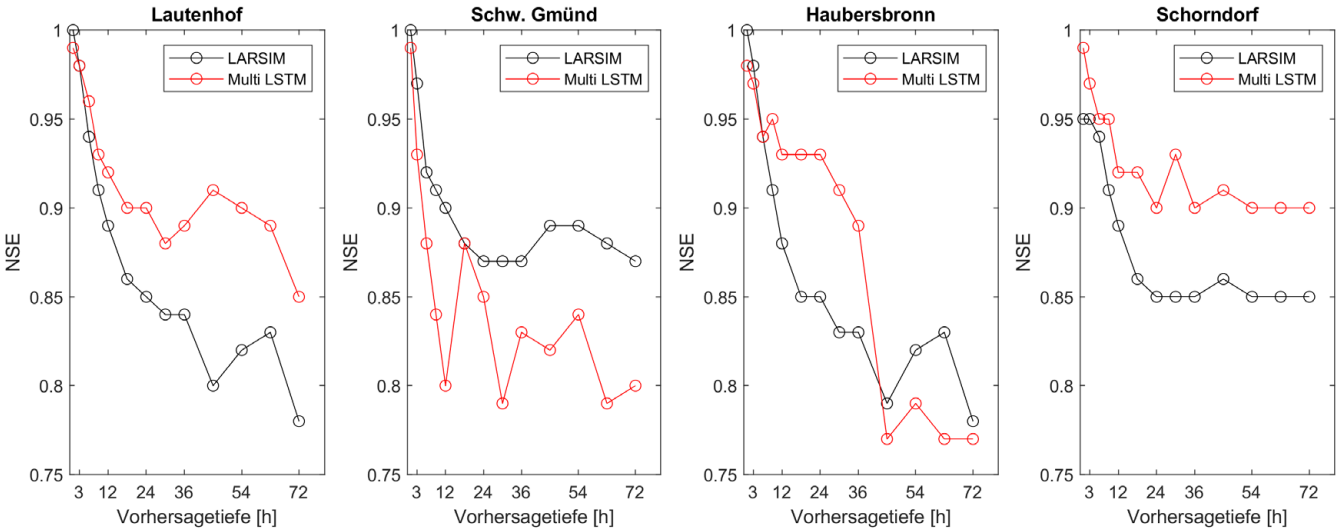

Abbildung 8
Mittlere NSE-Werte von LARSIM und multi-LSTM als Funktion der Vorhersagetiefe für Vorhersagetests von 21 ausgewählten Hochwasserereignissen im Vergleichszeitraum.
Average NSE values for LARSIM and multi-LSTM forecasts as a function of forecast depth for 21 selected flood events in the comparison period.

3.2.2 LARSIM vs. multi-LSTM

Wie in Kapitel 2.4.3 beschrieben, erfolgte der Vergleich von LARSIM und multi-LSTM bezüglich ihrer Vorhersagequalität anhand der 21 ausgewählten Hochwasserereignisse im Vergleichszeitraum, und als Funktion der Vorhersagetiefe. Die Ergebnisse sind in Abbildung 8 dargestellt. Wie zu erwarten nimmt mit zunehmender Vorhersagetiefe die Vorhersagequalität für alle Pegel und beide Modelle ab. Dies liegt nicht an der abnehmenden Qualität der Antriebsdaten, da, wie in Kapitel 2.4 erläutert, auch im Vorhersagezeitraum Messdaten verwendet wurden. Vielmehr hängt es mit dem abnehmenden Informationsgehalt der nur bis zum Vorhersagezeitpunkt verfügbaren Abflussmessungen zusammen, und wie gut jedes Modell den Gebietszustand aus diesen Messungen extrahieren und für die Vorhersage nutzen kann. Die Vorhersagequalität für beide Modelle bleibt selbst für die 72-Stunden-Vorhersage über einem NSE-Wert von 0,75 und kann damit insgesamt als hoch bezeichnet werden. Im direkten Vergleich der Modelle zeigt sich ein insgesamt ausgeglichenes Bild. An den Pegeln Lautenhof und Schorndorf liefert das LSTM bessere Vorhersagen, am Pegel Schwäbisch Gmünd LARSIM, am Pegel Haubersbronn wechselt die Vorhersagequalität mit der Vorhersagetiefe.

Zum weiteren Vergleich von LARSIM und multi-LSTM sind in Abbildung 9 Vorhersagen für das jeweils höchste im Testzeitraum beobachtete Hochwasser an jedem Untersuchungspegel dargestellt, für jeweils drei charakteristische Vorhersagezeitpunkte: vor dem Ereignis, zu Ereignisbeginn, und am Scheitelabfluss. Am Pegel Lautenhof wird die Hochwasserspitze durch LARSIM korrekt vorhergesagt, aber durch das multi-LSTM leicht unterschätzt. Das Ereignis zeigt jedoch, dass beide Modelle auch mehrgipflige Ereignisse vorhersagen können. An den weiteren Pegeln wird der Hochwasserverlauf zu allen Vorhersagezeitpunkten von beiden Modellen im Wesentlichen korrekt vorhergesagt.

4 Zusammenfassung und Ausblick

Das Ziel der vorliegenden Studie war eine Untersuchung des Potenzials maschineller Lernverfahren für die hydrologische Simulation und Vorhersage. Dazu wurden Long Short-Term Memory Netzwerke (LSTMs) genutzt. Sie besitzen im Gegensatz zu einfachen neuronalen Netzwerken ein "Gedächtnis", was sie für die Simulation von Zeitreihen prädestiniert. Für vier exemplarische Pegel in Baden-Württemberg wurden LSTM-Modelle auf Basis hydro-meteorologischer Eingangsgrößen aufgestellt und mehrjährige stündliche Abflusssimulationen und Abfluss-Vorhersagetests auf Basis gemessener meteorologischer Antriebsdaten erstellt. Diese wurden mit Abflussmessungen an den Pegeln sowie mit Abflusssimulationen und -vorhersagen des etablierten prozessbasierten Wasserhaushaltsmodelles LARSIM verglichen.

Bezüglich des Aufbaus von LSTM-Modellen für Simulation und Vorhersage lassen sich die wichtigsten Ergebnisse wie folgt darstellen: Die Simulationsqualität verbessert sich deutlich, wenn man neben den ursprünglichen Antriebsdaten zusätzlich auch aggregierte Zeitreihen verwendet, da darin Informationen über die langfristige Dynamik relevanter Wasserhaushaltsgrößen enthalten sind. Die Wiedergabe von Hochwasserabflüssen verbessert sich wiederum deutlich, wenn man die Standard-Zielfunktion MSE (mittlerer quadratischer Fehler) um einen Term erweitert, der den höchsten aufgetretenen Fehler zusätzlich gewichtet.

Bezüglich der Nutzung von LSTMs für Langfristsimulationen zeigt sich, dass sie sehr gut geeignet und qualitativ mit LARSIM-Simulationen vergleichbar sind: Beide Modelle erreichen – mit Ausnahme von LARSIM an einem Pegel – Nash-Sutcliffe Effizienzen (NSE) von 0,80 oder besser. Bei der Interpretation der Modellgüten ist zu beachten, dass sie mit Daten aus einem Zeitraum berechnet wurden, der für LARSIM einen (kleinen) Teil des Kalibrierungzeitraums umfasst, und für die LSTM-Modelle

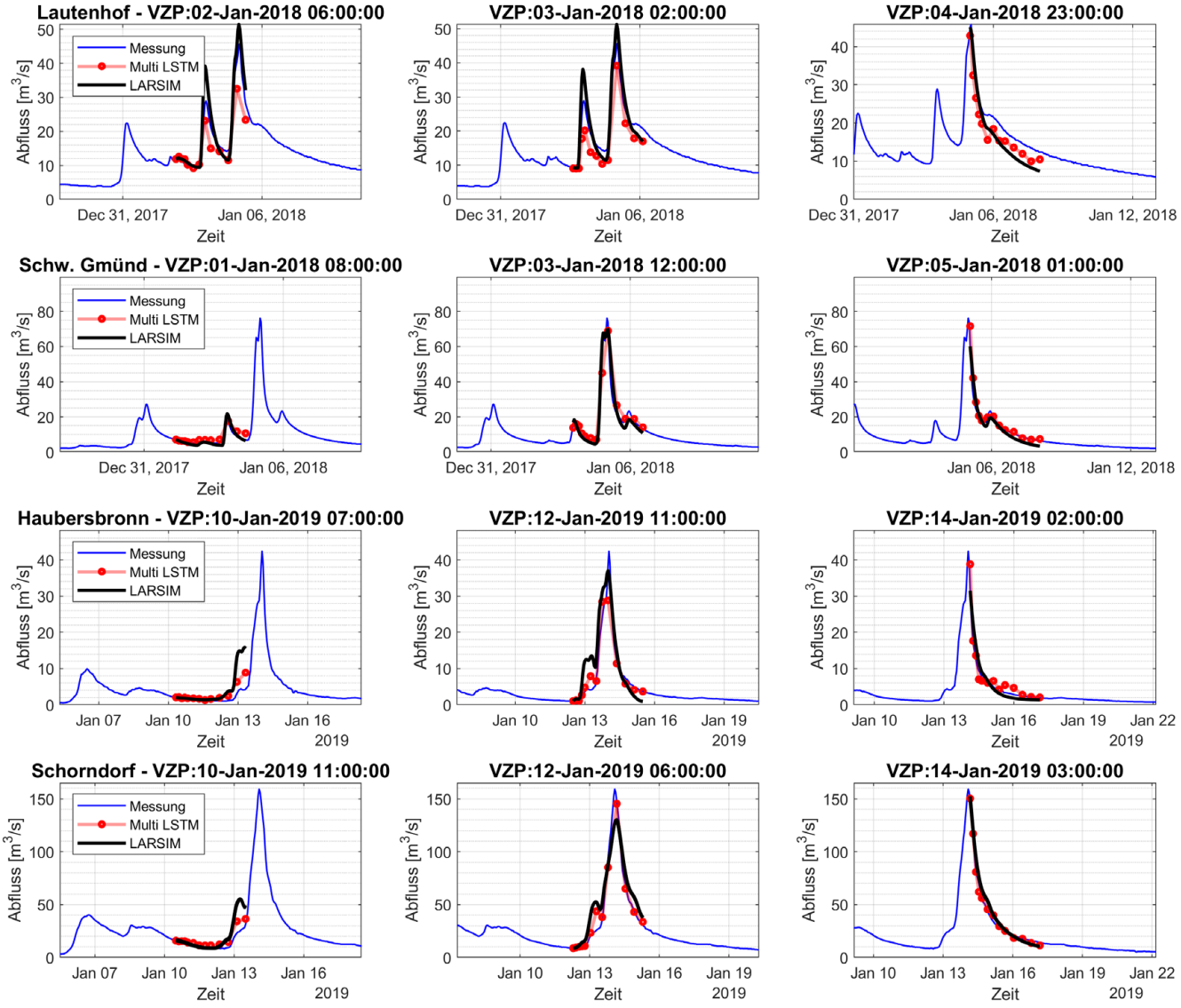

Abbildung 9
Vorhersagen mit LARSIM und multi-LSTM für das jeweils höchste beobachtete Hochwasserereignis an allen Testpegeln (Zeilen). Gezeigt sind 72 Stunden-Vorhersagen zu jeweils drei charakteristischen Vorhersagezeitpunkten (Spalten): vor dem Ereignis, zu Ereignisbeginn, am Scheitelabfluss. Die roten Punkte für das multi-LSTM markieren die berechneten Vorhersagetiefen.
Forecasts by LARSIM and multi-LSTM for the highest observed flood event at each test gauge (rows). 72-hour forecasts at three characteristic forecast times (columns): before the event, at the onset of the event, and at the flood peak. The red points for the multi-LSTM are the calculated forecast depths.

aus Gründen der Vergleichbarkeit mit LARSIM den Zeitraum des Hyperparameter- Trainings. Der Grund dafür liegt darin, dass LARSIM bereits vor dieser Studie umfassend und mit allen verfügbaren Daten für die Hochwasservorhersage kalibriert wurde, und keine weiteren Daten für eine völlig unabhängige Validierung vorlagen. Es ist davon auszugehen, dass die Gütemasse beider Modelle für einen Validierungsdatensatz etwas unter den hier gezeigten liegen. Bewertet man die Modelle nur im Hinblick auf die korrekte Wiedergabe von Hochwasser-Scheitelabflüssen, zeigen auch hier LSTMs und LARSIM gute Übereinstimmungen mit den Messwerten. Bei beiden Modellen zeigt sich eine gewisse Tendenz zur Unterschätzung der Scheitel.

Für die Vorherhersage wurden zwei LSTM-Varianten untersucht: rekursive LSTMs, die ihre Modellausgabe als Eingangsgröße für den nächsten Zeitschritt verwenden, und multi-LSTMs, bei denen für jede Vorhersagetiefe ein separates Modell aufgestellt wird. Die multi-LSTMs stellten sich als die klar bessere Variante heraus, da die bei rekursiven LSTMs auftretende Fehlerfortpflanzung die Vorhersagequalität verschlechtert. Bezüglich der Nutzung für die Vorhersage zeigte sich, dass sowohl LSTMs, als auch LARSIM eine hohe Vorhersagequalität erreichen. Bei der Vorhersage der höchsten Hochwasserereignisse an den Untersuchungspegeln lag der NSE-Wert für beide Modelle selbst für die 72-Stunden-Vorhersage über 0,75. Bei der Interpretation der Ergebnisse ist zu beachten, dass es sich bei den hier durchgeführten Modellläufen um Vorhersagetests mit gemessenen Antriebsdaten handelt. Für Modellläufe mit vorhergesagten Antriebsdaten ist eine niedrigere Vorhersagequalität zu erwarten.

Insgesamt lässt sich schlussfolgern, dass LSTMs in den hier gezeigten Untersuchungen für die Simulation und Vorhersage des Abflusses an Pegeln qualitativ ähnlich gute Ergebnisse erzielen wie das etablierte prozessbasierte Modell LARSIM. Es konnte damit gezeigt werden, dass Methoden des maschinellen Lernens ein großes Potenzial für die hydrologische Simulation und Vorhersage haben.

Um dieses Potenzial weiter auszuloten und auszuschöpfen, bieten sich weitere Untersuchungen an:

A) Verwendung von meteorologischen Vorhersagen für das Training der datenbasierten Modelle. Ausgereifte prozessbasierte Modelle wie LARSIM reagieren bei geeigneter Kalibrierung in der Regel robust auf Unsicherheiten oder systematische Fehler der meteorlogischen Vorhersagen. Inwieweit dies auch auf die hier abgeleiteten multi-LSTMs zutrifft muss noch geklärt werden. Eine Möglichkeit für die bestmögliche Anpassung der multi-LSTMs mit realen meteorologischen Vorhersagen besteht darin, die datenbasierten Modelle mit kombinierten Mess- und Vorhersagedaten zu trainieren, so dass sie systematische Fehler der Wettervorhersage modellintern kompensieren können.

B) In dieser Studie wurde für jedes Untersuchungsgebiet ein separates LSTM aufgestellt. Für große, genestete Einzugsgebiete ist es aber – analog zu prozessbasierten Modellen – denkbar, über das Gewässernetz sequentiell verknüpfte LSTMs aufzustellen, womit ggf. Rechenzeiten verkürzt und die Modellqualität verbessert werden kann.

C) Die Kombination von Prozesswissen und dem Lernen aus Daten in hybriden Modellen wird derzeit als der vielversprechendste Ansatz für die Weiterentwicklung hydrologischer Modelle angesehen (REICHSTEIN et al., 2019; KARPATNE et al., 2017; RAISSI et al., 2019). Hier sind unter anderem folgende Varianten denkbar: Nutzung von ML-Modellen als Postprozessor prozessbasierter Modelle, Nutzung von Zustandsgrößen prozessbasierter Modelle als Eingangsgrößen für ML-Modelle, Trainieren von ML-Modellen als recheneffizientes Surrogat prozessbasierter Modelle, um z. B. Ensemble-Wettervorhersagen effizient nutzen zu können.

Zweifellos sollten die Möglichkeiten von datenbasierten Modellen in der Hydrologie zukünftig weiter untersucht und in die wasserwirtschaftliche Praxis überführt werden. Dabei müssen auch die Grenzen und Probleme dieser Ansätze benannt, analysiert und soweit möglich behoben werden. Insbesondere sollen und können prozessbasierte Modelle nicht vollständig durch datenbasierte Modelle ersetzt werden, zumal viele analytische Fragestellungen mit datenbasierten Modellen (noch) nicht beantwortet werden können. Vielmehr hat sich in den datenbasierten Modellen und insbesondere in LSTM eine wichtige Ergänzung der prozessbasierten hydrologischen Modellierung ergeben, der zukünftig mehr Aufmerksamkeit geschenkt werden sollte.

Conclusions and Outlook

The aim of this study was to investigate the potential of machine learning methods for hydrological simulation and forecasting. Long Short-Term Memory Networks (LSTMs) were used for this purpose. In contrast to simple neural networks, they have a "memory", which makes them ideal candidates for time-series simulations. LSTM models based on hydro-meteorological input variables were set up for four exemplary gauging stations in Baden-Württemberg and applied for multi-year hourly streamflow simulations and forecast tests (72-hour forecasts based on measured meteorological forcing data). These were compared with streamflow measurements at the gauges and with streamflow simulations and forecasts from the process-based water balance model LARSIM.

Regarding the construction of LSTM models for simulation and forecasting, the most important outcomes were: The simulation quality improves significantly if, in addition to the original forcing data, aggregated time series are also used, as they contain information about the long-term dynamics of relevant water balance variables. Further, the representation of flood events is significantly improved when the standard objective function MSE (mean square error) is expanded to include a term that additionally weights the highest error.

Regarding the use of LSTMs for long-term simulations, it was shown that they are very suitable and qualitatively comparable to LARSIM simulations: Both models achieve – with the exception of LARSIM at one catchment – Nash-Sutcliffe efficiencies (NSE) of 0.8 or better. When interpreting the model qualities, it should be noted that the NSE values were calculated with data from a period that includes a (small) part of the calibration period of LARSIM and the period of hyperparameter training for the LSTM models (for comparability reasons with LARSIM). The reason for this is that LARSIM had already been extensively calibrated with all available data for flood forecasting before this study, and no further data was available for a completely independent validation. It can be assumed that the quality measures of both models for a validation dataset are slightly lower than those shown here. If the models are only evaluated regarding the correct representation of flood events, LSTMs and LARSIM also show good agreement with the measured values. In both models there is a tendency to underestimate the peaks.

For short-term forecasting, two LSTM variants were investigated: Recursive LSTMs, which use their model output as an input for the next time step, and multi-LSTMs, where separate models are set up for each forecast depth. The multi-LSTMs were clearly superior, most likely as the error propagation that occurs in the Recursive LSTMs worsens the forecast quality. Considering their usability for forecasting, it was shown that both LSTMs and LARSIM achieve a high forecast quality: When forecasting the highest flood events at the stations, the NSE values for both models were above 0.75, even for the 72-hour forecast. When interpreting these results, it should be kept in mind that the experiments were carried out with measured meteorological forcing data; a somewhat lower prediction quality should be expected under operational conditions, i.e. when using forecasted forcing data.

Overall, it can be concluded that the LSTMs in this study for the simulation and forecasting of streamflow achieve qualitatively similar results than the established process-based model LARSIM. We were thus able to show that machine learning methods have great potential for hydrological simulation and forecasting.

In order to further explore and exploit this potential, further investigation topics are possible: A) Use of meteorological forecasts for training the data-based models. When properly

calibrated, sophisticated process-based models such as LARSIM usually respond robustly to uncertainties or systematic errors in meteorological forecasts. It still needs to be explored to what extent this also applies to the multi-LSTMs. One possibility for an optimal adaptation of multi-LSTMs to real meteorological forecasts is to train them with combined measurement and forecast data so that they can learn and compensate for systematic errors in the weather forecast. B) In this study, separate LSTMs were established for each catchment. For large, nested catchment areas, however, it is conceivable – analogous to process-based models – to set up LSTMs linked sequentially across the river network, which may shorten calculation times and improve the model quality. C) The combination of process knowledge and learning from data in hybrid models is currently viewed as the most promising approach for the further development of hydrological models (REICHSTEIN et al., 2019; KARPATNE et al., 2017; RAISSI et al., 2019). The following variants are conceivable here, amongst others: use of ML models as post-processors of process-based models, using the state variables of process-based models as input variables for ML models, training of ML models as a computationally efficient surrogate of process-based models, for example for efficient ensemble forecasting.

Undoubtedly, the possibilities of data-based models in hydrology should be further investigated in the future and implemented in water management practice. The limitations and problems of these approaches must also be identified, analysed and, if possible, remedied. In particular, in our opinion, process-based models should not and cannot be completely replaced by data-based models, especially since many analytical questions cannot (yet) be answered with data-based models. Rather, we see data-based models and especially LSTMs as an important addition to the process-based hydrological modelling that should receive more attention in the future.

Erklärung zur Datenverfügbarkeit

Die im Rahmen der Studie verwendeten Daten und die erzeugten Ergebnisse können auf ordnungsgemäße Anfrage bei den Verfassern zur Verfügung gestellt werden.

Anschriften der Verfasser

Orhan Delil Tanrikulu, M.Sc.
PD Dr.-Ing. Uwe Ehret
Dr. rer. nat. Ralf Loritz
Karlsruher Institut für Technologie (KIT)
Institut für Wasser und Umwelt
Kaierstraße 12, 76131 Karlsruhe
orhan.tanrikulu@kit.edu
uwe.ehret@kit.edu
ralf.loritz@kit.edu

Dr.-Ing. Ingo Haag
HYDRON Ingenieurgesellschaft für
Umwelt und Wasserwirtschaft mbH
Ritterstraße 9, 76137 Karlsruhe
ingo.haag@hydron-gmbh.de

Dipl.-Ing. Ute Badde
LUBW Landesanstalt für
Umwelt Baden-Württemberg
Griesbachstraße 1, 76185 Karlsruhe
ute.badde@lubw.bwl.de

Literaturverzeichnis

ADDOR, N., A. J. NEWMAN, N. MIZUKAMI & M. P. Clark (2017): The CAMELS data set: catchment attributes and meteorology for large-sample studies, Hydrol. Earth Syst. Sci., 21(10), 5293-5313.

BREMICKER, M. (2000): Das Wasserhaushaltsmodell Larsim – Modellgrundlagen und Anwendungsbeispiele. Freiburger Schriften zur Hydrologie, Institut für Hydrologie, Universität Freiburg, Band 11.

BREMICKER, M., G. BRAHMER, N. DEMUTH, F.-K. HOLLE & I. HAAG (2013): Räumlich hoch aufgelöste LARSIM Wasserhaushaltsmodelle für die Hochwasservorhersage und weitere Anwendung. – KW Korrespondenz Wasserwirtschaft 6 (9), 509–519.

FRAME, J., F. KRATZERT, D. KLOTZ, M. GAUCH, G. SHALEV, O. GILON, L. QUALLS, H. GUPTA & G. NEARING (2022): Deep learning rainfall-runoff predictions of extreme events. Hydrology and Earth System Sciences, 26, 3377–3392.

FROCHTE, J. (2020): Maschinelles Lernen. Grundlagen und Algorithmen in Python. Hanser Verlag. ISBN 978-3-446-46144-4.

GAUCH, M., F. KRATZERT, D. KLOTZ, D. NEARING, J. LIN & S. HOCHREITER (2021): Rainfall-runoff prediction at multiple timescales with a single Long Short-Term Memory network. Hydrology and Earth System Sciences, 25, 2045–2062.

HAAG, I. & A. LUCE (2008): LARSIM-WT: an integrated water-balance and heat-balance model to simulate and predict stream water temperatures. Hydrological Processes 22, 1046-1056.

HAAG, I., M. JOHST, A. SIEBER & M. BREMICKER (2021): Leitfaden zur Kalibrierung von räumlich hoch aufgelösten LARSIM-Wasserhaushaltsmodellen für den operationellen Einsatz in der Hochwasservorhersage. 2. Auflage, Stand: 6. April 2021. Herausgegeben von der LARSIM-Entwicklergemeinschaft. https://larsim.info/.

HAAG, I., J. KRUMM., D. AIGNER, A. STEINBRICH & M. WEILER (2022): Simulation von Hochwasserereignissen in Folge lokaler Starkregen mit dem Wasserhaushaltsmodell LARSIM. Hydrologie & Wasserbewirtschaftung, 66, (1), 6-27, DOI: 10.5675/HyWa_2022.1_1.

HAAG, I., K. TELTSCHER & D. AIGNER. (2023): 2-Grad-Ziel für unsere Bäche – Wassertemperatur und Beschattung. KLIWA-Kurzbericht 2023: 44 S. https://www.kliwa.de/publikationen-kurzberichte.htm.

HEATON, J. (2008): Introduction to neural networks for java, 2nd edition (2nd). Heaton Research, Inc, ISBN 9781604390087.

HEBB, D.O. (1949): The Organization of Behavior: A Neuropsychological Theory, New York:Wiley.

HOCHREITER, S. & J. SCHMIDHUBER (1997): Long Short-Term Memory. Neural Computation, 9 (8), 1735–1780.

HU, C., Q. WU, H. LI, S. JIAN, N. LI & Z. LOU (2018): Deep learning with a long short term memory networks approach for rainfall-runoff simulation. Water, 10 (11), 1543.

HUNT, K.M.R., G.R. MATTHEWS, F. PAPPENBERGER & C. PRUDHOMME (2022): Using a long short-term memory (LSTM) neural network to boost river streamflow forecasts over the western United States, Hydrol. Earth Syst. Sci., 26(21), 5449-5472.

ISHIKAWA, M., I. HAAG, J. KRUMM, K. TELTSCHER & A. LORKE (2021): The effect of stream shading on the inflow characteristics in a downstream reservoir. River Research and Applications, Volume37, Issue7, S. 943-954.

JOSEPH, V.R. (2022), Optimal ratio for data splitting, Statistical Analysis and Data Mining: The ASA Data Science Journal, 15(4), 531-538.

KANG, K.-W., C.-Y. PARK & J.-H. KIM (1993). Neural network and its application to rainfall-runoff forecasting, Korean journal of hydrosciences, 4, 1-9.

KARPATNE, A., G. ATLURI, J.H. FAGHMOUS, M. STEINBACH, A. BANERJEE, A. GANGULY, S. SHEKHAR, N. SAMATOVA & V. KUMAR (2017): Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data, IEEE Transactions on Knowledge and Data Engineering, 29(10), 2318-2331.

KRATZERT, F., D. KLOTZ, C. BRENNER, K. SCHULZ & M. HERRNEGGER (2018): Rainfall–runoff modelling using long short-term memory (LSTM) networks. Hydrology and Earth System Sciences, 22 (11), 6005–6022.

KRATZERT, F., D. KLOTZ, M. HERRNEGGER, A.K. SAMPSON, S. HOCHREITER, & G.S. NEARING (2019): Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning, Water Resources Research, 55(12), 11344-11354.

KRATZERT, F., M. GAUCH, G. NEARING, S. HOCHREITER & D. KLOTZ (2021): Niederschlags-Abfluss-Modellierung mit Long Short-Term Memory (LSTM), Österreichische Wasser- und Abfallwirtschaft, 73(7), 270-280.

LAMB, A., A. GOYAL, Y. ZHANG, S. ZHANG, A. COURVILLE & Y. BENGIO (2016): Professor forcing: A new algorithm for training recurrent networks. https://doi.org/10.48550/arXiv.1610.09038.

LEG – LARSIM Entwicklergemeinschaft (2023): Das Wasserhaushaltsmodell LARSIM – Modellgrundlagen und Anwendungsbeispiele. Stand 6. April 2023. Abrufbar auf http://www.larsim.info/dokumentation/LARSIM-Dokumentation.pdf.

LUCE, A., I. HAAG & M. BREMICKER (2006): Einsatz von Wasserhaushaltsmodellen zur kontinuierlichen Abflussvorhersage in Baden-Württemberg. – Hydrologie und Wasserbewirtschaftung 50 (2), 58–66.

LUDWIG, P., F. EHMELE, M.J. FRANCA, S. MOHR, A. CALDAS-ALVAREZ, J.E. DANIELL, U. EHRET, H. FELDMANN, M. HUNDHAUSEN, P. KNIPPERTZ, K. KÜPFER, M. KUNZ, B. MÜHR, J.G. PINTO, J. QUINTING, A.M. SCHÄFER, F. SEIDEL & C. WISOTZKY (2023): A multi-disciplinary analysis of the exceptional flood event of July 2021 in central Europe – Part 2: Historical context and relation to climate change. Natural Hazards and Earth System Sciences, 23(4), 1287-1311.

MAIER, H.R., S. GALELLI, S. RAZAVI, A. CASTELLETT, A. RIZZOLI, I.N. ATHANASIADIS, M. SÀNCHEZ-MARRÈ, M. ACUTIS, W. WU & G.B. HUMPHREY (2023): Exploding the myths: An introduction to artificial neural networks for prediction and forecasting, Environmental Modelling & Software, 167, 105776.

MCCULLOCH, W.S. & W. PITTS (1943): A logical calculus of the ideas imminent in nervous activity. Bulletin and Mathematical Biophysics 5, 115–133.

MORETTI, G., K. TELTSCHER, J. REGENAUER, J. LIER, M. SEIBERT & I. HAAG. (2022): Nachkalibrierung der LARSIM-Wasserhaushaltsmodelle Baden-Württemberg. HYDRON GmbH im Auftrag der LUBW (unveröffentlicht).

MULVANY, T.J. (1851): On the use of self-registering rain and flood gauges in making observations of the relations of rain fall and flood discharges in a given catchment. Transactions of the Institution of Civil Engineers of Ireland , Vol. IV, pt. II, pp. 18-33, 1851, 4 (2).

NEARING, G.S., D. KLOTZ, J.M. FRAME, M. GAUCH, O. GILON, F. KRATZERT, A.K. SAMPSON, G. SHALEV & S. NEVO (2022): Technical note: Data assimilation and autoregression for using near-real-time streamflow observations in long short-term memory networks, Hydrol. Earth Syst. Sci., 26(21), 5493-5513.

NEVO, S., E. MORIN, A. GERZI ROSENTHAL, A. METZGER, C. BARSHAI, D. WEITZNER, D. VOLOSHIN, F. KRATZERT, G. ELIDAN, G. DROR, G. BEGELMAN, G. NEARING, G. SHALEV, H. NOGA, I. SHAVITT, L. YUKLEA, M. ROYZ, N. GILADI, N. PELED LEVI &,Y. MATIAS (2022): Flood forecasting with machine learning models in an operational framework. Hydrology and Earth System Sciences, 26 (15), 4013–4032.

NEWMAN, A.J., M.P. CLARK, K. SAMPSON, A. WOOD, A., L.E. HAY, A. BOCK, R.J. VIGER, D. BLODGETT, L. BREKKE, J.R. ARNOLD, T. HOPSON & Q. DUAN (2015): Development of a large-sample watershed-scale hydrometeorological dataset for the contiguous USA: dataset characteristics and assessment of regional variability in hydrologic model performance. Hydrol. Earth Syst. Sci., 19, 209-223. DOI:10.5194/hess-19-209-2015.

RAISSI, M., P. PERDIKARIS & G.E. KARNIADAKIS (2019): Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations, Journal of Computational Physics, 378, 686-707.

REICHSTEIN, M., G. CAMPS-VALLS, B. STEVENS, M. JUNG, J. DENZLER, N. CARVALHAIS & PRABHAT (2019): Deep learning and process understanding for data-driven Earth system science, Nature, 566, 195-204, 10.1038/s41586-019-0912-1.

ROSENBLATT, F. (1958): The Perceptron: A probabilistic model for information storage and organization in the brain, Psychological Review, vol. 65, pp. 386–408.

RUMELHART, D.E., G.E. HINTON & R.J. WILLIAMS (1986): Learning internal representations by error propagation. In: Rumelhart, D.E., McClelland, J.L. (Eds.), Parallel Distributed Processing. MIT Press, Cambridge.

SAHOO, B., R. JHA, A. SINGH et al. (2019): Long short-term memory (LSTM) recurrent neural network for low-flow hydrological time series forecasting. Acta Geophys., 67, 1471–1481.

STAHL, K., M. WEILER, D. FREUDIGER, I. KOHN, J. SEIBERT, M. VIS, K. GERLINGER & M. BÖHM (2016): Abflussanteile aus Schnee- und Gletscherschmelze im Rhein und seinen Zuflüssen vor dem Hintergrund des Klimawandels. Abschlussbericht an die Internationale Kommission für die Hydrologie des Rheingebietes (KHR), CHR 00-2016 2016.

THIREL, G., K. GERLINGER, C. PERRIN, G. DROGUE, B. RENARD & J.-P. WAGNER (2019): Quels futurs possibles pour les débits des affluents français du Rhin (Moselle, Sarre, Ill)? La Houille Blanche 5-6: 140–149. https://doi.org/10.1051/lhb/2019039.

WANG, Q., Y. MA, K. ZHAO & Y. TIAN (2022): A comprehensive survey of loss functions in machine learning. Annals of Data Science, 9 (2), 187–212.

XU, T. & F. LIANG (2021): Machine learning for hydrologic sciences: An introductory overview. WIREs Water 8, (5) (e1533).

Nicolas Dalla Valle & Simon Paul Seibert

Praxistransfer: Abflussscheitel-Füllen-Copulas für die Bemessung von Rückhaltebauwerken

From theory to practice: How to use peak discharge and volume copulas for sizing flood control reservoirs

Die Bemessung von Rückhaltebauwerken erfordert Bemessungsganglinien definierter Jährlichkeit, welche meist auf Basis maximaler Abflussscheitelwerte ermittelt werden. Für Rückhaltebauwerke ist aber vor allem auch das Volumen der Ganglinien relevant. Copulas bieten eine vergleichsweise einfache und flexible Möglichkeit, Hochwasserscheitel und -füllen gemeinsam statistisch zu betrachten und die Auswahl geeigneter Bemessungsereignisse zu objektivieren, sie werden aber in der wasserwirtschaftlichen Praxis in Deutschland bisher kaum genutzt. Als Beitrag zum Praxistransfer zeigt dieser Artikel ein mögliches Vorgehen auf, wie Copulas bei der Speicherbemessung angewendet werden können. An einem Beispieldatensatz wird die Ableitung geeigneter Randverteilungen für Scheitel und Volumina und die Auswahl von Copulamodellen diskutiert. Die angepasste Copula wird dann verwendet, um beobachtete und modellierte Ereignisse hinsichtlich der Jährlichkeit ihrer Scheitel-Volumen-Kombination einzuordnen. Da bisher kaum kommerzielle Software für die Umsetzung der vorgestellten Untersuchungen zur Verfügung steht, werden als Anhang für die einzelnen Arbeitsschritte Codebeispiele in der freien Programmiersprache *R* bereitgestellt.

Schlagwörter: Bemessung, Rückhaltebecken, Talsperren, Extremwertstatistik, Copulas Abflussscheitel, Ereignisabflussvolumina, Hochwasserfüllen, NA-Modellierung

Dimensioning of flood protection measures requires design hydrographs of defined return period. These are usually defined by maximum peak flow, although flood volume may be more significant. Copulas are a relatively simple and flexible tool for combined statistical analyses of peak flows and flood volumes. They can be used to rationalise the choice of design floods. However, despite their widespread use in science, copulas are still rarely used for practical applications in Germany. In order to foster the use of copulas this paper provides a hands-on example for dam dimensioning. Using gauge data we discuss the estimation of marginal distributions for peak flow and flood volume and the choice of a suitable copula model. Using the fitted copula we estimate the combined return period of peak flow and flood volume of historical events. Moreover, we use the copula to judge the results from a rainfall-runoff-model. In order to be able to reproduce the evaluations and transfer them to other use cases, executable R-scripts are available as an appendix.

Keywords: Dimensioning, dam structures, extreme value statistic, copula, peak flow, flood volume, rainfall runoff modeling

1 Einleitung

Für die Bemessung von Speichern und Talsperren nach DIN 19700 (DIN 2004a, 2004b) sind Scheitel und Volumina und damit Ganglinien und deren Form von Bemessungshochwassern definierter Auftretenswahrscheinlichkeit erforderlich. Neben den Abflussscheiteln sind dabei Informationen zu den Abflussfüllen und zu Kombinationen möglicher Scheitel und Füllen relevant. Trotz der Bedeutung der Füllen spielen diese in der Bemessungspraxis bis heute oft nur eine nachgeordnete Rolle: Bestehende Empfehlungen (DWA-M 552, 2012) beschränken sich weitgehend auf die Auswertung und statistische Einordnung der Abflussscheitel.

In der Praxis werden die benötigten Ganglinien meist durch Niederschlag-Abfluss-Modellierung (NA-Modellierung) erzeugt, da Messdaten extremer Ereignisse in der Regel fehlen. Allerdings bestehen in der Kalibrierung (und Evaluation) hydrologischer NA-Modelle ähnliche Defizite wie in der Extremwertstatistik: Hier kommen üblicherweise verschiedene Gütekriterien, visuelle Vergleiche simulierter und gemessener Ganglinien (GUPTA et al., 2009; CROCHEMORE et al., 2009) oder Gegenüberstellungen mit Hochwasserabflussquantilen zum Einsatz, eine dezidierte Auswertung und statistische Einordnung simulierter Ereignisabflussvolumina (im Folgenden: "Hochwasserfüllen") wird i. d. R. aber nicht durchgeführt.

Zur statistischen Einordnung simulierter und auch beobachteter Hochwasserfüllen sind zunächst Auswertungen von Hochwasserfüllen und (multivariate) Analysen über mögliche Kombinationen von Scheiteln und Füllen relevant. Die Ermittlung von Auftretenswahrscheinlichkeiten von Hochwasserfüllen erfordert Verfahren zur Abtrennung von relevanten Niederschlag-Abfluss-Ereignissen aus kontinuierlichen Zeitreihen sowie belastbare Vorgehensweisen zur Ermittlung ihrer theoretischen Extremwertverteilung analog dem Vorgehen bei Abflussscheiteln. Für die Bemessung von Speichern sind zudem Extrapolationsverfahren erforderlich, um Ereignisse mit sehr geringer Auftretenswahrscheinlichkeit wie 1.000- oder 10.000-jährliche Ereignisse zu bestimmen. Für die Ermittlung von Scheitelwerten geringer Jährlichkeit haben sich dazu in der Praxis pragmatische Vorgehensweisen wie die Skalierung des HQ_{100} mit einem empirischen Faktor oder die Konvention von KLEEBERG & SCHUMANN (2001) etabliert und werden mangels Alternativen weiter genutzt, obwohl z. B. letztere nicht mehr zur Anwendung empfohlen wird (SCHUMANN & FISCHER, 2023). Auch zur Ableitung extremer Niederschläge existieren entsprechende Verfahren z. B. PEN-LAWA (VERWORN & KUMMER, 2003), FISCHER & SCHUMANN (2018) oder KOUTSOYIANNIS et. al. (1998). Für die Extrapolation von Hochwasserfüllen fehlt bisher allerdings ein etabliertes Vorgehen.

Multivariate Fragestellungen und Herausforderungen bestehen in der wasserwirtschaftlichen Praxis nicht nur bei der Speicher-

bemessung. Auch für die Ermittlung von Lastfallkombinationen beim Aufeinandertreffen von Gewässern z. B. zur Dimensionierung von Schöpfwerken, zur Ermittlung von Planungsgrößen für Hochwasserschutzmaßnahmen oder der statistischen Einordnung von Abflussscheiteln, die sich aus der Überlagerung verschiedener Zuflüsse ergeben, ist die kombinierte Betrachtung verschiedener Einflussgrößen erforderlich. Hier behilft sich die Praxis bisher überwiegend mit sehr pragmatischen Ansätzen wie der Mündungsformel (RP STUTTGART, 2012) oder dem "Quantil-Differenzen-Ansatz" (BENDER, 2015). Diese Vorgehensweisen sind einfach anwendbar, lassen aber das hydrologische Regime der Zuflüsse außer Acht.

Diesen vereinfachten Vorgehensweisen wird in der wissenschaftlichen Fachliteratur das Potenzial von Copulaansätzen ("Copula" lat. binden, verbunden) (SKLAR, 1959, 1997) entgegengehalten, die eine gemeinsame statistische Modellierung von Zufallsvariablen erlauben. Copulas ermöglichen die flexible Konstruktion multivariater Verteilungsfunktionen, da ihre univariaten Randverteilungen unabhängig voneinander und vom gewählten Copulamodell ermittelt werden können, während die Copulafunktion die Abhängigkeitsstrukturen der betrachteten Größen abbildet. Dadurch können mit Copulas auch komplexe, höher dimensionale Abhängigkeitsstrukturen statistisch abgebildet und modelliert werden. In der wissenschaftlichen hydrologischen Fachliteratur finden sich zahlreiche Anwendungsfälle (TOOTOONCHI et al., 2022; SALVADORI et al., 2016) und es bestehen auch schon weiterführende Untersuchungen, z. B. zur Berücksichtigung der angesichts von Landschafts- und Klimawandel gerade für die Bemessung von Rückhaltebauwerken sehr relevanten und in vielen Zeitreihen beobachteten Instationarität (LI et al., 2023; BENDER et al., 2014).

Trotz ihres großen Potenzials sind multivariate Untersuchungen in der hydrologischen Praxis in Deutschland allerdings bisher kaum verbreitet. Anknüpfend an Arbeiten von BENDER (2015) und BENDER et al. (2018) besteht daher das Ziel dieses Artikels darin, Copulaansätze für die Bemessung von Speichern und Rückhaltebauwerken für die Praxis zugänglicher zu machen. Dazu werden anhand eines Beispieldatensatzes in einem ersten Schritt Randverteilungen für Scheitel und Füllen ermittelt und dann verschiedene Copulamodelle getestet. Die angepassten Copulamodelle werden genutzt, um beobachtete und mit NA-Modellen simulierte Hochwasserereignisse bzgl. ihrer Kombination von Abflussscheitel und -volumen statistisch einzuordnen.

Um die fachliche Diskussion der vorgestellten Methode in der Fachgemeinschaft zu stimulieren, ihre Verbreitung in der Bemessungspraxis und die Entwicklung von Standards zu fördern und um dem Mangel an kommerzieller Software entgegenzuwirken, wird im Anhang ein Praxisbeispiel mit lauffähigen, ausführbaren Skripten in der frei verfügbaren Programmiersprache *R* (R CORE TEAM, 2021) bereitgestellt. Die Skripte können auf Anfrage auch über die Autoren bezogen werden.

Methodisch gegliedert ist der Artikel in eine Beschreibung des Datensatzes (Kap. 2.1), die Ermittlung der Randverteilungen von Abflussscheiteln (Kap. 2.3) und Hochwasserfüllen (Kap. 2.4), die Parametrisierung und Evaluation unterschiedlicher Copulas (Kap. 2.5) sowie in eine knappe Erläuterung der Grundlagen zur Anwendung ereignisbasierter NA-Modelle (Kap. 2.6). Ergebnisse und Diskussion finden sich in Kapitel 3, der Artikel schließt mit einem Ausblick.

2 Material und Methoden

2.1 Daten

Als Beispieldatensatz für diesen Beitrag verwenden wir Aufzeichnungen des Gesamtzuflusses zu einem Speicher mit großem alpinem Einzugsgebiet (Fläche ca. 1.140 km^2). Das Gebiet war in den letzten 25 Jahren vermehrt von größeren Hochwassern betroffen (insb. Mai 1999 und August 2005). Für den Speicherzufluss liegen kontinuierliche Abflussbeobachtungen ab dem Abflussjahr (AJ) 1950 vor, ab dem Jahr 1959 stündliche Momentanwerte, vorher Tageshöchstwerte und unregelmäßige Einzelmessungen bei Hochwasser.

Für eine gemeinsame Auswertung der Scheitel und Füllen müssen die mit den Scheitelwerten korrespondierenden Füllen ermittelt werden, also Ereignisabflussvolumina bzw. Hochwasserfüllen. Zur Ableitung wurde automatisiert ausgehend vom Jahreshöchstwert der Teil der Ganglinie horizontal abgetrennt, der über einem Grenzwert von 50 m^3/s liegt. Dieses Vorgehen wird ausführlich in Kapitel 3.1 diskutiert. Um Artefakte in der Abtrennung zu vermeiden, wurden alle resultierenden Ganglinien visuell überprüft. Die resultierenden Jahresserien von Scheitel und Füllen sind grafisch (Abb. 1) und in den Codebeispielen im Anhang hinterlegt.

Das MHQ der Serie liegt bei rund 330 m^3/s, die korrespondierende mittlere Hochwasserfülle (MHV) bei 22,1 Mio. m^3. Die grafische Gegenüberstellung zeigt eine hohe Varianz der einzelnen Stichproben (der Variationskoeffizient der Scheitelwerte beträgt 0,55, der der Füllen 1,04). Dies spiegelt sich auch in einem mittleren Bestimmtheitsmaß von $R^2 = 0{,}76$ und einer Rangkorrelation nach Kendall (τ) von 0,72 wider.

Da beide Reihen lediglich zur Illustration des Vorgehens dienen, wird an dieser Stelle auf eine Diskussion der Datenqualität, Repräsentativität und Homogenität verzichtet und angenommen, dass alle erforderlich statistischen Kriterien erfüllt werden. Die für eine extremwertstatistische Auswertung erforderlichen Kriterien und Methoden für ihre Prüfung werden z. B. in DWA-M 552 (2012) und DWA-M 552 (2024, im Gelbdruck) erläutert.

2.2 Software

Alle statistischen Berechnungen wurden mit der Statistiksoftware R (Version 4.1.1) durchgeführt. Für die Extremwertstatistik wurde das Paket *lmomco* (Version 2.3.7) (ASQUITH, 2021) und für die Erstellung der Copulas das Paket *copula* (Version 1.0-1) (HOFERT et al., 2020) verwendet. Für die Erstellung von Abb. 3 wurde auch das Paket *copBasic* (Version 2.2.3) (ASQUITH, 2024) genutzt. Einige ergänzende Untersuchungen wurden mit dem Paket *VineCopula* (Version 2.5.0) vorgenommen (NAGLER et al., 2023). Um die Anwendung der Methoden zu vereinfachen, werden im Anhang für die einzelnen Bearbeitungsschritte lauffähige Codebeispiele dargelegt, die in *R* ausgeführt werden können.

2.3 Extremwertstatistik für die Hochwasserscheitel

Zur Ermittlung von Hochwasserquantilen werden üblicherweise auf Grundlage einer Stichprobe (Jahresserie) die Parameter einer Extremwertverteilung geschätzt. Da dieses Vorgehen in der Praxis allerdings nicht immer umsetzbar ist, wird nachfolgend eine

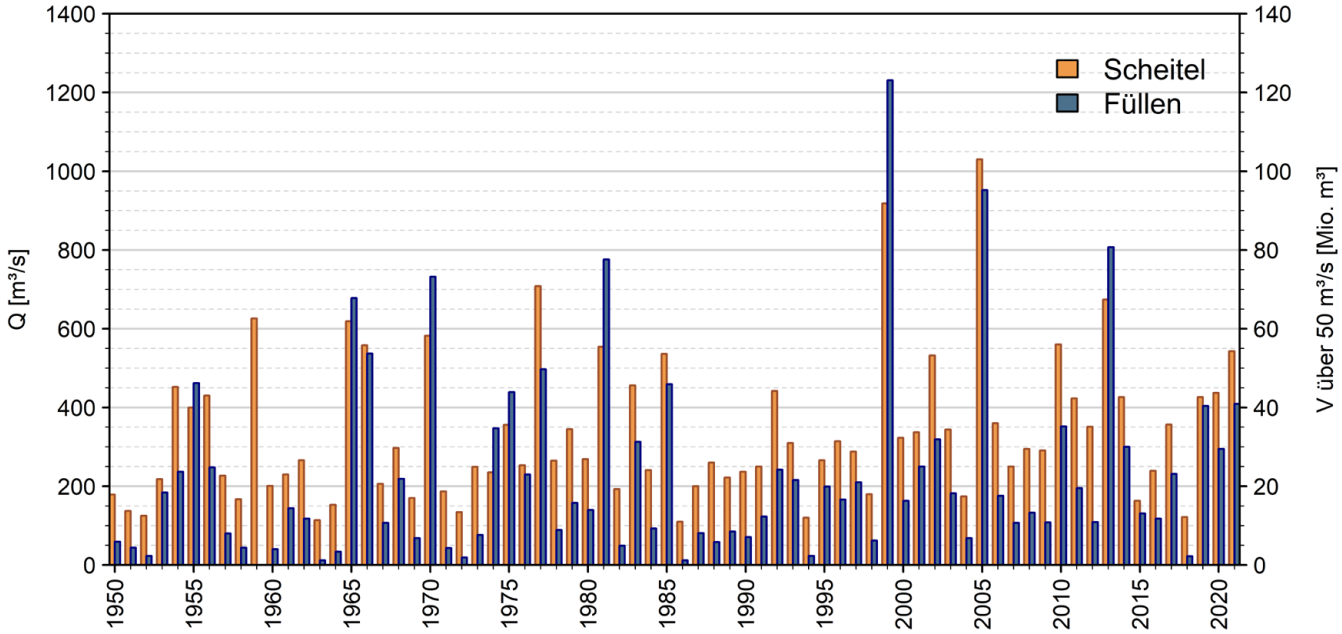

Abbildung 1
Beispieldatensatz: Serie der Jahreshöchstwerte des Abflusses und korrespondierender Ereignisfüllen (Hochwasserfüllen) über einem Abfluss von 50 m^3/s (Abflussjahre 1950 – 2021).
Example data set: Series of annual peak flow maxima und corresponding flood event volumes above 50 m^3/s (hydrological years 1950 – 2021).

Alternative dargelegt: Sie geht davon aus, dass zur Ermittlung einer Randverteilung für die Hochwasserscheitel eine Verallgemeinerte Extremwertverteilung (GEV) durch Parameteroptimierung an bereits vorliegende Hochwasserquantile für die Wiederkehrzeiten 5 a, 10 a, 20 a, 50 a, 100 a, 1.000 a und 10.000 a angepasst werden kann. Dadurch kann eine zu Bestandswerten passende, kontinuierliche Randverteilung erzeugt werden, die auch den Extrembereich bis $HQ_{10.000}$ abdeckt. Zudem erlaubt dieses Vorgehen den in der Praxis üblicherweise vorliegenden methodischen Bruch zwischen der Ermittlung von Quantilen ≤HQ100 und >HQ100 zu überwinden. Vor- und Nachteile dieses Vorgehens werden in Kapitel 3.2 erläutert und diskutiert. Das Ergebnis der Anpassung zeigt Abbildung 2 (links). Ein Codebeispiel für das Vorgehen kann Anhang 1 entnommen werden.

2.4 Extremwertstatistik für die Hochwasserfüllen

Für die extremwertstatistische Auswertung der Hochwasserfüllen wurde zunächst eine Verallgemeinerte Extremwertverteilung (GEV) an die korrespondierenden Füllen angepasst. Dieses Vorgehen führte bei Verwendung des Beispieldatensatzes aber zu nicht plausiblen Füllen für Ereignisse mit sehr geringer Überschreitungswahrscheinlichkeit. Daher wurde für die Ermittlung einer Randverteilung der Füllen der im Rahmen des MUNSTAR-Projekts (JUNGHÄNEL et al., 2022) für die Extremwertstatistik von Niederschlägen herausgearbeitete Ansatz auf Basis von KOUTSOYIANNIS (2004a, b) und KOUTSOYIANNIS et al. (1998) übernommen. Dieses Vorgehen unterstellt eine Heavy-tailed-Verteilung, wie sie für Niederschlagsmengen meist vorliegt, und ist daher vermutlich besser geeignet als der frühere KOSTRA-Ansatz (MALITZ & ERTEL, 2015). Zur Umsetzung wird die Verallgemeinerte Extremwertstatistik (GEV) mit einem vorgegebenen Formparameter von 0,1 auf Basis von L-Momenten an die Jahresserie der Füllen angepasst (zur Verdeutlichung der Unterschiede wird in Abb. 2, rechts zusätzlich eine "reguläre" GEV ohne vorgegebenen Formparameter an die Daten angepasst). Die ermittelte Verteilungsfunktion wird auch für die Abschätzung von 1.000- und 10.000-jährlichen Füllen ($HV_{1.000}$, $HV_{10.000}$) verwendet (Abb. 2). Analog zu den Abflussscheiteln wird so eine kontinuierliche Verteilungsfunktion für den gesamten Wertebereich ermittelt, die als Randverteilung für die Copula verwendet werden kann. Das Vorgehen wird in Kapitel 3.3 diskutiert. Ein Codebeispiel für die Anpassung findet sich in Anhang 2.

2.5 Copulas

In der Literatur ist eine große Zahl an Copulamodellen beschrieben (GENEST & FAVRE, 2007), die sich in ihrer Form und Abhängigkeitsstruktur mitunter stark unterscheiden. Ebenso wie bei den univariaten Verteilungsfunktionen für die Extremwertstatistik gibt es kein theoretisch "bestes" Copulamodell. In hydrologischen Untersuchungen werden häufig einparametrische, archimedische Copulas verwendet wie beispielsweise die Clayton-, Frank- oder Gumbel-Copula (TOOTOONCHI et al., 2022; GENEST & FAVRE, 2007 und CHOWDARY et al., 2011; DE MICHELE et al., 2005; POULIN et al., 2007; SRAJ et al., 2014). Diese unterscheiden sich in der Abhängigkeitsstruktur an den Rändern der multivariaten Verteilungen (Abb. 3). Die Gumbel-Copula unterstellt eine enge Abhängigkeit am oberen Rand, d. h. bei den hohen Wertebereichen der (beiden) Variablen. Hydrologisch ist dies gegeben, wenn sehr hohe Abflussscheitel meist auch sehr hohe Volumina aufweisen. Umgekehrt steht die Clayton-Copula für enge Abhängigkeiten an den unteren Rändern der beiden Verteilungen. Für den Extrembereich bedeutet dies, dass hier sowohl Hochwasserereignisse mit mittlerem Scheitel und sehr hohen Volumina als auch umgekehrt, Ereignisse mit sehr großem Scheitel und mittleren Volumina auftreten können. Die Frank-Copula ist geeignet, wenn keine gewichtete Abhängigkeit besteht. Da der Fokus bei

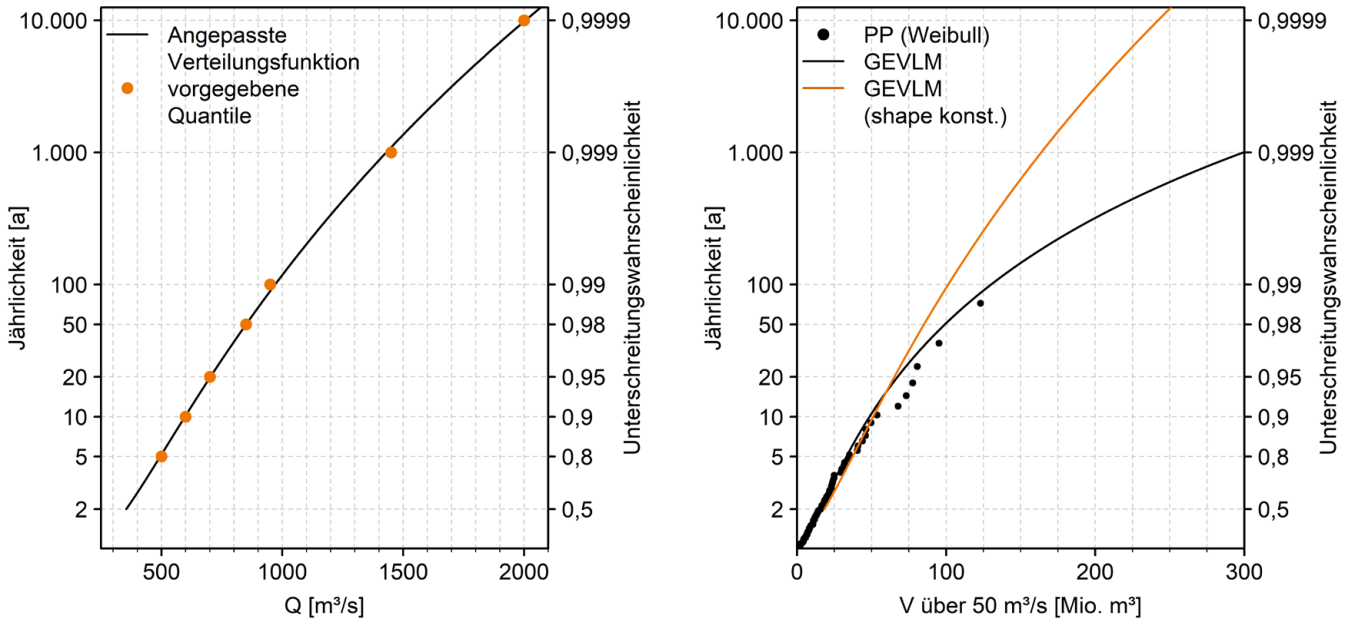

Abbildung 2
Links: Anpassung der Verallgemeinerten Extremwertverteilung (GEV) an bestehende Hochwasserquantile für den betrachteten Pegel (Codebeispiel in Anhang 1). Rechts: An die Jahresserie der Hochwasserfüllen angepasste Verallgemeinerte Extremwertverteilung (GEV) mit fixiertem Formparameter von 0,1 (orange, Codebeispiel in Anhang 2) und mit freier Parameteranpassung (schwarz). Dargestellt sind außerdem die Plotting Positions (PP) der Beobachtungsdaten nach Weibull.
Left: Generalised extreme value distribution (GEV) fitted to given flood quantiles for the examined gauge (code example in appendix 1). Right: Generalised extreme value distribution (GEV) fitted to the annual series of flood volumes with fixed shape parameter of 0.1 (orange, code example in appendix 2) and with free parameter estimation (black). The black dots are the Weibull plotting positions.

der Bemessung auf dem oberen Rand liegt, sollte die gewählte Copulafunktion insbesondere in diesem Bereich eine gute Übereinstimmung zeigen.

Neben den angepassten Copulas können auch multivariate empirische Unterschreitungswahrscheinlichkeiten für die einzelnen Datenpaare (H_i) angegeben werden. Anders als im univariaten Fall sind dafür aber unterschiedliche Definitionen denkbar, da in einem Datenpaar Werte mit abweichenden univariaten empirischen Unterschreitungswahrscheinlichkeiten kombiniert sein können (SALVADORI et al., 2016). In der Copulaliteratur (GENEST & FAVRE, 2007) wird die Berechnung nach FISHER & SWITZER (1985, 2001) empfohlen. Danach ist H für jedes Datenpaar der Quotient aus der Anzahl der Datenpaare, bei denen beide Werte kleiner oder gleich den Werten des betrachteten Datenpaares sind und der Anzahl n-1 der Beobachtungen.

Statistisch lässt sich die Abhängigkeit zweier Größen u. a. durch die Rangkorrelation nach Kendall (τ) ausdrücken. Der Copulaparameter (θ) von z. B. Clayton-, Frank- oder Gumbel-Copula kann direkt und robust aus τ abgeleitet werden, weshalb dieses Vorgehen häufig anderen Möglichkeiten zur Parameterschätzung wie z. B. dem maximum pseudo likelihood-Schätzer (MPL) vorgezogen wurde (GENEST & FAVRE, 2007; BENDER et al., 2013) und auch in dieser Untersuchung verwendet wird.

Da das gewählte Copulamodell eine Annahme über die Abhängigkeitsstruktur der betrachteten Größen vorgibt, kann die Wahl der Copula ähnlich sensitiv wie die Wahl einer Verteilungsfunktion in der univariaten Extremwertstatistik sein und sollte gewissenhaft erfolgen und fachlich begründet werden (POULIN et al., 2007). Letzteres ist in bestehenden Studien leider nicht immer der Fall (TOOTOONCHI et al., 2022; MICHIELS & DE SCHEPPER, 2008). Da sich bis heute keine Empfehlungen auf Basis systematischer Vergleiche verschiedener Copulafamilien für wasserwirtschaftliche Fragestellungen etabliert und Tests anderer Copulafamilien mit dem R-Paket *VineCopula* keinen signifikanten Mehrwert ergeben haben (NAGLER et al., 2023), werden auch in dieser Untersuchung beispielhaft die Frank-, Clayton- und Gumpel-Copula betrachtet.

Zur Beurteilung der Anpassungsgüte der Copula an die Beobachtungsdaten werden zunächst die statistischen Parameter Sn und Rn nach GENEST et al. (2009, 2013) berechnet. Beide Werte basieren auf der Abweichung zwischen den H_i der Beobachtungen und den an die Beobachtungen angepassten Copulamodellen. Ihre Werte sollten daher möglichst klein sein. Beide Verfahren erlauben außerdem die Prüfung der Nullhypothese, dass die H_i einem der angepassten Copulamodelle entstammen. Ein geeignetes Copulamodell sollte deshalb einen p-Wert aufweisen, der nicht zur Ablehnung der Nullhypothese führt (z. B. mehr als 0,05) (GENEST et al., 2013). Die Ermittlung von Sn und Rn sowie die Parametrisierung einer Copula können mit dem Beispielcode in Anhang 3 nachvollzogen werden.

Zudem wurden Kendall-Plots (K-Plots) (GENEST & BOIES, 2003; GENEST & FAVRE, 2007) erstellt. Diese sind von Quantil-Quantil-Plots inspiriert und können als multivariate Erweiterung der Plotting-Position-Plots verstanden werden. Hier werden die H_i der Beobachtungen und die Erwartungswerte der H_i auf Basis

der angepassten Copula gegen die theoretische Unterschreitungswahrscheinlichkeit aufgetragen, die vorliegen würde, wenn die untersuchten Größen unabhängig voneinander wären ($W_{i,n}$). Dabei steht *i* für den Rang der einzelnen Datenpaare und *n* für die Anzahl der Beobachtungen bzw. der verfügbaren Datenpaare. Zur Ermittlung der Erwartungswerte der H_i für die angepassten Copulas können wiederholt Serien von zufälligen Datenpaaren (im dargelegten Beispiel 1.000 Wiederholungen) aus der Copula gezogen und für diese die H_i berechnet werden. Damit das Ergebnis mit den Beobachtungen vergleichbar bleibt, muss die Zahl der Datenpaare pro Serie der Anzahl der Beobachtungspaare entsprechen. Der Erwartungswert der einzelnen H_i ergibt sich dann als Mittelwert aus den Wiederholungen. Liegt der Verlauf der H_i der Beobachtungen (überwiegend) nahe am Verlauf der Werte aus einer angepassten Copula, bildet letztere die Abhängigkeitsstruktur der Randverteilungen gut ab. In Analogie zur üblichen Darstellung bei der Anpassung von Verteilungsfunktionen in der univariaten Statistik (Abb. 2) ist auch bei dieser Darstellung eine Logarithmierung und Angabe von Jährlichkeiten möglich (Abb. 4, rechts), um besonders den Extrembereich betrachten zu können. Ein Beispielcode für die Erstellung eines K-Plots ist in Anhang 4 gegeben.

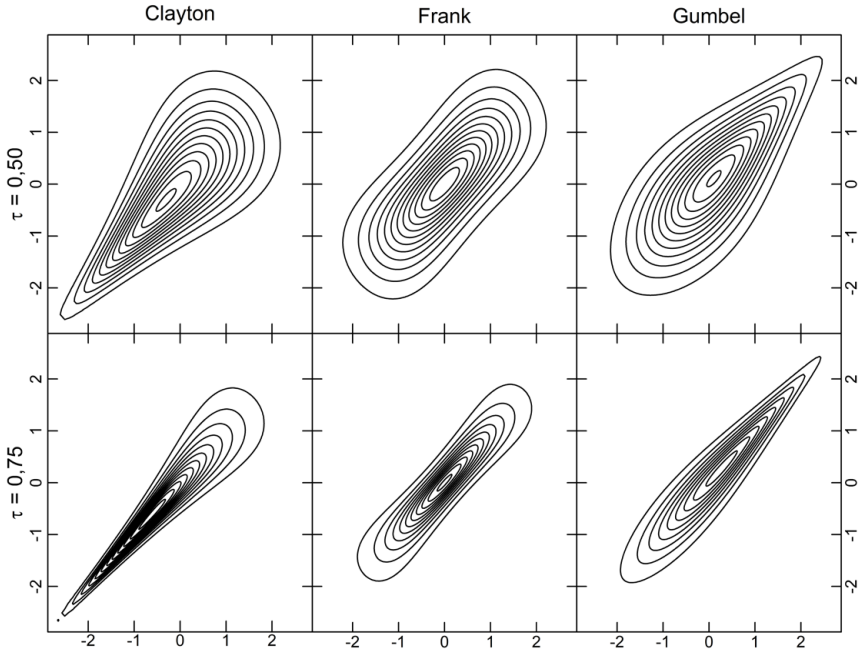

Abbildung 3
Schematische Darstellung der Abhängigkeitsstruktur verschiedener Copulamodelle in Abhängigkeit von der Rangkorrelation nach Kendall (τ) der beiden untersuchten Größen. Für die Darstellung wird für beide Randverteilungen eine Standardnormalverteilung angenommen, deren Wertebereiche auf den Achsen dargestellt sind.
Schematic representation of the dependence structure of different copula models based on Kendall's rank correlation coefficient (τ) of the two analysed variables. A standard normal distribution was assumed for both marginal distributions.

Zuletzt werden Streudiagramme z. B. analog CHOWDHARY et al. (2011) generiert, indem aus den angepassten Copulas durch zufälliges "Ziehen" 100.000 synthetische Datenpaare erzeugt wurden. Diese Kombinationen von Unterschreitungswahrscheinlichkeiten aus Scheitel und Füllen, die auf Basis der jeweiligen Copula gemeinsam auftreten können, lassen sich anhand der in Kapitel 2.3 und 2.4 ermittelten Randverteilungen in den tatsächlichen Wertebereich übertragen. Zusätzlich lassen sich in dem Diagramm Beobachtungsdaten darstellen. Der Vorteil dieser Visualisierung ist, dass die Güte des Gesamtmodells, bestehend aus Randverteilungen und Copula, beurteilt werden kann. Zudem lassen sich im Streudiagramm Isolinien gleicher Unterschreitungswahrscheinlichkeiten bzw. Jährlichkeiten darstellen. Ein Codebeispiel für die Erstellung des Streudiagramms ist in Anhang 5 hinterlegt.

Als letztes Kriterium werden die größten Ereignisse der verfügbaren Stichprobe analysiert (z. B. hinsichtlich Genese, Einfluss von Schneeschmelze, Bedeutung mehrgipfliger Ganglinien, usw.) und das hydrologische Regime des Einzugsgebiets berücksichtigt, um die Auswahl der Copula auch mit physikalischen Argumenten zu stützen.

2.6 Ableitung von Bemessungsganglinien mit Niederschlag-Abfluss-Modellen

Die Bemessung von Speichern erfordert Ganglinien, die beim Fehlen geeigneter Messdaten extremer Ereignisse üblicherweise mit NA-Modellen generiert werden. Für die Ermittlung wird i. d. R. einem inzwischen 40 Jahre alten "Bemessungskonzept" (DVWK Regeln 112, 1982; DVWK Regeln 113, 1984)) gefolgt und ein Ensemble aus Ganglinien generiert. Aus diesem müssen dann plausible Vertreter ausgewählt werden, was häufig schwierig und subjektiv ist. Als Anwendungsbeispiel wird daher zusätzlich gezeigt, wie Copula-Auswertungen verwendet werden können, um Ergebnisse von NA-Modellen hinsichtlich der kombinierten statistischen Unterschreitungswahrscheinlichkeit von Scheitel und Füllen einzelner Ganglinien einzuordnen. Dadurch können begründet aus dem Ensemble für die Bemessung geeignete Ganglinien ausgewählt werden. Zum besseren Verständnis wird im Folgenden kurz das "Bemessungskonzept" erläutert.

Zur Ermittlung eines Ganglinienensembles wird ein NA-Modell mit (ggf. abgeminderten) Bemessungsniederschlägen z. B. KOSTRA (MALITZ & ERTEL, 2015; JUNGHÄNEL et al., 2017; JUNGHÄNEL et al., 2022) oder PEN-LAWA (VERWORN & KUMMER, 2003) unterschiedlicher Dauerstufen und mit einer synthetischen zeitlichen Niederschlagsverteilung durchgerechnet. In der Regel wird dabei angenommen, dass das gesamte Einzugsgebiet gleichzeitig, einheitlich und vollflächig überregnet wird. Aus den Berechnungen ergibt sich dann für jeden Fließquerschnitt ein maximaler Abfluss pro Dauerstufe. Gegebenenfalls wird die Kalibrierung des Modells noch so modifiziert, dass der größte dieser maximalen Abflüsse einem Referenzwert z. B. einem statistisch abgeleiteten Hochwasserquantil entspricht. Die zugehörige Dauerstufe ist die maßgebliche Dauerstufe (für den Scheitel) des Fließquerschnitts. Bei diesem Vorgehen wird von einer Gleichsetzung der Jährlichkeiten von Niederschlagssumme und Abflussscheitel ausgegangen, d. h. im nachkalibrierten Modell erzeugt

ein hundertjährlicher Bemessungsniederschlag an einem Pegel in der maßgeblichen Dauerstufe einen hundertjährlichen Abflussscheitel. Die Gültigkeit dieser Annahme wird dann meist auch für alle anderen Fließquerschnitte im Modell unterstellt. Die korrekte Abbildung der Abflussfülle wird bei diesem Vorgehen nicht berücksichtigt.

Maßgeblich für die Bemessung von Speichern und Talsperren sind i. d. R. die Zuflussganglinien, die im Speicher das größte Rückhaltevolumen erfordern. Häufig handelt es sich dabei nicht um die Ganglinien mit dem größten Zuflussscheitel, sondern um Ganglinien mit einem kleineren Scheitel, aber größerer Dauer und damit höherer Fülle. Während die maßgebliche Dauerstufe für den Scheitel allein von den Eigenschaften des Einzugsgebiets bestimmt wird, hängt die maßgebliche Dauerstufe für die Speicherbemessung zusätzlich von den Eigenschaften und der Steuerung des Bemessungsbauwerks ab.

Für das gewählte Beispiel wurden Zuflussganglinien analog des oben skizzierten Vorgehens mit dem NA-Modell LARSIM (Large Area Runoff Simulation Modell, LUDWIG & BREMICKER, 2006) erzeugt. Das Modell wurde entsprechend bestehender Standards durch Änderung der programminternen Abflussbildungskomponente (BAF-Wert der Abflussbeiwertfunktion) so nachkalibriert, dass bei Niederschlägen einer bestimmten Jährlichkeit an allen relevanten Pegel in der dort maßgeblichen Dauerstufe Abflussscheitel der gleichen Jährlichkeit erreicht werden.

3 Ergebnisse und Diskussion

3.1 Ereignisabgrenzung und Ermittlung der Hochwasserfüllen

In unserem Beispiel verwenden wir mit Jahreshöchstwerten des Abflusses korrespondierende Füllen (Hochwasserfüllen). Ein wesentliches Hemmnis für die Anwendung von Extremwertstatistik für Hochwasserfüllen (und analog für die Anwendung von Scheitel-Füllen-Copulas) besteht bisher in dem Fehlen von Konventionen, Werkzeugen und Erfahrungen in der Ermittlung der Hochwasserfüllen. Anders als für die Scheitelwerte liegen standardmäßig keine Jahresserien korrespondierender Füllen vor. Für die Berechnung der Füllen werden Ganglinien benötigt. In der Praxis führt dies oft zu Einschränkungen, da hochaufgelöste Ganglinien – im Gegensatz zu Scheitelwerten – häufig erst seit wenigen Jahrzehnten vorliegen. Für historische Zeiträume sind häufig gar keine Ganglinien verfügbar oder diese liegen nur in sehr grober zeitlicher Auflösung vor, z. B. als "Tagesmittelwerte", die auf Basis weniger Einzelmessungen abgeleitet wurden.

Für die Abgrenzung von Ereignissen in kontinuierlichen Zeitreihen existiert bis heute kein operationell einsetzbares Verfahren, das unter verschiedensten Bedingungen automatisiert belastbare Ergebnisse liefert (SEIBERT et al., 2016; DWA-M 552, 2024, im Gelbdruck). Für Einzelfallbetrachtungen wie die hier skizzierte Speicherbemessung ist es vertretbar, die Ganglinie vereinfachend bei einem Grenzwert horizontal abzutrennen – insofern das Ergebnis plausibilisiert und ggf. korrigiert wird. Der Grenzwert von 50 m^3/s wurde hier gewählt, weil diese Menge am betrachteten Speicher in allen relevanten Szenarien dauerhaft abgegeben werden kann. Das Volumen unter dem Grenzwert ist daher für die Bemessung der Rückhalteräume im Speicher nicht relevant. Ein derartiger Grenzwert für die Abtrennung der Ganglinie sollte bei der Bemessung von Rückhaltebauwerken immer vorliegen, z. B. der gewünschte Drosselabfluss, sodass sich das Vorgehen auch auf andere Bauwerke übertragen lässt. Der Grenzwert muss so niedrig gewählt werden, dass nicht zu viele Ereignisse aus der Auswertung ausgeschlossen werden, weil ihre Scheitelwerte unter dem Grenzwert liegen. Das kann insbesondere bei der Verwendung von Jahresserien ein Problem sein, da diese häufig auch kleine Scheitelwerte aus hochwasserarmen Jahren beinhalten. Eine Lösung könnte die Verwendung partieller Serien darstellen, da der Grenzwert für die Abgrenzung der Serien dann auch für die Füllenberechnung verwendet werden kann. Weitere Einschränkungen können sich bei der Führung von Sicherheitsnachweisen nach DIN 19700 (DIN 2004a, 2004b) ergeben: Hier muss dann gewährleistet sein, dass der Grenzwert auch die Kriterien der unterschiedlichen Entlastungsszenarien (n-Fall vs. (n-1)-Fall) einhält, die für die verschiedenen Bemessungshochwasser vorgesehen sind.

Die Abgrenzung von Ereignissen über einen konstanten Grenzwert ist für den dargestellten Anwendungsfall hinreichend, im Allgemeinen aber stark vereinfachend. An Pegeln oder in Einzugsgebieten mit ausgeprägtem Jahresgang in der Abflussdynamik oder wenn Hochwasserereignisse unterschiedlicher Genese auftreten sind andere Ereignisabgrenzungsverfahren vorzuziehen (BLUME et al., 2007; TARASOVA et al., 2018). Hier sind weitere Bemühungen wünschenswert, um einen Praxistransfer der vorliegenden Forschungsansätze zu erreichen und Erfahrungen und operationell einsetzbare Werkzeuge verfügbar zu machen.

3.2 Extremwertstatistik für die Hochwasserscheitel

Durch die Anpassung der Verallgemeinerte Extremwertverteilung (GEV) an bestehende Hochwasserquantile wurde für den Beispieldatensatz eine kontinuierliche Randverteilung für die Copula-Auswertung ermittelt. Aus statistischer Perspektive ist dieses Vorgehen nicht korrekt. Stattdessen sollte die Randverteilung durch Anpassung an die Jahreshöchstwerte ermittelt werden. Hier wurde als Beispiel aber bewusst ein abweichendes, alternatives Vorgehen gewählt, das Copula-Auswertungen ermöglicht, die konkordant zu bestehenden Bemessungswerten sind. Es ist damit auch auf den in der Praxis sehr relevanten Fall anwendbar, dass eine Füllenstatistik und Copula-Auswertung zu bestehenden Scheitelquantilen ergänzt werden sollen. Auch ermöglicht diese Vorgehensweise die Ableitung und Anpassung einer stetigen Randverteilung an Bemessungswerte, die mit unterschiedlichen Methoden ermittelt wurden oder gar nicht auf einer extremwertstatistischen Auswertung der Jahresserie beruhen, weil bei der Ermittlung z. B. verschiedene Kriterien der Informationserweiterung berücksichtigt wurden.

Aus langjähriger Erfahrung am Bayerischen Landesamt für Umwelt ist dies in der Anwendungspraxis häufig der Fall und auch konsistent mit bestehenden Praxisempfehlungen (vgl. DWA-M 552 (2012), DWA-M 552 (2024, im Gelbdruck)). Bei der Speicherbemessung gilt dies bspw. für Werte mit sehr geringer Überschreitungswahrscheinlichkeit wie dem $HQ_{1.000}$ oder $HQ_{10.000}$ nahezu immer. Diese werden aufgrund der sonst erforderlichen Extrapolation i. d. R. per Konvention z. B. nach dem inzwischen veralteten (SCHUMANN & FISCHER, 2023) Ansatz von KLEEBERG & SCHUMANN (2001) oder als Vielfaches des HQ_{100} ermittelt.

Mit dem Vorgehen konnte in dem gewählten Beispiel eine gute Anpassung an alle vorgegebenen bemessungsrelevanten Hochwasserquantile erreicht werden. Andererseits wurde für den Bereich häufiger Ereignisse (kleiner ca. HQ5) keine gute Übereinstimmung erzielt, wie z. B. an der Clayton-Copula ersichtlich wird (Abb. 5), deren mit der Randverteilung umgerechnete Punktwolke aus zufälligen Datenpaaren im Bereich kleiner HQ5 gegen die Messwerte verschoben ist. Da für die hier diskutierte Anwendung der Speicherbemessung vor allem besonders seltene Ereignisse relevant sind, ist die Abweichung bei häufigen Ereignissen hier nicht problematisch und vertretbar. Bei der Ableitung der Randverteilungen sollte grundsätzlich auf eine belastbare Abbildung des relevanten Datenbereichs geachtet werden. Nach Möglichkeit sollten immer auf Grundlage einer Stichprobe die Parameter der Verteilung neu geschätzt und dann die Quantile bestimmt werden.

3.3 Extremwertstatistik für die Hochwasserfüllen

Auch für die extremwertstatistische Auswertung der Hochwasserfüllen fehlt es bisher an geeigneten Handreichungen und Erfahrung. Das DWA-Merkblatt zur Ermittlung von Hochwasserwahrscheinlichkeiten DWA-M 552 (2012) bspw. sparte das Thema explizit aus. In der fortgeschriebenen Fassung wird nun für die statistische Auswertung der Füllen ein analoges Vorgehen zur Scheitelstatistik empfohlen (DWA-M 552, 2024, im Gelbdruck). Damit besteht auch für die Auswertung der Füllen die Problematik, dass zwischen einer Vielzahl mathematischer Modelle gewählt werden muss, ohne dass eine Unterscheidung auf Basis physikalischer Kriterien möglich ist. Eine besondere Problematik besteht darüber hinaus – ebenso wie bei der Scheitelstatistik – in der Extrapolation auf Hochwasserfüllen mit sehr geringen Unterschreitungswahrscheinlichkeiten. Hier bestehen nach unserer Kenntnis bisher keine systematischen Untersuchungen und es sind ähnliche Unsicherheiten wie bei der Scheitelstatistik zu erwarten.

Für den gewählten Beispieldatensatz ergaben sich bei der freien Anpassung einer Verteilungsfunktion für extreme Hochwasser unplausible Abflussfüllen, z. B. ein 10.000-jährliches Hochwasservolumen von 650 Mio. m^3, dies entspricht 125 % des mittleren Jahresabflusses oder einem Effektivniederschlag im Einzugsgebiet von nahezu 600 mm, was deutlich über den 72 h-PEN-Werten für das Einzugsgebiet liegt. Es war deshalb notwendig, eine Einschränkung bei der Extrapolation einzuführen.

Da Hochwasserfüllen niederschlagsgetrieben sind und etablierte Alternativen fehlen, erschien eine Übertragung von Ansätzen zur Ermittlung von Niederschlagsmengen mit geringen Überschreitungswahrscheinlichkeiten sinnvoll. Ein möglicher Weg wäre daher das für die Erstellung des KOSTRA2020-Datensatzes verwendete Vorgehen basierend auf KOUTSOYIANNIS (2004a, b) und KOUTSOYIANNIS et al. (1998), bei dem eine Verallgemeinerte Extremwertverteilung (GEV) mit festem Formparameter von 0,1 an die Stichprobe angepasst wird (JUNGHÄNEL et al., 2022). Unter der Annahme, dass extreme Füllen nur bei sehr langen und extremen Niederschlagsereignissen auftreten und bei diesen Ereignissen im Einzugsgebiet ein konstanter (End-)Abflussbeiwert erreicht wird, erscheint die Übertragung des Formparameters für die Abschätzung von Hochwasserfüllen in erster Näherung gerechtfertigt. Die Untersuchungen beziehen sich allerdings nur auf Niederschlagsmengen mit Jährlichkeiten bis 100 a. Da die maximalen Niederschlagsmengen und damit auch die Hochwasserfüllen physikalisch begrenzt sind, ist der feste Formfaktor für größere Jährlichkeiten als ca. 100 a möglicherweise nicht mehr gerechtfertigt.

Die Übertagung des Koutsoyiannis-Ansatzes auf die Hochwasserfüllen des Beispieldatensatzes, wie in Kapitel 2.3 beschrieben, ermöglicht die Anpassung einer Verteilungsfunktion an die Jahresserie der Hochwasserfüllen. Die ermittelte Verteilungsfunktion deckt den gesamten Wertebereich ab und ist als Randverteilung für die Copula-Auswertung geeignet. Diese sehr pragmatische Herangehensweise sollte in der Zukunft jedoch durch systematische Untersuchungen geprüft werden. Analog z. B. zu KLEEBERG & SCHUMANN (2001) werden auch für die Füllenstatistik Konventionen zur Abschätzung extremer Hochwasserfüllen benötigt. Entsprechende Untersuchungen könnten einen wichtigen Beitrag zur wasserwirtschaftlichen Praxis liefern und auch Optionen zur Ableitung von Regionalisierungsprodukten für Hochwasserfüllen eröffnen.

3.4 Anpassung und Auswahl der Copula-Modelle

Zur Ermittlung der am besten geeigneten Copula wurden Gütemaße berechnet, sowie Kendall-Plots (K-Plots) und Copula-Streudiagramme erzeugt. Die Auswertung der Gütemaße (Tab. 1) zeigt, dass die Gumbel-Copula sowohl die kleinsten Sn- als auch Rn-Werte wie auch die größten p-Werte aufweist und damit von den betrachteten am besten abschneidet. Die Frank- und Clayton-Copula weisen bei Rn ähnliche Werte auf, allerdings ist der p-Wert der Clayton-Copula sogar unterhalb des gewählten Grenzwerts von 0,05, so dass als Alternative zur Gumbel-Copula anhand der Gütemaße die Frank-Copula plausibler erscheint. Dort ist Sn auch deutlich geringer als bei der Clayton-Copula. Insgesamt sind die Gütekriterien jedoch schwer interpretierbar, da Erfahrungen über "gute" Wertebereiche fehlen und die Relevanz von Unterschieden auf den Nachkommastellen kaum beurteilt werden kann.

Ähnliches gilt für die Kendall-Plots (K-Plots). Die Grafik zeigt, dass der Verlauf der aus beobachteten Daten generierten H_i

Tabelle 1
Statistische Gütemaße Sn und Rn für die Anpassung der drei untersuchten Copulamodelle an die verfügbaren Daten. Die ermittelten p-Werte basieren auf 100.000 Wiederholungen. Der Copulaparameter wurde durch Inversion von Kendall's τ geschätzt..
Goodness of fit measures Sn and Rn for the three copula models that were fitted to the data. The given p-values are based on 100,000 random samples. The copula parameter was estimated by inversion of Kendall's τ.

Copula	Sn		Rn	
	Wert	p-Wert	Wert	p-Wert
Gumbel	0,011	0,95	0,120	0,33
Frank	0,019	0,27	0,278	0,08
Clayton	0,037	<0,01	0,256	0,01

(schwarze Punkte in Abb. 4, links) am besten zu den H_i der an die Beobachtungen angepassten Gumbel- und Frank-Copulas passt. Allerdings sind die visuellen Unterschiede gering. Zudem stellt sich die Frage, inwiefern die H_i der größten Ereignisse belastbare Entscheidungskriterien im Extrembereich darstellen (das Problem tritt analog auch bei der Interpretation der plotting positions in der univariaten Extremwertstatistik auf). Im unteren Wertebereich bis etwa Wi:n ≤ 0,4 sind die angepassten Copulamodelle graphisch nicht unterscheidbar. Auffällig ist, dass die Abweichungen zur Clayton-Copula am größten sind und diese, analog zu den Ergebnissen der Gütemaße, am ungünstigsten scheint. Die Unterscheidbarkeit der angepassten Copulas im K-Plot erhöht sich deutlich, wenn man in Analogie zu Darstellungen aus der univariaten Extremwertstatistik die y-Achse logarithmiert und die Jährlichkeit aufträgt (Abb. 4, rechts). In dieser Darstellung wird deutlich, dass die Gumbel-Copula für Hochwasser mit großen Wiederkehrzeiten von den betrachteten Copulamodellen die beste Anpassung bietet, während sich die Modelle im Bereich kleiner ca. HQ2 kaum voneinander unterscheiden.

In den Streudiagrammen werden Unterschiede in der Copula-Form, den Isolinien und der statistischen Einordnung historischer Ereignisse besonders deutlich. Bei der Gumbel-Copula (Abb. 5, rechts) nimmt die Streuung der synthetisch generierten Daten (graue Punktwolke) zu den Extremen hin ab, da dort eine hohe Abhängigkeit der beiden Variablen unterstellt wird. Die Isolinien (blaue Linien) gleicher Jährlichkeit beschreiben innerhalb der Punktwolke eine relativ spitze 90°-Kurve und verlaufen davor und danach waagrecht bzw. senkrecht. Aufgrund der gewählten Abhängigkeitsstruktur werden die Ereignisse der Jahre 1999 und 2005 als 100- bis 200-jährliche Ereignisse eingeordnet. Bei der Frank-Copula (Abb. 5, Mitte) ist die Punktdichte entlang der Winkelhalbierenden durchgehend hoch, da keine gewichtete Abhängigkeit unterstellt wird. Die Isolinien der kleinen Jährlichkeiten verlaufen dadurch etwa kreisförmig, wobei sich mit zunehmender Jährlichkeit ein ausgeprägtes Plateau einstellt. Entsprechend wird Lastfällen wie dem 100-jährlichen Ereignis ein sehr weiter Wertebereich zugeordnet. Besonders auffällig ist, dass dem 1999er und dem 2005er Ereignis, im Gegensatz zur Gumbel-Copula eine Jährlichkeit von 500 a bis 1.000 a zugeordnet wird. Die Clayton-Copula (Abb. 5, links) ist sehr ähnlich zur Frank-Copula mit dem Unterschied, dass die Abhängigkeit im kleinen Wertebereich höher und die Streuung der Punktwolke dort geringer ist. Der Verlauf der Isolinien ist ähnlich zur Frank-Copula. Auffällig ist allerdings, dass die kleineren Beobachtungen nicht mehr im Bereich der synthetischen Punktwolke liegen. Dies ist auf die verwendete Randverteilung für die Scheitel zurückzuführen, die im Bereich häufiger Hochwasser keine gute Anpassung an die Beobachtungen geliefert hat (siehe Diskussion in Kap. 3.2). Den historischen Ereignissen werden durch die Clayton-Copula noch größere Jährlichkeiten von mehr als 1.000 a zugeordnet.

Unter Berücksichtigung der verschiedenen Auswertungen wird hier die Gumbel-Copula als beste Vertreterin eingestuft. Ein starker positiver Zusammenhang zwischen Scheitel und Fülle bei seltenen Hochwassern – wie ihn die Gumbel-Copula unterstellt – erscheint aus hydrologischer Sicht in diesem Einzugsgebiet grundsätzlich plausibel. Diese Einschätzung deckt sich auch mit anderen publizierten Studien, in denen ebenfalls die Gumbel-Copula als beste Vertreterin identifiziert wurde z. B. bei

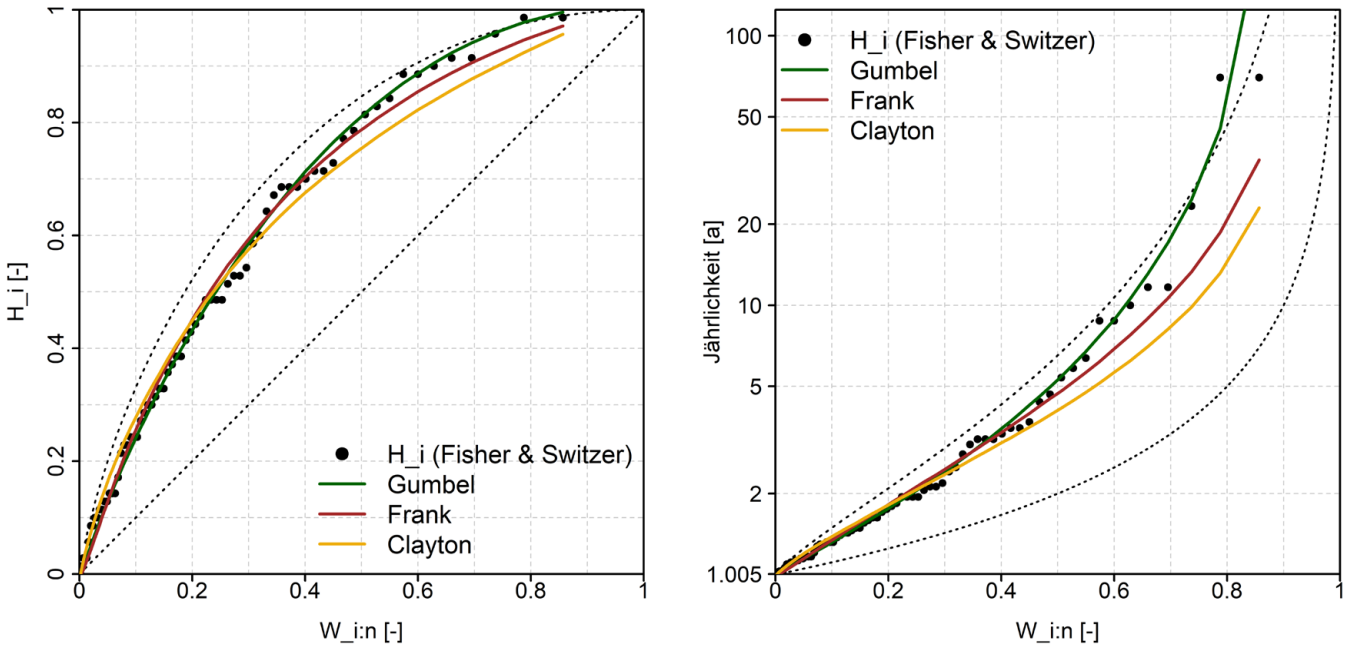

Abbildung 4
Kendall-Plot (K-Plot) der multivariaten empirischen Unterschreitungswahrscheinlichkeiten (H_i) der Beobachtungen und Erwartungswert der H_i aus an die Daten angepassten Gumbel-, Frank- und Clayton-Copulas basierend auf 1.000 Wiederholungen. Die gepunktete Linie beschreibt den theoretischen Verlauf der H_i, wenn die Scheitel und Füllen perfekt korreliert wären, die gestrichelte Linie den Verlauf, wenn beide unabhängig wären. Die rechte Darstellung zeigt den gleichen K-Plot mit logarithmierter y-Achse und Darstellung der Jährlichkeit.
Kendall-Plot (K-Plot) of multivariate empirical non-exceedance probability (H_i) of observation data and expected value of H_i calculated from fitted Gumbel, Frank and Clayton copulas based on 1,000 iterations. The dotted line shows the theoretical curve if peak flow and flood volume were perfectly correlated, the dashed line shows the theoretical curve, if both were independent. The right figure shows the same K-Plot with a logarithmised y-axis and return periods.

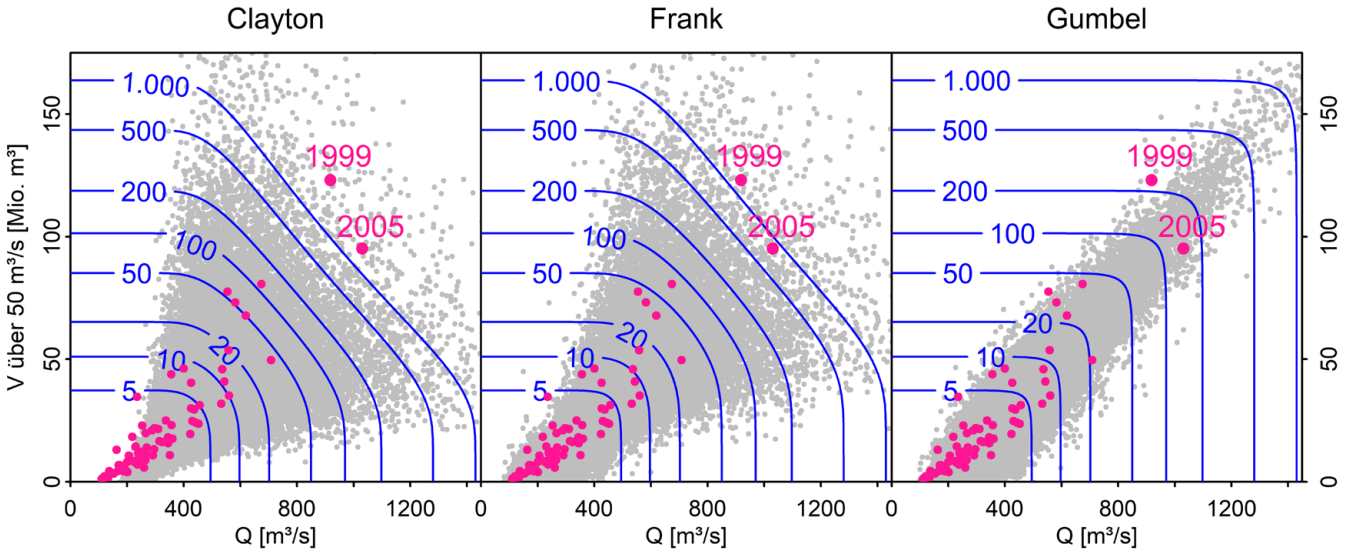

Abbildung 5
Zufällig aus der Copula gezogene und auf Basis von Randverteilungen für die Scheitel und Füllen in den realen Wertebereich umgerechnete Wertepaare für an die Beobachtungen angepasste Clayton-, Frank-, und Gumbel-Copulas (graue Punktwolke, jeweils 100.000 Punkte). Die blauen Linien sind Isolinien gleicher Jährlichkeit. Die pinken Punkte zeigen die Beobachtungen, auf deren Basis die Copula abgeleitet wurde. Die Hochwasser von 1999 und 2005 sind gesondert markiert. R-Code zum Erstellen ähnlicher Abbildungen kann den Anhängen 3 bis 5 entnommen werden.
Randomly sampled data pairs from the fitted Clayton, Frank and Gumbel copulas that were transformed into the real data range using the estimated marginal distributions for peak flow and flood volume (grey point clouds). The blue lines are isolines of similar return period. The pink points show the observations used to derive the marginal distributions and the copula parameter. The 1999 and 2005 floods are highlighted. R code to create similar figures is given in appendices 3 to 5.

Standsicherheitsuntersuchungen an Talsperren (DE MICHELE et al., 2005) oder für bivariate Häufigkeitsanalysen von Scheiteln und Füllen (SRAJ et al., 2014). In der Literatur finden sich aber auch Untersuchungen, in denen anderen Modellen der Vorzug gegeben wurde, z. B. der Clayton-Copula (CHOWDHARY et al., 2011). Viele weitere Copula-Familien sind bspw. auch über das R-Paket *VineCopula* (NAGLER et al., 2023) einfach zugänglich. Testweise Auswertung der dort verfügbaren Copulas haben jedoch keinen relevanten Mehrwert ergeben und nur neue Fragen aufgeworfen, z. B.: Ist der Einsatz zwei-parametrischer Copulas angesichts geringer Stichprobenumfänge gerechtfertigt? Sind radialsymmetrische Korrelationsstrukturen hydrologisch plausibel? – wodurch die Ansätze letztlich wieder verworfen wurden.

Im Hinblick auf die Bemessung zeigt die Auswertung vor allem deutlich, dass die Verwendung einzelner historischer Ereignisse für die Bauwerksdimensionierung sehr einseitig sein kann. Während das Hochwasser von 1999 zwar etwa einen 100-jährlichen Scheitel aufweist, lieferte es eine etwa 200-jährliche Fülle. Das Hochwasser im Jahr 2005 hingegen hatte nach dieser Auswertung einen mehr als 100-jährlichen Scheitel, aber eine weniger als 100-jährliche Fülle. Bemessungen auf Basis historischer Daten können daher stark von der verfügbaren Datengrundlage (Stichprobe) abhängen und der Verlauf der Isolinien zeigt eindrücklich, dass auch zahlreiche andere Kombinationen mit gleicher Auftretenswahrscheinlichkeit möglich sind.

3.5 Einordnung der Modellierungsergebnisse in den statistischen Kontext

Analog zur Einordnung historischer Hochwasser können auf Basis von Copula-Auswertungen auch Ganglinien aus der NA-Modellierung hinsichtlich ihrer kombinierten Jährlichkeit von Scheiteln und Füllen beurteilt werden. Dazu werden die mit dem NA-Modell generierten Kombinationen von Scheiteln und Füllen über dem Grenzwert von 50 m^3/s für die drei häufigsten Bemessungsjährlichkeiten 100 a (grün), 1.000 a (rot) und 10.000 a (gelb) in dem Streudiagramm auf Basis der Gumbel-Copula ergänzt (Abb. 6). Die verschiedenen Punkte kennzeichnen dabei jeweils die Modellergebnisse für unterschiedliche Niederschlagsdauerstufen der entsprechenden Jährlichkeit. Die Position der Datenpunkte hinsichtlich des Scheitels ergibt sich dabei durch die Kalibrierung des Modells auf den HQ_{100}-Scheitel nach dem in Kapitel 2.6 beschriebenen Vorgehen ("Bemessungskonzept"). Mit zunehmender Dauerstufenlänge steigen Abflussscheitel und Volumen, wobei der Verlauf der grünen Punkte bzgl. des Abflussscheitels auf dem Niveau von ca. 1.000 m^3/s sein Maximum (D = 72 h, Punkt B) erreicht (maßgebendes Ereignis). Konzeptionell sollte das verursachende Niederschlagsereignis bzgl. seiner Länge etwa mit dem 1 bis 2-fachen der Konzentrationszeit des Gebietes (hier ca. 72 h bis 96 h) übereinstimmen (DYCK & PESCHKE, 1995) und der Scheitel nahe an der Isolinie der entsprechenden Jährlichkeit liegen. Letzteres ist hier nicht der Fall: Das Ereignis ist hinsichtlich seiner Kombination von Scheitel- und Fülle eher vergleichbar mit dem Hochwasser von 1999, dem die Copulaauswertung eine kombinierte Jährlichkeit von etwa 200 a zuordnet. Das simulierte 48 h-Ereignis (Punkt A) liegt näher an der Isolinie und entspricht einem 50 bis 100-jährlichen Ereignis. Für die Simulation der 1.000- (Punkte C und D) und 10.000-jährlichen Ereignisse (Punkte E und F) zeigt sich ein ähnliches Bild. Keines der mit den Annahmen des "Bemessungskonzepts" simulierten Ereignisse liefert demnach eine plausible Scheitel-Füllen-Kombination, die exakt der geforderten Jährlichkeit entspricht.

Da die Copula-Auswertung einer bedeutenden Unsicherheit unterliegt, können prinzipiell auch abseits der Isolinien liegende Ereignisse als Bemessungsereignisse geeignet sein. Diese Art der Auswertung kann aber verwendet werden, um die Wahl geeigneter Ganglinien deutlich einzuschränken.

Dies ist insbesondere im Hinblick auf die für die Bemessung von Speichern wichtige Fülle relevant, da NA-Modelle üblicherweise nur auf die Scheitel kalibriert werden, Volumenstatistiken weitgehend fehlen und sich aus dem Anstieg der (KOSTRA-) Niederschlagshöhe mit zunehmender Dauerstufe unweigerlich kontinuierlich höhere Abflussvolumina ergeben. In dem dargelegten Beispiel veranschaulichen die Isolinien eindrucksvoll, dass vom Modell jenseits des Abflussscheitelmaximums in allen betrachteten Jährlichkeiten in den längeren Dauerstufen Füllen mit erheblich größeren Jährlichkeiten erreicht werden und analoge Muster und Verläufe wurden in internen Auswertungen auch bei anderen Daten, Speichern und NA-Modellen gefunden. Es bildet sich beim Abflussvolumen, anders als beim Abflussscheitel kein lokales Maximum und das in der Praxis häufig angewandte Vorgehen, alle verfügbaren Dauerstufen zu betrachten, würde eine erhebliche Überdimensionierung des zu bemessenden Bauwerks bedeuten.

Das gezeigte Beispiel verdeutlicht das große Potenzial von Copula-Auswertungen für die Praxis: Sie erlauben die Identifizierung von z. B. 100-jährlichen Kombinationen von Scheiteln und Füllen und im Umkehrschluss den Ausschluss nicht plausibler Lastfälle. Plausible Fälle sind durch Datenpunkte gekennzeichnet, die im "kritischen" Bereich der Isolinien liegen. Gemeint ist damit der gekrümmte Teil, in dem die Isolinien innerhalb der Punktwolke und nicht vertikal oder senkrecht verlaufen (Abb. 6). Im vorliegenden Beispiel werden die Füllen der entsprechenden Jährlichkeiten in ähnlichen Dauerstufen erreicht wie die Scheitel. Dadurch fällt der "kritische" Bereich der Isolinien klein aus, es sind aber verschiedene Kombinationen möglich. Der "kritische" Bereich wird allerdings auch von der Copulaformulierung bestimmt und ist z. B. bei Frank-Copulas erheblich größer als bei der Gumbel-Copula.

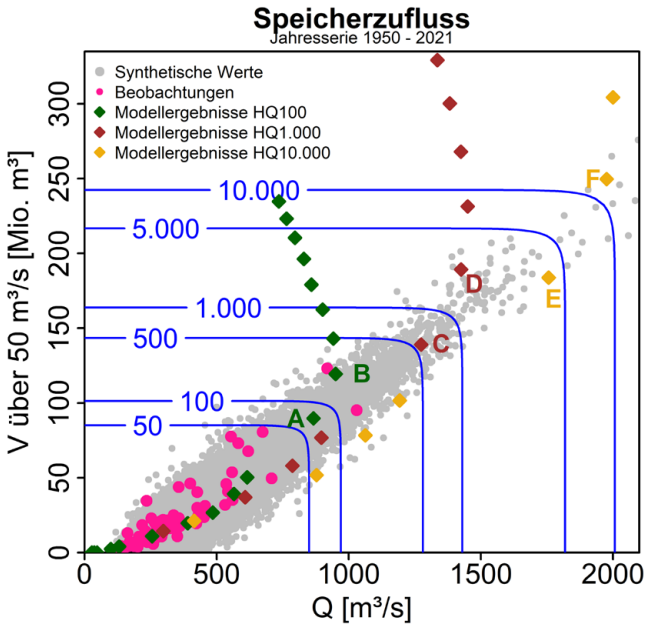

Abbildung 6
Copula-Streudiagramm aus Abbildung 5 rechts (Gumbel) in dem zusätzlich die Ergebnisse der Niederschlag-Abfluss-Simulation dargestellt sind. Die einzelnen Punkte repräsentieren die verschiedenen Dauerstufen der einzelnen Jährlichkeiten. Einzelne Ereignisse sind besonders markiert: A: 100 a, 48 h; B: 100 a, 72 h; C: 1.000 a, 48 h; 1.000 a, 72 h; E: 10.000 a, 48 h; F: 10.000 a, 72 h.
Copula scatter plot from figure 5 right (Gumbel) including the rainfall runoff simulations. The points represent different precipitation durations of the same return period. Some events are marked separately: A: 100 a, 48 h, B: 100 a, 72 h; C: 1,000 a, 48 h; 1,000 a, 72 h; E: 10,000 a, 48 h; F: 10,000 a, 72 h.

Die Auswertung zeigt auch, dass die ohnehin problematischen Annahmen des "Bemessungskonzepts", wie z. B. die Gleichsetzung der Jährlichkeiten von Niederschlag und Abfluss, aber auch die synthetischen Niederschlagsverläufe insbesondere bei der Simulation sehr langer Dauerstufen zu Ergebnissen führen können, die nicht mehr plausibel und für die Bemessung ungeeignet sind. Eine Diskussion und Fortschreibung des "Bemessungskonzeptes" erscheint uns daher sinnvoll und dringlich.

Zusammenfassung und Schlussfolgerungen

Die Bemessung von Rückhaltebauwerken erfordert Ganglinien, wobei nicht nur die Jährlichkeit des Abflussscheitels, sondern auch die der Fülle relevant ist. Über die statistischen Eigenschaften von Hochwasserfüllen ist – im Gegensatz zu den Scheitelwerten – jedoch meist nichts bekannt. In der Konsequenz ist die Auswahl von Ganglinien für Bemessungsaufgaben schwierig und subjektiv. Hinzu kommt, dass Scheitel und Volumen der Ganglinie nicht unabhängig voneinander sind, und deshalb bei der Bemessung geeignete Kombinationen beider Größen benötigt werden.

Zur gemeinsamen statistischen Betrachtung von Scheitel und Volumen bieten sich Auswertungen auf Basis von Copulas an. Diese erlauben, im Grunde analog zur univariaten Extremwertstatistik, die multivariate Betrachtung verschiedener Variablen. Sie sind heute in der hydrologischen Forschung, aber noch nicht in der wasserwirtschaftlichen Praxis etabliert. Letzteres liegt auch daran, dass bisher Handreichungen für die praktische Umsetzung fehlen und kaum kommerzielle Softwareprodukte verfügbar sind.

Als Beitrag zum Praxistransfer der Methodik zeigt der Artikel, wie Copulaansätze für die Bemessung von Speichern und Rückhaltebauwerken praktisch nutzbar gemacht werden können. Dazu werden anhand eines Musterdatensatzes die Randverteilungen für Scheitel und Füllen abgeleitet, verschiedene Copulamodelle verglichen und historische Hochwasserereignisse des Datensatzes statistisch eingeordnet. Als zusätzliches Anwendungsbeispiel wird gezeigt, wie NA-Modellsimulationen für eine Speicherbemessung hinsichtlich ihrer statistischen Eigenschaften eingeordnet und Bemessungsganglinien begründet ausgewählt bzw. ausgeschlossen werden können.

Um die fachliche Diskussion der vorgestellten Methode in der deutschsprachigen Fachgemeinschaft zu stimulieren und ihre Verbreitung in der Bemessungspraxis und die Entwicklung von Standards zu fördern, beinhaltet der Artikel zudem ein Praxisbeispiel mit Testdaten. Der Anhang enthält lauffähige, ausführbare

Skripte in der frei verfügbaren Programmiersprache *R*, die zum Nachvollziehen des Vorgehens oder als Blaupause für ähnliche Fragestellungen verwendet werden können.

Da sich analog zur univariaten Extremwertstatistik die Wahl des Copulamodells sensitiv auf das Ergebnis auswirken kann, werden verschiedene Ansätze zur Auswahl von Copulas und der Beurteilung ihrer Anpassung vorgestellt und diskutiert. Unter der Berücksichtigung von sowohl statistischen, als auch hydrologischen Kriterien wird in dem vorgestellten Fallbeispiel die Gumbel-Copula als beste Vertreterin ausgewählt. Neu eingeführt werden Copula-Streudiagrammen mit Isolinien gleicher Jährlichkeit. Mit diesen lassen sich nicht nur historische Ereignisse einordnen, sondern auch Simulationen hydrologischer Modelle und im Umkehrschluss nicht plausible Lastfälle ausschließen. Der Einsatz von Copulas bietet daher vielfältige und erhebliche Potenziale mit großem Mehrwert für die Anwendungspraxis.

Herausfordernd ist zum gegenwärtigen Zeitpunkt noch die Ableitung von Serien von Hochwasserfüllen. Für Einzelfälle wie die Bemessung von Rückhaltebauwerken konnten mit einem festen Schwellenwert zur Ereignisabgrenzung vereinfachend plausible Ergebnisse erzielt werden. Eine breitere, automatisierte Anwendung, z. B. für viele Pegel ist damit gegenwärtig aber bedauerlicherweise noch nicht möglich. Ursächlich hierfür sind jedoch nicht die Copulas, sondern das Fehlen belastbarer Techniken und praxistauglicher Werkzeuge zur Ereignisabgrenzung. Diesbezüglich besteht entsprechender Handlungs- und Entwicklungsbedarf. Gleiches gilt für die Etablierung wissenschaftlich anerkannter Standards für die Füllenstatistik.

Perspektivisch ist beabsichtigt Copulas auch für andere Fragestellungen zu testen und anzuwenden, z. B. zur Ermittlung von Überlagerungswahrscheinlichkeiten von Wellen an Flussmündungen, der Bemessung von Schöpfwerken oder den Einsatz der Methode zur Erzeugung synthetischer Ganglinien. Da sich Copula-Ansätze leicht auf höherdimensionale Ebenen übertragen lassen, können damit auch weitere Gebiets- oder Einzugsgebietseigenschaften wie Wellenanlaufzeiten, die Vorfeuchte im Gebiet oder die Jahreszeit berücksichtigt werden. Auch der Einsatz von Copulas zur Kalibrierung oder Validierung hydrologischer Modelle für Bemessungsaufgaben erscheint vielversprechend.

Summary and Conclusions

The design of retention structures does not only require discharge values with defined return periods but also flood volumes with defined statistical properties. The latter are typically unknown as available methods and recommendations focus on peak flow. In consequence, the selection of design hydrographs is often difficult and subjective. For instance, the inflow hydrograph with the highest peak may not necessarily represent the critical case for the design of retention volumes and vice versa. As peak and volume of the hydrograph are dependent, appropriate combinations of both parameters are vital for robust design purposes. Copulas offer techniques to jointly consider both variables. Such multivariate analyses are well established in hydrological research, although not yet adopted in water engineering practice.

One goal of the study was to demonstrate how to apply a copula approach to real-world data. We achieved this by deriving marginal distributions for annual series of peak flow and flood volumes using a sample data set and by fitting and comparing different copula models. Additionally, we used the results to assess historical flood events und to evaluate the results of a rainfall-runoff model.

To stimulate discussions about the use of copulas within the community, to foster their integration into design practices, and to promote the development of standards, the article also includes a practical example with test data and an appendix with executable scripts in the freely available programming language R.

Erklärung zur Datenverfügbarkeit

Alle während dieser Studie erzeugten oder analysierten Daten sind in diesem publizierten Artikel und in den dazugehörigen ergänzenden Informationsdateien enthalten.

Danksagung

Unser Dank gilt den zwei anonymen Gutachtern und Editoren, die durch ihre Hinweise ganz wesentlich zur Verbesserung des Artikels beigetragen haben.

Anschriften der Verfasser

Nicolas Dalla Valle
Dr. Simon Paul Seibert
Bayerisches Landesamt für Umwelt
Bürgermeister-Ulrich-Str. 160, 86179 Augsburg
nicolas.dallavalle@lfu.bayern.de
simon.seibert@lfu.bayern.de

Literaturverzeichnis

ASQUITH, W. (2021): lmomco. L-Moments, Censored L-Moments, Trimmed L-Moments, L-Comoments, and Many Distributions. Version 2.3.7. R-Paket.

ASQUITH, W. (2024): General Bivariate Copula Theory and Many Utility Functions. Version 2.2.3. R-Paket.

BELZILE, L. (2023): Liouville Copulas. Version 1.0.6. R-Paket.

BENDER, J. (2015): Zur Ermittlung von hydrologischen Bemessungsgrößen an Flussmündungen mit Verfahren der multivariaten Statistik: Universität Siegen (Mitteilungen des Forschungsinstituts Wasser und Umwelt der Universität Siegen, Bd. 9).

BENDER, J., T. WAHL, C. MUDERSBACH & J. JENSEN (2013): Flood Frequency Analysis at River Confluences – Univariate vs. Multivariate Extreme Value Statistics. Proceedings of the International Conference on Water Resources and Environment Research (ICWRER) 2013, 316–329.

BENDER, J., T. WAHL & J. JENSEN (2014): Multivariate design in the presence of non-stationarity. Journal of Hydrology, 514, 123–130.

BENDER, J., J. JENSEN, C. MUDERSBACH, B. KLEIN & B. ROTHE (2018): Multivariate Wahrscheinlichkeiten: ein Mehrgewinn – nicht nur für die Wissenschaft. Korrespondenz Wasserwirtschaft, 11(3): 160–165.

BLUME, T., E. ZEHE & A. BRONSTERT (2007): Rainfall-runoff response, event-based runoff coefficients and hydrograph separation. Hydrological Sciences Journal 52(5): 843–862.

CHOWDHARY, H., L.A. ESCOBAR & V.P. SINGH (2011): Identification of suitable copulas for bivariate frequency analysis of flood peak and flood volume data. Hydrology Research, 42(2-3): 193–216.

CROCHEMORE, L., C. PERRIN, V. ANDRÉASSIAN, U. EHRET, S.P. SEIBERT, S. GRIMALDI et al. (2014): Comparing expert judgement and numerical criteria for hydrograph evaluation. Hydrological Sciences Journal 60(3): 402–423.

DE MICHELE, C., G. SALVADORI, M. CANOSSI & A. PETACCIA (2005): Bivariate statistical approach to check adequacy of dam spillway. Journal of Hydrologic Engineering, 10(1): 50–57.

DIN (Deutsches Institut für Normung e. V.) (2004a): DIN 19700-10. Stauanlagen. Teil 10: Gemeinsame Festlegungen. Berlin:Beuth.

DIN (Deutsches Institut für Normung e. V.) (2004b): DIN 19700-11. Stauanlagen. Teil 11: Talsperren. Berlin:Beuth.

DVWK REGELN 112 (1982): Arbeitsanleitung zur Anwendung von Niederschlag-Abfluß-Modellen in kleinen Einzugsgebieten. Teil I: Analyse. Deutscher Verband für Wasserwirtschaft und Kulturbau e. V., Paul Parey: Hamburg, Berlin.

DVWK REGELN 113 (1984): Arbeitsanleitung zur Anwendung von Niederschlag-Abfluß-Modellen in kleinen Einzugsgebieten. Teil II: Synthese. Deutscher Verband für Wasserwirtschaft und Kulturbau e. V., Paul Parey: Hamburg, Berlin.

DWA-M 552 (2012): Merkblatt DWA-M 552: Ermittlung von Hochwasserwahrscheinlichkeiten. Deutsche Vereinigung für Wasserwirtschaft, Abwasser und Abfall e. V., Hennef.

DWA-M 552 (2024, im Gelbdruck): Merkblatt DWA-M 552.: Stochastische und deterministische Wege zur Ermittlung von Hochwasserwahrscheinlichkeiten. Deutsche Vereinigung für Wasserwirtschaft, Abwasser und Abfall e. V., Hennef.

DYCK, S. & G. PESCHKE (1995): Grundlagen der Hydrologie. Berlin: Verlag für Bauwesen.

FISCHER, S. & A.H. SCHUMANN (2018): Berücksichtigung von Starkregen in der Niederschlagsstatistik. Hydrologie & Wasserbewirtschaftung, 62(4): 248–256.

FISHER, N.I. & P. SWITZER (1985): Chi-plots for assessing dependence. Biometrika, 72(2): 253–256.

FISHER, N.I. & P. SWITZER (2001): Graphical assessment of dependence: Is a picture worth a 100 tests? The American Statistician, 55(3): 233–239.

GENEST, C. & J.-C. BOIES (2003): Detecting independence with Kendall plots. The American Statistician, 57(4): 275–284.

GENEST, C. & A.-C. FAVRE (2007): Everything you always wanted to know about Copula Modeling but were afraid to ask. Journal of Hydrological Engineering, 12(4): 347–368.

GENEST, C., B. RÉMILLARD & D. BEAUDOIN (2009): Goodness-of-fit-tests for copulas: A review and a power study. Insurance: Mathematics and Economics, 44(2): 199–214.

GENEST, C., W. HUANG & J-M. DUFOUR (2013): A regularized goodness-of-fit test for copulas. Journal de la Société Française de Statistique, 154(1): 64–77.

GUPTA, H.V., H. KLING, K.K. YILMAZ & G.F. MARTINEZ (2009): Decomposition of the mean squared error and NSE performance criteria. Implications for improving hydrological modelling. Journal of Hydrology 377(1-2): 80–91.

HOFERT, M., I. KOJADINOVIC, M. MAECHLER, J. YAN, J.G. NEŠLEHOVÁ & R. MORGER (2020): copula. Multivariate Dependence with Copulas. Version 1.0-1. R-Paket.

JUNGHÄNEL, T., H. ERTEL & T. DEUTSCHLÄNDER (2017): KOSTRA-DWD-2010R. Bericht zur Revision der koordinierten Starkregenregionalisierung und -auswertung des Deutschen Wetterdienstes in der Version 2010. Offenbach: Eigenverlag DWD.

JUNGHÄNEL, T., F. BÄR, T. DEUTSCHLÄNDER, U. HABERLANDT, I. OTTE, B. SHEHU, H. STOCKEL, K. STRICKER, L.-B. THIELE & W. WILLEMS (2022): Methodische Untersuchungen zur Novellierung der Starkregenstatistik für Deutschland (MUNSTAR). Synthesebericht. Offenbach: Eigenverlag DWD.

KLEEBERG, H.-B. & A.H. SCHUMANN (2001): Ableitung von Bemessungsabflüssen kleiner Überschreitungswahrscheinlichkeiten. Wasserwirtschaft, 91(2): 90–95.

KOUTSOYIANNIS, D., D. KOZONIS & A. MANETAS (1998): A mathematical framework for studying rainfall intensity-duraction-frequency. Journal of Hydrology, 206(1-2): 118–135.

KOUTSOYIANNIS, D. (2004a): Statistics of extremes and estimation of extreme rainfall: I. Theoretical investigation. Hydrological Sciences Journal, 49(4): 575–590.

KOUTSOYIANNIS, D. (2004b): Statistics of extremes and estimation of extreme rainfall: II. Empirical investigation of long rainfall records. Hydrological Sciences Journal, 49(4), 591–610.

LI, R., L. XIONG, C. JIANG, W. LI & C. LIU (2023): Quantifying multivariate flood risk under nonstationary condition. Natural Hazards, 116(1), 1161–1187.

LUDWIG, K. & M. BREMICKER (2006): The water balance model LARSIM. Freiburger Schriften zur Hydrologie, Bd. 22. Universität Freiburg.

MALITZ, G. & H. ERTEL (2015): KOSTRA-DWD-2010. Starkniederschlagshöhen für Deutschland (Bezugszeitraum 1951 – 2010). Abschlussbericht. Offenbach: Eigenverlag DWD.

MICHIELS, F. & A. DE SCHEPPER (2008): A copula test space model. How to avoid the wrong copula choice. Kybernetika, 44(6): 864-878.

NAGLER, T., U. SCHEPSMEIER, J. STOEBER, E.C. BRECHMANN, B. GRAELER, T. ERHARDT, C. ALMEIDA, A. MIN, C. CZADO, M. HOFMANN M. KILLICHES, H. JOE & T. VATTER (2023): VineCopula. Statistical Inference of Vine Copulas. Version 2.5.0. R-Paket.

ÖBMLFUW (Österreichisches Bundesministerium für Land- und Forstwirtschaft, Umwelt und Wasserwirtschaft) (2011): Leitfaden. Verfahren zur Abschätzung von Hochwasserkennwerten. Wien: Eigenverlag.

POULIN, A., D. HUARD, A.-C. FAVRE & S. PUGIN (2007): Importance of tail dependence in bivariate frequency analysis. Journal of Hydrologic Engineering, 12(4): 394–403.

R CORE TEAM (2021): R. A language and environment für statistical computing. R Foundation for Statistical Computing. Wien, Österreich.

RP STUTTGART (2012): Hochwassergefahrenkarte Baden-Württemberg – Beschreibung der Vorgehensweise zur Erstellung von Hochwassergefahrenkarten in Baden-Württemberg.

SALVADORI, G., F. DURANTE, C. DE MICHELE, M. BERNARDI & L. PETRELLA (2016): A mulitvariate copula-based framework for dealing with hazard szenarios and failure probabilities. Water Resources Research, 52(5): 3701–3721.

SCHUMANN, A.H. & S. FISCHER (2023): Sind Bemessungsasbflüsse nach dem Kleeberg/Schumann-Verfahren noch begründet? WASSERWIRTSCHAFT, 113(10): 10–14.

SEIBERT, S.P., U. EHRET & E. ZEHE (2016): Disentangling timing and amplitude errors in streamflow simulations, Hydrology and Earth System Science, 20(9): 3745–3763.

SKLAR, A. (1959), "Fonctions de répartition à n dimensions et leurs marges", Publ. Inst. Statist. Univ. Paris, 8: 229–231.

SKLAR, A. (1997): Random variables, distribution functions, and copulas – a personal look backward and forward. In: L. Rüschendorf, B. Schweizer, M. Taylor (Hrsg.): Distributions With Fixed Marginals & Related Topics.

SRAJ, M., N. BEZAK & M. BRILLY (2014): Bivariate flood frequency analysis using the copula function: a case study of the Litija station on the Sava River. Hyrological Processes, 29(2): 225–238.

TARASOVA, L., S. BASSO, M. ZINK & R. MERZ (2018): Exploring Controls on Rainfall-Runoff Events. 1. Time Series-Based Event Separation and Temporal Dynamics of Event Runoff Response in Germany. Water Resources Research 54 (10): 7711–7732.

TOOTOONCHI, F., M. SADEGH, J.O. HAERTER, O. RÄTY, T. GRABS & C. TEUTSCHBEIN (2022): Copulas for hydroclimatic analysis: A practice-oriented overview. WIREs Water, e1579: 1–28.

VERWORN, H.-R. & I. KUMMER (2003): Praxisrelevante Extremwerte des Niederschlags (PEN). Abschlussbericht. Institut für Wasserwirtschaft, Hydrologie und landwirtschaftlichen Wasserbau. Universität Hannover. Hannover.