

Alexander Dolich, Uwe Ehret & Ralf Loritz

Lernen oder Kalibrieren? Potenziale und Grenzen datengetriebener und konzeptioneller hydrologischer Modelle im deutschlandweiten Vergleich

To Learn or to calibrate? Potentials and limitations of data-driven and conceptual hydrological models in a Germany-wide comparison

Diese Studie untersucht die Leistungsfähigkeit regional trainierter datengetriebener „Long Short-Term Memory (LSTM)“-Netzwerke im Vergleich zu einzugsgebietsweise kalibrierten konzeptionellen HBV-Modellen (Hydrologiska Byråns Vattenbalansavdelning) zur Niederschlag-Abfluss-Simulation in Deutschland. Grundlage bildet der deutschlandweite CAMELS-DE-Datensatz (Catchment Attributes and MEteorology for Large-sample Studies-DEutschland), der hydrometeorologische Zeitreihen und Einzugsgebietseigenschaften für 1.314 ausgewählte Einzugsgebiete bereitstellt. Das LSTM-Ensemble erreicht einen Median-Nash-Sutcliffe-Effizienzwert (NSE) von 0,86 und übertrifft damit das HBV-Modell mit seinem Median-NSE von 0,72 in über 99 % der Einzugsgebiete. Beide Modelle unterschätzen Hochwasserscheitel, wobei das LSTM-Modell hierbei geringere Fehler und eine bessere zeitliche Genauigkeit erzielt. Besonders in großen, topographisch heterogenen und niederschlagsreichen Einzugsgebieten erreicht das LSTM-Modell hohe Simulationsgüten.

In unbeobachteten Einzugsgebieten erzielt das LSTM-Modell durch regionales Training einen Median-NSE von 0,69. Bereits geringe Mengen an Trainingsdaten verbessern die Modellgüte deutlich, mit einem Anstieg des Median-NSE auf 0,81. Aufgrund seiner massenerhaltenden Struktur reagiert das HBV-Modell empfindlicher auf erhebliche Massenbilanzfehler in den Eingangsdaten als das LSTM-Modell. Der systematische Vergleich der beiden Ansätze ermöglicht es, Einzugsgebiete mit problematischer Datenqualität besser zu identifizieren. Die Ergebnisse zeigen, dass sowohl konzeptionelle Modelle, aufgrund ihrer hohen Interpretierbarkeit, Extrapolationsfähigkeit und bewährten Anwendungspraxis, als auch datengetriebene LSTM-Modelle, dank ihrer überlegenen Simulationsgüte und Anwendbarkeit, in unbeobachteten Einzugsgebieten jeweils spezifische Stärken aufweisen und sich daher in der Niederschlag-Abfluss-Simulation optimal ergänzen können.

Schlagwörter: Maschinelles Lernen, Simulation, LSTM, HBV, CAMELS-DE

This study examines the performance of regionally trained data-driven Long Short-Term Memory (LSTM) networks in comparison to catchment-calibrated conceptual HBV models for rainfall-runoff simulation in Germany. It is based on the Germany-wide CAMELS-DE dataset, which provides hydrometeorological time series and catchment area characteristics for 1,314 selected catchments. The LSTM ensemble achieves a median Nash-Sutcliffe efficiency (NSE) of 0.86, outperforming the HBV model with its median NSE of 0.72 in over 99 % of the catchments. Both models underestimate flood peaks, with the LSTM simulations exhibiting smaller errors and better temporal accuracy. The LSTM achieves high simulation quality, especially in large, topographically heterogeneous, and precipitation-rich catchments. In ungauged catchments, the LSTM model achieves a median NSE of 0.69 through regional training. Even small amounts of training data significantly improve the simulation quality; the median NSE increases to 0.81. Due to its mass-conserving structure, the HBV model is more sensitive to significant mass balance errors in the input data than the LSTM model. A systematic comparison of the two approaches makes it possible to better identify catchments with problematic data quality. The results show that conceptual models like HBV, due to their high interpretability, extrapolation capability and proven practical application, as well as data-driven LSTM models, due to their superior simulation quality and applicability in ungauged basins, have specific strengths and can therefore complement each other optimally in rainfall-runoff simulation.

Keywords: Machine learning, simulation, LSTM, HBV, CAMELS-DE

1. Einleitung

Prozessbasierte und konzeptionelle hydrologische Modelle sind seit vielen Jahren der Standard in der Niederschlag-Abfluss-Simulation und -Vorhersage. In jüngerer Zeit gewinnen jedoch datengetriebene Methoden, insbesondere aus dem Bereich des „Deep Learning“, zunehmend an Bedeutung (NEARING et al., 2021). Diese neuen Ansätze werden bislang vor allem in der Forschung entwickelt und eingesetzt, halten aber inzwischen auch Schritt für Schritt Einzug in die hydrologische Praxis und die operationelle Hochwasservorhersage (NEARING et al., 2024; TANRIKULU et al., 2024; EHRET et al., 2025). Besonders hervorzuheben sind dabei Modelle auf Basis von „Long Short-Term Memory (LSTM)“-Netzwerken (HOCHREITER & SCHMIDHUBER, 1997).

Einen entscheidenden Beitrag zur Entwicklung dieser Modelle in der Hydrologie leisteten die Arbeiten von KRATZERT et al. (2018, 2019c). Sie zeigten, dass ein einzelnes LSTM-Netzwerk, trainiert auf einer Vielzahl von Einzugsgebieten mit meteorologischen Zeitreihen wie Niederschlag, Strahlung und Temperatur und statischen Einzugsgebietsmerkmalen wie Topographie, Bodenbeschaffenheit und Landnutzung, die Simulationsgenauigkeit von klassischen hydrologischen Modellen deutlich übertreffen kann.

Der rasante Fortschritt von Deep Learning Methoden in der Hydrologie ist dabei eng mit der Verfügbarkeit großer, harmonisierter Datensätze verknüpft. Da datengetriebene Modelle wie LSTMs für den Trainingsprozess große Mengen an hydrometeorologischen Daten benötigen, ist ihr Erfolg erst durch Datensätze mit

Hunderten von Einzugsgebieten möglich, wie sie beispielsweise durch die CAMELS-(Catchment Attributes and Meteorology for Large-Sample Studies)-Initiativen erstellt werden. Diese Datensätze vereinen für zahlreiche Einzugsgebiete detaillierte hydrologische und meteorologische Zeitreihen mit statischen Einzugsgebietseigenschaften, etwa zu Bodenbeschaffenheit oder Landnutzung. Besonders der CAMELS-US Datensatz (NEWMAN et al., 2015; ADDOR et al., 2017), der konsistente Daten für 671 Einzugsgebiete in den USA bereitstellt, hat die Entwicklung und Anwendung von Machine-Learning Methoden in der Hydrologie entscheidend vorangetrieben (KRATZERT et al., 2018, 2019c; FRAME et al., 2021).

Obwohl Deutschland über eine sehr umfangreiche hydrometeorologische Messinfrastruktur verfügt, fehlte bislang ein konsistenter, bundesweiter Datensatz. Dies erschwerte überregionale Studien und machte die Zusammenstellung und Aufbereitung hydrologischer Daten aufwendig. Mit der Veröffentlichung von CAMELS-DE (DOLICH et al., 2025; LORITZ et al., 2024) stand erstmals eine einheitliche Datenbasis für ganz Deutschland zur Verfügung. Dies eröffnete neue Möglichkeiten für datengetriebene Modellierungsansätze zur Abflusssimulation sowie für vertiefte Analysen hydrologischer Prozesse auf regionaler und nationaler Ebene. Bevor LSTM-basierte Simulationen und Vorhersagen in der Praxis umfassend eingesetzt werden können, ist es jedoch entscheidend, die Potenziale und Limitierungen dieser Modelle besser zu verstehen und systematisch zu evaluieren (BASTE et al., 2025; HEUDORFER et al., 2025).

Diese Studie vergleicht ein regional trainiertes Ensemble von LSTM-Netzwerken mit konzeptionellen HBV-Modellen (Hydrologiska Byråns Vattenbalansavdelning), die individuell pro Einzugsgebiet kalibriert wurden. Grundlage hierfür bildet der CAMELS-DE-Datensatz. Ziel ist es, die jeweiligen Stärken und Schwächen beider Modellansätze hinsichtlich ihrer Fähigkeit zur Simulation von Niederschlag-Abfluss-Prozessen auf Einzugsgebietsebene systematisch herauszuarbeiten. Dadurch sollen die Potenziale und Limitierungen datengetriebener Modelle im Vergleich zu traditionellen konzeptionellen Ansätzen erkennbar werden. Die Studie trägt somit dazu bei, eine differenzierte Einschätzung der Eignung von LSTM-Netzwerken und HBV-Modellen für spezifische hydrologische Fragestellungen und Anwendungen zu ermöglichen. Konkret werden folgende Forschungsfragen behandelt:

- 1.) Wie gut können regional trainierte LSTM-Netzwerke im Vergleich zu einzugsgebietsweise kalibrierten HBV-Modellen die Abflussdynamik und insbesondere Hochwasserereignisse in deutschen Einzugsgebieten simulieren?
- 2.) Welche Einzugsgebietseigenschaften und hydrologischen Bedingungen beeinflussen die Simulationsgüte von LSTM-Netzwerken und HBV-Modellen besonders stark?
- 3.) Wo liegen die jeweiligen Grenzen dieser Ansätze, insbesondere in Einzugsgebieten mit fehlerhaften, limitierten oder fehlenden Abflussdaten?

2. Datengrundlage und Methoden

Der folgende Abschnitt beschreibt den CAMELS-DE-Datensatz, welcher die Datengrundlage für die Modellierungen darstellt. Darüber hinaus stehen die beiden Modellansätze des LSTM-Netzwerks und des HBV-Modells im Fokus. Außerdem werden

die verwendeten Methoden zur Evaluierung der Simulationsergebnisse erklärt.

2.1. CAMELS-DE

Die vorliegende Studie verwendet den CAMELS-DE-Datensatz, der hydrometeorologische Zeitreihendaten in täglicher Auflösung sowie umfangreiche Einzugsgebietseigenschaften für 1.582 Einzugsgebiete aus 13 Bundesländern in Deutschland bereitstellt. Die Zeitreihen umfassen den Zeitraum von 1951 bis 2020 und beinhalten Abfluss- und Wasserstandsdaten sowie meteorologische Variablen wie Niederschlag, Lufttemperatur, relative Luftfeuchtigkeit und Strahlung. Die meteorologischen Eingangsgrößen wurden aus dem Datensatz HYRAS-DE (Hydrometeorologischer Rasterdatensatz für Deutschland, RAZAFIMAHARO et al., 2020) des Deutschen Wetterdienstes (DWD) für die jeweiligen Einzugsgebiete extrahiert und räumlich aggregiert. Neben den Zeitreihendaten enthält CAMELS-DE detaillierte, gebiets-spezifische Informationen zu Topographie, Landbedeckung, Bodeneigenschaften, Hydrogeologie und zum menschlichen Einfluss im Einzugsgebiet. Ergänzt wird dies durch statistische Kennwerte zum hydrologischen Regime sowie zu klimatischen Bedingungen. Weitere Informationen zu Aufbau, Aufbereitung und Qualitätssicherung des Datensatzes finden sich im zugehörigen Datensatzartikel (LORITZ et al., 2024) sowie in der offiziellen Datensatzbeschreibung (DOLICH et al., 2025).

2.2. LSTM-Netzwerk

„Long Short-Term Memory (LSTM)“-Netzwerke (HOCHREITER & SCHMIDHUBER, 1997) sind eine spezielle Klasse rekurrenter neuronaler Netze (RNNs), die für die Verarbeitung von Zeitreihendaten entwickelt wurden und sich insbesondere durch die Fähigkeit auszeichnen, Langzeitabhängigkeiten in sequenziellen Daten zu modellieren. Dies wird durch die Integration von sogenannten Speicherzellen („memory cells“) erreicht, die einen expliziten Zellzustand über längere Zeitintervalle hinweg aufrechterhalten können. Die Zellen sind mit verschiedenen steuernden „Gates“ ausgestattet: Das „Forget-Gate“ reguliert, welche Informationen aus dem vorherigen Zellzustand verworfen werden, während das „Input-Gate“ entscheidet, welche neuen Informationen hinzugefügt werden. Das „Output-Gate“ bestimmt schließlich, welche Informationen aus dem Zellzustand für die nächsten Layer oder als Output ausgegeben werden. Eine ausführliche, auch mathematische Beschreibung eines LSTM-Netzwerks liefern KRATZERT et al. (2019c).

Das Training eines LSTM-Netzwerks erfolgt auf Basis einer gegebenen Zeitreihe, wobei zu jedem Zeitschritt sowohl dynamische Eingangsdaten, z. B. meteorologische Variablen, als auch statische Prädiktoren, z. B. Einzugsgebietseigenschaften, bereitgestellt werden. Das Netzwerk generiert auf dieser Grundlage Simulationen für die Zielvariable, hier den spezifischen Abfluss. Der Trainingsprozess basiert auf dem Vergleich zwischen simuliertem und beobachtetem Wert mittels einer gewählten Verlustfunktion („Loss“), in dieser Studie der Nash-Sutcliffe-Effizienz, bei der der quadrierte Fehler mit der Standardabweichung des Abflusses gewichtet wird (KRATZERT et al., 2019c). Über „Backpropagation Through Time“ (BPTT) werden die Netzwerkgewichte iterativ so angepasst, dass der Fehler über alle Trainingsbeispiele hinweg minimiert wird, hier mittels des Adam-Optimierungsalgorithmus. Eine hohe Anzahl an Trainingszyklen und ein umfangreicher Trainingsdatensatz sind essentiell für die Generalisierungsfähigkeit und die Leistung des Modells. Ein Trainingszyklus, oder eine

Epoche, definiert dabei die einmalige, vollständige Verarbeitung des gesamten Trainingsdatensatzes, bei der die internen Modellgewichte schrittweise zur Optimierung der Simulation angepasst werden.

LSTM-Netzwerke sind dabei aus mehreren Gründen besonders für die Modellierung hydrologischer Zeitreihen geeignet. Sie können nichtlineare Zusammenhänge und insbesondere langfristige Abhängigkeiten in sequenziellen Daten erfassen, wie sie für hydrologische Prozesse charakteristisch sind. Entscheidend für die Erfassung saisonaler hydrologischer Prozesse ist dabei der Parameter der Sequenzlänge, welcher den zeitlichen Umfang der meteorologischen Eingangsdaten bestimmt. In der hydrologischen Modellierung wird die Sequenzlänge meist auf 365 Tage festgelegt, in der Annahme, dass diese Zeitperiode saisonale Prozesse wie Schneeakkumulation und -schmelze vollständig abbildet (KRATZERT et al., 2018; DE LA FUENTE et al., 2024).

Durch die Kombination von dynamischen meteorologischen Eingangsdaten, z. B. Niederschlag, Temperatur und Strahlung, mit statischen Einzugsgebietseigenschaften, etwa Topographie, Landnutzung und Bodeneigenschaften, sind LSTM-Modelle in der Lage, sowohl lokale als auch regionale hydrologische Verhaltensmuster zu erfassen und eine robuste Generalisierung über unterschiedliche Einzugsgebiete hinweg zu erreichen (KRATZERT et al., 2019b). Darüber hinaus ermöglichen es die internen Speichermechanismen der LSTM-Zellen, hydrologisch relevante Systemzustände wie etwa Schneespeicher oder Bodenfeuchte implizit zu repräsentieren, was eine hydrologische Interpretation der Modellzustände und damit auch der Modellierungsergebnisse erlaubt (KRATZERT et al., 2019a; LEES et al., 2022). Im Gegensatz zu anderen Disziplinen, in denen aktuell vorrangig Transformer-Architekturen für die Zeitreihenmodellierung eingesetzt werden (VASWANI et al., 2017), gelten LSTM-Netzwerke in der Hydrologie weiterhin als Stand der Technik bzgl. datengetriebener Modellierung und erzielen aktuell die besten Ergebnisse (LIU et al., 2024).

2.2.1. Netzwerkaufbau und Training

Beim regionalen Training von LSTM-Netzwerken werden diese nicht mit den Daten eines einzelnen Einzugsgebiets, sondern gleichzeitig mit den Daten zahlreicher Einzugsgebiete trainiert. Dieser Ansatz erhöht die Vielfalt und die Anzahl der Trainingsbeispiele deutlich, was zu einer verbesserten Generalisierungsfähigkeit und Robustheit des Modells führt (KRATZERT et al., 2019c; 2024). Durch die Einbindung statischer Einzugsgebietseigenschaften kann das Netzwerk Ähnlichkeiten und Unterschiede zwischen verschiedenen Einzugsgebieten erkennen und für die Modellierung nutzen. Ein wesentliches Merkmal des regionalen Trainings ist zudem die Möglichkeit, LSTM-Modelle auch in unbeobachteten Einzugsgebieten – also solchen ohne verfügbare Abflussmessungen im Trainingszeitraum – direkt einzusetzen.

Um Unsicherheiten aufgrund unterschiedlicher Initialisierungen der Modellgewichte zu adressieren, kam in dieser Studie ein Ensemble-Ansatz zur Anwendung: Es wurden zehn unabhängig initialisierte, regional trainierte LSTM-Netzwerke aufgesetzt. Aus den täglichen Simulationen aller Ensemblemitglieder berechnet das Modell für jeden Zeitschritt den Medianwert, auf dessen Basis sämtliche Leistungsmetriken, wie etwa der Nash-Sutcliffe-Effizienzwert (NSE), ermittelt werden. Der Trainingszeitraum für das LSTM-Modell umfasst den 1. Oktober 1970 bis 31. Dezember 1999, die Validierungsphase den 1. Oktober 1965 bis 30. Septem-

ber 1970 und der Testzeitraum erstreckt sich vom 1. Januar 2000 bis zum 31. Dezember 2020.

Als dynamische Eingangsvariablen dienen der mittlere Gebietsniederschlag und die räumliche Standardabweichung der Niederschlagsrasterwerte im Einzugsgebiet, die mittlere Strahlung sowie der Mittelwert der minimalen und maximalen Lufttemperatur im Einzugsgebiet. Die Zielvariable ist der spezifische Abfluss. Statische Prädiktoren sind Einzugsgebietsfläche, mittlere Höhe, Anteile an Ton, Schluff und Sand in der oberen Bodenschicht, Anteile verschiedener Landnutzungsklassen, mittlerer Gebietsniederschlag, Saisonalität des Niederschlags, Schneeanteil sowie Statistiken zu Starkniederschlags- und Trockenperioden.

Zusätzlich zum regional trainierten LSTM-Modell wurden für jedes einzelne Einzugsgebiet separate LSTM-Modelle mit denselben Hyperparametern wie das regionale Modell trainiert. Dies ermöglicht einen direkten Vergleich zum HBV-Modell, welches ebenfalls einzugsgebietsweise kalibriert wurde. Die zentralen Hyperparameter lauten: Anzahl LSTM-Layer = 1, hidden size = 128, Anzahl Epochen = 20, Lernrate = 0,001, Dropout = 0,4, Batchgröße = 256, Sequenzlänge = 365 Tage, Optimierungsalgorithmus = Adam. Die gewählte Implementierung und Parametrisierung wurde bereits im Rahmen von Benchmark-Studien umfassend getestet (ACUÑA ESPINOZA et al., 2024a). Der verwendete Code ist öffentlich verfügbar (ACUÑA ESPINOZA et al., 2024b).

2.2.2. Aktuelle Probleme und Limitierungen von LSTM-Modellen

Obwohl LSTM-Netzwerke inzwischen erfolgreich für die hydrologische Modellierung eingesetzt werden, bestehen weiterhin spezifische Herausforderungen und Limitierungen. Ein fundamentales Problem rein datengetriebener Ansätze liegt darin, dass diese Modelle nur sehr eingeschränkt auf Bedingungen und Ereignisse außerhalb des Trainingsdatenbereichs extrapolieren können (ACUÑA ESPINOZA et al., 2025b). Bei der Anwendung von LSTM-Modellen in der Hydrologie beschränkt sich dieses Defizit jedoch nicht nur auf die Extrapolation: Extremereignisse werden häufig unterschätzt, auch wenn diese innerhalb des Bereichs der Trainingsdaten liegen. Dies liegt darin begründet, dass die Modelle auf die dominierenden Durchschnittswerte der Verteilung der Trainingsdaten optimiert werden, also beispielsweise den mittleren Abfluss, und seltene Extremereignisse deshalb nur eine schwache Gewichtung erfahren.

Bei LSTM-Netzwerken besteht weiterhin das Risiko einer sogenannten Sättigung der Aktivierungsfunktion: Werden hohe Eingangsvariablen erreicht, wie etwa Niederschlagsmengen jenseits der Trainingsdaten oder langanhaltende Niederschlagsereignisse, führt dies ab einem bestimmten Punkt nicht mehr zu einer entsprechend erhöhten Abflussreaktion des Modells. Dieses Verhalten kann insbesondere bei der Vorhersage extremer Hochwasserereignisse oder bei der Bemessung problematisch sein (BASTE et al., 2025).

Darüber hinaus bleibt die Interpretierbarkeit von LSTM-Modellen eingeschränkt. Trotz der erfolgreichen Anwendung bei der Vorhersage hydrologischer Prozesse erschwert ihre interne Struktur ein tieferes Verständnis der zugrunde liegenden Prozesse. Zusätzlich fehlen physikalische Randbedingungen wie beispielsweise die Einhaltung der Massenbilanz, was die physikalische

Plausibilität der Ergebnisse reduziert und dadurch deren Anwendung in Studien mit sich ändernden Randbedingungen, wie zum Beispiel Klimawandelstudien, erschwert (ACUÑA ESPINOZA et al., 2024a; FENG et al., 2024). Es existieren zwar Ansätze aus der hybriden Modellierung, also dem Versuch, datengetriebene Modelle mit Prozessverständnis zu erweitern, wie beispielsweise ein massenerhaltendes LSTM-Modell (HOEDT et al., 2021) oder eine Kopplung von LSTM- und Konzeptmodell (ACUÑA ESPINOZA et al., 2024a). Dies kann zwar Extrapolationsfähigkeit schaffen und die Interpretierbarkeit der Modelle erhöhen, jedoch führte diese Einschränkung der Freiheitsgrade des rein datengetriebenen Modells bisher meist zu einer verminderten Modellleistung (KRATZERT et al., 2024).

2.3. HBV-Modell

Das HBV-Modell (Hydrologiska Byråns Vattenbalansavdelning; BERGSTRÖM & FORSMAN, 1973; SEIBERT, 2005) ist ein weit verbreitetes konzeptionelles Modell zur hydrologischen Abflussmodellierung, das am Schwedischen Meteorologischen und Hydrologischen Institut (SMHI) entwickelt wurde. Als klassisches hydrologisches Linearspeichermodell simuliert das HBV-Modell die wesentlichen hydrologischen Zustandsvariablen und Prozesse eines Einzugsgebiets, darunter Schneewasseräquivalent, Bodenwasser, Grundwasserspeicherung, Evapotranspiration, Schnell- und Basisabfluss sowie den Gesamtabfluss. Das Modell basiert auf einer Reihe von miteinander verbundenen Speicherkomponenten, deren Zuflüsse und Abflüsse kontinuierlich bilanziert werden.

Aufgrund dieses strukturellen Aufbaus erfüllt das HBV-Modell die Massenbilanz: Jeder in das System eingehende Niederschlag wird entweder in einem der Speicher gehalten, z. B. Schnee, Bodenwasser oder Grundwasser, oder er verlässt das Einzugsgebiet durch Evapotranspiration oder als Abfluss. Trotz seiner einfachen Struktur erlaubt das HBV-Modell durch die explizite Abbildung hydrologischer Prozesse und die Einhaltung der Massenerhaltung eine physikalische Interpretation der Simulationsergebnisse (SEIBERT & BERGSTRÖM, 2022). Der Code des hier verwendeten HBV-Modells ist öffentlich verfügbar (ACUÑA ESPINOZA et al., 2024b) und basiert auf SEIBERT (2005) und FENG et al. (2022).

2.3.1. Modellkalibrierung

In dieser Studie werden die HBV-Simulationen aus CAMELS-DE verwendet (LORITZ et al., 2024). Das HBV-Modell benötigt mit täglichem Niederschlag, Lufttemperatur und potenzieller Evapotranspiration vergleichsweise wenige meteorologische Eingangsvariablen und ist damit besonders geeignet für großflächige Anwendungen, auch in datenarmen Regionen (BERGSTRÖM, 2006). Die hier verwendeten Ergebnisse basieren auf einer räumlich auf die Einzugsgebietsfläche aggregierten („lumped“) Variante des klassischen HBV-Modells, die ein einheitliches Training für alle Einzugsgebiete in CAMELS-DE ermöglicht. Da die potenzielle Evapotranspiration (PET) nicht Bestandteil des im CAMELS-DE genutzten HYRAS-Datensatzes ist, wurde diese mithilfe einer modifizierten Hargreaves-Formel auf Basis von Temperatur und Niederschlag bestimmt (DROOGERS & ALLEN, 2002; CLERC-SCHWARZENBACH et al., 2024). Die Hargreaves-Methode war am besten geeignet, weil sie in einer Voruntersuchung unter verschiedenen Ansätzen, u. a. Penman und Priestley-Taylor, in den meisten Einzugsgebieten den geringsten Massenbilanzfeh-

ler und die besten Simulationsergebnisse mit einem HBV-Modell lieferte (LORITZ et al., 2024).

Das HBV-Modell wurde für jedes Einzugsgebiet individuell kalibriert, ein grundlegender Unterschied zum regional trainierten LSTM-Ansatz dieser Studie. Der Kalibrierungszeitraum ist dabei identisch mit dem Trainingszeitraum des LSTM-Modells, 1. Oktober 1970 bis 31. Dezember 1999. Für die Kalibrierung kam der „Differential Evolution Adaptive Metropolis (DREAM)“-Algorithmus zum Einsatz, ein auf MCMC (Markow-Chain-Monte-Carlo) basierender Optimierungsansatz, der die automatische Schätzung der Modellparameter erlaubt (VRUGT, 2016). Die Implementierung erfolgte unter Anwendung des Python-Paket „SPOTPY“ (HOUSKA et al., 2015). Zur Bewertung der Kalibrationsgüte diente, analog zum LSTM-Training, die Nash-Sutcliffe-Effizienz (NSE).

Das HBV-Modell wurde ausgewählt, da es trotz geringer Datenanforderungen in vielen verschiedenen hydrologischen Regionen Anwendung findet, relativ wenige Parameter benötigt und in zahlreichen Studien zuverlässige Ergebnisse geliefert hat. Die Ergebnisse des HBV-Modells ließen sich potenziell durch die akribische Betrachtung einzelner Einzugsgebiete und eine spezifische Anpassung der Modellarchitektur verbessern, was allerdings mit erheblichem Aufwand verbunden wäre. Außerdem ist anzumerken, dass es sich um nur einen einzigen kalibrierten Parametersatz handelt, während für das LSTM-Modell ein Ensemble aus 10 Mitgliedern verwendet wurde. Auch für das HBV-Modell wäre prinzipiell ein Ensembleansatz zu bevorzugen, da hierdurch eine Abschätzung der Modellunsicherheit sowie robustere Abflusssimulationen möglich wären (BEVEN & BINLEY, 1992). Die in dieser Studie vorgestellten Ergebnisse des HBV-Modells stehen stellvertretend für die Klasse konzeptioneller hydrologischer Modelle und dienen als Vergleichsgrundlage für die Bewertung des LSTM-Ansatzes in der Abflussmodellierung.

2.4. Evaluierung der Modellergebnisse

Im Folgenden werden die in dieser Studie verwendeten Gütemaße und Methoden zur Evaluierung der Simulationsergebnisse von LSTM- und HBV-Modell erläutert.

2.4.1. Simulationsgüte

Zur quantitativen Bewertung der Simulationsgüte kamen verschiedene Leistungsmetriken zum Einsatz. Eine vollständige Übersicht der eingesetzten Metriken mit deren Definitionen und Referenzen ist in Tabelle 1 aufgelistet. Die Auswahl umfasst sowohl generelle als auch abflussbereichsspezifische Kennwerte, um unterschiedliche Aspekte der Simulationsgüte differenziert abzubilden. Für die Gesamtbewertung der Simulationsgüte wurden die Nash-Sutcliffe-Effizienz (NSE) und der Kling-Gupta-Effizienzwert (KGE) herangezogen. Der NSE-Wert ist eine weit verbreitete Standardmetrik in der Hydrologie, die insbesondere hohe Abflüsse stärker gewichtet. Der KGE-Wert ergänzt diese Bewertung durch die gleichzeitige Berücksichtigung von Korrelation, Bias und Variabilität und ermöglicht dadurch eine ausgewogene Gesamtbeurteilung. Um die Modellleistung in verschiedenen Bereichen des Abflussregimes gezielt zu analysieren, wurden zusätzlich drei Metriken basierend auf Segmenten der Abflussdauerlinie eingesetzt:

- „Flow Duration Curve High-Segment Volume Error (FHV)“ zur Bewertung hoher Abflüsse,

Tabelle 1

Übersicht über die zur Abschätzung der Modellleistung verwendeten Metriken, inklusive Kurzbeschreibung, Wertebereich und Referenz. Tabelle basierend auf KRATZERT et al. (2021).

Overview of the metrics used to estimate model performance, including a brief description, value range, and reference. Table based on KRATZERT et al. (2021).

Metrik	Beschreibung	Wertebereich	Referenz
NSE	Nash-Sutcliffe-Effizienz: Maß für die Übereinstimmung zwischen beobachteten und simulierten Werten	$(-\infty, 1]$ Werte nahe 1 bevorzugt	NASH & SUTCLIFFE (1970)
KGE	Kling-Gupta-Effizienz: Kombination von Bias, Korrelation und Variabilität, um die Modellleistung umfassender zu bewerten	$(-\infty, 1]$ Werte nahe 1 bevorzugt	GUPTA et al. (2009)
FHV	Bias der 2 % höchsten Abflüsse (in %)	$(-\infty, \infty)$ Werte nahe 0 bevorzugt	YILMAZ et al. (2008)
FMS	Bias der Abflussdauerkurve (flow duration curve, FDC) zwischen dem 20 %- und dem 80 %-Perzentil (in %)	$(-\infty, \infty)$ Werte nahe 0 bevorzugt	YILMAZ et al. (2008)
FLV	Bias der 30 % niedrigsten Abflüsse (in %)	$(-\infty, \infty)$ Werte nahe 0 bevorzugt	YILMAZ et al. (2008)
Peak timing	Mittlere absolute Zeitdifferenz in Tagen zwischen beobachteten und simulierten Scheitelabflüssen	$(0, \infty)$ Werte nahe 0 bevorzugt	KRATZERT et al. (2021)

- „Flow Duration Curve Mid-Segment Error (FMS)“ für mittlere Abflüsse,
- „Flow Duration Curve Low-Segment Volume Error (FLV)“ für niedrige Abflüsse.

Darüber hinaus diente der „Peak Timing Error“ zur Erfassung der zeitlichen Genauigkeit bei Hochwasserereignissen. Dieser gibt an, wie gut die Modelle den Zeitpunkt von Abflussspitzen bestimmen. Hierbei ist zu beachten, dass die hydrologischen Prozesse, vor allem in kleinen Einzugsgebieten, meist schneller ablaufen als in den täglichen Daten repräsentiert, wodurch die Aussagekraft dieser Metrik hier eingeschränkt ist. Alle Metriken wurden mit Hilfe der Python-Bibliothek „NeuralHydrology“ (KRATZERT et al., 2019a) berechnet.

2.4.2. Bestimmung der Jährlichkeit von Hochwasserspitzen

Zur Bewertung der Modellgüte im Hinblick auf die Simulation extremer Abflussspitzen wurden die Hochwasserereignisse im Testzeitraum auf Grundlage ihrer statistischen Jährlichkeit (HQ) klassifiziert. Da der CAMELS-DE Datensatz keine offiziellen Wiederkehrwerte enthält, ließ sich die Jährlichkeit nur auf Basis der in CAMELS-DE enthaltenen Abflusszeitreihen statistisch ermitteln. Es ist dabei zu beachten, dass die so bestimmten HQ-Werte von den behördlich festgelegten Hochwasserkennwerten abweichen können und ausschließlich zur internen Bewertung der Modellleistung im Rahmen dieser Studie dienen. Die Berechnung erfolgte in mehreren Schritten:

Zunächst wurden für jedes Einzugsgebiet die jährlichen Maximalabflüsse aus dem Zeitraum 1951 bis 2020 extrahiert. Anschließend diente die Maximum-Likelihood-Methode dazu, eine Gumbel-Verteilung an diese jährlichen Maxima anzupassen, um eine inter- und extrapolationsfähige Extremwertverteilung zu erhalten. Aus der angepassten Verteilung wurden daraufhin die Abflüsse für die Wiederkehrintervalle HQ_2 , HQ_5 , HQ_{25} , HQ_{50} und HQ_{100} interpoliert bzw. extrapoliert. Für den Testzeitraum, die Jahre 2000 bis 2020, ließen sich mit der Funktion „find_peaks“ aus der Python-Bibliothek „scipy“ (VIRTANEN et al., 2020) die relevan-

ten Abflussspitzen in den beobachteten Zeitreihen identifizieren und anschließend den zuvor bestimmten HQ-Klassen zuordnen. Die entsprechende Modellsimulation wurde dann mit den beobachteten Werten verglichen.

Die Berechnung des prozentualen Fehlers zwischen simuliertem und beobachtetem Scheitelabfluss ermöglicht es, diese zu bewerten, jeweils in Abhängigkeit von der geschätzten Jährlichkeit des beobachteten Ereignisses. Dieses Vorgehen orientiert sich an der Methodik von ACUÑA ESPINOZA et al. (2025b), erweitert diese jedoch um die explizite Betrachtung der Vorzeichenrichtung des Fehlers. Die Verwendung des prozentualen Fehlers erlaubt eine differenzierte Einschätzung, ob das jeweilige Modell Hochwasserscheitel einer bestimmten Jährlichkeit systematisch über- oder unterschätzt.

2.4.3. Berechnung des Massenbilanzfehlers

Zur Untersuchung, wie das LSTM- und das HBV-Modell mit Fehlern in der Wasserbilanz einzelner Einzugsgebiete umgehen, wurde die Abweichung der Massenbilanz für jedes hydrologische Jahr, d. h. von 1. Oktober bis zum 30. September, im Testzeitraum der Jahre 2000 bis 2020, berechnet und anschließend gemittelt. Jahre, für die mehr als 10 % der Daten fehlen, wurden von der Analyse ausgeschlossen. Die Berechnungen basieren auf täglichen Werten des gemessenen Abflusses, des gemessenen Niederschlags sowie der berechneten potenziellen Verdunstung, welche über die modifizierte Hargreaves-Formel aus Temperatur- und Niederschlagsdaten abgeschätzt wurde (DROOGERS & ALLEN, 2002). Zunächst diente die folgende Gleichung dazu, den absoluten jährlichen Massenbilanzfehler zu berechnen:

$$\text{Fehler}_{\text{abs}} \text{ (mm yr}^{-1}\text{)} = P - (Q + ET) \quad (\text{Gl. 1})$$

Hierbei entspricht P dem Jahresniederschlag, Q dem spezifischen Abfluss und ET der potenziellen Verdunstung. Anschließend wurde der relative Fehler wie folgt berechnet:

$$\text{Fehler}_{\text{rel}} \text{ (%) } = \frac{\text{Fehler}_{\text{abs}}}{P} \times 100 \quad (\text{Gl. 2})$$

Der für die einzelnen hydrologischen Jahre berechnete relative Fehler wird wiederum gemittelt und als Maß für die langfristige Bilanzgüte der Daten verwendet.

Der relative Massenbilanzfehler erlaubt eine normierte Bewertung des Wasserhaushalts im Verhältnis zum Niederschlagsinput. Ein positiver relativer Fehler deutet darauf hin, dass das Einzugsgebiet mehr Wasser über den Niederschlag erhält, als über Abfluss und Verdunstung abgegeben wird. Ein negativer Fehler hingegen weist darauf hin, dass mehr Wasser das System verlässt, als durch Niederschlag eingebracht wurde.

Es ist dabei zu berücksichtigen, dass die Analyse auf der potenziellen Verdunstung basiert, welche die tatsächliche Verdunstung typischerweise überschätzt. Außerdem sind auch die für den HYRAS-Niederschlag verwendeten Stationsniederschläge mit Unsicherheiten zum Beispiel durch Windfehler, Benetzungsfehler und Verdunstungsverluste behaftet, wodurch eine systematische Unterschätzung des Niederschlags zu erwarten ist (RAUTHE et al., 2013). Aufgrund dieser Aspekte ist systematisch mit einem negativen Massenbilanzfehler zu rechnen.

3. Ergebnisse

Im Folgenden werden die Abfluss-Simulationsergebnisse des LSTM-Netzwerks und des HBV-Modells präsentiert, analysiert und gegenübergestellt. Die Analyse betrachtet dabei lediglich Einzugsgebiete, die im Trainingszeitraum 1970 bis 1999 über mindestens 20 % und im Testzeitraum 2000 bis 2020 über mindestens 10 % der möglichen Daten verfügen, was zu einer Auswahl von 1.314 Einzugsgebieten führt. Dies stellt sicher, dass einerseits der Datenumfang ausreichend ist, um das HBV-Modell zu kalibrieren, und andererseits ausreichend Daten im Testzeitraum zur Verfügung stehen, um die Modellgüte zu bewerten.

Alle im folgenden beschriebenen Ergebnisse beziehen sich auf diese Auswahl von 1.314 Einzugsgebieten, solange nicht anders angegeben. Weiterhin werden in Kapitel 3.1 die Ergebnisse des

einzugsgebietsweise trainierten („single-basin“) LSTM-Modells vorgestellt und mit den Ergebnissen des HBV-Modells und des regional trainierten LSTM-Modells verglichen. Die Ergebnisse des LSTM-Modells im weiteren Verlauf des Kapitels beziehen sich dabei, wenn nicht anders angegeben, stets auf das regional trainierte LSTM-Modell, da das regionale LSTM-Training dem aktuellen Stand der Wissenschaft entspricht.

3.1. Vergleich der Simulationsgüte

Wie aus Abbildung 1 ersichtlich, erreicht das regional trainierte LSTM-Modell einen Median-NSE-Wert von 0,86 über alle 1.314 ausgewählten Einzugsgebiete, während sich für das HBV-Modell ein Median-NSE-Wert über alle Einzugsgebiete von nur 0,72 ergibt. Die Verbesserung des LSTM-Modells im Vergleich zum CAMELS-DE-Benchmark von $NSE = 0,84$ (LORITZ et al., 2024) kommt dabei durch die Verwendung des LSTM-Ensemble-Ansatzes zustande. Im direkten Vergleich ist zu erkennen, dass das regional trainierte LSTM-Modell in annähernd allen Einzugsgebieten, insgesamt 99,3 %, eine bessere Simulationsgüte als das HBV-Modell erzielen konnte (Abb. 1b). Im Median liegt der NSE-Wert des LSTM-Netzwerks dabei in den einzelnen Einzugsgebieten um 0,13 höher als der des HBV-Modells.

Zusätzlich sind in Abbildung 1a die Ergebnisse der einzugsgebietsweise trainierten „single-basin“ LSTM-Modelle zu sehen. Durch die ausschließliche Verwendung der Trainingsdaten eines einzelnen Einzugsgebiets sinkt die Modellgüte im Vergleich zum regional trainierten LSTM-Modell auf einen Median-NSE-Wert von 0,73 und liegt dadurch im Bereich des HBV-Modells. Weiterhin erreichen die single-basin-LSTM-Modelle in lediglich 54 % der Einzugsgebiete eine höhere Modellgüte als das HBV-Modell, was die im Schnitt annähernd gleichwertige Modellgüte zum HBV-Modell verdeutlicht.

In Tabelle 2 sind die Ergebnisse aller in Kapitel 2.4.1 beschriebenen Gütemaße zu sehen. Wie zuvor dargestellt, liegt der Median-NSE-Wert des LSTM-Modells mit 0,86 höher als der Wert des HBV-Modells von 0,72. Der Unterschied beim KGE-Wert dagegen ist

geringer, das LSTM-Modell kommt hier auf einen Wert von 0,80, während das HBV-Modell einen KGE-Wert von 0,74 erreicht. Beide Modelle erzielen negative FHV-Werte, was bedeutet, dass sie hohe Abflüsse im Mittel unterschätzen. Mit einem Wert von -15,0 % liegt die Unterschätzung der Spitzenabflüsse durch das HBV-Modell höher als beim LSTM-Modell mit -13,5 %. Eine detaillierte Analyse der Fehler bei der Simulation von Hochwasserscheitelabflüssen erfolgt in Kapitel 3.3. Der Peak Time Error gibt Auskunft darüber, wie gut Scheitelabflüsse zeitlich getroffen werden. Dabei zeigt sich, dass das LSTM-Modell einen mittleren absoluten zeitlichen Versatz zwischen beobachteten und simulierten Scheitelabflüssen von 0,28 Tagen aufweist, während der Versatz beim HBV-Modell mit einem Wert von 0,58 Tagen ungefähr doppelt so hoch liegt. Weiterhin ist in Tabelle 2 die Anzahl der Einzugsgebiete aufgelistet, in denen der NSE-Wert ≤ 0 ist. Bei einem so gerin-

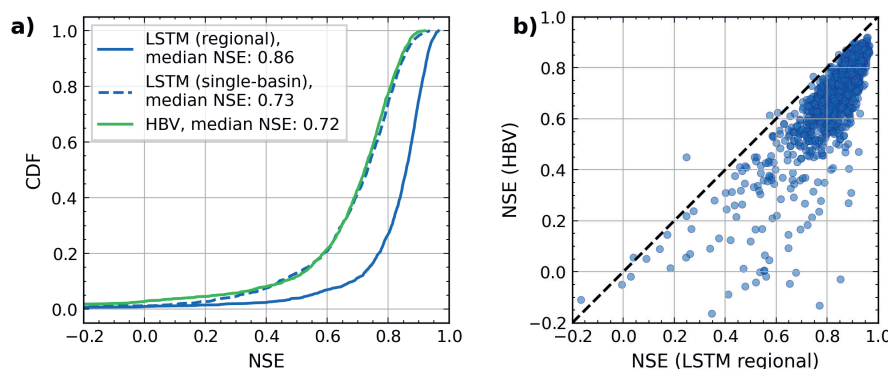


Abbildung 1

- Kumulative Verteilung der Nash-Sutcliffe Efficiency (NSE) für die 1.314 ausgewählten Einzugsgebiete des CAMELS-DE Datensatzes, für das HBV-Modell (grün), das regional trainierte LSTM-Modell (blau) und das einzugsgebietsweise trainierte LSTM-Modell (dunkelblau gestrichelt).
- NSE-Werte des HBV-Modells und des regional trainierten LSTM-Modells im direkten Vergleich für jedes Einzugsgebiet als Punktwolke.

Tabelle 2

Ergebnisse unterschiedlicher Metriken für die Evaluierung der Modellleistung von LSTM- und HBV-Modell. Abgebildet sind die jeweiligen Medianwerte für die 1.314 ausgewählten Einzugsgebiete. Eine Beschreibung der Metriken ist Tabelle 1 zu entnehmen.
Results of different metrics for evaluating the performance of the LSTM and HBV models. The respective median values for the 1,314 selected catchments are shown. A description of the metrics can be found in Table 1.

Modell	NSE (-)	KGE (-)	FHV (%)	FMS (%)	FLV (%)	Peak timing (Tage)	NSE ≤ 0
LSTM	0,86	0,80	-13,5	-12,1	38,6	0,28	12
HBV	0,72	0,74	-15,0	-6,8	-65,1	0,58	36

gen Wert kann davon ausgegangen werden, dass die Modelle gänzlich versagen und keine sinnvollen Abflüsse simulieren, was auf Probleme in den meteorologischen Eingangsdaten oder den Abflussdaten hindeutet. Im Fall des LSTM-Modells ist dies in 12 Einzugsgebieten der Fall, während das HBV-Modell in insgesamt 36 Einzugsgebieten einen negativen NSE-Wert aufweist.

Da die Abflussdaten in CAMELS-DE von den einzelnen Bundesländern geliefert wurden und es daher Unterschiede bezüglich Datenqualität und -quantität geben kann, z. B. bezüglich der Länge der Zeitreihen, erfolgt eine Analyse der Modellergebnisse auch nach Bundesländern. Diese Analyse zeigt zunächst, dass das LSTM-Modell in allen Bundesländern einen höheren Median-NSE-Wert erreicht als das HBV-Modell (Abb. 2a). Auch die Darstellung der Simulationsergebnisse auf einer Karte (Abb. 2b) bestätigt dieses Bild. Je stärker der Blauton der Punkte auf der Karte, desto ausgeprägter ist der Vorteil des LSTM- gegenüber dem HBV-Modell. Der überwiegend helle Blauton der Punkte lässt darauf schließen, dass der NSE-Wert des LSTM-Modells meist im Bereich 0,1 bis 0,2 höher liegt als der des HBV-Modells.

Aus der Abbildung wird deutlich, dass das LSTM-Modell in keinem Bundesland überproportional bessere Ergebnisse liefert als das HBV-Modell. Es existieren jedoch einzelne Einzugsgebiete, in denen das LSTM-Modell deutlich bessere Ergebnisse erzielt als das HBV-Modell. Abbildung 2a zeigt außerdem, dass beide Modelle in einer Großzahl der Bundesländer ähnlich gut funktionieren. Eine Ausnahme hiervon ist vor allem Brandenburg, wo beide Modelle deutlich schlechter abschneiden als im Rest Deutschlands: So liegt der Median-NSE für das LSTM-Modell bei 0,73, für das HBV-Modell bei 0,45. Auch in Sachsen-Anhalt ist die Modellgüte zwar geringer als im Rest Deutschlands, allerdings ist dieser Unterschied weitaus weniger ausgeprägt als in Brandenburg.

3.2. Zusammenhang zwischen statischen Einzugsgebietseigenschaften und Modellgüte

Um zu bewerten, ob Zusammenhänge zwischen bestimmten statischen Einzugsgebietseigenschaften im CAMELS-DE Datensatz und dem NSE-Wert der LSTM- und HBV-Simulationen existieren, wurden die Zusammenhänge analysiert und für das

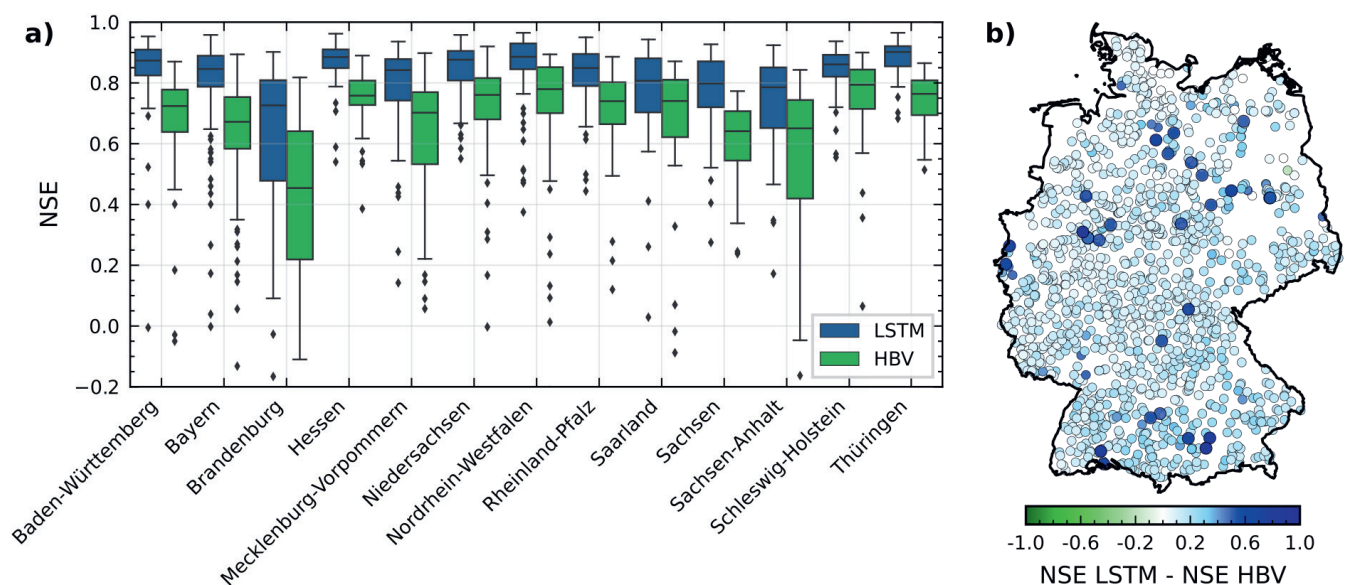
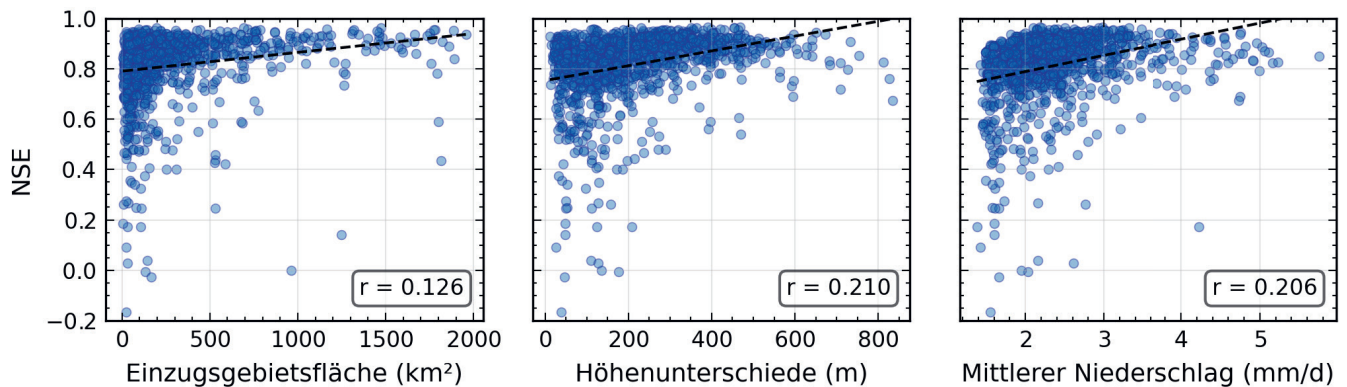


Abbildung 2

- a) Boxplots des NSE-Werts für die Ergebnisse des LSTM- und des HBV-Modells, gruppiert nach den 13 Bundesländern in CAMELS-DE.
b) Karte mit den Positionen der Stationen in CAMELS-DE, eingefärbt nach der Differenz zwischen den NSE-Werten für LSTM- und HBV-Modell. Je höher der Wert und damit je blauer der Punkt auf der Karte, desto höher ist der NSE-Wert des LSTM- im Vergleich zum HBV-Modell.
a) Box plots of the NSE for the results of the LSTM and HBV models, grouped according to the 13 federal states in CAMELS-DE.
b) Map showing the locations of the stations in CAMELS-DE, coloured according to the difference between the NSE values of LSTM and HBV model. The higher the value and thus the bluer the point on the map, the higher the NSE of the LSTM compared to the HBV model.

**Abbildung 3**

Zusammenhang zwischen dem NSE-Wert des LSTM-Modells und den folgenden Einzugsgebietseigenschaften: Einzugsgebietsfläche, Höhenunterschied im Einzugsgebiet und mittlerer Niederschlag für die Jahre 1951 bis 2020. Jeder Punkt stellt dabei eins der 1.314 ausgewählten Einzugsgebiete dar. Zusätzlich sind die Regressionsgerade sowie der r-Wert dieser Gerade für den Zusammenhang zwischen NSE und Einzugsgebietseigenschaft abgebildet.

Correlation between the NSE of the LSTM model and the following catchment attributes: catchment area, elevation difference within the catchment, and average precipitation for the years 1951 to 2020. Each point represents one of the 1,314 selected catchments. In addition, the regression line and the r-value of this line are shown for the correlation between NSE and catchment attribute.

LSTM-Modell, wie in Abbildung 3 zu sehen, als Punktwolken visualisiert. Zusätzlich fand eine Berechnung der Regressionsgeraden sowie des r-Wertes als Gütemaß der Regression statt. Damit soll abschätzbar sein, ob bestimmte Einzugsgebietseigenschaften die Güte der LSTM-Simulation negativ oder positiv beeinflussen. Die in Abbildung 3 dargestellten statischen Einzugsgebietseigenschaften sind eine Auswahl, bei der ein gewisser Zusammenhang erkennbar ist. Die p-Werte für die Korrelationen liegen im Bereich von $1,6 \times 10^{-14}$ bis $6,8 \times 10^{-6}$ und weisen damit auf eine hohe statistische Signifikanz hin.

Der NSE-Wert steigt mit zunehmender Einzugsgebietsgröße an, Simulationen mit dem LSTM-Modell funktionieren in großen Einzugsgebieten tendenziell also besser. Weiterhin gibt es auch einen positiven Zusammenhang mit den Höhenunterschieden im Einzugsgebiet: Das LSTM-Modell funktioniert in Einzugsgebieten mit ausgeprägter Topographie besser als in eher flachen Einzugsgebieten. Der Höhenunterschied wurde dabei aus der Differenz zwischen dem 95. und 5. Perzentil der Einzugsgebietshöhe berechnet. Des Weiteren sind die Ergebnisse des LSTM-Modells in trockenen Einzugsgebieten mit einem niedrigen langjährigen mittleren Niederschlag schlechter als in feuchten Gebieten mit größeren Niederschlagsmengen. Das HBV-Modell zeigt keinen Zusammenhang zwischen dem NSE-Wert und den in Abbildung 3 für das LSTM-Modell gezeigten Einzugsgebietseigenschaften; seine r-Werte liegen im Bereich zwischen -0,02 und 0,02.

3.3. Untersuchung der Simulation von Hochwasserscheitelabflüssen

Zunächst wird in Abbildung 4a der prozentuale Fehler aller simulierten Abflusswerte, nicht nur der Extreme, gezeigt. Hieraus wird ersichtlich, dass der mittlere prozentuale Fehler in diesem Fall für beide Modelle nahe Null ist, es also im Mittel keine systematische Unter- oder Überschätzung der Abflüsse gibt. Die Streuung des prozentualen Fehlers ist beim HBV- größer als beim LSTM-Modell. Um die Leistung der Modelle in Bezug auf die Simulation von Hochwasserscheitelabflüssen zu untersuchen, wurden diese Spitzen im Testzeitraum, wie in Kapitel 2.4.2 beschrieben, identifiziert. In Abbildung 4b sind Boxplots des prozentualen Fehlers für verschiedene Kategorien von Jährlichkeiten für das LSTM-

und das HBV-Modell dargestellt. Dabei unterschätzen beide Modelle systematisch die Abflussspitzen, wie aus dem negativen prozentualen Fehler hervorgeht. Das HBV-Modell unterschätzt die Spitzen dabei stärker als das LSTM-Modell, der Median des prozentualen Fehlers liegt für alle Kategorien zwischen 15 und 20 % geringer als der des LSTM-Modells.

Außerdem ist ein klares Muster darin zu erkennen, dass mit zunehmender Jährlichkeit des Hochwassers die Unterschätzung der simulierten Scheitelabflüsse zunimmt. Die höchsten Hochwasserscheitel werden also am stärksten unterschätzt. Abbildung 4c zeigt den Anteil an Fällen, in denen das LSTM- bzw. das HBV-Modell die Hochwasser der unterschiedlichen Jährlichkeitskategorien über- bzw. unterschätzt. Hier ergibt sich ein ähnliches Bild wie in Abbildung 4b: Der größte Teil der Hochwasserscheitel aller Kategorien wird unterschätzt, während das LSTM-Modell die Hochwasserspitzen etwas häufiger überschätzt als das HBV-Modell. Die größten Hochwasser, HQ_{50} bis HQ_{100} und $> HQ_{100}$, werden vom HBV-Modell fast immer unterschätzt, während im Falle des LSTM-Modells zumindest ein kleiner Teil der Spitzen überschätzt wird.

3.4. Einfluss von Massenbilanzfehlern auf die Simulationsergebnisse

Um zu überprüfen, wie gut die Modelle mit unterschiedlichen Massenbilanzfehlern umgehen können, wurde dieser Fehler, wie in Kapitel 2.4.3 beschrieben, für jedes Einzugsgebiet für den Testzeitraum von 2000 bis 2020 berechnet. In Abbildung 5 ist der Zusammenhang zwischen dem Massenbilanzfehler bzw. verschiedenen Klassen von Massenbilanzfehlern und dem Median-NSE-Wert des LSTM- und des HBV-Modells dargestellt. Dabei wurden die NSE-Werte für die jeweiligen Kategorien des Massenbilanzfehlers berechnet. Anhand der Abbildung lässt sich zunächst feststellen, dass der überwiegende Teil der 1.314 Einzugsgebiete einen negativen Massenbilanzfehler aufweist. Dies bedeutet, dass der Abfluss und die potenzielle Verdunstung den Einzugsgebietsniederschlag im langjährigen Mittel übersteigen. Dies ist zu erwarten, da die potenzielle Verdunstung größer ist als die tatsächliche Verdunstung.

Die NSE-Werte des LSTM-Modells sind im Median in den Kategorien mit einem Massenbilanzfehler zwischen +25 % und -50 % relativ stabil. Das HBV-Modell erzielt dagegen geringere NSE-Werte in der Kategorie +25 % bis 0 %. Die Anzahl der Einzugsgebiete in dieser Kategorie ist mit 43 zwar relativ gering, das HBV-Modell scheint allerdings mit Einzugsgebieten mit einem

positiven Massenbilanzfehler größere Probleme zu haben als mit negativen Fehlern. Für größere Massenbilanzfehler bis -100 % sinken sowohl der Median-NSE-Wert des LSTM- als auch des HBV-Modells auf einen Wert von 0,43 bzw. 0,17. Für die größte Massenbilanzfehlerkategorie von -100 % bis -200 % steigt der Median-NSE-Wert des LSTM-Modells stark auf einen Wert von

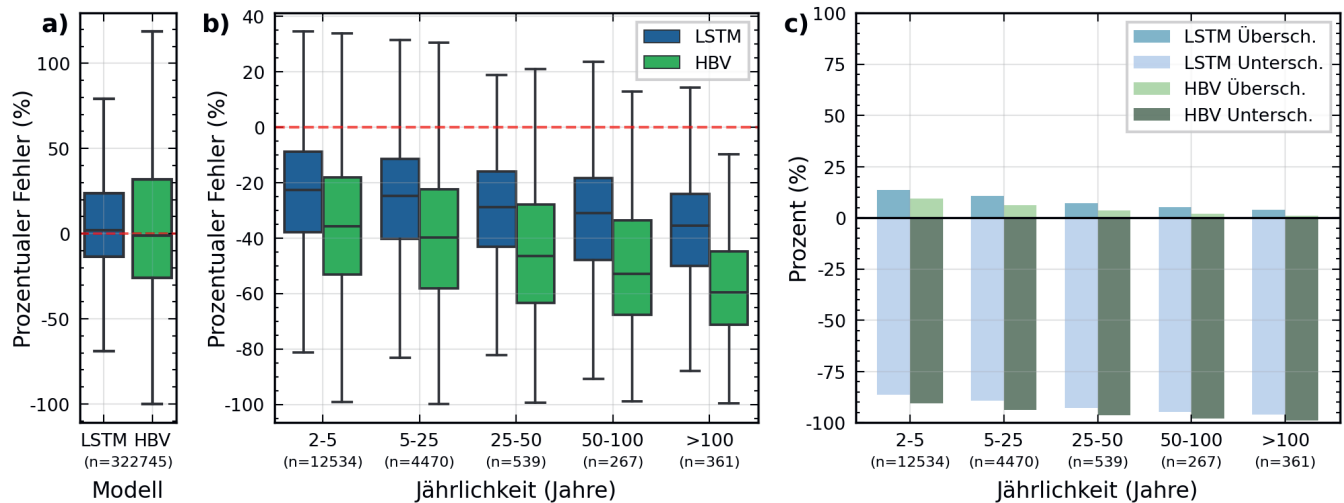


Abbildung 4

- a) Boxplots des prozentualen Fehlers des LSTM- und des HBV-Modells für alle Abflüsse im Testzeitraum, die Jahre 2000 bis 2020, in CAMELS-DE.
b) Prozentualer Fehler des LSTM- und des HBV-Modells bei der Vorhersage von Abflussspitzen im Testzeitraum. Die Abflussspitzen wurden dabei nach ihren Jährlichkeiten in die Kategorien HQ₂ bis HQ₅, HQ₅ bis HQ₂₅, HQ₂₅ bis HQ₅₀, HQ₅₀ bis HQ₁₀₀ und > HQ₁₀₀ eingeteilt.
c) Prozentsatz der Fälle, in denen das LSTM- bzw. das HBV-Modell die Abflussspitzen in denselben Kategorien wie b) überschätzt bzw. unterschätzt hat.
- a) Box plots of the percentage error of the LSTM and HBV model for all discharges in the test period, the years 2000 to 2020, in CAMELS-DE.
b) Percentage error of the LSTM and HBV model in predicting discharge peaks in the test period. The peak discharges were classified according to their return periods into the categories HQ₂ to HQ₅, HQ₅ to HQ₂₅, HQ₂₅ to HQ₅₀, HQ₅₀ to HQ₁₀₀ and >HQ₁₀₀.
c) Percentage of cases in which the LSTM and HBV model overestimated or underestimated the peak discharges, in the same categories as b).

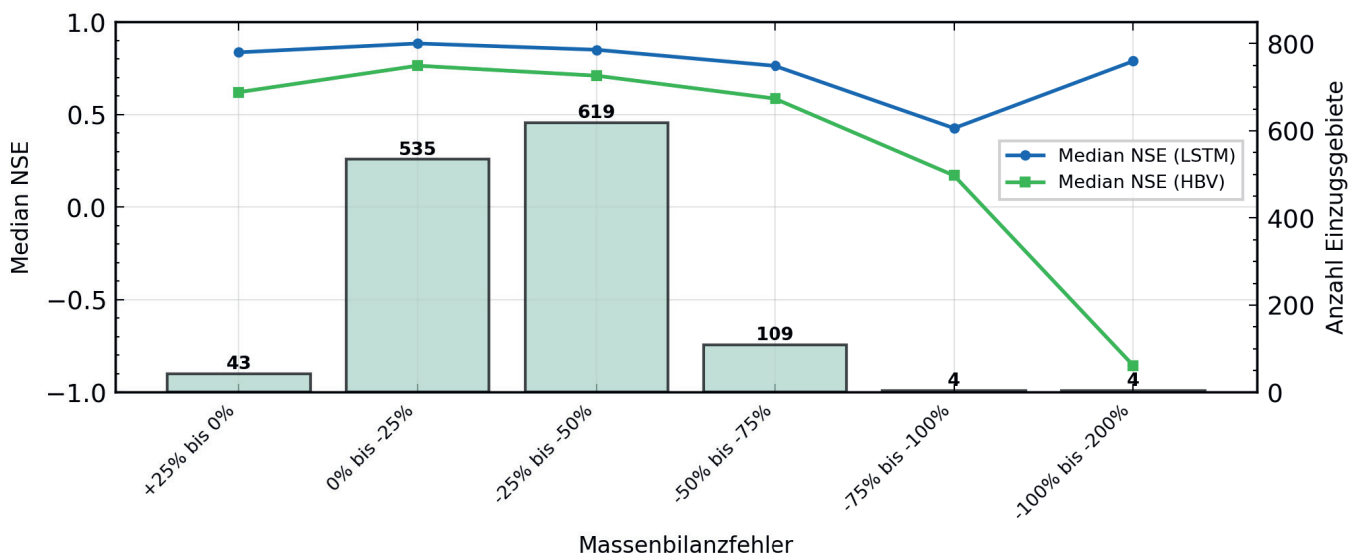


Abbildung 5

Zusammenhang zwischen den Massenbilanzfehlern der einzelnen Einzugsgebiete und deren NSE-Wert für LSTM- und HBV-Modell. Als Balken ist die Anzahl der Einzugsgebiete nach deren Massenbilanzfehler dargestellt, wobei Kategorien von Massenbilanzfehlern zur Gruppierung dienen. Zusätzlich ist der Median-NSE-Wert des LSTM- (blau) sowie des HBV-Modells (grün) für die unterschiedlichen Massenbilanzfehler-Kategorien abgebildet.

Relationship between the mass balance errors of the individual catchments and their NSE for LSTM and HBV model. The bars represent the number of catchments according to their mass balance errors, with divisions into categories of mass balance errors. In addition, the median NSE of the LSTM (blue) and HBV models (green) for the different mass balance error categories is shown.

0,79 an, während der Median-NSE des HBV-Modells auf einen Wert von -0,86 fällt. Das LSTM-Modell ist hier also in der Lage, vom Massenbilanzfehler unbeeinflusst den Abfluss zu simulieren, während das HBV-Modell bei größer werdenden Massenbilanzfehlern zunächst an Leistung verliert und ab einem gewissen Punkt nicht mehr funktioniert. Die Anzahl der Einzugsgebiete in Kategorien mit sehr großen Massenbilanzfehlern ist allerdings gering.

3.5. Ergebnisse in Einzugsgebieten mit limitierten Trainingsdaten

Der Datensatz CAMELS-DE beinhaltet insgesamt 114 Einzugsgebiete, für die im Trainingszeitraum keine Abflussdaten vorliegen, während im Testzeitraum Daten verfügbar sind. Anhand dieser Einzugsgebiete kann die Leistung eines Modells in unbeobachteten Einzugsgebieten getestet werden. Das HBV-Modell wurde automatisiert für jedes Einzugsgebiet anhand der Daten im Trainingszeitraum kalibriert, weshalb mit dem HBV-Modell in Einzugsgebieten ohne Daten im Trainingszeitraum keine Simulationen ohne zusätzliche Regionalisierung möglich sind (POOL et al., 2021; FENG et al., 2024). Das Training des LSTM-Modells erfolgte regional mit den Daten aller Einzugsgebiete. Es ist somit in der Lage, auch Simulationen in unbeobachteten Einzugsgebieten zu erstellen, bzw. aus den meteorologischen und statischen Eingangsvariablen einen Abfluss zu schätzen, unabhängig von der Verfügbarkeit von Abflussdaten im Einzugsgebiet.

Die Qualität der Ergebnisse des LSTM-Modells in Einzugsgebieten ohne Trainingsdaten ist in Abbildung 6 zu sehen. Außerdem sind auch Boxplots des NSE-Werts für Einzugsgebiete mit 0 bis 20 % verfügbaren Daten im Trainingszeitraum sowie mit mindestens 20 % der möglichen Trainingsdaten abgebildet. Da auch in dieser Analyse mindestens 10 % der möglichen Daten im Testzeitraum vorliegen müssen, entspricht die dritte Kategorie der Auswahl an 1.314 Einzugsgebieten im vorangegangenen Teil der Studie. Aus Abbildung 6 ist ersichtlich, dass das LSTM-Modell zwar im Vergleich zu beobachteten Einzugsgebieten niedrigere NSE-Werte erzielt, in unbeobachteten Einzugsgebieten wird aber weiterhin ein Median-NSE-Wert von 0,69 erreicht. Das ist etwas niedriger als der Median-NSE-Wert des HBV-Modells in beobachteten Einzugsgebieten, bei denen mindestens 20 % der möglichen Abflussdaten im Trainingszeitraum verfügbar waren, der 0,72 betrug. Diese Ergebnisse sind übereinstimmend mit den Ergebnissen von KRATZERT et al. (2019b) und HEUDORFER et al. (2025).

Einschränkend lässt sich jedoch hinzufügen, dass die Varianz der NSE-Werte in unbeobachteten Einzugsgebieten deutlich höher liegt und 10 dieser Einzugsgebiete, d. h. 8,8 %, einen NSE-Wert < 0 erzielen. In einigen unbeobachteten Einzugsgebieten kann das LSTM-Modell

also keine sinnvollen Abflusssimulationen erstellen. Weiterhin wird aus Abbildung 6 ersichtlich, dass schon eine geringe Menge an Trainingsdaten, 0 bis 20 % der möglichen Datenmenge, die Leistung des LSTM-Modells deutlich auf einen Median-NSE-Wert von 0,81 erhöht und sich auch die Varianz der NSE-Werte stark verringert. Der Anteil der Einzugsgebiete mit einem NSE-Wert < 0 bleibt dabei mit einem Wert von 9,2 % aber relativ konstant. Bei einer weiteren Verfügbarkeit von Trainingsdaten, ≥ 20 % der möglichen Daten, nimmt die Qualität der Simulationen, wie in den vorangegangenen Kapiteln beschrieben, weiter zu und der Anteil der Einzugsgebiete mit einem NSE-Wert < 0 beträgt lediglich 0,9 %.

Im Falle des HBV-Modells sind die Ergebnisse ähnlich, hier steigt der Median-NSE-Wert von 0,63 in der Kategorie der Einzugsgebiete mit bis zu 20 % verfügbaren Abflussdaten im Trainingszeitraum auf einen Wert von 0,72 für Einzugsgebiete mit mehr als 20 % der Daten im Trainingszeitraum an. Auch die Varianz der NSE-Werte nimmt ab, während der Anteil von Einzugsgebieten mit einem NSE-Wert < 0 von 12,6 % auf 2,4 % fällt.

4. Diskussion

Im Folgenden werden die in Kapitel 3 vorgestellten Ergebnisse im Hinblick auf die in Kapitel 1 formulierten Forschungsfragen diskutiert.

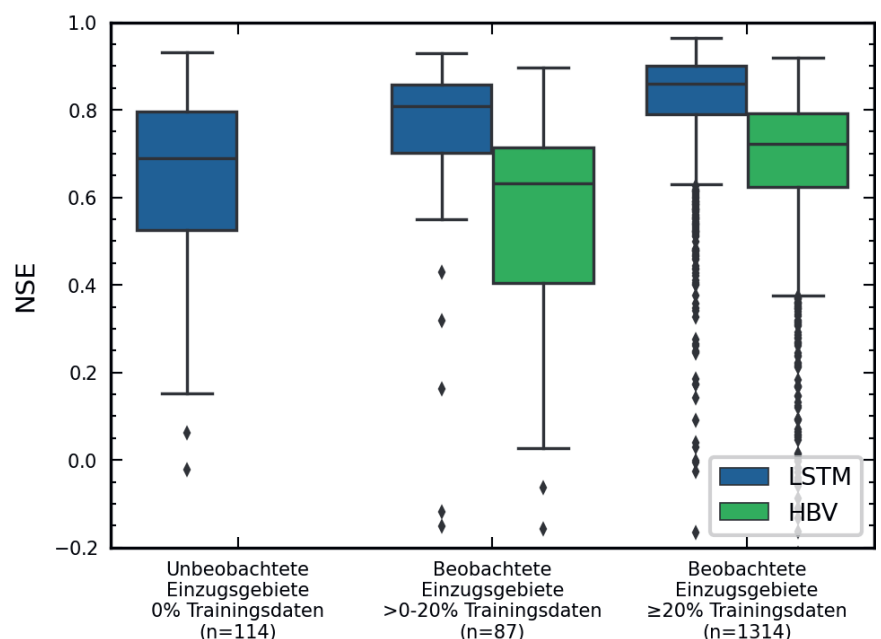


Abbildung 6

Boxplots der NSE-Werte des LSTM-Modells für unbeobachtete Einzugsgebiete, also ohne Abflussdaten im Trainingszeitraum, sowie des LSTM- und des HBV-Modells für beobachtete Einzugsgebiete mit > 0 bis 20 % verfügbaren Abflussdaten im Trainingszeitraum und beobachtete Einzugsgebiete mit mindestens 20 % verfügbaren Abflussdaten im Trainingszeitraum. Da auch hier alle Einzugsgebiete mindestens über 10 % der Abflussdaten im Testzeitraum verfügen müssen, entspricht die letzte Kategorie derselben Auswahl an Einzugsgebieten wie im vorangegangenen Teil der Studie.

Box plots of the NSE values of the LSTM model for ungauged catchments, i.e. without discharge data in the training period, as well as of the LSTM and HBV models for gauged catchments with > 0 to 20 % available discharge data in the training period and gauged catchments with at least 20 % available discharge data in the training period. Since all catchments must also have at least 10 % of the discharge data in the test period, the last category corresponds to the selection of catchments in the previous part of the study.

1.) Wie gut können regional trainierte LSTM-Netzwerke im Vergleich zu einzugsgebietsweise kalibrierten HBV-Modellen die Abflussdynamik und insbesondere Hochwasserereignisse in deutschen Einzugsgebieten simulieren?

Wie in Kapitel 3.1 gezeigt, erzielt das LSTM-Ensemble mit einem Median-NSE-Wert von 0,86 eine bessere globale Modellgüte als das HBV-Modell mit einem Median-NSE-Wert von 0,72 und übertrifft dieses in nahezu allen untersuchten Einzugsgebieten. Diese Ergebnisse verdeutlichen, dass regional trainierte LSTM-Netzwerke die allgemeine Abflussdynamik in Deutschland besser simulieren können als einzugsgebietsweise kalibrierte HBV-Modelle. Dennoch erzielt auch das HBV-Modell in den meisten Gebieten gute Ergebnisse. Die Leistung des HBV-Modells ließe sich in einzelnen Gebieten potenziell durch spezifische Anpassungen der Modellstruktur, Einbindung lokalen hydrologischen Wissens, alternative Kalibrierungsverfahren sowie durch die Anwendung von Ensembles weiter verbessern. Eine solche Vorgehensweise wäre jedoch mit einem wesentlich höheren zeitlichen und methodischen Aufwand verbunden als der Einsatz eines regional trainierten Standard-LSTM-Modells. Diese Befunde stimmen mit Studien aus Großbritannien (LEES et al., 2021) und den USA (KRATZERT et al., 2019c) überein, welche ebenfalls eine bessere globale Performance datengetriebener LSTM-Modelle im Vergleich zu klassischen konzeptionellen Modellen zeigen.

Die Metriken FHV (Flow Duration Curve High-Segment Volume Error) und Peak Timing Error geben Aufschluss darüber, wie gut hohe Abflüsse und Hochwasserscheitel simuliert werden. Hier erzielt das LSTM-Modell ebenfalls bessere Ergebnisse, mit einem FHV von -13,5 % gegenüber -15,0 % beim HBV-Modell und einem Peak Timing Error von 0,28 Tagen gegenüber 0,58 Tagen beim HBV-Modell. Wie in Kapitel 3.3 gezeigt, unterschätzen beide Modelle jedoch systematisch hohe Abflüsse und Hochwasserspitzen, wobei diese Unterschätzung beim HBV-Modell stärker ausgeprägt ist und insbesondere bei zunehmender Jährlichkeit größer wird. Die hier beobachtete bessere Güte des LSTM-Netzwerks bei der Simulation von Hochwasserereignissen könnte darin begründet sein, dass dieses während des regionalen Trainings von extremen Ereignissen aus zahlreichen Einzugsgebieten lernen kann, während das HBV-Modell jeweils auf die Daten eines einzelnen Einzugsgebiets beschränkt ist.

Gleichzeitig widersprechen diese Befunde teilweise Studien von FRAME et al. (2022) und ACUÑA ESPINOZA et al. (2025b), welche zeigten, dass konzeptionelle Modelle bei extremen Hochwasserereignissen ähnliche oder sogar bessere Ergebnisse erzielen können als LSTM-Netzwerke. Diese Diskrepanz könnte auf Unterschiede in der Modellstruktur, das Kalibrierungsverfahren oder die spezifischen hydro-klimatischen Bedingungen der untersuchten Regionen zurückzuführen sein.

Darüber hinaus zeigten KRATZERT et al. (2024) und BASTE et al. (2025), dass typische Trainingsverfahren von LSTM-Netzwerken durch eine Sättigung der Zellzustände bzw. der Aktivierungsfunktionen eine Obergrenze der simulierten Abflusswerte bedingen, welche unterhalb der Maxima der Trainingsdaten liegt. Ein weiterer Grund für die Unterschätzung von extremen Abflüssen liegt darin begründet, dass Extremereignisse wie Starkregen oder Hochwasser in den Trainingsdaten aufgrund ihres im Verhältnis seltenen Auftretens unterrepräsentiert sind, wodurch das LSTM-Modell während des Trainings vor allem für mittlere

Bedingungen optimiert wird und mittlere Abflüsse eine stärkere Gewichtung erhalten als Extremereignisse. Diese Aspekte limitieren die Fähigkeit der LSTM-Netzwerke, Extremwerte sinnvoll zu simulieren, auch wenn diese innerhalb der Trainingsdaten liegen. Solche Einschränkungen bei der Simulation von Extremereignissen sind insbesondere für operationelle Hochwasservorhersagen und bei der Bestimmung von Bemessungsabflüssen kritisch und derzeit Gegenstand aktiver Forschung.

Abschließend bleibt festzuhalten, dass die Simulation extremer Abflussereignisse sowohl für datengetriebene als auch für konzeptionelle Modelle nach wie vor eine Herausforderung darstellt (HARLIN, 1992; MUÑOZ-CASTRO et al., 2025). Dennoch bietet die Anwendung regional trainierter LSTM-Netzwerke in Deutschland insgesamt eine vielversprechende und potenziell leistungsfähigere Alternative oder Ergänzung zu konzeptionellen hydrologischen Modellierungsansätzen, vor allem aufgrund der hohen allgemeinen Simulationsgüte.

2.) Welche Einzugsgebietseigenschaften und hydrologischen Bedingungen beeinflussen die Simulationsgüte von LSTM-Netzwerken und HBV-Modellen besonders stark?

In Kapitel 3.2 wurde gezeigt, dass bestimmte statische Einzugsgebietseigenschaften einen Einfluss auf die Simulationsgüte des LSTM-Netzwerks haben, während für das HBV-Modell kein klarer Zusammenhang erkennbar ist. Für das LSTM-Modell gilt, dass es tendenziell in größeren Einzugsgebieten bessere Ergebnisse liefert als in kleineren Einzugsgebieten. Dies kann damit zusammenhängen, dass die tägliche Auflösung der Daten und des Modells in schnell reagierenden, kleinen Einzugsgebieten den Zusammenhang zwischen Niederschlag und Abfluss und deren Dynamik nicht optimal erfassen und abbilden kann. Außerdem spielen konvektive Niederschlagsereignisse in kleinen Einzugsgebieten oft eine wichtige Rolle für die Abflussbildung, wodurch schon geringe räumliche Fehler in der Niederschlagsdatenerfassung erhebliche Auswirkungen auf die Modellgenauigkeit haben. In größeren Einzugsgebieten ist der Einfluss einzelner lokaler Gewitter auf den Gesamtgebietsabfluss hingegen geringer, da sich solche Ereignisse über die Fläche räumlich mitteln und lokale Fehler weniger stark auf die simulierte Abflussdynamik durchschlagen.

Das LSTM-Modell erzielt außerdem in Einzugsgebieten mit ausgeprägter Topographie bessere Ergebnisse als in flachen Einzugsgebieten. Ein wesentlicher Grund hierfür ist die klare hydraulische Struktur in Gebieten mit stärkeren Höhenunterschieden: Einzugsgebietsgrenzen sind topographisch eindeutiger definiert und die Fließwege des gefallenen Niederschlags verlaufen strukturierter. Beispielsweise ist in Mittelgebirgsregionen der Zusammenhang zwischen Niederschlag und resultierender Abflussreaktion meist unmittelbarer sowie zeitlich weniger verzögert und gedämpft als in flachen Tieflandgebieten, wie sie etwa im Norden Deutschlands vorkommen. In diesen flacheren Gebieten erschweren geringe topographische Gradienten, bedeutendere Grundwasserflussprozesse und Unsicherheiten bei der Gebietsabgrenzung die Zuordnung zwischen Niederschlagsereignissen und Abflussreaktionen. Zusätzlich sind diese Gebiete in Deutschland häufig stark durch menschliche Eingriffe geprägt, beispielsweise durch den Einsatz von Bauwerken wie Schleusen, Wehren und anderen wasserwirtschaftlichen Anlagen, was den natürli-

chen Zusammenhang zwischen Niederschlag und Abfluss weiter schwächt und die Simulationsgüte verringert.

Bei weiterer Betrachtung der Simulationsgüte in den einzelnen Bundesländern ergibt sich ein relativ einheitliches Bild. In allen Bundesländern erzielt das LSTM-Modell basierend auf dem NSE-Wert bessere Ergebnisse als das HBV-Modell, die Ausprägung der Differenz zum HBV-Modell variiert dabei aber. Die Simulationsgüte ist lediglich in Brandenburg und in geringerem Maße in Sachsen-Anhalt im Vergleich zu den anderen Bundesländern deutlich verringert. Auch die Varianz der NSE-Werte für das LSTM- und das HBV-Modell ist in diesen Bundesländern erhöht. In Brandenburg befinden sich zudem 5 der Einzugsgebiete, in denen das LSTM-Modell einen NSE-Wert < 0 aufweist, was 41,7 % aller Einzugsgebiete mit negativem NSE-Wert entspricht. Für das HBV-Modell trifft dies auf 8 Einzugsgebiete in Brandenburg zu, d. h. 25 % der Gebiete mit negativem NSE-Wert.

Dies kann als Indiz für Qualitätsprobleme in den CAMELS-DE Daten für Brandenburg gelten. Generell stellt sich nun die Frage, ob abweichende hydrologische Bedingungen – beispielsweise bedingt durch die geringe Topographie, die vorherrschenden sandigen Böden und die starke menschliche Beeinflussung in Brandenburg – zu der verringerten Leistung und den Leistungsschwankungen zwischen den Bundesländern führen, oder ob diese Aspekte vor allem durch unterschiedliche Datenqualitäten in den Bundesländern zu erklären sind.

Allgemein lässt sich sagen, dass CAMELS-DE zwar für ganz Deutschland Daten in einheitlichem Format bereitstellt; da jedes Bundesland jedoch ein eigenes Pegelmessnetz betreibt, kann je nach verfügbaren Mitteln für Betrieb und Qualitätskontrolle auch die Quantität und Qualität der Abflussdaten unterschiedlich sein. Auch die meteorologischen Daten sind nur eingeschränkt konsistent, da der HYRAS-Datensatz auf interpolierten Bodenstationsdaten des Deutschen Wetterdienstes beruht, wobei das Messnetz und die Anzahl der Stationen im Zeitraum von 1951 bis 2020 stark variierten (LORITZ et al., 2024). Somit sind die Qualität und die Unsicherheit der meteorologischen Eingangsdaten sowohl zeitlich als auch räumlich nicht konstant und variieren auch von Einzugsgebiet zu Einzugsgebiet.

Trotzdem wurde gezeigt, dass anhand von CAMELS-DE ein robustes regionales LSTM-Netzwerk mit guter Gesamtsimulationsgüte trainiert werden kann, welches in einem großen Teil der Bundesländer und unter unterschiedlichen hydrologischen Bedingungen, auch im Vergleich mit einem konzeptionellen Modell, gute Ergebnisse erzielt. Auch das einzugsgebietsweise kalibrierte HBV-Modell hat in einem Großteil der Einzugsgebiete in CAMELS-DE zufriedenstellende Simulationsergebnisse geliefert, was als Indikator für eine überwiegend zufriedenstellende Datenqualität von CAMELS-DE gesehen werden kann.

3.) Wo liegen die jeweiligen Grenzen dieser Ansätze, insbesondere in Einzugsgebieten mit fehlerhaften, limitierten oder fehlenden Abflussdaten?

Sowohl das LSTM- als auch das HBV-Modell sind auf qualitativ hochwertige Eingangsdaten angewiesen. Dabei benötigt das LSTM-Modell zusätzlich eine möglichst große Menge an Trainingsdaten (KRATZERT et al., 2024), um den Zusammenhang zwischen meteorologischen Variablen wie Niederschlag und dem

resultierenden Abfluss zuverlässig zu erlernen. In Abbildung 1a wurde gezeigt, dass LSTM-Modelle, die nicht regional, sondern anhand einzelner Einzugsgebiete trainiert wurden, eine dem HBV-Modell ähnliche Simulationsgüte erzielen. Diese Ergebnisse lassen darauf schließen, dass die im Vergleich zum HBV-Modell erhöhte Anzahl der Parameter bzw. Modellgewichte des LSTM-Netzwerks nicht direkt ausschlaggebend für dessen höhere Simulationsgüte ist. Vielmehr führen das regionale Modelltraining und die Fähigkeit des LSTM-Modells, große Datensätze zu verarbeiten und von einer Vielzahl von Einzugsgebieten gleichzeitig zu lernen, zu dessen erhöhter Leistung.

Dieser Vorteil des regional trainierten LSTM-Modells verdeutlicht allerdings auch, dass für das Erreichen einer optimalen Simulationsgüte ein großer Umfang an Trainingsdaten benötigt wird. Bei Studien in nur einem oder wenigen Einzugsgebieten ist also keine generelle Steigerung der Simulationsgüte durch den Einsatz von LSTM-Netzwerken zu erwarten. Für solche lokalen Studien ließe sich jedoch ein LSTM-Modell mit großen Datensätzen regional trainieren und lokal anwenden. So gingen beispielsweise JOHNEN et al. (2025) vor, um Abflusssimulationen für sechs Einzugsgebiete im Harz zu erstellen, welche nicht Teil von CAMELS-DE sind. Die Simulationsgüte des LSTM-Modells ließ sich hier erhöhen, indem das Training nicht nur anhand dieser sechs Einzugsgebiete des Untersuchungsgebiets erfolgte, sondern zusätzlich der gesamte CAMELS-DE Datensatz für ein regionales Modelltraining verwendet wurde. Ein erneutes Training des regionalen LSTM-Modells ausschließlich mit den Einzugsgebieten des Untersuchungsgebiets („fine-tuning“) ermöglichte es, die Simulationsgüte weiter zu erhöhen.

Das HBV-Modell hingegen erfordert eine ausreichende Datenmenge für die individuelle Kalibrierung der Modellparameter je Einzugsgebiet. Konzeptionelle hydrologische Modelle ließen sich zwar ebenfalls global kalibrieren, allerdings nimmt dabei die Simulationsgüte in den einzelnen Einzugsgebieten ab (KRATZERT et al., 2024). Wie in der vorangegangenen Diskussion gezeigt, bietet CAMELS-DE für die Mehrheit der untersuchten Einzugsgebiete hydrometeorologische Daten in ausreichendem Umfang und meist hoher Qualität, sodass sowohl LSTM- als auch HBV-Modell in der Regel sinnvolle Simulationsergebnisse erzielen. Allerdings können sowohl meteorologische als auch hydrologische Eingangsdaten fehlerhaft oder unvollständig sein. Ein zentrales Problem stellen hierbei Massenbilanzfehler dar. Wie zu erwarten und in Kapitel 3.4 dargelegt, erzielt das LSTM-Netzwerk in Einzugsgebieten mit größeren Massenbilanzfehlern insgesamt robustere Ergebnisse als das HBV-Modell. Da das HBV-Modell aufgrund seiner konzeptionellen Struktur auf Massenerhaltung ausgelegt ist, reagiert es empfindlich auf Eingabefehler: Ein positiver Massenbilanzfehler führt beispielsweise dazu, dass sich die Speicher des Modells kontinuierlich auffüllen, was eine realistische Abflusssimulation zunehmend erschwert oder gar unmöglich macht. Ein stark negativer Massenbilanzfehler bedeutet wiederum, dass im Modell langfristig zu wenig Wasser verfügbar ist, wodurch Abflüsse systematisch unterschätzt werden.

Dies verdeutlicht zugleich die Bedeutung der Datensatzlänge in CAMELS-DE von bis zu 70 Jahren: Während Anpassungen der Kalibrierung oder der Initialisierung solche Fehler für kurze Zeiträume von wenigen Jahren oft kompensieren können, zeigen sich die Auswirkungen ungenauer Eingangsdaten in längeren Zeitreihen ab 20 Jahren deutlich stärker und können vom HBV-

Modell nicht mehr ausgeglichen werden. Es ist hierbei anzumerken, dass Versionen des HBV-Modells existieren, in denen Parameter für eine Korrektur des Niederschlags und der Evapotranspiration kalibriert werden (BERGSTRÖM, 1992). Mit diesen Parametern lässt sich die Simulationsgüte des HBV-Modells bei Eingangsdaten mit Massenbilanzfehlern potenziell verbessern, allerdings würde hierdurch auch die Interpretierbarkeit abnehmen. Die hier verwendete HBV-Version verfügt nicht über solche Korrekturparameter.

Die massenerhaltende Struktur des HBV-Modells ermöglicht grundsätzlich eine bessere physikalische Interpretierbarkeit der Ergebnisse, vorausgesetzt, alle relevanten Eingangs- und Ausgangsgrößen (Niederschlag, Abfluss und Evapotranspiration) werden korrekt erfasst. Vor allem die Abschätzung der Evapotranspiration stellt jedoch eine erhebliche Unsicherheitsquelle dar, da diese Variable großskalig kaum messbar ist und somit meist nur empirisch approximiert oder indirekt ermittelt werden kann. Unterschiedliche Methoden zur Berechnung der potenziellen Evapotranspiration können dabei stark voneinander abweichende Ergebnisse liefern (CLERC-SCHWARZENBACH et al., 2024; LORITZ et al., 2024).

LSTM-Netzwerke benötigen die Evapotranspiration nicht explizit als Eingangsvariable, erhalten aber alle Variablen, die typischerweise für deren Abschätzung verwendet werden. Das Netzwerk kann somit Verdunstungsprozesse indirekt erfassen und deren Relevanz für die Abflusssimulation datengetrieben bestimmen oder vernachlässigen. Da LSTM-Modelle grundsätzlich keine expliziten physikalischen Annahmen oder Randbedingungen erfüllen müssen, abgesehen von Zusammenhängen zwischen Inputs und Outputs, sind sie nicht per se massenerhaltend. Dieser Umstand ist einerseits ein Kritikpunkt hinsichtlich der physikalischen Interpretierbarkeit, stellt andererseits jedoch eine Stärke dar: Die größere Flexibilität ermöglicht es LSTM-Netzwerken, selbst bei unvollständigen, fehlerhaften oder unsicheren Eingangsdaten sinnvolle und robuste Abflusssimulationen zu generieren, solange ein Zusammenhang zwischen den meteorologischen Eingangsvariablen und dem Abfluss besteht. Diese Annahme wird durch die Ergebnisse von FRAME et al. (2022) gestützt, die zeigen, dass eine forcierte Massenerhaltung in LSTM-Modellen zu einer deutlichen Verschlechterung der Modellleistung im Vergleich zum klassischen, rein datengetriebenen LSTM-Ansatz führen kann.

Bezüglich limitierten Abflussdaten wurde in Kapitel 3.5 gezeigt, dass das LSTM-Modell durch das regionale Modelltraining auch Simulationen in Einzugsgebieten ohne jegliche Abflussdaten im Trainingszeitraum erstellen kann. Prinzipiell ist es möglich, auch ein HBV-Modell in unbeobachteten Einzugsgebieten anzuwenden. Hierfür sind jedoch Regionalisierungsmethoden nötig, bei denen beispielsweise Parameter von ähnlichen, beobachteten Einzugsgebieten transferiert werden (POOL et al., 2021) oder ein globaler Parametersatz mithilfe von Deep-Learning-Hybridmethoden kalibriert wird (FENG et al., 2022).

In der Basiskonfiguration des HBV-Modells, wie es diese Studie verwendet, werden die Parameter für jedes Einzugsgebiet separat kalibriert, wodurch keine Simulationen in gänzlich unbeobachteten Einzugsgebieten erstellt werden können. Die Ergebnisse des LSTM-Modells liegen, mit einem Median-NSE-Wert von 0,69 in unbeobachteten Einzugsgebieten, in einem ähnlichen Bereich wie der Median-NSE-Wert von 0,72, den das HBV-Modell

in beobachteten Einzugsgebieten mit mindestens 20 % verfügbaren Abflussdaten im Trainingszeitraum erreicht. Die Simulationen des LSTM-Netzwerks haben also auch in unbeobachteten Einzugsgebieten eine gute Qualität, allerdings ist die Simulationsgüte deutlich geringer als in Einzugsgebieten mit mehr als 20 % verfügbaren Trainingsdaten, wo der NSE-Wert 0,86 beträgt. Bei Betrachtung der Ergebnisse für Einzugsgebiete mit > 0 bis 20 % verfügbaren Abflussdaten im Trainingszeitraum wird klar, dass schon eine begrenzte Menge an Trainingsdaten die Simulationsgüte deutlich verbessert. So steigt der Median-NSE-Wert für das LSTM-Modell deutlich auf einen Wert von 0,80 und auch die Varianz der NSE-Werte nimmt deutlich ab. Das HBV-Modell erreicht für Einzugsgebiete mit limitierten Abflussdaten, bis zu 20 %, im Median einen NSE-Wert von 0,63.

Die Ergebnisse zeigen, dass sowohl das LSTM- als auch das HBV-Modell für einen Großteil der Einzugsgebiete auch bei einer begrenzten Menge an Trainingsdaten sinnvolle Simulationen erstellen können. Diese Erkenntnisse stimmen mit den Arbeiten von SEIBERT & BEVEN (2009), POOL et al. (2017) und ETTER et al. (2018) überein, in denen gezeigt wurde, dass schon unsichere und stark begrenzte Mengen an Abflussmessungen ausreichen, um die Simulationsgüte konzeptioneller hydrologischer Modelle zu verbessern. Auch das LSTM-Modell profitiert davon, wenn zumindest ein geringer Umfang an Trainingsdaten für das Einzugsgebiet zur Verfügung steht, allerdings ist der Umfang der Trainingsdaten in einzelnen Einzugsgebieten weniger entscheidend als im Falle des HBV-Modells. Für beide Modelle nimmt vor allem auch die Varianz der NSE-Werte mit der steigenden Verfügbarkeit von Trainingsdaten ab, was jedoch auch damit zu tun haben kann, dass die Daten von Einzugsgebieten mit kurzen Zeitreihendaten an sich fehlerbehafteter sein könnten als in Einzugsgebieten mit langjährig verfügbaren Pegelraten.

Die Tatsache, dass das LSTM-Modell auch in gänzlich unbeobachteten Einzugsgebieten für den Großteil der Einzugsgebiete gute Simulationen erzeugen kann, spricht dafür, dass zumindest ein gewisses Maß an Generalisierung durch das regionale Training des LSTM-Netzwerks erreicht wird. Ob das Modell dabei in der Lage ist, anhand der statischen Einzugsgebietseigenschaften erlernte Informationen zwischen ähnlichen Einzugsgebieten zu transferieren, oder ob lediglich allgemeingültige Beziehungen zwischen Niederschlag und Abfluss angewandt werden, steht dabei zur Debatte (HEUDORFER et al., 2025). Der Einsatz von LSTM-Netzwerken für die Abflusssimulation in unbeobachteten Einzugsgebieten birgt in jedem Fall großes Potenzial (KRATZERT et al., 2019b), was auch die Ergebnisse dieser Studie zeigen.

Die stark erhöhte Varianz der NSE-Werte im Vergleich zu beobachteten Einzugsgebieten und der Umstand, dass bei Simulationen in tatsächlich unbeobachteten Einzugsgebieten, also ohne Datenverfügbarkeit im Testzeitraum, keine Validierung der Simulationsergebnisse möglich ist, macht eine Abschätzung der Unsicherheit des simulierten Abflusses unabdingbar. Dies kann durch probabilistische Simulationen erreicht werden; im Falle von LSTM-Modellen bietet sich dabei der Einsatz von „Mixture Density“-Netzwerken an, welche die Momente von Verteilungen des Abflusses anstelle von deterministischen Abflusswerten für jeden Zeitschritt simulieren (KLOTZ et al., 2022).

Weiterhin bleibt das Problem der fehlenden Extrapolationsfähigkeit von datengetriebenen Ansätzen bestehen. Simulationen

außerhalb des Wertebereichs der Trainingsdaten sind zwar theoretisch möglich, aber mit hohen Unsicherheiten belastet. Ansätze aus der hybriden Modellierung, bei denen datengetriebene Modelle um physikalische Prozesse erweitert oder mit konzeptionellen hydrologischen Modellen kombiniert werden, können dieses Problem lösen. Allerdings führt dieses Vorgehen bisher meist zu einer Verringerung der allgemeinen Simulationsgüte im Vergleich zu rein datengetriebenen Methoden (KRATZERT et al., 2024).

Ein weiterer Ansatz, der dieses Problem indirekt zu lösen versucht, ist die Erweiterung der Trainingsdaten, sodass aus dem Extrapolationsproblem ein Interpolationsproblem wird. So untersuchen beispielsweise MARTEL et al. (2025) den Nutzen von LSTM-Modellen für Klimaänderungsstudien. Hierbei erzielt ein LSTM-Modell, welches mit Daten aus Einzugsgebieten in diversen Klimazonen trainiert wurde, deutlich bessere Ergebnisse als konzeptionelle Modelle oder ein LSTM-Modell, welches lediglich mit Einzugsgebieten im Untersuchungsgebiet trainiert wurde. Dieser Ansatz ist zwar vielversprechend, geht jedoch auch mit erhöhtem Datenbedarf einher, der durch die weiter fortschreitende Verfügbarkeit von großen, frei verfügbaren hydrometeorologischen Datensätzen wie CAMELS erfüllt wird. Außerdem steigen hierdurch die Komplexität des Modells und der Ressourcenbedarf für das Training. Zusätzlich kann der Ansatz, die Trainingsdaten zu erweitern, zwar die Ergebnisse in einem Untersuchungsgebiet verbessern, das Extrapolationsproblem an sich wird dadurch aber nicht gelöst.

Zusammenfassend lässt sich sagen, dass sowohl das LSTM- als auch das HBV-Modell auch bei einem geringen Umfang von Trainingsdaten in den meisten Fällen gute Simulationen erstellen können, wobei beide Modelle von umfangreichen Trainingsdaten stark profitieren. Dies deckt sich mit den Untersuchungen von STAUDINGER et al. (2025). Das LSTM-Modell hat die Möglichkeit, in Bezug auf die Massenbilanz fehlerhafte Eingangsdaten zu kompensieren. Das HBV-Modell stößt hier an seine Grenzen, bietet in diesem Aspekt allerdings eine bessere hydrologische Interpretierbarkeit, mit dem Nachteil, dass die Verdunstung abgeschätzt werden muss, was mit großen Unsicherheiten verbunden ist. Insgesamt erweist sich der Vergleich der beiden Modellkonzepte in dieser Studie als hilfreich, um Einzugsgebiete mit fehlerhaften Daten oder beispielsweise durch Bauwerke regulierte Abflüsse zu identifizieren, die kaum zu modellieren sind, aber auch, um die Schwächen des jeweiligen Ansatzes durch die Stärken des anderen auszugleichen.

5. Zusammenfassung und Ausblick

Diese Studie zeigt, dass datengetriebene, regional trainierte LSTM-Netzwerke bei der Niederschlag-Abfluss-Simulation in Deutschland eine bessere Simulationsgüte erzielen als konzeptionelle HBV-Modelle. Insbesondere die Simulation von Hochwasserscheitelabflüssen gelingt mit dem LSTM-Modell präziser und zeitlich genauer. Dennoch besitzen beide Modellansätze spezifische Stärken: Während das HBV-Modell durch seine hohe Interpretierbarkeit, Extrapolationsfähigkeit und langjährig bewährte Praxis überzeugt und damit zu einem besseren Systemverständnis beitragen kann, punktet das regional trainierte LSTM-Netzwerk mit der in nahezu allen Einzugsgebieten höheren Simulationsgüte und mit der einfachen Anwendbarkeit in unbeobachteten Einzugsgebieten.

Die Ergebnisse legen daher nahe, dass LSTM-Modelle konzeptionelle Modelle nicht ersetzen, sondern als leistungsfähige Ergänzung in die hydrologische Praxis integriert werden sollten. Da sich LSTM-Modelle mit denselben Eingangsdaten wie konzeptionelle Modelle betreiben lassen, ist auch der praktische Aufwand hierfür überschaubar. Weiterhin wurde die Bedeutung des regionalen Trainings gezeigt. LSTM-Modelle benötigen zwar einen großen Umfang an Trainingsdaten, welche in Deutschland jedoch in ausgezeichneter Qualität und Quantität vorhanden sind. Die Antwort auf die Titelfrage „Lernen oder kalibrieren?“ lautet somit klar: „Lernen und kalibrieren!“

Gleichzeitig zeigt sich, dass für den operationellen Einsatz von LSTM-Netzwerken in der Hochwasserprognose, insbesondere in kleinen und schnell reagierenden Einzugsgebieten, die tägliche Auflösung des hier verwendeten CAMELS-DE Datensatzes nicht ausreichend ist. Um datengetriebene Modelle gezielt für die Kurzfrist-Hochwasservorhersage in Deutschland entwickeln und trainieren zu können, wurde inzwischen der CAMELS-DE-1h Datensatz in stündlicher Auflösung erstellt (DOLICH et al., 2026). Erste Ansätze zur Nutzung solcher hochaufgelösten Daten zeigten sich in der Entwicklung der „multi-frequency LSTM (MF-LSTM)“-Architektur durch ACUÑA ESPINOZA et al. (2025a), die speziell für Simulationen und Vorhersagen auf Stundenbasis konzipiert wurde. Mit dem stündlichen CAMELS-DE-1h-Datensatz kann dieses Modell zukünftig umfassend getestet, weiterentwickelt und deutschlandweit eingesetzt werden.

Conclusion and Outlook

Overall, this study shows that data-driven LSTM networks achieve better simulation quality than the conceptual HBV model in rainfall-runoff simulation in Germany. In particular, the simulation of flood peak discharges is more precise and temporally accurate with the LSTM model. Nevertheless, both modelling approaches have their own specific strengths: the HBV model is characterized by its high interpretability, extrapolation capability and long-standing proven practice, thereby contributing to a better understanding of the system, while the regionally trained LSTM network demonstrates higher simulation quality in almost all catchments and can directly be applied in ungauged catchments. The results, therefore, suggest that LSTM models should not replace conceptual models, but should be integrated into hydrological practice as a powerful tool to supplement them. Since LSTM models can be operated with the same input data as conceptual models, the practical effort involved is also manageable. Furthermore, the importance of regional training has been demonstrated. Although LSTM models require a large amount of training data, this data is available in Germany in excellent quality and quantity. The answer to the question in the title “To learn or to calibrate?” is therefore: “Learn and calibrate!”

At the same time, it is clear that the daily resolution of the CAMELS-DE dataset used in this study is insufficient for operational use of LSTM networks in flood forecasting, especially in small and fast-responding catchments. In order to develop and train data-driven models specifically for short-term flood forecasting in Germany, the CAMELS-DE-1h dataset with hourly resolution was created (DOLICH et al., 2026). Initial approaches to using such high-resolution data can be seen in the development of the multi-frequency LSTM (MF-LSTM) architecture by ACUÑA ESPINOZA et al. (2025a), which was specifically designed for hourly

simulations and forecasts. The corresponding hourly CAMELS-DE dataset allows comprehensive testing, further development and the Germany-wide use of this model in the future.

Daten- und Code-Verfügbarkeit

Der CAMELS-DE Datensatz wurde im Zuge dieser Veröffentlichung unter anderem um die Ensemble-LSTM-Ergebnisse erweitert (v1.1.0) und ist zusammen mit einer ausführlichen Datensatzbeschreibung frei verfügbar unter DOLICH et al. (2025). Die verwendete LSTM-Architektur sowie der vollständige Code zur Trainings- und Testdurchführung sind im Python-Paket „Hy2DL“ enthalten, welches ebenfalls frei verfügbar ist (ACUÑA ESPINOZA et al., 2024b).

Danksagung

Ein großer Dank geht an das CAMELS-DE Team, ohne dessen Initiative, Arbeit und Expertise CAMELS-DE und damit diese Studie nicht möglich gewesen wären.

Die Autoren bedanken sich außerdem für die Unterstützung durch das Land Baden-Württemberg in Form des Service bwHPC (Baden-Württemberg High Performance Computing).

Anschriften der Verfasser

Alexander Dolich, M.Sc.
PD Dr.-Ing. Uwe Ehret
Dr. rer. nat. Ralf Lortz
Karlsruher Institut für Technologie (KIT)
Institut für Wasser und Umwelt
Kaiserstraße 12, 76131 Karlsruhe
alexander.dolich@kit.edu
uwe.ehret@kit.edu
ralf.lortz@kit.edu

Literaturverzeichnis

- ACUÑA ESPINOZA, E., R. LORITZ, M. ÁLVAREZ CHAVES, N. BÄUERLE & U. EHRET (2024a): To bucket or not to bucket? Analyzing the performance and interpretability of hybrid hydrological models with dynamic parameterization. – *Hydrology and Earth System Sciences*, 28, (12), 2705–2719. DOI: 10.5194/hess-28-2705-2024.
- ACUÑA ESPINOZA, E., A. MANOJ, M. ÁLVAREZ CHAVES, M. MÄLICKÉ & S. BASTE (2024b): KIT-HYD/Hy2DL: Hybrid Hydrological modeling using Deep Learning methods. <https://zenodo.org/doi/10.5281/zenodo.12800460>, zuletzt aufgerufen am 16.7.2025.
- ACUÑA ESPINOZA, E., F. KRATZERT, D. KLOTZ, M. GAUCH, M. ÁLVAREZ CHAVES, R. LORITZ & U. EHRET (2025a): Technical note: An approach for handling multiple temporal frequencies with different input dimensions using a single LSTM cell. – *Hydrology and Earth System Sciences*, 29(6), 1749–1758. DOI: 10.5194/hess-29-1749-2025.
- ACUÑA ESPINOZA, E., R. LORITZ, F. KRATZERT, D. KLOTZ, M. GAUCH, M. ÁLVAREZ CHAVES & U. EHRET (2025b): Analyzing the generalization capabilities of a hybrid hydrological model for extrapolation to extreme events. – *Hydrology and Earth System Sciences*, 29(5), 1277–1294. DOI: 10.5194/hess-29-1277-2025.
- ADDOR, N., A.J. NEWMAN, N. MIZUKAMI & M.P. CLARK (2017): The CAMELS data set: catchment attributes and meteorology for large-sample studies. – *Hydrology and Earth System Sciences*, 21(10), 5293–5313. DOI: 10.5194/hess-21-5293-2017.
- BASTE, S., D. KLOTZ, E.A. ESPINOZA, A. BÁRDOSSY & R. LORITZ (2025): Unveiling the Limits of Deep Learning Models in Hydrological Extrapolation Tasks: Catchment hydrology/Modelling approaches. – *Hydrology and Earth System Sciences*, 29(21), 5871–5891. DOI: 10.5194/hess-29-5871-2025.
- BERGSTRÖM, S. (1992): The HBV Model: Its Structure and Applications. Sveriges meteorologiska och hydrologiska institut (SMHI).
- BERGSTRÖM, S. (2006): Experience from applications of the HBV hydrological model from the perspective of prediction in ungauged basins. – ANDREASSIAN, V., A. HALL, N. CHAHINIAN & J. SCHAAKE (Hrsg.): Large Sample Basin Experiments for Hydrological Model Parameterization: Results of the Model Parameter Experiment—MOPEX. Iahs Aish Publications, 307. 97–107.
- BERGSTRÖM, S. & A. FORSMAN (1973): Development of a conceptual deterministic rainfall-runoff model. – *Hydrology Research*, 4(3), 147–170. DOI: 10.2166/nh.1973.0012.
- BEVEN, K. & A. BINLEY (1992): The future of distributed models: Model calibration and uncertainty prediction. – *Hydrological Processes*, 6(3), 279–298. DOI: 10.1002/hyp.3360060305.
- CLERC-SCHWARZENBACH, F., G. SELLERI, M. NERI, E. TOTH, I. VAN MEERVELD & J. SEIBERT (2024): Large-sample hydrology – a few camels or a whole caravan? – *Hydrology and Earth System Sciences*, 28(17), 4219–4237. DOI: 10.5194/hess-28-4219-2024.
- DE LA FUENTE, L.A., M.R. EHSANI, H.V. GUPTA & L.E. CONDON (2024): Toward interpretable LSTM-based modeling of hydrological systems. – *Hydrology and Earth System Sciences*, 28(4), 945–971. DOI: 10.5194/hess-28-945-2024.
- DOLICH, A., E.A. ESPINOZA, U. EHRET, M. KRAFT, J. BONDY, J. MEUER, & R. LORITZ (2026): CAMELS-DE-1h: Hourly hydro-meteorological time series, weather forecasts, and attributes for 1611 catchments in Germany. GFZ Data Services, DOI: 10.5880/fidgeo.2026.045.
- DOLICH, A., E.A. ESPINOZA, P. EBELING, B. GUSE, J. GÖTTE ... R. LORITZ (2025): CAMELS-DE: Hydrometeorological time series and attributes for 1582 catchments in Germany. <https://zenodo.org/doi/10.5281/zenodo.16755906>.
- DROOGERS, P. & R.G. ALLEN (2002): Estimating Reference Evapotranspiration Under Inaccurate Data Conditions. – *Irrigation and Drainage Systems*, 16(1), 33–45. DOI: 10.1023/A:1015508322413.
- ETTER, S., B. STROBL, J. SEIBERT & H.J.I. VAN MEERVELD (2018): Value of uncertain streamflow observations for hydrological modelling. – *Hydrology and Earth System Sciences*, 22(10), 5243–5257. DOI: 10.5194/hess-22-5243-2018.
- FENG, D., J. LIU, K. LAWSON & C. SHEN (2022): Differentiable, Learnable, Regionalized Process-Based Models With Multiphysical Outputs can Approach State-Of-The-Art Hydrologic Prediction Accuracy. – *Water Resources Research*, 58(10), e2022WR032404. DOI: 10.1029/2022WR032404.
- FENG, D., H. BECK, J. DE BRUIJN, R.K. SAHU, Y. SATOH ... C. SHEN (2024): Deep dive into hydrologic simulations at global scale: harnessing the power of deep learning and physics-informed differentiable models (δ HBV-globe1.0-hydroDL). – *Geoscientific Model Development*, 17(18), 7181–7198. DOI: 10.5194/gmd-17-7181-2024.
- FRAME, J.M., F. KRATZERT, A. RANEY, M. RAHMAN, F.R. SALAS & G.S. NEARING (2021): Post-Processing the National Water Model with Long Short-Term Memory Networks for Streamflow Predictions and Model Diagnostics. – *JAWRA, Journal of the American Water Resources Association*, 57(6), 885–905. DOI: 10.1111/1752-1688.12964.
- FRAME, J.M., F. KRATZERT, D. KLOTZ, M. GAUCH, G. SHALEV ... G.S. NEARING (2022): Deep learning rainfall-runoff predictions of extreme events. – *Hydrology and Earth System Sciences*, 26(13), 3377–3392. DOI: 10.5194/hess-26-3377-2022.
- 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. DOI: 10.1016/j.jhydrol.2009.08.003.
- HARLIN, J. (1992): Modelling the Hydrological Response of Extreme Floods in Sweden. – *Hydrology Research*, 23(4), 227–244. DOI: 10.2166/nh.1992.0016.

- HEUDORFER, B., H.V. GUPTA & R. LORITZ (2025): Are Deep Learning Models in Hydrology Entity Aware? – Geophysical Research Letters, 52(6), e2024GL113036. DOI: 10.1029/2024GL113036.
- HOCHREITER, S. & J. SCHMIDHUBER (1997): Long Short-Term Memory. – Neural Computation, 9(8), 1735-1780. DOI: 10.1162/neco.1997.9.8.1735.
- HOEDT, P.-J., F. KRATZERT, D. KLOTZ, C. HALMICH, M. HOLZLEITNER ... G. KLAMBAUER (2021): MC-LSTM: Mass-Conserving LSTM. – PLMR, Proceedings of Machine Learning Research, 4275-4286. <https://proceedings.mlr.press/v139/hoedt21a.html>, zuletzt aufgerufen am 25.11.2025.
- HOUSKA, T., P. KRAFT, A. CHAMORRO-CHAVEZ & L. BREUER (2015): SPOTting Model Parameters Using a Ready-Made Python Package. – PLOS ONE, 10(12), e0145180. DOI: 10.1371/journal.pone.0145180.
- JOHNEN, G., A. NIEMANN, P. NISTAHL & A. HUTWALKER (2025): Deep Learning zur Optimierung der Trinkwasserversorgung aus Talsperren. – Wasser und Abfall, 27(5), 12-17. DOI: 10.1007/s35152-025-1982-z.
- KLOTZ, D., F. KRATZERT, M. GAUCH, A. KEEFE SAMPSON, J. BRANDS-TETTER ... G. NEARING (2022): Uncertainty estimation with deep learning for rainfall-runoff modeling. – Hydrology and Earth System Sciences, 26(6), 1673-1693. DOI: 10.5194/hess-26-1673-2022.
- 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. DOI: 10.5194/hess-22-6005-2018.
- KRATZERT, F., M. HERRNEGGER, D. KLOTZ, S. HOCHREITER & G. KLAMBAUER (2019a): NeuralHydrology – Interpreting LSTMs in Hydrology. – SAMEK, W., G. MONTAVON, A. VEDALDI, L.K. HANSEN & K.-R. MÜLLER (Hrsg.): Lecture Notes in Computer Science. Springer International, 11700, 347-362. http://link.springer.com/10.1007/978-3-030-28954-6_19, zuletzt aufgerufen am 23.4.2025.
- KRATZERT, F., D. KLOTZ, M. HERRNEGGER, A.K. SAMPSON, S. HOCHREITER & G.S. NEARING (2019b): Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning. – Water Resources Research, 55(12), 11344-11354. DOI: 10.1029/2019WR026065.
- KRATZERT, F., D. KLOTZ, G. SHALEV, G. KLAMBAUER, S. HOCHREITER & G. NEARING (2019c): Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. – Hydrology and Earth System Sciences, 23(12), 5089-5110. DOI: 10.5194/hess-23-5089-2019.
- KRATZERT, F., D. KLOTZ, S. HOCHREITER & G.S. NEARING (2021): A note on leveraging synergy in multiple meteorological data sets with deep learning for rainfall-runoff modeling. – Hydrology and Earth System Sciences, 25(5), 2685-2703. DOI: 10.5194/hess-25-2685-2021.
- KRATZERT, F., M. GAUCH, D. KLOTZ & G. NEARING (2024): HESS Opinions: Never train a Long Short-Term Memory (LSTM) network on a single basin. – Hydrology and Earth System Sciences, 28(17), 4187-4201. DOI: 10.5194/hess-28-4187-2024.
- LEES, T., M. BUECHEL, B. ANDERSON, L. SLATER, S. REECE, G. COXON & S.J. DADSON (2021): Benchmarking data-driven rainfall-runoff models in Great Britain: a comparison of long short-term memory (LSTM)-based models with four lumped conceptual models. – Hydrology and Earth System Sciences, 25(10), 5517-5534. DOI: 10.5194/hess-25-5517-2021.
- LEES, T., S. REECE, F. KRATZERT, D. KLOTZ, M. GAUCH ... S.J. DADSON (2022): Hydrological concept formation inside long short-term memory (LSTM) networks. – Hydrology and Earth System Sciences, 26(12), 3079-3101. DOI: 10.5194/hess-26-3079-2022.
- LIU, J., Y. BIAN, K. LAWSON & C. SHEN (2024): Probing the limit of hydrologic predictability with the Transformer network. – Journal of Hydrology, 637, 131389. DOI: 10.1016/j.jhydrol.2024.131389.
- LORITZ, R., A. DOLICH, E. ACUÑA ESPINOZA, P. EBELING, B. GUSE ... L. TARASOVA (2024): CAMELS-DE: Hydro-meteorological time series and attributes for 1582 catchments in Germany. – Earth System Science Data, 16(12), 5625-5642. DOI: 10.5194/essd-16-5625-2024.
- MARTEL, J.-L., F. BRISSETTE, R. ARSENAULT, R. TURCOTTE, M. CASTAÑEDA-GONZALEZ ... L.-P. CARON (2025): Assessing the adequacy of traditional hydrological models for climate change impact studies: a case for long short-term memory (LSTM) neural networks. – Hydrology and Earth System Sciences, 29(13), 2811-2836. DOI: 10.5194/hess-29-2811-2025.
- MUÑOZ-CASTRO, E., B.J. ANDERSON, P.C. ASTAGNEAU, D.L. SWAIN, P.A. MENDOZA & M.I. BRUNNER (2026): How well do hydrological models simulate streamflow extremes and drought-to-flood transitions? – Hydrology and Earth System Sciences, 30(3), 825-848. DOI: 10.5194/hess-30-825-2026.
- NASH, J.E. & J.V. SUTCLIFFE (1970): River flow forecasting through conceptual models part I — A discussion of principles. – Journal of Hydrology, 10(3), 282-290. DOI: 10.1016/0022-1694(70)90255-6.
- NEARING, G.S., F. KRATZERT, A.K. SAMPSON, C.S. PELISSIER, D. KLOTZ ... H.V. GUPTA (2021): What Role Does Hydrological Science Play in the Age of Machine Learning? – Water Resources Research, 57(3), e2020WR028091. DOI: 10.1029/2020WR028091.
- NEARING, G., D. COHEN, V. DUBE, M. GAUCH, O. GILON, Y. MATIAS (2024): Global prediction of extreme floods in ungauged watersheds. – Nature, 627(8004), 559-563. DOI: 10.1038/s41586-024-07145-1.
- NEWMAN, A.J., M.P. CLARK, K. SAMPSON, A. WOOD, L.E. HAY ... Q. DUAN (2015): Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: Data set characteristics and assessment of regional variability in hydrologic model performance. – Hydrology and Earth System Sciences, 19(1), 209-223. DOI: 10.5194/hess-19-209-2015.
- POOL, S., D. VIVIROLI & J. SEIBERT (2017): Prediction of hydrographs and flow-duration curves in almost ungauged catchments: Which runoff measurements are most informative for model calibration? – Journal of Hydrology, 554, 613-622. DOI: 10.1016/j.jhydrol.2017.09.037.
- POOL, S., M. VIS & J. SEIBERT (2021): Regionalization for Ungauged Catchments — Lessons Learned From a Comparative Large-Sample Study. – Water Resources Research, 57(10), e2021WR030437. DOI: 10.1029/2021WR030437.
- RAUTHE, M., H. STEINER, U. RIEDIGER, A. MAZURKIEWICZ & A. GRATZKI (2013): A Central European precipitation climatology, Part I: Generation and validation of a high-resolution gridded daily data set (HYRAS). – Meteorologische Zeitschrift, 22(3), 235-256. DOI: 10.1127/0941-2948/2013/0436.
- RAZAFIMAHARO, C., S. KRÄHENMANN, S. HÖPP, M. RAUTHE & T. DEUTSCHLÄNDER (2020): New high-resolution gridded dataset of daily mean, minimum, and maximum temperature and relative humidity for Central Europe (HYRAS). – Theoretical and Applied Climatology, 142, (3-4), 1531-1553. DOI: 10.1007/s00704-020-03388-w.
- SEIBERT, J. (2005): HBV Light Version 2, User's Manual. Department of Physical Geography and Quaternary Geology, Stockholm University, Stockholm, https://www.geo.uzh.ch/dam/jcr:c8afa73c-ac90-478e-a8c7-929eed7b1b62/HBV_manual_2005.pdf, zuletzt aufgerufen am 23.4.2025.
- SEIBERT, J. & S. BERGSTRÖM (2022): A retrospective on hydrological catchment modelling based on half a century with the HBV model. – Hydrology and Earth System Sciences, 26(5), 1371-1388. DOI: 10.5194/hess-26-1371-2022.
- SEIBERT, J. & K.J. BEVEN (2009): Gauging the ungauged basin: how many discharge measurements are needed? – Hydrology and Earth System Sciences, 13(6), 883-892. DOI: 10.5194/hess-13-883-2009.

- STAUDINGER, M., A. HERZOG, R. LORITZ, T. HOUSKA, S. POOL ... U. EHRET (2025): How well do hydrological models learn from limited discharge data? A comparison of process- and data-driven models. – *Hydrology and Earth System Sciences*, 29(19), 5005–5029. DOI: 10.5194/hess-29-5005-2025.
- TANRIKULU, O.D., U. EHRET, I. HAAG, R. LORITZ & U. BADDE (2024): Untersuchungen zum Potenzial maschineller Lernverfahren für die hydrologische Simulation und Vorhersage am Beispiel von LSTM und LARSIM in Baden-Württemberg. – *Hydrologie und Wasserbewirtschaftung*, 68(3), 130-145. DOI: 10.5675/HyWa_2024.3_1.
- VASWANI, A., N. SHAZEER, N. PARMAR, J. USZKOREIT, L. JONES ... I. POLO-SUKHIN (2017): Attention Is All You Need. arXiv, <https://arxiv.org/abs/1706.03762>, zuletzt aufgerufen am 16.07.2025.
- VIRTANEN, P., R. GOMMERS, T.E. OLIPHANT, M. HABERLAND, T. REDDY ... Y. VÁZQUEZ-BAEZA (2020): SciPy 1.0: fundamental algorithms for scientific computing in Python. – *Nature Methods*, 17(3), 261-272. DOI: 10.1038/s41592-019-0686-2.
- VRUGT, J.A. (2016): Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation. – *Environmental Modelling & Software*, 75, 273-316. DOI: 10.1016/j.envsoft.2015.08.013.
- YILMAZ, K.K., H.V. GUPTA & T. WAGENER (2008): A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model. – *Water Resources Research*, 44(9), W09417. DOI: 10.1029/2007WR006716.