TUM SCHOOL OF
ENGINEERING AND DESIGN

TECHNISCHE UNIVERSITÄT MÜNCHEN

Deep learning and hybrid modeling of global vegetation and hydrology

Basil Kraft, M.Sc.

Vollständiger Abdruck der von der *TUM School of Engineering and Design* der *Technischen Universität München* zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz	Prof. Dr. rer. nat. Niklas Boers
Prüfer der Dissertation	1. Prof. Dr. rer. nat. habil. Marco Körner
	2. Prof. Dr. rer. nat. Markus Reichstein

Die Dissertation wurde am 24.01.2022 bei der *Technischen Universität München* eingereicht und durch die *TUM School of Engineering and Design* am 02.06.2022 angenommen.

Acknowledgments

Throughout my PhD studies, I received tremendous support from my supervisors, colleagues, friends, and family.

I thank Markus for always being supportive and for giving me the space to follow my interests. Markus taught me to balance perfectionism and pragmatism and how to keep an eye on the target. His broad interdisciplinary knowledge, curiosity, and ability to manage the chaos is truly inspiring.

Martin taught me to first listen, and then talk. He encouraged me to believe in my skills and convinced me again and again that I am on the right path. I will keep our "whiteboard-brainstorming-sessions" in good memory.

Whether it is machine learning, data science, or mathematics, Marco always knew the right answer. He helped me to develop technical skills and taught me that details matter.

I thank all my colleagues from the MPI for the inspiring discussions during day and night, for sharing their knowledge, and for motivating me. Special thanks to Nuno, Sujan, Simon, Christian, and José for the in-depth discussions and collaboration. Thank you, Uli, for the supply of datasets and drinks, and for the many good and bad jokes.

As a member of the IMPRS-gBGC graduate school, I experienced tremendous support from Steffi, John, and Stefanie. You are awesome!

I always enjoyed visiting my colleagues at the CVRG group in Munich, and I will keep our summer schools in good memory.

During my PhD studies, I had the chance to visit Devis Tuia's lab at the EPFL. I enjoyed this time very much and want to express my gratitude for this opportunity.

I want to thank Jake, Antonios, Jasper, Tina, Tiana, Jeff, Shane, and Santiago for giving me a reason to take a break from work.

Barbara and Ueli supported me throughout my entire life, regardless of my choices. I never took this unconditional support for granted, and I will never forget how much you gave me. Thank you!

Elena, you helped me through the hard times, and made the good times even better. I am very, very happy to have you, and I am tremendously grateful for your support.

Summary

Background

In the face of climate change, a better understanding and representation of processes at the land surface is needed, for example, to improve projections of the carbon balance of terrestrial ecosystems with increasing temperatures. While expert models provide valuable insights into land surface processes, these models are often coarse-grained, inflexible, and biased towards prior knowledge. Nevertheless, physically-based models are still the pivotal tool for performing long-term projections and to gather scientific insights into and challenge existing knowledge of land surface processes on the global scale.

As an alternative to expert models, machine learning provides a more flexible and data-adaptive pathway to model Earth observation data: In the past decade, particularly deep learning approaches outperformed expert models in many domains in and outside of the Earth sciences. This success is rooted in the ability of deep neural networks to learn highly non-linear representations of structured data in an end-to-end setting, *i.e.*, with minimal expert interaction.

A major challenge in modeling vegetation and hydrology are the complex temporal interactions (*dynamic memory effects*) of the involved processes. While the representation of memory effects in physically-based models is still an issue, specialized neural network architectures, such as recurrent neural networks (RNNs), can learn them from data. Nevertheless, deep neural networks are still not widely used in global-scale land surface modeling. On the one hand, this is owed to their missing physical consistency, resulting in poor model performance when conducting out-of-distribution predictions. This is, for example, the case when performing long-term predictions into a warmer climate regime. On the other hand, the missing physical interpretability limits trust in these models and hampers scientific understanding.

There are, however, approaches to gather scientific insights using machine learning via *explanations*. Within the of field *explainable machine learning*, a range of approaches were developed to visualize and describe model properties and decisions. Such methods can be used by domain experts for the discovery of new knowledge or linkages, i.e., to improve scientific understanding or to challenge existing theories. Explainable machine learning is also gathering momentum in the Earth sciences, where the presence of complex, non-linear processes often justifies the usage of machine learning algorithms.

Explainable machine learning does not solve the problem of physical inconsistency and commonly only provides qualitative insights. Recently, neural networks were successfully combined with physically-based modeling in so-called *hybrid models*. From a machine learning perspective, adding prior knowledge increases the model robustness by constraining the solution space to physically plausible solutions. From the physical perspective, hybrid modeling allows learning uncertain or less known processes in physically-based models from data, which can decrease model

biases. Hybrid models also provide additional insights: The outputs of the neural network, which are then used as an input to the physical equations, can be directly interpreted as latent variables and coefficients. However, hybrid models are still experimental and their applicability for dynamic, large-scale modeling has not been explored yet.

Research questions (RQs)

The overarching goal of this thesis is to assess the potential of deep neural networks to represent dynamic memory effects in Earth observation data, and to identify pathways to account for and identify them. More specifically, the research questions are:

RQ1 Can recurrent neural networks learn global-scale ecosystem behavior?

RQ2 Can dynamic memory effects in Earth observations be identified using explanatory approaches?

RQ3 What is the promise of global-scale hybrid modeling and what are its challenges and opportunities?

RQ1 The first research question addresses the applicability of RNNs to large-scale ecosystem modeling. While RNNs have been used in regional studies and for modeling spatially sparse global data, a systematic assessment for global scale, spatially continuous data is missing. In the context of the overarching goal of this thesis, I focus on the representation of dynamic memory effects under heterogeneous conditions. This question is specifically addressed in Chapter 2, but is also further explored throughout Chapter 3. In Chapter 2, a long short-term memory (LSTM) model, a commonly used RNN architecture, is employed to emulate a global physically-based dynamic land surface model. The model represents dynamic memory effects of precipitation on evapotranspiration via a soil moisture state. By using model outputs from a physically-based model, issues with data quality and observability can be ruled out. By evaluating the model simulations in a controlled setting, the capability of an RNN to capture memory effects and to represent spatial patterns is explored and discussed.

RQ2 The second question targets the identifiability of memory effects using explanations. It is addressed in Chapter 3, where I present a novel permutation-based explanatory approach to quantify memory effects, *i.e.*, the impact of antecedent environmental conditions on the current system behavior. The model-agnostic approach allows to qualitatively assess memory effects by comparing different models that account for consecutively larger memory via sequential block-permutation. As a proof-of-concept, the method is used to quantify memory effects of climate variations on vegetation state using global Earth observation data.

RQ3 In Chapter 4, I assess the applicability of the hybrid approach to end-to-end large-scale environmental modeling. Hybrid modeling was only used in a few small-scale experiments so far and the feasibility for the representation of more complex and diverse modeling settings has not yet been assessed. As a proof-of-concept, I present a dynamic, global-scale hybrid model of the

hydrological cycle. To account for dynamic memory effects, an LSTM is employed to simulate latent hydrological variables and coefficients, which are then used as inputs to simple hydrological balance equations.

Key results

RQ1 An LSTM network was able to represent the global vegetation dynamics to a satisfying degree. While there were minor discrepancies between the simulations of the physically-based model and the LSTM predictions, the major temporal and spatial patterns aligned well. These results imply that LSTMs are well suited to learn hydrological and vegetation dynamics without using prior knowledge. This is a significant finding, as it allows to use RNNs as a an out-of-the-box solution to model global-scale land surface processes, as done in Chapter 3 and 4.

RQ2 The patterns of vegetation state memory effects discovered with the novel block-permutation aligned well with results from existing (linear and local) approaches and agreed with prior knowledge. Compared to existing approaches, the proposed method allows using data-adaptive models that can learn complex temporal interactions across scales, e.g., RNNs.

RQ3 The hybrid model yielded reasonable latent variables and coefficients, which were evaluated and compared to estimates from physically-based models. The model does not only yield interpretable hydrological quantities, but it also showed improved local adaptivity compared to physically-based approaches. The demonstration of the feasibility of hybrid modeling for large-scale dynamical modeling of Earth system processes is a novelty, opening avenues for a broad application in the Earth sciences. Further research is needed to better understand the strengths and weaknesses of the hybrid approach, which are discussed in detail within this thesis.

Conclusion

This work assesses the potential of RNNs for vegetation and hydrological modeling at the global scale and demonstrates avenues towards explainable and interpretable approaches, with special emphasis on the representation of dynamic memory effects. I show that RNNs are capable to learn heterogeneous land surface processes and demonstrate that explainable machine learning can be used to gather qualitative insights into ecosystem processes. I also show that more detailed insights are possible through hybrid modeling. The presented approaches are not intended to replace physically-based models, but rather, they provide alternative, data-driven insights into the Earth system. In this thesis, I identify a range of questions and challenges that need to be addressed, such as the problem of equifinality in hybrid modeling. This work is a first step towards data-driven yet interpretable environmental modeling and introduces methods that may find broad application in Earth system modeling.

Zusammenfassung

Hintergrund

Angesichts des Klimawandels sind ein besseres Verständnis und eine präzisere Darstellung der Prozesse an der Landoberfläche erforderlich, um beispielsweise die Prognosen für die Kohlenstoffbilanz terrestrischer Ökosysteme bei steigenden Temperaturen zu verbessern. Expertenmodelle bieten zwar wertvolle Einblicke in die Prozesse an der Landoberfläche, doch sind diese Modelle oft grobkörnig, unflexibel und durch möglicherweise falsches oder unvollständiges Vorwissen beeinflusst. Dennoch sind physikalisch basierte Modelle nach wie vor das zentrale Instrument für die Durchführung langfristiger Projektionen und die Gewinnung wissenschaftlicher Erkenntnisse über die Prozesse an der Landoberfläche im globalen Maßstab.

Als Alternative zu Expertenmodellen bietet das maschinelle Lernen einen flexibleren Weg zur Modellierung von Erdbeobachtungsdaten. In den letzten Jahren haben insbesondere Ansätze des Deep Learning Expertenmodelle in vielen Bereichen innerhalb und außerhalb der Geowissenschaften abgelöst. Dieser Erfolg beruht auf der Fähigkeit tiefer neuronaler Netze, hochgradig nichtlineare Repräsentationen strukturierter Daten in einem “end-to-end” Setting zu erlernen, also mit minimaler Experteninteraktion.

Eine große Herausforderung bei der Modellierung von Vegetation und Hydrologie sind die komplexen zeitlichen Wechselwirkungen (dynamische Memoryeffekte) der involvierten Prozesse. Während die Repäsentation solcher Memoryeffekte in physikalisch basierten Modellen immer noch ein Problem darstellt, können spezialisierte neuronale Netzarchitekturen, wie zum Beispiel rekurrente neuronale Netze (RNNs), diese aus Daten erlernen. Dennoch werden tiefe neuronale Netze bei der Modellierung von Landoberflächen im globalen Maßstab noch nicht breit eingesetzt. Das liegt zum einen an ihrer fehlenden physikalischen Konsistenz, die zu einer schlechten Modellleistung bei der Durchführung von Vorhersagen außerhalb der Verteilung der Trainingsdaten führt. Dies ist zum Beispiel der Fall, wenn langfristige Vorhersagen in ein wärmeres Klimaregime gemacht werden. Andererseits schränkt die fehlende physikalische Interpretierbarkeit das Vertrauen in diese Modelle ein und erschwert es, wissenschaftliche Erkenntnisse abzuleiten.

Es gibt jedoch Ansätze, wissenschaftliche Erkenntnisse mit Hilfe von maschinellem Lernen über *Explanations* zu gewinnen. Im Bereich des erklärbaren maschinellen Lernens wurde eine Reihe von Ansätzen zur Visualisierung und Beschreibung von Modelleigenschaften und -entscheidungen entwickelt. Solche Methoden können von Domänenexperten zum Entdecken neuer Fakten verwendet werden, also um das wissenschaftliche Verständnis zu verbessern oder bestehende Theorien zu überprüfen. Das erklärbare maschinelle Lernen gewinnt auch in den Geowissenschaften an Bedeutung, wo das Vorhandensein komplexer, nichtlinearer Prozesse häufig den Einsatz von Algorithmen des maschinellen Lernens rechtfertigt.

Das erklärbare maschinelle Lernen löst nicht das Problem der physikalischen Inkonsistenz

und liefert in der Regel nur qualitative Erkenntnisse. Kürzlich wurden neuronale Netze erfolgreich mit der physikalischen Modellierung in sogenannten Hybridmodellen kombiniert. Aus der Perspektive des maschinellen Lernens erhöht das Vorwissen die Robustheit der Modelle, indem es den Lösungsraum auf physikalisch plausible Lösungen einschränkt. Aus physikalischer Sicht ermöglicht die hybride Modellierung das Erlernen von unsicheren oder weniger bekannten Prozessen in physikalisch basierten Modellen aus Daten, was die Anpassungsfähigkeit an die beobachteten Daten erhöht. Hybride Modelle liefern auch zusätzliche Erkenntnisse: Die Ausgaben des neuronalen Netzes, die dann als Eingabe für die physikalischen Gleichungen verwendet werden, können direkt als latente Variablen und Koeffizienten interpretiert werden. Hybride Modelle befinden sich jedoch noch im Versuchsstadium, und ihre Anwendbarkeit für die dynamische, groß angelegte Modellierung wurde noch nicht untersucht.

Forschungsfragen (*Research Questions*, RQs)

Das übergeordnete Ziel dieser Arbeit ist es, das Potenzial tiefer neuronaler Netze zur Darstellung dynamischer Memoryeffekte in Erdbeobachtungsdaten zu bewerten und Wege zu finden, diese zu berücksichtigen und zu identifizieren. Die Forschungsfragen lauten im Speziellen:

RQ1 Können rekurrente neuronale Netze das Verhalten von Ökosystemen im globalen Maßstab lernen?

RQ2 Können dynamische Memoryeffekte in Erdbeobachtungen mit Hilfe von Methoden des erklärbaren maschinellen Lernens identifiziert werden?

RQ3 Was verspricht die hybride Modellierung auf globaler Ebene und was sind ihre Herausforderungen und Chancen?

RQ1 Die erste Forschungsfrage bezieht sich auf die Anwendbarkeit von RNNs für die Modellierung von Ökosystemen im großen Maßstab. Während RNNs in regionalen Studien verwendung fanden, fehlt eine systematische Bewertung für räumlich kontinuierliche Daten auf globaler Ebene. Im Rahmen des übergeordneten Ziels dieser Arbeit konzentriere ich mich auf die Darstellung von dynamischen Memoryeffekten unter heterogenen Bedingungen. Diese Frage wird speziell in Kapitel 2 behandelt, wird aber auch in Kapitel 3 weiter untersucht. In Kapitel 2 wird ein Long Short-Term Memory (LSTM) Modell, eine häufig verwendete RNN Architektur, zur Emulation eines globalen physikalisch basierten dynamischen Landoberflächenmodells eingesetzt. Das Modell stellt dynamische Memoryeffekte von Niederschlägen auf die Evapotranspiration über einen Bodenfeuchtespeicher dar. Durch die Verwendung von Modellsimulationen eines physikalisch basierten Modells können Probleme mit der Datenqualität und der Beobachtbarkeit ausgeschlossen werden. Durch die Evaluierung der Modellsimulationen in einer kontrollierten Umgebung wird die Fähigkeit eines RNN zur Erfassung von Memoryeffekten und zur Darstellung räumlicher Muster veranschaulicht und diskutiert.

RQ2 Die zweite Frage zielt auf die Identifizierbarkeit von Memoryeffekten durch *Explanations*. Sie wird in Kapitel 3 behandelt, in dem ich einen neuartigen permutationsbasierten Methode

vorstelle, mit dem sich die Auswirkungen vorangegangener Umweltbedingungen auf das aktuelle Systemverhalten (Memoryeffekte) quantifizieren lassen. Im Vergleich zu bestehenden Ansätzen erlaubt die vorgeschlagene Methode die Verwendung von datenadaptiven Modellen, die komplexe zeitliche Interaktionen über Skalen hinweg erlernen können, wie beispielsweise LSTMs. Als Machbarkeitsstudie wurde die Methode zur Quantifizierung der Memoryeffekte von Klimaschwankungen auf Vegetation unter Verwendung globaler Erdbeobachtungsdaten getestet.

RQ3 In Kapitel 4 bewerte ich die Anwendbarkeit des hybriden Ansatzes für die dynamische Umweltmodellierung in großem Maßstab. Die hybride Modellierung wurde bisher nur in einigen wenigen Experimenten in kleinem Maßstab eingesetzt, und die Durchführbarkeit für die Darstellung komplexerer und vielfältigerer Modellierungssituationen wurde noch nicht bewertet. Zur Evaluierung das Konzepts stelle ich ein hybrides Modell des Wasserkreislaufs auf globaler Ebene vor. Um dynamische Memoryeffekte zu berücksichtigen, wird ein LSTM verwendet, um latente hydrologische Variablen und Koeffizienten zu simulieren, die dann in einfachen hydrologischen Bilanzgleichungen verwendet wurden.

Schlüsselergebnisse

RQ1 Ein LSTM war in der Lage, die globale Vegetationsdynamik in zufriedenstellendem Maße darzustellen. Es gab zwar geringfügige Diskrepanzen zwischen den Simulationen des physikalisch basierten Modells und den Vorhersagen von LSTM, aber die wichtigsten zeitlichen und räumlichen Muster stimmten gut überein. Diese Ergebnisse deuten darauf hin, dass LSTMs geeignet sind, hydrologische und vegetationsbezogene Dynamiken ohne Verwendung von Vorwissen zu erlernen. Dies ist eine wichtige Erkenntnis die es rechtfertigt, RNNs zur Modellierung von Landoberflächenprozessen im globalen Maßstab zu verwenden, wie etwa in den Kapiteln 3 und 4 dieser Dissertation.

RQ2 Die mit der neuartigen Block-Permutation entdeckten Muster der Memoryeffekte des Vegetationszustandes stimmten gut mit den Ergebnissen anderer (linearer und lokaler) Ansätze sowie auch mit dem bestehenden Vorwissen überein. Im Vergleich zu bestehenden Ansätzen erlaubt die vorgeschlagene Methode die Verwendung von datenadaptiven Modellen, die komplexe zeitliche Interaktionen über Skalen hinweg erlernen können, wie zum Beispiel RNNs.

RQ3 Das hybride Modell lieferte plausible latente Variablen und Koeffizienten, die bewertet und mit Schätzungen aus physikalisch basierten Modellen verglichen wurden. Das Modell simulierte nicht nur interpretierbare hydrologische Größen, sondern wies auch eine verbesserte lokale Anpassungsfähigkeit im Vergleich zu physikalisch basierten Ansätzen auf. Der Nachweis der Durchführbarkeit hybrider Modelle für die großmaßstäbliche dynamische Modellierung von Erdsystemprozessen ist ein Novum und eröffnet Wege für eine breite Anwendung in den Geowissenschaften. Weitere Forschung ist notwendig, um die Stärken und Schwächen des hybriden Ansatzes besser zu verstehen, die in dieser Arbeit ausführlich diskutiert werden.

Fazit

Diese Arbeit bewertet das Potenzial von RNNs für die Modellierung von Vegetation und Hydrologie auf globaler Ebene und zeigt Wege zu erklärbaren und interpretierbaren Ansätzen auf, wobei der Schwerpunkt auf der Darstellung dynamischer Memoryeffekte liegt. Ich zeige, dass RNNs in der Lage sind, heterogene Landoberflächenprozesse zu erlernen, und demonstriere, dass erklärbares maschinelles Lernen verwendet werden kann, um qualitative Erkenntnisse über Ökosystemprozesse zu gewinnen. Des Weiteren zeige ich, dass hybride Modellierung aufgrund der hohen Interpretierbarkeit zum Prozessverständnis beitragen kann. Die vorgestellten Ansätze sollen physikalisch basierte Modelle nicht ersetzen, sondern vielmehr alternative, datengetriebene Einblicke in das System Erde ermöglichen. Darüber hinaus werden eine Reihe von Fragen und Herausforderungen identifiziert, die angegangen werden müssen, wie beispielsweise das Problem der Äquifinalität bei der hybriden Modellierung. Diese Arbeit ist ein erster Schritt in Richtung datengesteuerter und dennoch interpretierbarer Umweltmodellierung und stellt Methoden vor, die in Zukunft eine breite Anwendung in den Geowissenschaften finden können.

Contents

1. Introduction

1.1. Motivation

The study of our planet in the so-called Earth sciences has changed drastically in the past decade: With unprecedented computational power, growing amounts of Earth observation data, advances in methodology, and an increase in interdisciplinary and global collaborations, the field is growing and flourishing (Bonan and Doney, 2018). Despite the progress, many challenges persist, among which the detailed understanding and predictability of the terrestrial ecosystems are of the most urgent ones in the face of climate change (Kawamiya et al., 2020). In the physically-based large-scale modeling of the Earth system, the representation of vegetation and hydrology is still a major source of uncertainty (Jia et al., 2019). While the reasons for these uncertainties are manifold, key issues are the high degree of abstraction and limited flexibility of physically-based models (Reichstein et al., 2019). Nevertheless, such expert models provide valuable insights and long-term projections (Fisher and Koven, 2020).

Due to the growing amounts of Earth observation data and computational resources, machine learning provides an alternative to physically-based approaches (Camps-Valls et al., 2020). In the past decade, deep neural networks replaced expert models in many domains, *e.g.*, natural language processing (Young et al., 2018), or computer vision (Voulodimos et al., 2018). The promise of deep learning is simple: Instead of relying on complicated expert systems, highly adaptive algorithms learn to solve complex tasks from data. In deep neural networks, this is achieved through a set of hierarchical, non-linear feature extractors (LeCun et al., 2015). In contrast to traditional machine learning (*i.e.*, shallow learning), deep learning systems can solve complex tasks in an end-to-end setting, meaning that a model is learned solely from data and with minimal expert intervention.

Today, deep learning is widely used with Earth observation data. Applications range from land cover classification to parameter retrieval, to gap-filling of data products used in environmental models or for diagnosis of the Earth system (Camps-Valls et al., 2021; Ma et al., 2019; Yuan et al., 2020; Zhu et al., 2017). Especially the capacity to learn temporal dependencies in complex time-series with model architectures such as the recurrent neural network (RNN) could have major benefits for dynamic land surface modeling: The ability to learn ecosystem dynamics (so-called *ecological memory effects*) from data rather than relying on incomplete or wrong prior knowledge has arguably a large potential (Bergen et al., 2019; Camps-Valls et al., 2021; Karpatne et al., 2019;

Reichstein et al., 2019; Zhu et al., 2017), particularly since current physically-based models still struggle to represent them (Ogle et al., 2015; Reichstein et al., 2018).

Despite its potential, the application of machine learning and deep neural networks in particular for the global-scale modeling of land surface dynamics is rather uncommon. There are two main reasons for this: First, a major drawback of representing environmental processes with statistical models is their *physical inconsistency*. The problem of inconsistency explicitly expresses when performing out-of-distribution prediction, *i.e.*, when the statistical properties of the training and the test data diverge. This is, for example, the case when performing long-term projections in a changing climate regime. Second, deep neural networks are *not physically interpretable*. This does not only limit the trust in them, but also makes it difficult to use them to gather scientific understanding (Reichstein et al., 2019; Roscher et al., 2020).

Therefore, we have models like RNNs that are well suited to represent ecological memory effects, but the possibilities to use them to gain scientific insights are limited. The overarching goal of this thesis is to contribute to filling this gap.

1.2. Ecosystem modeling

Terrestrial ecosystems consist of living organisms and their physical environment on the land surface[1]. They are complex dynamical systems, which expresses in a highly non-linear interplay between system components across magnitudes of temporal and spatial scales (Bonan, 2015). Terrestrial ecosystems are an interface between the land surface and the atmosphere through their exchange of energy, momentum, gases, and aerosols. The terrestrial vegetation has been acknowledged to play a pivotal role in the mitigation (or local amplification) of climate change (Heimann and Reichstein, 2008). However, the complex processes and feedback loops of vegetation-atmosphere interactions are still not well understood on the global scale (Friedlingstein et al., 2014; Jia et al., 2019; Walker et al., 2021).

1.2.1. Ecosystem models

Ecosystem modeling deals with the numerical representation of ecosystems with the aim to better understand their functioning, but also to simulate their behavior under future—or more generally: different—conditions (Hall and Day, 1977). Small-scale ecosystem modeling involves targeted measurements and experiments, whereas large-scale models rely on Earth observation data, and the possibilities for performing experiments are very limited (Carpenter et al., 1995). Thus, the process of model development can be seen as a reverse-engineering of a system with incomplete, infrequent, and uncertain observations with very limited capabilities of interaction—altogether, a

[1] Within this thesis, I use the term *ecosystem* as a synonym for *terrestrial ecosystem*. Although ecosystems include all forms of life, I focus on vegetation.

difficult undertaking. Nowadays, a plethora of models and model types exist, sometimes focusing on a specific part (*e.g.*, hydrology or vegetation) or attempting to represent the entirety of the land surface (so-called *land surface models*, LSMs)[2].

It is evident that an exact representation of such complex systems is not feasible. Consequently, an ecosystem model is a heavily abstracted representation of the reality, and the level of detail is constrained by many factors such as the availability, resolution, and quality of observations, expert knowledge, and computational capacity (Fisher and Koven, 2020).

An accurate representation of land surface processes has a range of applications with high societal relevance. However, ecosystem models are still affected by large uncertainties and as a consequence, key questions regarding the consequences of climate change are yet to be answered. It is, for example, still not clear how much of the anthropogenic carbon emissions can be mitigated by terrestrial ecosystems in the future, or how vegetation reacts to a warming climate or an increase in atmospheric carbon concentration (Kawamiya et al., 2020).

1.2.2. Challenges in ecosystem modeling

As a consequence of these uncertainties, state-of-the-art ecological and hydrological models show large disagreements across scales (Bonan and Doney, 2018; Fisher and Koven, 2020; Haddeland et al., 2011; Schellekens et al., 2017). The persisting uncertainties can be attributed to two major factors: *data* and *system complexity* (Bonan, 2015; Reichstein et al., 2019). To better understand the role of these factors, we first take a closer look at a high-level mathematical representation of a dynamical system. Therefore, we use a discrete-time state-space notation: An dynamical system can be characterized by its state

$$\boldsymbol{s}_t = f(\boldsymbol{s}_{t-1}, \boldsymbol{x}_t) \tag{1.1}$$

at time t and an evolution function f that describes how the system state is altered by external factors (forcings) $\boldsymbol{x}_t$ in interaction with the previous state $\boldsymbol{s}_{t-1}$. The output function g further allows to define a system response (output)

$$\boldsymbol{y}_t = g(\boldsymbol{s}_t) \quad , \tag{1.2}$$

which depends on the state $\boldsymbol{s}_t$ (and sometimes also on the forcings $\boldsymbol{x}_t$, omitted here). In land surface modeling, the evolution function f and the output function g are usually non-linear due to the

[2] I use the term *ecosystem modeling* to loosely refer to the representation of the land surface with emphasis on vegetation, as the term *land surface modeling* denotes a component of an Earth system model (ESM) that resolves the coupled fluxes of carbon, energy, and water (Fisher and Koven, 2020). Due to the strong links between vegetation and the hydrological cycle (Humphrey et al., 2018; Jung et al., 2017), I do not specifically differentiate between the modeling of vegetation and hydrology.

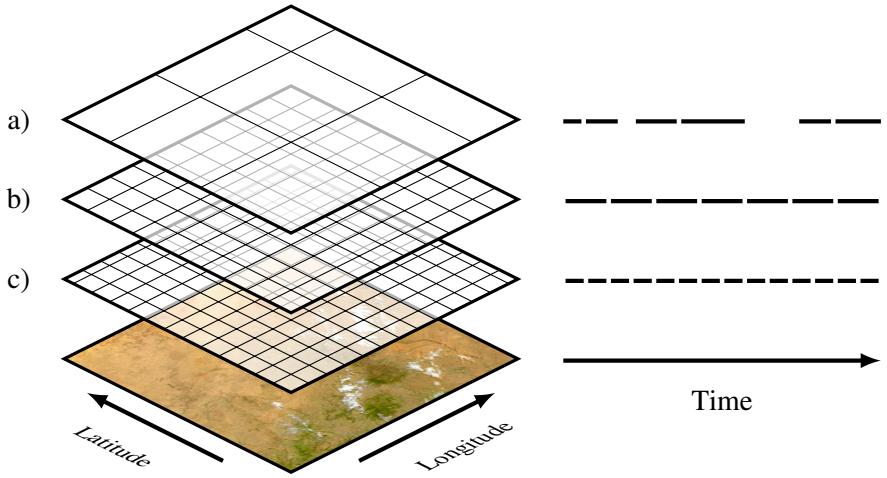


Figure 1.1.: Global Earth observation data products are often composed of different satellite overpasses, aggregated in space and time. The different spatial and temporal resolutions render the integration from different sources (here: a, b, and c, with different spatial (left) and temporal (right) resolutions) difficult and mask sub-resolution processes in unforeseen ways. RGB Image source: MODIS (Vermonte, 2015).

complexity of the involved processes (Camps-Valls et al., 2020). In the next two sections, we take a closer look at the relevance of *data* (Section 1.2.3) and *system complexity* (Section 1.2.4) in ecosystem modeling.

1.2.3. Earth observation data

The high-level mathematical representation in Equation 1.1 and 1.2 underlines the relevance of *data*. To force a dynamic model, we require external dynamic factors $\boldsymbol{x}$ (the *forcings*), which are seen as independent from the system. We also need observations of the model outputs $\boldsymbol{y}$, either for model tuning, validation, or both. It can also be an advantage to have measurements of system states, either to as a further means for validation, or to directly use the observations instead of modeling them.

Earth observation data is evolving quickly in terms of quantity and quality (Guo et al., 2015). With decades-long measurements available on different spatial and temporal scales, the community can draw from a wide range of observational products. Still, data is a limiting factor in many aspects. Even though we know the most important forcings that drive ecosystem behavior, we will never have a complete, accurate, and precise representation of them. Rather, we rely on spatially and temporally aggregated or sparse observations from space or sophisticated data assimilation products, both affected by biases and uncertainties. The aggregated Earth observation datasets may mask small-scale patterns and are difficult to use in practice due to their heterogeneous

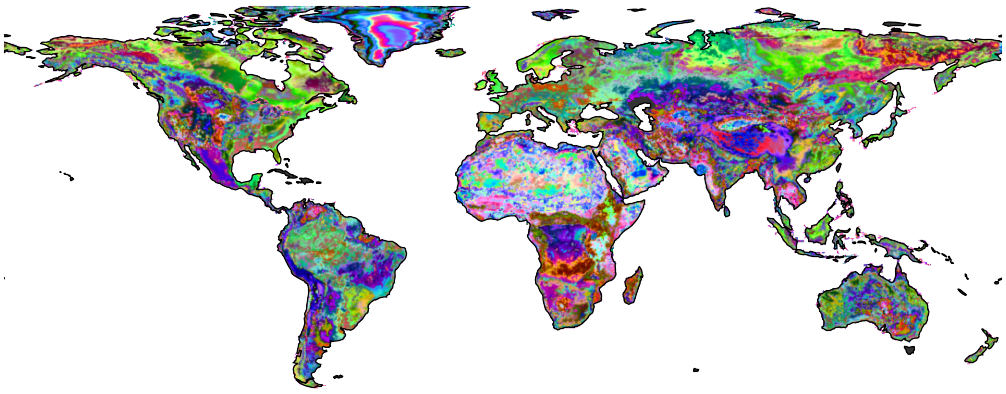


Figure 1.2.: The heterogeneity of the land-surface: RGB composite of the dimensionality-reduced datasets of soil properties, land-cover, and elevation highlights local and regional land-surface gradients. The reduction is based on the t-SNE algorithm (Hinton and Roweis, 2002).

resolution and coverage (Figure 1.1). These shortcomings affect both the model simulations and the conclusions drawn from the model parameters and simulations.

Another challenge is the measurability of certain processes or states on a global scale. The observation of soil moisture with satellite or airborne remote sensing, for example, is limited to the top soil layers (Mohanty et al., 2017) and the direct measurement of the carbon budget of ecosystems is restricted to site-level (Baldocchi et al., 2001; Jung et al., 2020). This restricted observability is one of the main constraints in ecosystem modeling (Bonan and Doney, 2018), and even with novel satellite missions with a higher spatial and temporal resolution, it seems that we can only—quite literally—scratch on the surface.

1.2.4. Modeling complex ecosystems

Next to data limitations, the vast *complexity* of ecosystems poses a major challenge in their numerical representation. An ecosystem is a dynamical and adaptive system, and its behavior (f in Equation 1.1) often seems complex and stochastic (Hantson et al., 2016; Reichstein et al., 2014). It has been acknowledged long ago that ecosystems are not the sum of loosely connected sub-components. Rather, ecosystems are shaped by the continuous interplay of chemical, physical, biological, and anthropogenic processes (Bianchi, 2020), which results in the heterogeneous land surface as we know it (Figure 1.2). Ecosystem processes span several magnitudes of temporal and spatial scales, from submillimeter and –second range (*e.g.*, photosynthesis) to centurial and global (*e.g.*, adaption of vegetation to warming climate) extent. These complex processes lead to highly non-linear interactions across time scales and result in lagged ecosystem responses to environmental changes, so-called *memory effects*.

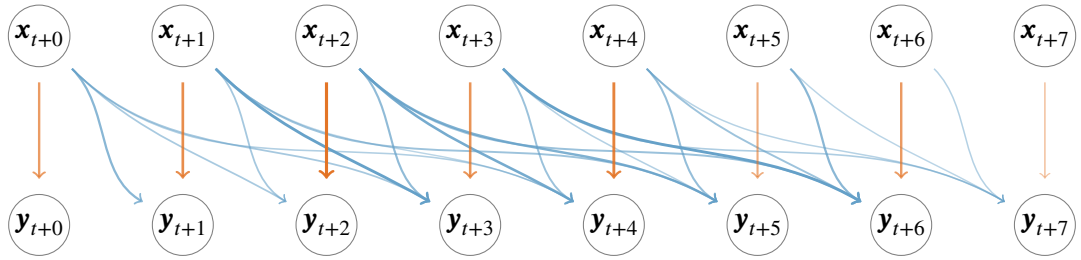


Figure 1.3.: In a dynamic system, the forcings (external inputs) $\boldsymbol{x}_t$ at time t can have a quasi-instantaneous impact on the output (system response) $\boldsymbol{y}_t$. The lagged impact of antecedent forcings $\boldsymbol{x}_t$ on $\boldsymbol{y}_{t+k}$ with time lag k are called *dynamic memory effects*, or *lag effects*.

In a dynamic system, the external factors (forcings) $\boldsymbol{x}_t$ can have a quasi-instantaneous impact on the response $\boldsymbol{y}_t$. In addition, $\boldsymbol{x}_t$ can have a delayed impact on $\boldsymbol{y}_{t+k}$ with time lag k (Figure 1.3). These lagged impacts are called *(dynamic) memory effects*. From Equation 1.1, it becomes evident that these memory effects can only be propagated through the system state $\boldsymbol{s}$. This is why we sometimes call this state the system's *memory*, as is encodes past events.

Memory effects are omnipresent in nature (Ogle et al., 2015): A long period of below-average precipitation, for example, may cause lower soil misture or even groundwater depletion, causing water stress in vegetation. This may expose vegetation to the risk of forest fires, diseases, or make the more vulnerable to insects outbreaks. It can take months, years or even decades for vegetation to recover from a drought and its consequences (Besnard et al., 2019). The representation of such processes is extremely challenging as it requires obserations of the forcings $\boldsymbol{x}$, a proper description of the processes f shaping the system state or observations thereof, and observations of the system responses $\boldsymbol{y}$ with the corresponding function g.

Ecosystem and hydrological models consist of a mixture of physical, semi-empirical, and empirical components (Fisher and Koven, 2020). The models are defined by their *structure* (the causal connections) and the process *parameterization* (the representation of a process). Uncertainties in the model structure may introduce errors that emerge from a mismatch between the causal structure of the model and the real-world system (Butts et al., 2004). This mismatch can introduce biases in parameter estimates and simulations due to compensation effects (Engelhardt et al., 2014; Gupta et al., 2012; Refsgaard et al., 2006). Similarly, the process parameterization introduces errors, owing to wrong or incomplete knowledge, data limitations, and aggregation effects. For example, the physics of snow crystals formation is well known on microscopic scales (Libbrecht, 2005), but we have neither the computational power nor the required small-scaled observations to feed the physical equations. Thus, the physically complex process of snow formation is parameterized such that it can be computed at the model scale. A common approach in global hydrological models is to simply assume all precipitation below a temperature threshold is snowfall (*e.g.*, Van Der Knijff

et al., 2010). Such heuristics can be found in all ecological and hydrological models, and while they work decently in many cases, they arguably do not do justice to the complexity of the real world.

The parameterization of a process implies that certain parameters need to be set (*e.g.*, the snow temperature threshold from above). Parameters can be retrieved from direct observations (*e.g.*, observed temperature threshold in experiments) or from prior knowledge (*e.g.*, freezing point of 0 °C). Furthermore, parameters can be found computationally by inverse modeling. In inverse modeling, model parameters are adjusted such that a model generates outputs close to observed values, given some input values (*e.g.*, Sood and Smakhtin, 2015). In physically-based modeling, parameterizations are often rigid and coarse-grained, which introduces a mismatch between simulations and observations.

Physically-based models remain the primary tool for advancing the understanding of processes on large scales and for performing long-term projections and scenarios (Bonan and Doney, 2018). However, advances in knowledge, data, and computational power did not directly translate into the reduction of model uncertainty as expected. It seems that increasing model complexity and resolution alone leads to a dead end, and new paradigms are needed (Reichstein et al., 2019).

1.3. Learning from data

1.3.1. Deep learning and neural networks

Deep learning has been the backbone of modern computer vision (Voulodimos et al., 2018) and natural language processing (Young et al., 2018) and enabled breakthroughs in several other research and engineering disciplines. The great success was fueled by a tremendous growth in computational power and data availability. Nowadays, data-driven algorithms outperform expert systems in various tasks (Goodfellow et al., 2016). This success is owed to the flexibility of deep learning algorithms to learn complex representation from structured data without relying on domain knowledge, but also on the availability of data (LeCun et al., 2015).

In the most common machine learning setting—called *supervised learning*—, we are given tuples of $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ with input features $\boldsymbol{x}_i \in X$ and the labels (or targets) $\boldsymbol{y}_i \in Y$, building the training set $\mathcal{D} = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), \dots, (\boldsymbol{x}_n, \boldsymbol{y}_n)\}$ of n samples. The goal in supervised learning is to learn a mapping from the input features to the targets by the function $f : X \to Y$ by minimizing a loss function $\ell : Y \times Y \to \mathbb{R}_+$. The function f is found by minimizing

$$f^* = \arg\min_{f \in \mathcal{H}} \sum_{i=1}^{N} \ell(f(\boldsymbol{x}_i), \boldsymbol{y}_i) \quad , \tag{1.3}$$

where $\mathcal{H}$ is the hypothesis space.

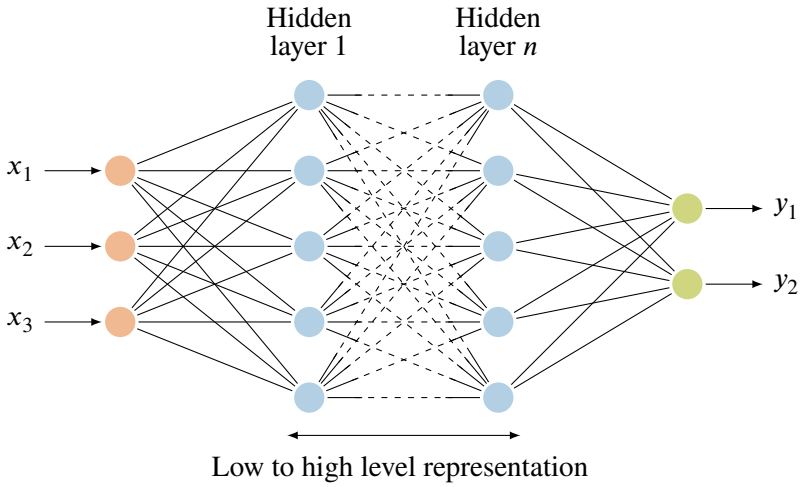


Figure 1.4.: A feed-forward neural network with n hidden layers. The sequential processing of the input features $\boldsymbol{x}$ leads to a high-level hidden representation at the layer n from which a linear mapping yields the predictions $\boldsymbol{y}$.

So-called shallow machine learning algorithms still require the user to provide hand-crafted input features, whereas deep learning algorithms are capable of learning high-level representations of the input data end-to-end. The representation is a learned mapping of the input achieved through a set of simpler, sequentially arranged non-linear transformations (Goodfellow et al., 2016), exemplified here with a fully-connected neural network in Figure 1.4. Neural networks are highly data-adaptive, and the performance scales well with increasing data size, which is not necessarily the case for other statistical learning algorithms (LeCun et al., 2015; Lipton et al., 2015).

A fully-connected neural network, as illustrated in Figure 1.4, is a universal function approximator (Csanád Csáji, 2001; Hornik et al., 1989) that requires minimal assumptions and minimal prior knowledge about the modeling problem. In other words, it has a minimal *inductive bias* (Baxter, 2000; Mitchell, 1980). It is often beneficial to encode some prior assumptions into the model architecture to improve generalizability and facilitate model training. A convolutional neural network (CNN), for example, introduces an inductive bias through the assumption of a spatially localized structure (Elsayed et al., 2020). A further example of an inductive bias is the selection of input features. Inductive biases restrict the hypothesis space $\mathcal{H}$ and ideally improve the data efficiency and generalizability of a model. The choice of assumptions often depends on domain knowledge, which seems sometimes trivial (*e.g.*, using a CNN for image classification), but usually requires expert knowledge and a careful model development process.

1.3.2. Recurrent neural networks

A number of deep learning-based approaches exist to deal with sequential data, such as Earth observation time series. A widely used concept is the RNN, which represents a system's state and learns its evolution in time from data. It is defined by a state transition function

$$\boldsymbol{h}_t = \sigma(\boldsymbol{W}_{\mathrm{x}}\boldsymbol{x}_t + \boldsymbol{W}_{\mathrm{h}}\boldsymbol{h}_{t-1} + \boldsymbol{b}_{\mathrm{h}}) \tag{1.4}$$

and an output function

$$\hat{\boldsymbol{y}}_t = \boldsymbol{W}_{\mathrm{y}}\boldsymbol{h}_t + \boldsymbol{b}_{\mathrm{y}} \quad . \tag{1.5}$$

The input $\boldsymbol{x}_t \in \mathbb{R}^M$ is the t^{th} element of a sequence $\boldsymbol{x} \in \mathbb{R}^{T\times M}$ of T elements with M input features. Equivalently, $\boldsymbol{y}_t \in \mathbb{R}^D$ is the t^{th} element of the labels $y \in \mathbb{R}^{T\times D}$ with label dimensionality D. The hidden state $\boldsymbol{h}_t \in \mathbb{R}^R$ represents the system state at time t and has dimensionality R, a hyperparameter corresponding to the number of recurrent neurons. The learned input weights $\boldsymbol{W}_{\mathrm{x}} \in \mathbb{R}^{R\times M}$ and recurrent weights $\boldsymbol{W}_{\mathrm{h}} \in \mathbb{R}^{R\times R}$ are multipied with the input and the previous hidden state, respectively, and a learned bias $\boldsymbol{b}_{\mathrm{h}} \in \mathbb{R}^R$ is added. The sigmoid activation function σ is used to introduce non-linearity. The hidden state $\boldsymbol{h}_t$ is mapped to one or multiple outputs $\hat{\boldsymbol{y}}_t$, which are compared to the observations $\boldsymbol{y}$ for supervision. This operation involves another set of learnable parameters, namely the output weights $\boldsymbol{W}_{\mathrm{y}} \in \mathbb{R}^{D\times R}$ and the output bias $\boldsymbol{b}_{\mathrm{y}} \in \mathbb{R}^D$. Overall, the model performs a mapping of the input space $X \in \mathbb{R}^{T\times M}$ to the label space $Y \in \mathbb{R}^{T\times D}$. The model parameters $\boldsymbol{\theta} = \{\boldsymbol{W}_{\mathrm{x}}, \boldsymbol{W}_{\mathrm{h}}, \boldsymbol{b}_{\mathrm{h}}, \boldsymbol{W}_{\mathrm{y}}, \boldsymbol{b}_{\mathrm{y}}\}$ are optimized using backpropagation through time (Goodfellow et al., 2016) with the aim to minimize the cost $\boldsymbol{J}$ in respect to a training dataset of n total $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ tuples and a loss function ℓ:

$$\underset{\boldsymbol{\theta}}{\mathit{minimize}}\ \boldsymbol{J}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_i) \quad . \tag{1.6}$$

Note that this concept is closely related to the discrete state-space notation introduced in Section 1.2.2: The system state is updated using an evolution function (Equation 1.4) and the state is mapped to the labels using an output function (Equation 1.5). Consequently, the hidden state h_t of the RNN can be understood as a complex system state containing relevant information to compute the future system behavior in interaction with the concurrent external factors $\boldsymbol{x}_t$. This loose and flexible concept provides a powerful approach to represent dynamic systems if process knowledge or observability of the system state is limited, as assumptions are minimal.

Basic RNNs as presented here are not often used in practice as they are not capable of learning long-term sequential dependencies (Lipton et al., 2015). Rather, more sophisticated architectures are used, such as the long short-term memory (LSTM) model (Hochreiter and Schmidhuber, 1997),

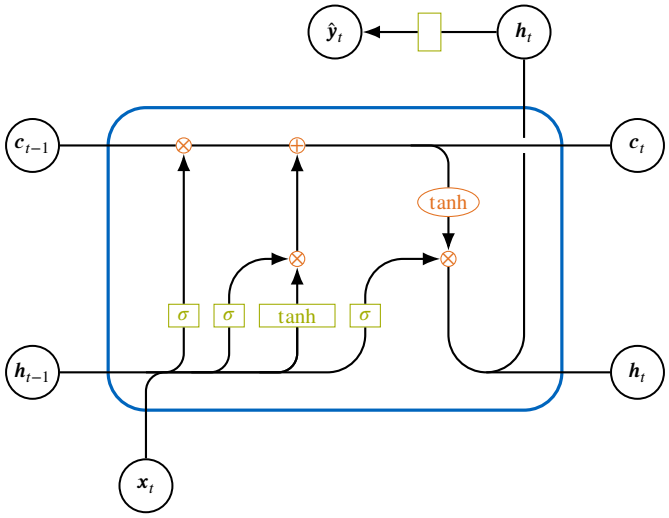


Figure 1.5.: The long short-term memory (LSTM) model with data input $\boldsymbol{x}_t$ at time t, the hidden state $\boldsymbol{h}_t$, cell state $\boldsymbol{c}_t$, and the model output $\hat{\boldsymbol{y}}_t$. The model contains four linear layers with non-linear activations (sigmoid σ and hyperbolic tangent tanh) and pointwise operators. The hidden state $\boldsymbol{h}_t$ is mapped to a prediction $\hat{\boldsymbol{y}}_t$ using another linear layer. The cell state $\boldsymbol{c}_t$ caries long-term information.

illustrated in Figure 1.5. I use the term RNN herein to refer to the general concept of the recurrent neural network.

When modeling heterogeneous ecohydrological processes, it is crucial to account for memory effects, *i.e.*, delayed system responses to antecedent conditions (Ogle et al., 2015). RNNs have been shown to be able to represent such temporal dependencies in regional studies (*e.g.*, rainfall-runoff modeling in Northern America, Kratzert et al., 2018). To properly represent processes of the land surface, a wide range of factors representing soil, vegetation, or other landscape features are relevant (Beck et al., 2016; Jung et al., 2020). Such factors may also change over time—for example, soils are in constant progression—, but assuming constant values for such slow processes is usually sufficient. A key advantage of a purely data-driven approach is the flexibility to use such variables without prescribing their interactions explicitly.

1.3.3. Deep learning for ecosystem modeling

Nowadays, deep learning is applied broadly in the Earth sciences. Two major reasons for its success are the growing availability of data and the ability to exploit spatio-temporal structures in data. Applications range from object detection, image recognition, and semantic representation, to anomaly/change detection and time-series regression (Zhu et al., 2017), for example for landscape prediction from climate (Requena-Mesa et al., 2018), rainfall-runoff modeling (Kratzert et al., 2018), crop field identification (Rußwurm and Körner, 2017), or modeling of global-scale ecosystem

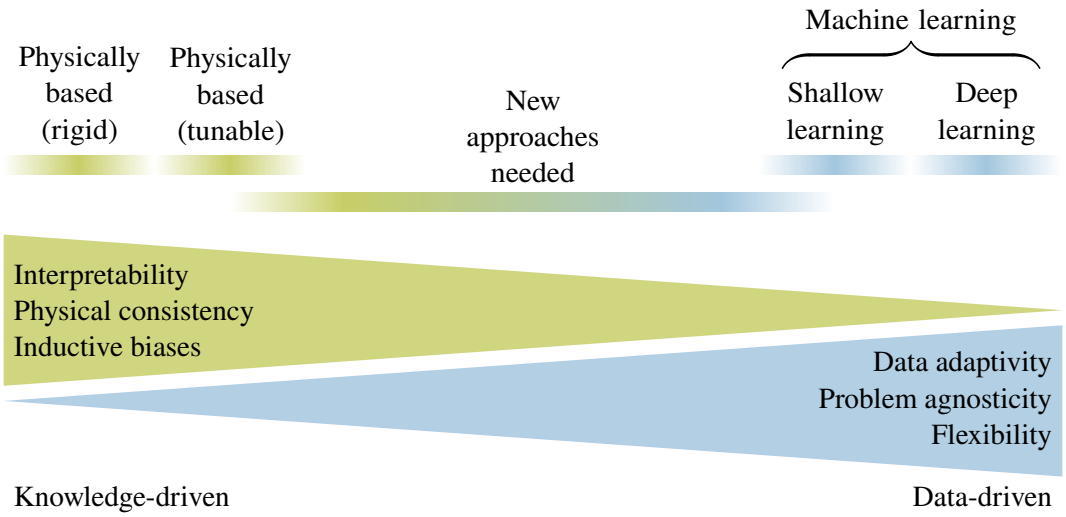


Figure 1.6.: Physically-based versus machine learning models: While physically-based models are primarily based on prior knowledge and offer interpretability, machine learning models are capable of learning from data in flexible ways, but in turn suffer from limited interpretability and physical consistency. New approaches are needed to close the gap between knowledge and data-driven methods.

responses (Kraft et al., 2021). For an extensive review of applications of deep learning in the Earth sciences, I refer to Camps-Valls et al. (2021).

So far, deep learning has only been marginally considered for scientific applications with the aim to improve system understanding or for making long-term predictions of complex systems (*e.g.*, meteorological forecasts or climate projection). This is owed to the low interpretability of deep neural networks. As a consequence, the capabilities for gaining scientific knowledge from, and including prior knowledge into them, are limited. This leaves a gap between physically-based models, which are both interpretable and physically consistent but prone to biases, and machine learning approaches, which can make use of large amounts of data more efficiently (Figure 1.6). Hence, the domain of Earth system modeling was largely untouched by the current developments in deep learning until recently (Reichstein et al., 2019). However, novel approaches provide an opportunity to gather scientific insights using explanatory approaches (Section 1.3.4). Furthermore, the fusion of the more rigid expert models and the flexible data-driven approaches in so-called *hybrid models* (Section 1.4) opens avenues for the combination of data- and knowledge-driven methods, and may motivate two communities to share ideas and knowledge.

1.3.4. Explainable machine learning

Machine learning models are broadly applied nowadays, but their low *interpretability* imposes challenges and limitations. Closely related, the limited *explainability* renders model debugging and

development difficult, limits our trust in them (Lipton, 2018), and hampers the usage of machine learning models for improving scientific understanding (Roscher et al., 2020). Therefore, they are often not considered for scientific applications with a focus on system understanding (Camps-Valls et al., 2020; Karpatne et al., 2018; Reichstein et al., 2019). Various approaches exist to tackle these shortcomings via *explanations*, subsumed under the term of explainable machine learning, or sometimes explainable artificial intelligence (XAI). The field is growing quickly (Arrieta et al., 2020; Biran and Cotton, 2017; Miller, 2019), but a concise terminology is still missing and basic terms, such as *transparency*, *interpretations*, and *explanations* are not well defined. For example, some authors use the terms *interpretability* and *explainability* interchangeably (*e.g.*, Miller, 2019), others use the terms distinctively (*e.g.*, Arrieta et al., 2020; Roscher et al., 2020), or state that an all-purpose definition does not exist (Rudin, 2019).

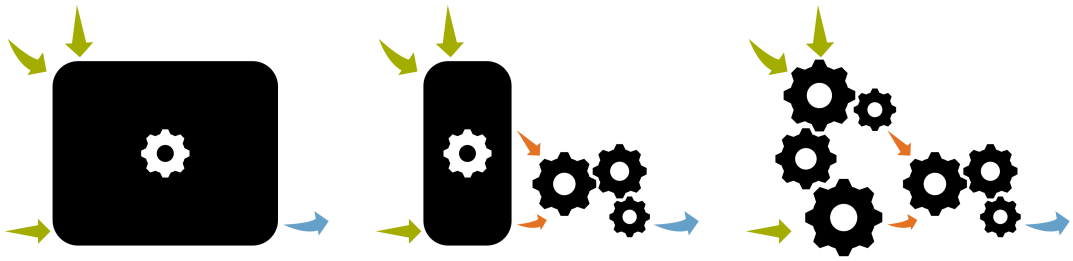
According to Roscher et al. (2020), a machine learning algorithm is considered *transparent* if the model functioning and training can be described and design choices can be motivated. In this sense, neural networks are not in-transparent *per se*, as we can write them down as a function, we can motivate our design choices, and we understand how the model optimization algorithms work.

Interpretations aim at presenting properties of a model in understandable terms and require a model and data (Roscher et al., 2020). To enhance the interpretability of neural networks, a set of tools and concepts have been developed (Arrieta et al., 2020; Lipton, 2018). A well-known example is the use of saliency maps, where input features (*e.g.*, pixels in image classification tasks) are highlighted to show which parts are most relevant for the model decision. Similarly, attention-based models allow to visualize what the model is focusing on (*e.g.*, Vig, 2019). Other approaches, such as local interpretable model-agnostic explanations (LIME, Ribeiro et al., 2016), use surrogate models to provide additional insights.

Explanations, which rely on interpretations, are "a set of statements usually constructed to describe a set of facts which clarifies the *cause*, *context*, and *consequences* of those facts" (Drake, 2018). Their purpose is to answer one of the "what?", "how?", and "why?" questions, *e.g.*, "Why did that event happen?" (Miller, 2019). Adadi and Berrada (2018) identify four use cases of explanations: *justification* (of a decision), *control* (to discover model vulnerabilities and flaws), *improvement* (of a model), and *discovery* (of new facts). From a scientific perspective, the discovery of new facts is arguably the most interesting use case. This can be achieved through a profound understanding of the machine learning approach (transparency), interpretations of the model behavior, and expert knowledge (Roscher et al., 2020).

1.4. Hybrid modeling

Hybrid modeling combines concepts from machine learning and physically-based modeling (Figure 1.7). The fusion of the two paradigms allows to build data-driven yet partially interpretable



(a) Machine learning (b) Hybrid modeling (c) Physically-based modeling

Figure 1.7.: Hybrid modeling (b) combines concepts from machine learning (a) and physically-based modeling (c). In end-to-end machine learning (a), a model learns a mapping from the input features to the output (observed labels) from data. In physically-based modeling (c), the processes are pre-described physically or conceptually in interpretable terms, or empirically from prior observations. Usually, physically-based models represent (non-observed) latent variables or coefficients, which emerge from the model and may be used as data products or to improve system understanding. In hybrid models (b), such latent variables or coefficients can be estimated by a machine learning algorithm in a flexible and data-adaptive way.

and physically consistent models (Camps-Valls et al., 2021; de Bézenac et al., 2019; Reichstein et al., 2019). After Reichstein et al. (2019), hybrid modeling denotes the replacement of a physical submodel with machine learning. More generally, in hybrid modeling, parts of a mechanistical or conceptual model are parameterized or replaced by machine learning.

Consider a process $y = q(x_1, x_2)$. In its simplest form, we can denote a hybrid model of q as an inner process

$$p = g_{\text{ML}}(x_1) \quad , \tag{1.7}$$

represented by a machine learning model g_{ML} that simulates a latent variable p, and an outer process

$$y = f_{\text{phys}}(p, x_2) \quad , \tag{1.8}$$

where f_{phys} is a physically-based function with a known structure. The variables x_1 and x_2 are the forcings and/or static variables, here they are univariate for simplicity.

Offline and online hybrid modeling

The parameterization may be performed *offline*, *i.e.*, the machine learning model g_{ML} is trained in advance and fixed at model run-time, or *online*, meaning that g_{ML} is trained end-to-end as a component of the entire model. In the offline mode, we deal with two separate modeling problems:

In the first step, a machine-learning model is trained to approximate the inner process (Equation 1.7). This requires observations or simulations of p, but they are often not available. Offline hybrid modeling is commonly used to emulate certain processes for computational efficiency (*e.g.*, Rasp et al., 2018). Hereinafter, I focus on online hybrid modeling, and *hybrid modeling* refers to the online case.

Parameters, variables, and coefficients

A precise differentiation between the parameters $\boldsymbol{\theta}_{\text{ML}}$, $\boldsymbol{\theta}_f$, and the intermediate outputs $\boldsymbol{p}$ (strictly speaking also parameters to f_{phys} that depend on data) and their conceptual meaning is important: Variables are quantities that can be observed independently from the experiment, which are to be brought into a relationship by a model. Parameters are constant quantities that define the model behavior, which "stand for inherent properties of nature" (Bard, 1974). Within the parametric hybrid modeling context, I use the term *ML parameter* to refer to the non-interpretable parameters of the machine learning model ($\boldsymbol{\theta}_{\text{ML}}$), whereas the term *physical parameters* is used to refer to constant, global parameters ($\boldsymbol{\theta}_f$) of the physically-based module. These parameters are found through model optimization and do not depend on data, *i.e.*, they are fixed after model training. Parameters that depend on data (*i.e.*, they are an output of g_{ML}) are referred to as *coefficients*. These learned coefficients can vary in space, time, or along any other data dimension. Outputs of g_{ML} that are physical quantities (*e.g.*, a flux such as evaporation or a state such as groundwater) are referred to as *latent variables* if they are not used for supervision. Note that this terminology is used in S4 Kraft et al. (2022), but not in S3 Kraft et al. (2020).

Parametric and non-parametric approaches

Hybrid modeling denotes a broad concept, which renders a categorization challenging. I differentiate between parametric and non-parametric hybrid modeling, where *(non-)parametric* refers to the physical submodel (Figure 1.8). If the model is non-parametric, only the parameters $\boldsymbol{\theta}_{\text{ML}}$ of the machine learning model g_{ML} need to be optimized. In this case, the optimization problem can be described as a minimization problem over a set of $i \in \{1, \dots, n\}$ training samples, denoted as

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}=(\theta_{\text{ML}})}{argmin} \sum_{i=1}^{n} \ell(f_{\text{phys}}(x_{1,i}, g_{\text{ML}}(x_{2,i}, \theta_{\text{ML}})), y_i) \quad , \tag{1.9}$$

here exemplified using the simple hybrid model defined in Equations 1.7-1.8. In the parametric case, the optimization is extended to parameters $\boldsymbol{\theta}_f$ of f_{phys}:

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}=(\theta_{\text{ML}},\theta_f)}{argmin} \sum_{i=1}^{N} \ell(f_{\text{phys}}(x_{1,i}, g_{\text{ML}}(x_{2,i}, \theta_{\text{ML}}), \theta_f), y_i) \quad . \tag{1.10}$$

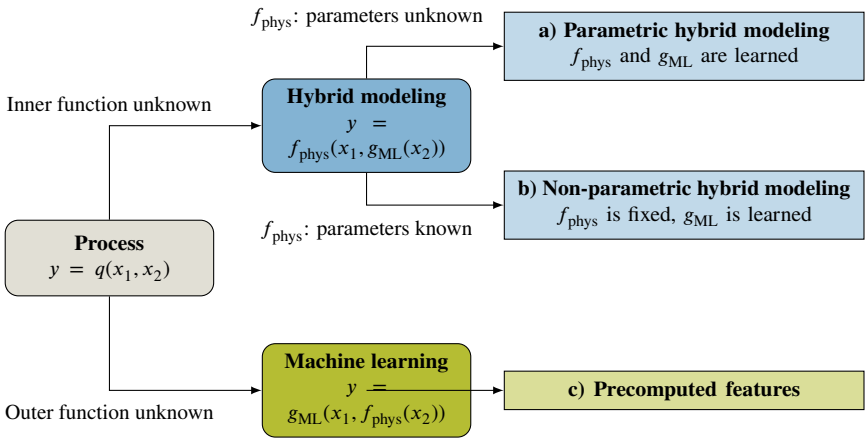


Figure 1.8.: Online hybrid modeling is the combination of machine learning and physically-based modeling in an end-to-end training setting. A hybrid model of the process q depending on x_1 and x_2 uses a machine learning model g_{ML} to represent subprocesses of q, providing latent variables as input to, or coefficients for a physically-based submodel f_{phys} (top branch). If f_{phys} has tunable parameters itself, we call the approach parametric (**a**), else non-parametric (**b**). If the output of a physically-based model f_{phys} is used as input to a machine learning model g_{ML} (**c**), we deal with classical machine learning with precomputed features (bottom branch).

Assuming that g_{ML} is a neural network, as is commonly the case, the optimization can be done by using standard gradient descent with backpropagation (Goodfellow et al., 2016), given that f_{phys} is differentiable. A parametric hybrid model can still be optimized by using backpropagation: The optimizer can concurrently update the parameters $\boldsymbol{\theta}_{\text{ML}}$ and $\boldsymbol{\theta}_f$.

Non-parametric hybrid modeling denotes a special case, where f_{phys} is purely physical (based on first principles) or parameters are known/fixed during model training. In this case, only the ML parameters are optimized. A prominent example of a non-parametric hybrid model—the first application of hybrid modeling in Earth sciences to my knowledge—combines a convolutional deep neural network (CDNN) with fluid dynamics to predict sea surface temperature (SST, de Bézenac et al., 2019). Based on k past fields of SST ($\{I_{t-k-1}, \cdots, I_t\}$), a deep convolutional neural networks simulates a motion field (equivalent to Equation 1.7), which is fed into physical equations of advection and diffusion (equivalent to Equation 1.8), yielding the next SST field, I_{t+1} (Figure 1.9). Instead of directly predicting the SST using the CDNN, physical process knowledge was used to restrict the hypothesis space. The authors could show that the model outperforms purely deep learning–based as well as a purely physically-based models.

The approach yields the coefficient $\hat{\boldsymbol{\omega}}_t$, the estimate of the motion field per time step t. This offers several opportunities: It allows to further regularize the model by adding soft constraints on

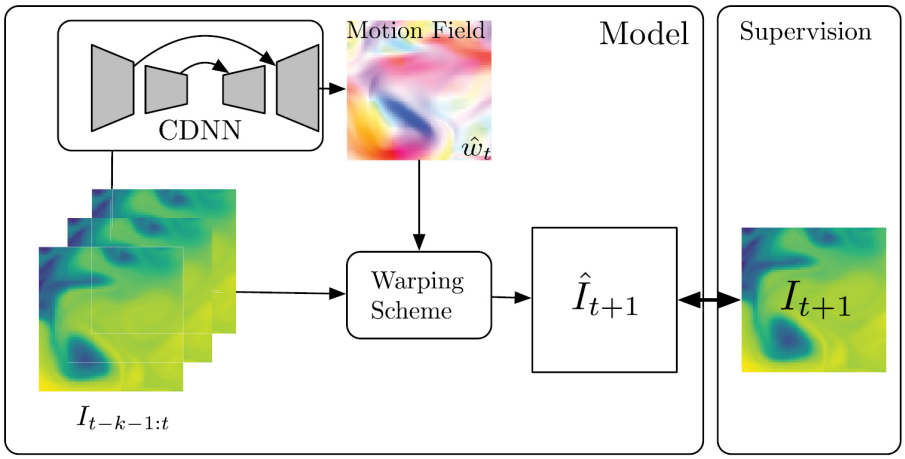


Figure 1.9.: Sea surface temperature (SST) prediction using a non-parametric hybrid model (de Bézenac et al., 2019): fields of SST are fed into a convolutional deep neural network (CDNN) to simulate a latent motion field further used in physical equations of diffusion and advection (warping scheme) to predict the next SST field.

$\hat{\boldsymbol{\omega}}_t$ to the loss function, *e.g.*, by assuming a certain spatial smoothness of $\hat{\boldsymbol{\omega}}_t$, or by penalizing large changes between $\hat{\boldsymbol{\omega}}_{t-1}$, and $\hat{\boldsymbol{\omega}}_t$ (*i.e.*, temporal smoothness). In addition, the latent variables can be used for quality control. When the model is applied to unseen data, the motion field may be used as a diagnostics tool for model reliability, *e.g.*, through expert validation. Furthermore, the motion field could be used as data product for other purposes, for example, as input to a physically-based model. Last but not least, the $\hat{\boldsymbol{\omega}}_t$ can help experts to better understand processes related to sea surface temperature or motion.

In the above example, no physical parameters needed to be tuned. However, in ecohydrological modeling, the processes can almost never be represented with pure physics. An example was provided in the introduction (Section 1.2): Snow formation is well understood physically, but a representation with first principles would require enormous computational resources and high-resolution data as input to the equations. Similar problems exist in other domains, such as oceanography or atmosphere modeling (*e.g.*, O'Gorman and Dwyer, 2018; Rasp et al., 2018). An example of parametric hybrid modeling is presented in Chapter 4.

Accounting for memory effects in hybrid modeling

For the representation of land surface processes, non-dynamical models may not suffice due to memory effects. As not all relevant system states can be observed and used as an input, a non-dynamical model cannot represent memory effects adequately. We can extend the non-dynamical

hybrid model (Equations 1.7 and 1.8) to introduce a system state, *i.e.*, to account for memory effects. To represent implicit, non-physical memory effects (which we may not be able to explicitly represent in a physical manner or of which we are not aware), we can use a recurrent neural network

$$\boldsymbol{h}_t = g_{\text{RNN}}(\boldsymbol{h}_{t-1}, [\boldsymbol{x}_t, \boldsymbol{s}_{t-1}]) \quad , \tag{1.11}$$

which takes the concatenated forcings $\boldsymbol{x}_t$ and the previous time step's *physical* states $\boldsymbol{s}_{t-1}$, together with its own hidden previous state $\boldsymbol{h}_{t-1}$ (see Section 1.3.2 for more details on RNNs) as inputs. At each time step, the RNN yields an updated hidden state $\boldsymbol{h}_t$. Note that this state is difficult to interpret and used by the RNN to account for memory effects. Furthermore, we use an output mapping function

$$\boldsymbol{p}_t = g_{\text{out}}(\boldsymbol{h}_t) \quad , \tag{1.12}$$

which yields (latent) variables and/or coefficients $\boldsymbol{p}_t$ required by the physical model

$$\boldsymbol{y}_t, \boldsymbol{s}_t = f_{\text{phys}}(\boldsymbol{p}_t, \boldsymbol{x}_t, \boldsymbol{s}_{t-1}) \quad . \tag{1.13}$$

In addition, the physical model f_{phys} takes forcings $\boldsymbol{x}_t$ and the past states $\boldsymbol{s}_{t-1}$ as input and yields the outputs $\boldsymbol{y}_t$ but also updates the physical states $\boldsymbol{s}_t$. Note that the two different model states $\boldsymbol{h}_t$ and $\boldsymbol{s}_t$ play a different role: The hidden state of the g_{RNN} is not physical, *i.e.*, we cannot link its values to real-world quantities. This state implicitly accounts for memory effects that are neglected in the physical model f_{phys}. The physical states $\boldsymbol{s}_t$ are interpretable, and we can constrain them to, for example, obey mass conservation laws. The above approach is just one way to account for memory effects in a hybrid model, but applications do, to my best knowledge, not yet exist beyond the studies presented in Chapter 4.

Challenges and opportunities

In physically-based modeling, coefficients are often assumed constant globally or across classes of similar instances. It is, for example, common to parameterize dynamic global vegetation models (DGVMs) by plant functional types (PFTs) instead of using spatially explicit coefficients (Sitch et al., 2003), *i.e.*, coefficients are defined per vegetation group. Furthermore, it is common practice to calibrate models per spatial or logical unit. For example, hydrological models are often calibrated per catchment (*e.g.*, Van Der Knijff et al., 2010), or catchments with similar properties are parameterized jointly (Beck et al., 2016). These approaches are sometimes sufficient, *e.g.*, if it makes sense to assume a constant value globally. However, the discretization often leads to model biases since the environmental processes are fine-grained and heterogeneous in reality (Reichstein et al., 2019). The flexibility of hybrid models to use spatially or temporally varying coefficients that

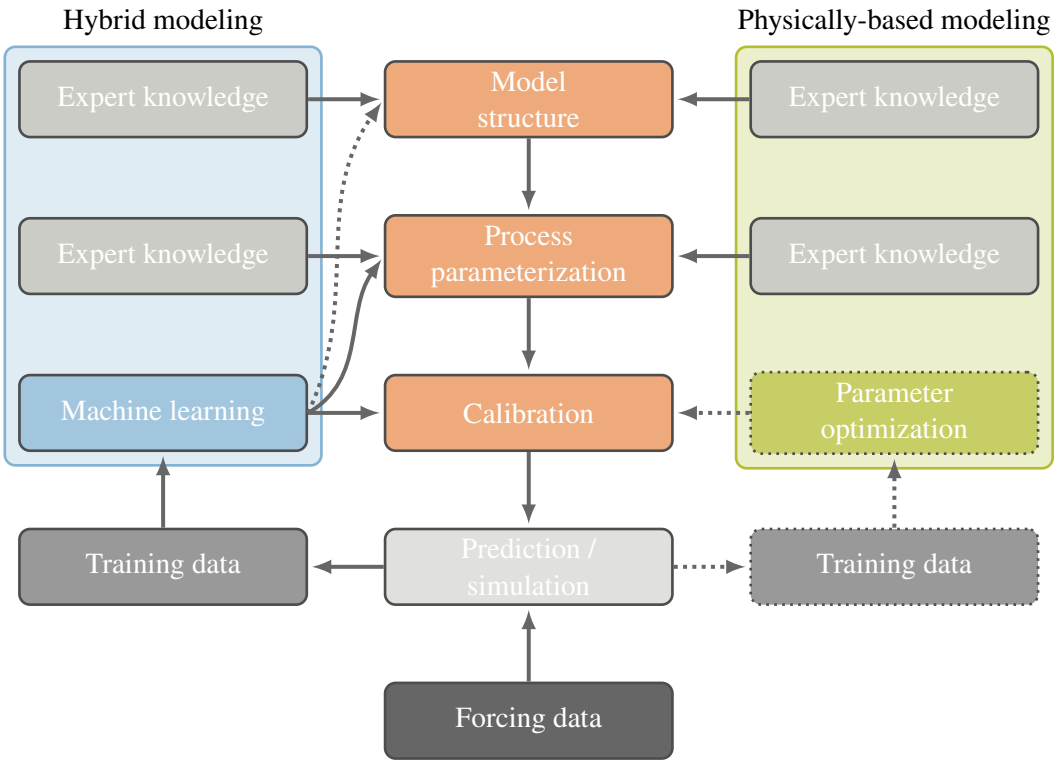


Figure 1.10.: Hybrid modeling versus physically-based modeling. The central column represents the modeling steps, from model structure (causal connections), to process parameterization (processes representation), to optional calibration (parameter tuning), and finally, the simulations based on forcing data. In physically-based modeling (right collumn), expert knowledge is used to design the model structure and to define the process parameterization. Parameters are either taken entirely from data and prior knowledge, or partially calibrated using optimization techniques. In hybrid modeling (left collumn), machine learning can be used in a more flexible way to parameterize processes and calibrate parameters, or to replace parts of the model structure.

are linked to data offers new pathways to improve environmental models: Not only is it possible to calibrate a physically-based model in a spatially and temporally explicit way, but we can also replace entire parameterizations or components of the model structure with a machine learning model as illustrated in Figure 1.10.

This flexibility of hybrid modeling offers a variety of opportunities for model development, *e.g.*, to replace uncertain processes with machine learning. But they also come at a cost: The modeling problem may already be underconstrained in physically-based models with only a few parameters, and this issue becomes even more severe when several coefficients are tuned simultaneously with a highly data-adaptive model. The lack of identifiability occurs when different parameter combinations lead to the same result. In environmental modeling, the term *equifinality* is more

common, referring to the case where identifiability is lacking (Beven and Freer, 2001). The issue of equifinality is an outstanding challenge in hybrid modeling.

1.5. Scope and thesis outline

In this thesis, I investigate the potential of deep learning and hybrid modeling for global-scale ecohydrological modeling. The main goal is to contribute to the exploration of the gap between physically-based modeling and machine learning, as illustrated in Figure 1.6.

The chapters are arranged along four key first-author publications, which focus on different aspects of the broader research question. Chapter 2 assesses the capability of RNNs to represent global ecosystem dynamics. This requires a model to be flexible enough to represent temporal ecohydrological dependencies in interaction with heterogeneous land surface conditions. To exclude potential confounding factors such as data quality and availability, the study is based on simulated data from a physically-based land surface model. In Chapter 3, the concept is applied to real-world observations. In addition, a model agnostic explanatory approach to identify temporal dependencies (*i.e.*, memory effects) is introduced and critically discussed. The explanatory approach can provide high-level qualitative insights into ecological memory effects and thereby contribute to populating the gap between the physically-based and the machine learning paradigm. Chapter 4 consists of two studies that represent a large step towards the combination of physically-based modeling and machine learning in a dynamic hybrid model at global scale. In Chapter 5, I summarize the contributions of this thesis to the current research landscape, discuss how the presented studies help to close the gap between physically-based modeling and machine learning, and what prospects arise from it.

More specifically, the research questions (RQs) are:

RQ1 Can recurrent neural networks learn global-scale ecosystem behavior? (Chapter 2)

RQ2 Can dynamic memory effects in Earth observations be identified using explanatory approaches? (Chapter 3)

RQ3 What is the promise of global-scale hybrid modeling and what are its challenges and opportunities? (Chapter 4)

1.6. List of publications

First-author publications

The following first-author publications are contained in this thesis.

S1 B. Kraft, S. Besnard, and S. Koirala (2021). "Emulating Ecological Memory with Recurrent Neural Networks." In: *Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science and Geosciences*. Ed. by G. Camps-Valls, D. Tuia, X. X. Zhu, and M. Reichstein. 1st edition. Hoboken, NJ: Wiley & Sons. ISBN: 978-1-119-64614-3[3]

A summary of the content and the original publication is provided in Chapter 2.

S2 B. Kraft, M. Jung, M. Körner, C. Requena Mesa, J. Cortés, and M. Reichstein (2019). "Identifying Dynamic Memory Effects on Vegetation State Using Recurrent Neural Networks." In: *Frontiers in Big Data* 2. ISSN: 2624-909X. DOI: 10.3389/fdata.2019.00031

A summary of the content and the original publication is provided in Chapter 3.

S3 B. Kraft, M. Jung, M. Körner, and M. Reichstein (2020). "Hybrid Modeling: Fusion of a Deep Learning Approach and a Physics-Based Model for Global Hydrological Modeling." In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. Vol. XLIII-B2-2020. Copernicus GmbH, pp. 1537–1544. DOI: 10.5194/isprs-archives-XLIII-B2-2020-1537-2020

A summary of the content and the original publication is provided in Chapter 4.

S4 B. Kraft, M. Jung, M. Körner, S. Koirala, and M. Reichstein (2022). "Towards hybrid modeling of the global hydrological cycle." In: *Hydrology and Earth System Sciences* 26.6, pp. 1579–1614. DOI: 10.5194/hess-26-1579-2022

A summary of the content and the original publication is provided in Chapter 4.

Second-author publications

The following second-author publications are not contained in this thesis.

1. C. Requena-Mesa, M. Reichstein, M. Mahecha, B. Kraft, and J. Denzler (2018). "Predicting Landscapes as Seen from Space from Environmental Conditions." In: *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, pp. 1768–1771. DOI: 10.1109/IGARSS.2018.8519427

2. M. Reichstein, S. Besnard, N. Carvalhais, F. Gans, M. Jung, B. Kraft, and M. Mahecha (2018). "Modelling Landsurface Time-Series with Recurrent Neural Nets." In: *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, pp. 7640–7643. DOI: 10.1109/IGARSS.2018.8518007

[3] Not peer-reviewed

2. Global ecosystem modeling using recurrent neural networks

This section is based on

B. Kraft, S. Besnard, and S. Koirala (2021). "Emulating Ecological Memory with Recurrent Neural Networks." In: *Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science and Geosciences*. Ed. by G. Camps-Valls, D. Tuia, X. X. Zhu, and M. Reichstein. 1st edition. Hoboken, NJ: Wiley & Sons. ISBN: 978-1-119-64614-3
Non-final author proof copy.

2.1. Study summary

RNNs have been broadly tested and applied on sequential modeling problems. Although various studies demonstrated their usefulness in the context of Earth observation data, global studies focusing on physical land-surface processes did not exist. The main reason for the lack of studies is the limited range of applications: Currently, long-term forecasting of land-surface processes using neural networks is not competitive to physically-based or sophisticated data assimilation approaches, and insights are limited due to the low interpretability of neural networks.

For studies such as Kraft et al. (2022), however, it is essential to know the capabilities and limitations of RNNs to represent land-surface processes under a broad range of conditions. Due to uncertainties and biases commonly present in Earth observation data, the study presented here uses simulations from a physically land-surface model to control for confounding factors, such as data deficiencies or incomplete predictors.

We used global simulations of daily evapotranspiration from the MATSIRO land-surface model, which uses meteorological forcings and a set of land-surface properties as input. The ecological memory is represented by a single latent state variable, which is soil moisture. To emulate the

processes related to evapotranspiration, we used an LSTM, which ideally learns the interactions of meteorological forcings and static variables across time scales.

We could show that an RNN is able to emulate the physically-based model when fed with the same forcings, which means that the interaction of meteorological forcings and static variables could be learned and generalized to unseen conditions. This study presents a proof-of-concept that underlines the data adaptivity of RNNs and the applicability to global Earth observation data.

Contribution The study was conducted in close cooperation with the co-authors. Data processing was equally done by all authors, while the model was implemented by me. All authors contributed equally to the analysis and writing.

2.2. Emulating ecological memory with recurrent neural networks

Please turn to the next page.

Chapter 18

Emulating Ecological Memory With Recurrent Neural Networks

Abstract

Ecosystem processes are driven both by contemporary and antecedent environmental and land surface conditions through *ecological memory effects*. This chapter provides an insight into the relevance of memory effects in the Earth system and presents an experimental case study to use an Recurrent Neural Network (RNN) model to emulate a physical model. In addition to introducing an experimental design suitable for such purposes, we demonstrate that an RNN is largely capable of learning the memory effects encoded in a physical model. A non-temporal fully connected model cannot reproduce such memory effects, especially during anomalous conditions (e.g. climate extremes).

18.1. Ecological memory effects: concepts and relevance

Ecological memory can be broadly defined as the encoding of past environmental conditions in the current ecosystem state that affects its future trajectory. The consequent effects, known as *memory effects*, are the direct influence of ecological memory on the current ecosystem functions [Peterson, 2002, Ogle et al., 2015]. Such memory effects are prevalent across several spatial and temporal scales. For example, at the seasonal scale, the variability of spring tem-

perature affects ecosystem productivity over the subsequent summer and autumn [Buermann et al., 2018]. Inter-annually, moisture availability over previous year is linked to contemporary ecosystem carbon uptake [Aubinet et al., 2018, Barron-Gafford et al., 2011, Ryan et al., 2015]. Furthermore, less frequent and large extreme events (*e.g.*, heat waves, frost, fires, insect outbreaks) can lead to short-term phenological changes [Marino et al., 2011] or long-term damage to the ecosystem with diverse effects on present and future ecosystem dynamics [Larcher, 2003, Lobell et al., 2012, Niu et al., 2014]. This evidence highlights the relevance of short to long-term temporal dependencies on past environmental conditions in terrestrial ecosystems. However, due to the large spectrum of the environmental conditions and their consequent effects on the ecosystem, quantifying and understanding the strength and persistence of memory effects is often challenging.

Ecological memory effects may comprise *direct* and *indirect* influences of *external* and *internal* factors [Ogle et al., 2015] that are either *concurrent* or *lagged* in time. For instance, a drought may directly decrease ecosystem productivity, with indirect concurrent effects on loss of biomass due to the drought-induced fire (t2 in Fig. 18.1). Additionally, ecosystems may not only be responding to concurrent factors, but also to the lagged effects of past environmental conditions. A drought event can further impact the ecosystem productivity for months to years through a direct but lagged effect. On the other hand, indirect lagged effects involve from external factors that affect the ecosystem productivity during a drought, *e.g.*, disturbances like tree mortality and deadwood accumulation (t3 in Fig. 18.1), which may lead to insect outbreaks with further influences on the ecosystem (t4 and t5 in Fig. 18.1).

Memory effects are not exclusive to ecosystem productivity, but encompass a large number of Earth system processes of carbon [Green et al., 2019] and water cycles [Humphrey et al., 2017]. A key variable that encodes memory effects in Earth system is the soil moisture. Soil moisture is controlled by instantaneous and long-term climate regimes, vegetation properties, soil hydraulic properties, topography, and geology. As such, soil moisture exhibits complex variabilities in space, time, and along soil depth. Owing to this central role but large complexity, most physical models are built around the parameterization of moisture storage, which in

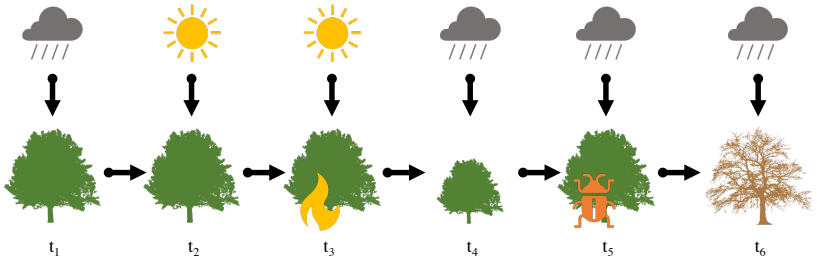


Figure 18.1: Schematic diagram illustrating the temporal forest dynamics during and post-disturbance: drought occurring in t1 and t2 conditions fire event in t2 and insect outbreaks in t4.

turn affects the responses of land surface to environmental conditions. Nevertheless, physical models have inherent uncertainties due to differences in structure and complexity and input data as well as unconstrained model parameters.

Several data-driven methods have therefore been developed to address the shortcomings of physical models for understanding Earth system processes as observed in the data. But, the data-driven methods may also be limited by data quality and availability. For example, the vegetation state over the land surface can be observed with satellite remote sensing. Yet, state variables such as soil moisture, which have imprints of ecological memory, are difficult to measure beyond meaningful soil depths and across larger scales. This poses a key challenge in capturing the memory effects using conventional data-driven methods. As such, dynamic statistical methods, such as RNNs [LeCun et al., 2015], may address these shortcomings, as they do not necessarily require measurements or observations of state variables. In this context, RNNs have a large potential for bringing the data-driven estimates on par with Earth system models with regards to capturing the ecological memory effects on land surface responses. This chapter focuses on this aspect and demonstrates the capabilities of RNNs to quantify memory effects with and without the use of state variables.

391

18.2. Data-driven approaches for ecological memory effects

18.2.1. A brief overview of memory effects

Conceptually, the memory effects on a system response Y_t, at time t, encompass the influences of forcing X_{t-k} in previous $k \geq 1$ time steps. As such, the memory effects propagate through time via the system state S_t, at every time step, which can be expressed as

$$S_t = f(S_{t-1}, X_t) \quad . \tag{18.1}$$

The response Y_t is, in turn, a function of S_t as

$$Y_t = g(S_t) \quad . \tag{18.2}$$

The S_t encodes all memory effects needed to compute Y_t, and it can be interpreted as the ecological memory. From a data-driven perspective, the memory S_t emerges solely from the effects of 'unobserved' previous states that are not directly encoded in any given observations [Jung et al., 2019]. For example, if instantaneous vegetation state (*e.g.*, vegetation greenness) and current climatic conditions (*e.g.*, air temperature or rainfall) are included in the observed state O_t, their effects are not necessarily encoded in S_t. Therefore, S_t can be mathematically expressed as

$$S_t = f(S_{t-1}, X_t, O_t) \quad . \tag{18.3}$$

392

18.2.2. Data-driven methods for memory effects

Following Equation 18.3, several data-driven statistical methods have been employed to account for ecological memory and quantify their effects on ecohydrological responses. Given the lack of observed state variables, a common practice is to use hand-designed features, such as lag or cumulative variables of past time-steps, in sequence-agnostic machine learning methods (*e.g.*, random forest, feed-forward networks) [Tramontana et al., 2016, Papagiannopoulou et al., 2017]. Although these methods generally work well, they do not capture the long-term dependencies of ecohydrological processes on past environmental conditions and interactions among different variables, as well as their complex temporal dynamics [Lipton et al., 2015]. Alternatively, Bayesian non-linear mixed-effects methods that consider joint probability distributions of different variables, have shown promising avenues to represent interactions and understand environmental and biological memory [Ogle et al., 2015, Liu et al., 2019].

Lastly, dynamic deep learning methods, such as RNNs, are capable of extracting temporal features. As such, they can represent ecosystem responses to past environmental conditions and capture ecological memory effects. In RNNs, analogous to Equation 18.1, a hidden state S_t is updated from the past state S_{t-1} and concurrent observations X_t. Owing to that, dynamic methods have been successfully applied in sequence learning (e.g., speech recognition) and land cover classification [Rußwurm and Körner, 2017].

With the increasing availability of remote sensing and climate data that span several decades, new avenues to employ temporally dynamic statistical methods like RNNs have been opened for exploring and understanding the known and unknown temporal dynamics of Earth system processes. In fact, such methods have already been applied to dynamically incorporate the effects of recent and past vegetation and climate dynamics on, for instance, ecosystem productivity [Reichstein et al., 2018], and the memory effects therein [Kraft et al., 2019]. Compared to static methods, the dynamic methods improve the prediction of seasonal dynamics of net carbon dioxide fluxes, with varying degrees of memory effects across different climate and ecosystem types [Besnard et al., 2019].

393

18.3. Case study: emulating a physical model using recurrent neural networks

As shown in previous studies, RNNs can potentially learn ecological memory [Reichstein et al., 2018, Besnard et al., 2019, Kraft et al., 2019]. It is, however, unclear under what conditions the RNNs can emulate the ecosystem responses, and to what extent the ecological memory play a role in defining these responses. Using RNNs for such questions in the real-world data is often challenging due to the data availability (*e.g.*, gaps in the remote sensing data), data uncertainty, and data inconsistency. Despite the limitations in data quality, the RNNs provide useful insights on ecosystem responses to past environmental conditions, albeit with inherent uncertainties. The validation of RNNs prediction would require more data including those from natural control and factorial experiments, but such data are hardly available.

To address this issue, we implement a series of experiments on a complete set of simulated data, *i.e.*, a simulation from a physical land surface model, to test whether—and to what extent—an RNN can learn ecological memory and simulate its effects on ecohydrological processes. The physical model simulation circumvent known limitations in measured Earth observation data, such as noise and biases, limited length of the time-series with potentially limited representation of the full range of environmental conditions, or incomplete set of variables. It should be noted that the physical model simulations are not the observed reality, but they provide a viable test bed for evaluating RNNs. Given the same input data, RNNs should be able to replicate the underlying processes included in the physical model. Such an exercise provides a robust assessment on the usefulness of dynamic statistical models for Earth system science. More specifically, in this upcoming sections, we demonstrate the capabilities of RNNs to:

1. emulate global spatio-temporal distributions of daily Evapotranspiration (ET) simulations from a physical land surface model;
2. quantify the effect of land surface states (e.g., soil moisture state) that are not directly provided as input to RNN;

3. evaluate capability of RNNs to capture the seasonal dynamics of ET under normal and extreme climatic conditions.

18.3.1. Physical model simulation data

The test data set for the RNN experiments was obtained from the simulations of a physically-based global land surface model, the MATSIRO [Takata et al., 2003, Koirala et al., 2014]. The MATSIRO is a land surface scheme of an Earth system model that simulates the water and energy budget over the land surface using physically-based representations of hydrological fluxes such as runoff, ET, and a cascade of storage components including snow, soil and groundwater. In the MATSIRO model, the hydrological fluxes are diagnosed based on the prognostic variation of hydrological storages. As such, memory effects of past climatic and environmental conditions on current fluxes are explicitly considered through their dependence on storage. In essence, the temporal variations of storage variables are constrained by physical mass balance equations, and can be represented as

$$S_t = f(S_{t-1}, X_t, Z_t) \quad , \tag{18.4}$$

where X_t represents the input drivers controlling the soil moisture S_t, such as precipitation, vegetation activity, and soil characteristics, and Z_t represents the output fluxes such as runoff and ET.

The output variables, Z_t, at any time, are non-linear and complex functions of climatic conditions and moisture storage, and thus include the memory effects of past conditions. Nevertheless, due to physical constraints of the mass balance equations, the model responses are mathematically tangible and depend exclusively on the input data and physical processes equations in the model. A brief overview of the input variables and their features are provided in Table 18.1.

Table 18.1: Data sets used in MATSIRO model simulation

	Variables	**Native resolution**	
		spatial	**temporal**
Spatial	Plant functional types, soil texture	0.5 degree	-
Spatial, seasonal and inter-annual	Rainfall, snowfall, air temperature, snowfall, downward shortwave and long-wave radiation, wind speed, specific humidity, surface pressure, Cloud cover, leaf area index	0.5 degree	daily

18.3.2. Experimental design

To assess the capability of an RNN to emulate the MATSIRO model, we implemented experiments to predict ET and its dependence on ecological memory provided through soil moisture. To do so, we also use the exact set of input variables (Table 18.1) from MATSIRO simulations.

Different RNN model setups were contrasted in a 2×2 factorial experiment design (Table 18.2). All RNN setups use at least the meteorological drivers and the static variables as inputs. We used the Long Short-Term Memory (LSTM) architecture [Hochreiter and Schmidhuber, 1997], capable of learning long-term dependencies and therefore accounting for ecological memory effects. If the temporal model without soil moisture ($\text{LSTM}_{\neg\text{SM}}$) is capable of learning the memory effects implicitly, its performance should be on par with a temporal model with soil moisture as an additional input (LSTM_{SM}), as ET is only dependent on soil moisture state in the MATSIRO model (*cf.* Section 18.3.1). In addition, two non-temporal models based on multiple Fully Connected (FC) layers were trained, one without soil moisture ($\text{FC}_{\neg\text{SM}}$), and one with soil moisture as input (FC_{SM}). While both models do not have access to past observations conceptually, the latter can use the concurrent soil moisture state. Contrasting the FC models allows to assess the local importance of soil moisture.

The predictions from four model set-ups were evaluated against the MATSIRO simulation at global and regional scales. At the grid-scale, the overall performances were evaluated using the Nash-Sutcliffe model efficiency coefficient (NSE) [Nash and Sutcliffe, 1970] and the Root

Table 18.2: Factorial experimental design: the four models are trained individually to assess the capability of an LSTM to learn ecological memory ($LSTM_{SM}$, with soil moisture *vs.* $LSTM_{\neg SM}$, without soil moisture as input) and to quantify the local relevance of soil moisture for ET (FC_{SM} *vs.* $FC_{\neg SM}$). The temporal models learn a mapping from the concurrent and past features $X_{\leq t}$ to the target Y_t, while the non-temporal models have access to the concurrent features X_t only. S_t is the ecosystem state, *i.e.*, soil moisture.

		model type			
		temporal		non-temporal	
model input	w/ SM	$LSTM_{SM}$	$Y_t = f(X_{\leq t}, S_t)$	FC_{SM}	$Y_t = f(X_t, S_t)$
	w/o SM	$LSTM_{\neg SM}$	$Y_t = f(X_{\leq t})$	$FC_{\neg SM}$	$Y_t = f(X_t)$

Mean Square Error (RMSE) [Omlin and Reichert, 1999]. Globally, the performance are also summarized across different temporal (daily, daily anomalies, daily seasonal cycle, interannual variation) scales. At the regional scale, our evaluation focused on the capability of LSTM to simulate temporal ET dynamics in two focus regions: the humid Amazon and semi-arid Australia [Boening et al., 2012]. In these two example cases, the mean seasonal cycle for the period 2001-2013 and seasonal anomalies observed during climate extreme events (2005 drought in the Amazon [Phillips et al., 2009] and the 2010 La Niña in Australia [Boening et al., 2012]) were evaluated. Table 18.3 summarizes the main features of the evaluations.

Table 18.3: Summary of the scope of the experiments.

	Objective	**Regions assessed**	**Period assessed**	**Input used**
Analysis 1	Use of RNNs for emulating physical models	global	2001-2013	Original input + soil moisture physical model outputs
Analysis 2	Simulating seasonal dynamics under normal and extreme conditions	Amazon basin and Australia	2001-2013, 2005 and 2010	Original input + soil moisture physical model outputs

18.3.3. RNN setup and training

As described in the previous section, two different RNN models were used: a temporal model (LSTM) and a non-temporal model (FC), *i.e.*, with stacked fully connected layers. All setups had the same input features as the MATSIRO model, and optionally soil moisture as added as an input variable (see Table 18.1). The models were trained on the MATSIRO ET simulations, with Mean Square Error (MSE) as a loss function.

The LSTM takes the multivariate time-series and static variables as input, which is followed by a hyperbolic tangent activation and a linear layer that maps the LSTM output at each time step to a single value: the predicted ET. The FC models consists of several fully connected layers, each followed by a non-linear activation function, except for the output layer, where no activation function is used. The FC model takes the static variables and only a single time-step of the time-series as input.

The final model architectures (Table 18.4) were selected using a hyper-parameter optimization approach: the Bayesian optimization hyper-band algorithm [Falkner et al., 2018]. The state-of-the-art optimization algorithm efficiently finds optimal hyper-parameters by combining an early stopping mechanism (dropping non-promising runs early) and a Bayesian sampling of promising hyper-parameters, wit h a surrogate loss model for the existing samples.To prevent over-fitting of the hyper-parameters, we used only every 6^{th} latitude/longitude grid-cell (approximately 3 % of the data) during hyper-parameter optimization. To avoid over-fitting of the residuals caused by temporal auto-correlation and to test how the model generalizes, the data were split into two sets: training data from 1981 to 1999 inclusive and test data from 2000 to 2013 inclusive. For both sets, an additional period of 5 years was used for model warm-up. For all four setups, the hyper-parameter optimization and model training were carried out independently.

398

Table 18.4: The model and training parameters from hyper-parameter optimization and their ranges searched. Both LSTM models (SM vs ¬SM) consist of several LSTM layers, followed by multiple fully connected layers. The non-temporal FC models consist of several stacked fully connected layers. In all setups, dropout was enabled for the input data and between all layers. Note that a dropout of 0.0 means that no dropout is applied.

Parameter	Search range	SM	¬SM
LSTM			
dropout (input)	(0.0, 0.5)	0.0	0.0
LSTM number of layers	(1, 3)	2	1
LSTM hidden size	(50, 300)	300	200
LSTM dropout	(0.0, 0.5)	0.4	0.3
FC number of layers	(2, 6)	3	5
FC hidden size	(50, 300)	300	300
FC activation	{ReLU, softplus, tanh}	ReLU	ReLU
FC dropout	(0.0, 0.5)	0.3	0.1
learning rate	(0.001, 0.0001)	0.001	0.001
weight decay	(0.01, 0.0001)	0.01	0.01
FC			
dropout (input)	(0.0, 0.5)	0.0	0.0
FC number of layers	(2, 6)	6	4
FC hidden size	(50, 600)	200	200
FC activation	{ReLU, softplus, tanh}	ReLU	ReLU
FC dropout	(0.0, 0.5)	0.0	0.0
learning rate	(0.001, 0.0001)	0.01	0.01
weight decay	(0.01, 0.0001)	0.001	0.001

18.4. Results and discussion

18.4.1. The predictive capability across scales

In this section, we evaluate the performances of the different RNN setups against the MATSIRO simulations. In general, it is evident that the LSTM model setups perform considerably better than the FC models (Fig. 18.2). In fact, outside the tropical humid regions, the LSTM models achieve a systematically higher predictive capacity than the FC models. The LSTM models have a higher median NSE ($LSTM_{SM}$: 0.98, $LSTM_{\neg SM}$: 0.97) and lower RMSE ($LSTM_{SM}$: 0.15, $LSTM_{\neg SM}$: 0.19) than the FC models (NSE of FC_{SM}: 0.93, $FC_{\neg SM}$: 0.89, and RMSE of FC_{SM}: 0.28, $FC_{\neg SM}$: 0.33). However, within the tropical humid regions, all setups have lower performance than in other regions (median NSE of 0.78, 0.75, 0.69, 0.57 and

median RMSE of 0.45, 0.48, 0.55, 0.61 for $LSTM_{SM}$, $LSTM_{\neg SM}$, FC_{SM}, and $FC_{\neg SM}$, respectively). This may be possibly associated with a larger variability in the water fluxes leading to a low signal-to-noise ratio in this region.

It can be hypothesized that an LSTM model can learn the ecological memory effect of soil moisture, even when soil moisture is not included as an input variable. Along this line, we find that the $LSTM_{SM}$ and $LSTM_{\neg SM}$ setups perform better compared to FC setups. This provides evidence that the two LSTM model architectures, with or even without soil moisture, are suitable for learning information content related to unseen state variable, such as soil moisture.

Yet, differentiating $LSTM_{SM}$ and $LSTM_{\neg SM}$ setups does not provide information on where the ecological memory of soil moisture is the strongest. We, therefore, plot the differences in term of predictive capacity between the model set-ups with and without soil moisture as an input variable (Figure 18.3). As expected, contrasting $LSTM_{\neg SM}$ with $LSTM_{SM}$ shows no substantial differences across the globe between the two LSTM models (first row of Fig. 18.3), suggesting no apparent memory effects of soil moisture on ET responses. On the other hand, the comparison of the FC models (second row of Fig. 18.3) suggested that the performances of the model with and without soil moisture can vary significantly in space. This is also reflected in the global model performance (NSE): While the 75^{th} percentile of the temporal ($LSTM_{\neg SM}$: 0.98) versus the non-temporal ($FC_{\neg SM}$: 0.95) models are similar, the 25^{th} percentile differs largely ($LSTM_{\neg SM}$: 0.94, $FC_{\neg SM}$: 0.86). This shows that the $LSTM_{\neg SM}$ model is capable of learning heterogeneous global dynamics, while the $FC_{\neg SM}$ model struggles in particular regions, which are, as we argue here, the ecosystems exhibiting strong memory effects. The differences in $FC_{\neg SM}$ model and the FC_{SM} setups were mostly apparent in water-limited regions. In these mostly semi-arid regions, the memory effects through soil moisture that are present, and influential [Koirala et al., 2014], in the MATSIRO simulations cannot be well reproduced by FC models, especially when soil moisture is not provided as an input variable.

We further investigated the performances of the model experiments across different temporal scales in training and test sets (Fig. 18.4). As it has been shown in Fig. 18.2, the two

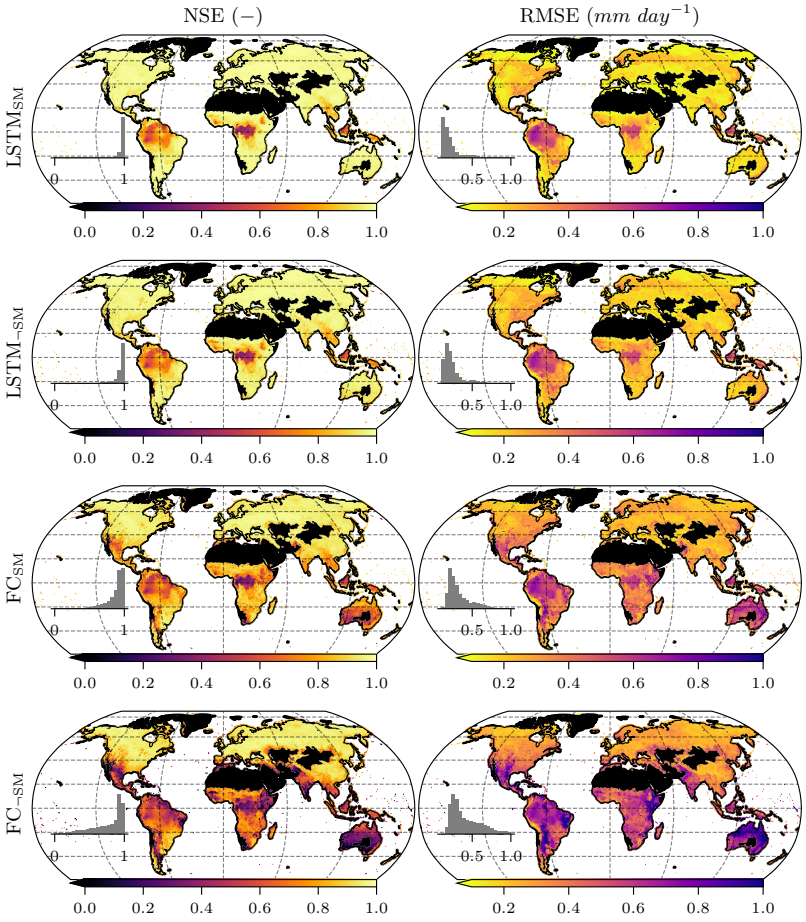


Figure 18.2: **Global distributions of performances of different model setups based on daily model predictions from the test dataset.** Nash-Sutcliffe model efficiency coefficient (NSE) is shown in the left and Root Mean Square Error (RMSE) in the right column for the temporal LSTM and non-temporal FC models with (SM) and without (¬SM) soil moisture, respectively. The inset histogram represents the global distribution of the metrics.

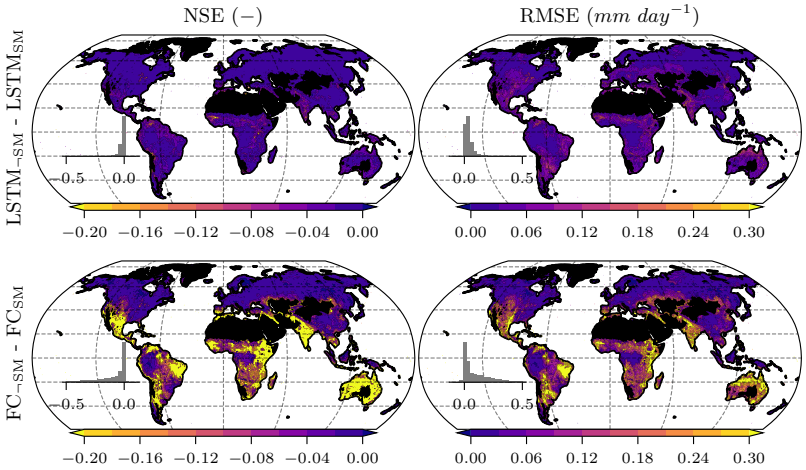


Figure 18.3: **Difference maps of Nash-Sutcliffe model efficiency coefficient (NSE) and Root Mean Square Error (RMSE) for the LSTM (first row) and FC (second row) models.** For the LSTM models, differences in NSE or RMSE were computed as $LSTM_{\neg SM}$ - $LSTM_{SM}$, while for the FC models, differences were computed as $FC_{\neg SM}$ - FC_{SM}. While the SM models receive the ecosystem state (soil moisture) as input, the ¬SM do not have access to the state. Red colors indicate that the SM model performs better than ¬SM. The inset histogram represents the global distribution of the differences.

LSTM models (shown in the blue box-plots) were able to learn the spatio-temporal daily patterns with NSE values close to 1 and a low variation across different grid-cells. We further found that the performance of LSTM models are relatively weaker for the predictions of daily and annual anomalies than that for mean daily seasonal cycle. The performances of the LSTM models were still good with a median NSE of 0.91 ($LSTM_{SM}$) and 0.88 ($LSTM_{\neg SM}$) for the anomalies.

The FC models performed worse than the LSTM models on the daily time series, particularly when soil moisture was not used as an input variable ($FC_{\neg SM}$). The decomposition of the daily time series into mean seasonal cycle and anomalies suggested that the lower performance of the FC models compared to the LSTM models, was mostly controlled by weaker performance with regards to anomalies (median NSE of 0.75 for FC_{SM} and 0.63 for $FC_{\neg SM}$).

The mean seasonal cycle was captured similarly well in the LSTM and FC models (median NSE from 0.97 to 1.00, where lowest is $FC_{\neg SM}$ and highest is $LSTM_{SM}$), although with a larger variability across grid-cells, with a 25^{th} to 75^{th} percentile of 0.95 to 1.00 (FC_{SM}) and 0.82 to 0.99 ($FC_{\neg SM}$) versus 1.00 to 1.00 ($LSTM_{SM}$) and 0.97 to 1.00 ($LSTM_{\neg SM}$). The model performances for anomalies were substantially lower for FC models compared to the LSTM models. These results suggest that ecological memory effects appear to be especially relevant for improving the model performance of capturing the daily and annual anomalies.

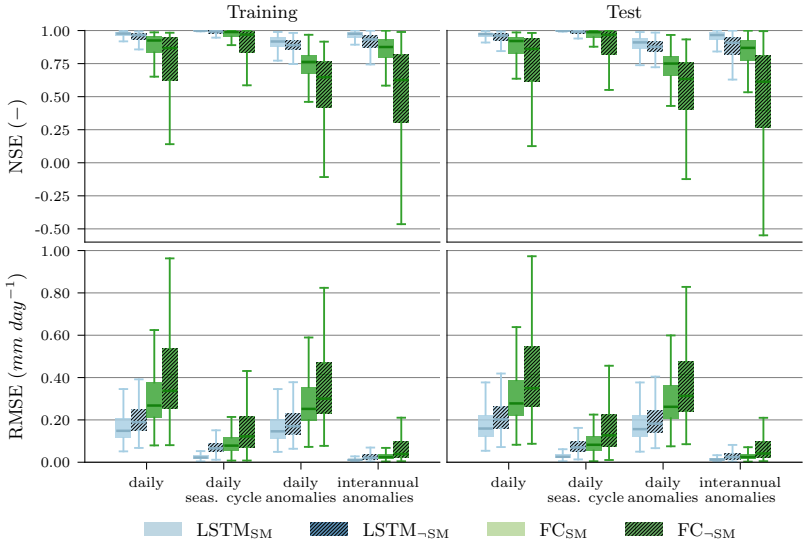


Figure 18.4: **Box and whisker plots showing grid-level model performances across time-scales (i.e. daily, daily seasonal cycle, daily anomalies, and annual anomalies) for the training and test sets.** Daily seasonal cycle are calculated as the mean of each day across different years, daily anomalies are calculated as the difference between daily raw estimates and the mean of each day and annual anomalies are calculated as the difference between mean annual and mean estimates within each grid-cell. Nash-Sutcliffe model efficiency coefficient (NSE) and Root Mean Square Error (RMSE) are shown. The whiskers represent the 1.5 · inter-quartile range (IQR) of the spatial variability of the model performances.

Surprisingly, the FC_{SM} model performed worse than the LSTM models, particularly for the

anomalies, even though the only relevant state variable for a given time step, SM_t, was known to the model. This contradiction may be associated with several factors. First, in MATSIRO simulation, the ET is based on the transient soil moisture with losses and gains within a day between the SM_{t-1} and SM_t. In the experiment here, SM_{t-1} was used as an input for the FC_{SM} model, and as such, one would expect some minor differences. Additionally, albeit hypothetical, the FC may not have enough capacity to extract high-level features for an instantaneous mapping from the concurrent time step of the input data, while the LSTM models can learn complex representations from a series of past time steps. Therefore, the LSTM can learn part of the ecological memory effects through temporal dynamics of soil moisture in addition to instantaneous soil moisture, compared to information used by the FC_{SM}. This also extends to potential utilization of distribution of input data by LSTM model, which has access to full global distribution of all the input data.

18.4.2. Prediction of seasonal dynamics

We have shown evidences of capabilities of RNN models in emulating a land surface model globally across different temporal scales. But, it is also worthwhile to analyze whether the model experiments can emulate a temporal dynamics in *normal* and *extreme/anomalous* climatic condition or not. This is an important factor, as extreme conditions are rare and only represent a fraction of the full data, that RNN models uses to learn about the dynamics.

In general, for the mean seasonal cycles of 2001-2013, the FC models is farther from the MATSIRO simulations in both the Amazon and Australian regions (Fig. 18.5, top row). But, not all the models perform well under all conditions. For example, in humid Amazon, the $LSTM_{SM}$ performs the best across all months, while other models performs relatively worse in drier condition (July-Dec). The mean seasonal variations of the residuals (second row) show that the LSTM models can better learn temporal dynamics of ET than FC models, as the residuals for these models (blue lines) is closer to zero over the entire year. The FC models have larger residuals, with particularly high values for $FC_{\neg SM}$ model, especially during the dry season in the Amazon basin and over the growing season in Australia (August to May).

404

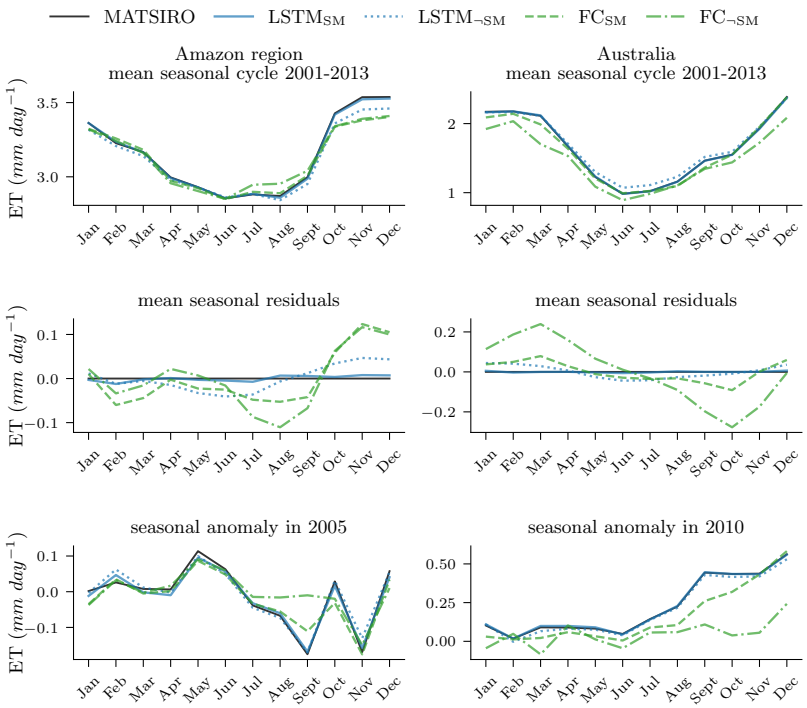


Figure 18.5: **Seasonal cycle (first row), seasonal variation of the residuals (second row) and seasonal anomaly (third row) in the Amazon region (first column) and Australia (second column).** Seasonal residuals were computed as ET residuals$_i$ = [ET MATSIRO$_i$ - mean(ET MATSIRO)] - [ET predicted$_i$ - mean(ET predicted)], where i is a monthly observation. Seasonal anomalies are shown for the years 2005 and 2010 for the Amazon region and Australia, respectively.

The high values in the seasonal patterns of residuals in Australia for the FC$_{\neg SM}$ experiment but not in the FC$_{SM}$ model suggested an apparent importance of soil moisture in controlling ET in this region.

We further investigate the performance of LSTM models under two extreme climatic conditions: 2005 drought in the Amazon, and the 2010 La Niña in Australia (Fig. 18.5, bottom

row). The $LSTM_{SM}$ (dashed blue line) and $LSTM_{\neg SM}$ (solid blue line) models can reproduce the MATSIRO simulation of strong seasonal anomalies even under the extreme conditions (second row). As also shown in the previous sections, the FC_{SM} model cannot reproduce the seasonal anomalies as well as the LSTM models.

18.5. Conclusions

This chapter provided an overview of ecological memory effects in the Earth system, along with a case for the application of a deep learning method, the RNNs, for representing ecological memory effects. The case study used the simulations of a physical model as a pseudo-observation to evaluate the capabilities of RNNs models to predict ET and ecological memory effects therein.

The LSTM model was able to capture the ecological memory effects inherent in the physical model. Moreover, the difference in the performances of the LSTM model with and without soil moisture state was found to be negligible. This appeared to be consistent from daily to annual temporal scales, and over most regions globally. This finding demonstrated that the LSTM, through its hidden states, is indeed able to learn the memory effects that are explicitly encoded in the state variables of a physical model.

We further found that the LSTM was able to predict the soil moisture-ET dynamics even during anomalous climatic conditions demonstrating that the predictions of the LSTM are general and universally applicable under wide range of environmental conditions. This was true for seasonal responses of ET to the 2005 dry spell in the Amazon, and 2010 La Niña event in Australia. The non-temporal FC models generally performed worse, especially with regards to anomalies when soil moisture was not given as input ($FC_{\neg SM}$). Under the assumption that the physical model is analogous to the reality, the poorer performance of the model can be interpreted as the importance of memory effects of soil moisture on ET. The relatively weaker performance of the FC model, that has access to soil moisture (FC_{SM}), compared to the LSTM architectures could not be explained conceptually. We hypothesize that access to the

distribution of the past climate observations in the LSTM models may be associated with its better performance.

In summary, our results based on simulations of a physical model demonstrated the usefulness of LSTM model architecture for learning the dynamics and the ecological memory of unobserved state variables, such as soil moisture. This justifies the need, and provides confidence, for use of dynamic statistical model, such as LSTM, when investigating temporally dependent ecohydrological processes using often limited observation-based data set. The coupling of dynamic data-driven methods either with physically-based models (i.e., hybrid modeling, or with complementary machine learning approaches (e.g., convolutional neural networks, paves the way for a better understanding of the known as well as unknown Earth system processes.

408

Bibliography

Marc Aubinet, Quentin Hurdebise, Henri Chopin, Alain Debacq, Anne De Ligne, Bernard Heinesch, Tanguy Manise, and Caroline Vincke. Inter-annual variability of Net Ecosystem Productivity for a temperate mixed forest: A predominance of carry-over effects? *Agricultural and Forest Meteorology*, 262:340–353, 2018. doi: 10.1016/j.agrformet.2018.07.024.

Greg A Barron-Gafford, Russell L Scott, G Darrel Jenerette, and Travis E Huxman. The relative controls of temperature, soil moisture, and plant functional group on soil co2 efflux at diel, seasonal, and annual scales. *Journal of Geophysical Research: Biogeosciences*, 116, 2011. doi: 10.1029/2010JG001442.

Simon Besnard, Nuno Carvalhais, M. Altaf Arain, Andrew Black, Benjamin Brede, Nina Buchmann, Jiquan Chen, Jan G. P. W. Clevers, Loïc P. Dutrieux, Fabian Gans, Martin Herold, Martin Jung, Yoshiko Kosugi, Alexander Knohl, Beverly E. Law, Eugénie Paul-Limoges, Annalea Lohila, Lutz Merbold, Olivier Roupsard, Riccardo Valentini, Sebastian Wolf, Xudong Zhang, and Markus Reichstein. Memory effects of climate and vegetation affecting net ecosystem CO_2 fluxes in global forests. *PLOS ONE*, 14:e0211510, 2019. doi: 10.1371/journal.pone.0211510.

Carmen Boening, Josh K. Willis, Felix W. Landerer, R. Steven Nerem, and John Fasullo. The 2011 la niña: So strong, the oceans fell. *Geophysical Research Letters*, 39, 2012. doi: 10.1029/2012GL053055.

Wolfgang Buermann, Matthias Forkel, Michael O'Sullivan, Stephen Sitch, Pierre Friedlingstein, Vanessa Haverd, Atul K. Jain, Etsushi Kato, Markus Kautz, Sebastian Lienert, Danica Lombardozzi, Julia E. M. S. Nabel, Hanqin Tian, Andrew J. Wiltshire, Dan Zhu, William K. Smith, and Andrew D. Richardson. Widespread seasonal compensation ef-

fects of spring warming on northern plant productivity. *Nature*, 562:110, 2018. doi: 10.1038/s41586-018-0555-7.

Stefan Falkner, Aaron Klein, and Frank Hutter. Bohb: Robust and efficient hyperparameter optimization at scale. *arXiv preprint arXiv:1807.01774*, 2018.

Julia K. Green, Sonia I. Seneviratne, Alexis M. Berg, Kirsten L. Findell, Stefan Hagemann, David M. Lawrence, and Pierre Gentine. Large influence of soil moisture on long-term terrestrial carbon uptake. *Nature*, 565:476–479, 2019. doi: 10.1038/s41586-018-0848-x.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9:1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

Vincent Humphrey, Lukas Gudmundsson, and Sonia I Seneviratne. A global reconstruction of climate-driven subdecadal water storage variability. *Geophysical Research Letters*, 44(5): 2300–2309, 2017. doi: 10.1002/2017GL072564.

Martin Jung, Christopher Schwalm, Mirco Migliavacca, Sophia Walther, Gustau Camps-Valls, Sujan Koirala, Peter Anthoni, Simon Besnard, Paul Bodesheim, Nuno Carvalhais, Frederic Chevallier, Fabian Gans, Daniel S. Groll, Vanessa Haverd, Kazuhito Ichii, Atul K. Jain, Junzhi Liu, Danica Lombardozzi, Julia E. M. S. Nabel, Jacob A. Nelson, Martijn Pallandt, Dario Papale, Wouter Peters, Julia Pongratz, Christian Rödenbeck, Stephen Sitch, Gianluca Tramontana, Ulrich Weber, Markus Reichstein, Philipp Koehler, Michael O'Sullivan, and Anthony Walker. Scaling carbon fluxes from eddy covariance sites to globe: Synthesis and evaluation of the FLUXCOM approach. *Biogeosciences Discussions*, pages 1–40, 2019. doi: https://doi.org/10.5194/bg-2019-368.

Sujan Koirala, Pat J.-F. Yeh, Yukiko Hirabayashi, Shinjiro Kanae, and Taikan Oki. Global-scale land surface hydrologic modeling with the representation of water table dynamics. *Journal of Geophysical Research: Atmospheres*, 119:75–89, 2014. doi: 10.1002/2013JD020398.

Basil Kraft, Jung Martin, Marco Körner, Christian Requena Mesa, José Cortés, and Markus

Reichstein. Identifying dynamic memory effects on vegetation state using recurrent neural networks. *Frontiers in Big Data*, 2:31, 2019. doi: 10.3389/fdata.2019.00031.

Walter Larcher. *Physiological Plant Ecology: Ecophysiology and Stress Physiology of Functional Groups*. Springer-Verlag, Berlin Heidelberg, 2003. ISBN 978-3-540-43516-7.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521:436, 2015. doi: 10.1038/nature14539.

Zachary C. Lipton, John Berkowitz, and Charles Elkan. A critical review of recurrent neural networks for sequence learning. *arXiv:1506.00019 [cs]*, 2015.

Yao Liu, Christopher R Schwalm, Kimberly E Samuels-Crow, and Kiona Ogle. Ecological memory of daily carbon exchange across the globe and its importance in drylands. *Ecology Letters*, 22:1806–1816, 2019. doi: 10.1111/ele.13363.

David B. Lobell, Adam Sibley, and J. Ivan Ortiz-Monasterio. Extreme heat effects on wheat senescence in India. *Nature Climate Change*, 2:186–189, 2012. doi: 10.1038/nclimate1356.

Garrett P. Marino, Dale P. Kaiser, Lianhong Gu, and Daniel M. Ricciuto. Reconstruction of false spring occurrences over the southeastern United States, 1901–2007: an increasing risk of spring freeze damage? *Environmental Research Letters*, 6:024015, 2011. doi: 10.1088/1748-9326/6/2/024015.

J Eamonn Nash and Jonh V Sutcliffe. River flow forecasting through conceptual models part i—a discussion of principles. *Journal of hydrology*, 10(3):282–290, 1970. doi: 10.1016/0022-1694(70)90255-6.

Shuli Niu, Yiqi Luo, Dejun Li, Shuanghe Cao, Jianyang Xia, Jianwei Li, and Melinda D. Smith. Plant growth and mortality under climatic extremes: An overview. *Environmental and Experimental Botany*, 98:13–19, 2014. doi: 10.1016/j.envexpbot.2013.10.004.

Kiona Ogle, Jarrett J. Barber, Greg A. Barron-Gafford, Lisa Patrick Bentley, Jessica M. Young, Travis E. Huxman, Michael E. Loik, and David T. Tissue. Quantifying ecological

memory in plant and ecosystem processes. *Ecology Letters*, 18:221–235, 2015. doi: 10.1111/ele.12399.

Martin Omlin and Peter Reichert. A comparison of techniques for the estimation of model prediction uncertainty. *Ecological modelling*, 115:45–59, 1999.

Christina Papagiannopoulou, DG Miralles, Wouter A Dorigo, NEC Verhoest, Mathieu Depoorter, and Willem Waegeman. Vegetation anomalies caused by antecedent precipitation in most of the world. *Environmental Research Letters*, 12:074016, 2017. doi: 10.1088%2F1748-9326%2Faa7145.

Garry D. Peterson. Contagious disturbance, ecological memory, and the emergence of landscape pattern. *Ecosystems*, 5:329–338, 2002. doi: 10.1007/s10021-001-0077-1.

Oliver L. Phillips, Luiz E. O. C. Aragão, Simon L. Lewis, Joshua B. Fisher, Jon Lloyd, Gabriela López-González, Yadvinder Malhi, Abel Monteagudo, Julie Peacock, Carlos A. Quesada, Geertje van der Heijden, Samuel Almeida, Iêda Amaral, Luzmila Arroyo, Gerardo Aymard, Tim R. Baker, Olaf Bánki, Lilian Blanc, Damien Bonal, Paulo Brando, Jerome Chave, Átila Cristina Alves de Oliveira, Nallaret Dávila Cardozo, Claudia I. Czimczik, Ted R. Feldpausch, Maria Aparecida Freitas, Emanuel Gloor, Niro Higuchi, Eliana Jiménez, Gareth Lloyd, Patrick Meir, Casimiro Mendoza, Alexandra Morel, David A. Neill, Daniel Nepstad, Sandra Patiño, Maria Cristina Peñuela, Adriana Prieto, Fredy Ramírez, Michael Schwarz, Javier Silva, Marcos Silveira, Anne Sota Thomas, Hans ter Steege, Juliana Stropp, Rodolfo Vásquez, Przemyslaw Zelazowski, Esteban Alvarez Dávila, Sandy Andelman, Ana Andrade, Kuo-Jung Chao, Terry Erwin, Anthony Di Fiore, Eurídice Honorio C, Helen Keeling, Tim J. Killeen, William F. Laurance, Antonio Peña Cruz, Nigel C. A. Pitman, Percy Núñez Vargas, Hirma Ramírez-Angulo, Agustín Rudas, Rafael Salamão, Natalino Silva, John Terborgh, and Armando Torres-Lezama. Drought sensitivity of the amazon rainforest. *Science*, 323:1344–1347, 2009. doi: 10.1126/science.1164033.

M. Reichstein, S. Besnard, N. Carvalhais, F. Gans, M. Jung, B. Kraft, and M. Mahecha. Modelling Landsurface Time-Series with Recurrent Neural Nets. In *IGARSS 2018 - 2018*

IEEE International Geoscience and Remote Sensing Symposium, pages 7640–7643, 2018. doi: 10.1109/IGARSS.2018.8518007.

Marc Rußwurm and Marco Körner. Temporal Vegetation Modelling Using Long Short-Term Memory Networks for Crop Identification from Medium-Resolution Multi-spectral Satellite Images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1496–1504, 2017. doi: 10.1109/CVPRW.2017.193.

Edmund M Ryan, Kiona Ogle, Tamara J Zelikova, Dan R LeCain, David G Williams, Jack A Morgan, and Elise Pendall. Antecedent moisture and temperature conditions modulate the response of ecosystem respiration to elevated co 2 and warming. *Global change biology*, 21: 2588–2602, 2015. doi: 10.1111/gcb.12910.

Kumiko Takata, Seita Emori, and Tsutomu Watanabe. Development of the minimal advanced treatments of surface interaction and runoff. *Global and Planetary Change*, 38:209–222, 2003. doi: 10.1016/S0921-8181(03)00030-4.

Gianluca Tramontana, Martin Jung, Christopher R. Schwalm, Kazuhito Ichii, Gustau Camps-Valls, Botond Ráduly, Markus Reichstein, M. Altaf Arain, Alessandro Cescatti, Gerard Kiely, Lutz Merbold, Penelope Serrano-Ortiz, Sven Sickert, Sebastian Wolf, and Dario Papale. Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms. *Biogeosciences*, 13:4291–4313, 2016. doi: 10.5194/bg-13-4291-2016.

3. Quantifying ecological memory effects using explanations

This section is based on

B. Kraft, M. Jung, M. Körner, C. Requena Mesa, J. Cortés, and M. Reichstein (2019). "Identifying Dynamic Memory Effects on Vegetation State Using Recurrent Neural Networks." In: *Frontiers in Big Data* 2. ISSN: 2624-909X. DOI: 10.3389/fdata.2019.00031

3.1. Study summary

This chapter contains a peer-reviewed study that demonstrates the capability of RNNs to represent ecological memory effects of climate variations on vegetation state on the global scale. Furthermore, a permutation-based approach to identify memory effects is introduced: The time-series data is blocked and permuted during model training to interrupt the sequential order, limiting the ecological memory the RNN can learn. The impact of memory effects (past meteorological forcings on vegetation state) is quantified by comparing the performance of models accounting for different temporal context. Previous studies derived their insights from simple linear or shallow machine learning models that require manually designed input features. In contrast, our approach requires minimal prior knowledge and profits from the data adaptiveness of RNNs.

The patterns we found fall in line with prior knowledge. The results agreed to the findings from previous studies in general, while local patterns differed. The presented approach constitutes a first try to use the flexibility of deep learning models to gather insight into ecological memory despite their low interpretability.

The approach can easily be applied to other domains and is especially useful when dealing with complex systems or limited process understanding. Still, the insights remain qualitative and

[1] https://creativecommons.org/licenses/by/4.0/

high-level, and the attribution of memory effects to specific meteorological variables is not possible using the presented approach, a challenge that may be addressed in future studies.

Contribution While the challenge to identify memory effects using RNNs was given by the supervisors, I took the major steps towards method development, always in close interaction with the co-authors. A co-author conducted a statistical test for the significance of the results on cell level. All co-authors contributed in meetings, where results and next steps were discussed regularly. The manuscript was written in collaboration.

3.2. Identifying dynamic memory effects using recurrent neural networks

Please turn to the next page.

frontiers in Big Data

ORIGINAL RESEARCH
published: 23 October 2019
doi: 10.3389/fdata.2019.00031

Identifying Dynamic Memory Effects on Vegetation State Using Recurrent Neural Networks

Basil Kraft[1,2]*, *Martin Jung*[1], *Marco Körner*[2], *Christian Requena Mesa*[1,3,4], *José Cortés*[1,5] *and Markus Reichstein*[1]

[1] Department of Biogeochemical Integration, Max Planck Institute for Biogeochemistry, Jena, Germany, [2] Department of Aerospace and Geodesy, Technical University of Munich, Munich, Germany, [3] German Aerospace Center (DLR), Institute of Data Science, Jena, Germany, [4] Department of Computer Science, Friedrich Schiller University, Jena, Germany, [5] Department of Geography, Friedrich Schiller University, Jena, Germany

OPEN ACCESS

Edited by:
Alexandra Konings,
Stanford University, United States

Reviewed by:
Youngryel Ryu,
Seoul National University, South Korea
Yi Yin,
California Institute of Technology, United States
Christina Papagiannopoulou,
Ghent University, Belgium

***Correspondence:**
Basil Kraft
bkraft@bgc-jena.mpg.de

Specialty section:
This article was submitted to Data-driven Climate Sciences, a section of the journal Frontiers in Big Data

Received: 18 April 2019
Accepted: 22 August 2019
Published: 23 October 2019

Citation:
Kraft B, Jung M, Körner M, Requena Mesa C, Cortés J and Reichstein M (2019) Identifying Dynamic Memory Effects on Vegetation State Using Recurrent Neural Networks. Front. Big Data 2:31. doi: 10.3389/fdata.2019.00031

Vegetation state is largely driven by climate and the complexity of involved processes leads to non-linear interactions over multiple time-scales. Recently, the role of temporally lagged dependencies, so-called memory effects, has been emphasized and studied using data-driven methods, relying on a vast amount of Earth observation and climate data. However, the employed models are often not able to represent the highly non-linear processes and do not represent time explicitly. Thus, data-driven study of vegetation dynamics demands new approaches that are able to model complex sequences. The success of Recurrent Neural Networks (RNNs) in other disciplines dealing with sequential data, such as Natural Language Processing, suggests adoption of this method for Earth system sciences. Here, we used a Long Short-Term Memory (LSTM) architecture to fit a global model for Normalized Difference Vegetation Index (NDVI), a proxy for vegetation state, by using climate time-series and static variables representing soil properties and land cover as predictor variables. Furthermore, a set of permutation experiments was performed with the objective to identify memory effects and to better understand the scales on which they act under different environmental conditions. This was done by comparing models that have limited access to temporal context, which was achieved through sequence permutation during model training. We performed a cross-validation with spatio-temporal blocking to deal with the auto-correlation present in the data and to increase the generalizability of the findings. With a full temporal model, global NDVI was predicted with R^2 of 0.943 and *RMSE* of 0.056. The temporal model explained 14% more variance than the non-memory model on global level. The strongest differences were found in arid and semiarid regions, where the improvement was up to 25%. Our results show that memory effects matter on global scale, with the strongest effects occurring in sub-tropical and transitional water-driven biomes.

Keywords: memory effects, lag effects, recurrent neural network (RNN), long short-term memory (LSTM) network, normalized difference vegetation index (NDVI)

1. INTRODUCTION

In the past decades, terrestrial ecosystems have been recognized to play a key role in the global carbon cycle as a sink of atmospheric CO_2, acting as a buffer for human carbon emissions (Bonan, 2015). Links between terrestrial carbon uptake to short- and midterm climate variations are still poorly understood and therefore, identifying driving mechanisms of vegetation state is crucial (Reichstein et al., 2013).

While the large-scale spatial distribution of vegetation mainly depends on climatologies, short-term dependencies of vegetation dynamics on climate variability are more complex (Papagiannopoulou et al., 2017a). This complexity expresses in dynamic interactions on multiple temporal scales, generating patterns that can only be understood and predicted considering antecedent ecosystem states and environmental conditions (Chave, 2013; De Keersmaecker et al., 2015; Seddon et al., 2016). These time-lagged impacts, so-called memory effects, have long been neglected, but have gained attention recently (Frank et al., 2015; Ogle et al., 2015).

Recently, different studies investigated memory effects to understand how vegetation reacts to climate on global level and how vulnerable ecosystems are toward weather extremes. Still, a profound comprehension of the involved processes is lacking (Ogle et al., 2015). Nevertheless, progress toward a better understanding was made. Seddon et al. (2016), for example, used an auto-regressive approach to model vegetation state as a function of temperature, water availability, cloud cover and the past vegetation state to determine sensitivity of vegetation toward and importance of the climate drivers. Similarly, De Keersmaecker et al. (2015) deployed a multiple linear regression model to analyze ecosystem resistance and resilience. They modeled anomalies of Normalized Difference Vegetation Index (NDVI), a proxy for vegetation state (Tucker, 1979), as a function of temperature anomalies, a drought index and past NDVI anomalies. Liu et al. (2018) used multiple linear regression to investigate the sensitivity of vegetation toward climate variability and to assess water memory effects. Wu et al. (2015) analyzed the impact of temperature, precipitation and solar short-wave irradiation on vegetation state, using a linear regression framework with lagged variables. In the mentioned studies, the learned model coefficients were linked to memory effects or the closely related ecosystem resilience. These studies provided important insights into memory effects, meteorological drivers of vegetation and its sensitivity toward environmental conditions. However, there is evidence that linear models are not able to adequately represent the temporal interactions inherent to ecosystem processes (Papagiannopoulou et al., 2017a). Thus, non-linear approaches that can cope with this complexity, are worthwhile exploring. To this end, Papagiannopoulou et al. (2017a) developed a Granger causality framework based on random forests to analyze the impact of climate drivers on anomalies of vegetation state and showed that non-linear approaches are needed to model vegetation dynamics. Other non-linear approaches to study global vegetation dynamics, however, have not been tested to our knowledge.

We take this opportunity to test the applicability of a state-of-the-art machine learning model to study global memory effects: Recurrent Neural Networks (RNNs). RNNs maintain a hidden state representing the system's memory (Werbos, 1990; Goodfellow et al., 2016). This memory evolves through time and is accessed for making predictions in interaction with concurrent factors. The model learns during training what share of information must be retained, forgotten and updated in order to predict the target variable and thus learns a complex representation of the modeled system. A widely used instance of the RNN model is the Long Short-Term Memory (LSTM) network (Hochreiter and Schmidhuber, 1997; Greff et al., 2017) that solves some of the shortcomings of the standard RNN. LSTMs have been proven to perform excellently on sequential data, for example in speech recognition (Graves et al., 2013), energy load forecasting (Marino et al., 2016), or crop field classification (Rußwurm and Körner, 2017, 2018). LSTMs model time explicitly and can learn interactions on multiple time-scales (Lipton et al., 2015; Reichstein et al., 2019) and can easily be extended in a modular fashion. Further, LSTMs allow the usage of raw time-series as input rather than lagged and aggregated features. For an introduction to Deep Learning and related terms we refer to Goodfellow et al. (2016), also available online (https://www.deeplearningbook.org/). For Deep Learning in the context of Earth system sciences, we recommend Reichstein et al. (2019).

In this study, we model NDVI using precipitation, temperature, short-wave irradiation and relative humidity, together with static variables representing land cover and soil properties as predictor variables. To quantify memory effects, we test and extend a time-series permutation approach that has been contemplated by Reichstein et al. (2018) and applied to CO_2 fluxes at site level by Besnard et al. (2019). By permuting the feature and target time-series in unison during model training, the model is restricted to learn instantaneous effects only, which allows to quantify the model improvement when including memory effects. Here, we extend this method by using a block-permutation approach: By successively permuting the time-series while keeping blocks of a given length in original order during training, we limit the access to past observations of meteorological time-series to a specific length. The different models are then analyzed and compared to improve our understanding of memory effects. We consider this study a "proof of concept" that introduces a new approach for using machine learning for process understanding.

2. MATERIALS AND METHODS

2.1. Vegetation Data (NDVI)

The Global Inventory Monitoring and Modeling System (GIMMS) NDVI 3g v1 (update of the NDVI 3g v0 dataset, Pinzon and Tucker, 2014) is a widely used, 15-daily global product based on data collected by the Advanced Very High Resolution Radiometer (AVHRR) that spans the period of July 1981 to December 2015. We used 33 years of the data from 1983 to 2015 (792 time-steps) in order to match the cross-validation scheme described later. To match other data used in this study and to

reduce noise as well as observations gaps, the NDVI data was aggregated from its original spatial resolution of 0.083 to 0.5°. Only non-interpolated observations with good quality were used, and pixel-time-steps were dropped if more than 50% missing data was present at aggregation level. Also, aggregated pixels with more than 50% missing data in the time dimension were rejected, which mainly removes high latitude regions. Finally, pixels with more than 20% water are dropped to exclude coastal areas, and such with more than 50% barren were removed to exclude deserts. This speeds up model training while only locations with a marginal vegetation signal are removed.

2.2. Explanatory Variables

A total of 27 explanatory variables were used of which 6 were dynamic and 21 static. The dynamic variables 2 m air temperature (mean, minimum, and maximum), 2 m relative humidity and incoming short-wave radiation from ERA-Interim (Dee et al., 2011) and precipitation from the Multi-Source Weighted-Ensemble Precipitation (MSWEP) global precipitation dataset version 2.0 (Beck et al., 2019) were temporally aggregated to match the 15-daily NDVI data. Static variables used are Available Water Capacity from the Harmonized World Soil Database version 1.1 (FAO/IIASA/ISRIC/ISSCAS/JRC, 2009) and Water Table Depth (Fan et al., 2013, provided by the Global Water Scarcity Information Service: http://glowasis.eu). In addition, Land Cover Fractions (LCF) for the classes *Water*, *Evergreen Needleleaf Forest*, *Evergreen Broadleaf Forest*, *Deciduous Needleleaf Forest*, *Deciduous Broadleaf Forest*, *Mixed Forest*, *Closed Shrublands*, *Open Shrublands*, *Woody Savannas*, *Savannas*, *Grasslands*, *Permanent Wetlands*, *Croplands*, *Urban and Built-up*, *Cropland/Natural vegetation mosaic*, *Snow and ice*, *Barren or Sparsely Vegetated* were derived from Moderate Resolution Imaging Spectroradiometer (MODIS) MCD12Q1 collection 5 (Friedl et al., 2010). Finally, C4 fractions for the classes *Croplands* and *Croplands/Natural Vegetation* mosaic were obtained from Monfreda et al. (2008). All data was aggregated to 0.5° resolution. For an analysis of the effect of using static variables as predictors on the model performance and patterns of memory effects, we refer the reader to the **Supplementary Material**, section 1.

2.3. Modeling Approach

To model global vegetation dynamics, we chose an RNN architecture. RNNs efficiently encode information seen at past time-steps. This property emerges from its hidden state h, representing the memory of the network (Goodfellow et al., 2016). Information is extracted context-based from the state $h^{(t-1)}$ and is used together with predictor $X^{(t)}$ to compute output $h^{(t)}$, which is also the input for the next time-step. An extensively reported issue with the standard RNN is the vanishing and exploding gradient problem (Pascanu et al., 2013), which limits its power to capture long-term dependencies. Thus, more complex models, such as the LSTM are used in practice to circumvent this issue (Greff et al., 2017).

The model architecture is illustrated in **Figure 1**. To find an optimal set of hyper-parameters for the model, we performed a grid search (searched range reported in brackets). The 27 predictor variables were standardized and each time-step was passed through a fully connected neural network with 2 (1–3) layers, each consisting of 128 (32–256) nodes. Dropout regularization of 0.1 (0.0–0.4) was applied after both layers. The output was used as input for a single (1–3) LSTM layer with a hidden size of 256 (32–512) nodes. A fully connected layer was attached to the output in order to map $h^{(t)}$ to $\hat{NDVI}^{(t)}$. We used a mini-batch size of 20 (10–100) and Adam optimizer (Kingma and Ba, 2014) with a learning rate of 0.001 (0.0001–0.1) and Mean Squared Error (MSE) as objective function. Early stopping was used as regularization to avoid over-fitting on the training data. The model was implemented in PyTorch v0.4 (Paszke et al., 2017).

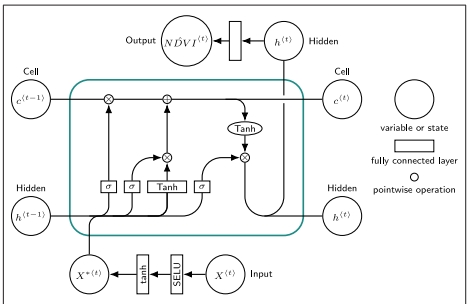


FIGURE 1 | The proposed model: At its core, it consists of a standard Long Short-Term Memory (LSTM) network. In addition, the input variables $X^{(t)}$ (a time-step t from the time-series concatenated with the static variables) are fed through two fully connected layers, yielding $X^{*(t)}$. $X^{*(t)}$ is concatenated to the hidden state of the last time step $h^{(t-1)}$ and then passed through the LSTM's internal layers. The cell $c^{(t)}$ bypasses the non-linear transformations to maintain long-term dependencies. The output $h^{(t)}$ is passed through a fully connected layer to map the LSTM's output to a single value, $\hat{NDVI}^{*(t)}$, the prediction for time-step t. Figure adapted from colah.github.io/posts/2015-08-Understanding-LSTMs.

2.4. Cross-Validation

To achieve a biased-reduced assessment of memory effects of climate variables on vegetation, we performed a k-fold cross-validation with spatial and temporal blocking. In a simple k-fold cross-validation, the data samples are randomly divided into k sets and each of them is used consecutively either for model training, validation or testing. Since most environmental variables are structured in space and time (Legendre, 1993), a random partitioning of the samples would possibly introduce a biased estimation of memory effects: Neglected covariates, as well as the model itself, often lead to residuals that are structured in space and time. The model can overfit the emerging residual dependency structure using predictor variables (Roberts et al., 2017) and as a consequence, we would overestimate memory effects of climate variables. Therefore, we performed a spatio-temporal cross-validation.

We subdivided the spatial and temporal domain into consecutive blocks and assigned all elements of a block to one of the cross-validation sets. The choice of the block

size is a trade-off between data limits, computational requirements and autocorrelation requirements (Roberts et al., 2017). Spatial blocking was done by subdividing the global raster into blocks of 5 × 5 pixels. Each 5 × 5 block was randomly assigned to one of 4 spatial folds. To account for temporal autocorrelation, the time-series were split into 4-folds of 9 years, overlapping by 1 year. The overlapping corresponds to the warmup period which is applied as the LSTM's state is initialized as zero and has to encode some of the time-series history first before becoming fully effective. The cross-validation scheme is illustrated in **Figure 2**.

Model training was done by iteratively using 2 spatial sets for training, 1 for validation and 1 for testing. For each of these combinations, 1 temporal block was used for validation and test while the other 3 were used for training. Note that we did not separate validation and test set in the temporal domain to not further reduce the sample size used for training, which is one of the above-mentioned trade-offs. As the model performance showed a low sensitivity toward the hyperparameters, we expect that this had a low impact on the results.

As the random assignment of the spatial blocks to the cross-validation sets may not be ideal (e.g., underrepresentation of some regions in the training set), anchor point of the spatial blocks and their assignment to the sets were varied randomly in 10 repetitions. For each of these repetitions, independent predictions for the test sets were retrieved. Each fold contained about 37% of the data for training (10,300,000 observations) and 6% for validation (1,650,000 observations). With the 4 folds from temporal, the 4 folds from spatial blocking and the 10 repetitions we ended up with 160 independent runs per model. We used the median of the 10 runs as final predictions.

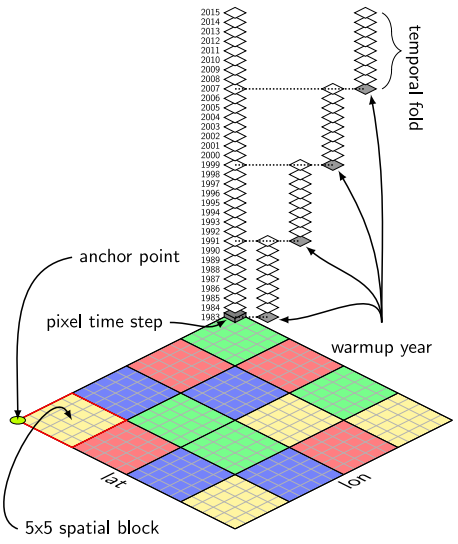


FIGURE 2 | Spatio-temporal cross-validation scheme: The 4 temporal folds consist of 9 years of 15 daily consecutive data, each overlapping by 1 year, the warmup period. While the temporal partitioning is fixed, the spatial blocking is random; consecutive blocks of 5 × 5 pixels are assigned to 1 of 4 spatial folds ( ). Each color represents one spatial fold. 2 of the 4 spatial folds are used for training, 1 for validating and 1 for testing. For a given setting (e.g., training: , validating: , testing:), 3 of the temporal folds are used for training and the remaining temporal fold is used for validation and testing. Both the spatial and temporal folds are iterated until each pixel time-step is predicted once (in the test set). The entire cross-validation is repeated 10 times with changing anchor point (such that the points covered by one 5 × 5 block are varying) and random assignment of the blocks to one of the spatial folds.

2.5. Model Evaluation

To assess the model's predictive performance, we used the Root Mean Squared Error (*RMSE*) and the R^2. We decomposed the raw time-series ($NDVI_{RAW}$) into the median seasonal cycle ($NDVI_{MSC}$) and the anomalies ($NDVI_{ANO}$). $NDVI_{MSC}$ was calculated pixel-wise as the median of the time-series across all years and $NDVI_{ANO}$ as the difference of $NDVI_{RAW}$ and $NDVI_{MSC}$. The decomposition was derived individually for the observations and the predictions. To quantify global model performance and memory effects, we used robust metrics based on R^2 and *RMSE*. First, we aggregated the observed and predicted time-series per hydro-climatic biome (b), as defined by Papagiannopoulou et al. (2018), by using the pixel-area weighted average (yielding R^2_b and $RMSE_b$). The biome-specific metrics were then aggregated to global level using the biome-area (A_b) weighted mean:

$$R^2_{global} = \frac{1}{A}\sum_{b=1}^{B} R^2_b * A_b$$

$$RMSE_{global} = \frac{1}{A}\sum_{b=1}^{B} RMSE_b * A_b$$

where A is the total area. This aggregation was done because $NDVI_{ANO}$ has a low signal-to-noise ratio compared to $NDVI_{MSC}$ and $NDVI_{RAW}$, which has two causes: First, $NDVI_{ANO}$ has a weaker signal (lower amplitude) than $NDVI_{MSC}$ and $NDVI_{RAW}$ in most cases. Second, $NDVI_{MSC}$ was calculated as the median over several years, which lowers the impact of noise while this is not the case for $NDVI_{RAW}$ and $NDVI_{ANO}$. In order to compare model performance among the different decompositions, we prefer a metric that corrects for this imbalance. R^2_{global} and $RMSE_{global}$ reflect how large-scale NDVI patterns are reproduced while keeping the impact of data noise low.

2.6. Identification of Memory Effects

To quantify memory effects, we trained multiple models with limited access to temporal context: During training, the dynamic features (climate variables) and the target (NDVI) time-series were permuted at each training step in unison,

keeping n antecedent elements in original order, referred to as model M_n (**Figure 3**). Validation and prediction were done on non-permuted time-series. We use the case $n = 1$ for illustration: Here, $NDVI_t$ is a function of $X = \{X_{t-1}, X_t\}$, which includes the instantaneous effect ($t \rightarrow t$) plus one past observation ($t-1 \rightarrow t$), hence memory of length $n = 1$, corresponding to 15 days. There are two special cases, the *full* memory model M_{full}, where no permutation is done and M_0, which is the *non-memory* model where the time-series are permuted randomly without blocking. To assess memory effects of different lengths, multiple models M_n with $n = \{full, 0, 1, 2, 3, 4, 5, 6\}$ corresponding to $\{full, 0, 15, 30, 45, 60, 75, 90\}$ days were computed. This choice was based on preliminary experiments, showing that the model performance was flattening after a lag of 90 days and the need to restrict the number of model runs. Note that although the permutation does destroy the order of the time-series before element $t - n$, the model can still learn from the distribution of the previous values.

We used the metric $Mem_n = R_n^2 - R_0^2$ to quantify memory effects, where n denotes the number of antecedent observations being included. *Mem* is the difference in R^2 between two models describing the impact of giving more temporal context on the model performance. For brevity, *Mem* refers to the total memory effects derived from M_{full} and M_0.

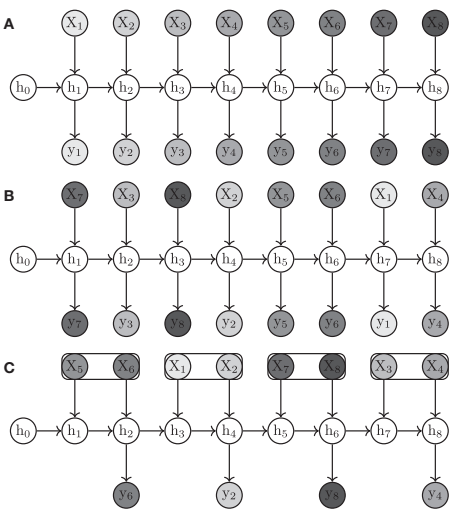


FIGURE 3 | Permutation scheme for M_{full} **(A)**, M_0 **(B)**, and M_1 **(C)**. For model M_n, permutation is done at each training step in order to restrict the memory n observations into past. The feature (X) and target (y) time-series are permuted in unison. h_t represents the hidden state of the model at time-step t. M_{full} learns the full memory effects as the time-series are not permuted, M_0 only learns instantaneous effects, the time-series are permuted randomly. For M_1, blocks of size 2 are permuted randomly and the first block starts at position randomly chosen from $\{0, \ldots, n\}$ to vary the elements covered by the blocks. During training, only the last element of each block is used in the loss calculation as n antecedent elements must be in original order, which is not the case for other elements.

To determine pixels of significant memory effects, we performed a permutation test. Our test statistic is the memory effect *Mem* and our null hypothesis was that *Mem* is equal to 0–meaning that on average, the models have the same performance. Each prediction can be labeled as coming from M_0 and M_{full}, and under the null hypothesis, they are exchangeable. For the permutation test, we permuted these labels 999 times (for all pixels simultaneously) and calculated each test statistic for each pixel at each permutation. The p-value is the proportion of test statistics that are as extreme as our observed test statistic. Since the permutation test was done on each pixel, we incurred in the multiple testing problem: As we perform thousands of simultaneous tests, it is more likely to observe significance just by chance. This was addressed by using the distribution of the maximum statistic to determine the threshold of significance at each pixel (Cortés et al. in preparation). At each permutation, we saved the maximum of the absolute value of the test statistic amongst all pixels, max(|*Mem*|). With the original data's maximum, these form the distribution of the maximum statistic. The threshold for significance at the pixel level was determined by the 90th percentile of this distribution.

3. RESULTS

3.1. Model Performance

First, we take a look at the global model performance of the full memory model M_{full} and the non-memory model M_0. Therefore, pooled—all pixels and time-steps combined—metrics *RMSE* and R^2 were calculated. M_{full} achieved an *RMSE* of 0.056 compared to model M_0 with an *RMSE* of 0.068. This is an error reduction of 14%. The R^2 increased by 2.8% from 0.916 to 0.943 from M_0 to M_{full}. As the global variability of NDVI is largely caused by spatial variability (68%), we also looked at the R^2 after removing the mean from each time-series. There, the improvement was 8.8% from 0.807 to 0.878.

The spatial variability of the model performance for M_{full} is illustrated in **Figure 4**. A high R^2 in terms of $NDVI_{RAW}$ and $NDVI_{MSC}$ is achieved in the northern temperate and boreal regions, eastern South America, as well as Savanna and Steppe ecosystems of Africa—regions of distinct seasonal NDVI signal. In contrast, rainforests and dry regions, where the seasonal cycle is less pronounced, show lower values of R^2, as errors take larger effects due to lower overall variance. For $NDVI_{ANO}$, R^2 is lower in general but achieves values between 0.25 and 0.4 in arid and semiarid regions. The *RMSE* of $NDVI_{RAW}$ and $NDVI_{MSC}$ is distributed more homogeneously, low values are found in arid regions due to the low vegetation signal.

3.2. Global Memory Effects

Global memory effects based on the aggregated R^2_{global} for $NDVI_{RAW}$, $NDVI_{MSC}$ and $NDVI_{ANO}$ are shown in **Table 1**. While M_{full} performs better in all cases, memory effects on $NDVI_{ANO}$ are stronger than on $NDVI_{MSC}$ in terms of absolute

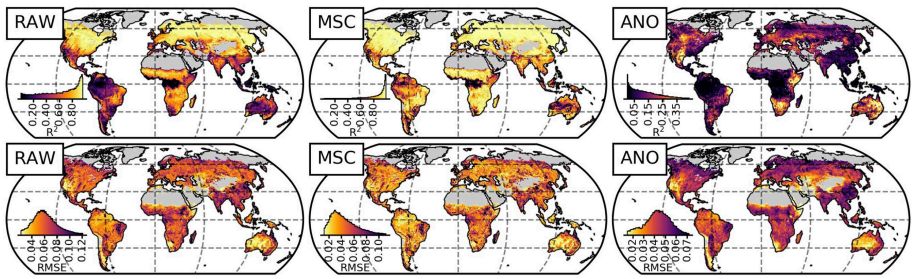


FIGURE 4 | Global model performance of the full memory model M_{full} based on NDVI time-series from 1984 to 2015 in terms of R^2 (top row) and *RMSE* (bottom row) for the raw time-series (RAW, left column), median seasonal cycle (MSC, center column) and anomalies (ANO, right column). The histograms are area-weighted.

TABLE 1 | Model performance of models M_{full} and M_0 for $NDVI_{RAW}$, $NDVI_{MSC}$, and $NDVI_{ANO}$.

		$NDVI_{RAW}$	$NDVI_{MSC}$	$NDVI_{ANO}$
P^2_{global}	M_0	0.848	0.881	0.323
	M_{full}	0.904	0.928	0.465
	% increase	6.3	6.3	30.6
	Mem	0.06	0.06	0.14
$RMSE_{global}$	M_0	0.025	0.017	0.018
	M_{full}	0.017	0.008	0.015
	% decrease	28.9	50.9	15.0

The metrics were calculated from area-weighted, per bioclimatic region aggregated time-series.

and relative increases of explained variance. Yet, note that for the seasonal cycle, the fraction of unexplained variance is halved from 12 to 7%, which is also reflected in the 50% decrease of the $RMSE_{global}$ in **Table 1**.

Figure 5 shows the spatial variability of memory effects. Significant effects were detected in transitional and sub-tropical biomes in general and—to a lower extent—mid-latitude water-driven climates, while the weak effects in temperate, boreal and rainforest climates were not significant on pixel basis. Accounting for antecedent climate conditions improves R^2 for $NDVI_{MSC}$ mainly in the tropical belt. However, these effects were not found to be significant. Finally, hotspots of significant memory effects for $NDVI_{ANO}$ are similar to those of $NDVI_{RAW}$, but more concentrated on arid and semiarid regions. Some areas show negative memory effects, especially in the case of $NDVI_{MSC}$. Note that a small number of pixels has negative correlations, which is not reflected by the R^2. However, these negative correlations are close to zero (not shown) and thus neglectable.

3.3. Biome-Specific Memory Effects

To understand how vegetation state is affected by antecedent climate under different environmental conditions, we take a look at biome-specific memory effects and how they change along climatic gradients.

First, we illustrate the predicted time-series for the different models with a regional example exhibiting strong memory effects (**Figure 6**): The Chobe National Park is located in Northern Botswana (∼ 19°S 24°E) and has a transitional water-driven climate with a distinct dry and wet season, the latter starting in October and ending in April. The selected area is—compared to its surroundings—only marginally affected by wildfires (see Fox et al., 2017 for further details). Both models, M_{full} and M_0 predict the overall patterns well, however, M_{full} performs considerably better. During low vegetation activity outside the raining season, the models perform equally with comparable variability of the error. In the rainy season when vegetation is active, the anomalies are stronger in general. Here, the full memory model M_{full} performs best, followed by M_1. The error variation of M_0 is larger during this period, whilst M_{full} errors have the lowest variation.

We further tested the impact of memory length on model performance on global as well as on biome level compared to baseline M_0 (**Figure 7**), based on the permutation approach. In general, the model performance is increasing with more temporal context in a saturating way. Even if the model performance is not strictly increasing in all cases with longer memory, a positive (asymptotic) relationship was found. Some biomes show a small drop in model performance with increasing memory length. We must keep in mind that the global MSE is minimized in model training. The different models may invest in reducing MSE in different regions as long as the global cost decreases, thus we only expect the global model performance to increase strictly, while regional discrepancies are expected. On global level, memory effects on $NDVI_{RAW}$, $NDVI_{MSC}$, and $NDVI_{ANO}$ are congruent. Transitional and sub-tropical biomes show strong yet highly variable memory effects on $NDVI_{MSC}$. Distinct memory effects on $NDVI_{ANO}$ are found in water-driven ecosystems.

Furthermore, we look at memory effects in the climate space of mean annual precipitation and temperature (**Figure 8**). For $NDVI_{RAW}$, we observe increasing memory effects with higher mean temperature, similar to $NDVI_{MSC}$. Below a threshold of around 14°C, memory effects are barely present. For $NDVI_{MSC}$, precipitation seems to play a minor role. In contrast, memory

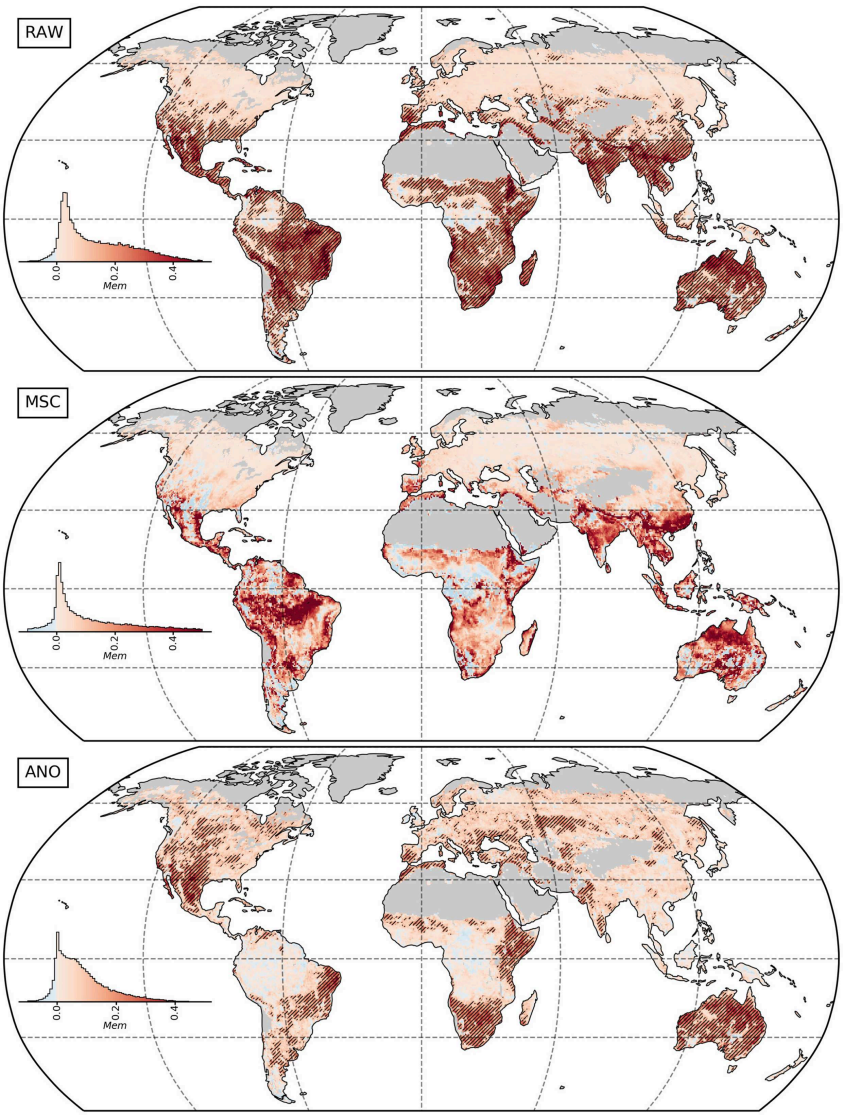


FIGURE 5 | Global memory effects based on NDVI time-series from 1984 to 2015 for the raw time-series (RAW, top), median seasonal cycle (MSC, center) and anomalies (ANO, bottom). Values represent the difference in R^2 between the full memory model (M_{full}) and the non-memory (M_0). Black striped areas ▨ indicate significant memory effects after accounting for the multiplicity ($\alpha = 0.1$). The histograms are area-weighted.

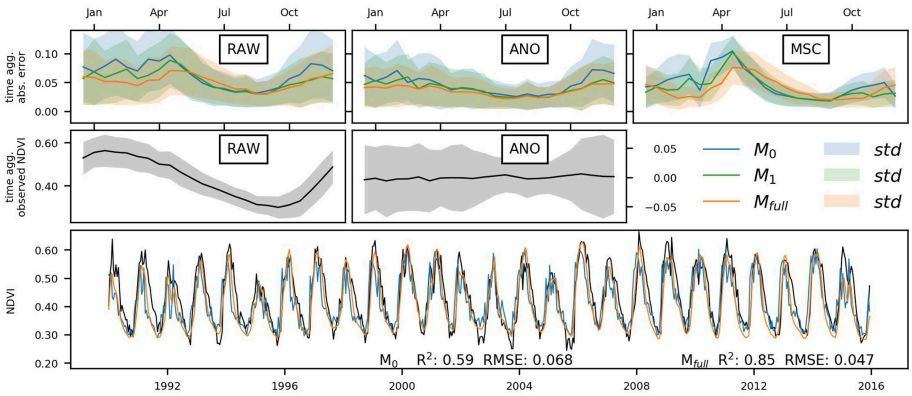


FIGURE 6 | Model predictions for a 10 × 5 pixel area located in the Chobe national park, Botswana (~ 19°S 24°E). The time-aggregated absolute error and its standard deviation over all pixels of M_0, M_1, M_{full} are shown in the top row, time aggregated observed NDVI in middle row (note that the MSC is contained in the RAW plot), and a subset of the observed and predicted NDVI time-series from 1990 to 2015 in the bottom row.

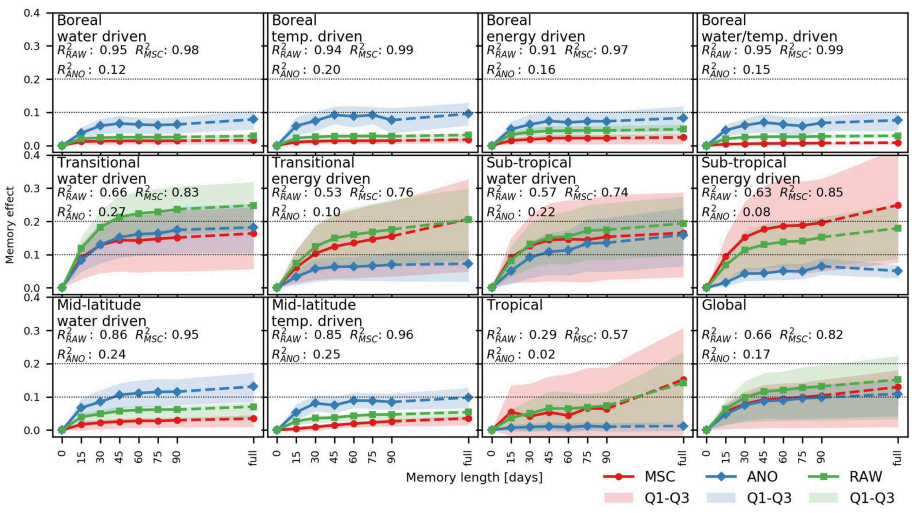


FIGURE 7 | Global and biome-specific mean and interquartile range (Q1–Q3) of memory effects (*Mem*) for $NDVI_{RAW}$ (mean removed), $NDVI_{MSC}$ (mean removed) and $NDVI_{ANO}$. The x axis represents the memory length, the number of days that are taken into account by the model. The shown R^2 reflect the performance of the full memory model (M_{full}). See Papagiannopoulou et al. (2018) for biome definition.

effects on $NDVI_{ANO}$ are higher with lower mean precipitation and higher temperatures. We see low memory effects above 700 mm annual precipitation and again, mean temperatures below 14°C.

Finally, we show the inter-biome mean and variation of memory effects per month separately for the Northern and Southern Hemisphere (**Figure 9**). In other words, this is the increase in explained variance across years per month from the

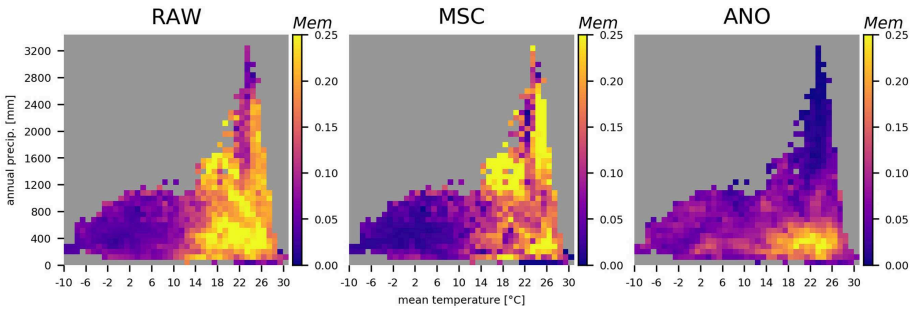


FIGURE 8 | Memory effects (*Mem*) in the climate space, annual precipitation vs. temperature for $NDVI_{RAW}$ (left), $NDVI_{MSC}$ (middle), and $NDVI_{ANO}$ (right). The spatial variability of $NDVI_{RAW}$ and $NDVI_{MSC}$ has been removed, i.e., the mean of each pixel's time-series has been subtracted prior to calculating *Mem*. Gray cells represent cases with <10 values.

non-memory model M_0 to the full memory M_{full} model. Note that we only display the results for $NDVI_{RAW}$, as $NDVI_{ANO}$ yields the same results and the approach is not applicable to $NDVI_{MSC}$. In boreal regions, the patterns are widely consistent, with small or no memory effects in winter, stronger effects in the start of the growing season and moderate effects at peak vegetation activity with a peak toward autumn. The transitional and sub-tropical water-driven biomes exhibit stronger memory effects in the Southern Hemisphere, with high values from December to May. The respective energy-driven regions show low memory effects in general. Furthermore, we see remarkable differences between the water and temperature-driven mid-latitudes: The water-driven regions show opposite patterns in Northern and Southern hemisphere, strongest memory effects occur in summer during the growing season. In temperature-driven regions, however, we see a distinct peak in the beginning of the growing season in spring and substantially lower memory effects during the remaining time of active vegetation. The topics, finally, show no memory effects of monthly variations.

4. DISCUSSION

4.1. Memory Effects on Vegetation State

We found memory effects on global scale with a bigger impact on the anomalies of vegetation state than on the seasonal cycle and generally lower impact on boreal and temperate climates and tropical rainforests. We detected large regional variations of memory effects and linked them to hydro-climatic biomes and climate gradients.

Our results shown in **Figure 7** suggest that sub-monthly, short-term memory effects play a dominant role while the impact of mid-term memory is weaker. For temperature and energy-driven ecosystems, lower memory effects on vegetation anomalies were found, which aligns with findings by Wu et al. (2015), Seddon et al. (2016), and Papagiannopoulou et al. (2017b). In water-driven regions, except for the boreal climates, we observe strong memory effects on vegetation anomalies, which was also found by the aforementioned studies.

We found evidence that ecosystems in colder climates with a mean temperature below 10–15°C are less affected by memory effects in general (**Figure 8**). Above this threshold, memory effects on the median seasonal cycle of vegetation state do not depend on mean annual precipitation, whereas effects on the anomalies are stronger with an annual rainfall below 700 mm. The strongest memory effects are found in sub-tropical and transitional ecosystems (**Figure 5**). The effect is similar for the seasonal cycle between energy and water-driven subregions, while the anomalies are much stronger affected by past climate in respective water-driven regions.

Sub-tropical water-driven regions—containing the arid and semiarid regions of the world (Papagiannopoulou et al., 2018)—are mainly shaped through patterns of precipitation, events that often occur in short pulses, followed by dry phases of varying length (Snyder and Tartowski, 2006). Through the limited water availability, vegetation dynamics in these regions largely depend on water storage in soils. Anomalies in soil moisture can last over several months (Koster et al., 2004), potentially leading to strong memory effects. While small precipitation pulses often cannot penetrate soil layers below 20–30 cm, clustered events in interaction with lower temperatures can refill deeper soil water storage. This resource can be accessed by deeper rooted plants, some even specialize on extracting water from different soil layers through the season (Schwinning and Ehleringer, 2001). This buffering of precipitation events in soil combined with large anomalies of precipitation can lead to strong memory effects, which is reflected in our results.

Similar patterns occur in transitional water-driven ecosystems, building the transition from arid and semiarid regions to humid climates. These ecosystems are still largely limited by water availability (Papagiannopoulou et al., 2018) and exhibit a higher vegetation density than sub-tropical regions. Arid and to a lower extent semiarid ecosystem are sparsely vegetated and thus, a generally low vegetation signal is observed.

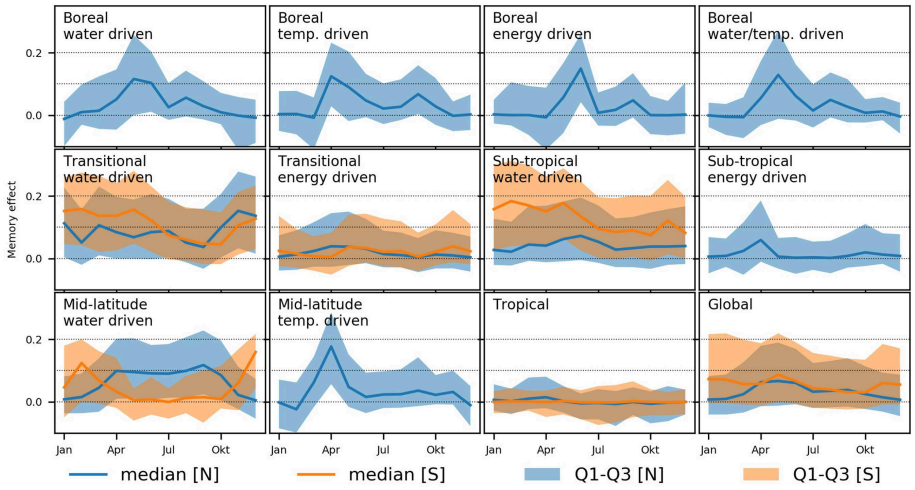


FIGURE 9 | Monthly biome-specific memory effects for the Northern [N] and Southern [S] Hemisphere. The metric reflects how much better the full memory model (M_{full}) explains the variance across years per month compared to the non-memory model (M_0). Cases with less that 10 pixels are not shown. See Papagiannopoulou et al. (2018) for biome definition.

The small variations in the NDVI and thus the low signal-to-noise ratio may mask effects that we try to identify. In arid and semiarid ecosystems of the Southern Hemisphere, the memory effects are occurring during active vegetation phase in the rainy season (**Figure 9**), which is also the case in the regional example shown in **Figure 6**. In the Northern Hemisphere, however, the link between precipitation and memory effect is less evident.

In the boreal and mid-latitude water-limited biomes, we see patterns of strong memory effects in spring (**Figure 9**). This is supposedly related to snowmelt and phenological effects of temperature. To determine when the snow cover disappeared or the top layer of the soil thawed, a certain amount of temporal context is needed, leading to relatively strong memory effects. Further, vegetation greenup timing in these ecosystems depends on the history of temperatures during previous months, which is often modeled as temperature sums in phenological models. In addition, some plants require chilling before warming effects can be effective (Migliavacca et al., 2012). Since the start of the growing season itself has a lagged impact on productivity after spring, e.g., as a consequence of more or less accumulated biomass, we see an impact of memory effects related to the spring vegetation dynamics lasting until around June. The length of memory effects (**Figure 7**) is similar for all boreal biomes with a maximum length of 15–30 days and stronger effects on vegetation anomalies than on the median seasonal cycle. This is counter-intuitive, as we would expect to see a strong dependency of the phenology on antecedent weather patterns due to the aforementioned cumulative temperature effects. However, the seasonal variations are well-predicted by both models ($R^2 > 0.95$), hence we see only small memory effects, even if a large fraction of the non-explained variance of the non-memory model is explained additionally by the full memory model. Moderate memory effects are observed in the remaining growing season, we expect that an increasing drought stress in boreal regions could alter the temporal dependencies in the future (Barichivich et al., 2014).

4.2. Time-Series Permutation Approach

An evaluation of the presented approach is challenging because there is no ground-truth of memory effects. However, we can assess the plausibility of the results in consideration of our understanding of ecosystem processes. We looked at biome-specific monthly memory effects and showed a regional example, where the full memory model performs best and a model with shorter memory length still performs better than the non-memory model. The differences in model performance were associated with periods of active vegetation, where predictions were better and more robust when including more memory. In contrast, dry seasons with barely any vegetation activity or winter periods in boreal regions are captured equally by all models. This suggests that the found memory effects are not just an artifact but are indeed linked to vegetation dynamics. Furthermore, we looked at the length of memory effects and found that models accounting for longer temporal context perform better. The found relationships between climate gradients

and memory effects align well with prior knowledge about ecosystem functioning.

Another way to evaluate the time-series permutation approach is a comparison with other studies. This turns out to be challenging as these studies (e.g., De Keersmaecker et al., 2015; Wu et al., 2015; Seddon et al., 2016; Papagiannopoulou et al., 2017b; Liu et al., 2018) use other predictor variables with different spatial and temporal resolution and different approaches (e.g., global vs. pixel-wise optimized). Due to some similarities in the study design and presentation of results, we can conduct a direct comparison to Wu et al. (2015): They employed a linear model with the lagged predictor variables temperature, precipitation and solar radiation to model monthly global NDVI on pixel basis. The authors used the regression coefficients to interpret drivers of and memory effect on vegetation state. Based on a visual inspection of the spatial model performance, our model (**Figure 4**) seems to perform better in terms of R^2, even if trained globally and spatio-temporal cross-validation was applied (see section 4.3.2 for further discussion). The found patterns of memory effects align in general, the same major hotspots are detected, yet our results indicate more wide-spread memory effects. It is possible that the regions we detected in addition are characterized by strong non-linear climate-vegetation interactions (Foley et al., 1998; Bonan, 2015; Papagiannopoulou et al., 2017a) and cannot be represented by a linear model as a consequence.

Papagiannopoulou et al. (2017a) (and the follow-up study Papagiannopoulou et al., 2017b) take a different approach based on a non-linear Granger causality framework: They quantified the model improvement from a model that uses past NDVI anomalies only compared to a model that uses climate variables in addition. The 4,571 (3,197 in the follow-up) climate variables include lagged and cumulative features and extreme indices. In a comparison, we must keep in mind that the reported "Granger causality on vegetation" may not be directly comparable to our memory effects metrics and that the temporal resolution of the time-series differ. While—based on a visual inspection—the main patterns of memory effects on the NDVI anomalies (Papagiannopoulou et al., 2017a) seem widely congruent with our findings (**Figure 4**, anomalies), the most striking difference are the significantly lower effects we found in the Sahel. Interestingly, this is also the region where the LSTM model performs worse than the pixel-wise trained random forest model. These discrepancies may attribute to the different resolutions of the time-series, or to the global vs. pixel-wise modeling approach (further discussed in section 4.3.2).

A drawback of the presented permutation approach is that we cannot attribute memory effects to single variables. Yet, we linked the strongest memory effects to water-limited ecosystems, which was also found by previous studies. We can conclude that, even though results are similar, we see regional differences, and that further development and discussion of the different approaches is needed.

Another way of identifying memory effects may be to apply the permutation approach after the training. In other words, the LSTM which has learned the dynamic effects in the data will be given a permuted time-series in the prediction. This resembles the permutation approach for studying variable importance with other machine learning approaches like random forests.

4.3. Advantages and Limitations

4.3.1. Data Limitations

Remote sensing data is inherently affected by errors related to data processing, the sensor, atmospheric effects and scene properties (Friedl et al., 2001). As a consequence, some regions—for example such with a complex topography—exhibit larger measurement errors, which affects the reliability of the results. Alike, the climatic reanalysis datasets used as predictor variables are affected by uncertainties linked to the underlying datasets and the modeling approach. A further limitation is the spatial and temporal resolution of the data. It is possible—yet not well-understood—that the temporal resolution (15 days) masks important short-term ecological processes that may propagate to longer temporal scales. Similarly, the spatial resolution of 0.5° integrates finer-grained local variations, leaving us with a smoothed signal.

Furthermore, the NDVI's dynamic range is limited since the signal saturates with dense vegetation. This poses an issue especially in dense forest areas like rainforests, where the NDVI shows little to no seasonality (Huete et al., 2006) and the anomalies mainly reflect noise. Thus, the results regarding rainforest areas should be taken with a grain of salt.

In addition, the model is limited by the choice of predictor variables: Ecosystem processes are highly complex and vegetation state depends on a vast number of factors, like nutrient availability (Fisher et al., 2012), human and natural disturbances (Reichstein et al., 2013; Trumbore et al., 2015), surface and sub-surface water flow (Koirala et al., 2017) and many others that are not included in the model. As a consequence, the interactions of the climate with those variables are neglected.

4.3.2. Global Modeling Approach

While previous studies looking into memory effects or related topics (e.g., De Keersmaecker et al., 2015; Wu et al., 2015; Seddon et al., 2016; Papagiannopoulou et al., 2017a,b; Liu et al., 2018) trained a model per pixel, we used a global modeling approach: A main objective of this study was to test the applicability of LSTMs to represent global vegetation dynamics. This choice was motivated by the great success of LSTMs in many other domains: LSTMs are dynamic models that are able to capture dependencies on multiple scales and—in theory—of unlimited length. LSTMs can be applied to raw time-series opposed to approaches that work on lagged and aggregated features (Lipton et al., 2015). This renders the approach fully data-driven, as no feature design choices are necessary. Furthermore, such a model can be easily extended in a modular fashion to include spatial context using Convolutional Neural Networks, for example. In this sense, the presented approach is generic. As such models can easily have thousands of parameters, they require large amounts of data to be trained. The length of satellite observation time-series (in our case ∼800 time-steps) is far away from being sufficient. With a global modeling approach, the dataset is much bigger and more adequate for a deep learning approach. Moreover, this approach achieves a unified global predictive model.

The global modeling approach was further motivated by the fact that the datasets are autocorrelated in space. We follow Roberts et al. (2017), who suggest that spatial cross-validation should be performed in all cases when dealing with environmental datasets. Especially for machine learning methods with high flexibility, overfitting is a problem that needs to be addressed. This choice, however, has a negative side-effect: The model's ability to adapt to local characteristics is limited and thus, some specificities cannot be learned. Rather, the model learns generalizable memory effects and therefore, the estimates of memory effects are conservative. In an effort to counter this issue, we included static variables that should help the model to implicitly link local differences to environmental conditions. In section 1 of the **Supplementary Material**, we showed that adding static variables improved model performance and made the predictions more robust. Furthermore, including these variables leads to a finer-grained picture of memory effects. This indicates that the global model learns specific local system behavior by linking it to actual local conditions rather than by "memorizing."

A further drawback of the global modeling scope is that the model—with the objective to reduce global loss—trades off different regions: To reduce the loss, the model may invest more of its capacity in better represented areas while neglecting under-represented regions. We expect that this is also the reason for the "negative" memory effects; from a theoretical point of view, knowing more about past environmental conditions cannot result in inferior predictions. We investigated this issue in section 2 in the **Supplementary Material**, where the globally trained model was compared to a model optimized for a single biome only. The memory effects and length were qualitatively similar. However, the geographic distribution of the memory effects on the median seasonal cycle showed substantial differences, while the patterns for the anomalies were more congruent. Thus, we recommend interpreting the memory effects regarding the median seasonal cycle with caution. This problem could be reduced by using higher resolution data and adding covariates that reflect these local variabilities better, e.g., human factors and additional soil properties.

4.4. Applications

RNNs are still rarely used to model Earth observation time-series. As shown here, RNNs are well-suited to model such data, as they are able to extract complex features from raw data with the benefit of rendering feature design unnecessary. Other than for diagnostic modeling, RNNs can also be used for upscaling of fluxes, gap-filling or benchmarking of physical models, for example. The time-series permutation approach presented here can easily be applied to other fields where a profound understanding of memory effects is pivotal, such as hydrology.

4.5. Conclusion

In this study, we have tested the applicability of an LSTM network to model Earth system variables using multivariate predictors. We used 33 years of climate variables together with static soil and land cover features to model 15 daily satellite based NDVI observations. The model was able to learn the global spatial and temporal variability of vegetation dynamics to a satisfying degree. This demonstrates the great capabilities of LSTMs, which are still rarely used in Earth system sciences, yet their potential is known from other disciplines.

Furthermore, we used a time-series permutation approach to identify memory effects of climate on vegetation state. Our results confirm findings from previous studies and highlight some new aspects of memory effects: While the geographic distribution widely agrees with other studies, we linked memory effects to climate gradients and took a closer look at their biome-specific temporal occurrence and length. The presented approach requires minimal prior knowledge of the domain and can be combined with powerful machine learning models. These properties render the approach into a useful tool that expands existing methods, possibly serving as a benchmark for approaches being able to do a more detailed analysis of variable contributions to memory effects.

DATA AVAILABILITY STATEMENT

All datasets analyzed for this study are included in the manuscript and the **Supplementary Files**.

AUTHOR CONTRIBUTIONS

BK conducted this study in the framework of his doctoral studies with the supervision of MR, MJ, and MK, who helped to conceive and plan the experiment, as well as discussing results on a regular base. BK performed the data processing, model setup and analysis, and wrote the manuscript with the help of MR, MJ, and MK. CR helped with regular critical comments and discussions and comments on the final draft. JC performed the test for significance for the global map of memory effects.

FUNDING

This research was funded by the International Max Planck Research School for Global Biogeochemical Cycles (IMPRSgBGC) and supported by the Max Planck Institute for Biogeochemistry.

ACKNOWLEDGMENTS

We want to thank Uli Weber from the MPI for Biogeochemistry for providing and preprocessing the datasets.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdata.2019.00031/full#supplementary-material

REFERENCES

Barichivich, J., Briffa, K. R., Myneni, R., Schrier, G. V. D., Dorigo, W., Tucker, C. J., et al. (2014). Temperature and snow-mediated moisture controls of summer photosynthetic activity in northern terrestrial ecosystems between 1982 and 2011. *Rem. Sens.* 6, 1390–1431. doi: 10.3390/rs6021390

Beck, H. E., Wood, E. F., Pan, M., Fisher, C. K., Miralles, D. G., van Dijk, A. I., et al. (2019). MSWEP V2 global 3-hourly 0.1° precipitation: methodology and quantitative assessment. *Bull. Am. Meteorol. Soc.* 100, 473–500. doi: 10.1175/BAMS-D-17-0138.1

Besnard, S., Carvalhais, N., Arain, M. A., Black, A., Brede, B., Buchmann, N., et al. (2019). Memory effects of climate and vegetation affecting net ecosystem CO_2 fluxes in global forests. *PLoS ONE* 14:e0211510. doi: 10.1371/journal.pone.0211510

Bonan, G. (2015). *Ecological Climatology: Concepts and Applications*. Cambridge: Cambridge University Press.

Chave, J. (2013). The problem of pattern and scale in ecology: what have we learned in 20 years? *Ecol. Lett.* 16, 4–16. doi: 10.1111/ele.12048

De Keersmaecker, W., Lhermitte, S., Tits, L., Honnay, O., Somers, B., and Coppin, P. (2015). A model quantifying global vegetation resistance and resilience to short-term climate anomalies and their relationship with vegetation cover. *Glob. Ecol. Biogeogr.* 24, 539–548. doi: 10.1111/geb.12279

Dee, D. P., Uppala, S. M., Simmons, A., Berrisford, P., Poli, P., Kobayashi, S., et al. (2011). The era-interim reanalysis: configuration and performance of the data assimilation system. *Q. J. Roy. Meteorol. Soc.* 137, 553–597. doi: 10.1002/qj.828

Fan, Y., Li, H., and Miguez-Macho, G. (2013). Global patterns of groundwater table depth. *Science* 339, 940–943. doi: 10.1126/science.1229881

FAO/IIASA/ISRIC/ISSCAS/JRC (2009). *Harmonized World Soil Database (Version 1.1)*. Rome; Laxenburg: FAO; IIASA.

Fisher, J. B., Badgley, G., and Blyth, E. (2012). Global nutrient limitation in terrestrial vegetation. *Glob. Biogeochem. Cycles* 26:GB3007. doi: 10.1029/2011GB004252

Foley, J. A., Levis, S., Prentice, I. C., Pollard, D., and Thompson, S. L. (1998). Coupling dynamic models of climate and vegetation. *Glob. Change Biol.* 4, 561–579. doi: 10.1046/j.1365-2486.1998.t01-1-00168.x

Fox, J. T., Vandewalle, M. E., and Alexander, K. A. (2017). Land cover change in northern botswana: the influence of climate, fire, and elephants on semi-arid savanna woodlands. *Land* 6:73. doi: 10.3390/land6040073

Frank, D., Reichstein, M., Bahn, M., Thonicke, K., Frank, D., Mahecha, M. D., et al. (2015). Effects of climate extremes on the terrestrial carbon cycle: concepts, processes and potential future impacts. *Glob. Change Biol.* 21, 2861–2880. doi: 10.1111/gcb.12916

Friedl, M. A., McGwire, K. C., and McIver, D. K. (2001). "An overview of uncertainty in optical remotely sensed data for ecological applications," in *Spatial Uncertainty in Ecology*, eds C. T. Hunsaker, M. F. Goodchild, M. A. Friedl, and T. J. Case (New York, NY: Springer), 258–283.

Friedl, M. A., Sulla-Menashe, D., Tan, B., Schneider, A., Ramankutty, N., Sibley, A., et al. (2010). Modis collection 5 global land cover: algorithm refinements and characterization of new datasets. *Rem. Sens. Environ.* 114, 168–182. doi: 10.1016/j.rse.2009.08.016

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. Cambridge, MA: MIT Press.

Graves, A., Mohamed, A.-R., and Hinton, G. (2013). "Speech recognition with deep recurrent neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Vancouver, BC: IEEE), 6645–6649.

Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., and Schmidhuber, J. (2017). LSTM: a search space odyssey. *IEEE Trans. Neural Netw. Learn. Syst.* 28, 2222–2232. doi: 10.1109/TNNLS.2016.2582924

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Huete, A., Running, S., and Myneni, R. (2006). "Monitoring rainforest dynamics in the amazon with modis land products," in *2006 IEEE International Symposium on Geoscience and Remote Sensing* (Denver, CO: IEEE), 263–265.

Kingma, D. P. and Ba, J. (2014). Adam: a method for stochastic optimization. *CoRR* abs/1412.6980.

Koirala, S., Jung, M., Reichstein, M., de Graaf, I. E., Camps-Valls, G., Ichii, K., et al. (2017). Global distribution of groundwater-vegetation spatial covariation. *Geophys. Res. Lett.* 44, 4134–4142. doi: 10.1002/2017GL072885

Koster, R. D., Dirmeyer, P. A., Guo, Z., Bonan, G., Chan, E., Cox, P., et al. (2004). Regions of strong coupling between soil moisture and precipitation. *Science* 305, 1138–1140. doi: 10.1126/science.1100217

Legendre, P. (1993). Spatial autocorrelation: trouble or new paradigm? *Ecology* 74, 1659–1673. doi: 10.2307/1939924

Lipton, Z. C., Berkowitz, J., and Elkan, C. (2015). A critical review of recurrent neural networks for sequence learning. *arXiv* 1506.00019.

Liu, L., Zhang, Y., Wu, S., Li, S., and Qin, D. (2018). Water memory effects and their impacts on global vegetation productivity and resilience. *Sci. Rep.* 8:2962. doi: 10.1038/s41598-018-21339-4

Marino, D. L., Amarasinghe, K., and Manic, M. (2016). "Building energy load forecasting using deep neural networks," In *42nd Annual Conference of the IEEE Industrial Electronics Society, IECON 2016* (Florence: IEEE), 7046–7051.

Migliavacca, M., Sonnentag, O., Keenan, T., Cescatti, A., O'keefe, J., and Richardson, A. (2012). On the uncertainty of phenological responses to climate change, and implications for a terrestrial biosphere model. *Biogeosciences* 9, 2063–2083. doi: 10.5194/bg-9-2063-2012

Monfreda, C., Ramankutty, N., and Foley, J. A. (2008). Farming the planet: 2. Geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000. *Glob. Biogeochem. Cycles* 22:GB1022. doi: 10.1029/2007GB002947

Ogle, K., Barber, J. J., Barron-Gafford, G. A., Bentley, L. P., Young, J. M., Huxman, T. E., et al. (2015). Quantifying ecological memory in plant and ecosystem processes. *Ecol. Lett.* 18, 221–235. doi: 10.1111/ele.12399

Papagiannopoulou, C., Gonzalez Miralles, D., Decubber, S., Demuzere, M., Verhoest, N., Dorigo, W. A., et al. (2017a). A non-linear granger-causality framework to investigate climate-vegetation dynamics. *Geosci. Model Dev.* 10, 1945–1960. doi: 10.5194/gmd-10-1945-2017

Papagiannopoulou, C., Gonzalez Miralles, D., Demuzere, M., Verhoest, N., and Waegeman, W. (2018). Global hydro-climatic biomes identified via multi-task learning. *Geosci. Model Dev.* 11, 4139–4153. doi: 10.5194/gmd-2018-92

Papagiannopoulou, C., Miralles, D. G., Dorigo, W. A., Verhoest, N. E. C., Depoorter, M., and Waegeman, W. (2017b). Vegetation anomalies caused by antecedent precipitation in most of the world. *Environ. Res. Lett.* 12:074016. doi: 10.1088/1748-9326/aa7145

Pascanu, R., Mikolov, T., and Bengio, Y. (2013). "On the difficulty of training recurrent neural networks," in *International Conference on Machine Learning* (Atlanta, GA), 1310–1318.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., et al. (2017). "Automatic differentiation in pytorch," in *Neural Information Processing Systems Workshop (NIPS-W)* (Long Beach, CA).

Pinzon, J. E. and Tucker, C. J. (2014). A non-stationary 1981–2012 avhrr ndvi3g time series. *Rem. Sens.* 6, 6929–6960. doi: 10.3390/rs6086929

Reichstein, M., Bahn, M., Ciais, P., Frank, D., Mahecha, M. D., Seneviratne, S. I., et al. (2013). Climate extremes and the carbon cycle. *Nature* 500:287. doi: 10.1038/nature12350

Reichstein, M., Besnard, S., Carvalhais, N., Gans, F., Jung, M., Kraft, B., et al. (2018). "Modelling landsurface time-series with recurrent neural nets," in *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium* (Valencia: IEEE), 7640–7643.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., et al. (2019). Deep learning and process understanding for data-driven earth system science. *Nature* 566:195. doi: 10.1038/s41586-019-0912-1

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* 40, 913–929. doi: 10.1111/ecog.02881

Rußwurm, M., and Körner, M. (2017). "Temporal vegetation modelling using long short-term memory networks for crop identification from medium-resolution multi-spectral satellite images," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (Honolulu, HI), 1496–1504.

Rußwurm, M., and Körner, M. (2018). Multi-temporal land cover classification with sequential recurrent encoders. *ISPRS Int. J. Geo Inform.* 7:129.

Schwinning, S. and Ehleringer, J. R. (2001). Water use trade-offs and optimal adaptations to pulse-driven arid ecosystems. *J. Ecol.* 89, 464–480. doi: 10.1046/j.1365-2745.2001.00576.x

Seddon, A. W., Macias-Fauria, M., Long, P. R., Benz, D., and Willis, K. J. (2016). Sensitivity of global terrestrial ecosystems to climate variability. *Nature* 531:229. doi: 10.1038/nature16986

Snyder, K. and Tartowski, S. (2006). Multi-scale temporal variation in water availability: implications for vegetation dynamics in arid and semi-arid ecosystems. *J. Arid Environ.* 65, 219–234. doi: 10.1016/j.jaridenv.2005.06.023

Trumbore, S., Brando, P., and Hartmann, H. (2015). Forest health and global change. *Science* 349, 814–818. doi: 10.1126/science.aac6759

Tucker, C. J. (1979). Red and photographic infrared linear combinations for monitoring vegetation. *Rem. Sens. Environ.* 8, 127–150. doi: 10.1016/0034-4257(79)90013-0

Werbos, P. J. (1990). Backpropagation through time: what it does and how to do it. *Proc. IEEE* 78, 1550–1560. doi: 10.1109/5.58337

Wu, D., Zhao, X., Liang, S., Zhou, T., Huang, K., Tang, B., et al. (2015). Time-lag effects of global vegetation responses to climate change. *Glob. Change Biol.* 21, 3520–3531. doi: 10.1111/gcb.12945

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

4. Hybrid modeling: combining data- and knowledge-driven approaches

This section is based on

B. Kraft, M. Jung, M. Körner, and M. Reichstein (2020). "Hybrid Modeling: Fusion of a Deep Learning Approach and a Physics-Based Model for Global Hydrological Modeling." In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. Vol. XLIII-B2-2020. Copernicus GmbH, pp. 1537–1544. DOI: 10.5194/isprs-archives-XLIII-B2-2020-1537-2020

and

B. Kraft, M. Jung, M. Körner, S. Koirala, and M. Reichstein (2022). "Towards hybrid modeling of the global hydrological cycle." In: *Hydrology and Earth System Sciences* 26.6, pp. 1579–1614. DOI: 10.5194/hess-26-1579-2022

4.1. Study summary

The following sections demonstrate the potential of global-scale hybrid modeling of the hydrological cycle. The basic concept is outlined in the first study (Kraft et al., 2020) and further developed and assessed in the second study (Kraft et al., 2022). By combining statistical modeling with physical knowledge, we developed a partially interpretable hybrid model that allows insights into the global water cycle. On the global level, the hybrid model performed on par with a set of physically-based models and achieved better local adaptivity. The improved adaptivity is a key strength of the hybrid approach and is enabled by the data-adaptivity of the RNN.

For the first time, a data-driven yet physically consistent partitioning of water storage components was achieved. The partitioning agreed with physically-based patterns, especially in regions where

[1] https://creativecommons.org/licenses/by/4.0/

the hydrological processes are better understood and more certain. In other regions, especially the transitional zones, the hybrid model diagnosed larger soil moisture and a lower groundwater variability compared to the physically-based models.

The successful implementation of a global-scale hybrid model gives rise to a number of applications, which are discussed in Chapter 5.

Contribution The idea for both studies was developed together with the co-authors in the framework of my doctoral studies. The success of the project was highly uncertain as similar approaches do not exist yet, and thus, expertise from all the co-authors was required. While the conceptual development of the approach was a team effort, I implemented the model and conducted the analysis of the result. During the model development, I acquired unique conceptual and technical expertise in the field of hybrid modeling. I took the lead in manuscript writing, but the co-authors contributed significantly, especially in the interpretation of the hydrological simulations.

4.2. Hybrid modeling: fusion of a deep learning approach and a physics-based model for global hydrological modeling

Please turn to the next page.

The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLIII-B2-2020, 2020
XXIV ISPRS Congress (2020 edition)

HYBRID MODELING: FUSION OF A DEEP LEARNING APPROACH AND A PHYSICS-BASED MODEL FOR GLOBAL HYDROLOGICAL MODELING

B. Kraft[1,2,*], M. Jung[1], M. Körner [2], M. Reichstein[1]

[1] Department of Biogeochemical Integration, MPI for Biogeochemistry, Jena, Germany
(bkraft, mjung, mreichstein)@bgc-jena.mpg.de
[2] Department of Aerospace and Geodesy, Technical University of Munich, Munich, Germany
marco.koerner@tum.de

KEY WORDS: Hybrid Modeling, Deep Learning, Hydrology, Global Modeling, LSTM

ABSTRACT:

Process-based models of complex environmental systems incorporate expert knowledge which is often incomplete and uncertain. With the growing amount of Earth observation data and advances in machine learning, a new paradigm is promising to synergize the advantages of deep learning in terms of data adaptiveness and performance for poorly understood processes with the advantages of process-based modeling in terms of interpretability and theoretical foundations: *hybrid modeling*. Here, we present such an end-to-end hybrid modeling approach that learns and predicts spatial-temporal variations of observed and unobserved (latent) hydrological variables globally. The model combines a dynamic neural network and a conceptual water balance model, constrained by the water cycle observational products of evapotranspiration, runoff, snow-water equivalent, and terrestrial water storage variations. We show that the model reproduces observed water cycle variations very well and that the emergent relations of runoff-generating processes are qualitatively consistent with our understanding. The presented model is—to our knowledge—the first of its kind and may contribute new insights about the dynamics of the global hydrological system.

1. INTRODUCTION

Process-based models of the Earth and its subsystems have been key to diagnose, predict, and understand environmental processes and change for decades. Such models are based on conceptualizations and abstractions of many individual processes according to expert understanding. They are forced, evaluated, and occasionally tuned using environmental observations. The rapidly growing amount of Earth observation data, however, does not necessarily translate into better process models, as process representations are predefined rather than learned from data. Due to advances in machine learning, complex patterns and relationships in multivariate datasets can now be recognized with high accuracy and further exploited. These models typically need large amounts of training data, while they are agnostic to the physical meaning and consistency among variables. It is, thus, promising to explore a synergistic combination of machine learning and process-based approaches for modeling in Earth system sciences (Reichstein et al., 2019). The hybrid approach is still in its infancy and we are aware of one application on Earth observation data only: de Bézenac et al. (2019) predicted future sea-surface temperature fields by using a convolutional encoder-decoder network to learn a motion field that was fed into a physical model of advection and diffusion.

We present an end-to-end global hybrid hydrological model that couples long short-term memory (LSTM, Hochreiter and Schmidhuber, 1997) networks with a traditional conceptual water balance model that is trained jointly on a set of water cycle observations: total water storage (TWS), runoff (Q), evapotranspiration (ET), and snow water equivalent (SWE). The model is forced by the meteorological variables precipitation, air temperature, and net radiation. From a deep learning perspective, the hybrid approach can be seen as a regularization of the neural network, constraining the solution space to physically plausible results. Furthermore, the hydrological states (pools) and fluxes (inflows and outflows) of the conceptual water balance model remain interpretable and are still largely data-driven, as they are informed by the neural network.

In this study, we provide a proof-of-concept and test the applicability of hybrid modeling to learn a representation of the global water cycle from data. We explore the robustness of the approach based on independent cross-validations which include the full training set-up.

2. GLOBAL DATASETS

2.1 Total Water Storage Anomalies (TWS)

The Gravity Recovery & Climate Experiment (GRACE) Mascon Equivalent Water Height RL06 with Coastal Resolution Improvement (CRI) v1 (Watkins et al., 2015; Wiese et al., 2016; Wiese et al., 2018) represents variations in global water storages, *i.e.*, groundwater, soil moisture, surface water, snow, and ice for land pixels. The product has a native spatial resolution of 3° but is delivered at 0.5°. For this study, all time series datasets were aggregated to 1° resolution, but still, the TWS data may not represent local grid-scale variabilities properly. The TWS data is available from April 2002 to June 2016 covering irregular, roughly monthly periods. As we observed some outliers in the dataset, observations $-500 > tws > 500$ were removed.

2.2 Evapotranspiration (ET)

Monthly ET data was retrieved from the global FLUXCOM-RS product (Jung et al., 2019; Tramontana et al., 2016), which is based on upscaling of FLUXNET (Baldocchi et al., 2001) eddy covariance data. The upscaling is achieved using an ensemble

* Corresponding author

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLIII-B2-2020-1537-2020 | © Authors 2020. CC BY 4.0 License. 1537

The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLIII-B2-2020, 2020
XXIV ISPRS Congress (2020 edition)

of machine learning models, each learning a mapping from remote sensing (RS) observations to the site-level fluxes, which can then be upscaled to global scale. The ET was derived from the latent energy estimates, assuming a constant latent heat of vaporization of $2.45\,\mathrm{MJ\,mm^{-1}}$.

2.3 Total Runoff (Q)

GRUN v1 is a global gridded dataset providing estimates of monthly total runoff with a native spatial resolution of 0.5° (Ghiggi et al., 2019). The authors used random forests to model local discharge observations from small catchments as a function of climate data and generalized the learned relationships to retrieve global estimates.

2.4 Snow Water Equivalent (SWE)

Daily SWE was retrieved from GlobSnow v2 (Luojus et al., 2014; Takala et al., 2011) and aggregated from 0.25° to 1° spatial resolution. The product only covers the Northern Hemisphere and pixel time-steps with no snow are encoded as missing values. As the absence snow is important information that we do not want to discard, the SWE product was enriched using $8\,\mathrm{d}$ MODIS snow cover fractions (SCF) disaggregated to daily using nearest neighbor (Hall and Riggs, 2016). SWE with missing data were set to 0 if: a) more than 24 consecutive days were missing for SWE and b) the mean SCF over ± 12 days was below 10 %. This gap-filling mainly assigned zero SWE to previously missing values in the Southern Hemisphere and Northern Summer. Note that some mountainous regions were masked out in the GlobSnow product. The SWE signal is known to saturate at 100–150mm (Larue et al., 2017).

2.5 Meteorological Forcing

As time-varying model inputs, we used three meteorological forcing datasets, each on daily resolution: Net radiation is obtained from the SYN1deg Ed3A product (Doelling, 2017) of the Clouds and the Earth's Radiant Energy Systems (CERES) program (Wielicki et al., 1996). The precipitation data was retrieved from the Global Precipitation Climatology Project daily 1° dataset (GPCP-1DD) v1.2 (Huffman et al., 2012). Air temperature was obtained from the CRUNCEP v8 dataset, a combined product of the observation-based Climate Research Unit (CRU) and the National Center for Environmental Prediction (NCEP) reanalysis data (Harris et al., 2014; Viovy, 2018).

2.6 Static Datasets

A number of static datasets were used to represent the spatial variability of surface and subsurface environmental conditions. To represent topography, we used the digital elevation model from GTOPO30 (DOI/USGS/EROS, 1997). Furthermore, we used variables from the soilgrids dataset (Hengl et al., 2017): absolute depth to bedrock and the average across all soil layers of bulk density, coarse fragments, clay, silt, and sand content. Land cover fractions were derived from the Globland30 dataset (Chen et al., 2015) for the classes water bodies, wetlands, artificial surfaces, tundra, permanent snow and ice, grasslands, barren, cultivated land, shrublands, and forests. In addition, a wetland dataset was used that contains fractions of groundwater-driven wetlands, regularly flooded wetlands, and the intersection of the them (Tootchi et al., 2019).

These 22 variables were aggregated from their mostly finer native spatial resolution to $\frac{1}{30}^\circ$ to keep information on the spatial variability inside a 1° model pixel. To reduce the size of the stacks (30 (lat. pixels) $\times$ 30 (lon. pixels) $\times$ 22 (variables) $=$ 19 800 values per model cell) and ultimately the number of parameters in the model, we reduced the dimensionality of the static variables in a pre-processing step. A simple convolutional autoencoder was used for this, consisting of an encoder network, a bottleneck layer, and a decoder network. The encoder layers extract features from the input stack with consecutively smaller capacity. The final representation is the bottleneck layer, with a vector of size 30. The decoder, which has the reverse structure of the encoder network, maps the bottleneck layer back to the input stack. By minimizing the reconstruction loss, the model is forced to find a low-dimensional representation of the stack.

2.7 Masking & Bioclimatic Regions

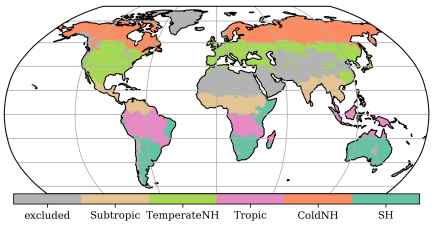


Figure 1. The masked out cells ('excluded') and the bioclimatic regions used for model evaluation: Cold Northern Hemisphere ('ColdNH'), Temperate Northern Hemisphere ('TemperateNH'), 'Tropic', 'Subtropic' and remaining Southern Hemisphere regions ('SH').

To retrieve valid land pixels with a clear signal of TWS, ET, and Q, cells with more than 50 % water bodies, 10 % permanent snow or ice, 10 % artificial surfaces, and 10 % bare land were removed. Further, regions with strong anthropogenic groundwater withdrawal were discarded, as the model does not account for these effects. After applying these criteria, the dataset consisted of 11 026 spatial samples. Note that some grid-cells were masked out further due to missing values in the SWE dataset, *e.g.*, some mountainous areas. The excluded cells are shown in Figure 1 along with five bioclimatic regions used in the model evaluation.

3. GLOBAL HYBRID HYDROLOGICAL MODELING

3.1 The Hybrid Hydrological Model

The hybrid model represents the major states and fluxes of the hydrological cycle (see Box 1). The model learns a mapping from the meteorological features (X) to the target variables (y). To predict y_t at time t, it has access to the present and past observations $X_{\leq t}$ and a set of static variables.

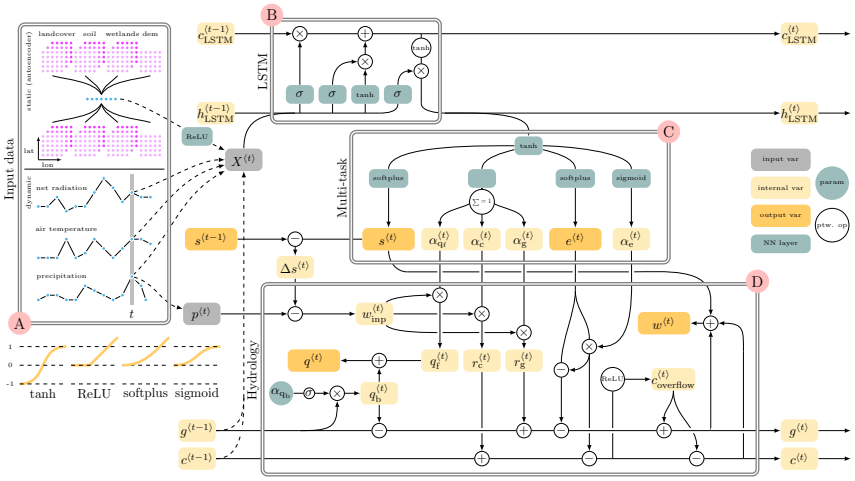
3.2 Self-Paced Multi-Task Learing

To combine the four loss terms corresponding to the target variables, we used self-paced task uncertainty weighing (Kendall et al., 2018), as done in state-of-the-art multi-task learning (*e.g.* Liebel and Körner, 2018). By optimizing an uncertainty term σ for each task (Equation 1), the different uncertainties inherent to the target variables are compensated dynamically.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLIII-B2-2020-1537-2020 | © Authors 2020. CC BY 4.0 License.

The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLIII-B2-2020, 2020
XXIV ISPRS Congress (2020 edition)

Box 1: The end-to-end hybrid hydrological model



A Input data
The meteorological time series (Section 2.5), encoded static variables (Section 2.6) and physically interpretable states groundwater (GW, g)[a] and cumulative water deficit (CWD, c) are fed into the LSTM.

B The LSTM layer
The LSTM updates the hidden states $h^{\langle t \rangle}_{\text{LSTM}}$ and $c^{\langle t \rangle}_{\text{LSTM}}$ at each time-step.

$$h^{\langle t \rangle}_{\text{LSTM}}, c^{\langle t \rangle}_{\text{LSTM}} = \text{LSTM}(h^{\langle t-1 \rangle}_{\text{LSTM}}, c^{\langle t-1 \rangle}_{\text{LSTM}}, X^{\langle t \rangle})$$

C Multi-task layer
The multi-task layer, comprising of independent feed-forward layers (NN), yields interpretable variables: evapotranspiration (ET, e), snow water equivalent (SWE, s), and fractions (α) defining how the liquid water input (w_{inp}) is partitioned into the fluxes of fast runoff ($w_{\text{inp}} \cdot \alpha_{\text{q}_\text{f}} \rightarrow q_\text{f}$), soil recharge ($w_{\text{inp}} \cdot \alpha_\text{c} \rightarrow r_\text{c}$), and groundwater recharge ($w_{\text{inp}} \cdot \alpha_\text{g} \rightarrow r_\text{g}$). The current w_{inp} is the precipitation (p) minus snow accumulation or plus snow melt (Δs). I addition, a fraction α_e determines the source pool from which e is taken from. If α_e=1, e is taken from the soil, if α_e=0, e is taken from the groundwater.

$$\begin{aligned}
e^{\langle t \rangle} &= \text{softplus}(\text{NN}(h^{\langle t \rangle}_{\text{LSTM}})) \\
s^{\langle t \rangle} &= \text{softplus}(\text{NN}(h^{\langle t \rangle}_{\text{LSTM}})) \\
\alpha^{\langle t \rangle}_{\text{q}_\text{f}}, \alpha^{\langle t \rangle}_\text{c}, \alpha^{\langle t \rangle}_\text{g} &\overset{\Sigma=1}{=} \text{softplus}(\text{NN}(h^{\langle t \rangle}_{\text{LSTM}})) \\
\alpha^{\langle t \rangle}_\text{e} &= \text{sigmoid}(\text{NN}(h^{\langle t \rangle}_{\text{LSTM}}))
\end{aligned}$$

D Water balance model
The hydrological block implements water balance equations. The physical state variables g and c are updated at each time-step using a combination of the above latent variables and variables derived here. When $c = 0$, the soil is fully water-saturated, negative values indicate a water deficit. If $c > 0$, the soil capacity is exceeded and overflow occurs (c_{overflow}). Note that for the model evaluation, c is transformed such that a deficit is denoted by positive values. The base runoff (Q_b, q_b) is defined as g times a learned global fraction $\alpha_{\text{q}_\text{b}}$. The total runoff (Q, q) is the sum of q_b and q_f. The total water storage (TWS, w) anomalies are calculated as the sum of s, g, and c, minus the mean of w to get the variation around 0. The units are mm for state variables and $\text{mm}\,\text{d}^{-1}$ for fluxes.

$$c^{\langle t \rangle} = c^{\langle t-1 \rangle} + \overbrace{\alpha^{\langle t \rangle}_\text{c}(p^{\langle t \rangle} - \Delta s^{\langle t \rangle})}^{r^{\langle t \rangle}_\text{c}} - e^{\langle t \rangle}\alpha^{\langle t \rangle}_\text{e}$$

$$c^{\langle t \rangle} = c^{\langle t \rangle} - \overbrace{\max(c^{\langle t \rangle}, 0)}^{c^{\langle t \rangle}_{\text{overflow}}}$$

$$q^{\langle t \rangle} = \overbrace{\alpha^{\langle t \rangle}_{\text{q}_\text{f}}(p^{\langle t \rangle} - \Delta s^{\langle t \rangle})}^{q^{\langle t \rangle}_\text{f}} + \overbrace{g^{\langle t-1 \rangle}\,\text{sigmoid}(\alpha_{\text{q}_\text{b}}) \cdot 0.01}^{q^{\langle t \rangle}_\text{b}}$$

$$g^{\langle t \rangle} = g^{\langle t-1 \rangle} - q^{\langle t \rangle}_\text{b} + \overbrace{\alpha^{\langle t \rangle}_\text{g}(p^{\langle t \rangle} - \Delta s^{\langle t \rangle})}^{r^{\langle t \rangle}_\text{g}} + c^{\langle t \rangle}_{\text{overflow}} - e^{\langle t \rangle}(1 - \alpha^{\langle t \rangle}_\text{e})$$

$$w^{\langle t \rangle} = s^{\langle t \rangle} + g^{\langle t \rangle} + c^{\langle t \rangle}$$

[a] (acronym, math. symbol)

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLIII-B2-2020-1537-2020 | © Authors 2020. CC BY 4.0 License.

The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLIII-B2-2020, 2020
XXIV ISPRS Congress (2020 edition)

$$\mathcal{L} = \sum_{i}^{n} \frac{1}{2 \cdot \sigma_i^2} \mathcal{L}_i + \log(\sigma_i) = \sum_{i}^{n} w_i \, \mathcal{L}_i + r_i \qquad (1)$$

where w_i is a weight for the task i of n total tasks, reciprocal to the task uncertainty σ_i and r_i is a regularization term to prevent the uncertainty from converging to infinity. In practice, the uncertainty is encoded as $s := \log(\sigma^2)$ to assert numerical stability and to have an unbound parameter s. Hence, $w = 0.5 \cdot \exp(-s)$ and $r = 0.5 \cdot s$.

We added a further constraints ($\mathcal{C}_\text{g}$) to penalize negative values for groundwater (GW). In preliminary experiments, we observed that the model can easily reach a loss $\mathcal{C}_\text{g} = 0$, and, thus, s converged to minus infinity. To prevent this, a constant of 0.1 was added: $\mathcal{C}_\text{g} = \text{mean}(-\min(\boldsymbol{g}, 0)) + 0.1$, where $\boldsymbol{g}$ is a simulated groundwater time series.

3.3 Model Selection & Training

The model was trained end-to-end and simultaneously on global observation-based products of TWS, SWE, ET, and Q using the backpropagation algorithm (Goodfellow et al., 2016). We used the root mean square error (RMSE) as the objective function. The model was implemented in PyTorch v1.4 (Paszke et al., 2017).

The time series were split into two periods, 2002-01 to 2008-12 for training and 2009-01 to 2014-12 for validation and testing. The feature time series were extended by selecting ten random years from the features of the respective periods for model spin-up to obtain steady physical model states (GW and soil cumulative water deficit (CWD)), before the actual evaluation period. Furthermore, a warmup period of one year was added to both time-ranges to have some temporal context even for the start of the periods. In addition, the samples were split into mutually exclusive regular grids for the hyperparameter (HP) optimization and the cross-validation (Fig. 2). These measures were taken to reduce overfitting due to spatial and temporal autocorrelation (Roberts et al., 2017).

For the model selection, we used the Bayesian optimization hyper-band (BOHB) algorithm (Falkner et al., 2018) from the *Ray.tune* framework (Liaw et al., 2018). BOHB is a state-of-the-art method for HP optimization that combines an early stopping mechanism (dropping non-promising runs) and a Bayesian surrogate model that suggests new HPs. Here, we used samples from one of the four spatial grids. To match the cross-validation scheme, the samples were split into five folds, of which three were used for training and one for validation. The final HPs are reported in Table 1. The remaining three grids were used to perform three independent cross-validations: in each, one fold was withheld for testing ($5\,\%$ of the grid-cells) and the remaining four folds ($20\,\%$ of the grid-cells) were iterated such that each fold was used for validation once. The test set predictions used for the model evaluation are referred to as $cv_{i,f}$, where $i \in \{1, 2, 3\}$ is the cross-validation and $f \in \{1, 2, 3, 4\}$ is the fold index.

3.4 Model Evaluation

First, the model fit was quantified regarding the temporal patterns aggregated by the bioclimatic regions (Figure 1) using the Pearson correlation coefficient (r) and the Nash–Sutcliffe model efficiency coefficient (NSE). The NSE ranges from $-\infty$ to 1, a negative NSE indicates that the model fit is worse than just taking the observed mean as prediction, 1 is a perfect fit (Nash and Sutcliffe, 1970). The evaluation was performed based on the test sets which have not been used for HP optimization or model training. From the three cross-validations, only one of the four runs was used and combined into one unified dataset, *i.e.*, $cv_{i\in\{1,2,3\},f=1}$. Then, we aggregated the time series per bioclimatic regions using the mean of all respective grid-cells. We then calculated r and NSE for each bioclimatic region.

Then, the robustness of the simulated latent variables was assessed. As the proposed hybrid model has a high degree of freedom compared to conceptual models, it is crucial to check if repeated runs lead to similar results. Robust model predictions increase the trust in the latent variable estimates. The robustness of the model was assessed using the simulations from the cross-validation. In addition, we assess the plausibility of the non-observed (latent) estimates based on our process understanding. For the evaluation of the latent variables, we cannot

Model architecture			
layer	num. layers	hidden size	dropout
static encoding	2 (1, 2)	100 (50, 100)	0.2 (0.0, 0.5)
LSTM	1 (−)	100 (50, 200)	−
task branches	1 (1, 3)	100 (50, 200)	0.2 (0.0, 0.5)
Optimizer parameters			
learning rate	10^{-2} (10^{-2}, 10^{-4})		
task weight learning rate	10^{-2} (10^{-2}, 10^{-4})		
weight decay	10^{-5} (10^{-2}, 10^{-5})		
grad. clipping	0.6 (0.1, 1)		

Table 1. Model architecture and optimizer hyperparameters with range limits searched in brackets (lower, upper). The static encoding layer extracts features of the static input which are fed into the LSTM together with the meteorological forcing time series. The single-layer LSTM is followed by multiple task branches. The learning rate defines the step size of the optimizer (with an independent learning rate for the task weights), weight decay adds L2 regularization (preventing large parameter values) and gradient clipping counteracts exploding gradients.

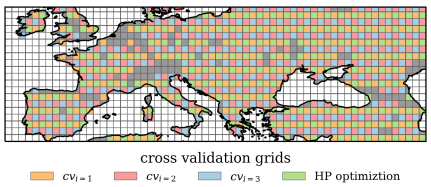


Figure 2. Regional example of the data splitting for the hyperparameter tuning and cross-validation. The grid-cells are split into four mutually exclusive, regular grids (colored). The grid-cells of each set are separated by a buffer to reduce the spatial autocorrelation between the samples. The samples of each grid were then split randomly into 5 sets of which one was used for testing and the remaining four were iterated such that each set was used as validation set once. One of the four grids was used for hyperparameter optimization. Following this scheme, three separate cross-validations ($cv_{i\in\{1,2,3\}}$) are performed, each yielding four predictions on the test set. Note that some grid-cells are masked out (grey), see Section 2.7 for more details.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLIII-B2-2020-1537-2020 | © Authors 2020. CC BY 4.0 License.

The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLIII-B2-2020, 2020
XXIV ISPRS Congress (2020 edition)

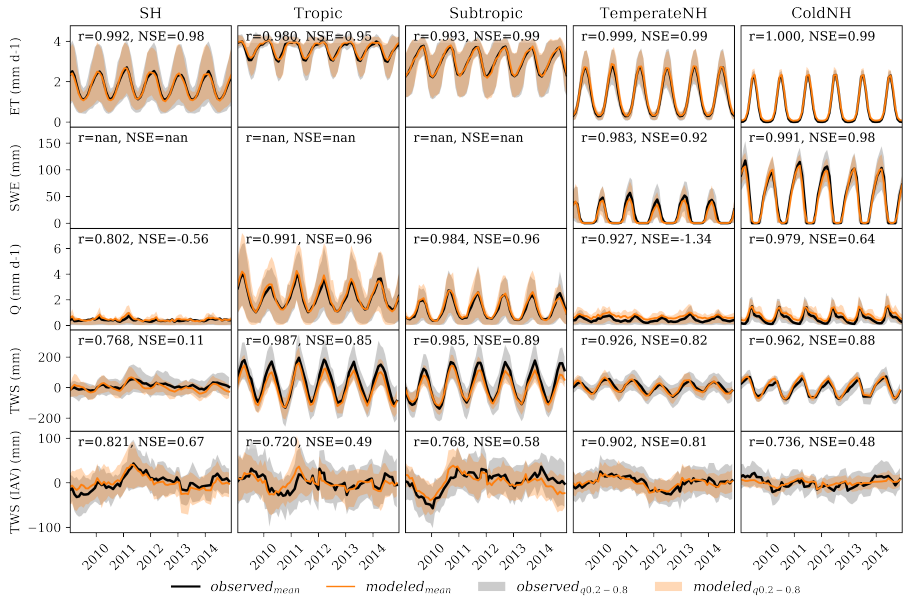


Figure 3. The model performance based on the test set by bioclimatic regions. The four target variables evapotranspiration (ET), snow water equivalent (SWE), runoff (Q), total water storage (TWS), as well as the TWS interannual variability (IAV) are shown. The TWS IAV is calculated as the deviation from the mean seasonal cycle, for observations and the predictions independently. The shaded areas indicate the $0.2 - 0.8$ quantiles of the spatial variablity. For each region and variable, the Pearson correlation coefficient (r) and the Nash–Sutcliffe model efficiency coefficient (NSE) are shown.

rely on a ground-truth. Rather, the patterns are confronted with domain knowledge. Exemplarily, we take a closer look at the liquid water input (w_{inp}) partitioning through fast runoff fraction (α_{q_f}), soil recharge fraction (α_c), and groundwater recharge fraction (α_g). These fractions are known to depend strongly on the water status of the soil (CWD) with, *e.g.*, more fractional runoff under wet conditions. As the fractions are learned from data and no constraints were imposed, we evaluated their relationship with CWD qualitatively and quantitatively using the Spearman's rank correlation coefficient (r_s).

4. RESULTS & DISCUSSION

4.1 Model Performance by Bioclimatic Regions

The observed and simulated time series and the model performance per bioclimatic region are shown in Figure 3. The hybrid model has learned the temporal patterns of the target variables. The seasonality was represented well with varying performance among bioclimatic regions and variables. Remember that ET and Q are upscaled from point measurements and products of machine learning algorithms themselves. The ET product, for example, is known to be affected by systematic biases due to biases in the underlying site measurements and an incomplete spatial sampling (Jung et al., 2020). For that reason, the trust in these variables, especially the interannual variability (IAV), is limited. Similarly, the SWE product is affected by biases due to a signal saturation above 100–150mm (Larue et al., 2017). Therefore, and also because TWS explicitly depends on all the other target variables, we use the observation-based TWS as the main reference for assessing the model performance.

The response of TWS to precipitation can be strongly delayed due to buffering effects of snow mass, soil moisture, or groundwater. This expresses in a lag between the seasonality of precipitation and TWS, but also single precipitation events cause a delayed response in the TWS (Humphrey et al., 2016). The model fit the seasonal patterns of TWS well, especially in the Tropics, Subtropics, and the Northern Hemisphere (NSE >0.8). In the temperate and more clearly in the cold Northern Hemisphere, the predictions exhibited a phase-shift compared to the observations. This means that the model struggled to discharge the input of water at an adequate pace. Similar phase-shifts can be observed in conceptual models (*e.g.* Schellekens et al. (2017) and Trautmann et al. (2018)) and the phenomenon is still under investigation. A reason for this mismatch could be a missing implementation of lateral fluxes between grid-cells or buffering effects of surface water storages like wetlands. In Figure 3, we also show the interannual variability (IAV) of TWS, calculated as the deviation from the mean seasonality. The IAV signal reflects how the model can deal with anomalous conditions, like strong precipitation events or droughts. The model was able to predict the timing and strength of the major TWS anomalies.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLIII-B2-2020-1537-2020 | © Authors 2020. CC BY 4.0 License.

The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLIII-B2-2020, 2020
XXIV ISPRS Congress (2020 edition)

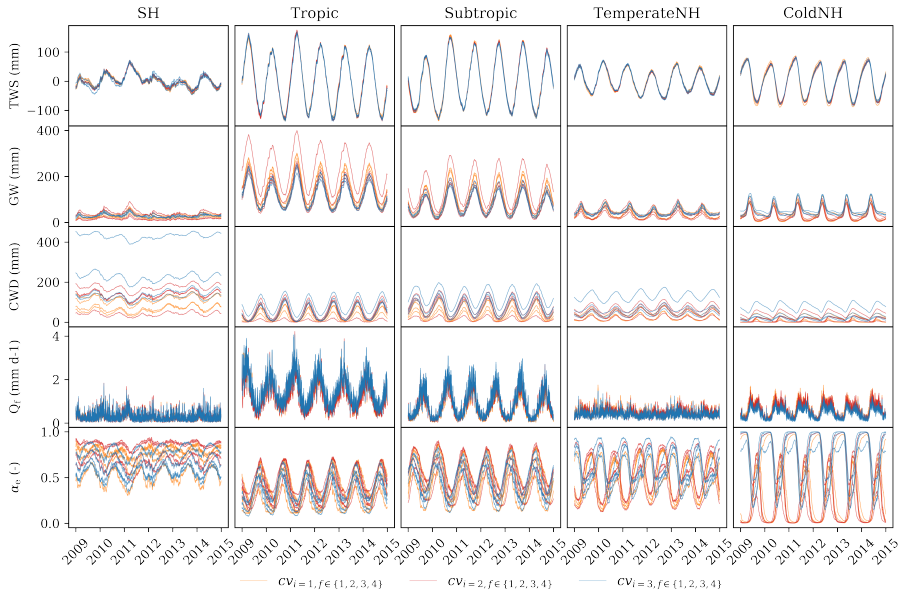


Figure 4. Regional mean time series of repeated model simulations of total water storage (TWS) and the latent variables groundwater (GW), soil cumulative water deficit (CWD), fast runoff (Q_f), and ET partitioning fraction (α_e), defining to what share evapotranspiration is extract form the soil versus groundwater. The lines represent the mean value of a single cross-validation test set. The lines are colored by cross-validation run index, *i.e.*, lines with the same color come from one cross-validation run and represent the same grid-cells. The repeated runs give an impression of the model robustness.

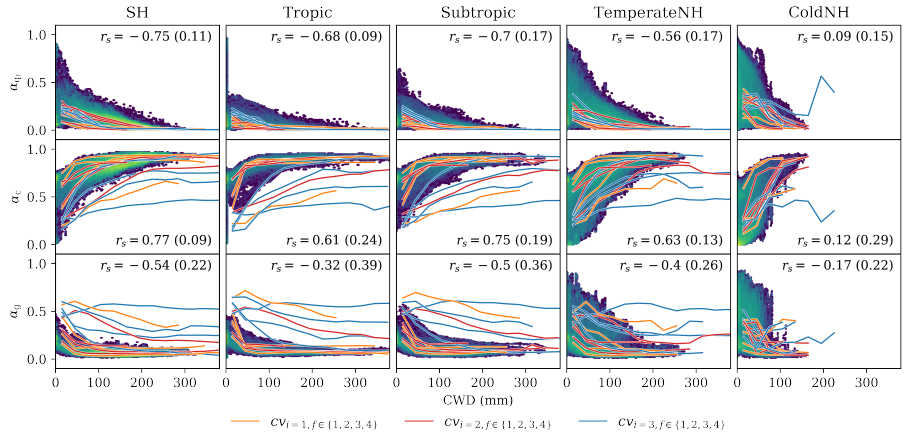


Figure 5. Density plot of the soil cumulative water deficit (CWD) versus the liquid water input (w_{inp}) partitioning fast runoff fraction (α_{q_f}), soil recharge fraction (α_e), and groundwater recharge fraction (α_g). The fractions define how much of w_{inp} goes into the respective fluxes. The relationships is quantified using the mean Spearman's rank correlation coefficient (r_s) over all folds, the standard deviation is shown in brackets. For the density plot, on single fold ($cv_{i=1,f=1}$) was used. The lines represent the binned median, *i.e.*, the median of the fractions over a range of CWD values, of the individual cross-validation test sets. The lines are colored by cross-validation run index, *i.e.*, lines with the same color come from one cross-validation run and represent the same grid-cells.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLIII-B2-2020-1537-2020 | © Authors 2020. CC BY 4.0 License.

The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLIII-B2-2020, 2020
XXIV ISPRS Congress (2020 edition)

4.2 Model Robustness & Latent Variables

A challenge in hybrid modeling is to find the right balance between constraining the model sufficiently to avoid equifinalities and to allow it enough flexibility to adapt to the data. This act of balance requires domain knowledge and a careful evaluation of the results. Based on a set of repeated model runs from the cross-validations, we assess the robustness of the simulations. While the RMSE varied only marginally (1.42 ± 0.03) and the target variables predictions were robust, the stability of the latent variable simulations was lower among cross-validation folds (Figure 4).

The robustness of the latent variable simulations varied among the bioclimatic regions. This indicates that the optimization problem was underconstrained under certain conditions and different pathways lead to a similar solution in terms of target variables. We take a closer look at the SH regions and note that the mean CWD varied substantially among the model runs. Note that, here, the snow mass is neglectable and thus, TWS is partitioned between GW and CWD. TWS, however, reflects the anomalies of the total water column and thus, the absolute values of GW and CWD are not constrained through this relationship. Thus, further constraints were added to the model: through the base runoff (Q_b) being a constant fraction of GW and the ET partitioning, the solution space is reduced. Similarly, the absolute values of CWD are constrained by the $CWD_{overflow}$ and the ET partitioning. Under certain conditions, however, these constraints are not sufficient: in a hydrological regime where soil moisture and groundwater are not limited, for example, the model fails to learn from which pool the ET is extracted. Likewise, if the soil is never or only rarely water-saturated and CWD overflow ($CWD_{overflow}$) does not occur, the mean CWD is not constrained.

In other regions, the simulations were more stable. In the Tropics and Subtropics, GW, CWD, and the ET partitioning fraction (α_e) were estimated more robustly, even if we see some outliers. In the TemperateNH and ColdNH regions, the GW simulations were rather stable, but we see a varying offset of CWD. Here, the model struggled again to yield robust estimates of α_e with even opposite seasonal patterns. This suggests overall that potential groundwater access by plants via ET is not well constrained in the current set-up.

The relationship between w_{inp} partitioning fractions and CWD and its robustness is shown in Figure 5. These patterns follow, to a certain degree, simple hydrological laws: if the soil is wet, for example, we expect to see a decrease in soil recharge fraction (α_c), an increase in groundwater recharge fraction (α_g), and a larger fast runoff fraction (α_{q_f}). Insofar, the patterns align with our prior knowledge. However, the fractions were not estimated robustly, which also reflects in rather large variations in r_s, especially in the cold Northern Hemisphere. There, the relationship was less pronounced, which could be caused by snowmelt dynamics adding complexity.

4.3 Limitations

The cross-validation scheme was designed to have global coverage and reduce spatial and temporal autocorrelation between samples of the training, validation and test set. Due to a limited amount of samples, we made a compromise between data limitations and autocorrelation requirements (Roberts et al., 2017). Similarly, aggregating the daily predictions to match the monthly target variables may introduce leakage, as the target variables can influence the feature time series (*e.g.* ET $\rightarrow$ precipitation). Further, we noted that some cross-validation runs did not converge ideally. Thus, the assessment of the robustness does not only reflect the model robustness, but also the robustness of the training process.

5. CONCLUSION

We presented a global end-to-end hybrid hydrological model that combines artificial neural networks and a conceptual model. To our knowledge, the presented approach is the first application of the hybrid approach to model global environmental systems. The approach opens doors to novel data-driven simulations, attribution, and diagnostic assessments of water cycle variations globally and is applicable to other fields. Our experiments have shown that a major challenge remains to sufficiently constrain the model to retrieve interpretable simulations of non-observed (latent) variables. Under certain conditions, the simulations are unstable but we can infer general patterns of the water cycle using this data-driven approach. Thus, further refinement of the model is required. This iterative process of model improvement, evaluation, and discussion is part of the scientific process that leads ultimately to a better understanding of the subject of investigation.

ACKNOWLEDGEMENTS

We want to thank the International Max Planck Research School for Global Biogeochemical Cycles (IMPRSgBGC) and the Max Planck Institute for Biogeochemistry for the funding and support of this project.

REFERENCES

Baldocchi, D., Falge, E., Gu, L., Olson, R., Hollinger, D., Running, S., Anthoni, P., Bernhofer, C., Davis, K., Evans, R., et al. (2001). "FLUXNET: A new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities." In: *Bulletin of the American Meteorological Society* 82.11, pp. 2415–2434. DOI: 10.1175/1520-0477(2001)082<2415:FANTTS>2.3.CO;2.

Chen, J., Chen, J., Liao, A., Cao, X., Chen, L., Chen, X., He, C., Han, G., Peng, S., Lu, M., et al. (2015). "Global land cover mapping at 30 m resolution: A POK-based operational approach." In: *ISPRS Journal of Photogrammetry and Remote Sensing* 103, pp. 7–27. DOI: 10.1016/j.isprsjprs.2014.09.002.

de Bézenac, E., Pajot, A., and Gallinari, P. (2019). "Deep learning for physical processes: Incorporating prior scientific knowledge." In: *Journal of Statistical Mechanics: Theory and Experiment* 2019.12, p. 124009. DOI: 10.1088/1742-5468/ab3195.

Doelling, D. (2017). *CERES Level 3 SYN1DEG-DAYTerra+Aqua HDF4 file - Edition 4A*. DOI: 10.5067/Terra+Aqua/CERES/SYN1degDay_L3.004A.

DOI/USGS/EROS (1997). *USGS 30 ARC-second Global Elevation Data, GTOPO30*. Boulder CO. DOI: 10.5065/A1Z4-EE71.

Falkner, S., Klein, A., and Hutter, F. (2018). "BOHB: Robust and efficient hyperparameter optimization at scale." In: arXiv: 1807.01774 [cs.LG,cs.ML].

Ghiggi, G., Humphrey, V., Seneviratne, S. I., and Gudmundsson, L. (2019). "GRUN: an observation-based global gridded runoff dataset from 1902 to 2014." In: *Earth System Science Data* 11.4, pp. 1655–1674. DOI: 10.5194/essd-11-1655-2019.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. Cambridge, MA, USA: MIT press. URL: http://deeplearningbook.org.

Hall, D. and Riggs, G. (2016). *Modis/Terra Snow Cover 8-Day L3 Global 0.05 Deg CMG*. Version 6. Boulder, Colorado, USA: NASA National Snow and Ice Data Center Distributed Active Archive Center. DOI: 10.5067/MODIS/MOD10C2.006.

Harris, I., Jones, P. D., Osborn, T. J., and Lister, D. H. (2014). "Updated high-resolution grids of monthly climatic observations–the CRU TS3. 10 Dataset." In: *International journal of climatology* 34.3, pp. 623–642. DOI: 10.1002/joc.3711.

The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLIII-B2-2020, 2020
XXIV ISPRS Congress (2020 edition)

Hengl, T., Jesus, J. M. de, Heuvelink, G. B., Gonzalez, M. R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., et al. (2017). "SoilGrids250m: Global gridded soil information based on machine learning." In: *PLoS ONE* 12.2. DOI: 10.1371/journal.pone.0169748.

Hochreiter, S. and Schmidhuber, J. (1997). "Long short-term memory." In: *Neural computation* 9.8, pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

Huffman, G., Bolvin, D., and Adler, R. (2012). "GPCP version 1.2 1-degree daily (1DD) precipitation data set." In: *World Data Center A, National Climatic Data Center, Asheville, NC*. DOI: 10.5065/d6d50k46.

Humphrey, V., Gudmundsson, L., and Seneviratne, S. I. (2016). "Assessing global water storage variability from GRACE: trends, seasonal cycle, subseasonal anomalies and extremes." In: *Surveys in Geophysics* 37.2, pp. 357–395. DOI: 10.1007/s10712-016-9367-1.

Jung, M., Koirala, S., Weber, U., Ichii, K., Gans, F., Camps-Valls, G., Papale, D., Schwalm, C., Tramontana, G., and Reichstein, M. (2019). "The FLUXCOM ensemble of global land-atmosphere energy fluxes." In: *Scientific data* 6.1, pp. 1–14. DOI: 10.1038/s41597-019-0076-8.

Jung, M., Schwalm, C., Migliavacca, M., Walther, S., Camps-Valls, G., Koirala, S., Anthoni, P., Besnard, S., Bodesheim, P., Carvalhais, N., et al. (2020). "Scaling carbon fluxes from eddy covariance sites to globe: synthesis and evaluation of the FLUXCOM approach." In: *Biogeosciences* 17.5, pp. 1343–1365. DOI: 10.5194/bg-17-1343-2020.

Kendall, A., Gal, Y., and Cipolla, R. (2018). "Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7482–7491. DOI: 10.1109/CVPR.2018.00781.

Larue, F., Royer, A., De Sève, D., Langlois, A., Roy, A., and Brucker, L. (2017). "Validation of GlobSnow-2 snow water equivalent over Eastern Canada." In: *Remote sensing of environment* 194, pp. 264–277. DOI: 10.1016/j.rse.2017.03.027.

Liaw, R., Liang, E., Nishihara, R., Moritz, P., Gonzalez, J. E., and Stoica, I. (2018). "Tune: A research platform for distributed model selection and training." In: arXiv: 1807.05118 [cs.LG].

Liebel, L. and Körner, M. (2018). "Auxiliary tasks in multi-task learning." In: arXiv: 1805.06334v2 [cs.CV].

Luojus, K., Pulliainen, J., Takala, M., Lemmetyinen, J., Kangwa, M., Eskelinen, M., Metsämäki, S., Solberg, R., Salberg, A.-B., Bippus, G., Ripper, E., Nagler, T., Derksen, C., Wiesmann, A., Wunderle, S., Hüsler, F., Fontana, F., and Foppa, N. (2014). *GlobSnow-2 Final Report — European space agency study contract report*. Helsinki: Finnish Meteorological Institute. URL: http://www.globsnow.info/docs/GlobSnow_2_Final_Report_release.pdf.

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

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). "Automatic differentiation in PyTorch." In: *Neural Information Processing Systems Workshop (NIPS-W)*. Long Beach, CA, USA.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., et al. (2019). "Deep learning and process understanding for data-driven Earth system science." In: *Nature* 566.7743, p. 195. DOI: 10.1038/s41586-019-0912-1.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., et al. (2017). "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure." In: *Ecography* 40.8, pp. 913–929. DOI: 10.1111/ecog.02881.

Schellekens, J., Dutra, E., Torre, A. M.-d. la, Balsamo, G., Dijk, A. van, Weiland, F. S., Minvielle, M., Calvet, J.-C., Decharme, B., Eisner, S., et al. (2017). "A global water resources ensemble of hydrological models: The eartH2Observe Tier-1 dataset." In: *Earth System Science Data* 9, pp. 389–413. DOI: 10.5194/essd-2016-55.

Takala, M., Luojus, K., Pulliainen, J., Derksen, C., Lemmetyinen, J., Kärnä, J.-P., Koskinen, J., and Bojkov, B. (2011). "Estimating northern hemisphere snow water equivalent for climate research through assimilation of space-borne radiometer data and ground-based measurements." In: *Remote Sensing of Environment* 115.12, pp. 3517–3529. DOI: 10.1016/j.rse.2011.08.014.

Tootchi, A., Jost, A., and Ducharne, A. (2019). "Multi-source global wetland maps combining surface water imagery and groundwater constraints." In: *Earth System Science Data* 11.1, pp. 189–220. DOI: 10.5194/essd-11-189-2019.

Tramontana, G., Jung, M., Schwalm, C. R., Ichii, K., Camps-Valls, G., Ráduly, B., Reichstein, M., Arain, M. A., Cescatti, A., Kiely, G., and al., et (2016). "Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms." In: *Biogeosciences* 13.14, pp. 4291–4313. ISSN: 1726-4189. DOI: 10.5194/bg-13-4291-2016.

Trautmann, T., Koirala, S., Carvalhais, N., Eicker, A., Fink, M., Niemann, C., and Jung, M. (2018). "Understanding terrestrial water storage variations in northern latitudes across scales." In: *Hydrology and Earth System Sciences* 22.7, pp. 4061–4082. DOI: 10.5194/hess-22-4061-2018.

Viovy, N. (2018). "CRUNCEP version 7-atmospheric forcing data for the community land model." In: *Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory, Boulder CO, USA*. DOI: 10.5065/PZ8F-F017.

Watkins, M. M., Wiese, D. N., Yuan, D.-N., Boening, C., and Landerer, F. W. (2015). "Improved methods for observing Earth's time variable mass distribution with GRACE using spherical cap mascons." In: *Journal of Geophysical Research: Solid Earth* 120.4, pp. 2648–2671. DOI: 10.1002/2014JB011547.

Wielicki, B. A., Barkstrom, B. R., Harrison, E. F., Lee III, R. B., Smith, G. L., and Cooper, J. E. (1996). "Clouds and the Earth's Radiant Energy System (CERES): An Earth Observing System Experiment." In: *Bulletin of the American Meteorological Society* 77.5, pp. 853–868. DOI: 10.1175/1520-0477(1996)077<0853:CATERE>2.0.CO;2.

Wiese, D. N., Yuan, D.-N., Boening, C., Landerer, F. W., and Watkins, M. M. (2018). *JPL GRACE Mascon Ocean, Ice, and Hydrology Equivalent Water Height Release 06 Coastal Resolution Improvement (CRI) Filtered*. PO.DAAC, CA, USA. Version 1.0. DOI: 10.5067/TEMSC-3MJC6.

Wiese, D. N., Landerer, F. W., and Watkins, M. M. (2016). "Quantifying and reducing leakage errors in the JPL RL05M GRACE mascon solution." In: *Water Resources Research* 52.9, pp. 7490–7502. DOI: 10.1002/2016WR019344.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLIII-B2-2020-1537-2020 | © Authors 2020. CC BY 4.0 License. 1544

4.3. Towards hybrid modeling of the global hydrological cycle

Please turn to the next page.

Hydrol. Earth Syst. Sci., 26, 1579–1614, 2022
https://doi.org/10.5194/hess-26-1579-2022

Hydrology and Earth System Sciences
Open Access
EGU

Towards hybrid modeling of the global hydrological cycle

Basil Kraft[1,2], **Martin Jung**[1], **Marco Körner**[2], **Sujan Koirala**[1], **and Markus Reichstein**[1]

[1]Department of Biogeochemical Integration, Max Planck Institute for Biogeochemistry, Germany
[2]Department of Aerospace and Geodesy, Technical University of Munich, Germany

Correspondence: Basil Kraft (bkraft@bgc-jena.mpg.de)

Received: 16 April 2021 – Discussion started: 25 May 2021
Revised: 25 January 2022 – Accepted: 31 January 2022 – Published: 23 March 2022

Abstract. State-of-the-art global hydrological models (GHMs) exhibit large uncertainties in hydrological simulations due to the complexity, diversity, and heterogeneity of the land surface and subsurface processes, as well as the scale dependency of these processes and associated parameters. Recent progress in machine learning, fueled by relevant Earth observation data streams, may help overcome these challenges. But machine learning methods are not bound by physical laws, and their interpretability is limited by design.

In this study, we exemplify a hybrid approach to global hydrological modeling that exploits the data adaptivity of neural networks for representing uncertain processes within a model structure based on physical principles (e.g., mass conservation) that form the basis of GHMs. This combination of machine learning and physical knowledge can potentially lead to data-driven, yet physically consistent and partially interpretable hybrid models.

The hybrid hydrological model (H2M), extended from Kraft et al. (2020), simulates the dynamics of snow, soil moisture, and groundwater storage globally at 1° spatial resolution and daily time step. Water fluxes are simulated by an embedded recurrent neural network. We trained the model simultaneously against observational products of terrestrial water storage variations (TWS), grid cell runoff (Q), evapotranspiration (ET), and snow water equivalent (SWE) with a multi-task learning approach.

We find that the H2M is capable of reproducing key patterns of global water cycle components, with model performances being at least on par with four state-of-the-art GHMs which provide a necessary benchmark for H2M. The neural-network-learned hydrological responses of evapotranspiration and grid cell runoff to antecedent soil moisture states are qualitatively consistent with our understanding and theory. The simulated contributions of groundwater, soil moisture, and snowpack variability to TWS variations are plausible and within the ranges of traditional GHMs. H2M identifies a somewhat stronger role of soil moisture for TWS variations in transitional and tropical regions compared to GHMs.

With the findings and analysis, we conclude that H2M provides a new data-driven perspective on modeling the global hydrological cycle and physical responses with machine-learned parameters that is consistent with and complementary to existing global modeling frameworks. The hybrid modeling approaches have a large potential to better leverage ever-increasing Earth observation data streams to advance our understandings of the Earth system and capabilities to monitor and model it.

1 Introduction

Physically based global hydrological models (GHMs) are an essential tool to understand, monitor, and forecast the water cycle, with an array of societal implications (Jiménez Cisneros et al., 2014). Yet, GHMs and land surface models face many challenges related to process representations and parameterizations, resulting in large uncertainties (Schellekens et al., 2017). The existing state-of-the-art GHMs still disagree across all spatial and temporal scales, which may be attributed to limited, biased, and uncertain data, the heterogeneity of considered processes, or a lack of process understanding (Haddeland et al., 2011; Beck et al., 2017). While global water cycle observations are increasing rapidly, a thorough integration with a GHM to overcome uncertainties is rarely facilitated due to the model complexity and computational expenses, even though some GHMs use data, e.g., river

Published by Copernicus Publications on behalf of the European Geosciences Union.

discharge, to calibrate model parameters (e.g., Van Beek et al., 2011).

Different pathways have been proposed to utilize additional Earth observation data in hydrological modeling. For instance, physically based models benefit from using spatially explicit parameters, which can be retrieved from Earth observation data. It is, for example, common to use spatiotemporally varying leaf area index as a model parameter (e.g., Van Der Knijff et al., 2010) to account for vegetation dynamics. Furthermore, upscaling of locally estimated or measured parameters to global scale – such as catchment parameters (Beck et al., 2016) or soil properties (Hengl et al., 2017) – can improve model accuracy. Using model–data integration approaches, it has been shown that relatively simple conceptual hydrological models can yield state-of-the-art performance when calibrated simultaneously on multiple observational data constraints (Trautmann et al., 2018), which opens new avenues for targeted, partially data-driven experiments to parameterize hydrological processes.

Other approaches to integrate additional observations and physically based models have been developed in the domain of data assimilation (McLaughlin, 2002; Reichle, 2008). While classic data assimilation aims to correct model states or provide initial conditions using additional observational data (Sun et al., 2016), promising concepts exist to learn time-varying model parameters from data (Moradkhani et al., 2005; Geer, 2021). If system understanding and out-of-sample performance (e.g., long-term prediction) are not central, then the use of (purely data-driven) deep learning approaches has been proposed and applied recently in hydrology, and experimental methods for gaining (so far only qualitative) insights exist (Shen et al., 2018).

Recently, it has been proposed to fuse process models with machine learning into one end-to-end modeling system in the so-called hybrid modeling approaches (Reichstein et al., 2019). The hybrid approaches aim at harvesting the information in Earth observation data efficiently by replacing uncertain parameters and processes with a machine learning model, while still maintaining model interpretability and physical consistency. Furthermore, the approach facilitates the incorporation and integration of information from multiple data sources, which is a bottleneck in GHMs. Hybrid modeling can be employed to improve the predictability of the Earth system or components thereof, such as sea surface temperature (de Bézenac et al., 2019) or subgrid atmospheric processes (Rasp et al., 2018). Alternatively, but not mutually exclusive, hybrid modeling can leverage the flexibility of machine learning models with the goal to retrieve data-driven, yet interpretable, physical coefficients and latent variables.

One of the key hydrological data products for diagnosing and understanding global land water cycle variations is total terrestrial water storage (TWS). The TWS is an observation-based rasterized product that integrates all water storage components and is used for calibration and validation of process-based models (Güntner et al., 2007; Schellekens et al., 2017; Trautmann et al., 2018; Scanlon et al., 2019) and in data-driven studies (Humphrey et al., 2016; Andrew et al., 2017; Rodell et al., 2018). An attribution of TWS variations to its components is still unclear, as current model simulations do not produce consistent spatiotemporal patterns due to uncertainties in the model structure and process description, forcing data, and parameter values (Güntner, 2008). Such attribution is not trivial, especially as contiguous observations of the storage components are not available separately on a global scale (e.g., groundwater) or limited (e.g., soil moisture, where satellite observations are only representative of the top soil layers). Thus, decomposition of TWS components is either done with large-scale hydrological modeling (Schellekens et al., 2017), locally using in situ data (e.g., Swenson et al., 2008), or with data-driven approaches without a strict constraint on physical consistency (Andrew et al., 2017).

This study aims to complement and bridge the previous global-scale hydrological modeling and observation-based syntheses by comprehensively evaluating the potential of hybrid modeling at the global scale. In particular, it provides a much-needed data-driven perspective on the global water cycle and its spatiotemporal variability based on carefully designed cross-validation analysis, together with a crucial consideration of the basic physical principle of mass conservation. To do so, we have further developed the model proposed by Kraft et al. (2020), especially with regards to model robustness and physical consistency. The overarching goal of this study is to provide a comprehensive description and assessment of the applicability of the hybrid modeling approach as a potential novel avenue for global hydrological simulation. Particular emphasis are put on benchmarking against and complementing state-of-the-art hydrological models and assessing the plausibility and interpretability of the machine-learning-based data-driven hydrological responses going beyond the typical focus on predictive skills. Furthermore, we examine the potential applications and limitations on a challenging use case of decomposing the contributions of different water storage components to the variations of TWS.

We first describe the datasets used, the hybrid hydrological model (H2M), and the model training and evaluation approach in Sect. 2. We then show the H2M performance in Sect. 3.1 and present the benchmarking against a set of GHM simulations from the eartH2Observe ensemble in Sect. 3.2. Section 3.3 provides the data-driven perspective on hydrological responses, followed by Sect. 3.4, which focuses on the TWS decomposition. Additional plausibility and interpretability of the H2M simulations are presented in Sects. 4.1 and 4.2. Last, we provide a more general assessment of the challenges and opportunities of the hybrid approach in Sect. 4.3.

2 Data and methods

2.1 Datasets

2.1.1 Meteorological forcing

A total of three time-varying meteorological datasets were used to force H2M as follows (Table 1):

i. Precipitation observations, obtained from the Global Precipitation Climatology Project dataset (GPCP-1DD) v1.2 (Huffman et al., 2012),

ii. Net radiation, provided by the SYN1deg Ed3A product (Doelling, 2017) of the Clouds and the Earth's Radiant Energy Systems (CERES) program (Wielicki et al., 1996), and

iii. Air temperature, obtained from CRUNCEP v8 dataset, a product of the observation-based Climatic Research Unit (CRU) and the National Centers for Environmental Prediction (NCEP) reanalysis data (Harris et al., 2014; Viovy, 2018).

To test the impact of the model forcings on the comparison with GHMs (Sect. 3.2), we carried out additional H2M simulation with forcing datasets from the WATCH Forcing Data–ERA-Interim (WFDEI) dataset (Weedon et al., 2014) in an independent setup (Appendix D).

2.1.2 Static variables

A set of temporally static variables was used to represent land surface characteristics as follows (Table 1):

i. The soil properties from the SoilGrids dataset (Hengl et al., 2017), including absolute depth to bedrock and the average (along depth) of the bulk density, coarse fragments, clay, silt, and sand (six variables in total).

ii. The land cover fractions from the GlobeLand30 dataset (Chen et al., 2015) for the 10 classes of water bodies, wetlands, artificial surfaces, tundra, permanent snow and ice, grasslands, barren, cultivated land, shrublands, and forests.

iii. The digital elevation model from GTOPO30 (DOI/USGS/EROS, 1997).

iv. The fractions of groundwater-driven wetlands, regularly flooded wetlands, and the intersection of them (Tootchi et al., 2019), i.e., a total of three variables.

These 20 static variables were spatially aggregated from their finer resolution to $1/30°$ to maintain sub-grid variations, yielding a block of 30 latitude cells times 30 longitude cells times 20 variables, i.e., a total of 18 000 values per 1° grid cell, which is the spatial resolution of the forcing data. Due to the high dimensionality of the static variables, the data were compressed in a preprocessing step using a simple convolutional auto-encoder, consisting of an encoder, a bottleneck layer, and a decoder. The encoder is a stack of consecutively smaller convolutional neural network (CNN) layers that reduce the input block to a vector of size 30, i.e., the bottleneck layer. This process is then reverted in the decoder model, mapping the vector back to the input data. The CNN model is optimized to reconstruct the input data but is forced to find a low-dimensional representation enforced by the bottleneck (e.g., Goodfellow et al., 2016). The resulting compressed dataset consists of 30 latent variables per grid cell that encode the original high-dimensional data (18 000), which is then used as an input to H2M (Sect. 2.2.2). Note that this preprocessing step was done independently from the training of H2M.

2.1.3 Observational constraints

In total, four observational hydrological variables were used to constrain H2M. The datasets were aggregated to a common spatial resolution of 1° (Table 1). Due to differences in temporal coverage of the data products, a common period of February 2002 to December 2014 was selected.

i. The monthly TWS observations from the Gravity Recovery and Climate Experiment (GRACE) Mascon Equivalent Water Height RL06 with Coastal Resolution Improvement (CRI) v1 (Watkins et al., 2015; Wiese et al., 2016, 2018) reflect vertically integrated variations in the water storage. These include the total variations in all storage components, including groundwater, soil moisture, surface water, biosphere-bound water, snow, and ice. To minimize the effect of outliers on the H2M performance, the TWS observations outside the range of -500 to 500 mm were excluded.

ii. Monthly ET estimates were obtained from the global FLUXCOM-RS product (Tramontana et al., 2016; Jung et al., 2019), which is based on machine-learning-driven estimates that are upscaled from site-level FLUXNET eddy covariance measurements (Baldocchi et al., 2001) to a global scale using a range of satellite-based drivers. ET was converted from latent energy estimates assuming a constant latent heat of vaporization of $2.45\,\mathrm{MJ\,mm^{-1}\,m^{-2}}$.

iii. Monthly Q estimates were obtained from the GRUN v1 dataset (Ghiggi et al., 2019). GRUN is based on an upscaling approach that correlates small catchment observations of Q to climate variability. The machine-learned relationships are then generalized to the global scale. Note that only catchments with an area similar to the spatial resolution of the meteorological forcings were used for the prediction, and thus, Q does not include larger routed streamflows and provides an estimate of gridded runoff.

Table 1. Dataset overview, including water cycle constraints, meteorological forcing, and static variables with their native and aggregated spatial and temporal resolution. We use upper case for state variables and lower case for fluxes in the mathematical notation.

	Acr.	Math. notation	Spatial resolution		Temporal resolution	Dataset	Resources
			Native	Agg.			
Water cycle constraints							
Terrestrial water storage	TWS	T	0.50°	1.00°	Monthly	GRACE Tellus JPL RL06M v1	Watkins et al. (2015), Wiese et al. (2018)
Evapotranspiration	ET	e	0.50°	1.00°	Monthly	FLUXCOM v1	Tramontana et al. (2016), Jung et al. (2019)
Grid cell runoff	Q	q	0.50°	1.00°	Monthly	GRUN v1	Ghiggi et al. (2019)
Snow water equivalent	SWE	S	0.25°	1.00°	Daily	GlobSnow v2	Takala et al. (2011), Luojus et al. (2014)
Meteorological forcing							
Precipitation	–	p	1.00°	1.00°	Daily	GPCP 1dd v1.2	Huffman et al. (2012)
Net radiation	–	r_{net}	1.00°	1.00°	Daily	CERES SYN1deg Ed4A	Wielicki et al. (1996), Doelling (2017)
Air temperature	–	T_{air}	0.50°	1.00°	Daily	CRUNCEP v8	Harris et al. (2014), Viovy (2018)
Static variables							
Soil properties	–	–	1/120°	1/30°	–	Soil grids v2	Hengl et al. (2017)
Land cover fractions	–	–	1/360°	1/30°	–	Globland30 v1	Chen et al. (2015)
Digital elevation model	–	–	1/120°	1/30°	–	GTOPO	DOI/USGS/EROS (1997)
Wetlands	–	–	1/240°	1/30°	–	Tootchi	Tootchi et al. (2019)

Note: Acr. – acronym; Agg. – aggregated.

iv. The daily SWE observations were obtained from the GlobSnow v2 product (Takala et al., 2011; Luojus et al., 2014). GlobSnow provides snow water equivalent in the Northern Hemisphere above 40° N, while the mostly snow-free Southern Hemisphere is not covered. In GlobSnow, the time steps with no snow are encoded as missing values. Thus, we gap-filled the GlobSnow product but only with zero values if (a) the snow cover fraction from MODIS (Hall and Riggs, 2016) was below 10 % and (b) the GlobSnow product had missing values in a window of ±12 d. The remaining missing values were not altered.

2.1.4 Global hydrological model ensemble

To evaluate the H2M simulations of TWS and its components, we selected the GHMs from the eartH2Observe ensemble (Schellekens et al., 2017) version WWR1. From the 10 available model simulations, we selected the ones which included groundwater storage: LISFLOOD (Van Der Knijff et al., 2010), W3RA (Van Dijk and Warren, 2010; Van Dijk et al., 2014), PCR-GLOBWB (Van Beek et al., 2011; Wada et al., 2014), and SURFEX-TRIP (Decharme et al., 2010, 2013).

As the models represent different water storages (Table 2), they were combined to conceptually match storages modeled in the H2M (see Sect. 2.2.1). Snow water equivalent (SWE) is available in all models and was used as is. Groundwater (GW) storage, conceptualized as all delayed storage components, is the sum of groundwater and surface storage (SStor), if available for a model. Soil moisture (SM) was combined with canopy interception (CInt), if available. Note that the H2M does not represent SM directly but the cumulative soil water deficit (CWD), but we consider the dynamics of negative CWD to correspond to SM, and thus, the terms are used interchangeably when talking about soil moisture dynamics.

The GHMs were aggregated spatially from 0.5° to match the 1.0° resolution of our simulations. Such spatial aggregations for model comparison are common practice in model intercomparison studies (e.g., Taylor et al., 2012). We expect the variations within four 0.5° cells to be small and thus assume that the 1.0° aggregation does not distort the modeled large-scale spatial patterns.

2.1.5 Data filtering

The data used for H2M were additionally filtered to remove regions with low variations in the hydrological cycle, high anthropogenic impact, and with known data limitations, using the following criteria:

Table 2. The terrestrial water storage (TWS) components as represented by the selected process models. While the hybrid hydrological model (H2M) represents snow water equivalent (SWE) explicitly, like the process models, the remaining TWS components are partitioned into soil cumulative water deficit (CWD) and groundwater (GW), which can be interpreted as fast and slow storage. To compare these components to the global hydrological models (GHMs), we calculated the storage as soil moisture plus canopy interception (CInt) if available and groundwater plus surface storage (SStor) if available, respectively. Note that CWD represents a deficit and, thus, it corresponds to negative soil water storage.

		−CWD (fast storage)		GW (slow storage)	
Model	SWE	SM	CInt	GW	SStor
LISFLOOD	✓	✓	×	✓	×
W3RA	✓	✓	×	✓	×
PCR-GLOBWB	✓	✓	✓	✓	✓
SURFEX-TRIP	✓	✓	✓	✓	✓

Note: SWE – soil water equivalent; CWD – cumulative soil water deficit; GW – groundwater; SM – soil moisture; CInt – canopy interception; SStor – surface storage.

1. grid cells with more than 50 % water bodies, more than 90 % permanent snow or ice, or more than 90 % bare land,
2. regions with more than 90 % artificial built-up surfaces,
3. regions with large groundwater withdrawals labeled as groundwater depletion under anthropogenic influence in Rodell et al. (2018),
4. grid cells with more than 50 % missing values in any of the time series of the observational constraints, and
5. mountainous areas, which are masked in GlobSnow.

After applying the filters, a total of 12 084 of 1° grid cells, covering roughly 80 % of the global land area, were selected.

2.2 The hybrid hydrological model (H2M)

The H2M consists of a dynamic neural network and a simple hydrological framework that represent the major water fluxes and changes in water storage (Fig. 1).

The H2M is set up as a global model, i.e., the same model is used to predict the full spatiotemporal domain, in contrast to separate models for each grid cell in a local setup. The H2M only considers the vertical flow/transport of the water through the system and does not include the lateral flow of either surface (river routing) or sub-surface water (groundwater flow).

The neural network (Sect. 2.2.2) yields a set of time-varying coefficients conditioned on the meteorological forcing and spatial properties derived from the static input variables. These coefficients (e.g., snowmelt factor) are then used in a set of hydrological equations that are introduced in Sect. 2.2.1. For inference (after the optimization of the neural network), the model can be applied to unseen data like any forward simulation model without further model tuning.

For the sake of consistency and clarity, α denotes the time-varying coefficients that are directly estimated by the neural network, and β denotes the global parameters that are learned as spatially constant. Throughout the paper, t is used as time index and i as the grid cell index. Uppercase variables are used for physical state variables. The code is available online (see the code and data availability section).

2.2.1 Hydrological components

In this section, we introduce the main hydrological components of the H2M.

Snow

Snow water equivalent is one of the water storages simulated by the H2M, and it is also constrained by the corresponding observation during model training.

Snow accumulation is precipitation p with air temperatures $T_{\text{air}} \leq 0\,^{\circ}\text{C}$, as follows:

$$s_{\text{acc},t,i} = p_{t,i} \cdot [T_{\text{air},t,i} \leq 0] \cdot \beta_{\text{snow}} \qquad (\text{in mm d}^{-1}). \quad (1)$$

The accumulation is scaled by a learned (optimized) global constant $0 < \beta_{\text{snow}} < 1$. The correction accounts for the known overestimation of solid precipitation due to over-correction for under catch of snowfall in gauge measurements (Decharme and Douville, 2006). Potential snowmelt is then calculated using a degree day approach, as follows:

$$s_{\text{melt},t,i} = \alpha_{\text{smelt},t,i} \cdot \max(T_{\text{air},t,i}, 0) \qquad (\text{in mm d}^{-1}). \quad (2)$$

Opposite to snow accumulation, s_{melt} occurs under the condition of $T_{\text{air}} > 0\,^{\circ}\text{C}$. The time-varying snowmelt coefficient α_{smelt} is estimated by the neural network module and mapped to positive values by applying the softplus activation function, i.e., $\text{Softplus}(x) = \log(1 + e^{x})$. The snow water equivalent is then updated using snow accumulation and melt, as follows:

$$S_{t,i} = \max(S_{t-1,i} + s_{\text{acc},t,i} - s_{\text{melt},t,i}, 0) \qquad (\text{in mm}). \quad (3)$$

Positive values of S are enforced by truncating negative values.

The temperature constraints on snowmelt and accumulation were introduced to avoid compensation effects between s_{acc} and s_{melt}. It must be noted that such constraints are needed despite the fact that the relationship between snowfall or snowmelt and air temperature at 2 m may not always be realistic due to the corresponding associations with atmospheric (for snowfall) and land surface conditions (for snowmelt). We argue that the constraint will reduce or ideally remove equifinality among the parameters, and thus increase identifiability. This would allow for a physical interpretation of the parameters and processes.

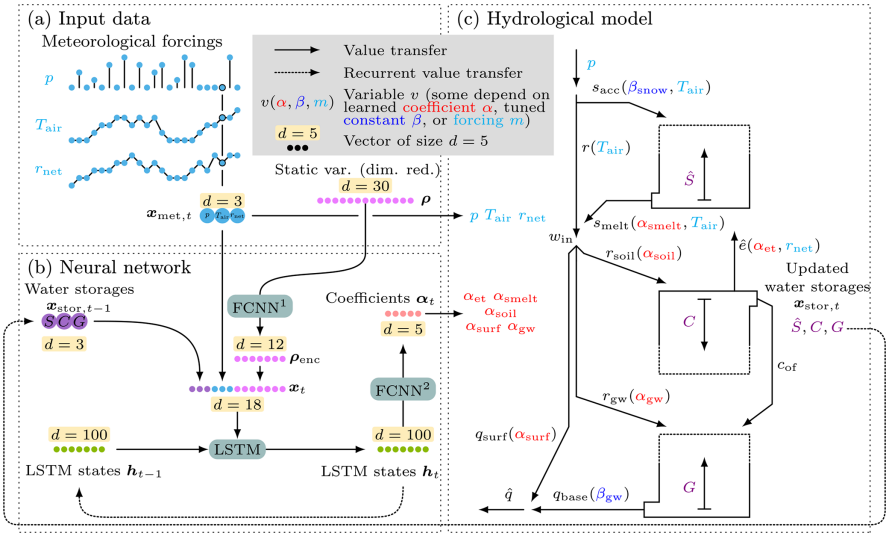


Figure 1. In the H2M, a **(b)** dynamic neural network (NN) simulates a set of time-varying coefficients that are used in a simple **(c)** hydrological model. The meteorological forcings $\boldsymbol{x}_{met,t}$ at time t are used as input **(a)** to the NN and to the physical equations. The NN contains a long short-term memory (LSTM) layer and two fully connected networks (FCNNs). The model maintains two sets of states, namely the (physical) water storages $\boldsymbol{x}_{stor}$ and the LSTM's internal (non-physical) state $\boldsymbol{h}$ (cell state omitted here). It is conditioned on additional inputs representing static land surface and soil properties $\boldsymbol{\rho}$ and the previous water storages $\boldsymbol{x}_{stor,t-1}$. The NN module yields five time-varying coefficients ($\boldsymbol{\alpha}$) which are used in the balance equations. In total, two global parameters ($\boldsymbol{\beta}$) are estimated independently from the data input directly by the optimizer. The location of usage in the balance equations is indicated in parentheses, $\hat{(\cdot)}$ denotes the variables that are constrained with observations, and upper case variables are storages. Forcings (cyan): p – precipitation; T_{air} – air temperature; r_{net} – net radiation. Water storages (purple): $\hat{S}$ – snow water equivalent; C – cumulative soil water deficit; G – groundwater. Time-varying coefficients (red): α_{soil} – soil recharge fraction; α_{gw} – groundwater recharge fraction; α_{surf} – surface runoff fraction; α_{smelt} – snowmelt coefficient; α_{et} – evaporative fraction. Learned global constants (blue): β_{snow} – snow undercatch correction constant; β_{gw}: baseflow constant. Water fluxes: r – rainfall; s_{acc} – snow accumulation; s_{melt} – snowmelt; w_{in} – liquid phase water input; r_{soil} – soil recharge; r_{gw} – groundwater recharge; $\hat{e}$ – evapotranspiration; c_{of} – overflow; q_{surf} – surface runoff; q_{base} – baseflow; $\hat{q}$ – total runoff.

Soil recharge, groundwater recharge, and surface runoff

The water input (in liquid form) w_{in} ($\mathrm{mm\,d^{-1}}$) is the sum of snowmelt and rainfall. It is partitioned into three fluxes, namely surface runoff, q_{surf}, soil recharge, r_{soil}, and groundwater recharge, r_{gw}.

The coefficients for the partitioning are estimated by the neural network module and mapped to the range $(0, 1)$ and naturally constrained to the sum of 1 by applying the softmax transformation; $\mathrm{Softmax}(\boldsymbol{x})_j = e^{x_j} / \sum_k^K e^{x_k}$ for the element j of K elements. The softmax transformation generalizes the logistic function to multiple dimensions. Note that the constraint ensures that the incoming water is neither lost nor generated during the partitioning, respecting the physical law for the conservation of mass.

From the partitioning coefficients, soil recharge r_{soil}, groundwater recharge r_{gw}, and surface runoff q_{surf} fluxes are then calculated as follows:

$$r_{soil,t,i} = \alpha_{soil,t,i} \cdot w_{in,t,i} \quad (\mathrm{in\,mm\,d^{-1}}), \tag{4}$$

$$r_{gw,t,i} = \alpha_{gw,t,i} \cdot w_{in,t,i} \quad (\mathrm{in\,mm\,d^{-1}}), \text{ and} \tag{5}$$

$$q_{surf,t,i} = \alpha_{surf,t,i} \cdot w_{in,t,i} \quad (\mathrm{in\,mm\,d^{-1}}), \tag{6}$$

respectively, where α_{soil}, α_{gw}, and α_{surf} are the partitioning coefficients of the total incoming water w_{in}. All partitioning parameters vary in both space and time.

Evapotranspiration and soil moisture

The total evapotranspiration is calculated as the product of the evaporative fraction α_{et} and net radiation r_{net} ($\text{MJ}\,\text{d}^{-1}\,\text{m}^{-2}$) converted to $\text{mm}\,\text{d}^{-1}$ assuming a latent heat of vaporization of $2.45\,\text{MJ}\,\text{mm}^{-1}\,\text{m}^{-2}$, as follows:

$$e_{t,i} = \alpha_{\text{et},t,i} \cdot \frac{r_{\text{net},t,i}}{2.45} \qquad (\text{in mm}\,\text{d}^{-1}). \tag{7}$$

The evaporative fraction is learned by the neural network and mapped to the range $(0, 1)$ by applying the sigmoid activation function of $\sigma(x) = 1/(1+e^{-x})$. Note that evapotranspiration is constrained by the corresponding observation during model training.

Once the evapotranspiration and soil recharge are calculated, the soil moisture is parameterized as the cumulative soil water deficit $C \geq 0$ as follows:

$$C^*_{t,i} = C_{t-1,i} + r_{\text{soil},t,i} - e_{t,i} \qquad (\text{in mm}), \tag{8}$$

$$c_{\text{of},t,i} = \text{Softplus}(C^*_{t,i}) \qquad (\text{in mm}\,\text{d}^{-1}), \text{ and} \tag{9}$$

$$C_{t,i} = C^*_{t,i} - c_{\text{of},t,i}, \qquad (\text{in mm}), \tag{10}$$

which has the benefit of having a physical saturation limit of 0. For the comparison with the GHMs (Sect. 3.2), we calculate soil moisture (mm) dynamics as $M = -C$. The state C is updated by addition of the soil recharge r_{soil}, subtraction of evapotranspiration e (Eq. 8), and leveling by the overflow mechanism (Eqs. 9 and 10). If C approaches 0, an overflow mechanism allows for direct discharge of excess soil moisture into the deeper groundwater storage. Due to the heterogeneity within a model cell, the overflow c_{of} starts already at values close to 0, which is achieved by using the softplus function.

Baseflow and groundwater

The baseflow is calculated as fraction of the past groundwater storage G_{t-1} via the learned global baseflow constant β_{gw} with the range $(0, 1)$, as follows:

$$q_{\text{base},t,i} = G_{t-1,i} \cdot \beta_{\text{gw}} \qquad (\text{in mm}\,\text{d}^{-1}). \tag{11}$$

Once the baseflow, groundwater recharge, and overflow of soil storage are calculated, the groundwater storage can be updated using a simple water balance, as follows:

$$G_{t,i} = G_{t-1,i} + c_{\text{of},t,i} + r_{\text{gw},t,i} - q_{\text{base},t,i} \quad (\text{in mm}). \tag{12}$$

In H2M, G represents an unconfined aquifer with an unlimited storage capacity.

Total runoff

The total runoff is simply calculated as the sum of the surface runoff q_{surf} (Eq. 6) and the baseflow q_{base} (Eq. 11), as follows:

$$q_{t,i} = q_{\text{surf},t,i} + q_{\text{base},t,i} \qquad (\text{in mm}\,\text{d}^{-1}). \tag{13}$$

We emphasize here that the neural network receives the state of water storage as inputs and is, thus, able to learn interactions of the water storages, the input variables, and the corresponding hydrological partitioning and outflow coefficients. Thus, the runoff generation and evapotranspiration processes do not only depend on the current and past meteorological condition and static variables but also on hydrological state, e.g., the soil water deficit. Therefore, we additionally use runoff as a data constraint during model training.

H2M storage components

For model training against GRACE, the variations in the modeled terrestrial water storage components are added to calculate the total terrestrial water storage as follows:

$$T^*_{t,i} = S_{t,i} + G_{t,i} + (-C_{t,i}) \qquad (\text{in mm}). \tag{14}$$

Note that $-C$ is used in Eq. (14) as C itself is defined as the water deficit. As the observations of the terrestrial water storage from GRACE represent the temporal variations, the mean of simulated storage were removed from each grid cell as follows:

$$T_{t,i} = T^*_{t,i} - \frac{1}{\mathcal{T}} \cdot \sum_{k=1}^{\mathcal{T}} T^*_{k,i} \qquad (\text{in mm}), \tag{15}$$

where k is the time step of $\mathcal{T}$ total steps. The TWS is constrained by observations during model training.

Note that H2M does not represent surface water storage – a fourth major component of TWS, dominant especially in and around large surface water bodies like rivers and lakes – explicitly. This will be considered in the discussion of the results.

Compared to physically based models, the H2M does not explicitly partition the sub-surface storages as soil moisture and groundwater storages. Rather, it is represented as GW and CWD. The partition is an emergent behavior of H2M constraints by the major hydrological fluxes. Negative CWD is loosely and conceptually interpreted as root zone soil moisture, as it serves as the moisture source for evapotranspiration. This is in fact consistent with the physical models, even though CWD does not have a continuous interaction with GW storage except during overflow in H2M.

GW storage represents all delayed residual liquid water storage with infinite capacity. It is constrained by the baseflow fraction and subsequently temporal variation of total runoff (Eq. 11), which leads to a delayed dynamics compared to CWD.

2.2.2 The neural network (NN) module

The NN module (Fig. 1b) consists of three consecutively arranged sub-modules employed for extractions of different features. Overall, the NN module learns the spatiotemporally varying coefficients of the hydrological model using meteorological and dimensionality-reduced static variables of

land (sub-)surface characteristics. The pseudocode of the NN module is presented in Appendix E, while the sub-modules are introduced here.

The first feed-forward (i.e., non-temporal) sub-module learns a compressed representation of the static variables (Eq. 16). This representation, together with meteorological input, is then fed into the second sub-module, a recursive long short-term memory (LSTM) model (Hochreiter and Schmidhuber, 1997), as shown in Eq. (17). The third sub-module (Eq. 18) transforms the outputs of the LSTM to a set of coefficients, which are then fed into the hydrological components. As the model weights are shared across all grid cells, the NN module learns from the global dynamics and not exclusively from each grid cell. For a comprehensive overview of the neural network architectures, see Goodfellow et al. (2016).

The first sub-module is a fully connected neural network (FCNN^1 in Fig. 1), with a single hidden layer and 150 nodes, as follows:

$$\boldsymbol{\rho}_{\text{enc},i} = f_{\text{FCNN}^1}(\boldsymbol{\rho}_i). \tag{16}$$

It takes the static encodings $\boldsymbol{\rho}$ (see Sect. 2.1.2) as inputs and transforms them into a more condensed form ($\boldsymbol{\rho}_{\text{enc}}$). This reduces the high dimensionality of static inputs from 30 to 12 values. Ideally, this lower-dimensional representation describes the most significant gradients of the land characteristics at the sub-grid scale (visualized in Fig. C2; Appendix C). Note that the static variables have already been compressed in a preprocessing step, and the transformation in this sub-module is optimized specifically for the parameterization of the hydrological components.

The second sub-module is an LSTM, a recurrent neural network (RNN) variant that updates its states dynamically using the previous states and the current input. LSTMs are broadly used in the Earth sciences due to their ability to learn temporal dynamics (Körner and Rußwurm, 2021), i.e., to represent memory effects that are present in hydrological observations (Kraft et al., 2019, 2021a; Humphrey et al., 2016). It has a hidden (in the sense of latent) state vector $\boldsymbol{h}$ whose length (100 in H2M) is a tunable hyperparameter. The hidden state is updated at each time step by using interactions of the previous states $\boldsymbol{h}_{t-1,i}$ and the current input $\boldsymbol{x}_{t,i}$, as follows:

$$\boldsymbol{h}_{t,i} = f_{\text{RNN}}(\boldsymbol{h}_{t-1,i}, \boldsymbol{x}_{t,i}). \tag{17}$$

A further cell state $\boldsymbol{c}$ was omitted here for simplicity. In H2M, $\boldsymbol{x}_{t,i}$ is a multivariate input consisting of concatenated current meteorological conditions $\boldsymbol{x}_{\text{met},t,i}$, antecedent physical states from the hydrological model $\boldsymbol{x}_{\text{stor},t-1,i}$, and the static features $\boldsymbol{\rho}_{\text{enc},i}$ from Eq. (16). The input allows the LSTM to learn interactions among the variables conditioned on static land properties like land cover type or elevation. In the optimization process, the RNN learns to maintain a memory of information from past time steps and is capable of updating, removing, and extracting information from its state.

In summary, the LSTM sub-module is similar to a physically based model – it takes the current inputs and static characteristics and updates the system state based on their interactions with the past state. It should be noted that neither its hidden state nor the update function is physically interpretable.

Last, the third sub-module linearly maps the LSTM output $\boldsymbol{h}$ to the coefficients $\boldsymbol{\alpha}$ of the hydrological components (FCNN^2 in Fig. 1), as follows:

$$\boldsymbol{\alpha}_{t,i} = f_{\text{FCNN}^2}(\boldsymbol{h}_{t,i}). \tag{18}$$

The vector $\boldsymbol{\alpha}$ contains five time-varying scalars corresponding to soil recharge fraction α_{soil}, groundwater recharge fraction α_{gw}, surface runoff fraction α_{surf}, snowmelt coefficient α_{smelt}, and evaporative fraction α_{et}.

2.3 Model training

This section introduces the necessary aspects of the model training and validation. First, we introduce the cross-validation setup, followed by the model training, and the loss function.

2.3.1 Cross-validation setup

We use k-fold cross-validation to validate the H2M against observations that were withheld during the training. In the cross-validation, the model is optimized first on a set of training grid cells and applied to a different set of test grid cells, i.e., spatial splitting. Specifically, the grid cells were first split into four sets of grids $g_l, l \in \{1, 2, 3, 4\}$, each consisting of every second grid cell in latitude and longitude direction with an offset O_l. The offsets of $O = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$ are chosen such that the selected grids did not overlap while covering the full spatial domain. This procedure asserts a minimum distance needed to avoid potential issues of spatial autocorrelation (Roberts et al., 2017) within each grid. Each grid was then randomly subdivided into five folds for cross-validation, with three folds for training and one each for validation and testing. The validation subset was used in early stopping, i.e., to stop the training after the validation loss increases over several consecutive iterations. After the training stop, the best model parameters are loaded and predictions are made on the test subset which are used as the final prediction. In the iteration through the folds, every fold is used once in the test set, and as such, a complete set of predictions for a grid cell that was not informed by its own observation is obtained for the respective grid.

In addition to the spatial splitting, the data were also split into calibration and validation time periods akin to the traditional approach. To do so, February 2002 to December 2008 was used for calibration, and January 2009 to December 2014 was used for validation and testing.

The hyperparameters of the NN (i.e., the number of layers and hidden nodes in the neural networks, the learning rate, weight decay, dropout, and gradient clipping) are determined on a single grid, and the cross-validation is only applied on the remaining three grids. For hyperparameter tuning, we employed the Bayesian optimization hyperband (BOHB) algorithm (Falkner et al., 2018) as implemented in the ray.tune framework (Liaw et al., 2018).

This setup was chosen to avoid over-fitting, which is needed due to the data adaptivity of neural networks. Note that the spatial splitting reduces the dependency between the cross-validation sets but does not completely remove it. In addition to the spatial and temporal splitting and the early stopping, we used weight decay (Loshchilov and Hutter, 2017) for regularization.

2.3.2 Training setup

As the neural networks and the hydrological equations are differentiable, standard gradient descent approaches with back-propagation can be used for optimizing the H2M (Goodfellow et al., 2016). We use a multi-task loss as optimization objective which is a recent concept in deep learning for multi-criteria model calibration (see below) and AdamW (Loshchilov and Hutter, 2017) as the optimizer.

Following a common practice in machine learning, the input variables and the observational data constraints are each z-transformed individually to follow a standard normal distribution, using the precomputed mean and standard deviations from the training set. For physical consistency, the corresponding non-transformed variables are used for the hydrological balance equations (see Sect. 2.2).

To obtain an equilibrium of physical and hidden states of H2M, a model spin-up is carried out with spin-up data of a 5-year duration, with each full year selected randomly from the training set. In each optimization iteration, the model is first forced by the spin-up data to retrieve steady states, which are then used as initial conditions during the full forward run with parameter updates (see the pseudocode in Appendix E).

2.3.3 Multi-task loss

The goal of the model optimization is to minimize the total loss, which consists of the following two major aspects:

1) The loss term is calculated as the sum of squared residuals for each z-transformed observational data constraint, as follows:

$$\mathcal{L}_v(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\phi}, \boldsymbol{\beta}) = \sum_{t=1}^{\mathcal{T}} \sum_{i=1}^{\mathcal{I}} ||y_{v,t,i} - \hat{y}_{v,t,i}||^2 \ , \quad v \in \{T, S, e, q\}. \quad (19)$$

Here, $y_{v,t,i}$ and $\hat{y}_{v,t,i}$ are the observed and predicted values of the variable v, respectively. The predictions depend on the input data $\boldsymbol{x}$, the neural network parameters $\boldsymbol{\phi}$, and the learned global constants $\boldsymbol{\beta}$. An additional loss term is employed to promote parameters that would lead to near-zero cumulative soil water deficit C (soil becomes saturated) at least occasionally, as follows:

$$\mathcal{L}_C(\boldsymbol{x}; \boldsymbol{\phi}, \boldsymbol{\beta}) = \sum_{t=1}^{\mathcal{T}} \sum_{i=1}^{\mathcal{I}} (p_{10}(\hat{C}_{t,i}) + b_c) \cdot w_c \ . \quad (20)$$

This term pushes the lower 10 percentile p_{10} of C towards zero. It was needed to reduce the state drift mostly related to spin-up with random years of data that resulted in non-interpretable offsets in C (Kraft et al., 2020). A bias $b_c = 0.1$ was added to prevent the loss from becoming zero, which would interfere with the multi-task loss weighting described below. The loss weight w_c was lowered consecutively during training such that the loss $\mathcal{L}_C$ had only an impact during the early training phase.

2) The task uncertainty term $\boldsymbol{\sigma}$ weights the individual losses dynamically, as follows:

$$\mathcal{L}_{\text{total}}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\phi}, \boldsymbol{\beta}, \boldsymbol{\sigma}) = \sum_{v \in \{T,S,e,q,C\}} \frac{1}{2 \cdot \sigma_v^2} \mathcal{L}_v + \log(\sigma_v) \ , \quad (21)$$

where $\boldsymbol{\sigma}$ is a vector of task-specific uncertainties used to give more or less weight to a particular loss term. The task-specific uncertainties are trained during optimization such that the emphasis on a specific task changes dynamically over the course of the model optimization. Note that $\log(\sigma_v)$ prevents the uncertainties from diverging to infinity. This approach, called self-paced multi-task weighting (Kendall et al., 2018), is advantageous as the weights do not need to be subjectively predefined. The weights are visualized in Fig. C1 in the Appendix.

Hence, the global optimization problem can be expressed as follows:

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}=(\boldsymbol{\phi},\boldsymbol{\beta},\boldsymbol{\sigma})}{\arg\min} \ \mathcal{L}_{\text{total}}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\phi}, \boldsymbol{\beta}, \boldsymbol{\sigma}) \ , \quad (22)$$

in which the parameters of the neural network $\boldsymbol{\phi}$, the global constants $\boldsymbol{\beta}$, and the task weights $\boldsymbol{\sigma}$ are all concurrently and simultaneously optimized.

2.4 Model evaluation and analysis

This section introduces the performance metrics, the spatial and temporal scales, and the methods used to decompose the TWS components.

2.4.1 Performance metrics

The quality of the model predictions was mainly assessed using the Nash–Sutcliffe model efficiency coefficient (NSE) as follows:

$$e_{\mathrm{NSE}} = 1 - \frac{\sum_{i=1}^{N}(m_i - o_i)^2}{\sum_{i=1}^{N}(o_i - \overline{o})^2} \ , \tag{23}$$

where m_i is the modeled value, o_i the observed value, N is the total number of data points, and $\overline{o}$ is the mean of observations (Nash and Sutcliffe, 1970). A NSE of $e_{\mathrm{NSE}} = 1$ indicates a perfect fit, while a NSE of $e_{\mathrm{NSE}} = 0$ ($e_{\mathrm{NSE}} < 0$) indicates that the predictive performance of the model is the same as (worse than) that of the mean. Additionally, the root mean square error (RMSE), the Pearson correlation coefficient (r), and the ratio of modeled versus observed standard deviation (SDR) were used for model performance evaluation.

2.4.2 Temporal and spatial scales

The performance of H2M was evaluated across different temporal scales. To do so, the observed and modeled time series were decomposed into the mean seasonal cycle (MSC) and the interannual variability (IAV) as follows:

$$v_{\mathrm{MSC},m} = \frac{1}{Y}\sum_{y=1}^{Y} v_{m,y} \ , \text{ and} \tag{24}$$

$$v_{\mathrm{IAV},m,y} = v_{m,y} - v_{\mathrm{MSC},m} \ , \tag{25}$$

where v is the observed or modeled time series, m is the month, and y is the year out of Y total years. Before calculating the model performance metrics for MSC and IAV, the linear trends were removed from the time series.

Spatially, the model performance is also evaluated across several scales to investigate robustness of the model for local- to global-scale variations. For the regional-scale analysis, we use continent-wise hydroclimatic biomes from Papagiannopoulou et al. (2018), a machine-learning-based dataset that accounts for climate–vegetation interactions. The number of classes was reduced by combining some of the similar sub-regions, e.g., transitional water-driven and transitional energy-driven or subtypes of boreal regions (Fig. 2). While aggregating the modeled variables to a regional scale, an area-weighted method was used to accommodate for differences in the grid area across the latitude.

For the global-scale performance, we calculate the metrics in two different ways that produce a single metric by a mapping function $f_{\mathrm{perf}} : \mathbb{R}^{\mathcal{T}} \times \mathbb{R}^{\mathcal{T}} \longmapsto \mathbb{R}$ that compares two sequences of length $\mathcal{T}$. The first, which we call the global performance, represents the performance of the globally aggregated variables and is defined as follows:

$$\mathcal{M}_{\mathrm{global}} = f_{\mathrm{perf}}\left(\left\{\mu_{\mathrm{m},t}\right\}_{t=1,\dots,\mathcal{T}}, \left\{\mu_{\mathrm{o},t}\right\}_{t=1,\dots,\mathcal{T}}\right). \tag{26}$$

The variables $\mu_{\mathrm{m},t}$ and $\mu_{\mathrm{o},t}$ represent the modeled and the observed weighted spatial mean for one time step t, respectively. Similar to regional-scale evaluations, these metrics reflect how the area-weighted globally aggregated time series compare. The global-scale signal are themselves useful indicators, as they are often used to characterize the Earth system and land surface processes, e.g., climatic changes (Pachauri et al., 2014), or to evaluate water-carbon relations (Jung et al., 2017; Humphrey et al., 2016).

In contrast, the global summary of the local performance is indicative of how the model performs locally all over the globe and is calculated as follows:

$$\mathcal{M}_{\mathrm{local}} = \mathrm{median}\left(\left\{f_{\mathrm{perf}}(m_{t,i}, o_{t,i})\right\}_{i=1,\dots,\mathcal{I}}\right). \tag{27}$$

Here, the performance is first calculated for the modeled (m) versus observed (o) time series per grid cell i. The resulting cell-wise metric is then reduced using the area-weighted median. The local metrics are useful because the positive and negative model errors and tendencies can compensate when aggregated over a large spatial extent (e.g., Jung et al., 2017).

2.4.3 Terrestrial water storage variations and decomposition

For the analysis on the decomposition of TWS (Sects. 3.4 and 4.2.2), we use the simulated variables SWE, GW, and CWD to assess their contributions to the TWS dynamics, seasonality, and interannual variability. Note that CWD represents a deficit of water in the soil. As a consequence, CWD shows opposite dynamics to water storages. In the following, we calculate the absolute,

$$\mathcal{A}_v = \sum_{t=1}^{\mathcal{T}} |v_t - \overline{v}| \ , \quad v \in \{-C, G, S\}, \tag{28}$$

and relative contribution (hereinafter simply contribution),

$$\mathcal{C}_v = \frac{\mathcal{A}_v}{\sum_{w\in\{-C,G,S\}}\mathcal{A}_w} \quad v \in \{-C, G, S\}, \tag{29}$$

for each component v, following Getirana et al. (2017). Here, $\overline{v}$ is the mean over the time series v. The contributions are calculated per grid cell for the time series and their MSC and IAV.

3 Results

We first assess the performance of H2M simulations against the four observational data constraints (TWS, SWE, Q, and ET) at different spatial and temporal scales. This is followed by a comparison and benchmarking of model performance of H2M TWS and SWE against the simulations from four GHMs in the eartH2Observe ensemble. As the hybrid modeling framework has been significantly developed since Kraft et al. (2020), the H2M performance needs to be re-evaluated here. After the evaluations, we present a closer analysis and interpretation of the parameters estimated by the neural network that define the hydrological responses and generation

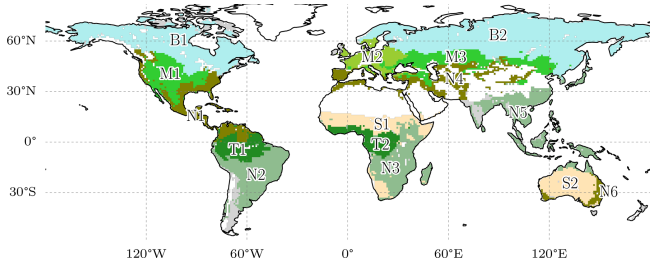


Figure 2. Continental hydro-climatic regions, adapted from Papagiannopoulou et al. (2018). Boreal – North America (B1) and Eurasia (B2); temperate – North America (M1), Europe (M2), and Asia (M3); transitional – North and Central America (N1), South America (S2), Africa (N3), Eurasia and North Africa (N4), Southeast Asia (N5), and Australia (N6); subtropical – Africa (S1) and Australia (S2); tropical – South America (T1) and Africa (T2).

of key hydrological fluxes in H2M. Finally, we present and compare the partitioning of TWS components.

An optimization run of a single cross-validation iteration takes 6 h, a forward run for all grid cells and the entire period from 2002 to 2014 takes about 15 min. Each model was run on a NVIDIA Tesla Volta V100 16 GB GPU (graphics card) with up to 10 CPU cores (Intel® Xeon®; 2.20 GHz) for data buffering and background tasks.

3.1 General model performance

For the assessment of the H2M performance, we only used grid cells from the test set and time steps from the test period of 2009 to 2014, which were not used during the model training and, hence, not seen by the neural network component of H2M.

The model reproduced the patterns of the observed variables well (Table 3). In general, the global signal (global performance; see Eq. 26) was reproduced better than the local cell-level signal (local performance; see Eq. 27). For both observational constraint variables TWS and SWE, a NSE $e_{\mathrm{NSE}} > 0.8$ and Pearson's correlation $r > 0.9$ on the global level and $e_{\mathrm{NSE}} > 0.5$ and $r > 0.8$ for the local level was achieved. The seasonal signals of $\mathrm{TWS_{MSC}}$ and $\mathrm{SWE_{MSC}}$ were modeled with high accuracy ($e_{\mathrm{NSE}} > 0.9$ on the global level; $e_{\mathrm{NSE}} = 0.7$ on the local level) while the interannual variability performance varied. The $\mathrm{TWS_{IAV}}$ was reproduced well with $e_{\mathrm{NSE}} = 0.54$ ($r = 0.8$) on the global level and with $e_{\mathrm{NSE}} = 0.26$ ($r = 0.67$) on the local level. The $\mathrm{SWE_{IAV}}$ performance was decent for the global signal ($e_{\mathrm{NSE}} = 0.22$; $r = 0.87$) but lower ($e_{\mathrm{NSE}} = 0.15$; $r = 0.64$) on the local level.

Both ET and Q are machine learning model based and not directly observed at global scale. The patterns were reproduced well in terms of the seasonality on the global level, while the local performance was lower. For the $\mathrm{ET_{IAV}}$, a low NSE is achieved on the global level ($e_{\mathrm{NSE}} = -0.17$) and on the cell level ($e_{\mathrm{NSE}} = -0.65$), while the correlation is still relatively good, with $r = 0.67$ on the global level and $r = 0.6$ on the local level. The SDR, the ratio of modeled and observed standard deviation, indicates that, on both the global and local levels the variability in the simulated $\mathrm{ET_{IAV}}$ signal is substantially larger than the reference data with SDR of 1.41 on the global level and SDR of 1.65 on the cell level (see Fig. A2 in the Appendix for spatial patterns). For Q, the performance is decent on the global level and lower on the local cell level. Also here, low values in terms of NSE are accompanied by relatively good correlation. Because the independent data for ET and Q are not direct observations, we focus on TWS and SWE in the following. Maps of mean simulated versus observed fluxes and the spatial patterns of the model performance are provided in Appendix A.

3.2 Benchmarking H2M against GHMs

For the quantitative benchmarking of H2M performance with the state-of-the-art GHMs from eartH2Observe (see Sect. 2.1.4), we use the common time period of 2009 to 2012 (not 2009–2014, as in the previous section) but all common grid cells between the GHMs and H2M. This is justified, as H2M has a negligible generalization error in space, i.e., the H2M performance is not systematically better in training grid cells. Similarly, we use the entire common time period (including the training data) for the qualitative assessment of the water cycle dynamics, as also in time, the generalization error was small. We note here that H2M was optimized with the datasets used for evaluation, while the GHMs have either been calibrated using catchment-level observational runoff data (LISFLOOD) or rely on prior parameter estimation (W3RA, SURFEX-TRIP, and PCR-GLOBWB) alone (Schellekens et al., 2017). The comparison presented here serves the purpose of performance benchmarking of the hybrid modeling approach rather than finding the best model.

The H2M modeling efficiency (i.e., the NSE) falls within the range of the GHMs in terms of the global performance

Table 3. The global (spatially averaged) and local (median cell level) model performance for the observational constraint variables terrestrial water storage (TWS), snow water equivalent (SWE), evapotranspiration (ET), and runoff (Q) and their decomposition into the mean seasonal cycle (MSC) and interannual variability (IAV). The Nash–Sutcliffe model efficiency (NSE), Pearson correlation (r), root mean square error (RMSE), and the ratio of modeled and observed standard deviation (SDR) are calculated for the test set, where values represent the mean across the 15 cross-validation runs. Positive values of SDR indicate that the modeled variance is larger than the observed. Note that, for the SWE, cells with constant 0 were dropped. The values were calculated for the test set in the range 2009 to 2014 on monthly timescale.

			TWS			SWE			ET			Q		
		Metric		MSC	IAV		MSC	IAV		MSC	IAV		MSC	IAV
Global	performance	NSE (–)	0.84	0.93	0.54	0.96	0.96	0.22	0.96	0.96	−0.11	0.75	0.78	0.47
		Pearson's r (–)	0.94	0.97	0.80	0.98	0.98	0.87	1.00	1.00	0.67	0.93	0.97	0.81
		SDR (–)	1.15	1.10	1.09	1.02	1.01	1.57	0.99	0.99	1.41	0.93	0.87	1.13
		RMSE (mm)	7.33	4.97	3.27	5.22	5.98	2.16	0.07	0.07	0.02	0.06	0.05	0.03
Local	performance	NSE (–)	0.54	0.70	0.26	0.58	0.74	0.15	0.79	0.87	−0.77	0.20	0.17	0.07
		Pearson's r (–)	0.82	0.93	0.67	0.89	0.96	0.64	0.95	0.98	0.60	0.80	0.91	0.62
		SDR (–)	0.98	1.09	0.95	0.91	0.92	0.97	1.03	1.01	1.65	0.98	0.97	1.04
		RMSE (mm)	42.80	22.59	28.72	15.49	13.13	10.60	0.27	0.22	0.14	0.44	0.31	0.27

(◇ in Fig. 3), although the performance varies less across the variables and temporal scales. However, H2M achieves a consistently higher local performance (boxes in Fig. 3). The TWS is reproduced slightly better by the PCR-GLOBWB, which, however, has a relatively low performance on the local scale. All models struggle to reproduce the SWE_{IAV} signal. The median NSE of H2M is on par with W3RA and SURFEX-TRIP, while the performance on spatially aggregated level is lower. A comparison of the model performance using the same forcings as in the eartH2Observe ensemble is provided in Appendix D.

While all models reproduce the global monthly and seasonal TWS (Fig. 4) relatively well, the results vary more substantially for the TWS_{IAV}. Here, the H2M, WR3A, and LISFLOOD models show the best agreement with the TWS observations (also see Fig. 3 of the model performance). The lower agreement of SURFEX-TRIP and PCR-GLOBWB on the global interannual scale can be attributed to the time periods 2005–2006 and 2008–2010, respectively. From Fig. B1 of the regional averages (Appendix B), it becomes evident that this low agreement on global level is mainly due to a low agreement in the tropical regions (T1 – S-AM tropical; T2 – AFR tropical).

The global SWE was well reproduced by H2M; in particular, the seasonal cycle showed better agreement than the GHMs, where the latter agreed well with the timing but not the magnitude (Fig. 5). The global interannual variability was not reproduced well by H2M, LISFLOOD, and PCR-GLOBWB. Interestingly, H2M performed best when forced by the same WFDEI data as in the GHM simulations (Fig. D1 in Appendix D). A regional model comparison of the time series are provided in Appendix B.

3.3 Hydrological responses in H2M

For the qualitative assessment of the hydrological responses, we use all grid cells, like in the previous section, and show the time range from 2003 to 2014 in time series plots. This involves the training data, but the impact is minimal due to a negligible generalization error. The H2M yields a set of data-driven, spatiotemporally varying coefficients that define the hydrological responses and generation of key hydrological fluxes. In particular, we focus on the following four parameters: α_{soil}, the fraction of throughfall that percolates into the soil, α_{gw}, the fraction that recharges the groundwater, α_{surf}, the fraction that runs off as surface runoff component, and α_{et}, the evaporative fraction (ratio of evapotranspiration to net radiation). Here, we analyze the spatiotemporal variability in the parameters and how they are associated with soil moisture condition defined by soil water deficit. In essence, these are analogous to stage–discharge relationships (Kumar, 2011) that are commonly used to characterize hydrological responses of river discharge at the catchment scale.

The partitioning of the liquid water input w_{inp} (rainfall plus snowmelt) by the fractions for soil recharge (α_{soil}), groundwater recharge (α_{gw}), and surface runoff (α_{surf}) was robust across cross-validation runs and showed a clear relationship to CWD (Fig. 6). With an increasing soil water deficit (larger CWD; drier soil), the soil recharge increases, while the groundwater recharge and surface runoff decrease. For a CWD below 200 mm, we observe a large spatiotemporal variation in the partitioning, evident through the relatively large difference between the 20th and 80th percentiles. The transition from larger soil recharge to larger groundwater recharge and surface runoff is exponentially decreasing, i.e., the change is faster with lower CWD (wetter soil). Above a CWD of 200 mm (dry soil), the partitioning is constant in space and time with α_{soil} converging to 1, while α_{gw} and α_{surf}

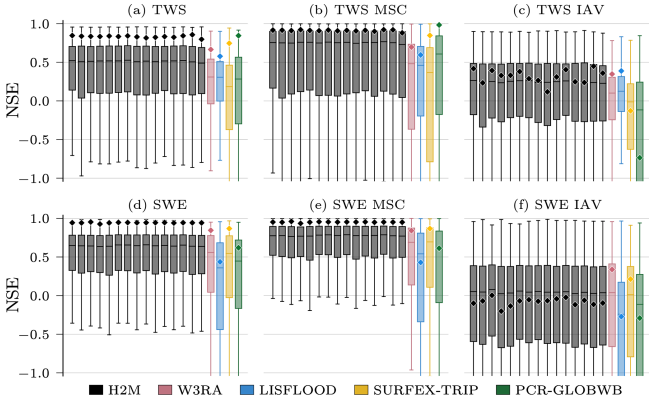


Figure 3. Global and local grid cell level Nash–Sutcliffe model efficiency coefficient (NSE) of the hybrid hydrological model (H2M) and the process-based global hydrological models (GHMs) for the terrestrial water storage (TWS) on top and the snow water equivalent (SWE) at the bottom. The gray bars represent individual cross-validation runs. The ⋄ markers show the global (spatially averaged signal) model performance, and the boxes represent the spatial variability of the local cell level performance. The y axis was cut at -1 due to some large negative NSE values. The panels show the model performance with respect to the full time series, the mean seasonal cycle (MSC), and the interannual variability (IAV). Note that, for SWE, only grid cells with at least 1 d of snow are shown, as the NSE is not defined if the observations are constant zero, which would lead to a comparison of different grid cells. The metrics are calculated from the complete common time range from 2009 to 2012 on a monthly timescale. Note that deviations from the numbers reported in Table 3 are due to different time ranges.

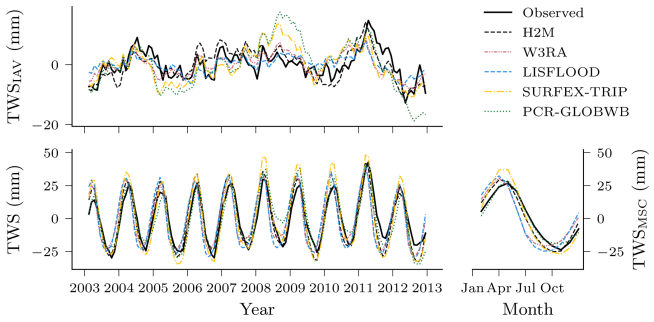


Figure 4. Comparison of the hybrid hydrological model (H2M) and a set of process-based global hydrological models (GHMs) of the terrestrial water storage (TWS), its mean seasonal cycle ($\mathrm{TWS_{MSC}}$), and its interannual variability ($\mathrm{TWS_{IAV}}$) for the global signal. The time series were aggregated using the cell-size-weighted mean across all grid cells. The regional time series are show in Appendix B, Fig. B1.

converge to 0. The relatively large variation under wet conditions (low CWD) in Fig. 6 can be attributed about equally to temporal and spatial variability. The groundwater recharge fraction α_{gw} shows a slightly larger temporal variability than the other fractions, and the contribution of the temporal component was generally a bit lower in the transitional regions.

In most hydroclimatic regions, α_{et} showed a negative relationship to CWD under dry conditions (magenta lines in Fig. 7), and no relationship in presence of precipitation or snowmelt (green lines in Fig. 7). The high latitude and tropical regions showed a less clear relationship and less variation in CWD in general. In all regions, α_{et} was close to 1 with large water input ($w_{\mathrm{in}} > 5$ mm). In arid (S1–2) and semiarid

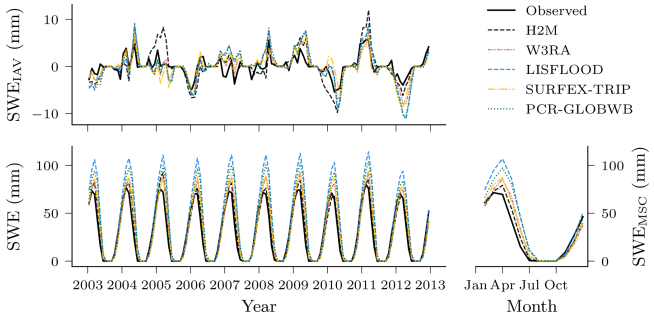


Figure 5. Comparison of the hybrid hydrological model (H2M) and a set of process-based global hydrological models (GHMs) of the snow water equivalent (SWE), its mean seasonal cycle (SWE_{MSC}), and its interannual variability (SWE_{IAV}) for the global signal. The time series were aggregated using the cell-size-weighted mean across all grid cells. The regional time series are show in Appendix B, Fig. B2.

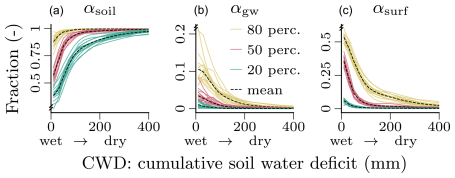


Figure 6. Relationship between the water input partitioning fractions for soil (α_{soil}; **a**), groundwater (α_{gw}; **b**) and surface runoff (α_{surf}; **c**), and the cumulative soil water deficit (CWD), as learned by the neural network. The figure shows the respective percentiles of the spatiotemporal conditional distribution $P(\alpha \mid C \in B_i)$, where C is the cumulative soil water deficit on the x axis discretized into $N = 10$ bins $B = \{[0, 40), \ldots, [360, 400)\}_{i=1,\ldots,N}$. The colored lines show the percentiles per cross-validation run, and the black dashed lines show the mean across the colored lines. The CWD dynamics correspond to negative soil moisture, i.e., larger CWD for drier soils, and thus, a larger CWD corresponds to smaller soil moisture. The plots are based on global daily cell time steps from 2009 to 2014. Note the differences in the y scale.

(N1–5) climates, α_{et} exhibits a large range with steep gradients, given low water input ($w_{in} = 0$ mm), decreasing with larger CWD (drier soil). The 10–90th percentile spread is large in most cases, which indicates that the relationship is modeled with a large spatiotemporal variability.

The H2M shows a large water balance surplus of 12.9 and 21.4 mm yr^{-1}, respectively, depending on the forcing dataset used (Table 4). The values are robust across cross-validation runs. The largest surplus occurs with the GPCP precipitation product, which is 9 mm yr^{-1} larger than WFDEI. The GHMs all show a lower ET and a larger Q trend than H2M.

The global parameters (β) were both estimated robustly, with a mean baseflow constant $\beta_{gw} = 0.008$ and a mean snow undercatch correction constant $\beta_{snow} = 0.77$ and a relative standard deviation of 6 % and 2 % across the 15 cross-validation runs, respectively.

3.4 Terrestrial water storage composition

In this section, we show the TWS partitioning into snow, soil moisture, and groundwater variations as simulated by H2M and compare it with the corresponding partitioning from the GHMs.

The spatial patterns of the TWS partitioning vary strongly among the models (Fig. 8). Some patterns are consistent, though. The TWS seasonality (Fig. 8, top) is dominated by SWE in the high latitudes in all model simulations. Furthermore, all models tend to attribute the TWS variability to soil moisture in hot arid and semiarid climates. In other regions, the models diverge substantially. Both W3RA and PCR-GLOBWB attribute stronger groundwater contributions in most tropical and mild climates, while LISFLOOD and SURFEX-TRIP do not show much variation outside cold, semiarid, and arid regions. In H2M, only the humid Amazon region and Southeast Asia show a distinct contribution from groundwater. For the TWS_{IAV} decomposition (Fig. 8; bottom), we see a rough agreement between the H2M, LISFLOOD, W3RA, and PCR-GLOBWB model in North America, Europe, and northern and central Asia. The latter two, again, show a stronger groundwater contribution, which extends to southern tropical and mild climates. The largest difference between H2M and the GHMs is the low H2M contribution of groundwater to TWS_{IAV} in Africa, which could also be seen in the TWS_{MSC} decomposition (Fig. 8; top).

Not only the spatial patterns of the TWS partitioning show large variations. Figure 9 illustrates the differences in amplitude and timing for the global time series and their decomposition into MSC and IAV. For the seasonal TWS signal,

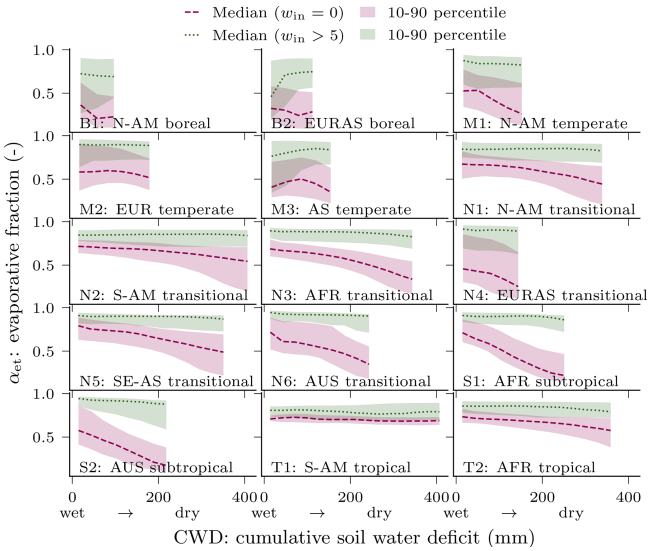


Figure 7. Relationship between evaporative fraction (α_{et}) and cumulative soil water deficit (CWD) for different hydroclimatic regions. The lines shows the respective percentiles of the spatiotemporal conditional distribution $P(\alpha_{et} \mid C \in B_i)$, where C is the cumulative soil water deficit on the x axis discretized into $N = 10$ bins $B = \{[0, 40), \ldots, [360, 400)\}_{i=1,\ldots,N}$. The lines represent the median, and the 10 to 90th percentile is displayed as a shaded area. The red colors depict conditions without water input, $P(\alpha_{et} \mid C \in B_i, w_{in} = 0)$, i.e., no precipitation or snowmelt, and green colors represent high water input larger than 5 mm, $P(\alpha_{et} \mid C \in B_i, w_{in} > 5)$. Note that the CWD minimum was subtracted per grid cell. To exclude cells with a low CWD variability, only the cells in the top 60 % maximum CWD were used. The CWD dynamics correspond to negative soil moisture, i.e., a larger CWD implies drier soils. The plots are based on global daily cell time steps from 2009 to 2014.

Table 4. Global yearly evapotranspiration (ET), grid cell runoff (Q), precipitation (Precip.), and storage change (Δ storage) over the period from 2003 to 2012 for the hybrid hydrological model (H2M) and a set of physically based global hydrological models (GHMs). The H2M was forced with the GPCP precipitation product (H2M) and the WFDEI data (H2M (WFDEI)) independently. The latter dataset is also used by the GHMs. The values for H2M and H2M (WFDEI) represent the mean ± the standard deviation across all cross-validation runs. Values from the common land-mask of all models were considered.

Model	ET (mm yr^{-1})		Q (mm yr^{-1})		Precip.* (mm yr^{-1})	Δ storage (mm yr^{-1})	
H2M	564	±6.7	274	±6.5	860	21.4	±1.1
H2M (WFDEI)	553	±6.0	285	±6.5	851	12.9	±1.0
W3RA	515		332		851	2.5	
LISFLOOD	468		397		851	−14.3	
SURFEX-TRIP	552		296		851	2.3	
PCR-GLOBWB	504		348		851	−1.3	

* GPCP for H2M or, otherwise, WFDEI.

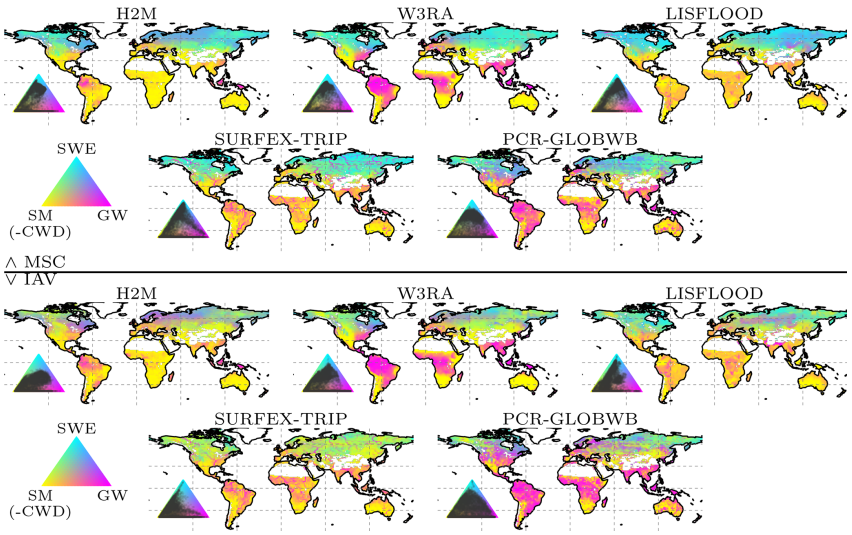


Figure 8. Terrestrial water storage (TWS) variation partitioning into soil moisture (SM, corresponding to negative modeled cumulative water deficit, CWD), groundwater (GW), and snow water equivalent (SWE) by the hybrid hydrological model (H2M) and a set of process-based global hydrological models (GHMs). The top panel shows the partitioning of the mean seasonal cycle (MSC), and the bottom panel shows the interannual variability (IAV). The map colors correspond to the mixture of the contributions of the three variables, and the inset ternary plots reflect the density of the map points projected onto the components. The contribution is calculated as the sum of the bias-removed absolute deviance of a component from the mean, divided by the contribution of all components. Note that surface storage is included in the groundwater component for the models SURFEX-TRIP and PCR-GLOBWB. The decomposition is done based on the years 2003 to 2012.

the amplitudes are qualitatively similar, and the main contribution comes from the snow. H2M, SURFEX-TRIP, and PCR-GLOBWB show a soil moisture slightly delayed to the snow seasonality and the groundwater peak setting in the late northern spring. W3RA shows very similar soil moisture and groundwater curves, which are slightly delayed to the snow seasonality, and LISFLOOD simulates groundwater and soil moisture in alternating cycles with only little variability. The IAV timings of the components are more consistent, but the amplitudes differ significantly across the models. The H2M attributes most TWS_{IAV} to variations in soil moisture, while groundwater dominates the signal for PCR-GLOBWB. Note that the groundwater component also includes the surface water storage for the latter. Also, SURFEX-TRIP and PCR-GLOBWB both show a large global negative IAV anomaly from 2005 to 2006 and a positive one from 2008 to 2010, which are not observed by GRACE.

4 Discussion

In this section, we briefly discuss the model performance and then assess the plausibility of a set of hydrological responses in H2M. We discuss the machine-learned relationship between CWD and runoff-generating processes, followed by an analysis of the CWD–α_{et} (evaporative fraction) relationship. Then, we shed some light on the contrast of TWS composition between H2M and GHM simulations. Finally, we discuss general challenges and opportunities of the hybrid approach.

4.1 Model performance

The H2M simulations have a good agreement with the TWS and SWE observations despite the data biases. While some GHMs performed well at the global scale, H2M shows evidences of data adaptability at the local scale. This can be attributed to the data-driven patterns injected through the neural networks.

The TWS seasonality was reproduced well by H2M, except for extremely arid climates, with a low signal-to-noise ratio in observation, resulting in poor NSE values but also small RMSE and decent Pearson's correlation. The largest errors occur in humid regions with a stark TWS seasonality and large runoff rates, e.g., the Amazon basin, central Africa, and

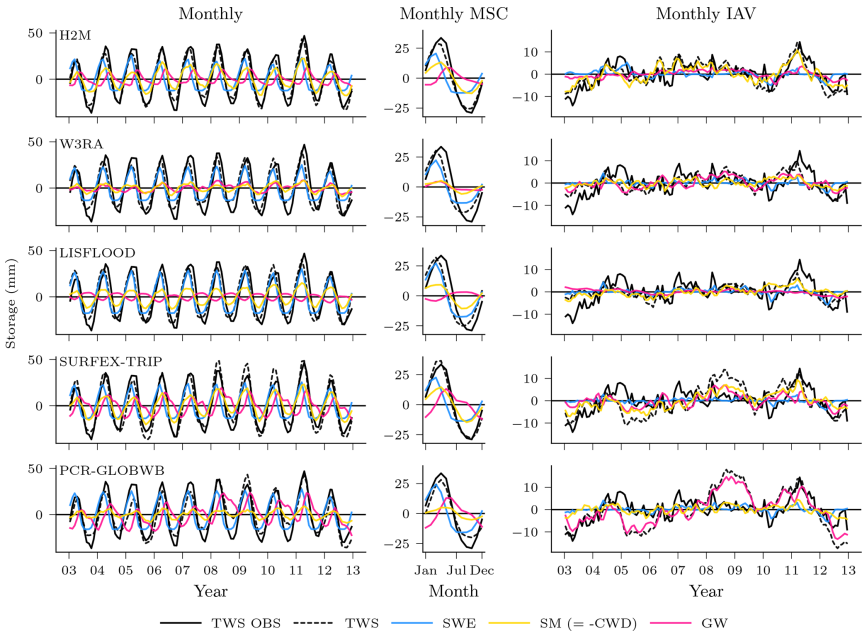


Figure 9. Global variability in the terrestrial water storage (TWS) and the components snow water equivalent (SWE), soil moisture (SM), and groundwater (GW) for the hybrid hydrological model (H2M) and the process-based global hydrological models (rows). Note that SM corresponds to negative modeled cumulative water deficit (CWD) in H2M. For reference, the TWS observations are shown (TWS OBS). The monthly signal (left) and its decomposition into the mean seasonal cycle (MSC; center) and the interannual variability (IAV; right) are arranged in columns. The time series represent the global signal, i.e., the data were aggregated using the cell-size-weighted average per time step, and only cell time steps present in all model simulations were used. The y scale is consistent in columns but varies across the signal components. The training and test period is shown for the complete years 2003 to 2012. Note that surface storage is included in the groundwater component for the models SURFEX-TRIP and PCR-GLOBWB.

Southeast Asia (Fig. A1). This may be related to the missing representations of lateral flow or surface water storage variations in general, which can be important TWS contributions in humid environments (Kim et al., 2009; Scanlon et al., 2019) but also to data biases. A near-perfect fit was achieved for the globally averaged SWE seasonality (Fig. 5), while the local performance varied strongly across regions with the poorest performance in extremely cold tundra (Fig. B2). The SWE_{IAV} is highly sensitive to the precipitation forcing data, which is highlighted by substantially better agreement with GlobSnow when H2M was forced with the WFDEI dataset (Fig. D1 in Appendix D).

In the hybrid modeling framework, the quality of the observational constraints is a major source of uncertainty. The data used in this study have well-documented deficiencies. The precipitation product, for example, shows large uncertainties in Africa due to limitations in density and quality of measurement sites (Sylla et al., 2013) and exhibits biases in snowfall estimates in the Northern Hemisphere due to overcorrection of snowfall under catch (Behrangi et al., 2016; Panahi and Behrangi, 2019). The GlobSnow SWE saturates above 120 mm and underestimates the interannual variability (Luojus et al., 2010). TWS quality is generally difficult to quantify, as an equivalent ground-based measurement does not exist, and its complex preprocessing has known impacts on the data quality (Scanlon et al., 2016). The machine-learning-based constraints of Q and ET are not directly observed, and thus, they are expected to have considerable global and regional uncertainties and biases (Ghiggi et al., 2019; Jung et al., 2020). This could lead to inconsistencies in the water balance (Trautmann et al., 2022). However, the multi-objective optimization may dampen the negative effects of biases, as the model can trade off the different constraints.

4.2 Model interpretability

In this section, we assess the model interpretability, i.e., the plausibility of the hydrological responses that emerge from the machine-learned coefficients which have not been prescribed a priori. We discuss the partitioning of water fluxes and their dependence on antecedent soil moisture condition and then evaluate the partitioning of water storage contributing to TWS dynamics.

4.2.1 Hydrological responses

The H2M-learned hydrological responses to soil moisture states that are consistent with the hydrological understanding, and the learned coefficients are estimated robustly across cross-validation runs. The fact that these patterns are an emerging behavior constrained by a basic physical constraint of mass balance, i.e., the relationships were not explicitly predefined, is an encouraging finding that justifies the usage and further investigation of the hybrid approach in general.

The partitioning of incoming water into surface runoff and recharge of the soil and groundwater shows a clear nonlinear response to soil dryness (Fig. 6). The fraction of surface runoff (α_{surf}) decreases rapidly with increasing dryness while soil recharge (α_{soil}) increases correspondingly. Groundwater recharge occurs under wet conditions and approaches zero with increasing soil dryness. This runoff-generating process response to soil moisture qualitatively matches the expected behavior implemented in GHMs (Bergström, 1995).

The H2M predicts a large spatiotemporal variability in the soil-moisture-dependent runoff–recharge partitioning, as indicated by different percentiles in Fig. 6. For example, under moist conditions, more than 50 % of water input (blue lines in Fig. 6) or hardly anything (yellow lines) can be directed to surface runoff. Such large variability in the response can be expected due to large variations of topography, soil, and vegetation properties that control the infiltration–runoff response. The H2M approach, therefore, appears to offer perspectives in capturing the large natural variability in the effective runoff-generating process response. Note that these processes have been challenging to parameterize in traditional GHMs (Döll and Flörke, 2005; Beck et al., 2016, 2017; Koirala et al., 2017), and thus, the hybrid approach can fill in critical process gaps and long-standing physical modeling challenges.

The learned relationship between evaporative fraction (α_{et}) and soil dryness (Fig. 7) is generally consistent with the demand–supply framework for evapotranspiration (Budyko, 1974). Under wet conditions, ET scales with atmospheric demand represented by net radiation, while evaporative fraction declines with increasing dryness, which is most clearly seen in the semi-arid regions of Australia and Africa. The learned relationship between α_{et} and soil moisture response functions appears to be rather gradual as opposed to an idealized piecewise function with a clear soil moisture threshold that is still frequently employed in process models (Seneviratne et al., 2010; Schwingshackl et al., 2017). However, an about-constant potential evaporative fraction was predicted when there was substantial rain (or snowmelt), independent of the soil moisture state (green lines in Fig. 7). This shows that the model implicitly accounts for wetting of the top soil layers, which alleviates water stress even though it represents soil moisture (expressed as negative CWD) as a single bucket. The specific response of evaporative fraction predicted by H2M varies substantially between regions and within regions indicated by the shading in Fig. 7. Vegetation storage capacity has long been identified as a key uncertainty in the process models in controlling soil moisture stress responses (Ichii et al., 2009). Our approach in H2M avoids such explicit parameterizations of relatively less understood physical processes, and its effectiveness is supported by better performance of H2M in simulating TWS variations in tropical and subtropical regions compared to GHMs (Sect. 3.2), despite its simple overall structure.

4.2.2 Terrestrial water storage composition

As reported previously (Andrew et al., 2017) and as presented here, the attribution of TWS variations is a challenge that is yet to be met in global hydrology. The fact that all models disagree largely with respect to the decomposition was the main motivation to use an alternative, data-driven hybrid approach. The decomposition patterns simulated by H2M are reasonable, although the ground truth for a quantitative assertion is missing. The H2M simulations agree with the GHM, especially in regions where the decomposition is well constrained, which is an encouraging finding. In the tropical and semi-arid to arid regions, the decomposition is less clear. Here, all models disagree, although the larger soil moisture variations versus smaller groundwater variation is a unique feature of the H2M simulations. This may indicate that H2M is under-constrained in these regions. Or, the differences could result from a more accurate representation of the involved processes due to the local adaptivity of H2M. Most likely, it is a combination of both.

The dominant contribution of the SWE to seasonal cycle of TWS in the high latitudes (Figs. 8 and 9), but a lower contribution to the interannual variability is consistent across models and has also been previously reported (e.g., Rangelova et al., 2007; Trautmann et al., 2018). It should be noted that the SWE_{IAV} was reproduced poorly by all models, reflecting large uncertainties in the input precipitation and SWE observations. Despite regional differences, the models also consistently attribute most of the TWS seasonal and interannual variability to soil moisture in arid and semi-arid regions (Fig. 8). The dominance of soil moisture is plausible in these regions, as the potential evapotranspiration is high, and precipitation is low and infrequent or strongly seasonal (Nicholson, 2011). Given the absence of secondary moisture sources, such as lateral flow and a lack of deep-rooted plants, most

of the storage variations occur within a shallow soil depth (Grayson et al., 2006).

In other regions, the partitioning between groundwater and soil moisture variability is less clear. On both the seasonal and interannual scales, groundwater contributions to TWS correlate with humidity at the global scale (Feddema, 2005). In the boreal humid regions of northwestern North America, Scandinavia, and northwestern Russia, as well as the northeastern Asian coast, the groundwater contribution to TWS is larger than that of soil moisture. Here, groundwater recharge is concentrated in spring, with large snowmelt (Fig. 9) cooccurring with low evaporative demand due to low temperatures, irradiation, and vegetation productivity, which results in a large water surplus (Jasechko et al., 2014). The boreal regions with stronger soil moisture contribution are the ones affected by permafrost, where most of the vertical movement is limited to the thawed topsoil, and horizontal baseflow is usually lower than in non-permafrost soils (Bui et al., 2020). Thus, the patterns diagnosed by H2M are plausible. It must be noted, however, that significant drainage of the surplus water happens via river flows and lateral transport, which are not represented in H2M.

The large groundwater contribution on both seasonal and interannual scales in humid regions has been diagnosed by all models. In the tropics, the largest difference between H2M and the GHMs is the larger soil moisture contribution in the African rainforest simulated by H2M. The lower groundwater variability is – to a certain extent – reasonable, as the central Amazon and Southeast Asian rainforests are the most humid regions globally, with the largest annual precipitation (Zelazowski et al., 2011) and a shallow plant rooting depth, while the African rainforest is relatively drier and has deeper plant roots (Yang et al., 2016; Fan et al., 2017). However, the soil moisture variability is only marginally larger in H2M, while it is mainly the low groundwater amplitude that makes the difference (Fig. B3 in Appendix B).

In the arid-to-wet transition regions of Africa, H2M diagnoses only marginal groundwater variability compared to larger amplitudes in the GHMs. The H2M resolves the water balance mainly using soil moisture variations, i.e., through soil recharge and evapotranspiration, while the soil overflow was negligible. While the patterns found by H2M are within those of GHMs in most regions, the notable strong soil moisture contribution in tropical savanna and humid subtropical climates is unique in H2M.

GHMs require a large number of parameters that are either empirically derived or based on remote sensing or statistical datasets, e.g., plant functional types, root zone depth, soil texture maps, or soil thermal and hydraulic properties. Often, the said parameters are uncertain and may not necessarily represent a process at spatial scale of GHMs (scale mismatch) or within grid or catchment variabilities (sub-grid to local heterogeneity). Thus, simple heuristics have been used to parameterize hydrological processes, which can, in reality, be of high complexity (Beck et al., 2016). It has been suggested that GHMs underestimate the land water storage capacity in general and that especially the variability in deeper soil is too low (Zeng et al., 2008). In addition, the link between deeper soil layers and plant transpiration through root water uptake is often not represented adequately in GHMs (Jackson et al., 2000), although such effects have been found to play an important role in below-surface water variability (e.g., Kleidon and Heimann, 2000; Koirala et al., 2017). Compared to the GHMs, H2M provides a novel avenue on which storage variations are less bound by, presumably, ad hoc prescriptions of the size of soil and other storages. The diagnosed patterns of soil and groundwater variations, therefore, emerge from observation-based variations in water storage and fluxes. The H2M approach that also implicitly learns the layering of the soil, thus, can be used to address uncertainties in the moisture storage capacities (Zeng et al., 2008; Scanlon et al., 2019) and plant rooting depth (Yang et al., 2016) used in GHMs, which are likely to have a strong influence on the TWS partitioning.

The smaller groundwater contribution in H2M is also potentially related to the missing mechanisms of capillary rise and root water uptake from the groundwater. Thus, the cumulative water deficit dynamics implicitly represent all the below-ground water that will be returned to the atmosphere by root water uptake and transpiration at some point. As a possible consequence, H2M diagnoses larger soil moisture in transitional and especially in the subtropical regions but, more evidently, smaller groundwater variability.

Finally, the missing (explicit) representation of surface water and river storage may cause biases in H2M simulations. Surface storage has been found to contribute significantly to the TWS variations (Güntner et al., 2007; Scanlon et al., 2019), and a proper representation thereof is desirable. Furthermore, lateral water influx across a cell via rivers is not represented and may have a significant impact on the TWS composition (Kim et al., 2009).

4.3 Challenges and opportunities

The data-driven character of the H2M offers a set of opportunities but is accompanied by challenges. The H2M makes use of observational data streams that are not typically used in GHMs. However, to retain the interpretability of the predicted coefficients, the model structure must be kept simple; the model flexibility needs to be compensated with a simple causal model structure. Still, the H2M offers a great opportunity to study the hydrological cycle from a different viewpoint that is strongly rooted in the observation-based datasets, which are growing in availability at an unprecedented rate in the era of Earth observation.

The hydrological pathways in H2M are rather simple compared to GHMs, but the model still expresses a high data adaptivity, as demonstrated. While GHMs usually represent a wide range of hydrological sub-processes (e.g., infiltration, preferential flow, and topographical runoff–run on), the

hybrid model integrates them to a few response functions, and the model complexity and interactions within are, so to speak, outsourced to the neural network. Still, missing representations of storage components (e.g., surface storage) and hydrological pathways (e.g., streamflows) limit the model flexibility and can, to a certain extent, corrupt the other latent variables as the model tries to accommodate for missing processes. Thus, the estimated coefficients in the current H2M implementation should be treated with some skepticism. At the same time, the relaxation of assumptions can be seen as an opportunity, as the prior knowledge used in GHMs may be wrong or incomplete. The impact of trading prior knowledge and model complexity with more flexibility and a data-driven approach on model uncertainties is a key aspect that needs further investigation.

As the model behavior emerges largely from the observational data constraints, the hybrid approach constitutes a novel technique for studying TWS variations. While purely data-driven approaches (see Andrew et al., 2017, for an overview) are generally useful as they provide insights independent from GHMs, they are based on strong qualitative assumptions (e.g., the temporal characteristics of the components at different depths), and they do not allow the incorporation of physical knowledge, principles, and constraints. GHMs themselves largely rely on prior knowledge, which may be false or incomplete, and the model parameterization is usually not resolved regionally, resulting in model uncertainties (Beck et al., 2016) which are eventually expressed in the disagreement among model simulations. The hybrid model can be seen as a compromise between the purely data-driven and the physically based approaches, as physical principles (e.g., mass conservation) are respected, but qualitative assumptions on the processes are still used.

Global hydrological models are often used for different tasks, such as the assessment of the water cycle at past and present, predictions for the future for evaluating implications of, e.g., land use changes by scenarios, and to gain process understanding. In principle and technically, a global hybrid hydrological model can be applied for the same tasks, while related simulations need to be interpreted with care. The strongest use case of H2M is the assessment of recent variations in the water cycle, since it can act as a physically consistent, yet data-adaptive, bridge between heterogeneous global data streams, and it complements traditional data assimilation approaches. Interpreting predictions too far into the past or future can be risky when factors that are not represented physically play a role that had little impact during the training period (e.g., permafrost melting and CO_2 fertilization of water use efficiency). Likewise, it could make sense to conduct scenarios of, for example, different land use if the conditions represented by the scenarios have been represented during the machine learning in some way, while there always remains the danger that learned relationships by the neural network are just statistical associations rather than causal relationships (shortcut learning; Geirhos et al., 2020). As we could show, gaining process understanding from the hybrid model can be feasible as the spatially and temporally varying coefficients learned by the neural network are plausible and partly very interesting. However, such uncovered patterns may rather represent hypotheses that should be tested with complementary approaches like physical process modeling, direct observations, or experiments.

Improving the model through a better representation of the process complexity is an obvious next step. Several processes were not explicitly represented, such as overland flow, soil moisture recharge from the groundwater through capillary rise, or snow sublimation. The under-complex representation of certain processes leads to biases and uncertainties. For example, estimating the baseflow parameterization on cell level could improve the representative power of the model, as has been shown by Beck et al. (2013). This is, however, challenging as an increasingly complex model needs to be complemented by additional data constraints or better physical processes in order to avoid parameter equifinality issues that lead to the same or similar model responses across a large range of parameter values. It is also possible that the decomposition into CWD and GW is not properly constrained under some circumstances, e.g., in ecosystems that are not water limited. Here, either the groundwater or the soil moisture may be restored as needed (due to frequent precipitation) to match the observation of terrestrial water storage. More research is needed to address these problems, particularly a complementary development of application-based models, as presented in this study, and smaller-scale, better constrained exercises to advance hybrid modeling can be a viable alternative.

Closely related to equifinality is the quantification of model (epistemic) and data (aleatoric) uncertainties. A proper representation of model uncertainties would enable a direct identification of equifinality and allow a targeted model development for uncertain processes. The implementation of such a mechanism could be built into the neural network, e.g., by using Bayesian deep learning (Wang and Yeung, 2020) or deep generative models (Goodfellow et al., 2016). Explicit consideration of data uncertainty will also be beneficial, either to propagate forcing data uncertainties through the model or to model the uncertainties of the observational constraint variables, which is not always provided. Data assimilation is a framework that allows representing such uncertainties (Reichle, 2008) and can even be extended to incorporate model parameter estimation (Moradkhani et al., 2005), i.e., learning physical processes as in the hybrid approach presented here. In contrast to data assimilation that often targets improving prediction skills, the goal of hybrid modeling is to develop a generalizable model, which can be applied beyond the specific forecasting task in data assimilation. Nevertheless, non-parametric machine learning approaches can also be included into data assimilation as discussed in Geer (2021).

The rapid development of novel products opens interesting opportunities, like a daily TWS product (Kvas et al.,

2019) can help to better constrain sub-monthly water processes. Furthermore, the upcoming Surface Water and Ocean Topography (SWOT) mission, which is targeted at observing surface water storage variations (Biancamaria et al., 2016), could be useful to solve current shortcomings of the H2M. In addition, parameters estimated by other approaches, such as the upscaled baseflow index (Beck et al., 2013), offer interesting independent constraints that allow the addition of further complexity to the model without increasing the uncertainty.

Finally, incorporating lateral interactions and flow between grid cells (e.g., large-scale groundwater flow and river routing) are outstanding but relevant challenges, as the paradigm of optimizing neural networks with randomized samples that are independent will likely not be sufficient in modeling connections and interactions between regions. Such endeavors would also allow for bringing in established global datasets of river discharge measurements such as provided by the Global Runoff Data Centre (GRDC; Fekete et al., 1999).

5 Conclusions

The present study demonstrates the strengths of combining machine learning and physical process understanding for global hydrological modeling. The main conclusions of this study are as follows:

1. The hybrid model is capable of obtaining similar performance to physically based models at global level but achieved better local adaptivity. This highlights the strengths of the hybrid approach, which can replace complex physical processes, integrate different datasets, and is highly data-adaptive due to the model parameterization by a neural network.

2. The model simulations were plausible and followed basic hydrological principles. This is partially due to the physical constraints, which force the model into physical consistency (e.g., conservation of mass), but is also emerging from the multiple data constraints.

3. The hybrid model partitioning of the terrestrial water storage into its components yielded plausible and interesting patterns. The agreement of the decomposition is generally high in regions where the physically based models are more consistent (e.g., temperate, semi-arid, and arid regions), but generally, the hybrid model shows a larger contribution by soil moisture.

4. Key opportunities and challenges in hybrid modeling to be addressed in the future are identification of equifinality, quantification of uncertainties, integration of multi-resolution datasets, and representation of cell neighborhood effects, such as lateral fluxes.

Hybrid modeling has the potential to advance the Earth sciences by providing an alternative perspective to knowledge-driven approaches. The data adaptivity can reveal the weaknesses and strengths of process-based models and provide important insights for water cycle attribution and diagnostics. The findings and methods of this study can be generalized to other spheres and scales across the Earth system, as long as sufficient data and process knowledge are available.

Appendix A: Spatial model performance

Overall, high NSE of TWS_{MSC} is achieved in most regions (Fig. A1). Low TWS_{NSE} hot spots are primarily found in some arid regions with little overall TWS variability, e.g., the Namib desert in southern Africa or the Gobi desert in eastern Asia. In terms of the RMSE, regions with larger variations in TWS dominate with the largest MSC error in the Amazon and less expressed in southeastern Asia. The correlation (r) was constantly well above 0.5 for TWS_{MSC}, except for the Gobi Desert, where the TWS variations are minimal. The TWS_{IAV} was also reproduced well in terms of r.

The SWE_{MSC} is reproduced well in terms of NSE and r, while NSE for SWE_{IAV} is low, especially in tundra regions (Fig. A1). The RMSE is also larger in high latitudes but more concentrated in regions with large seasonal amplitudes.

The average patterns of states (TWS and SWE) and fluxes (ET and Q) were reproduced well in general (Fig. A2). The model underestimates the variability in TWS in central Amazon, West Africa, and India. These patterns align well with the occurrence of large rivers (e.g., Amazon, Ganges, Mississippi, Niger, or Yenisei) and may be caused by missing representation of river routing. The SWE is overestimated in the extremely cold regions of North America and Northeast Asia and underestimated in tundra regions. Average Q is largely underestimated in Central Africa and slightly overestimated in northwestern Eurasia, central Amazon, and coastal regions of Australia and East Asia. ET, finally, is underestimated by the model, prominently in most of sub-Saharan Africa and East Brazil, while no major biases are present in other regions.

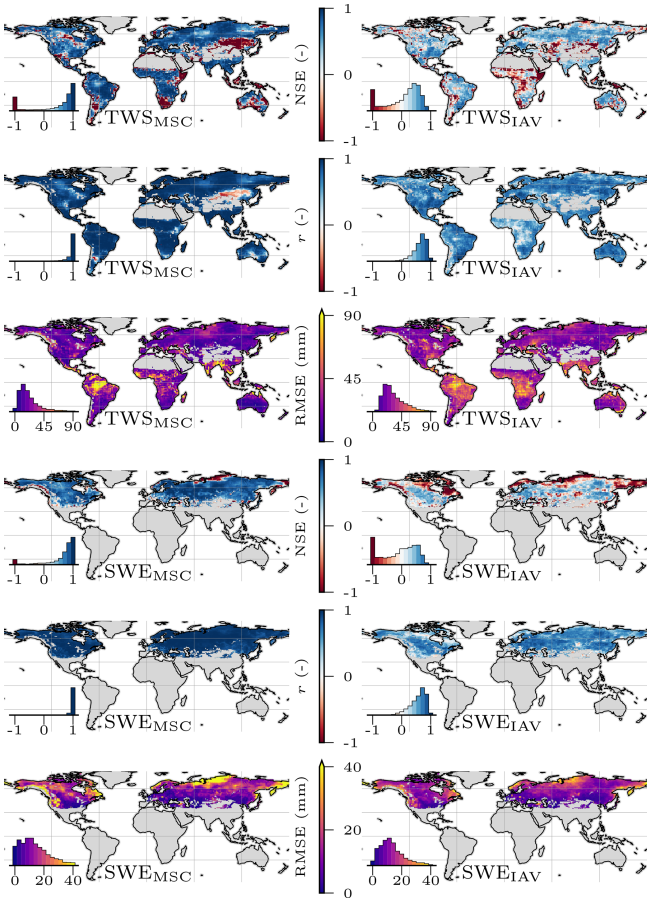


Figure A1. Local model performance for terrestrial water storage (TWS) and snow water equivalent (SWE) on the mean seasonal cycle (MSC) and the interannual variability (IAV) within the test period (2009 to 2014). The Nash–Sutcliffe model efficiency (NSE), Pearson correlation (r), and root mean square error (RMSE) are shown. The inset plots show the cell-area-weighted histogram of the map values.

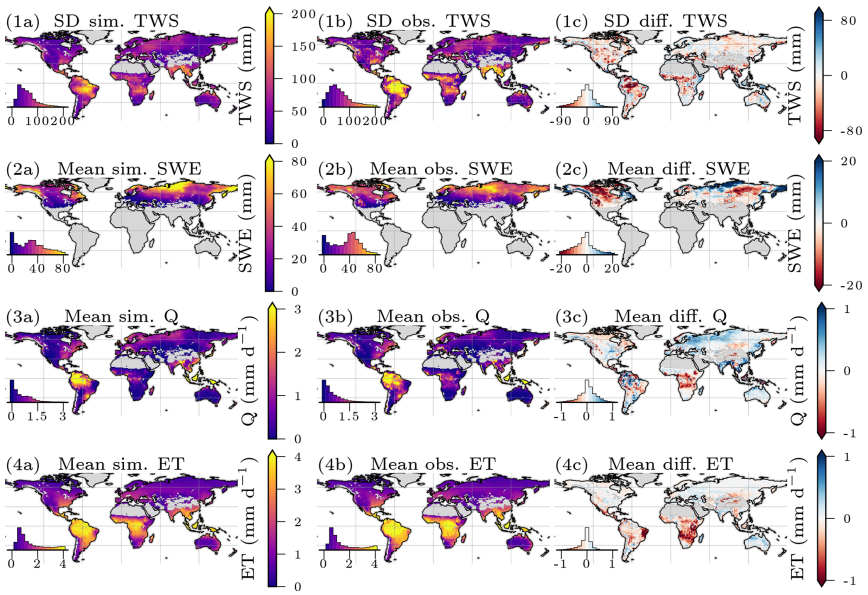


Figure A2. Mean **(a)** simulated, **(b)** observed, and **(c)** the difference of simulated minus observed (positive means simulated is larger) terrestrial water storage (TWS; **1a–c**), snow water equivalent (SWE; **2a–c**), total runoff (Q; **3a–c**), and evapotranspiration (ET; **4a–c**). Note that, for the TWS, the standard deviation is shown as the values represent variations around the mean. The inset histograms represent the map value distributions, and the mean for the test period (2009 to 2014) is shown.

Appendix B: Regional comparison of simulated time series

On a regional scale, most models reproduced the TWS_{MSC} well ($e_{NSE} > 0.5$), while the TWS_{IAV} performance varied ($e_{NSE} < 0.5$) (Fig. B1). The variation between models was larger in terms of IAV, especially in transitional and tropical zones. Especially the TWS_{IAV} seems to be reproduced poorly in certain regions by all models, e.g., temperate Asia (M3), transitional Africa (N3), Eurasia (N4), and Southeast Asia (N5). In the high latitudes, we observe a phase difference of the simulated TWS compared to the observations for all models except the PCR-GLOBWB.

Most models manage to reproduce the SWE_{MSC} well, with an $e_{NSE} > 0.5$, while the SWE_{IAV} performance is more variant and lower in general (Fig. B2). We note a phase difference between the model simulations and observations that is most notable in the boreal regions, indicating that the models either accumulate too much snow during winter or do not manage to discharge it in spring or both. The phase difference is less expressed in H2M and lowest in PCR-GLOBWB. The SWE_{IAV} varies strongly across different regions. The SWE_{IAV} has strong seasonal variations, with opposite patterns in different regions that cancel each other out on a global level. This is evident on the regional anomalies and results in low variability at the global scale. In general, all models reproduce the sign of anomalies better than the amplitudes.

The regional-scale seasonal anomalies of simulated soil moisture (corresponding to negative CWD in H2M) and GW show a more detailed picture of the model variabilities (Fig. B3). The global-scale SM amplitude of H2M is larger than the one of the GHMs (although close to the SURFEX-TRIP model), while the GW variations are smaller in H2M. The largest discrepancies between H2M and the GHMs are in the northern (N1) and southern (N2) America transitional, the Australian subtropical (S2), and the African tropical (T2) regions. However, also the within GHM variation is large in most regions. The model simulations agree relatively well in the temperate regions (M1–3) and in the Africa (N3), Eurasia (N4), and Australia (N6) transitional zones.

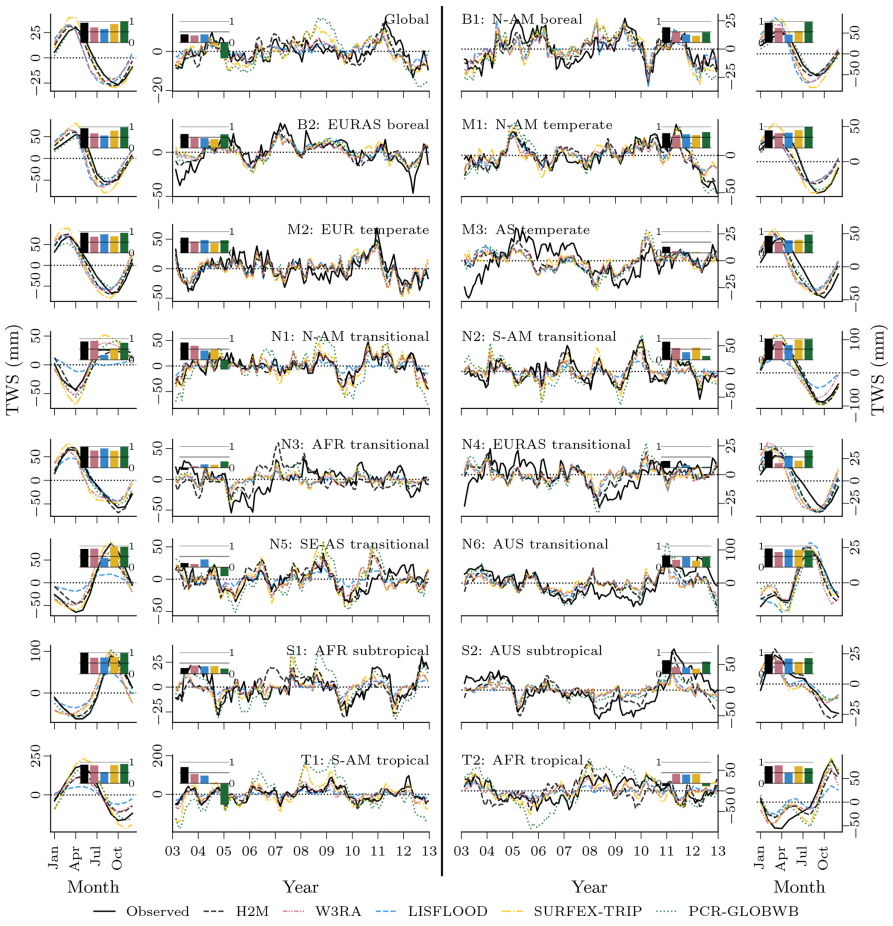


Figure B1. Comparison of the hybrid hydrological model (H2M) and a set of process-based global hydrological models (GHMs) of the terrestrial water storage mean seasonal cycle (TWS_{MSC}; outer columns) and interannual variability (TWS_{IAV}; center columns) in millimeters for hydro-climatic regions (Fig. 2). The time series were aggregated using the cell-size-weighted mean across all grid cells in the respective region. The inset axes show the Nash–Sutcliffe model efficiency (NSE) of each model with the same color-coding as the time series. Note that the *y* scale differs between plots.

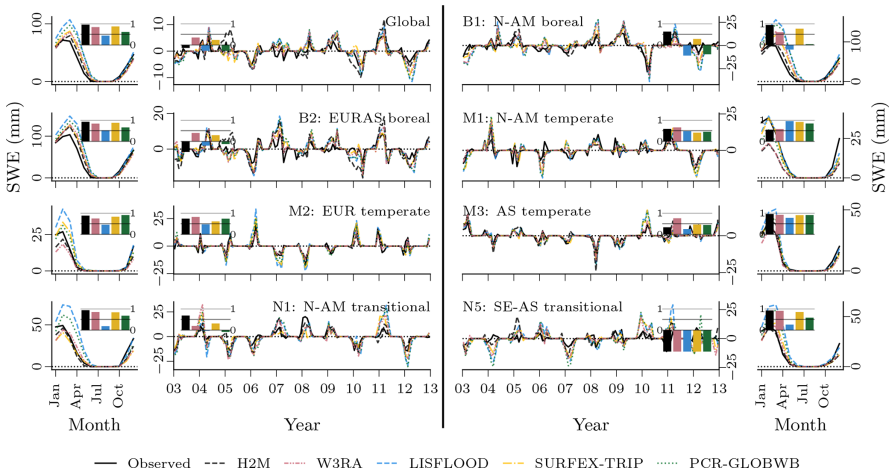


Figure B2. Comparison of the hybrid hydrological model (H2M) and a set of process-based global hydrological models (GHMs) of the snow water equivalent mean seasonal cycle (SWE_{MSC}; outer columns) and interannual variability (SWE_{IAV}; center columns) in millimeters for hydro-climatic regions (Fig. 2). The time series were aggregated using the cell-size-weighted mean across all grid cells in the respective region. The inset axes show the Nash–Sutcliffe model efficiency (NSE) of each model with the same color-coding as the time series. Note that regions without snow dynamics are not included. The *y* scale differs between plots.

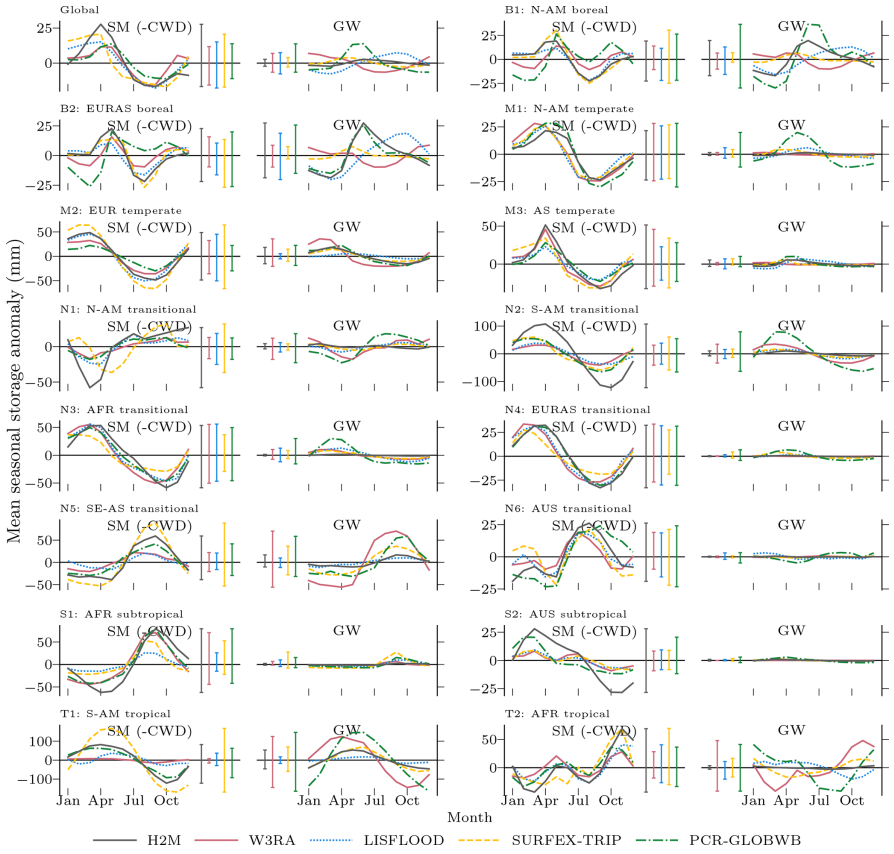


Figure B3. Global and regional mean seasonal anomalies of soil moisture (SM) and groundwater (GW) for the hybrid model (H2M) and the process-based global hydrological models. Note that SM corresponds to negative modeled cumulative water deficit (CWD). Ranges from the minimum to the maximum value per model are shown next to the seasonal cycle as vertical lines. The regions are shown in Fig. 2. Surface storage is included in the groundwater component for the models SURFEX-TRIP and PCR-GLOBWB. The plots are based on global daily cell time steps from 2009 to 2014. Note that the y scale is consistent within, but differs across, regions.

Appendix C: Model optimization

The model optimization within the cross-validation setting is shown in Fig. C1. The learning process was stable in most cases, and a smooth model convergence was achieved. Only one run (fold 2; CV2) was unstable as the training collapsed. Due to the early stopping mechanism, however, the model from the best validation loss is restored and used for the test set prediction. The loss and weight ($w = 1/2\sigma^2$, where σ is the task uncertainty, see Sect. 2.3.3) distributions at optimum across cross-validation runs were stable (bottom row of box plots in Fig. C1). The generalization loss from the training to the validation loss is minimal, although a slightly larger spread of the validation losses can be observed. The largest generalization error occurred with SWE. Note that the training and validation sets are not only split in space but also in time. This could indicate that snow dynamics are less stable over time and change due to, for example, a warming climate.

The task weights were stable across cross-validation runs. The weights are difficult to interpret, as they do not directly translate to inverse variable uncertainty (Kendall et al., 2018) but also depend on the variable variance (although the loss is calculated on standardized data). From the box plots in Fig. C1, we can see that variables with a lower loss are given more weight, except for the CWD loss (a soft constraint that avoids CWD drift in early training), which reaches the optimum at 0.1 relatively quickly. It is possible that the lower weight of TWS is caused by its dependency on the other variables, i.e., if the model tries too hard to improve TWS, other variable losses decrease.

Part of the model tuning involved optimization of the sub-network FCNN^1 (Fig. 1), extracting features from the static variables which are then fed into the recurrent neural network. We visualized the outputs ($\boldsymbol{\rho}_{\text{enc}}$ in Fig. 1) of the FCNN^1 to obtain an impression of the most relevant gradients within the static variables. For visualization, the 12 activations were reduced to three dimensions using t-SNE (t-distributed stochastic neighbor embedding; Hinton and Roweis, 2002). The resulting map (Fig. C2) reveals patterns that seem very familiar, as the components align with patterns of biomass, vegetation type, and aridity. Note that the t-SNE algorithm is non-deterministic and can yield vastly different results depending on chosen hyperparameters. Also, the reduction to three dimensions only reveals the major gradients and does not represent the entire variability.

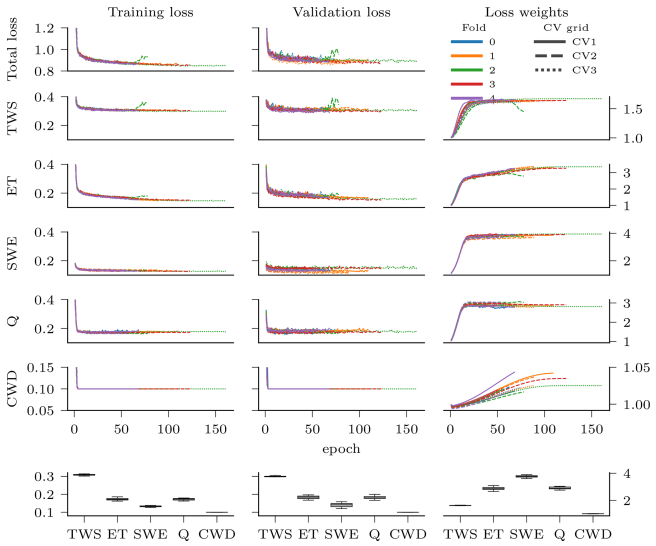


Figure C1. Model training process for the cross-validation runs. The left and central columns represent the unweighted total and variable-specific mean squared error (MSE) loss. The right column shows how the task weights developed over training time. The x axis represents the number of iterations through the training set (epochs). The bottom row contains the column-wise distribution of the variables losses (or weights) at the end of the model optimization. Note that, for the soft constraint on CWD, a bias of 0.1 was added, i.e., 0.1 is the optimum.

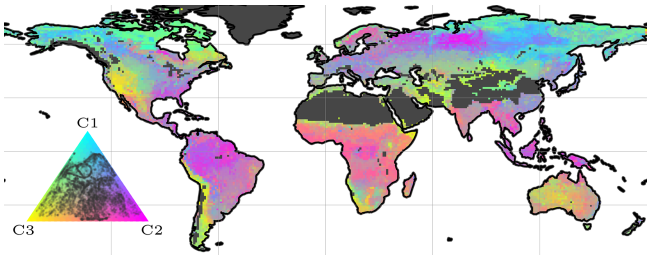


Figure C2. The t-distributed stochastic neighbor (t-SNE) reduction to three dimensions (C1–3) of static variable encoding (originally 12 dimensions; ρ_{enc} in Fig. 1) of one cross-validation run. The encoding is a low-level representation of the static inputs, i.e., soil and land-cover properties, learned by a neural network. The inset ternary plots show the distribution of the map values.

Appendix D: Model forcing with WFDEI

To test the impact of the forcing datasets, the model was trained on the WFDEI forcings (Weedon et al., 2014) as used in the eartH2Observe ensemble. The performance (Fig. D1), with respect to TWS, was almost identical with slightly larger NSE on the global signal and lower NSE on local level when using WFDEI. The NSE of SWE was larger with WFDEI, especially for the IAV. Due to the similar performance, we conclude that the impact of the forcings is negligible, and the results are robust in regards to them.

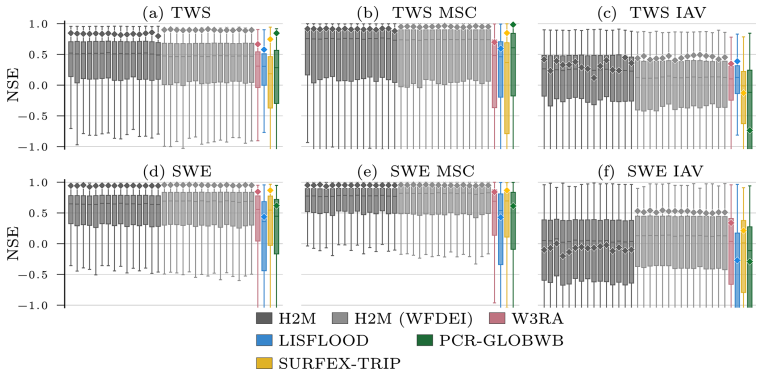


Figure D1. Global and local grid cell level Nash–Sutcliffe model efficiency coefficient (NSE) of the hybrid hydrological model (H2M) and the process-based global hydrological models (GHMs) for the terrestrial water storage (TWS) on top and the snow water equivalent (SWE) at the bottom. The gray bars represent the cross-validation runs using the forcings described in Sect. 2.1.1 (dark gray; H2M), and using the WFDEI forcings as used in the eartH2Observe ensemble (light gray; H2M (WFDEI)). The ⋄ markers show the global (spatially averaged signal) model performance, and the boxes represent the spatial variability in the local cell-level performance. The y axis was cut at -1 due to some large negative NSE values. The panels show the model performance in respect to the full time series, the mean seasonal cycle (MSC), and the interannual variability (IAV). Note that, for SWE, only grid cells with at least 1 d of snow are shown, as the NSE is not defined if the observations are constant zero, which would lead to a comparison of different grid cells. The metrics are calculated from the complete common time range from 2009 to 2012 on monthly timescale. Note that deviations from the numbers reported in Table 3 are due to different time ranges.

Appendix E: Model pseudocode

The pseudocode in Fig. E1 shows the model optimization process.

$\phi, \beta, \sigma \leftarrow initialize()$ # Initialize model weights ϕ, global constants β, task uncertainties σ
while not converged **do**
  $cells \leftarrow sample_{\text{cells}}(gridcells, n)$ # sample n gridcells
  $m_{\text{sim}} \leftarrow meteo[cells]$ # select cells from forcings
  $m_{\text{spinup}} \leftarrow sample_{\text{spinup}}(m_{\text{sim}}, 5)$ # sample 5 random years
  $m \leftarrow concat(m_{\text{spinup}}, m_{\text{sim}})$ # concatenate
  $\rho \leftarrow static[cells]$ # select cells from static
  $y \leftarrow target[cells]$ # select cells from targets
  $c, h \leftarrow zeros(100)$ # initialize LSTM hidden states
  $s \leftarrow zeros(3)$ # initialize physical storages
  $loss \leftarrow 0.0$ # initialize loss
  $\rho_{\text{enc}} \leftarrow FCNN^1(\rho)$ # compress static encodings
  for $t \in \{1, \dots, T\}$ **do**
    $c, h \leftarrow LSTM(c, h, s, m[t], \rho_{\text{enc}})$ # update LSTM states
    $\alpha \leftarrow FCNN^2(h)$ # get coefficients
    $s, f \leftarrow hydro(s, m[t], \alpha, \beta)$ # run phys. model, get storages s and fluxes f
    $\hat{y} \leftarrow collect(s, f)$ # collect target variables
    if $t \notin spinup$ **then**
      $loss \leftarrow loss + MSE(\hat{y}, y[t], \sigma)$ # add weighted loss to previous loss
    end if
  end for
  $\phi, \beta, \sigma \leftarrow update(\phi, \beta, \sigma, loss)$ # update parameters
end while

Figure E1. The training loop of the hybrid hydrological model.

Code and data availability. The H2M and its training are implemented in PyTorch 1.5 (Paszke et al., 2019), an open-source deep learning framework for the Python programming language (https://www.python.org/, Python Core Team, 2022). The simulated hydrological data and the code are available at https://doi.org/10.17617/3.65 (Kraft et al., 2021b). The code is also available on GitHub (https://github.com/bask0/h2m; Kraft, 2022). Note that we cannot share the data used as model input, but all datasets are referenced in the paper.

Author contributions. The study was conceptualized by all the authors. BK implemented the model and performed the data analysis in close collaboration with the co-authors. All authors contributed to the paper.

Competing interests. The contact author has declared that neither they nor their co-authors have any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Acknowledgements. We thank the International Max Planck Research School for Global Biogeochemical Cycles (IMPRS-gBGC) and the Max Planck Institute for Biogeochemistry, for funding and supporting this project. In addition, we thank Uli Weber, for the data preprocessing, and our colleagues at the MPI for Biogeochemistry and TU Munich, for the stimulating discussions. We are very grate-

ful to the reviewers Lieke Melsen, Derek Karssenberg, the anonymous referees, and Albrecht Weerts, the editor, for helping us improve the paper with their comments and suggestions.

Financial support. The article processing charges for this open-access publication were covered by the Max Planck Society.

Review statement. This paper was edited by Albrecht Weerts and reviewed by Derek Karssenberg and three anonymous referees.

References

Andrew, R., Guan, H., and Batelaan, O.: Estimation of GRACE water storage components by temporal decomposition, J. Hydrol., 552, 341–350, https://doi.org/10.1016/j.jhydrol.2017.06.016, 2017.

Baldocchi, D., Falge, E., Gu, L., Olson, R., Hollinger, D., Running, S., Anthoni, P., Bernhofer, C., Davis, K., Evans, R., Fuentes, J., Goldstein, A., Katul, G., Law, B., Lee, X., Malhi, Y., Mayers, T., Munger, W., Walt, O., Paw U, K. T., Pilegaard, K., Schmid, H. P., Valentini, R., Verma, S., Vesla, T., Wilson, K., and Wofsy, S.: FLUXNET: A new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities, B. Am. Meteorol. Soc., 82, 2415–2434, https://doi.org/10.1175/1520-0477(2001)082<2415:FANTTS>2.3.CO;2, 2001.

Beck, H. E., van Dijk, A. I., Miralles, D. G., de Jeu, R. A., Bruijnzeel, L. S., McVicar, T. R., and Schellekens, J.: Global patterns in base flow index and recession based on streamflow observations from 3394 catchments, Water Resour. Res., 49, 7843–7863, https://doi.org/10.1002/2013WR013918, 2013.

Beck, H. E., van Dijk, A. I., De Roo, A., Miralles, D. G., McVicar, T. R., Schellekens, J., and Bruijnzeel, L. A.: Global-scale regionalization of hydrologic model parameters, Water Resour. Res., 52, 3599–3622, https://doi.org/10.1002/2015WR018247, 2016.

Beck, H. E., van Dijk, A. I. J. M., de Roo, A., Dutra, E., Fink, G., Orth, R., and Schellekens, J.: Global evaluation of runoff from 10 state-of-the-art hydrological models, Hydrol. Earth Syst. Sci., 21, 2881–2903, https://doi.org/10.5194/hess-21-2881-2017, 2017.

Behrangi, A., Christensen, M., Richardson, M., Lebsock, M., Stephens, G., Huffman, G. J., Bolvin, D., Adler, R. F., Gardner, A., Lambrigtsen, B., and Fetzer, E.: Status of high-latitude precipitation estimates from observations and reanalyses, J. Geophys. Res.-Atmos., 121, 4468–4486, https://doi.org/10.1002/2015JD024546, 2016.

Bergström, S.: The HBV model, in: Computer Models of Watershed Hydrology, edited by: Singh, V. P., Water Resources Publications, Colorado, USA, 443–476, ISBN 978-1887201742, 1995.

Biancamaria, S., Lettenmaier, D. P., and Pavelsky, T. M.: The SWOT mission and its capabilities for land hydrology, Surv. Geophys., 37, 307–337, https://doi.org/10.1002/2015WR017952, 2016.

Budyko, M. I.: Climate and life, vol. 18, Academic Press, 1 edn., ISBN 978-0121394509, 1974.

Bui, M. T., Lu, J., and Nie, L.: A Review of Hydrological Models Applied in the Permafrost-Dominated Arctic Region, Geosciences, 10, 401, https://doi.org/10.3390/geosciences10100401, 2020.

Chen, J., Chen, J., Liao, A., Cao, X., Chen, L., Chen, X., He, C., Han, G., Peng, S., Lu, M., Zhang, W., Tong, X., and Mills, J.: Global land cover mapping at 30 m resolution: A POK-based operational approach, ISPRS J. Photogramm., 103, 7–27, https://doi.org/10.1016/j.isprsjprs.2014.09.002, 2015.

de Bézenac, E., Pajot, A., and Gallinari, P.: Deep learning for physical processes: Incorporating prior scientific knowledge, J. Stat. Mech.-Theory E., 2019, 124009, https://doi.org/10.1088/1742-5468/ab3195, 2019.

Decharme, B. and Douville, H.: Uncertainties in the GSWP-2 precipitation forcing and their impacts on regional and global hydrological simulations, Clim. Dynam., 27, 695–713, https://doi.org/10.1007/s00382-006-0160-6, 2006.

Decharme, B., Alkama, R., Douville, H., Becker, M., and Cazenave, A.: Global evaluation of the ISBA-TRIP continental hydrological system. Part II: Uncertainties in river routing simulation related to flow velocity and groundwater storage, J. Hydrometeorol., 11, 601–617, https://doi.org/10.1175/2010JHM1212.1, 2010.

Decharme, B., Martin, E., and Faroux, S.: Reconciling soil thermal and hydrological lower boundary conditions in land surface models, J. Geophys. Res.-Atmos., 118, 7819–7834, https://doi.org/10.1002/jgrd.50631, 2013.

Doelling, D.: CERES Level 3 SYN1DEG-DAYTerra+Aqua HDF4 file – Edition 4A, https://doi.org/10.5067/Terra+Aqua/CERES/SYN1degDay_L3.004A, 2017.

DOI/USGS/EROS: USGS 30 ARC-second Global Elevation Data, GTOPO30, https://doi.org/10.5065/A1Z4-EE71, 1997.

Döll, P. and Flörke, M.: Global-Scale estimation of diffuse groundwater recharge: model tuning to local data for semi-arid and arid regions and assessment of climate change impact, https://d-nb.info/1054768056/34 (last access: 3 March 2021), 2005.

Falkner, S., Klein, A., and Hutter, F.: BOHB: Robust and efficient hyperparameter optimization at scale, arXiv preprint, https://arxiv.org/abs/1807.01774 (last access: 9 March 2022), 2018.

Fan, Y., Miguez-Macho, G., Jobbágy, E. G., Jackson, R. B., and Otero-Casal, C.: Hydrologic regulation of plant rooting depth, P. Natl. Acad. Sci. USA, 114, 10572–10577, https://doi.org/10.1073/pnas.1712381114, 2017.

Feddema, J. J.: A revised Thornthwaite-type global climate classification, Phys. Geogr., 26, 442–466, https://doi.org/10.2747/0272-3646.26.6.442, 2005.

Fekete, B. M., Vörösmarty, C. J., and Grabs, W.: Global, composite runoff fields based on observed river discharge and simulated water balances, https://csdms.colorado.edu/wiki/Data:GRDC (last access: 9 March 2022), Global Runoff Data Centre Koblenz [data], 1999.

Geer, A.: Learning earth system models from observations: machine learning or data assimilation?, Philos. T. Roy. Soc. A, 379, 20200089, https://doi.org/10.1098/rsta.2020.0089, 2021.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A.: Shortcut learning in deep neural networks, Nature Machine Intelligence, 2, 665–673, https://doi.org/10.1038/s42256-020-00257-z, 2020.

Getirana, A., Kumar, S., Girotto, M., and Rodell, M.: Rivers and floodplains as key components of global terrestrial water storage variability, Geophys. Res. Lett., 44, 10–359, https://doi.org/10.1002/2017GL074684, 2017.

Ghiggi, G., Humphrey, V., Seneviratne, S. I., and Gudmundsson, L.: GRUN: an observation-based global gridded runoff dataset from 1902 to 2014, Earth Syst. Sci. Data, 11, 1655–1674, https://doi.org/10.5194/essd-11-1655-2019, 2019.

Goodfellow, I., Bengio, Y., and Courville, A.: Deep Learning, MIT press, http://www.deeplearningbook.org (last access: 9 March 2022), ISBN 9780262035613, 2016.

Grayson, R. B., Andrew, W., Walker, J. P., Kandel, D. G., Costelloe, J. F., and Wilson, D. J.: Controls on patterns of soil moisture in arid and semi-arid systems, in: Dryland ecohydrology, edited by: D'Odorico, P. and Porporato, A., Springer, Dordrecht, the Netherlands, 109–127, https://doi.org/10.1007/1-4020-4260-4_7, 2006.

Güntner, A.: Improvement of global hydrological models using GRACE data, Surv. Geophys., 29, 375–397, https://doi.org/10.1007/s10712-008-9038-y, 2008.

Güntner, A., Stuck, J., Werth, S., Döll, P., Verzano, K., and Merz, B.: A global analysis of temporal and spatial variations in continental water storage, Water Resour. Res., 43, W05416, https://doi.org/10.1029/2006WR005247, 2007.

Haddeland, I., Clark, D. B., Franssen, W., Ludwig, F., Voß, F., Arnell, N. W., Bertrand, N., Best, M., Folwell, S., Gerten, D., Gomes, S., Gosling, S. N., Hagemann, S., Hanasaki, N., Harding, R., Heinke, J., Kabat, P., Koirala, Sujan amd Oki, T., Polcher, J., Stacke, T., Viterbo, P., Weedon, G. P., and Yeh, P.: Multimodel estimate of the global terrestrial water balance: setup and first results, J. Hydrometeorol., 12, 869–884, https://doi.org/10.1175/2011JHM1324.1, 2011.

Hall, D. and Riggs, G.: Modis/Terra Snow Cover 8-Day L3 Global 0.05 Deg CMG, https://doi.org/10.5067/MODIS/MOD10C2.006, 2016.

Harris, I., Jones, P. D., Osborn, T. J., and Lister, D. H.: Updated high-resolution grids of monthly climatic observations–the CRU TS3. 10 Dataset, Int. J. Climatol., 34, 623–642, https://doi.org/10.1002/joc.3711, 2014.

Hengl, T., de Jesus, J. M., Heuvelink, G. B., Gonzalez, M. R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., and Kempen, B.: SoilGrids250m: Global gridded soil information based on machine learning, PLoS ONE, 12, e0169748, https://doi.org/10.1371/journal.pone.0169748, 2017.

Hinton, G. and Roweis, S. T.: Stochastic neighbor embedding, in: NIPS, vol. 15, Citeseer, 833–840, http://papers.nips.cc/paper/2276-stochastic-neighbor-embedding (last access: 9 March 2022), 2002.

Hochreiter, S. and Schmidhuber, J.: Long short-term memory, Neural Comput., 9, 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735, 1997.

Huffman, G., Bolvin, D., and Adler, R.: GPCP version 1.2 1-degree daily (1DD) precipitation data set, World Data Center A, National Climatic Data Center, Asheville, NC, https://doi.org/10.5065/d6d50k46, 2012.

Humphrey, V., Gudmundsson, L., and Seneviratne, S. I.: Assessing global water storage variability from GRACE: trends, seasonal cycle, subseasonal anomalies and extremes, Surv. Geophys., 37, 357–395, https://doi.org/10.1007/s10712-016-9367-1, 2016.

Ichii, K., Wang, W., Hashimoto, H., Yang, F., Votava, P., Michaelis, A. R., and Nemani, R. R.: Refinement of rooting depths using satellite-based evapotranspiration seasonality for ecosystem modeling in California, Agric. Forest Meteorol., 149, 1907–1918, https://doi.org/10.1016/j.agrformet.2009.06.019, 2009.

Jackson, R. B., Schenk, H., Jobbagy, E., Canadell, J., Colello, G., Dickinson, R., Field, C., Friedlingstein, P., Heimann, M., Hibbard, K., Kicklighter, D. W., Kleidon, A., Neilson, R. P., Parton, W. J., Sala, O. E., and Sykes, M. T.: Belowground consequences of vegetation change and their treatment in models, Ecol. Appl., 10, 470–483, https://doi.org/10.1890/1051-0761(2000)010[0470:BCOVCA]2.0.CO;2, 2000.

Jasechko, S., Birks, S. J., Gleeson, T., Wada, Y., Fawcett, P. J., Sharp, Z. D., McDonnell, J. J., and Welker, J. M.: The pronounced seasonality of global groundwater recharge, Water Resour. Res., 50, 8845–8867, https://doi.org/10.1002/2014WR015809, 2014.

Jiménez Cisneros, B. E., Oki, T., Arnell, N. W., Benito, G., Cogley, J. G., Döll, P., Jiang, T., Mwakalila, S. S., Fischer, T., Gerten, D., Hock, R., Kanae, S., Lu, X., Mata, L. J., Pahl-Wostl, C., Strzepek, K. M., Su, B., and van den Hurk, B.: Freshwater resources, in: Climate change 2014: impacts, adaptation, and vulnerability. Part A: global and sectoral aspects. Contribution of working group II to the fifth assessment report of the intergovernmental panel on climate change, edited by: Field, C. B., Cambridge University Press, 229–269, https://doi.org/10.1017/CBO9781107415379.008, 2014.

Jung, M., Reichstein, M., Schwalm, C. R., Huntingford, C., Sitch, S., Ahlström, A., Arneth, A., Camps-Valls, G., Ciais, P., Friedlingstein, P., Gans, F., Ichii, K., Jain, A. K., Kato, E., Papale, D., Poulter, B., Raduly, B., Rödenbeck, C., Tramontana, G., Viovy, N., Wang, Y.-P., Weber, U., Zaehle, S., and Zeng, N.: Compensatory water effects link yearly global land CO_2 sink changes to temperature, Nature, 541, 516–520, https://doi.org/10.1038/nature20780, 2017.

Jung, M., Koirala, S., Weber, U., Ichii, K., Gans, F., Camps-Valls, G., Papale, D., Schwalm, C., Tramontana, G., and Reichstein, M.: The FLUXCOM ensemble of global land-atmosphere energy fluxes, Scientific Data, 6, 1–14, https://doi.org/10.1038/s41597-019-0076-8, 2019.

Jung, M., Schwalm, C., Migliavacca, M., Walther, S., Camps-Valls, G., Koirala, S., Anthoni, P., Besnard, S., Bodesheim, P., Carvalhais, N., Chevallier, F., Gans, F., Goll, D. S., Haverd, V., Köhler, P., Ichii, K., Jain, A. K., Liu, J., Lombardozzi, D., Nabel, J. E. M. S., Nelson, J. A., O'Sullivan, M., Pallandt, M., Papale, D., Peters, W., Pongratz, J., Rödenbeck, C., Sitch, S., Tramontana, G., Walker, A., Weber, U., and Reichstein, M.: Scaling carbon fluxes from eddy covariance sites to globe: synthesis and evaluation of the FLUXCOM approach, Biogeosciences, 17, 1343–1365, https://doi.org/10.5194/bg-17-1343-2020, 2020.

Kendall, A., Gal, Y., and Cipolla, R.: Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7482–7491, https://doi.org/10.1109/CVPR.2018.00781, Salt Lake City, UT, USA, 18–22 June 2018.

Kim, H., Yeh, P. J.-F., Oki, T., and Kanae, S.: Role of rivers in the seasonal variations of terrestrial water stor-

age over global basins, Geophys. Res. Lett., 36, L17402, https://doi.org/10.1029/2009GL039006, 2009.

Kleidon, A. and Heimann, M.: Assessing the role of deep rooted vegetation in the climate system with model simulations: mechanism, comparison to observations and implications for Amazonian deforestation, Clim. Dynam., 16, 183–199, https://doi.org/10.1007/s003820050012, 2000.

Koirala, S., Jung, M., Reichstein, M., de Graaf, I. E., Camps-Valls, G., Ichii, K., Papale, D., Ráduly, B., Schwalm, C. R., Tramontana, G., and Carvalhais, N.: Global distribution of groundwater-vegetation spatial covariation, Geophys. Res. Lett., 44, 4134–4142, https://doi.org/10.1002/2017GL072885, 2017.

Körner, M. and Rußwurm, M.: Recurrent Neural Networks and the Temporal Component, in: Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science and Geosciences, edited by: Camps-Valls, G., Tuia, D., Zhu, X. X., and Reichstein, M., pp. 105–119, Wiley & Sons, 1st edn., ISBN 978-1-119-64614-3, 2021.

Kraft, B.: H2M model code, GitHub [code], https://github.com/bask0/h2m, last access: 21 March 2021.

Kraft, B., Jung, M., Körner, M., Requena Mesa, C., Cortés, J., and Reichstein, M.: Identifying dynamic memory effects on vegetation state using recurrent neural networks, Frontiers in Big Data, 2, 31, https://doi.org/10.3389/fdata.2019.00031, 2019.

Kraft, B., Jung, M., Körner, M., and Reichstein, M.: Hybrid modeling: Fusion of a deep learning approach and a physics-based model for global hydrological modeling, Int. Arch. Photogramm., 43, 1537–1544, https://doi.org/10.5194/isprs-archives-XLIII-B2-2020-1537-2020, 2020.

Kraft, B., Besnard, S., and Koirala, S.: Emulating Ecological Memory with Recurrent Neural Networks, Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science, and Geosciences, 269–281, Wiley & Sons, ISBN 978-1-119-64614-3, https://doi.org/10.1002/9781119646181.ch18, 2021a.

Kraft, B., Jung, M., Körner, M., Koirala, S., and Reichstein, M.: Daily model simulations, Edmond [data set], https://doi.org/10.17617/3.65, 2021b.

Kumar, A.: Stage-Discharge Relationship, in: Encyclopedia of Snow, Ice and Glaciers. Encyclopedia of Earth Sciences Series, edited by: Singh, V., Singh, P., and Haritashya, U., Springer, Dordrecht, https://doi.org/10.1007/978-90-481-2642-2_537, 2011.

Kvas, A., Behzadpour, S., Ellmer, M., Klinger, B., Strasser, S., Zehentner, N., and Mayer-Gürr, T.: ITSG-Grace2018: Overview and evaluation of a new GRACE-only gravity field time series, J. Geophys. Res.-Sol. Ea., 124, 9332–9344, https://doi.org/10.1029/2019JB017415, 2019.

Liaw, R., Liang, E., Nishihara, R., Moritz, P., Gonzalez, J. E., and Stoica, I.: Tune: A research platform for distributed model selection and training, arXiv preprint, https://arxiv.org/abs/1807.05118 (last access: 9 March 2022), 2018.

Loshchilov, I. and Hutter, F.: Decoupled weight decay regularization, arXiv preprint, https://arxiv.org/abs/1711.05101v3 (last access: 9 March 2022), 2017.

Luojus, K., Pulliainen, J., Takala, M., Derksen, C., Rott, H., Nagler, T., Solberg, R., Wiesmann, A., Metsamaki, S., Malnes, E., and Bojkov, B.: Investigating the feasibility of the GlobSnow snow water equivalent data for climate research purposes, in: 2010 IEEE International Geoscience and Remote Sensing Symposium, pp. 4851–4853, IEEE, Honolulu, HI, USA, https://doi.org/10.1109/IGARSS.2010.5741987, 2010.

Luojus, K., Pulliainen, J., Takala, M., Lemmetyinen, J., Kangwa, M., Eskelinen, M., Metsämäki, S., Solberg, R., Salberg, A.-B., Bippus, G., Ripper, E., Nagler, T., Derksen, C., Wiesmann, A., Wunderle, S., Hüsler, F., Fontana, F., and Foppa, N.: GlobSnow-2 Final Report – European space agency study contract report, http://www.globsnow.info/docs/GlobSnow_2_Final_Report_release.pdf (last access: 3 March 2021), 2014.

McLaughlin, D.: An integrated approach to hydrologic data assimilation: interpolation, smoothing, and filtering, Adv. Water Resour., 25, 1275–1286, https://doi.org/10.1016/S0309-1708(02)00055-6, 2002.

Moradkhani, H., Sorooshian, S., Gupta, H. V., and Houser, P. R.: Dual state–parameter estimation of hydrological models using ensemble Kalman filter, Adv. Water Resour., 28, 135–147, https://doi.org/10.1016/j.advwatres.2004.09.002, 2005.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I – A discussion of principles, J. Hydrol., 10, 282–290, https://doi.org/10.1016/0022-1694(70)90255-6, 1970.

Nicholson, S. E.: Dryland Climatology, Cambridge University Press, Cambridge, UK, https://doi.org/10.1017/CBO9780511973840, 2011.

Pachauri, R. K., Allen, M. R., Barros, V. R., Broome, J., Cramer, W., Christ, R., Church, J. A., Clarke, L., Dahe, Q., Dasgupta, P., Dubash, N. K., Edenhofer, O., Elgizouli, I., Field, C. B., Forster, P., Friedlingstein, P., Fuglestvedt, J., Gomez-Echeverri, L., Hallegatte, S., Hegerl, G., Howden, M., Jiang, K., Jimenez Cisneroz, B., Kattsov, V., Lee, H., Mach, K. J., Marotzke, J., Mastrandrea, M. D., Meyer, L., Minx, J., Mulugetta, Y., O'Brien, K., Oppenheimer, M., Pereira, J. J., Pichs-Madruga, R., Plattner, G.-K., Pörtner, H.-O., Power, S. B., Preston, B., Ravindranath, N. H., Reisinger, A., Riahi, K., Rusticucci, M., Scholes, R., Seyboth, K., Sokona, Y., Stavins, R., Stocker, T. F., Tschakert, P., van Vuuren, D., and van Ypserle, J.-P.: Climate change 2014: synthesis report. Contribution of Working Groups I, II and III to the fifth assessment report of the Intergovernmental Panel on Climate Change, IPCC, ISBN 978-92-9169-143-2, 2014.

Panahi, M. and Behrangi, A.: Comparative analysis of snowfall accumulation and gauge undercatch correction factors from diverse data sets: In situ, satellite, and reanalysis, Asia-Pac. J. Atmos. Sci., 56, 1–14, https://doi.org/10.1007/s13143-019-00161-6, 2019.

Papagiannopoulou, C., Miralles, D. G., Demuzere, M., Verhoest, N. E. C., and Waegeman, W.: Global hydro-climatic biomes identified via multitask learning, Geosci. Model Dev., 11, 4139–4153, https://doi.org/10.5194/gmd-11-4139-2018, 2018.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library, in: Advances in Neural Information Processing Systems, vol. 32, Curran Associates, Inc., https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf (last access: 9 March 2022), 2019.

Python Core Team: Python: A dynamic, open source programming language, Python Software Foundation, https://www.python.org/, 9 March 2022.

Rangelova, E., Van der Wal, W., Braun, A., Sideris, M., and Wu, P.: Analysis of Gravity Recovery and Climate Experiment time-variable mass redistribution signals over North America by means of principal component analysis, J. Geophys. Res.-Earth, 112, F03002, https://doi.org/10.1029/2006JF000615, 2007.

Rasp, S., Pritchard, M. S., and Gentine, P.: Deep learning to represent subgrid processes in climate models, P. Natl. Acad. Sci. USA, 115, 9684–9689, https://doi.org/10.1073/pnas.1810286115, 2018.

Reichle, R. H.: Data assimilation methods in the Earth sciences, Adv. Water Resour., 31, 1411–1418, https://doi.org/10.1016/j.advwatres.2008.01.001, 2008.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat: Deep learning and process understanding for data-driven Earth system science, Nature, 566, 195, https://doi.org/10.1038/s41586-019-0912-1, 2019.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., and Dormann, C. F.: Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure, Ecography, 40, 913–929, https://doi.org/10.1111/ecog.02881, 2017.

Rodell, M., Famiglietti, J., Wiese, D., Reager, J., Beaudoing, H., Landerer, F. W., and Lo, M.-H.: Emerging trends in global freshwater availability, Nature, 557, 651–659, https://doi.org/10.1038/s41586-018-0123-1, 2018.

Scanlon, B. R., Zhang, Z., Rateb, A., Sun, A., Wiese, D., Save, H., Beaudoing, H., Lo, M. H., Müller-Schmied, H., Döll, P., van Beek, R., Swenson, S., Lawrence, D., Croteau, M., and Reedy, R. C.: Tracking seasonal fluctuations in land water storage using global models and GRACE satellites, Geophys. Res. Lett., 46, 5254–5264, https://doi.org/10.1029/2018GL081836, 2019.

Scanlon, B. R., Zhang, Z., Save, H., Wiese, D. N., Landerer, F. W., Long, D., Longuevergne, L., and Chen, J.: Global evaluation of new GRACE mascon products for hydrologic applications, Water Resour. Res., 52, 9412–9429, https://doi.org/10.1002/2016WR019494, 2016.

Schellekens, J., Dutra, E., Martínez-de la Torre, A., Balsamo, G., van Dijk, A., Sperna Weiland, F., Minvielle, M., Calvet, J.-C., Decharme, B., Eisner, S., Fink, G., Flörke, M., Peßenteiner, S., van Beek, R., Polcher, J., Beck, H., Orth, R., Calton, B., Burke, S., Dorigo, W., and Weedon, G. P.: A global water resources ensemble of hydrological models: the eartH2Observe Tier-1 dataset, Earth Syst. Sci. Data, 9, 389–413, https://doi.org/10.5194/essd-9-389-2017, 2017.

Schwingshackl, C., Hirschi, M., and Seneviratne, S. I.: Quantifying spatiotemporal variations of soil moisture control on surface energy balance and near-surface air temperature, J. Climate, 30, 7105–7124, https://doi.org/10.1175/JCLI-D-16-0727.1, 2017.

Seneviratne, S. I., Corti, T., Davin, E. L., Hirschi, M., Jaeger, E. B., Lehner, I., Orlowsky, B., and Teuling, A. J.: Investigating soil moisture–climate interactions in a changing climate: A review, Earth-Sci. Rev., 99, 125–161, https://doi.org/10.1016/j.earscirev.2010.02.004, 2010.

Shen, C., Laloy, E., Elshorbagy, A., Albert, A., Bales, J., Chang, F.-J., Ganguly, S., Hsu, K.-L., Kifer, D., Fang, Z., Fang, K., Li, D., Li, X., and Tsai, W.-P.: HESS Opinions: Incubating deep-learning-powered hydrologic science advances as a community, Hydrol. Earth Syst. Sci., 22, 5639–5656, https://doi.org/10.5194/hess-22-5639-2018, 2018.

Sun, L., Seidou, O., Nistor, I., and Liu, K.: Review of the Kalman-type hydrological data assimilation, Hydrolog. Sci. J., 61, 2348–2366, https://doi.org/10.1080/02626667.2015.1127376, 2016.

Swenson, S., Famiglietti, J., Basara, J., and Wahr, J.: Estimating profile soil moisture and groundwater variations using GRACE and Oklahoma Mesonet soil moisture data, Water Resour. Res., 44, W01413, https://doi.org/10.1029/2007WR006057, 2008.

Sylla, M., Giorgi, F., Coppola, E., and Mariotti, L.: Uncertainties in daily rainfall over Africa: assessment of gridded observation products and evaluation of a regional climate model simulation, Int. J. Climatol., 33, 1805–1817, https://doi.org/10.1002/joc.3551, 2013.

Takala, M., Luojus, K., Pulliainen, J., Derksen, C., Lemmetyinen, J., Kärnä, J.-P., Koskinen, J., and Bojkov, B.: Estimating northern hemisphere snow water equivalent for climate research through assimilation of space-borne radiometer data and ground-based measurements, Remote Sens. Environ., 115, 3517–3529, https://doi.org/10.1016/j.rse.2011.08.014, 2011.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An overview of CMIP5 and the experiment design, B. Am. Meteorol. Soc., 93, 485–498, https://doi.org/10.1175/BAMS-D-11-00094.1, 2012.

Tootchi, A., Jost, A., and Ducharne, A.: Multi-source global wetland maps combining surface water imagery and groundwater constraints, Earth Syst. Sci. Data, 11, 189–220, https://doi.org/10.5194/essd-11-189-2019, 2019.

Tramontana, G., Jung, M., Schwalm, C. R., Ichii, K., Camps-Valls, G., Ráduly, B., Reichstein, M., Arain, M. A., Cescatti, A., Kiely, G., Merbold, L., Serrano-Ortiz, P., Sickert, S., Wolf, S., and Papale, D.: Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms, Biogeosciences, 13, 4291–4313, https://doi.org/10.5194/bg-13-4291-2016, 2016.

Trautmann, T., Koirala, S., Carvalhais, N., Eicker, A., Fink, M., Niemann, C., and Jung, M.: Understanding terrestrial water storage variations in northern latitudes across scales, Hydrol. Earth Syst. Sci., 22, 4061–4082, https://doi.org/10.5194/hess-22-4061-2018, 2018.

Trautmann, T., Koirala, S., Carvalhais, N., Güntner, A., and Jung, M.: The importance of vegetation in understanding terrestrial water storage variations, Hydrol. Earth Syst. Sci., 26, 1089–1109, https://doi.org/10.5194/hess-26-1089-2022, 2022.

Van Beek, L., Wada, Y., and Bierkens, M. F.: Global monthly water stress: 1. Water balance and water availability, Water Resour. Res., 47, W07517, https://doi.org/10.1029/2010WR009792, 2011.

Van Der Knijff, J., Younis, J., and De Roo, A.: LISFLOOD: a GIS-based distributed model for river basin scale water balance and flood simulation, Int. J. Geogr. Inf. Sci., 24, 189–212, https://doi.org/10.1080/13658810802549154, 2010.

Van Dijk, A. and Warren, G.: The Australian water resources assess- ment system, version 0.5, 3, http://www.clw.csiro.au/publications/waterforahealthycountry/2010/wfhc-awras-evaluation-against-observations.pdf (last access: 3 March 2021), 2010.

van Dijk, A. I. J. M., Renzullo, L. J., Wada, Y., and Tregoning, P.: A global water cycle reanalysis (2003–2012) merging satellite gravimetry and altimetry observations with a hydrological multi-model ensemble, Hydrol. Earth Syst. Sci., 18, 2955–2973, https://doi.org/10.5194/hess-18-2955-2014, 2014.

Viovy, N.: CRUNCEP version 7-atmospheric forcing data for the community land model, Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory, Boulder CO, USA, https://doi.org/10.5065/PZ8F-F017, 2018.

Wada, Y., Wisser, D., and Bierkens, M. F. P.: Global modeling of withdrawal, allocation and consumptive use of surface water and groundwater resources, Earth Syst. Dynam., 5, 15–40, https://doi.org/10.5194/esd-5-15-2014, 2014.

Wang, H. and Yeung, D.-Y.: A survey on Bayesian deep learning, ACM Comput. Surv., 53, 1–37, https://doi.org/10.1145/3409383, 2020.

Watkins, M. M., Wiese, D. N., Yuan, D.-N., Boening, C., and Landerer, F. W.: Improved methods for observing Earth's time variable mass distribution with GRACE using spherical cap mascons, J. Geophys. Res.-Sol. Ea., 120, 2648–2671, https://doi.org/10.1002/2014JB011547, 2015.

Weedon, G. P., Balsamo, G., Bellouin, N., Gomes, S., Best, M. J., and Viterbo, P.: The WFDEI meteorological forcing data set: WATCH Forcing Data methodology applied to ERA-Interim reanalysis data, Water Resour. Res., 50, 7505–7514, https://doi.org/10.1002/2014WR015638, 2014.

Wielicki, B. A., Barkstrom, B. R., Harrison, E. F., Lee III, R. B., Smith, G. L., and Cooper, J. E.: Clouds and the Earth's Radiant Energy System (CERES): An Earth Observing System Experiment, B. Am. Meteorol. Soc., 77, 853–868, https://doi.org/10.1175/1520-0477(1996)077<0853:CATERE>2.0.CO;2, 1996.

Wiese, D. N., Landerer, F. W., and Watkins, M. M.: Quantifying and reducing leakage errors in the JPL RL05M GRACE mascon solution, Water Resour. Res., 52, 7490–7502, https://doi.org/10.1002/2016WR019344, 2016.

Wiese, D. N., Yuan, D.-N., Boening, C., Landerer, F. W., and Watkins, M. M.: JPL GRACE Mascon Ocean, Ice, and Hydrology Equivalent Water Height Release 06 Coastal Resolution Improvement (CRI) Filtered, PO.DAAC, CA, USA, https://doi.org/10.5067/TEMSC-3MJC6, 2018.

Yang, Y., Donohue, R. J., and McVicar, T. R.: Global estimation of effective plant rooting depth: Implications for hydrological modeling, Water Resour. Res., 52, 8260–8276, https://doi.org/10.1002/2016WR019392, 2016.

Zelazowski, P., Malhi, Y., Huntingford, C., Sitch, S., and Fisher, J. B.: Changes in the potential distribution of humid tropical forests on a warmer planet, Philos. T. Roy. Soc. A, 369, 137–160, https://doi.org/10.1098/rsta.2010.0238, 2011.

Zeng, N., Yoon, J.-H., Mariotti, A., and Swenson, S.: Variability of basin-scale terrestrial water storage from a PER water budget method: The Amazon and the Mississippi, J. Climate, 21, 248–265, https://doi.org/10.1175/2007JCLI1639.1, 2008.

5. Synthesis

In this chapter, I evaluate and discuss the contribution of this thesis in the context of the current scientific discourse. In Section 5.1, I summarize how the presented work contributes to answering the research questions raised in Section 1.5, followed by a broader reflection of the potential and limitations of the presented approaches in Section 5.2 and an outlook in Section 5.3.

5.1. Contribution of this thesis

5.1.1. [RQ1] Can recurrent neural networks learn global-scale ecosystem behavior?

The potential of recurrent neural network to learn ecosystem behavior under a wide range of conditions has been demonstrated specifically in Chapters 2 and 3, but is a key topic throughout this thesis. Although the applicability of RNNs to model Earth observation time series has been shown before (*e.g.*, Kratzert et al., 2018; Reddy and Prasad, 2018), global studies were previously not available. The LSTM is an easy-to-use model architecture that can serve as a benchmark for both machine learning and physically-based modeling. As shown in Chapter 2, an LSTM can almost perfectly emulate a physically-based model without spending much effort on model development. In Chapter 3, the approach was applied to real observations. I could show that an LSTM is able to perform at least on a par with a random forest model that received carefully designed, hand-crafted input features and was trained on a pixel level (Papagiannopoulou et al., 2017). In conclusion, I demonstrated the suitability of RNNs to emulate global ecosystem behavior using Earth observation data. The RNNs were able to represent complex temporal interactions (memory effects), which legitimizes the usage of such models in studies like S3 Kraft et al. (2020) or S4 Kraft et al. (2022).

5.1.2. [RQ2] Can dynamic memory effects in Earth observations be identified using explanatory approaches

The explanatory approach presented in Chapter 3 was a first effort to gather insights into memory effects using RNNs, which offers several advantages compared to traditional approaches like physically-based modeling, or more rigid data-driven approaches: A global model, such as used in Chapter 3 is less prone to overfitting, while training many local models with dozens of input

features as done in other studies is problematic. Furthermore, an LSTM can capture long-term dependencies in time series and is highly data adaptive. Finally, in comparison to physically-based models or shallow learning algorithms, the inductive biases are kept at a minimal level. This is especially useful in domains such as ecology and hydrology, where the process knowledge and observability are limited. However, the data adaptivity of the LSTMs comes at the cost of limited control and interpretability, and thus, our method is considered as a complementary extension to existing approaches. Explainable machine learning should be used with care (Rudin, 2019) and permutation-based approaches to quantify feature importance remain problematic. This is especially the case when the input features are collinear or autocorrelated (Hooker and Mentch, 2019), a constraint we have with environmental data (Roberts et al., 2017). Further experiments with synthetic data could shed light on this issue. With awareness of these limitations in mind, a purely data-driven and assumption-free assessment of memory effects is a novelty and offers an alternative perspective on a hard problem in ecology and hydrology.

5.1.3. [RQ3] What is the promise of global-scale hybrid modeling and what are its challenges and opportunities?

The explicit incorporation of physical knowledge into machine learning models has great potential for the Earth sciences. While current physically-based models struggle to make use of the richness and quantity of Earth observation data, hybrid models offer an alternative pathway to incorporate diverse data streams while being—to a certain degree—physically consistent and interpretable. Within this thesis, I demonstrated the feasibility of representing complex land surface processes constrained through multiple Earth observation products in Chapter 4. These studies represent a first step towards global-scale hybrid modeling that will have an impact on future efforts to improve numerical representations of Earth system processes. This goes beyond the domain of hydrological and vegetation modeling, and it is worthwhile to explore incorporating deep neural networks into Earth system models. Another contribution of this work is the application of *dynamic* hybrid modeling, which allows accounting explicitly and implicitly for memory effects.

Although I showed that a hybrid model achieves improved data adaptivity, I do not argue that traditional, knowledge-driven approaches are a method of the past. Rather, insights from hybrid modeling can be used in physically-based modeling and vice-versa. It is preferable to represent well-understood processes with physically-based approaches whenever possible, as these approaches still offer better interpretability and control.

The hybrid approach may solve some of the problems involved in environmental modeling. Still, several methodological aspects are yet to be tackled. One of the major challenges is the problem of equifinality, which could be identified and then reduced in a targeted way by using uncertainty-aware models. Further challenges and issues are discussed in more detail in the next section.

5.2. Reflection and future prospects

In this section, I discuss three major aspects presented in this thesis in a broader scientific context:

Section 5.2.1 — Deep learning has been shown to be capable of approximating complex system behavior across research domains. I outline how this potential can be harvested for vegetation and hydrology modeling, specifically in the context of the challenges described in Section 1.2.

Section 5.2.2 — Physically-based modeling is still the main tool for running long-term predictions and scenarios due to a lack of physical consistency of machine learning models. I discuss approaches to combine data- and knowledge-driven modeling and contrasted them to hybrid modeling.

Section 5.2.3 — Hybrid models are, compared to machine learning models, partially interpretable. What is the prospect of this data-driven yet interpretable approach for ecosystem modeling? What are the main challenges?

5.2.1. Deep learning for ecosystem modeling

The capability of RNNs to learn global-scale environmental processes across heterogeneous conditions with minimal assumptions has been demonstrated in the Chapters 2 and 3. In the broader context, RNNs have been shown to be suitable to model Earth observation data in various tasks. For example, Kratzert et al. (2018) showed that an LSTM can compete with sophisticated expert models in catchment-level runoff modeling. Besnard et al. (2019) used LSTM to model ecosystem-atmosphere CO_2 fluxes, and Haider et al. (2019) used LSTMs to predict wheat production. RNNs have also been used for classification tasks, for example for crop identification (Rußwurm and Körner, 2017). Today, a wide range of deep learning–based model types exist that can deal with sequential data, each with its own strengths and weaknesses (Ang et al., 2020), but RNNs still are a reliable and easy-to-use architecture. In the following, I discuss how RNNs can tackle current challenges in ecological modeling and outline a selection of applications.

Tackling the challenges of vegetation and hydrological modeling

Uncertainties and biases in Earth observations affect both machine learning and physically-based models, but in different ways. Neural networks can deal well with noisy labels in general, given enough training data (*e.g.*, Rolnick et al., 2018) and that the model is regularized properly to avoid overfitting on the noise. Similarly, physically-based models can handle noisy labels, as they are also either tuned using optimization techniques or manually tweaked to match the average validation signal. Neural networks can handle (non-trivial) biases owed to their high data-adaptivity, given

that the systematic errors are present in the training and the unseen test dataset. Physically-based models are prone to observation biases, especially in the input data, as biases propagate through the model and directly affect the simulations and parameter estimates (Beck et al., 2016; Döll et al., 2003).

Further data limitations emerge from the sparseness of the observations and the aggregation into coarse temporal and spatial grids. Data aggregation leads to information loss and potentially introduces biases (Colin et al., 2018), but the unified data format facilitates data processing and handling. Working with data in their native resolution is still not very common, but model approaches in deep learning exist to handle non-uniform input and output data, subsumed under the term *data fusion* (Zhu et al., 2017). Some data fusion techniques can deal with specific challenges, such as combining spatial and temporal context by extracting features independently from different satellite products (Benedetti et al., 2018), or dealing with multi-resolution or irregularly sampled time series (Singh et al., 2019). These approaches have the potential to make use of raw data products with multifaceted features.

The complexity of ecosystem processes poses a major challenge in environmental modeling. This is arguably the main motivation for deep learning over physically-based approaches. Deep learning models can learn highly non-linear processes, and in the case of LSTMs, interactions across different temporal scales (Lipton et al., 2015). The black-box character allows feeding the model with informative input features even if their exact interaction with other variables is not well understood. Nevertheless, it is still beneficial to perform a careful feature selection due to the increase in training data that is required with high-dimensional input (*curse of dimensionality*, Verleysen and François, 2005) and an increased risk for learning non-causal relationships (*shortcut learning*, Geirhos et al., 2020).

Applications of LSTMs in large-scale environmental modeling

Deep learning models are commonly applied to either make predictions into the future (*e.g.*, Haider et al., 2019; Kratzert et al., 2018), or to create data products (*e.g.*, Contractor and Roughan, 2021; Rußwurm and Körner, 2017). Such applications are being developed and improved constantly. But how can large-scale ecological modeling benefit from this potential?

Due to the low prior knowledge requirements and the ability to learn temporal interactions across time scales, RNNs can serve as a benchmark for physically-based models. In physically-based modeling, it often remains unclear whether the model itself or data deficiencies are the limiting factors. An RNN fed with the same data may help to identify model weaknesses and to identify regions where data quality is a limiting factor. Without adding physical constraints, such as presented in Chapter 4, however, an RNN can compensate data biases, which needs to be considered. The usage of RNNs or deep learning models in general as a benchmark is—to my knowledge—not common practice in physically-based modeling. LSTMs can also be used for

data cleaning or enhancement to provide improved products for physically-based modeling, for example, for gap-filling of environmental time series (*e.g.*, Contractor and Roughan, 2021; Huang and Hsieh, 2020).

Another application of RNNs in large-scale ecological modeling is the emulation of existing physically-based models. Neural networks are often computationally more efficient than physically-based models (*e.g.*, Krasnopolsky, 2020; Rasp et al., 2018). A deep learning–based emulator can serve as a flexible and fast tool for testing hypotheses or can replace certain computationally expensive parts of a model. Deep neural networks could further be used to simulate specific processes within a physically-based model. Land-surface models, for example, rely on estimates of future leaf area index (LAI), a measure of vegetation density, which is commonly simulated using dynamic global vegetation models (DGVMs). However, DGVMs still show substantial biases in their LAI projections (Murray-Tortarolo et al., 2013). It is worthwhile to test the application of LSTMs or other machine learning approaches for this purpose.

As previously discussed, a major issue with machine learning modeling is the lack of interpretability. In Chapter 3, I demonstrated an approach to identify memory effects based on permutation-based explanations to gather system understanding. Further approaches, like feature visualization and clustering have been proposed to interpret LSTMs (Pérez-Suay et al., 2020). The field of explainable machine learning is evolving quickly, and many new approaches are in development (Arrieta et al., 2020; Gunning et al., 2019; Langer et al., 2021; Molnar, 2019). Explainable approaches can be used to gather insights into ecosystem functioning, but further research is needed to explore and discuss the applicability of explanations to the Earth sciences.

This—presumably incomplete—list of applications justifies a broader investigation of the capabilities of RNNs and other sequential models in the context of large-scale environmental modeling. Some open questions are whether LSTMs can maintain and access information across hundreds or thousands of time steps, and to what extent they are suitable to not only represent coarse dynamics (*e.g.*, seasonal signals), but also the—usually more interesting—anomalous, small-scale system behavior. Further research regarding the explainability of LSTMs for system understanding is needed beyond the proof-of-concept provided here.

5.2.2. Integrating prior knowledge and observations

I discussed the advantages of deep learning approaches, in particular LSTMs, for modeling ecosystem processes. As discussed within this thesis (Section 1.3 and Chapter 4), and more broadly in Reichstein et al. (2019) and Camps-Valls et al. (2021), plain machine learning models lack physical consistency. This can lead to implausible and non-robust results caused by observational biases or out-of-sample extrapolation (Camps-Valls et al., 2020). Recently, approaches to make data-driven models physically (more) consistent have been introduced, mainly with the goal to improve the predictability of Earth system processes. These approaches can be subsumed under

the term *physics informed (*or *guided) machine learning*. Adding physical knowledge to a machine learning model can be seen as a regularization technique, but more specifically, it reduced the hypothesis space by introducing inductive biases. From the range of current methods that combine physical and data-driven modeling, I want to highlight three general approaches and contrast them to hybrid modeling: regularization via loss functions (soft constraints), mass conserving neural networks (hard constraints), and data assimilation.

Physics-based loss functions

A high-level approach for introducing physical knowledge to machine learning models is through physics-based loss functions. The approach introduces soft constraints by penalizing physically inconsistent results, but inconsistent results are still possible. An illustrative example is provided in Karpatne et al. (2018): The authors used physical knowledge to predict lake temperature profiles: Water density increases with depth, and temperature is closely related to water density. If the density at a depth d_1 is larger than the density at depth $d_2 > d_1$, the physical law of increasing density with increasing depth is violated, and thus, this solution is penalized in the loss function. A side-product of the regularization is an additional means to assess model consistency: When the model is applied to new data, the regularization term can be used to diagnose physical inconsistencies, even if labels are not available. Similarly, we could add qualitative knowledge, for example, about the smoothness of a target time series or spatial field (which can also be motivated by physics) to penalize non-plausible solutions.

Physics-based loss functions are a relatively cheap and flexible way to introduce physical knowledge to machine learning models. Soft constraints can be set on top of existing gradient-based machine learning models without architectural changes. The soft constraints can, however, not assert hard physical constraints and do not provide additional insights exceeding the aforementioned consistency checks. In contrast, hybrid models can enforce hard physical constraints, such as conservation of mass, and provide additional insights via latent variables and coefficients. Hybrid modeling, however, requires in-depth expert knowledge and may impose wrong or incomplete prior system understanding. When aiming for improved out-of-sample prediction, the benefit of using hybrid modeling over physics-based loss functions needs to be investigated further and is certainly problem-specific. If model interpretation is required—either for system understanding or for an in-depth assessment of the model reliability—hybrid modeling is the method of choice.

Mass and energy conservation

In the past years, several approaches have been proposed to assert conservation laws in neural networks, for example, for energy conservation (Zhong et al., 2021). These approaches assume closed physical systems and are, thus, too restrictive for many applications in environmental

modeling. A less restrictive approach has been proposed by Hoedt et al. (2021): the mass conserving LSTM (MC-LSTM). The MC-LSTM allows differentiating between mass inputs, a quantity that must be conserved, and auxiliary inputs, which are used to control the mass distribution within and withdrawal from the system. The authors showed that the MC-LSTM is applicable for river rainfall-runoff modeling: While the MC-LSTM did not outperform non-mass conserving architectures in general, it performed best in modeling extreme events, which is, especially in runoff modeling but also in ecology, challenging and of high interest.

The MC-LSTM is more restrictive than a basic LSTM. It uses physical knowledge to add an inductive bias, which can increase the robustness and generalizability. Compared to a hybrid approach as used in Chapter 4, the MC-LSTM is less restrictive, which may be an advantage in certain scenarios. The MC-LSTM is not bound by potentially wrong or simplistic physical constraints as is the case in hybrid modeling. Still, the MC-LSTM approach faces similar problems as a hybrid model when applied to real-world problems, as not all mass inputs or outputs may be observable. In rainfall-runoff modeling, for example, evapotranspiration is a key process of water withdrawal, but high-quality measurements are not available at the catchment level. Thus, the water balance is not closed if evapotranspiration is ignored. Hoedt et al. (2021) introduced a "trash cell" that can discharge exceeding mass to account for fluxes that could not be observed, which violates conservation of mass. For such scenarios, where observations are missing or uncertain, it may be an advantage to explicitly encode physical knowledge, *i.e.*, to use hybrid modeling.

Neural networks with build-in conservation laws can be seen as a special case of hybrid modeling where the physical constraints are encoded into a general-purpose model architecture. While this approach allows using the same architecture for different use-cases—*e.g.*, rainfall-runoff or traffic modeling, as in Hoedt et al. (2021)—, the interpretability is very limited. In contrast to the hybrid hydrological model presented in S4 Kraft et al. (2022), the MC-LSTM does not provide physically interpretable coefficients or latent variables.

Data assimilation

Data assimilation combines simulations from physically-based models with observations to generate optimal estimates of geophysical states (Reichle, 2008). Data assimilation uses advanced Kalman filtering techniques to find the most likely system state, given a state-generating model, a transfer function from the states to the observations, and the observations (Evensen, 2009). The statistical framework of data assimilation allows quantifying uncertainties in the state and output estimates by propagating data and model uncertainties. Data assimilation is broadly applied in atmospheric and oceanic modeling (Carrassi et al., 2018), for example to provide initial conditions in meteorological forecasts (Huang et al., 2009) or to provide reanalysis datasets of past geophysical states (*e.g.*, Hersbach et al., 2020).

Data assimilation is especially useful if a system can be well described with physically-based

approaches, such as in atmosphere modeling. The inclusion of model errors remains a major challenge, and commonly, the model is assumed to be perfect ("perfect model assumption"). Approaches that account for model errors exist (Howes et al., 2017), but such assimilation systems are intended to quantify the errors rather than to reduce them by updating the model accordingly. The joint assimilation of states and parameters estimation in so-called *dual stateparameter estimation* has been proposed, but seems to be still rather experimental (*e.g.*, Moradkhani et al., 2005). In addition, the parameters are updated based on data but not learned as a function of other variables. There is no pure inference or generalization to unseen conditions as in hybrid modeling. Nevertheless, the quantification of uncertainties within a well-defined framework remains a striking advantage of data assimilation. The field of data assimilation is evolving quickly, especially due to the growing amounts of data and computational resources. In the future, machine learning and data assimilation may be integrated for data-adaptive, uncertainty-aware systems (Geer, 2021).

5.2.3. Challenges in hybrid modeling

So far, I have discussed the potential of deep learning approaches, specifically of LSTMs, for large-scale vegetation and hydrology modeling. I provided a brief overview of approaches for integrating knowledge- and data-driven modeling. Together with hybrid modeling, these methods populate the space between pure machine learning and physically-based modeling (Figure 5.1). Data assimilation allows quantifying uncertainties but still largely relies on physically-based modeling. Physics-guided machine learning, on the other side, allows including high-level soft or hard physical constraints, leading to flexible, yet not interpretable black-box models. Hybrid modeling probably is the most flexible approach as specific processes can be replaced with machine learning models. This can lead to rather rigid or rather flexible solutions, depending on the model architecture.

As claimed and demonstrated before, hybrid models are partially interpretable. While the machine learning algorithms themselfes remain non-interpretable, they yield interpretable quantities. However, these estimates may be uncertain or biased. Furthermore, questions regarding the generalizability and best practices in model development need investigation. In the following, I highlight the current challenges in hybrid modeling and outline methodological pathways to improve model reliability and robustness.

Quantification of uncertainties

The interpretability of latent variables and coefficients in a hybrid model can be undermined by a major issue: equifinality. If the solution found by optimization is not unique, the interpretation of the respective quantities is problematic (Beven and Freer, 2001). For illustration, I use a simple hybrid model of snow mass estimation: Instead of modeling snow mass S as a function of air

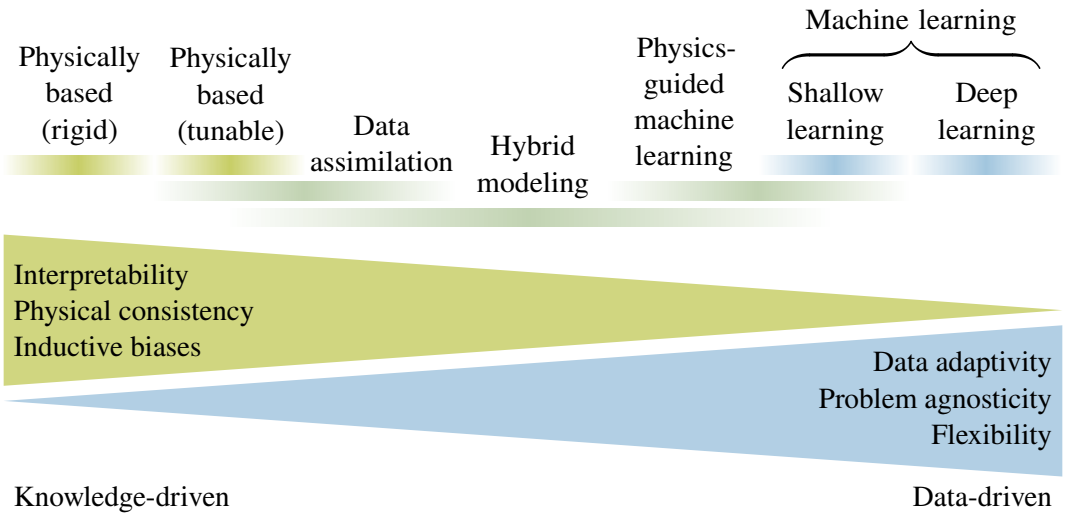


Figure 5.1.: Revisiting Figure 1.6. Different approaches attempt to close the gap between knowledge- and data-driven modeling.

temperature T_{air} and precipitation p directly, we estimate snow fall s_{fall} and melt s_{melt} separately, using f_{fall} and f_{melt}, respectively, both machine learning models. Now, we train a hybrid model $S_t = S_{t-1} + s_{\text{fall},t} - s_{\text{melt},t} = f_{\text{fall}}(T_{\text{air},t}, p_t) - f_{\text{melt}}(T_{\text{air},t}, p_t)$. It is obvious that infinite solutions exist, and we would draw wrong conclusions if we were to interpret the quantities of snow fall and melt. So, how can we deal with the problem of equifinality?

The first part of the answer is to further constrain the problem using prior knowledge or additional data. We know, for example, that snow melt is only happening above the freezing point and snow fall below. Further, we could limit $s_{\text{melt}} \leq S$ to avoid negative values, and $s_{\text{fall}} \leq p$ to assert that snowfall is not exceeding precipitation. These constraints are physically flawless, but we would likely see that the observations themselves contain biases that violate physical laws, possibly due to measurement errors or masking (*e.g.*, the average daily temperature may be negative, but positive temperatures during the day still caused snow melt). Implementing physical knowledge and still not constraining the solution too strictly towards prior knowledge is part of model development.

In the process of model development, the ability to identify uncertainties would be highly beneficial and facilitate the reduction of equifinalities. This is the second part of the answer: Identification of model uncertainties in the estimated physical parameters, coefficients, and variables is *the* key challenge in hybrid modeling. Several approaches exist to quantify uncertainties in neural networks, but it remains unclear whether these methods are directly applicable to hybrid modeling. A straight-forward method to quantify uncertainty in neural networks is the Monte Carlo dropout method (Gal and Ghahramani, 2016), where random nodes of the network are deactivated during training and inference. In inference, an uncertainty estimate is achieved through repeated forward

runs with dropout, allowing to quantify uncertainties in the learned physical coefficients and the downstream responses. Another approach worthwhile exploring is based on conditional generative adversarial models (cGANs, Goodfellow et al., 2020; Mirza and Osindero, 2014): Compared to the non-hybrid setting, where the labels are directly simulated conditioned on some factors, we could simulate the latent variables and coefficients and plug them into a physically-based model.

Other approaches exist to quantify uncertainties in deep neural networks (*e.g.*, Abdar et al., 2021). How these methods behave in a hybrid setting and whether they are sufficient to identify equifinalities needs investigation.

Generalizability

While S3 Kraft et al. (2020) was a proof-of-concept that demonstrated the applicability of hybrid modeling on the global scale, S4 Kraft et al. (2022) focused on the interpretability of the coefficients and latent variables. A next step in hybrid modeling is the investigation of the generalizability into unseen conditions, for example, for long-term predictions and scenarios. To this date, it is not clear under which conditions a more rigid, physically-based model is more reliable for extrapolation, and when a less restrictive model, as, for example, mass or energy-conserving methods (Section 5.2.2), perform better. Such questions need to be answered domain-specific, since the data and knowledge constraints vary between different applications.

Model development

Deep neural networks of various forms have been studied for decades. Many aspects of model development, optimization, and evaluation have been investigated systematically, or at least heuristics are known from previous experiments. For example, so-called *skip connections* enabled the training of very deep architectures (He et al., 2015) in certain domains. It is important to better understand the role and impact of such architectural choices, for example, by performing experiments (Orhan and Pitkow, 2018) or visualization (Li et al., 2018). Next to architectural choices, the optimization process is a key ingredient for deep learning and the choice of the optimizer (*e.g.*, Schmidt et al., 2021; Schneider et al., 2019) or the regularization technique (*e.g.*, Kukačka et al., 2017; Nusrat and Jang, 2018; Zaremba et al., 2015) can have a substantial impact on the model performance and robustness.

Due to the novelty of hybrid modeling, best practices and theoretically backed rules for model development and optimization do not exist. Thus, it is crucial to investigate the impact of the model structure and training procedure on the model performance and simulated variables and coefficients, especially the interplay of the physical parameters and the ML parameters, as well as the impact of the physical constraints on the model optimization. Similar to the efforts in deep learning, a combination of experiments, theoretical work, and visualizations can help to build

better and more reliable models. This process can be supported with approaches from explainable machine learning. Examples are visualizing loss landscapes (Li et al., 2018) or challenging model robustness with adversarial (yet physically consistent) examples (Goodfellow et al., 2015).

Combining diverse data sources

In contrast to physically-based models, hybrid models can be fed with additional datasets that are affecting the processes we seek to represent without explicit process knowledge. This may include auxiliary variables that interact with the environmental processes, even if we are not able to describe a process physically. Methods from data fusion (Zhu et al., 2017) could be applied to combine datasets with different resolutions to make full use of the information present in the observations.

An issue that needs further investigation is the weighting of losses in the context of multitask learning, as done in Chapter 4. The automatic uncertainty weighting (Kendall et al., 2018) was robust in the presented studies, but a more detailed investigation of the approach and a comparison to other methods is needed.

5.3. Outlook

As demonstrated and discussed within this thesis, deep learning methods have the potential to advance the predictability and understandability of vegetation and hydrology at large scales. The limited interpretability, however, hampers model trust and interpretability. Thus, methods from explainable machine learning should be considered to support the development of more robust and trustworthy models and, ultimately, to improve system understanding. Still, the lack of physical consistency and explainability is a major limitation of conventional deep learning approaches, which motivates the usage of physics-guided machine learning and hybrid modeling.

Hybrid modeling may find broad application in Earth sciences in the future. Besides the application demonstrated here, other domains can benefit from the data-adaptive paradigm. A goal for the near future is the incorporation of the approach in Earth system models, which are the primary tool to inform society and decision-makers about the long-term impact of human actions on climate and environment. A step in this direction will be taken by coupling the carbon and water cycle in an uncertainty-aware hybrid model that builds upon the work presented in Chapter 4, for which I acquired funding from the German Federal Ministry of Economics and Technology. At the same time, hybrid modeling can be used in domain-specific applications, for example, in rainfall-runoff modeling, modeling of ecosystem-atmosphere fluxes, or the prediction of forest fires. In parallel to application-based experiments, there is a need for methodological studies that investigate and overcome the current challenges (*e.g.*, uncertainty quantification).

List of Figures

Bibliography

Abdar, M., F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, V. Makarenkov, and S. Nahavandi (2021). "A Review of Uncertainty Quantification in Deep Learning: Techniques, Applications and Challenges." In: *Information Fusion* 76, pp. 243–297. ISSN: 15662535. DOI: 10.1016/j.inffus.2021.05.008. arXiv: 2011.06225.

Adadi, A. and M. Berrada (2018). "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)." In: *IEEE Access* 6, pp. 52138–52160. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2018.2870052.

Ang, J.-S., K.-W. Ng, and F.-F. Chua (2020). "Modeling Time Series Data with Deep Learning: A Review, Analysis, Evaluation and Future Trend." In: *2020 8th International Conference on Information Technology and Multimedia (ICIMU)*, pp. 32–37. DOI: 10.1109/ICIMU49871.2020.9243546.

Arrieta, A. B., N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera (2020). "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI." In: *Information Fusion* 58, pp. 82–115. ISSN: 1566-2535. DOI: 10.1016/j.inffus.2019.12.012.

Baldocchi, D., E. Falge, L. Gu, R. Olson, D. Hollinger, S. Running, P. Anthoni, C. Bernhofer, K. Davis, R. Evans, J. Fuentes, A. Goldstein, G. Katul, B. Law, X. Lee, Y. Malhi, T. Meyers, W. Munger, W. Oechel, K. T. P. U, K. Pilegaard, H. P. Schmid, R. Valentini, S. Verma, T. Vesala, K. Wilson, and S. Wofsy (2001). "FLUXNET: A New Tool to Study the Temporal and Spatial Variability of Ecosystem-Scale Carbon Dioxide, Water Vapor, and Energy Flux Densities." In: *Bulletin of the American Meteorological Society* 82.11, pp. 2415–2434. ISSN: 0003-0007, 1520-0477. DOI: 10.1175/1520-0477(2001)082<2415:FANTTS>2.3.CO;2.

Bard, Y. (1974). *Nonlinear Parameter Estimation*. New York: Academic Press. ISBN: 0-12-078250-2.

Baxter, J. (2000). "A Model of Inductive Bias Learning." In: *Journal of Artificial Intelligence Research* 12, pp. 149–198. ISSN: 1076-9757. DOI: 10.1613/jair.731.

Beck, H. E., A. I. J. M. van Dijk, A. de Roo, D. G. Miralles, T. R. McVicar, J. Schellekens, and L. A. Bruijnzeel (2016). "Global-Scale Regionalization of Hydrologic Model Parameters." In: *Water Resources Research* 52.5, pp. 3599–3622. ISSN: 1944-7973. DOI: 10.1002/2015WR018247.

Benedetti, P., D. Ienco, R. Gaetano, K. Ose, R. G. Pensa, and S. Dupuy (2018). "M3 Fusion: A Deep Learning Architecture for Multiscale Multimodal Multitemporal Satellite Data Fusion." In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 11.12, pp. 4939–4949. ISSN: 2151-1535. DOI: 10.1109/JSTARS.2018.2876357.

Bergen, K. J., P. A. Johnson, M. V. de Hoop, and G. C. Beroza (2019). "Machine Learning for Data-Driven Discovery in Solid Earth Geoscience." In: *Science* 363.6433. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aau0323.

Besnard, S., N. Carvalhais, M. A. Arain, A. Black, B. Brede, N. Buchmann, J. Chen, J. G. P. W. Clevers, L. P. Dutrieux, F. Gans, M. Herold, M. Jung, Y. Kosugi, A. Knohl, B. E. Law, E. Paul-Limoges, A. Lohila, L. Merbold, O. Roupsard, R. Valentini, S. Wolf, X. Zhang, and M. Reichstein (2019). "Memory Effects of Climate and Vegetation Affecting Net Ecosystem CO2 Fluxes in Global Forests." In: *PLOS ONE* 14.2, e0211510. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0211510.

Beven, K. and J. Freer (2001). "Equifinality, Data Assimilation, and Uncertainty Estimation in Mechanistic Modelling of Complex Environmental Systems Using the GLUE Methodology." In: *Journal of Hydrology* 249.1, pp. 11–29. ISSN: 0022-1694. DOI: 10.1016/S0022-1694(01)00421-8.

Bianchi, T. S. (2020). "The Evolution of Biogeochemistry: Revisited." In: *Biogeochemistry*. ISSN: 1573-515X. DOI: 10.1007/s10533-020-00708-0.

Biran, O. and C. Cotton (2017). "Explanation and Justification in Machine Learning: A Survey." In: *IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI),* p. 6.

Bonan, G. (2015). *Ecological Climatology: Concepts and Applications*. Third. Cambridge: Cambridge University Press. ISBN: 978-1-107-04377-0. DOI: 10.1017/CBO9781107339200.

Bonan, G. and S. C. Doney (2018). "Climate, Ecosystems, and Planetary Futures: The Challenge to Predict Life in Earth System Models." In: *Science* 359.6375. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aam8328.

Butts, M. B., J. T. Payne, M. Kristensen, and H. Madsen (2004). "An Evaluation of the Impact of Model Structure on Hydrological Modelling Uncertainty for Streamflow Simulation." In: *Journal of Hydrology*. The Distributed Model Intercomparison Project (DMIP) 298.1, pp. 242–266. ISSN: 0022-1694. DOI: 10.1016/j.jhydrol.2004.03.042.

Camps-Valls, G., D. H. Svendsen, J. Cortés-Andrés, Á. Moreno-Martínez, A. Pérez-Suay, J. Adsuara, I. Martín, M. Piles, J. Muñoz-Marí, and L. Martino (2020). "Living in the Physics and Machine Learning Interplay for Earth Observation." In: *arXiv:2010.09031 [physics, stat]*. arXiv: 2010.09031 [physics, stat].

Camps-Valls, G., D. Tuia, X. X. Zhu, and M. Reichstein (2021). *Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science and Geosciences*. 1st edition. Hoboken, NJ: Wiley. ISBN: 978-1-119-64614-3.

Carpenter, S. R., S. W. Chisholm, C. J. Krebs, D. W. Schindler, and R. F. Wright (1995). "Ecosystem Experiments." In: *Science*. DOI: 10.1126/science.269.5222.324.

Carrassi, A., M. Bocquet, L. Bertino, and G. Evensen (2018). "Data Assimilation in the Geosciences: An Overview of Methods, Issues, and Perspectives." In: *WIREs Climate Change* 9.5, e535. ISSN: 1757-7799. DOI: 10.1002/wcc.535.

Colin, B., M. Schmidt, S. Clifford, A. Woodley, and K. Mengersen (2018). "Influence of Spatial Aggregation on Prediction Accuracy of Green Vegetation Using Boosted Regression Trees." In: *Remote Sensing* 10.8, p. 1260. DOI: 10.3390/rs10081260.

Contractor, S. and M. Roughan (2021). "Efficacy of Feedforward and LSTM Neural Networks at Predicting and Gap Filling Coastal Ocean Timeseries: Oxygen, Nutrients, and Temperature." In: *Frontiers in Marine Science*. DOI: http://dx.doi.org/10.3389/fmars.2021.637759.

Csanád Csáji, B. (2001). "Approximation with Artificial Neural Networks." MA thesis. Hungary: Faculty of Sciences, Eötvös Loránd University.

de Bézenac, E., A. Pajot, and P. Gallinari (2019). "Deep Learning for Physical Processes: Incorporating Prior Scientific Knowledge." In: *Journal of Statistical Mechanics: Theory and Experiment* 2019.12, p. 124009. ISSN: 1742-5468. DOI: 10.1088/1742-5468/ab3195.

Döll, P., F. Kaspar, and B. Lehner (2003). "A Global Hydrological Model for Deriving Water Availability Indicators: Model Tuning and Validation." In: *Journal of Hydrology* 270.1, pp. 105–134. ISSN: 0022-1694. DOI: 10.1016/S0022-1694(02)00283-4.

Drake, J. (2018). *Introduction to Logic by Jess Drake*. 1st edition. ED-TECH PRESS. ISBN: 978-1-78882-358-6.

Elsayed, G. F., P. Ramachandran, J. Shlens, and S. Kornblith (2020). "Revisiting Spatial Invariance with Low-Rank Local Connectivity." In: *arXiv:2002.02959 [cs, stat]*. arXiv: 2002.02959 [cs, stat].

Engelhardt, I., J. G. D. Aguinaga, H. Mikat, C. Schüth, and R. Liedl (2014). "Complexity vs. Simplicity: Groundwater Model Ranking Using Information Criteria." In: *Groundwater* 52.4, pp. 573–583. ISSN: 1745-6584. DOI: 10.1111/gwat.12080.

Evensen, G. (2009). *Data Assimilation: The Ensemble Kalman Filter*. Springer Science & Business Media. ISBN: 978-3-642-03711-5.

Fisher, R. A. and C. D. Koven (2020). "Perspectives on the Future of Land Surface Models and the Challenges of Representing Complex Terrestrial Systems." In: *Journal of Advances in Modeling Earth Systems* 12.4, e2018MS001453. ISSN: 1942-2466. DOI: 10.1029/2018MS001453.

Friedlingstein, P., M. Meinshausen, V. K. Arora, C. D. Jones, A. Anav, S. K. Liddicoat, and R. Knutti (2014). "Uncertainties in CMIP5 Climate Projections Due to Carbon Cycle Feedbacks." In: *Journal of Climate* 27.2, pp. 511–526. ISSN: 0894-8755, 1520-0442. DOI: 10.1175/JCLI-D-12-00579.1.

Gal, Y. and Z. Ghahramani (2016). "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning." In: *International Conference on Machine Learning*. PMLR, pp. 1050–1059.

Geer, A. J. (2021). "Learning Earth System Models from Observations: Machine Learning or Data Assimilation?" In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 379.2194, p. 20200089. DOI: 10.1098/rsta.2020.0089.

Geirhos, R., J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann (2020). "Shortcut Learning in Deep Neural Networks." In: *Nature Machine Intelligence* 2.11, pp. 665–673. ISSN: 2522-5839. DOI: 10.1038/s42256-020-00257-z. arXiv: 2004.07780.

Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep Learning*. Illustrated edition. Cambridge, Massachusetts: The MIT Press. ISBN: 978-0-262-03561-3.

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio (2020). "Generative Adversarial Networks." In: *Communications of the ACM* 63.11, pp. 139–144. ISSN: 0001-0782. DOI: 10.1145/3422622.

Goodfellow, I. J., J. Shlens, and C. Szegedy (2015). "Explaining and Harnessing Adversarial Examples." In: *arXiv:1412.6572 [cs, stat]*. arXiv: 1412.6572 [cs, stat].

Gunning, D., M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang (2019). "XAI—Explainable Artificial Intelligence." In: *Science Robotics* 4.37. ISSN: 2470-9476. DOI: 10.1126/scirobotics.aay7120.

Guo, H.-D., L. Zhang, and L.-W. Zhu (2015). "Earth Observation Big Data for Climate Change Research." In: *Advances in Climate Change Research*. Special Issue on Advances in Future Earth Research 6.2, pp. 108–117. ISSN: 1674-9278. DOI: 10.1016/j.accre.2015.09.007.

Gupta, H. V., M. P. Clark, J. A. Vrugt, G. Abramowitz, and M. Ye (2012). "Towards a Comprehensive Assessment of Model Structural Adequacy." In: *Water Resources Research* 48.8. ISSN: 1944-7973. DOI: 10.1029/2011WR011044.

Haddeland, I., D. B. Clark, W. Franssen, F. Ludwig, F. Voß, N. W. Arnell, N. Bertrand, M. Best, S. Folwell, D. Gerten, S. Gomes, S. N. Gosling, S. Hagemann, N. Hanasaki, R. Harding, J. Heinke, P. Kabat, S. Koirala, T. Oki, J. Polcher, T. Stacke, P. Viterbo, G. P. Weedon, and P. Yeh (2011). "Multimodel Estimate of the Global Terrestrial Water Balance: Setup and First Results." In: *Journal of Hydrometeorology* 12.5, pp. 869–884. ISSN: 1525-7541, 1525-755X. DOI: 10.1175/2011JHM1324.1.

Haider, S. A., S. R. Naqvi, T. Akram, G. A. Umar, A. Shahzad, M. R. Sial, S. Khaliq, and M. Kamran (2019). "LSTM Neural Network Based Forecasting Model for Wheat Production in Pakistan." In: *Agronomy* 9.2, p. 72. DOI: 10.3390/agronomy9020072.

Hall, C. A. S. and J. W. Day, eds. (1977). *Ecosystem Modeling in Theory and Practice: An Introduction with Case Histories*. 1st edition. New York: Wiley. ISBN: 978-0-471-34165-9.

Hantson, S., A. Arneth, S. P. Harrison, D. I. Kelley, I. C. Prentice, S. S. Rabin, S. Archibald, F. Mouillot, S. R. Arnold, P. Artaxo, D. Bachelet, P. Ciais, M. Forrest, P. Friedlingstein, T. Hickler, J. O. Kaplan, S. Kloster, W. Knorr, G. Lasslop, F. Li, S. Mangeon, J. R. Melton, A. Meyn, S. Sitch, A. Spessa, G. R. van der Werf, A. Voulgarakis, and C. Yue (2016). "The Status and Challenge of Global Fire Modelling." In: *Biogeosciences* 13.11, pp. 3359–3375. ISSN: 1726-4170. DOI: 10.5194/bg-13-3359-2016.

He, K., X. Zhang, S. Ren, and J. Sun (2015). "Deep Residual Learning for Image Recognition." In: *arXiv:1512.03385 [cs]*. arXiv: 1512.03385 [cs].

Heimann, M. and M. Reichstein (2008). "Terrestrial Ecosystem Carbon Dynamics and Climate Feedbacks." In: *Nature* 451.7176, pp. 289–292. ISSN: 1476-4687. DOI: 10.1038/nature06591.

Hersbach, H., B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, A. Simmons, C. Soci, S. Abdalla, X. Abellan, G. Balsamo, P. Bechtold, G. Biavati, J. Bidlot, M. Bonavita, G. D. Chiara, P. Dahlgren, D. Dee, M. Diamantakis, R. Dragani, J. Flemming, R. Forbes, M. Fuentes, A. Geer, L. Haimberger, S. Healy, R. J. Hogan, E. Hólm, M. Janisková, S. Keeley, P. Laloyaux, P. Lopez, C. Lupu, G. Radnoti, P. de Rosnay, I. Rozum, F. Vamborg, S. Villaume, and J.-N. Thépaut (2020). "The ERA5 Global Reanalysis." In: *Quarterly Journal of the Royal Meteorological Society* 146.730, pp. 1999–2049. ISSN: 1477-870X. DOI: 10.1002/qj.3803.

Hinton, G. and S. Roweis (2002). "Stochastic Neighbor Embedding." In: *Neural Information Processing Systems*. Vol. 15, pp. 857–864.

Hochreiter, S. and J. Schmidhuber (1997). "Long Short-Term Memory." In: *Neural Computation* 9.8, pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735.

Hoedt, P.-J., F. Kratzert, D. Klotz, C. Halmich, M. Holzleitner, G. Nearing, S. Hochreiter, and G. Klambauer (2021). "MC-LSTM: Mass-Conserving LSTM." In: *arXiv:2101.05186 [cs, stat]*. arXiv: 2101.05186 [cs, stat].

Hooker, G. and L. Mentch (2019). "Please Stop Permuting Features: An Explanation and Alternatives." In: *arXiv:1905.03151 [cs, stat]*. arXiv: 1905.03151 [cs, stat].

Hornik, K., M. Stinchcombe, and H. White (1989). "Multilayer Feedforward Networks Are Universal Approximators." In: *Neural Networks* 2.5, pp. 359–366. ISSN: 0893-6080. DOI: 10.1016/0893-6080(89)90020-8.

Howes, K. E., A. M. Fowler, and A. S. Lawless (2017). "Accounting for Model Error in Strong-Constraint 4D-Var Data Assimilation." In: *Quarterly Journal of the Royal Meteorological Society* 143.704, pp. 1227–1240. ISSN: 1477-870X. DOI: 10.1002/qj.2996.

Huang, I.-H. and C.-I. Hsieh (2020). "Gap-Filling of Surface Fluxes Using Machine Learning Algorithms in Various Ecosystems." In: *Water* 12.12, p. 3415. DOI: 10.3390/w12123415.

Huang, X.-Y., Q. Xiao, D. M. Barker, X. Zhang, J. Michalakes, W. Huang, T. Henderson, J. Bray, Y. Chen, Z. Ma, J. Dudhia, Y. Guo, X. Zhang, D.-J. Won, H.-C. Lin, and Y.-H. Kuo (2009). "Four-Dimensional Variational Data Assimilation for WRF: Formulation and Preliminary Results." In: *Monthly Weather Review* 137.1, pp. 299–314. ISSN: 1520-0493, 0027-0644. DOI: 10.1175/2008MWR2577.1.

Humphrey, V., J. Zscheischler, P. Ciais, L. Gudmundsson, S. Sitch, and S. I. Seneviratne (2018). "Sensitivity of Atmospheric CO2 Growth Rate to Observed Changes in Terrestrial Water Storage." In: *Nature* 560.7720, pp. 628–631. ISSN: 1476-4687. DOI: 10.1038/s41586-018-0424-4.

Jia, G., E. Shevliakova, P. Artaxo, N. De Noblet-Ducoudré, R. Houghton, J. House, K. Kitajima, C. Lennard, A. Popp, A. Sirin, R. Sukumar, and L. Verchot (2019). "Land-Climate Interactions." In: *Climate Change and Land (IPCC Special Report)*, pp. 131–247.

Jung, M., M. Reichstein, C. R. Schwalm, C. Huntingford, S. Sitch, A. Ahlström, A. Arneth, G. Camps-Valls, P. Ciais, P. Friedlingstein, F. Gans, K. Ichii, A. K. Jain, E. Kato, D. Papale, B. Poulter, B. Raduly, C. Rödenbeck, G. Tramontana, N. Viovy, Y.-P. Wang, U. Weber, S. Zaehle, and N. Zeng (2017). "Compensatory Water Effects Link Yearly Global Land CO2 Sink Changes to Temperature." In: *Nature* 541.7638, pp. 516–520. ISSN: 1476-4687. DOI: 10.1038/nature20780.

Jung, M., C. Schwalm, M. Migliavacca, S. Walther, G. Camps-Valls, S. Koirala, P. Anthoni, S. Besnard, P. Bodesheim, N. Carvalhais, F. Chevallier, F. Gans, D. S. Goll, V. Haverd, P. Köhler, K. Ichii, A. K. Jain, J. Liu, D. Lombardozzi, J. E. M. S. Nabel, J. A. Nelson, M. O'Sullivan, M. Pallandt, D. Papale, W. Peters, J. Pongratz, C. Rödenbeck, S. Sitch, G. Tramontana, A. Walker, U. Weber, and M. Reichstein (2020). "Scaling Carbon Fluxes from Eddy Covariance Sites to Globe: Synthesis and Evaluation of the FLUXCOM Approach." In: *Biogeosciences* 17.5, pp. 1343–1365. ISSN: 1726-4170. DOI: 10.5194/bg-17-1343-2020.

Karpatne, A., I. Ebert-Uphoff, S. Ravela, H. A. Babaie, and V. Kumar (2019). "Machine Learning for the Geosciences: Challenges and Opportunities." In: *IEEE Transactions on Knowledge and Data Engineering* 31.8, pp. 1544–1554. ISSN: 1558-2191. DOI: 10.1109/TKDE.2018.2861006.

Karpatne, A., W. Watkins, J. Read, and V. Kumar (2018). "Physics-Guided Neural Networks (PGNN): An Application in Lake Temperature Modeling." In: *arXiv:1710.11431 [physics, stat]*. arXiv: 1710.11431 [physics, stat].

Kawamiya, M., T. Hajima, K. Tachiiri, S. Watanabe, and T. Yokohata (2020). "Two Decades of Earth System Modeling with an Emphasis on Model for Interdisciplinary Research on Climate (MIROC)." In: *Progress in Earth and Planetary Science* 7.1, p. 64. ISSN: 2197-4284. DOI: 10.1186/s40645-020-00369-5.

Kendall, A., Y. Gal, and R. Cipolla (2018). "Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7482–7491.

Kraft, B., M. Jung, M. Körner, S. Koirala, and M. Reichstein (2022). "Towards hybrid modeling of the global hydrological cycle." In: *Hydrology and Earth System Sciences* 26.6, pp. 1579–1614. DOI: 10.5194/hess-26-1579-2022.

Kraft, B., M. Jung, M. Körner, and M. Reichstein (2020). "Hybrid Modeling: Fusion of a Deep Learning Approach and a Physics-Based Model for Global Hydrological Modeling." In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. Vol. XLIII-B2-2020. Copernicus GmbH, pp. 1537–1544. DOI: 10.5194/isprs-archives-XLIII-B2-2020-1537-2020.

Kraft, B., S. Besnard, and S. Koirala (2021). "Emulating Ecological Memory with Recurrent Neural Networks." In: *Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science and Geosciences*. Ed. by G. Camps-Valls, D. Tuia, X. X. Zhu, and M. Reichstein. 1st edition. Hoboken, NJ: Wiley & Sons. ISBN: 978-1-119-64614-3.

Kraft, B., M. Jung, M. Körner, C. Requena Mesa, J. Cortés, and M. Reichstein (2019). "Identifying Dynamic Memory Effects on Vegetation State Using Recurrent Neural Networks." In: *Frontiers in Big Data* 2. ISSN: 2624-909X. DOI: 10.3389/fdata.2019.00031.

Krasnopolsky, V. (2020). "Using Machine Learning for Model Physics: An Overview." In: *arXiv:2002.00416 [physics, stat]*. arXiv: 2002.00416 [physics, stat].

Kratzert, F., D. Klotz, C. Brenner, K. Schulz, and M. Herrnegger (2018). "Rainfall–Runoff Modelling Using Long Short-Term Memory (LSTM) Networks." In: *Hydrology and Earth System Sciences* 22.11, pp. 6005–6022. ISSN: 1027-5606. DOI: 10.5194/hess-22-6005-2018.

Kukačka, J., V. Golkov, and D. Cremers (2017). "Regularization for Deep Learning: A Taxonomy." In: *arXiv:1710.10686 [cs, stat]*. arXiv: 1710.10686 [cs, stat].

Langer, M., D. Oster, T. Speith, H. Hermanns, L. Kästner, E. Schmidt, A. Sesing, and K. Baum (2021). "What Do We Want from Explainable Artificial Intelligence (XAI)? – A Stakeholder Perspective on XAI and a Conceptual Model Guiding Interdisciplinary XAI Research." In: *Artificial Intelligence* 296, p. 103473. ISSN: 0004-3702. DOI: 10.1016/j.artint.2021.103473.

LeCun, Y., Y. Bengio, and G. Hinton (2015). "Deep Learning." In: *Nature* 521.7553, pp. 436–444. ISSN: 1476-4687. DOI: 10.1038/nature14539.

Li, H., Z. Xu, G. Taylor, C. Studer, and T. Goldstein (2018). "Visualizing the Loss Landscape of Neural Nets." In: *arXiv:1712.09913 [cs, stat]*. arXiv: 1712.09913 [cs, stat].

Libbrecht, K. G. (2005). "The Physics of Snow Crystals." In: *Reports on Progress in Physics* 68.4, pp. 855–895. ISSN: 0034-4885. DOI: 10.1088/0034-4885/68/4/R03.

Lipton, Z. C. (2018). "The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability Is Both Important and Slippery." In: *Queue* 16.3, pp. 31–57. ISSN: 1542-7730. DOI: 10.1145/3236386.3241340.

Lipton, Z. C., J. Berkowitz, and C. Elkan (2015). "A Critical Review of Recurrent Neural Networks for Sequence Learning." In: *arXiv:1506.00019 [cs]*. arXiv: 1506.00019 [cs].

Ma, L., Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson (2019). "Deep Learning in Remote Sensing Applications: A Meta-Analysis and Review." In: *ISPRS Journal of Photogrammetry and Remote Sensing* 152, pp. 166–177. ISSN: 0924-2716. DOI: 10.1016/j.isprsjprs.2019.04.015.

Miller, T. (2019). "Explanation in Artificial Intelligence: Insights from the Social Sciences." In: *Artificial Intelligence* 267, pp. 1–38. ISSN: 0004-3702. DOI: 10.1016/j.artint.2018.07.007.

Mirza, M. and S. Osindero (2014). "Conditional Generative Adversarial Nets." In: *arXiv:1411.1784 [cs, stat]*. arXiv: 1411.1784 [cs, stat].

Mitchell, T. M. (1980). *The Need for Biases in Learning Generalizations*. Tech. rep.

Mohanty, B. P., M. H. Cosh, V. Lakshmi, and C. Montzka (2017). "Soil Moisture Remote Sensing: State-of-the-Science." In: *Vadose Zone Journal* 16.1, vzj2016.10.0105. ISSN: 1539-1663. DOI: 10.2136/vzj2016.10.0105.

Molnar, C. (2019). *Interpretable Machine Learning*.

Moradkhani, H., S. Sorooshian, H. V. Gupta, and P. R. Houser (2005). "Dual State–Parameter Estimation of Hydrological Models Using Ensemble Kalman Filter." In: *Advances in Water Resources* 28.2, pp. 135–147. ISSN: 0309-1708. DOI: 10.1016/j.advwatres.2004.09.002.

Murray-Tortarolo, G., A. Anav, P. Friedlingstein, S. Sitch, S. Piao, Z. Zhu, B. Poulter, S. Zaehle, A. Ahlström, M. Lomas, S. Levis, N. Viovy, and N. Zeng (2013). "Evaluation of Land Surface Models in Reproducing Satellite-Derived LAI over the High-Latitude Northern Hemisphere. Part I: Uncoupled DGVMs." In: *Remote Sensing* 5.10, pp. 4819–4838. DOI: 10.3390/rs5104819.

Nusrat, I. and S.-B. Jang (2018). "A Comparison of Regularization Techniques in Deep Neural Networks." In: *Symmetry* 10.11, p. 648. DOI: 10.3390/sym10110648.

Ogle, K., J. J. Barber, G. A. Barron-Gafford, L. P. Bentley, J. M. Young, T. E. Huxman, M. E. Loik, and D. T. Tissue (2015). "Quantifying Ecological Memory in Plant and Ecosystem Processes." In: *Ecology Letters* 18.3, pp. 221–235. ISSN: 1461-0248. DOI: 10.1111/ele.12399.

O'Gorman, P. A. and J. G. Dwyer (2018). "Using Machine Learning to Parameterize Moist Convection: Potential for Modeling of Climate, Climate Change, and Extreme Events." In: *Journal of Advances in Modeling Earth Systems* 10.10, pp. 2548–2563. ISSN: 1942-2466. DOI: 10.1029/2018MS001351.

Orhan, A. E. and X. Pitkow (2018). "Skip Connections Eliminate Singularities." In: *arXiv:1701.09175 [cs]*. arXiv: 1701.09175 [cs].

Papagiannopoulou, C., D. G. Miralles, S. Decubber, M. Demuzere, N. E. C. Verhoest, W. A. Dorigo, and W. Waegeman (2017). "A Non-Linear Granger-causality Framework to Investigate Climate–Vegetation Dynamics." In: *Geoscientific Model Development* 10.5, pp. 1945–1960. ISSN: 1991-959X. DOI: 10.5194/gmd-10-1945-2017.

Pérez-Suay, A., J. E. Adsuara, M. Piles, L. Martínez-Ferrer, E. Díaz, A. Moreno-Martínez, and G. Camps-Valls (2020). "Interpretability of Recurrent Neural Networks in Remote Sensing." In: *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, pp. 3991–3994. DOI: 10.1109/IGARSS39084.2020.9323898.

Rasp, S., M. S. Pritchard, and P. Gentine (2018). "Deep Learning to Represent Subgrid Processes in Climate Models." In: *Proceedings of the National Academy of Sciences* 115.39, pp. 9684–9689. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1810286115.

Reddy, D. S. and P. R. C. Prasad (2018). "Prediction of Vegetation Dynamics Using NDVI Time Series Data and LSTM." In: *Modeling Earth Systems and Environment* 4.1, pp. 409–419. ISSN: 2363-6211. DOI: 10.1007/s40808-018-0431-3.

Refsgaard, J. C., J. P. van der Sluijs, J. Brown, and P. van der Keur (2006). "A Framework for Dealing with Uncertainty Due to Model Structure Error." In: *Advances in Water Resources* 29.11, pp. 1586–1597. ISSN: 0309-1708. DOI: 10.1016/j.advwatres.2005.11.013.

Reichle, R. H. (2008). "Data Assimilation Methods in the Earth Sciences." In: *Advances in Water Resources*. Hydrologic Remote Sensing 31.11, pp. 1411–1418. ISSN: 0309-1708. DOI: 10.1016/j.advwatres.2008.01.001.

Reichstein, M., M. Bahn, M. D. Mahecha, J. Kattge, and D. D. Baldocchi (2014). "Linking Plant and Ecosystem Functional Biogeography." In: *Proceedings of the National Academy of Sciences* 111.38, pp. 13697–13702. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1216065111.

Reichstein, M., S. Besnard, N. Carvalhais, F. Gans, M. Jung, B. Kraft, and M. Mahecha (2018). "Modelling Landsurface Time-Series with Recurrent Neural Nets." In: *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, pp. 7640–7643. DOI: 10.1109/IGARSS.2018.8518007.

Reichstein, M., G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and Prabhat (2019). "Deep Learning and Process Understanding for Data-Driven Earth System Science." In: *Nature* 566.7743, pp. 195–204. ISSN: 1476-4687. DOI: 10.1038/s41586-019-0912-1.

Requena-Mesa, C., M. Reichstein, M. Mahecha, B. Kraft, and J. Denzler (2018). "Predicting Landscapes as Seen from Space from Environmental Conditions." In: *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, pp. 1768–1771. DOI: 10.1109/IGARSS.2018.8519427.

Ribeiro, M. T., S. Singh, and C. Guestrin (2016). ""Why Should I Trust You?": Explaining the Predictions of Any Classifier." In: *arXiv:1602.04938 [cs, stat]*. arXiv: 1602.04938 [cs, stat].

Roberts, D. R., V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillera-Arroita, S. Hauenstein, J. J. Lahoz-Monfort, B. Schröder, W. Thuiller, D. I. Warton, B. A. Wintle, F. Hartig, and C. F. Dormann (2017). "Cross-Validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure." In: *Ecography* 40.8, pp. 913–929. ISSN: 1600-0587. DOI: 10.1111/ecog.02881.

Rolnick, D., A. Veit, S. Belongie, and N. Shavit (2018). "Deep Learning Is Robust to Massive Label Noise." In: *arXiv:1705.10694 [cs]*. arXiv: 1705.10694 [cs].

Roscher, R., B. Bohn, M. F. Duarte, and J. Garcke (2020). "Explainable Machine Learning for Scientific Insights and Discoveries." In: *IEEE Access* 8, pp. 42200–42216. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.2976199.

Rudin, C. (2019). "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." In: *Nature Machine Intelligence* 1.5, pp. 206–215. ISSN: 2522-5839. DOI: 10.1038/s42256-019-0048-x.

Rußwurm, M. and M. Körner (2017). "Temporal Vegetation Modelling Using Long Short-Term Memory Networks for Crop Identification from Medium-Resolution Multi-spectral Satellite Images." In: *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE Computer Society, pp. 1496–1504. ISBN: 978-1-5386-0733-6. DOI: 10.1109/CVPRW.2017.193.

Schellekens, J., E. Dutra, A. Martínez-de la Torre, G. Balsamo, A. van Dijk, F. Sperna Weiland, M. Minvielle, J.-C. Calvet, B. Decharme, S. Eisner, G. Fink, M. Flörke, S. Peßenteiner, R. van Beek, J. Polcher, H. Beck, R. Orth, B. Calton, S. Burke, W. Dorigo, and G. P. Weedon (2017). "A Global Water Resources Ensemble of Hydrological Models: The eartH2Observe Tier-1 Dataset." In: *Earth System Science Data* 9.2, pp. 389–413. ISSN: 1866-3508. DOI: 10.5194/essd-9-389-2017.

Schmidt, R. M., F. Schneider, and P. Hennig (2021). "Descending through a Crowded Valley - Benchmarking Deep Learning Optimizers." In: *arXiv:2007.01547 [cs, stat]*. arXiv: 2007.01547 [cs, stat].

Schneider, F., L. Balles, and P. Hennig (2019). "DeepOBS: A Deep Learning Optimizer Benchmark Suite." In: *arXiv:1903.05499 [cs, stat]*. arXiv: 1903.05499 [cs, stat].

Singh, B. P., I. Deznabi, B. Narasimhan, B. Kucharski, R. Uppaal, A. Josyula, and M. Fiterau (2019). "Multi-Resolution Networks For Flexible Irregular Time Series Modeling (Multi-FIT)." In: *arXiv:1905.00125 [cs, eess, stat]*. arXiv: 1905.00125 [cs, eess, stat].

Sitch, S., B. Smith, I. C. Prentice, A. Arneth, A. Bondeau, W. Cramer, J. O. Kaplan, S. Levis, W. Lucht, M. T. Sykes, K. Thonicke, and S. Venevsky (2003). "Evaluation of Ecosystem Dynamics, Plant Geography and Terrestrial Carbon Cycling in the LPJ Dynamic Global Vegetation Model." In: *Global Change Biology* 9.2, pp. 161–185. ISSN: 1365-2486. DOI: 10.1046/j.1365-2486.2003.00569.x.

Sood, A. and V. Smakhtin (2015). "Global Hydrological Models: A Review." In: *Hydrological Sciences Journal* 60.4, pp. 549–565. ISSN: 0262-6667. DOI: 10.1080/02626667.2014.950580.

Van Der Knijff, J. M., J. Younis, and A. P. J. D. Roo (2010). "LISFLOOD: A GIS-based Distributed Model for River Basin Scale Water Balance and Flood Simulation." In: *International Journal of Geographical Information Science* 24.2, pp. 189–212. ISSN: 1365-8816. DOI: 10.1080/13658810802549154.

Verleysen, M. and D. François (2005). "The Curse of Dimensionality in Data Mining and Time Series Prediction." In: *Computational Intelligence and Bioinspired Systems*. Ed. by J. Cabestany, A. Prieto, and F. Sandoval. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, pp. 758–770. ISBN: 978-3-540-32106-4. DOI: 10.1007/11494669_93.

Vermonte, E. (2015). *MOD09A1 MODIS/Terra Surface Reflectance 8-Day L3 Global 500m SIN Grid V006*. https://doi.org/10.5067/MODIS/MOD09A1.006.

Vig, J. (2019). "Visualizing Attention in Transformer-Based Language Representation Models." In: *arXiv:1904.02679 [cs, stat]*. arXiv: 1904.02679 [cs, stat].

Voulodimos, A., N. Doulamis, A. Doulamis, and E. Protopapadakis (2018). "Deep Learning for Computer Vision: A Brief Review." In: *Computational Intelligence and Neuroscience* 2018, e7068349. ISSN: 1687-5265. DOI: 10.1155/2018/7068349.

Walker, A. P., M. G. De Kauwe, A. Bastos, S. Belmecheri, K. Georgiou, R. F. Keeling, S. M. McMahon, B. E. Medlyn, D. J. P. Moore, R. J. Norby, S. Zaehle, K. J. Anderson-Teixeira, G. Battipaglia, R. J. W. Brienen, K. G. Cabugao, M. Cailleret, E. Campbell, J. G. Canadell, P. Ciais, M. E. Craig, D. S. Ellsworth, G. D. Farquhar, S. Fatichi, J. B. Fisher, D. C. Frank, H. Graven, L. Gu, V. Haverd, K. Heilman, M. Heimann, B. A. Hungate, C. M. Iversen, F. Joos, M. Jiang, T. F. Keenan, J. Knauer, C. Körner, V. O. Leshyk, S. Leuzinger, Y. Liu, N. MacBean, Y. Malhi, T. R. McVicar, J. Penuelas, J. Pongratz, A. S. Powell, T. Riutta, M. E. B. Sabot, J. Schleucher, S. Sitch, W. K. Smith, B. Sulman, B. Taylor, C. Terrer, M. S. Torn, K. K. Treseder, A. T. Trugman, S. E. Trumbore, P. J. van Mantgem, S. L. Voelker, M. E. Whelan, and P. A. Zuidema (2021). "Integrating the Evidence for a Terrestrial Carbon Sink Caused by Increasing Atmospheric CO2." In: *New Phytologist* 229.5, pp. 2413–2445. ISSN: 1469-8137. DOI: 10.1111/nph.16866.

Young, T., D. Hazarika, S. Poria, and E. Cambria (2018). "Recent Trends in Deep Learning Based Natural Language Processing [Review Article]." In: *IEEE Computational Intelligence Magazine* 13.3, pp. 55–75. ISSN: 1556-6048. DOI: 10.1109/MCI.2018.2840738.

Yuan, Q., H. Shen, T. Li, Z. Li, S. Li, Y. Jiang, H. Xu, W. Tan, Q. Yang, J. Wang, J. Gao, and L. Zhang (2020). "Deep Learning in Environmental Remote Sensing: Achievements and Challenges."

In: *Remote Sensing of Environment* 241, p. 111716. ISSN: 0034-4257. DOI: 10.1016/j.rse.2020.111716.

Zaremba, W., I. Sutskever, and O. Vinyals (2015). "Recurrent Neural Network Regularization." In: *arXiv:1409.2329 [cs]*. arXiv: 1409.2329 [cs].

Zhong, Y. D., B. Dey, and A. Chakraborty (2021). "Benchmarking Energy-Conserving Neural Networks for Learning Dynamics from Data." In: *arXiv:2012.02334 [cs, eess, math]*. arXiv: 2012.02334 [cs, eess, math].

Zhu, X. X., D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer (2017). "Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources." In: *IEEE Geoscience and Remote Sensing Magazine* 5.4, pp. 8–36. ISSN: 2168-6831. DOI: 10.1109/MGRS.2017.2762307.

A. License agreement Chapter 2

The book chapter in Chapter 2 is republished within this thesis under the *terms and conditions* listed below. Full reference:

> B. Kraft, S. Besnard, and S. Koirala (2021). "Emulating Ecological Memory with Recurrent Neural Networks." In: *Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science and Geosciences*. Ed. by G. Camps-Valls, D. Tuia, X. X. Zhu, and M. Reichstein. 1st edition. Hoboken, NJ: Wiley & Sons. ISBN: 978-1-119-64614-3

Terms and Conditions

1 Description of Service; Defined Terms. This Republication License enables the User to obtain licenses for republication of one or more copyrighted works as described in detail on the relevant Order Confirmation (the "Work(s)"). Copyright Clearance Center, Inc. ("CCC") grants licenses through the Service on behalf of the rightsholder identified on the Order Confirmation (the "Rightsholder"). "Republication", as used herein, generally means the inclusion of a Work, in whole or in part, in a new work or works, also as described on the Order Confirmation. "User", as used herein, means the person or entity making such republication.

2 The terms set forth in the relevant Order Confirmation, and any terms set by the Rightsholder with respect to a particular Work, govern the terms of use of Works in connection with the Service. By using the Service, the person transacting for a republication license on behalf of the User represents and warrants that he/she/it (a) has been duly authorized by the User to accept, and hereby does accept, all such terms and conditions on behalf of User, and (b) shall inform User of all such terms and conditions. In the event such person is a "freelancer" or other third party independent of User and CCC, such party shall be deemed jointly a "User" for purposes of these terms and conditions. In any event, User shall be deemed to have accepted and agreed to all such terms and conditions if User republishes the Work in any fashion.

3 Scope of License; Limitations and Obligations.

3.1 All Works and all rights therein, including copyright rights, remain the sole and exclusive property of the Rightsholder. The license created by the exchange of an Order Confirmation (and/or any invoice) and payment by User of the full amount set forth on that document includes only those rights expressly set forth in the Order Confirmation and in these terms and conditions, and conveys no other rights in the Work(s) to User. All rights not expressly granted are hereby reserved.

3.2 General Payment Terms: You may pay by credit card or through an account with us payable at the end of the month. If you and we agree that you may establish a standing account with CCC, then the following terms apply: Remit Payment to: Copyright Clearance Center, 29118 Network Place, Chicago, IL 60673-1291. Payments Due: Invoices are payable upon their delivery to you (or upon our notice to you that they are available to you for downloading). After 30 days, outstanding amounts will be subject to a service charge of 1-1/2% per month or, if less, the maximum rate allowed by applicable law. Unless otherwise specifically set forth in the Order Confirmation or in a separate written agreement signed by CCC, invoices are due and payable on "net 30" terms. While User may exercise the rights licensed immediately upon issuance of the Order Confirmation, the license is automatically revoked and is null and void, as if it had never been issued, if complete payment for the license is not received on a timely basis either from User directly or through a payment agent, such as a credit card company.

3.3 Unless otherwise provided in the Order Confirmation, any grant of rights to User (i) is "one-time" (including the editions and product family specified in the license), (ii) is non-exclusive and non-transferable and (iii) is subject to any and all limitations and restrictions (such as, but not limited to, limitations on duration of use or circulation) included in the Order Confirmation or invoice and/or in these terms and conditions. Upon completion of the licensed use, User shall either secure a new permission for further use of the Work(s) or immediately cease any new use of the Work(s) and shall render inaccessible (such as by deleting or by removing or severing links or other locators) any further copies of the Work (except for copies printed on paper in accordance with this license and still in User's stock at the end of such period).

3.4 In the event that the material for which a republication license is sought includes third party materials (such as photographs, illustrations, graphs, inserts and similar materials) which are identified in such material as having been used by permission, User is responsible for identifying, and seeking separate licenses (under this Service or otherwise) for, any of such third party materials; without a separate license, such third party materials may not be used.

3.5 Use of proper copyright notice for a Work is required as a condition of any license granted under the Service. Unless otherwise provided in the Order Confirmation, a proper copyright notice will read substantially as follows: "Republished with permission of [Rightsholder's name], from [Work's title, author, volume, edition number and year of copyright]; permission conveyed through Copyright Clearance Center, Inc. " Such notice must be provided in a reasonably legible font size and must be placed either immediately adjacent to the Work as used (for example, as part of a by-line or footnote but not as a separate electronic link) or in the place where substantially all other credits or notices for the new work containing the republished Work are located. Failure to include the required notice results in loss to the Rightsholder and CCC, and the User shall be liable to pay liquidated damages for each such failure equal to twice the use fee specified in the Order Confirmation, in addition to the use fee itself and any other fees and charges specified.

3.6 User may only make alterations to the Work if and as expressly set forth in the Order Confirmation. No Work may be used in any way that is defamatory, violates the rights of third parties (including such third parties' rights of copyright, privacy, publicity, or other tangible or intangible property), or is otherwise illegal, sexually explicit or obscene. In addition, User may not conjoin a Work with any other material that may result in damage to the reputation of the Rightsholder. User agrees to inform CCC if it becomes aware of any infringement of any rights in a Work and to cooperate with any reasonable request of CCC or the Rightsholder in connection therewith.

4 Indemnity. User hereby indemnifies and agrees to defend the Rightsholder and CCC, and their respective employees and directors, against all claims, liability, damages, costs and expenses, including legal fees and expenses, arising out of any use of a Work beyond the scope of the rights granted herein, or any use of a Work which has been altered in any unauthorized way by User, including claims of defamation or infringement of rights of copyright, publicity, privacy or other tangible or intangible property.

5 Limitation of Liability. UNDER NO CIRCUMSTANCES WILL CCC OR THE RIGHTSHOLDER BE LIABLE FOR ANY DIRECT, INDIRECT, CONSEQUENTIAL OR INCIDENTAL DAMAGES (INCLUDING WITHOUT LIMITATION DAMAGES FOR LOSS OF BUSINESS PROFITS OR INFORMATION, OR FOR BUSINESS INTERRUPTION) ARISING OUT OF THE USE OR INABILITY TO USE A WORK, EVEN IF ONE OF THEM HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. In any event, the total liability of the Rightsholder and CCC (including their respective employees and directors) shall not exceed the total amount actually paid by User for this license. User assumes full liability for the actions and omissions of its principals, employees, agents, affiliates, successors and assigns.

6 Limited Warranties. THE WORK(S) AND RIGHT(S) ARE PROVIDED "AS IS". CCC HAS THE RIGHT TO GRANT TO USER THE RIGHTS GRANTED IN THE ORDER CONFIRMATION DOCUMENT. CCC AND THE RIGHTSHOLDER DISCLAIM ALL OTHER WARRANTIES RELATING TO THE WORK(S) AND RIGHT(S), EITHER EXPRESS OR IMPLIED, INCLUDING WITHOUT LIMITATION IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. ADDITIONAL RIGHTS MAY BE REQUIRED TO USE ILLUSTRATIONS, GRAPHS, PHOTOGRAPHS, ABSTRACTS, INSERTS OR OTHER PORTIONS OF THE WORK (AS OPPOSED TO THE ENTIRE WORK) IN A MANNER CONTEMPLATED BY USER; USER UNDERSTANDS AND AGREES THAT NEITHER CCC NOR THE RIGHTSHOLDER MAY HAVE SUCH ADDITIONAL RIGHTS TO GRANT.

7 Effect of Breach. Any failure by User to pay any amount when due, or any use by User of a Work beyond the scope of the license set forth in the Order Confirmation and/or these terms and conditions, shall be a material breach of the license created by the Order Confirmation and these terms and conditions. Any breach not cured within 30 days of written notice thereof shall result in immediate termination of such license without further notice. Any unauthorized (but licensable) use of a Work that is terminated immediately upon notice thereof may be liquidated by payment of the Rightsholder's ordinary license price therefor; any unauthorized (and unlicensable) use that is not terminated immediately for any reason (including, for example, because materials containing the Work cannot reasonably be recalled) will be subject to all remedies available at law or in equity, but in no event to a payment of less than three times the Rightsholder's ordinary license price for the most closely analogous licensable use plus Rightsholder's and/or CCC's costs and expenses incurred in collecting such payment.

8 Miscellaneous.

8.1 User acknowledges that CCC may, from time to time, make changes or additions to the Service or to these terms and conditions, and CCC reserves the right to send notice to the User by electronic mail or otherwise for the purposes of notifying User of such changes or additions; provided that any such changes or additions shall not apply to permissions already secured and paid for.

8.2 Use of User-related information collected through the Service is governed by CCC's privacy policy, available online here: https://marketplace.copyright.com/rs-ui-web/mp/privacy-policy

8.3 The licensing transaction described in the Order Confirmation is personal to User. Therefore, User may not assign or transfer to any other person (whether a natural person or an organization of any kind) the license created by the Order Confirmation and these terms and conditions or any rights granted hereunder; provided, however, that User may assign such license in its entirety on written notice to CCC in the event of a transfer of all or substantially all of User's rights in the new material which includes the Work(s) licensed under this Service.

8.4 No amendment or waiver of any terms is binding unless set forth in writing and signed by the parties. The Rightsholder and CCC hereby object to any terms contained in any writing prepared by the User or its principals, employees, agents or affiliates and purporting to govern or otherwise relate to the licensing transaction described in the Order Confirmation, which terms are in any way inconsistent with any terms set forth in the Order Confirmation and/or in these terms and conditions or CCC's standard operating procedures, whether such writing is prepared prior to, simultaneously with or subsequent to the Order Confirmation, and whether such writing appears on a copy of the Order Confirmation or in a separate instrument.

8.5 The licensing transaction described in the Order Confirmation document shall be governed by and construed under the law of the State of New York, USA, without regard to the principles thereof of conflicts of law. Any case, controversy, suit, action, or proceeding arising out of, in connection with, or related to such licensing transaction shall be brought, at CCC's sole discretion, in any federal or state court located in the County of New York, State of New York, USA, or in any federal or state court whose geographical jurisdiction covers the location of the Rightsholder set forth in the Order Confirmation. The parties expressly submit to the personal jurisdiction and venue of each such federal or state court. If you have any comments or questions about the Service or Copyright Clearance Center, please contact us at 978-750-8400 or send an e-mail to support@copyright.com.